Master´s Thesis

Geography

Development Geography

MODELING THE SPATIAL DISTRIBUTION OF *CULEX* AND *STEGOMYIA* MOSQUITOES COLLECTED IN THE TAITA HILLS, KENYA IN 2016, WITH NOTES ON OTHER GENERA

Ruut Uusitalo

2017

Supervisors: Mika Siljander, Petri Pellikka, Lorna Culverwell, Kristian Forbes

UNIVERSITY OF HELSINKI

DEPARTMENT OF GEOSCIENCES AND GEOGRAPHY

DIVISION OF GEOGRAPHY

P.O. Box 64 (Gustaf Hällströmin katu 2a)

FI-00014 University of Helsinki Finland

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiedekunta/Osasto Fakultet/Sektion – Faculty	Laitos/Institution– Department	
Faculty of Science	Department of Geosciences and Geography	
Tekijä/Författare – Author		
Ruut Uusitalo		
Työn nimi / Arbetets titel – Title		
Modeling the spatial distribution of *Culex* and *Stegomyia* mosquitoes collected in the Taita Hills, Kenya in 2016, with notes on other genera		
Oppiaine /Läroämne – Subject		
Development geography		
Työn laji/Arbetets art – Level	Aika/Datum – Month and year	Sivumäärä/ Sidoantal – Number of pages
Master´s Thesis	May 2017	103 pp. + appendices

Tiivistelmä/Referat – Abstract

Mosquitoes are arguably amongst the most economically and socially important animals on the planet due to their ability to act as vectors for pathogens, including parasites and viruses, from animals to humans, or between humans. Mosquito-borne diseases (MBDs), are contracted following infection by one or more mosquito borne viruses (MBVs) or parasites, including dengue virus (DENV), chikungunya virus (CHIKV), Zika virus (ZIKV), West Nile virus (WNV), yellow fever virus (YFV) and malaria, and annually cause more than one million human deaths (WHO 2016). MBDs are contracted after an infected mosquito transfers one or more pathogens in the course of blood feeding from one host to another. Three important genera which act as vectors for many pathogens are *Anopheles*, *Culex* and *Stegomyia* and they are most problematic in the tropical and subtropical regions of Asia, South America and Africa (WHO 2016).

Among vector-borne diseases (VBDs), MBDs have the strongest dependence on environmental factors. These factors have either direct or indirect impact on mosquito presence and abundance as mosquitoes are dependent on habitat suitability (Franklin & Miller 2010; Rasheed *et al*. 2013). This study will utilize species distribution modeling (SDM) to investigate the relationship between environmental, anthropogenic and distance factors on the occurrence of mosquito species. It forms part of an ongoing *Wildlife screening project*, led by Prof. Olli Vapalahti, which aims to screen mosquitoes, rodents and bats for new and known viruses in Kenya. The absence of previous studies of the geographical distribution and habitat suitability patterns of mosquito species over the Taita Hills region in southeastern Kenya, justifies the need for this research.

This project has three main objectives: 1) to investigate which mosquito genera are distributed in the Taita Hills, and how they are distributed, 2) to examine which factors best explain the presence of *Culex* and *Stegomyia* mosquitoes, 3) to test whether any of the available statistical regression models can reliably estimate the distribution of *Culex* and *Stegomyia* mosquitoes, and to build predictive maps for estimations created by the most reliable models.

Biological, Geographic Information Systems (GIS) and statistical methods were combined in the study. Data consists of occurrence, environmental, anthropogenic, distance and biological data. The specimens were collected from 122 locations from January–March 2016 throughout the Taita Hills. Environmental, anthropogenic and distance data were acquired from the satellite and aerial imagery and produced in ArcMap. The biomod2 package, intended for ensemble forecasting of species distributions in R, was used to generate models.

After multicollinearity of the environmental, anthropogenic and distance factors was pruned, the best estimating predictor variables were selected. The factors that best estimated the distribution of *Culex* were slope, human population density, NDVI, distance to roads and elevation. This resulted in six reliable models with accurate estimation values. Multivariate adaptive regression splines (MARS) resulted area under the curve (AUC)- value of 0.806, and a traditional Generalized linear model(GLM) brought an AUC- value of 0.730 with high statistical significance rates, both above the value for a good model fit (AUC $\geq$0.7); thus ensuring a reliable estimation.

Five environmental, anthropogenic and distance factors best estimated the distribution of *Stegomyia*: mean radiation in January–March, human population density, NDVI, distance to roads and mean temperature in January–March. By these predictors, biomod2 resulted in highest AUC- values for generalized boosted model (hereafter GBM) and random forest (RF) with AUC- value of 0.708 for each. Hence, reliable estimations resulted for both *Culex* and *Stegomyia*, which are visualized by the probability of presence maps in the Results chapter. The results may be used as a guide for public health officials in the Taita region regarding the distribution, favorable habitats and prevention strategies of *Culex* and *Stegomyia* mosquitoes, which are capable of transmitting mosquito-borne infections.

Avainsanat – Nyckelord – Keywords

species distribution modeling, mosquito-borne diseases, biomod2, predictive maps, mosquitoes, ecological statistical modelling

Säilytyspaikka – Förvaringställe – Where deposited

University of Helsinki, Kumpula Science Library

Muita tietoja – Övriga uppgifter – Additional information

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiedekunta/Osasto Fakultet/Sektion – Faculty	Laitos/Institution– Department	
Matemaattis-luonnontieteellinen tiedekunta	Geotieteiden ja maantieteen laitos	
Tekijä/Författare – Author Ruut Uusitalo		
Työn nimi / Arbetets titel – Title Modeling the spatial distribution of *Culex* and *Stegomyia* mosquitoes collected in the Taita Hills, Kenya in 2016, with notes on other genera		
Oppiaine /Läroämne – Subject Maantiede		
Työn laji/Arbetets art – Level Pro gradu	Aika/Datum – Month and year Toukokuu 2017	Sivumäärä/ Sidoantal – Number of pages 103 + liitteet

Tiivistelmä/Referat – Abstract

Hyttyset ovat yksi taloudellisesti ja sosiaalisesti merkittävimmistä eläinlajeista planeetallamme, sillä ne kykenevät välittämään taudinaiheuttajia, kuten loisia tai viruksia, eläimistä ihmisiin ja ihmisistä toisiin. Hyttysten levittämät taudit syntyvät yhden tai useamman hyttysen levittämän viruksen tai loisen aiheuttamana tartuntana. Tällaisia tartuntatauteja ovat dengue virus (DENV), chikungunya virus (CHIKV), Zika virus (ZIKV), malaria, Länsi-Niilin virus ja keltakuume, jotka ovat aiheuttaneet vuosittain yli miljoona kuolemaa maailmanlaajuisesti (WHO 2016). Hyttysten levittämät sairaudet syntyvät, kun tartunnan saanut hyttynen siirtää yhden tai useamman taudinaiheuttajan isännästä toiseen veren imemisen aikana. Kolme hyttyssukua; *Anopheles*, *Culex* ja *Stegomyia (Aedes)*, toimivat merkittävimpinä taudinaiheuttajien välittäjinä synnyttäen ongelmallisimman tilanteen erityisesti Aasian, Etelä-Amerikan ja Afrikan trooppisilla ja subtrooppisilla alueilla (WHO 2016).

Vektorien välittämistä taudeista, hyttysten levittämät taudit ovat läheisimmin yhteydessä ihmistoimintaan liittyviin tekijöihin sekä ympäristötekijöihin. Ympäristötekijöillä on joko suora tai epäsuora vaikutus hyttysten esiintymiseen, sillä hyttyset ovat riippuvaisia suotuisasta elinympäristöstä (Franklin & Miller 2010; Rasheed *et al.* 2013). Tämä tutkimus hyödyntää lajilevinneisyysmallinnusta hyttyshavaintojen, ympäristömuuttujien ja ihmistoimintaan liittyvien muuttujien välisten suhteiden tarkastelussa. Tämä tutkimus on osa prof. Olli Vapalahden luotsaamaa *Villieläinten seulonta*-projektia, jonka tavoitteena on löytää uusia lajeja ja etsiä mahdollisia viruksia jyrsijöistä, lepakoista ja hyttysistä Keniassa. Hyttyslajien maantieteelliseen levinneisyyteen ja elinympäristöyhteyksiin liittyvien aiempien tutkimusten puuttuminen vahvistaa tarvetta lisätutkimukselle Taita Hillsin alueella Kaakkois-Keniassa.

Tutkimuksella on kolme päätavoitetta: 1) tutkia, mitä hyttyssukuja Taita Hillsin alueella esiintyy, ja miten kerättyjen hyttyssukujen levinneisyys sijoittuu alueellisesti 2) tarkastella, mitkä tekijät selittävät parhaiten *Culex* ja *Stegomyia* hyttysten levinneisyyttä, 3) antaa vastaus hypoteesiin; voiko jokin tilastollinen malli ennustaa uskottavasti *Culex* ja *Stegomyia* hyttysten levinneisyyttä. Mahdollisten luotettavien mallien avulla on lisäksi tarkoitus ennustaa hyttyslajien levinneisyyttä ennustekartoin.

Tässä tutkimuksessa yhdistettiin biologisia, tilastollisia, ja paikkatietojärjestelmiin perustuvia tutkimusmetodeita. Tutkimusaineisto sisältää havaintoaineiston, ympäristöaineiston, ihmistoimintaan ja etäisyyksiin perustuvan aineiston sekä biologisen aineiston. Näytteitä kerättiin yhteensä 122 sijainnista Taita Hillsin alueella tammi-maaliskuussa 2016. Ympäristöaineisto sekä ihmistoimintaan ja etäisyyksiin perustuvat aineistot saatiin satelliitti- ja ilmakuvista, ja ne tuotettiin ja muokattiin ArcMap- ohjelmassa. Analyysissä käytettiin biomod2-ohjelmapakettia, joka on lajilevinneisyyden ennustamiseen tarkoitettu alusta R-ohjelmointiympäristössä.

Selittävien muuttujien eli ennustemuuttujien korrelaatioiden testauksen jälkeen parhaiten ennustavat muuttujat valittiin lopulliseen malliin. Parhaiten *Culex*in levinneisyyttä ennustavia tekijöitä olivat rinnekaltevuus, asukastiheys, NDVI, etäisyys tiehen sekä korkeus. Tämä tuotti 6 luotettavaa ennustemallia korkeilla ennustearvoilla. Multivariate adaptive regression splines (MARS) tuotti AUC(Area under curve)-arvon 0.806, ja perinteinen yleistetty lineaarinen malli(GLM) tuotti AUC-arvon 0.730 tilastollisesti merkitsevillä arvoilla. Kumpikin malli sai hyvän mallin sovittamisen ylittävän AUC-arvon (AUC $\geq$0.7), ja tuotti näin luotettavan ennusteen *Culex* ja *Stegomyia* hyttysten lajilevinneisyydelle.

Stegomyia- hyttysten levinneisyyttä ennusti parhaiten viisi ennustemuuttujaa mukaan lukien keskisäteily, asukastiheys, NDVI, etäisyys tiehen sekä keskilämpötila. Näillä muuttujilla, korkeimmat AUC-arvot tuotti yleistetty luokittelupuumenetelmä (GBM) ja satumetsä(RF), AUC-arvoilla 0.708. Kummallekin hyttyssuvulle, *Culex*ille ja *Stegomyia*lle syntyi luotettavia levinneisyysennusteita, jotka esitetään todennäköisyyskarttoina Results-osiossa. Tutkimuksen tuloksia voidaan hyödyntää terveysviranomaisten ohjenuorana hyttysperäisiä tauteja levittävien *Culex* ja *Stegomyia* hyttysten suotuisten elinympäristöjen kartoittamisessa, sekä niiden esiintymiseen ja tautien ehkäisyyn liittyvien strategioiden tukena Taita Hillsin alueella.

Avainsanat – Nyckelord – Keywords

Lajilevinneisyysmallinnus, SDM, hyttysten levittämät taudit, biomod2, ennustekartat, hyttyset, ekologinen tilastollinen mallinnus

Säilytyspaikka – Förvaringställe – Where deposited

Helsingin yliopisto, Kumpulan tiedekirjasto

Muita tietoja – Övriga uppgifter – Additional information

Contents

List of Figures

List of tables

Abbreviations

AIC	Akaike information criterion
An.	Anopheles
ANN	Artificial neural networks
AUC	Area under the curve
CDC	Centers for Disease Control and Prevention
CHIKV	Chikungunya virus
CTA	Classification tree analysis
Cx	Culex
DEM	Digital elevation model
DENV	Dengue virus
GAM	Generalized additive model
GARP	Hybrid methods
GBM	Generalized boosted model
GLM	Generalized linear model
GIS	Geographic information systems
GPS	Global positioning systems
Lt	Lutzia
MARS	Multiple adaptive regression splines
Maxent	Maximum entropy modeling
MBD	Mosquito-borne disease
MODIS	Moderate Resolution Imaging Spectroradiometer
MTI	Mosquito Taxonomic Inventory
NDVI	Normalized difference vegetation index
RF	Random forest
RMSE	Root mean squared error

ROC	Receiver operating characteristic
SAC	Spatial autocorrelation
SDM	Species distribution modeling
SPOT	Satellite Pour l'Observation de la Terre
St.	Stegomyia
ZIKV	Zika virus
YFV	Yellow Fever virus
VBD	Vector-borne disease
WHO	World Health Organization
WNV	West Nile virus

1. Introduction

Mosquitos (Diptera: Culicidae) are important vectors for numerous potentially deadly diseases, which cause millions of deaths each year. The most deadly of these are chikungunya, West Nile virus, malaria, yellow fever, Zika virus and dengue virus of which malaria alone caused 438 000 deaths worldwide in 2015 (WHO 2016a). Additionally, the incidence of dengue has risen 30-fold in the past three decades (WHO 2016c). It is estimated that 390 million dengue cases occur each year, of which 96 million are present with clinical symptoms (Bhatt *et al.* 2013).

The Afrotropical region is at high risk of many mosquito-borne diseases and Sub-Saharan Africa carries high share of the global malaria burden with 88% of total cases and 90% of deaths (WHO 2016a). Recently, new outbreaks of mosquito-borne diseases have been reported in many African countries; in Angola, Uganda and the Democratic Republic of the Congo new Yellow fever outbreaks were reported in 2016 (WHO 2016b). In addition to this, Africa is the second most affected continent for Dengue fever, with 16% of global dengue infections, thus, Southeastern African countries including Kenya, have been identified as risk areas for dengue infections (CDC 2012).

The first laboratory-confirmed dengue outbreak occurred in Kenya in the early 1980s (CDC 2012). An outbreak of non–malarial illness was reported in 2013, it was estimated that 13 % of participants were actually infected with dengue virus but were diagnosed with malaria (Sharp 2015). In Kenya, as elsewhere in Africa, dengue is commonly misdiagnosed as malaria based on the lack of laboratory- based diagnostic testing (Attaway *et al.* 2014). In 2016, the Ministry of Health of Kenya alerted the WHO of an outbreak of chikungunya virus in Mandera East in northern Kenya (WHO 2016b). These events call for improved medical tests and diagnoses, and also the geographical association of disease vectors in Kenya.

One way to estimate disease risk, is by utilizing the species distribution modeling (hereafter SDM) for vector species (Franklin & Miller 2010). Among vector-borne diseases (hereafter VBDs), mosquito-borne diseases (hereafter MBDs) have the strongest correlation with environmental factors (Rasheed *et al.* 2013). Habitat suitability has both a direct and indirect impact on mosquito presence and abundance (Franklin & Miller 2010). Environmental and

anthropogenic disturbances such as climate change, urbanization, deforestation and pollution represent crucial factors in the emergence of both mosquito species and MBDs. Studying mosquitoes through SDM has many advantages. When the species whose potential habitat is predicted are of medical or veterinary importance, the results of SDM may serve public health goals and support epidemiological studies (Franklin & Miller 2010). Public health officials in the Taita Hills may find the species- environment concentration useful when planning interventions, to indicate the areas where the virus vectors are likely to be found, or where they will potentially spread to in the near future.

This study is a part of ongoing multidisciplinary *Wildlife screening*- project led by Professor Olli Vapalahti from the Department of Virology at the University of Helsinki, and implemented by researchers with biological and virological backgrounds. The purpose of the project is to investigate the prevalence and distribution of Dengue virus in mosquito and human hosts in the Taita Hills in southeastern Kenya. The thesis has an interdisciplinary background linked to public health concern, ecological gradient analysis and biogeography in addition to GIS. Data collections were implemented in January-March in 2016 resulting in a sample size of approximately 4000 mosquitoes. Specimens were analyzed at the University of Helsinki. As part of the project, my thesis study focuses on estimating the distribution of virus vectors *Culex* and *Stegomyia*.

Several studies have used SDM when modeling the distribution of mosquito species. Recent SDM mosquito studies have mostly used Maximum entropy model (Maxent) as a model algorithm (Mughini-Gras *et al.* 2014; Sallam *et al.* 2016; Fatima *et al.* 2016; Larson *et al.* 2010; Conley *et al.* 2014). Additionally, in a few mosquito studies, the random forest model has been used to estimate the distributions of mosquito species (Ibañez-Justicia *et al.* 2015; Kwon *et al.* 2015). In this study, *presence-absence data* of genus observations was used in the biomod2 package, an ensemble platform for SDM in R statistical computing software (Thuiller *et al.* 2016). Biomod2 has been used in several studies, including modeling eastern mosquitofish (*Gambusia affinis*) distributions and distributions of invasive freshwater bivalve species (Murphy *et al.* 2015; Gama *et al.* 2016). Biomod2 has not been previously used in modeling mosquito distributions, therefore this study may demonstrate additional applications in this research area.

2. Research objectives

The primary aim of this study is to determine which mosquito genera are present over the Taita Hills in southeastern Kenya. The purpose is to estimate the distribution of *Culex* and *Stegomyia* mosquitoes and to find environmental, anthropogenic and distance factors, which are linked to their distribution. Statistical and GIS tools enable a more detailed assessment between these factors and mosquito presences. The main questions posed by this study are:

1) Which mosquito genera are present in the Taita Hills and how are they distributed?

2) Which environmental, anthropogenic and distance factors have the strongest influence on the presence of *Culex* and *Stegomyia* mosquitoes?

3) Can any of the statistical models reliably estimate the distribution of *Culex* and *Stegomyia* mosquitoes? H0= Models cannot estimate the distribution of *Culex* and *Stegomyia* accurately. H1=Models are able to reliably estimate *Culex* and *Stegomyia* distributions.

The first objective of the study is to investigate which mosquito genera were present in the Taita Hills and how are their presence and absence distributed. The second objective is to investigate which environmental, anthropogenic and distance factors have the strongest effect on the distribution of *Culex* and *Stegomyia* mosquitoes. Based on these factors, the statistical models will be created. The third objective of the study is to approve or reject the null hypothesis; whether any of the statistical models run by using biomod2 are valid to estimate *Culex* and *Stegomyia* distributions. The models that result in the highest prediction accuracy values with statistical significance rates are selected for best assessing the presence of *Culex* and *Stegomyia*. The third objective aims to estimate favorable habitats where *Culex* and *Stegomyia* mosquitoes may potentially be spreading to and to present them by predictive maps. Predictive distribution maps are produced for specific areas by a method of interpolation, consisting of areas where mosquitoes haven´t been collected.

3. Theoretical background

This chapter gives more accurate perspective of mosquitoes, role as vectors of viruses and parasites and their habitat characteristics. Due to the scientific diversity of this study, themes from biology, ecology, GIS and virology are combined. This chapter goes through the reasons why it is important to study mosquitoes and their distributions and introduces the key concepts of the study.

3.1 Mosquitoes

At time of writing, 3554 species of mosquitoes are recognized, worldwide (MTI 2017). They are distinguished from the other true flies (those which have one pair of wings) by their specially adapted mouthpart, the proboscis, which enables the females of many species to feed on blood and all individuals to feed on plant saps (AMCA 2014). Only female mosquitoes are able to suck blood, therefore, the females are in focus when discovering and examining MBDs (AMCA 2014). The life cycle, the existence as a virus vector, the taxonomy and the habitat suitability of mosquitoes all play a significant role when understanding their activity and effect on their habitats and on humans.

The mosquito life cycle is broken down into four stages. The first stage is as an egg, which hatches into a larva, which then goes through four larval instars (Becker *et al.* 2010). The larval stage is followed by pupal stage and final transformation occurs from a pupa into an adult mosquito (Becker *et al.* 2010). Adult mosquitoes then lay eggs, which continue the life cycle. The larval stage is significant as well as the adult stage, since the habitats that larvae are adapted to develop within, or conditions that they are adapted to tolerate, directly impact the possible distribution of mosquito species.

Each species has its own habitat or so-called niche, which consists of a unique, n-dimensional array of environmental tolerances and resource needs (Drew *et al.* 2011). The suitable area for mosquito species habitat varies and change based on many factors. In addition to study the potential mosquito habitats, it is imperative to identify the main characters, activity and life cycle of mosquitoes. A minimum of seven mosquito genera were collected in the Taita region during the fieldtrip, but only *Culex* and *Stegomyia* are modelled

by spatial distribution. In the following, the main characteristics and activities of these genera are introduced.

3.1.1 *Culex* and *Stegomyia*

Culex (Linnaeus 1758) is a large and important genus of mosquitoes including 770 species divided into 26 subgenera, which tend to hibernate over the cold months and to breed during the summer months (Figure 1; MTI 2017; Mosquito World 2017b). They are unicolorous mosquitoes but some species have markings on the legs and pale spots on the wings (MTI 2017). Eggs are laid as rafts on the surface of standing water such as in ground water but they also can be found in leaf axils, tree-holes, rock-holes and crab-holes (MTI 2017). After *Culex* mosquitoes hatch, they stay close and do not travel more than a few hundreds of meters from the location (Mosquito World 2017b).

Figure 1. *Culex* (1.) and *Stegomyia* (2.) females. *Culex* mosquitoes are frequently unicolorous, brown species, while *Stegomyia* are dark mosquitoes with white scale patches.

Adult *Culex* mosquitoes mainly bite at night and are aggressive biters, feeding on humans and animals, including birds and reptiles (MTI 2017). Female *Culex* mosquitoes need the protein obtained from blood meals in order to develop eggs (Mosquito World 2017b). *Culex* mosquitoes are amongst the most ubiquitous species on the planet, occurring in the tropics to cool temperate regions in all zoogeographical regions excluding the extreme northern latitudes (MTI 2017).

Mosquito taxonomy is a contested subject, with two schools of thought as to the placement of many genera that are frequently referred to as "*Aedes*". This thesis follows the classification of Reinert *et al*. (2009), which considers *Stegomyia* Theobald 1901 as a separate genus from *Aedes* Meigen 1818, rather than as a subgenus, which has been put forward several times for convenience. Thus *Stegomyia* is a moderately sized genus with 128 species divided among 8 subgenera (MTI 2017). Immature *Stegomyia* hatch in small collections of water including rock and tree holes, bamboo internodes, leaf axils and, artificial containers (MTI 2017). Mating of *Stegomyia* species occurs in the immediate vicinity of the larval habitats and adults usually fly an average of 400 m, which indicates that people are moving the virus within and between communities (MTI 2017; WHO 2017b). They are mainly diurnal mosquitoes, thus are more active biters during the day rather than at night, and will feed on a number of species, including humans.

Stegomyia albopicta is originally a forest species but has become adapted to rural, suburban and urban human environments (WHO 2017c). It has spread from Asia to Africa, the Americas, and Europe, mainly aided by the international trade in used car tyres, where eggs are deposited when containing rainwater (WHO 2017c). *Stegomyia* may occur also at high elevations from the tropics to the arctic worldwide (MTI 2017).

3.1.2 *Culex* and *Stegomyia* as virus vectors

Many species belonging to *Culex* and *Stegomyia* are well-known disease vectors (WHO 2017). Species of *Culex* transmit West Nile virus, Japanese encephalitis, bird malaria and filarial worms, amongst other things (WHO 2017c). The most prevalent species is *Culex pipiens*, which is also the main carrier of West Nile virus (Mosquito World 2017b). West Nile Virus is found worldwide, excluding the northern latitudes in the continents of North America, West Asia, The Middle East, Europe and Africa (WHO 2017c).

Species of *Stegomyia* genus are capable of transmitting yellow fever, dengue and helminths causing Brugian and Bancroftian filariasis (MTI 2017). A mosquito acquires the virus or parasite when feeding on the blood of an infected person (WHO 2017b). The virus infects the mosquito mid-gut spreading to the salivary glands over an incubation period of 8- 12 days, after which the virus can be transmitted to humans during feeding (WHO 2017b). *Stegomyia*

aegypti is a vector of urban yellow fever and dengue fever viruses and is widely distributed through the tropical and subtropical regions worldwide (MTI 2017). Other important vectors of *Stegomyia* are *St. africanus* and *St. bromeliae* which transmit yellow fever virus in Africa, and also *St. albopictus* which transmits dengue virus in the Americas, Europe and Africa (MTI 2017). In addition to *Culex* and *Stegomyia,* several other mosquito genera were collected in the Taita Hills, which are briefly introduced in the next chapter.

3.1.3 Other mosquito genera collected in Taita

More than five other mosquito genera were collected from the study area including *Anopheles, Aedimorphus*, *Uranotaenia*, *Eretmapodites* and *Lutzia*. Issues with the available mosquito keys meant that a lot of species belonging to the tribe *Aedini* were not identified in time for inclusion herein. Each genus has its own specific characteristics, activities and connections to mosquito-borne infections.

Anopheles includes a total of 475 species, of which 30–40 species transmit malaria (MTI 2017; CDC 2012). Most *Anopheles* are crepuscular, being active at dawn and dusk, or nocturnal, being active at night (CDC 2012). They prefer temperate, subtropical or tropical areas at elevations from coastal areas to mountain terrains, and are distributed worldwide excluding the majority of the Pacific Islands and Antarctica (MTI 2017; CDC 2012). *Anopheles* mosquitoes are the sole vectors of human malarial parasites in addition to which they are also vectors of microfilariae and encephalitis viruses (MTI 2017). The primary malaria vectors in Africa, *Anopheles gambiae* and *An. funestus*, are two of the most efficient malaria vectors in the world (CDC 2012).

Aedimorphus Theobald 1903 consists of 66 species which inhabit temporary and semi-permanent fresh-water ground pools as well as swamps, artificial containers, wells, puddles, rock holes and animal footprints (Figure 2; MTI 2017). Some female *Aedimorphus* prefer feeding on humans or on non-human hosts either during the daytime or at night (MTI 2017). Many *Aedimorphus* species transmit pathogens causing diseases such as Japanese encephalitis virus, Rift Valley fever or West Nile virus in humans and animals (MTI 2017). Most *Aedimorphus* species occur in the Afrotropical Region, but they are also found in Australasian and Oriental Regions, Central America and the Papuan area (MTI 2017).

Figure 2. Five genera in addition to *Culex* to *Stegomyia* were confirmed from Taita. 1. *Anopheles*. 2. *Aedimorphus*. 3. *Uranotaenia*. 4. *Eretmapodites*. *Lutzia tigripes* is absent in the photos.

Uranotaenia Lynch Arribálzaga 1891 includes 270 species divided into two subgenera (Peyton 1972). They are small, delicate mosquitoes whose larvae live in a range of habitats such as ground waters, rock holes, leaf axils or artificial containers (MTI 2017). Female *Uranotaenia* rarely feed on humans but has connections to potential transmission of pathogens that cause *Flaviviruses* including West Nile virus (WNV) and nounané virus (NOUV) (Pachler *et al.* 2014; Junglen *et al.* 2009). Many *Uranotaenia* species are attracted to light and rarely rest inside houses (MTI 2017). They occur in the Afrotropical and Oriental Regions, Australia, Europe and the Middle East and in the Americas (MTI 2017).

Eretmapodites Theobald 1901 include 48 species, and they are mainly forest mosquitoes, but they may occur in banana plantations (MTI 2017). The larvae inhabit water contained in snail shells, containers, tree-holes and in bamboo stumps (MTI 2017). *Eretmapodites* can attack humans even if they prefer other hosts, and they are able to transmit pathogens causing

diseases such as Rift Valley fever virus or yellow fever virus (MTI 2017). *Eretmapodites* are distributed only in the Afrotropical Region (MTI 2017).

Lutzia Theobald 1903 is a genus of large mosquitoes including eight species (MTI 2017). The larvae are mainly found in ground-water habitats, but also elsewhere, such as in tree holes and in artificial containers (MTI 2017). *Lutzia* females feed upon animals and occasionally humans, but there is no evidence of connection to transmission of the pathogens (MTI 2017). Members of *Lutzia* occur in the Neotropical, Australasian and Afrotropical Regions and in Japan (MTI 2017). Only one species of *Lutzia*, *Lt. tigripes*, is known from the Afrotropical region, thus all mention of *Lutzia* in this study refers to this species.

In addition to these seven genera, 300 mosquitoes were collected from the Taita Hills, but have not yet been identified to genus. Thus, they are removed from the further modeling process. As land-use changes worldwide, mosquitoes are forced to adapt to their new environment. Understanding preferred habitats for mosquitoes gives us more ideas on how the species react to changes in environment. In the next chapter, we discuss this phenomenon from a global point of view.

3.1.4 Changing habitats

All mosquitoes need an aquatic habitat which they can use as a breeding site (Figure 3). Larvae mainly inhabit stagnant water bodies such as tanks, water deposits and containers, flower pots, swimming pools, pet bowls, drains, leaf axils, tree holes and gutters due to the high nutritional level of water (MTI 2017; WHO 2017c) The rise in human population densities and mean temperature has led to increasing global changes in land-use, and suitable breeding sites for the mosquitoes are becoming more common. This phenomenon mainly occurs in the developing world, but can also be observed in developed countries in Europe and in the Americas.

In Pakistan, one of the major reasons for the distribution of mosquitoes into new areas is proved to be urbanization and the developing infrastructure such as new road networks (Fatima *et al.* 2016). When a country is developing, the mobility of people and goods via roads and other infrastructure increases; this opens up a possibility for mosquitoes to spread

into new areas. In Kenya, dengue transmission is facilitated by the humid and rainy climate on the coast and by expanding road and rail transportation networks, which are favorable opportunities for mosquitoes (Attaway *et al.* 2014). The same phenomenon can also be observed in Europe. In The Netherlands, mosquito breeding sites are shifting even more and more from forest environments to unforested environments (Ibañez-Justicia *et al.* 2015). This fact brings new challenges for monitoring and controlling mosquito species distributions and thus, mosquito-borne infections in Europe.

Figure 3 A. Large numbers of *Culex* larvae and adults, were collected from open water environments. The pond was located at an altitude of 1700 m close to Yale, the second highest top in the Taita Hills. B. A tree hole situated in the Ngangao montane forest. C. *Stegomyia aegypti* larvae were found in old car tyres filled with water.

In order to understand which factors are playing significant role in the distribution of different mosquito species to new areas, we can model the spatial distributions of species of which we have biological and occurrence data. In the following sections, the focus is on describing the characteristics of the SDM process and the environmental, anthropogenic and distance factors on which the distribution of mosquitoes are dependent.

3.2 Methodological framework

Species distribution models are great tools when studying suitable species habitats, as they discover connections between environmental factors and the distribution of a plant or an animal species (Franklin & Miller 2010). SDM is closely linked to habitat selection theory, which suggests that animals behave ideally and that they can identify the best quality habitats which are available (Drew *et al.* 2011). The theory expects that as population size increases in the habitat, the quality of habitat decreases, as low-quality areas are more likely to be used as population density increases (Drew *et al.* 2011). Thus, it is assumed that temporal variation in habitat use will be further in low- quality habitats which are only occupied at high population densities (Drew *et al.* 2011).

SDM uses statistical associations, which are seen as empirical models optimizing precision and reality (Kienast *et al.* 2012). They are based on observed data which can be affected by biotic interactions, disturbances and dispersal limitations (Franklin & Miller 2010). Based on the statistically built model, SDM extrapolates species distribution data both in space and time (Franklin & Miller 2010). This enables prediction of the spatial distributions of species without any biological data because environmental factors are spatially auto-correlated (Drew *et al.* 2011).

The principle of SDM is to construct a response variable and a set of predictor variables in such a way that the relations between variables are exposed (Guisan *et al.* 2006). By applying ecological techniques to SDMs, vector-borne diseases can be modeled as the insect vectors transmitting diseases such as dengue virus are strongly influenced by environmental conditions (Drew *et al.* 2011). When constructing a species distribution model, it is important to pay attention to 1) the selection of either *presence-only* or *presence-absence* data, 2) the modeling process steps, 3) the selection of a suitable model algorithm and 4) the selection of appropriate environmental data (Drew *et al.* 2011).

3.2.1 The species distribution modeling process

In Figure 4, the process of SDM is performed stepwise by Franklin & Miller (2010). When building a predictive distribution model, the first stages are to assume a conceptual model of the expected species- environment relationships (Guisan *et al.* 2006). Later on, model fitting

and validation are conducted by assessment of response functions and model parameters (Guisan *et al.* 2006).

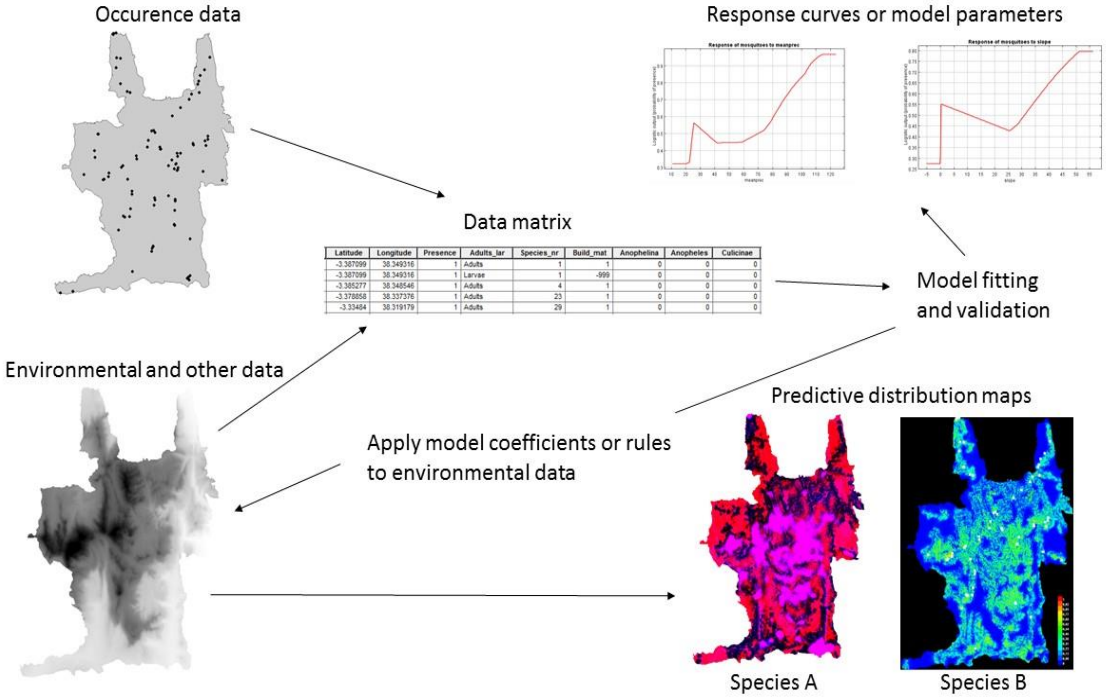


Figure 4. The SDM process is introduced stepwise (Franklin & Miller 2010). Occurrence, environmental, anthropogenic and distance data in each collection site are combined to the same data matrix. Afterwards, highly correlated explanatory variables are removed and models are run, resulting in response curves and maps for estimating the species distributions.

The processing steps include data compilation and model creation, calibration and validation. Both occurrence data and environmental data from each collection point are compiled to the same dataset (Franklin & Miller 2010). Occurrence data consists of the observations regarding the mosquito species locations. Satellite and aerial images are used to derive the explanatory variables including mean precipitation in January–March, land cover, mean temperature in January–March, mean radiation in January–March, mean relative humidity in January–March, building design, human population density, distance to houses and distance to roads, normalized difference vegetation index (NDVI), slope, and elevation variables.

After occurrence, environmental, anthropogenic and distance data have been aggregated, model analysis methods are selected and response functions or model parameters are validated in R. Some of modeling methods can merely be applied to binary (presence-absence) response variables, including generalized linear model (GLM) and generalized

additive model (GAM) (Franklin & Miller 2010). On the contrary, maximum entropy modeling (Maxent) is represented as a model algorithm using presence-only data (Smith 2012). Afterwards, model parameters are validated and model coefficients and rules are applied to environmental data. By predictive maps produced by R- statistical computing software, the models can be visualized and the hypothesis is tested.

3.2.2 Factors affecting the distribution of mosquitoes

Several factors control the distribution of mosquito species. Based on results from a regression study (Austin 1971), indirect and direct factors are either proximal (causal), having more importance as predictors under more optimal conditions for the species; or distal (surrogate), being important close to the limits of a species distribution (Austin 2002). Distal factors are resources of regulators which are correlated with species distributions and thus are easier to observe than proximal factors (Austin 2002).

Environmental gradients that affect the occurrence of species includes direct, indirect and resource factors (Austin 1980). Direct factors are those gradients that have a direct physiological effect on species presence such as temperature and precipitation (Austin 1980). In contrast, an indirect factor has no direct effect on species occurrence; being a distal variable such as elevation, it affects species occurrence as a result of location-specific correlations (Franklin & Miller 2010; Austin 1980).

The variation of insect populations may be detected on a small scales due to their small size as well as their sensitivity to environmental gradients. As environment and population size change, species distribution may vary over time. Environmental gradients such as wind and water flow or the presence of unsuitable patches can constrain the direction of mosquito population spread (Schowalter 2011). The directions in which mosquito populations can spread are restricted by gradients in moisture, in chemical concentrations and in temperature (Schowalter 2011).

This study focuses on discovering connections between species distribution of mosquitoes and human population density, building design, distance to houses, distance to roads, land

cover, precipitation, elevation, temperature, relative humidity, slope, NDVI and radiation. The importance of these factors is due to their direct or indirect impact on mosquito survival, reproduction and development, which have a great influence on mosquito presence and abundance (CDC 2012).

3.2.3 Possible advantages in the use of SDM

There are two main purposes for producing predictive maps. At first, maps which estimate potential species distributions are used to interpolate or to fill in the gaps in geographical information about where a species occurs, for those species whose distribution is in equilibrium with environmental conditions (Franklin & Miller 2010). Furthermore, the maps are used to extrapolate or estimate suitable environmental conditions compatible with survival in the areas where the species could be present, based on known locations in the present or the past (Franklin & Miller 2010).

Additionally, a considerable advantage of SDM is its ability to interpolate or extrapolate from point observations over space and forecast the occurrence of a species for locations which lack survey data (Franklin & Miller 2010). The relationship between a species and its abiotic and biotic environment based on observations can be realized using SDM (Franklin & Miller 2010). In this study, modeling mosquito species distributions is implemented by the biomod2 package of R, which aims at ensemble forecasting of species distributions where methodological uncertainties in models and species-environment relationships are taken into account (Thuiller *et al.* 2016).

4. Study area

This chapter aims to introduce the study area in the geographic and climatic context. Kenya is located on the equator in eastern Africa between Somalia and Tanzania, and has 45.9 million residents (CIA 2016). The Kenyan climate varies from tropical on the coast to arid in the interior (CIA 2016). There are two annual wet and dry seasons, and the total rainfall varies greatly across the country (Kaplan *et al.*1976: vii).

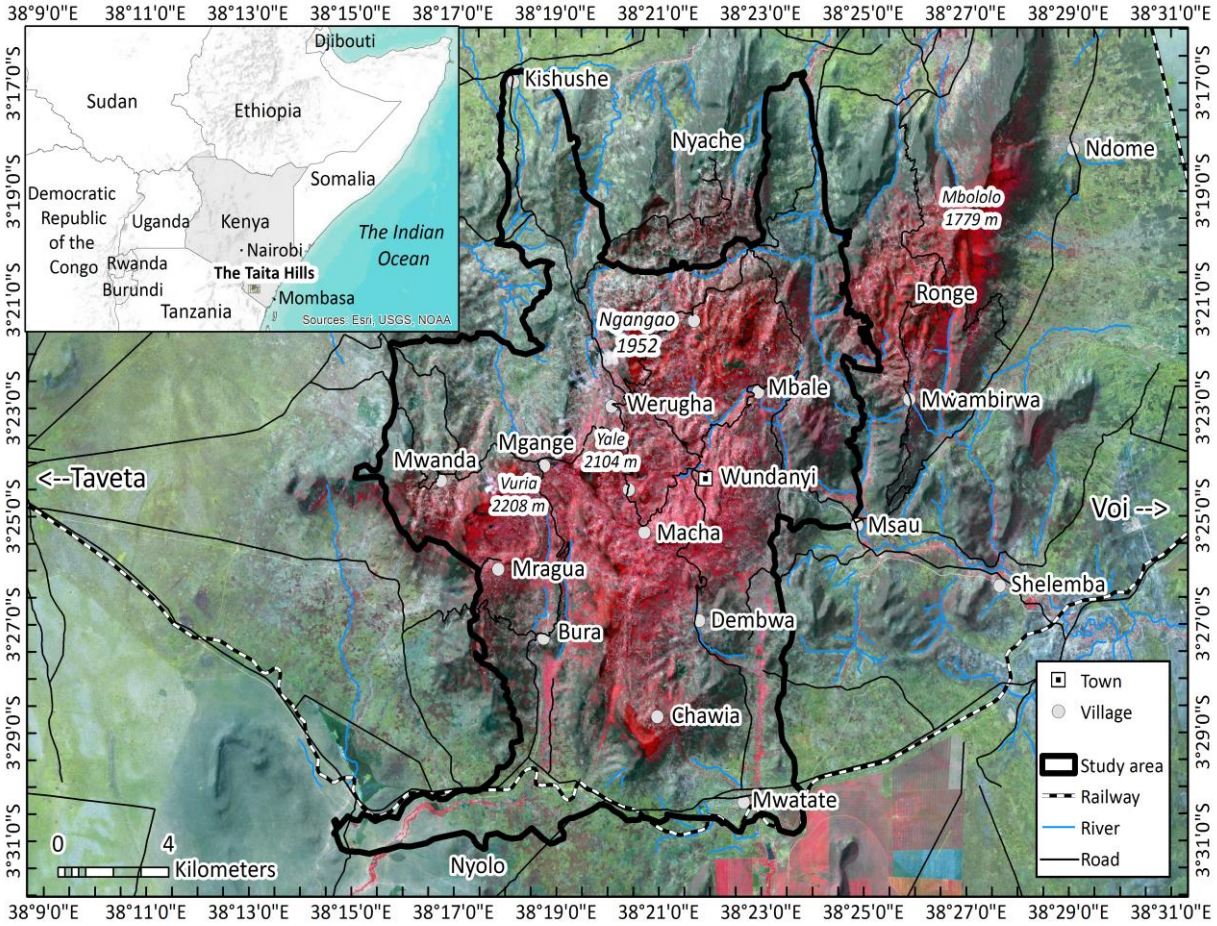


Figure 5. The Taita Hills is located 150 kilometers from Mombasa and is surrounded by the plains in southeastern Kenya.

The Taita Hills (03°20'S, 38°15'E), is a part of the Taita–Taveta county in southeastern Kenya (Figure 5). This district covers an area of 1000 square kilometers. The study area is 286 km² and covers mainly mountainous parts of the region, but also part of the surrounding plains. The Taita Hills region is located about 150 kilometers inland from the Indian Ocean, therefore is part of the Eastern Arc Mountains which begin in eastern Tanzania and end in southeastern Kenya (Salminen 2004).

The altitude of the Taita Hills varies from 600–900 meters up to 2200 meters above sea level, constituting the first barrier for the moist winds blowing from the Indian Ocean (Salminen 2004). Thus, the Taita Hills receive a high levels of rainfall (on average 1330–1910 mm/a), which create many forming a suitable habitats for mosquito species to take advantage of, particularly on the southern and eastern where the most of the rain falls (Salminen 2004). The

Hills are located in the inter-tropical convergence zone, with long rains occurring from March–May and short rains in November and -December (Maeda 2011).

The demographic patterns have changed during recent decades and the population across the entyre Taita–Taveta county is approximately 300 000 people (Msagha 2004). The main source of livelihood in the area is intensive agriculture, which is simultaneously a major reason for deforestation (Soini 2005). Approximately half of the montane forests in the Taita Hills have been converted to agricultural land between 1955 and 2004 (Pellikka *et al.* 2009).

The Taita Hills is isolated from the surrounding mountainous areas by the Tsavo plains at elevation of 700 meters (Salminen 2004). Dengue and Zika virus vectors *Stegomyia aegypti* and *St. albopicta* are often thought to be a problem merely on the plains and on the coastal areas in Kenya (Attaway *et al.* 2014). This study justifies the argument that these virus vectors have spread westward from coastal Kenya to the Taita Hills, which is located in the middle of developing infrastructure between new railway and road constructions. Development of these infrastructures may, open up favorable opportunities for mosquitoes to expand their distributions.

5. Materials

Constructing SDM data requires several steps; defining the study area, selecting the relevant sampling method, and selecting and editing occurrence and environmental data were all crucial parts of creating the models. The quality and accuracy of each step is important and affects the results. This chapter concentrates on considering the materials and data used and produced in the modeling process including study area, sampling method, occurrence data and environmental, anthropogenic and distance data.

5.1 Sampling design

Spatial sampling design is closely linked to the occurrence data, as the purpose is to evaluate the population mean of a variable such as abundance or biomass (Franklin & Miller 2010). The sampling evaluates the extent of a variable in space or its value at unmeasured locations

in spatial interpolation process, estimating the value of an unsampled location as an average of its neighbors weighted by their inverse distance from the new location (Franklin & Miller 2010). In this study, the measurement scale of the sampling was categorical and continuous, as the amount of species was calculated for each location.

The sampling method of this study has a systematic background. Systematic samples produce better data than random samples for evaluating SDMs regarding their species-environment relationships (Franklin & Miller 2010). In addition, it is emphasized that a reliable sampling method results in a more accurate SDM production for the specific research goals (Barnhart *et al.* 2014). The collections consisted of six transects of roads that all lead to Werugha village at an elevation of 1700m. The sampling sites were chosen within an altitude range of 100 meters from one another. This sampling method enables more reliable results in spatial interpolation, as systematic samples minimize the distance from any point in the study area to a sample point (Franklin & Miller 2010). Additionally, the collection locations with absence points were registered into occurrence data. Other mosquito data collected during the first two weeks of February 2016 (without a systematic sampling method), was added to the occurrence data, as the collection locations included the study area.

5.2 Occurrence data

The occurrence data proposes either detection or non-detection of a species throughout a study region (Drew *et al.* 2011). Occurrence of so-called biological data is measured at nominal (e.g. presence/absence), ordinal (e.g. ranked abundance) and ratio (e.g. abundance, richness) levels, and that affects the types of model algorithms to use and the measurement level of the SDM product e.g. probability or suitability of occurrence (Franklin & Miller 2010). In this study, nominal level was used, as both *presence* and *absence* data were recorded.

5.2.1 Mosquito collections

At each of the collection locations the procedure was identical; locations of the mosquito collections were saved as GPS points (Garmin Map 64S) and the basic habitat information

was filled manually into a pre-made collection sheet at every collection site. This provided georeferenced data for the mosquito species and their habitats.

The characteristics of collections sites and equipment used for the collections varied depending on whether adults or larvae were collected in the locations. Both adult and immature mosquito life stages were collected from each location, where possible. Adults were collected by using commercially available prokopacks (manufactured by The John W. Hock Company), which consist of an aspirator attached to a 12 V battery, and collection cups which trap the mosquitoes that have been aspirated into them during use. Collection cups were monitored by eye so that they held a minimum of around 10 specimens, to a maximum number that meant the specimen quality was not compromised by overcrowding in the cups. Once full, a lid was placed over the collection cup and then it was transferred into a cool box lined with ice blocks in order to prevent excess stress, and therefore mosquitoes from dying or knocking into each other, before arriving at the laboratory. Other equipment used for adult collections were light traps, which collected mosquitoes overnight (The John W. Hock Company). Adults were mainly collected from inside houses; around toilets, bathrooms and in the septic tanks, and also resting on vegetation, including banana plants.

Figure 6 A. *Stegomyia aegypti* larvae which have been mounted on a microscope slide to aid their identification. Note that the last few segments of each specimen have been removed in order to best view the necessary characters from the side, and mounted away from the rest of the specimen. B. Larvae were reared to adults in zip lock bags filled with water from the collection habitats.

Immature life stages (Figure 6) were collected with a dipper and aquarium net and sorted using a cut off pipette and a water bowl. Larvae were stored in zip lock bags, which facilitated

transport to the research station. They were frequently found in water tanks and septic tanks containing dark water rich in nutrients. Larvae were also collected from the forests and plantation fields from vegetation such as three-holes and banana plants. Large numbers of larvae were recovered from small stagnant ponds or in discarded tyres filled with rain–water. Moreover, ditches located in the shade were popular locations for larvae. In contrast, direct sunshine proved to be unsuitable for larvae, so few collections were made in open pools. Both larvae and adults were collected in each location, with larvae being more rarely collected than adult mosquitoes. Therefore, fewer larval collections were made in comparison to adult collections. After field collections, the next step was to process the specimens in the laboratory.

Mosquitoes were collected over the course of seven weeks in different locations in the Taita Hills during the dry season. During the first collection period, two weeks in February 2016 when collections were made with others from the research group, the aim was to collect competent dengue and Zika vectors, namely *Stegomyia aegypti* and other members of the genus. During this collection period, mosquitoes were collected from the forests and banana plantations, and from the villages and city centres. While collection–methods remained constant throughout the project, the collection locations during this time were chosen at random rather than along a specified transect.

During the second collection period in February and -March 2016, the sampling method was developed to ensure compatibility with model development. Suitable habitats for a variety of species, including some mosquitoes, are typically roadsides, as they are attracted by human activity and air pollution caused by vehicles (Schowalter 2011). Therefore, mosquitoes were collected mainly from dwellings next to roads. We drove to the location and walked from house to house as far as mosquito adults or larvae were observed.

Collection locations were determined systematically, resulting in 122 collection sites including both larvae and adult collections (Figure 7). Excluding the northeastern part of the Taita Hills, the mountain area was well represented in the aggregation of mosquito collection locations. The final sample size consisted of 3130 mosquitoes including 73 presence locations of *Culex*, 28 of *Stegomyia*, nine of *Lutzia*, four of *Anopheles*, 2 two of *Aedimorphus*, two of

Eretmapodites and one of *Uranotaenia.* In some studies, it is proposed that 30–100 locations of species presence are needed for achieving acceptable efficiency in species distribution models (Franklin & Miller 2010). Thus, only *Culex* and *Stegomyia* genera are used for modelling in this study.

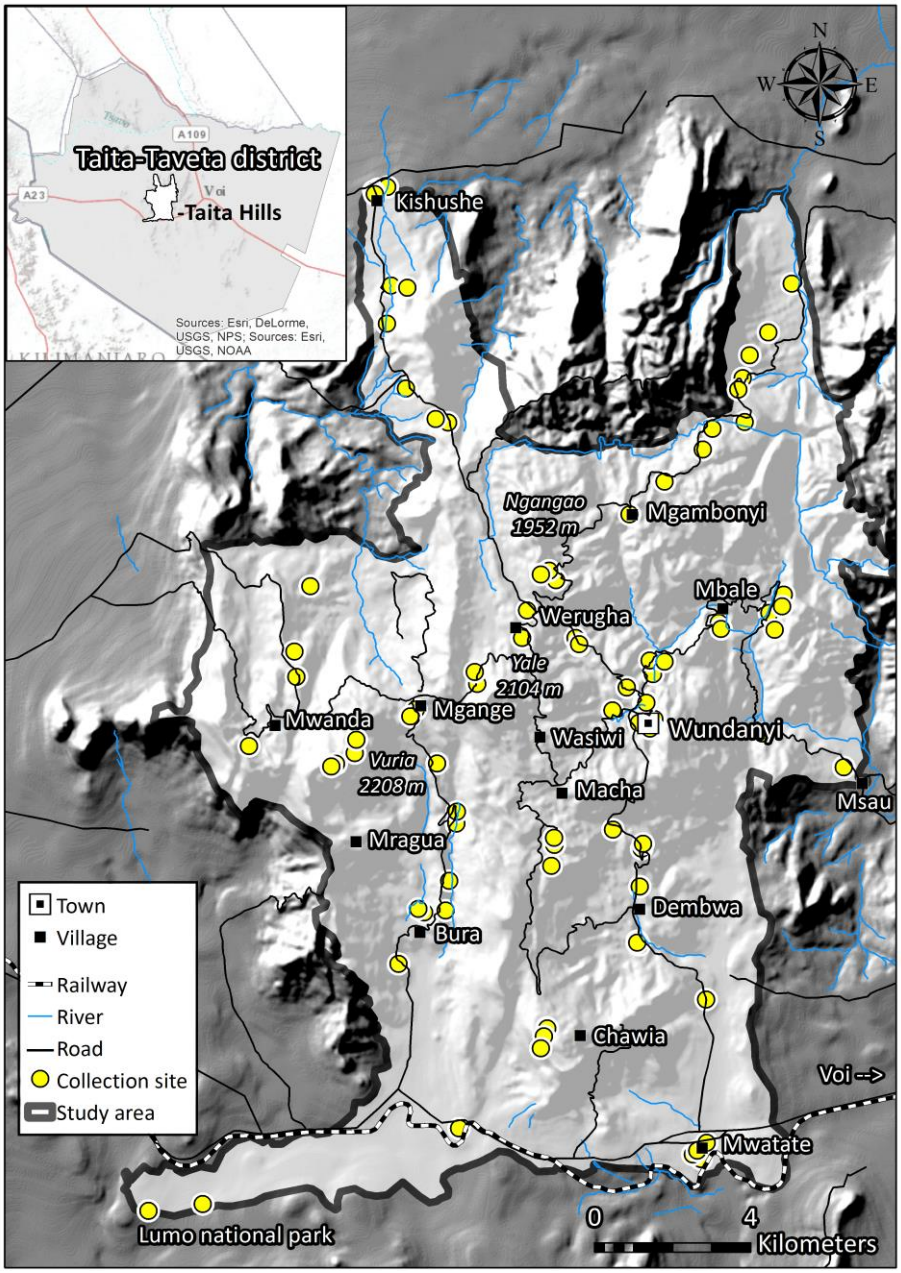


Figure 7. Occurrence data includes 122 collection points over the Taita Hills mainly following the sampling method where collection sites were located within an altitude range of 100 meters from one another.

5.2.2 Species processing and identification

After collecting the mosquitoes, it was essential to process mosquitoes in the proper way, using specific steps. The early sorting of mosquitoes was performed by PhD- student Lorna

Culverwell from the Department of Virology. A myriad of subtle differences in the shape, coloring and ratio of morphological structures were crucial when differentiating species. At first, the different species were sorted from each other by using a low-powered stereo dissecting microscope (Figure 8).While some structures were easily visible, others required a more powerful microscope to see clearly, so some specimens had to be left for later identification. So far as possible the specimens were sorted by genera, subgenera and species on return to the University of Helsinki in Finland, but some remain to be identified.

Following collection, adults were sorted into tubes of 80% ethanol, for later DNA and morphological examination, and then stored first at -20ºC and subsequently at -70 ºC before being transported to Finland for species identifications. Immature collections were split so that some of them were killed by dropping them into sub-boiling water, and transferring them into 80% ethanol, while others were link-reared from a larva through to adult, keeping the exuviae (skin) that remained following the transition from larva to pupa for each specimen and then again from pupa to adult. These exuviae and whole adults were then slide mounted, to facilitate adult identification. Identifications were made following available keys for the region (Service 1991; Huang 2000; Huang 2004).

Adult mosquitoes were processed individually, while the larvae were reared to adults in the laboratory in the research center. Adult mosquitoes were frozen for 20 minutes before further processing using a stereo microscope. Mosquitoes were stored either in ethanol (80% or 100% (Abs)) or RNA-*later* to preserve the RNA or DNA, as necessary. Absolute ethanol was used for mosquitoes that were intended to be used for DNA study. RNA-*later* was used for the mosquitoes intended for virus isolation (Figure 8).

Females that had recently had a blood meal were of particular interest due to the heightened possibility that they would be carrying one or more parasite or virus, which will be screened for at a later point. Most adult female mosquitoes were processed using one of two methods, depending on their intended purpose: 1. Mosquitoes for virus detection were placed in individual 0.5 ml tubes containing RNA-*later* and crushed with a toothpick. Most of the blood-fed and gravid females were intended for virus detection due to the increased significance in the boarder study of virus detection, than species identification. 2. Other

mosquitoes were placed into a tube containing 100% of ethanol in order to preserve DNA. Neither method is perfect for morphological examination at a later point, since storing mosquitoes in any reagent increases the chance that scales present on the surface of the body are rubbed off, which complicates later identification.

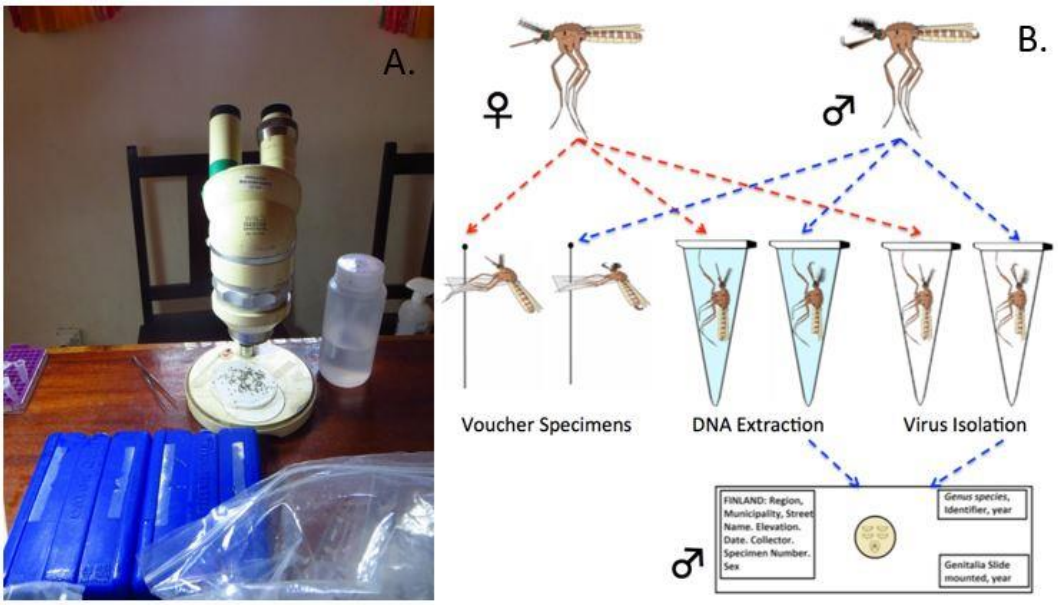


Figure 8. A. A microscope was vital to sort mosquitoes by their appearance and by sex. B. The procedure of specimen processing. Mosquitoes were initially sorted by sex and appearance, and subsequently they were placed into either a tube of RNA-later with the purpose of virus isolation, into a tube of ethanol with the purpose of DNA extraction (Culverwell, pers.comm), and a small number were mounted onto cardboard points and pinned for morphological examination. Males of some species also had genitalia removed and dissected onto a microscope slide, to facilitate species identification.

In collections with high numbers of mosquitoes of the same species, first 10 males and females were placed in individual tubes containing ethanol in order to be identified and analyzed later and acted as vouchers for the rest of that sex/species in the collection. Thereafter, pools of 20 males and females (but not mixed sexes), of the same species were crushed in 1.5 ml eppendorf tubes containing RNA-*later* and then stored at -20ºC. Each tube was labeled and recorded on the relevant collection form. The specimens in RNA- *later* required a continuous freezing chain, so were stored at -20ºC initially, during transportation to University of Nairobi, specimens were stored in a cool box with a freezer block, and then stored at -70 C° before transportation to Finland, where they were also stored at -70 C° before further processing.

The specimens in ethanol were identified using the earlier stated keys, and the specimens in RNA-*later* were further pooled, had RNA extracted and were screened by polymerase chain

reaction (PCR) for viruses in spring 2017 (Service 1991; Huang 2000; Huang 2004). Specimens were identified by Lorna Culverwell during autumn 2016 and spring 2017. Mosquitoes can be identified to various taxonomic levels, the main ones being Family, Subfamily, Tribe, Genus, Subgenus and Species, from the broader to the more specific levels, respectively (MTI 2017). The aim was to identify the specimens as accurately as possible; however, for modeling purposes, genus level identifications were used due to time restraints for more specific identifications. The identification keys that are available for this region are out of date and do not include all species, so the extra work required for species level identification was not possible before the end of the deadline period. Identifications from the mosquitoes from the Taita Hills revealed seven named genera, with several others visible but as yet not identified. *Culex* and *Stegomyia* will be used in the modeling process, which will be introduced next.

5.3 Environmental, anthropogenic and distance data

In this chapter, both environmental and other variables related to modeling are explained more precisely. Explanatory variables including environmental, anthropogenic and distance data were obtained either straight from the satellite and aerial imagery or by creating them from the satellite imagery in ArcMap (Version 10.3.1). There were 12 environmental and other variables explicating the connections between variables. All 12 variables obtained from different data sources are represented in Table 1. The selection of these variables to predict mosquito distributions was based on the literature or selected by observation in the field such as building design.

As mosquito collections were implemented during January–March 2016, we used the environmental data of the same time of year if available. Environmental variables included mean precipitation (January–March), mean relative humidity (January–March), land cover and normalized difference vegetation index (NDVI), mean temperature (January–March), mean radiation (January–March), digital elevation model (DEM), and slope. Other variables consisted of human population density, distance to houses, distance to roads, and building material of the houses, all of which provide important background information for the species–habitat relationship. The data were derived from either satellite or aerial imagery and the values for each collection point were received by using the Extract Multi-Values to

Points- tool in ArcMap. This way, the data matrix was filled with the values of environmental and other data for each collection point.

Table 1. The range of values in explanatory variables in the collection sites, the data source and description.

Environmental, anthropogenic or distance factor	Min	Max	Avg	Description of data
Distance to houses (m)	0	1270	52	Mean Euclidean distance to houses derived by ArcMap from the building data (Heiskanen *et al.* 2015).
Elevation (m)	694	2079	1330	Mean elevation was obtained from DEM. DEM was acquired from a SPOT satellite image for the year 2003, which utilized a 20 metre planimetric resolution (Clark *et al.* 2005).
Distance to roads (m)	0	927	127	Mean Euclidean distance to roads. Road data obtained from University of Helsinki (Broberg *et al.* 2004).
Mean precipitation (mm)	20	113	47	Mean precipitation was obtained from long-term mean precipitation grids, which were interpolated from monthly available meteorological data and surrounding areas using ANUSPLIN software (Hutchinson 1995; Erdogan *et al.* 2011).
Mean radiation (kWh/m²)	176	228	216	Irradiance, solar radiation energy received on a given surface area in a given time (kWh/m²), was calculated from the DEM using an ArcInfo AML macro (shortwavc.aml) (Esri Inc., Redlands, CA, USA) (Kumar *et al.* 1997, Zimmermann 2000).
Mean relative humidity (%)	71	94	77	Mean monthly relative humidity derived from Virtanen (2015).
Mean temperature (C°)	15	25	21	Mean monthly temperature in derived from Virtanen (2015).
Slope (°)	0	43	11	Mean slope degree derived from DEM.
Human population density (persons/km²)	0	9090	1260	Derived from Heiskanen *et al.* (2015) and modified by digitizing more houses in the study area in QGIS. Used 6 habitants per house as Msagha, J. (2004) in Pellikka *et al.* (2004).
NDVI	-0.4	0.2	-0.2	NDVI derived from Sentinel-2A satellite imagery (2016).
Land cover	-	-	-	Derived from the SPOT 4 image from October 2011 and identified by Heikinheimo (2015).
Building design	-	-	-	Obtained by digitizing the houses of mosquito collection sites in QGIS and by classifying houses by observation at field and by Google Earth.

5.3.1 Mean precipitation, mean temperature and mean relative humidity

Precipitation, temperature and relative humidity all affect the spread and reproduction of mosquitoes. Mosquito larval habitats always require water and moisture, which is why a large number of mosquitoes are found in locations with high precipitation or humidity, if the other required habitat conditions are suitable (Mosquito World 2017a). For instance, *Stegomyia aegypti* responded well to variations in rainfall, presenting high variability in precipitation–facilitated species spread (Fatima *et al.* 2016). In this study, precipitation was obtained from long-term mean precipitation grids, which were interpolated from monthly available meteorological data and surrounding areas using ANUSPLIN software (Hutchinson 1995; Erdogan *et al.* 2011). Temperature and relative humidity data were obtained from Virtanen (2015). Mean precipitation, mean temperature and mean relative humidity for each collection point from January–March was calculated by using the Field Calculator- tool in ArcMap and by inserting values into the data matrix with the Extract Multi-Values to Points- tool.

Data collection was scheduled for the period immediately after the rainy season, thus not the period of highest precipitation. Average precipitation at all collection points during January to -March was 46.72 mm per month (Figure 9). Additionally, temperature is seen to be an important variable in the reproduction of mosquitoes, with daytime temperature range affecting mosquito presence and cold winter temperatures affecting mosquito mortality (Fatima *et al.* 2016).

Regarding the longevity and the standing variety of temperatures for *Stegomyia aegypti* and *St. albopicta* species, *St. albopicta* has a greater longevity than *St. aegypti* (Fatima *et al.* 2016). Nevertheless, it was estimated that *St. aegypti* tolerated a wider range of temperatures; the optimal survival temperature of *St. aegypti* was 21°C, while *St. albopicta* was seen to have much wider range for optimal survival, between 20–30°C (Brady *et al.* 2013). Among the collection locations in the Taita Hills, the average temperature from January–March was 21°C, which also justifies the suitable habitat of these two virus vectors (Brady *et al.* 2013). The Taita Hills is a mountainous area where humid wind raises the relative humidity. The average relative humidity in the mosquito collection locations in January–March was 77%, which is proportionally high.

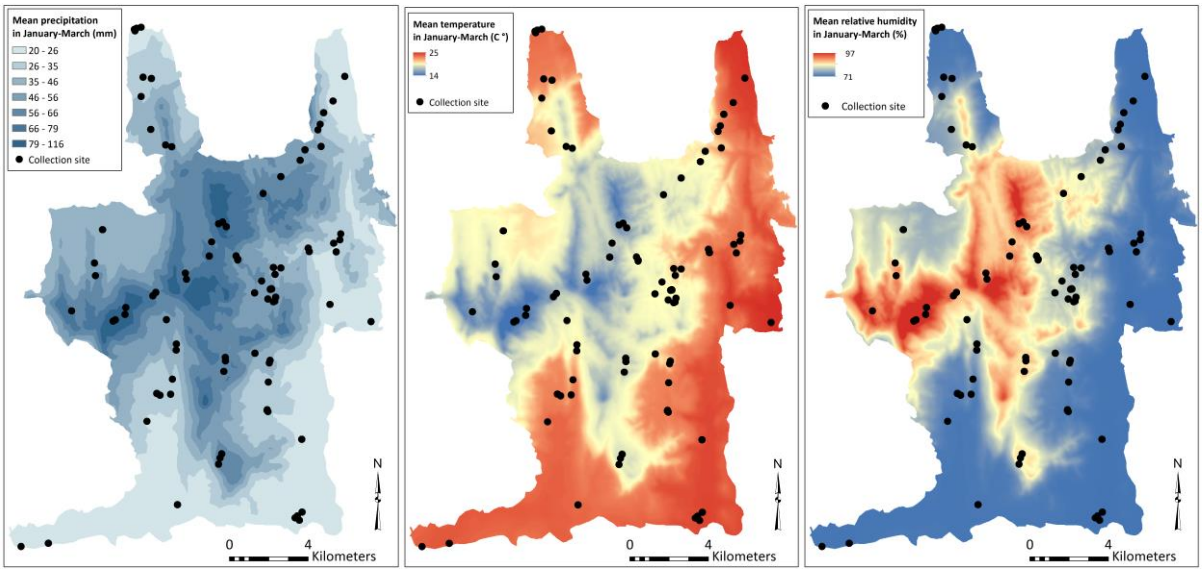


Figure 9. A. Mean precipitation (mm) in January–March in the study area varies between 20–120 mm per month. The mountain areas receive the highest precipitation rates. B. Mean temperature varies from 14 C° up to 25 C°. Temperature is highest on the plateau and lowest in the mountain areas. C. Mean relative humidity ranges between 71 and 97 per cent being greatest at high altitudes.

5.3.2 Land cover and normalized difference vegetation index (NDVI)

Land cover and vegetation are assessed to be important factors influencing the occurrence of mosquitoes. If land-use changes occur due to the urbanization and deforestation, it may positively affect the mosquito species richness. According to Fatima *et al.* (2016), long-term modifications to landscape and rising temperatures have affected the dengue vector, *Stegomyia aegypti´*s range shift. The primary focus of discovering land cover and NDVI data in this study was to concentrate on finding water surfaces and locations with rich vegetation that create suitable conditions for mosquito habitats.

Land cover data were derived from the SPOT satellite image from October 2011 and identified by Heikinheimo (2015). Land cover is an important factor in explaining species–habitat relationship. Ten different land cover types exist in Taita including cropland, scrubland, thicket, woodland, plantation and broadleaved forest, bare soil and built up areas, as well as bare rock and water areas. The most general land cover type in the collection locations was cropland, which is characterized by cultivated terrestrial area (Figure 10).

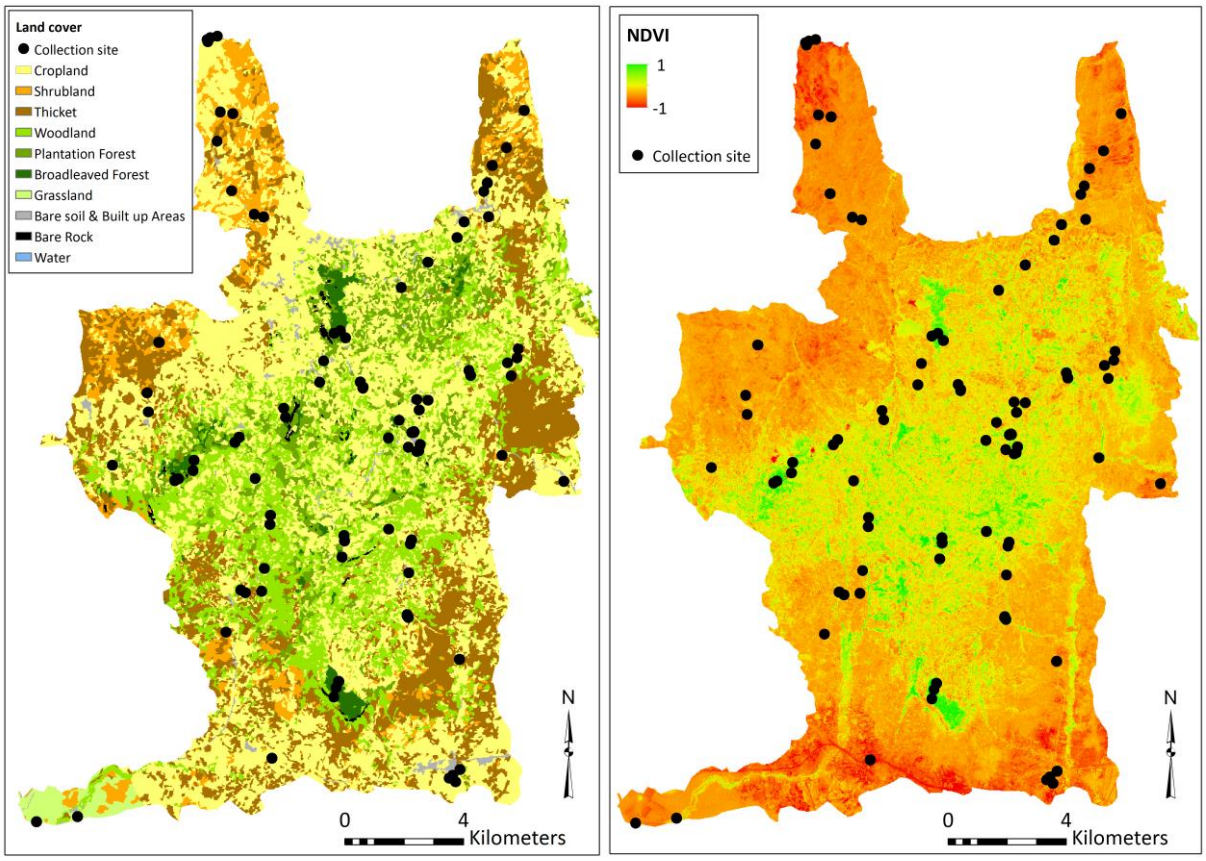


Figure 10. A. The most common land cover types in Taita are cropland and woodland. Broadleaved forest occurs in the biodiversity hotspots. B. NDVI values range between -1 and 1. Greenest area in the north is in Ngangao broadleaved forest and greenest area in the south is Chawia broadleaved forest. The areas with poor vegetation are located on the plateau.

As land cover variable is a classification-graded variable and models do not always accept a classification-graded variable being a part of the model, NDVI was instead used in the modeling. NDVI was derived from Sentinel-2A satellite imagery (2016) by using Image Analysis Tool in ArcMap and the NDVI values of each collection point were inserted to the data matrix (Tucker 1979). Mean NDVI value of all collection points was -0.2 (Figure 10). NDVI is a measure of plant greenness and the values range between -1 and 1. As value 1 is an indicator of maximum amount of green plants and -1 is an indicator of minimum amount of greenness, we can conclude that most collection points were not covered by high amounts of green plants (Tucker 1979). The collection locations with the highest NDVI values were clearly visible in the data being Ngangao and Chawia forests, which are identified as the hotspots of the biodiversity in Taita.

5.3.3 Elevation, slope and mean radiation

Topographical aspects and solar radiation are indirect factors affecting mosquito distributions. Behind these indirect factors are other factors such as precipitation, temperature and wind, which are the main influencing factors in mosquito presence and absence. In some studies, mosquito species richness has been highest on plateaus but decreased towards foothills, low-montane and mid-range montane habitats, while increasing again in montane areas at high elevations (Eisen *et al.* 2008). Thus, it has been justified that some species reach their "cool end" in low montane and mid- range montane habitats and, on the contrary, some mosquito species are cold-adapted (Eisen *et al.* 2008). In this study, mosquito collections were conducted within an altitude range of 670 m up to 2208 m (the top of Vuria). Similar findings were observed as some mosquito species were not collected above 1000 meters. Contrarily, some mosquito species which were collected above 1500 meters were not found at lower altitudes.

Topographical factors, elevation and slope, were used as explanatory factors in the species distribution model. Elevation and slope angles were obtained from a DEM dataset, which consists of an ordered array of grids representing distribution of elevations in a landscape (Franklin & Miller 2010). DEM was acquired from a SPOT satellite image for the year 2003, which utilized a 20 meter planimetric resolution (Clark *et al.* 2005). DEM was obtained from scanned Survey of Kenya 1:50 000 scale topographic map from which the contours lines were digitized and Topo to Raster function in ArcGIS 10.3.1 (Clark *et al.* 2005). Many topographical variables are derived from DEM, such as; slope, slope aspect, specific catchment area, slope curvature and hillslope position (Franklin & Miller 2010). Mean elevation was obtained from DEM dataset. Elevation was calculated for each collection point from DEM by the Extract Multi-Values to Points-tool in ArcMap. Average elevation of all collection points was 1328 meters, which indicates that the area is highly mountainous (Figure 11). Slope was derived from DEM by the Slope-tool in ArcMap, and the value of slope angle at each collection point was added to the data matrix. Average slope angle of all collection points was 11 degrees (Figure 11).

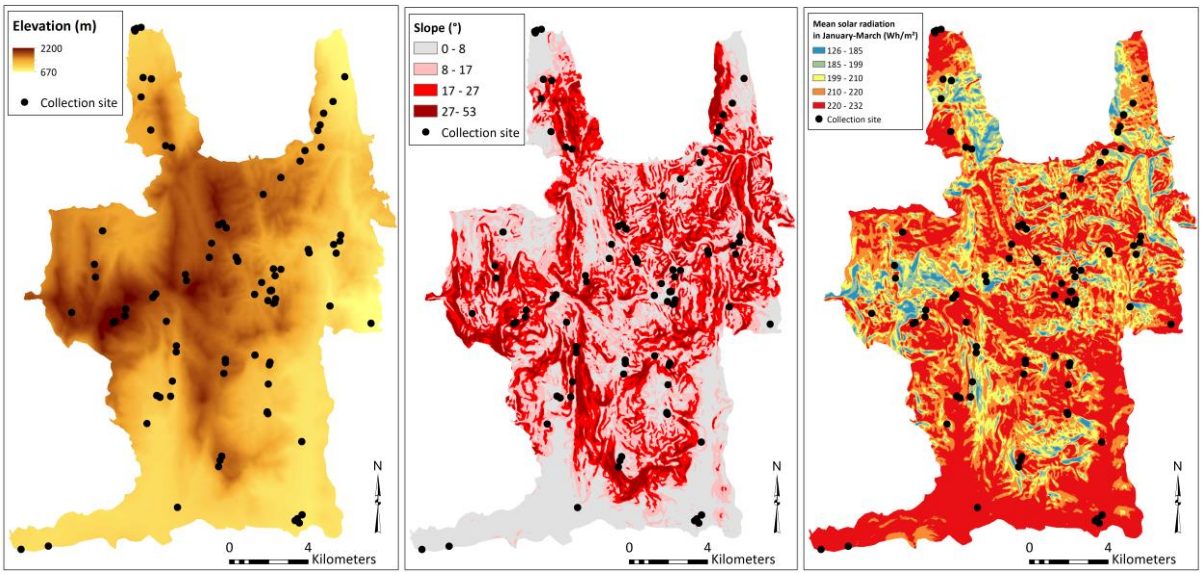


Figure 11.A. In Taita, the altitude ranges between 670 meters and 2200 meters. The surrounding area is characterized by plateau. B. The mountain areas can be recognized by the values of slope angles. C. Solar radiation values are highest on the plateau and lowest in the forestall areas.

In additional to topographical variables, radiation influence is essential to the mosquito habitats. Mean radiation data were obtained from shortwave calculation program (SHORTWAVC.AML) (Kumar *et al.* 1997). Mean radiation for January–March was calculated by the Field Calculator from the satellite imagery and values were added to the mosquito data matrix in the same way as elevation and slope variables. The average mean radiation in January–March was 216 kWh/m² (Figure 11). Furthermore, non-environmental drivers also influence mosquito presence and abundance.

In addition to environmental variables, there exist anthropogenic and distance factors which influence the habitat suitability of mosquitoes. Anthropogenic drivers affect mosquito abundance when mosquitoes use humans and animals as their primary hosts, positioning human and animal health at risk for mosquito-borne diseases. Anthropogenic and distance data consists of variables of human population density, distance to houses, distance to roads, and building design, which will be introduced next.

5.3.4 Human population density and distance variables

It is established that areas with high population density, carry a high dengue risk; nevertheless, many high-density population regions may have a low risk for dengue (Attaway

et al. 2014). Partly based on this interest, we took anthropogenic and other factors into the model process, including variables of human population density, distance to houses, distance to roads and building material of houses. Through genetic markers, it was justified that humans have formed the evolutionary distribution and history of *Stegomyia aegypti* across time and space (Brown *et al.* 2014). This study aims at presenting the influence of human population density (persons/km²) and distance from homes and roads (m) on the observation of specific mosquito species. Furthermore, the interest of these predictor variables is based on the dependence of specific species observations in large human settlements.

Adding the human population density variable to the model brought an interesting aspect to the study, due to disease control patterns in the Taita Hills area. General human population density data were obtained from the University of Helsinki, from an earlier study about predictive modeling of human population distribution in Kenya (Heiskanen *et al.* 2015). Data were modified by digitizing more houses in the study area in QGIS and by giving the value of 6 for each house, as that has been stated to be the average number of residents per house in the Taita Hills (Broberg *et al.* 2004). Subsequently, the human population density was calculated for each collection point by using the Spatial Analyst Extention- and Kernel Density-tools in ArcMap. The average human population density of collection points was 1258 residents per km² (Figure 12).

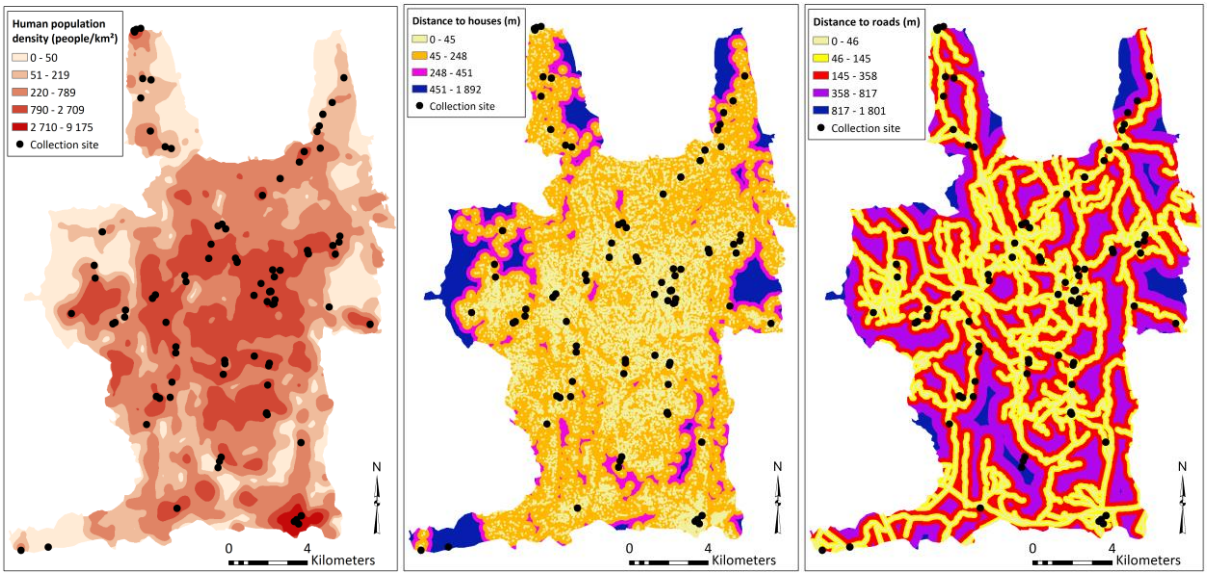


Figure 12. A. Locations of the largest villages in Taita can be recognized by the highest population densities. B. The locations with more than 300 meters to nearest house are situated either in the forestall area or in the national parks, such as Tsavo West national park in the west and Lumo national park in the southwest. C. The forestal areas are the longest distance (800–1800m) from the nearest roads.

Moreover, distance to houses and distance to roads were thought to be crucial factors for mosquito presence or absence, as mosquito prevalence is affected by infrastructure and human settlements. In other studies, these variables are not used as explanatory variables, and this adds to the interest to study their influence on the occurrence.

Distance to roads values were obtained from road data derived from University of Helsinki and calculated with the Mean Euclidean Distance-tool in ArcMap. Distance to houses values were obtained from the building data (Heiskanen *et al.* 2015) and calculated by mean Euclidean distance tool in ArcMap. An additional raster dataset for variables distance to houses and distance to roads were created. The tool calculated average distance from each collection site to houses or roads in meters. Finally, with the Extract Multi-Values to Points-tool, exact average distance values from collection points were added to the data matrix. Average distance to houses from each collection point was 52 meters and average distance to roads from each collection location was 127 meters (Figure 12). This implies that mosquito collections were mainly implemented among human settlements, but collections were only to some extent affected by road locations.

5.3.5 Building design

During the fieldtrip, we made an interesting observation relating to the building design and mosquito habitat suitability. In Taita, building design is manifested in two different types either cabins and huts built using traditional methods with tiles and mud, or modern buildings built using modern design or material, including cement or by EPS panels (Figure 13; Nduire 2016).

There are structural differences between two building designs, as traditionally designed buildings have more air holes than modernly built houses. Moist air does not stay inside the buildings, but evaporates away due to the air gaps. On the contrary, in more modern buildings, moisture remains inside the structures because there are neither air holes nor a developed ventilation system. As the moisture remains inside the building, it can condense and form for example a small puddle constituting a breeding site for mosquitoes. This was experienced especially in bathrooms where moist air could not circulate and often resulted in collections with high numbers of mosquitoes. Thus, a majority of mosquitoes collected from

inside buildings were located in modern- style houses, and a minority were located in the traditionally built houses. Mosquitoes were not frequently found in traditionally built houses or huts.

Figure 13. A. A traditional house built by clay in Kishushe in the Taita Hills. B. EPS panels or cement, have become popular building materials in the area.

The connection between building design and the abundance of mosquitoes in developing countries has not been studied so far. However, similar observations among bat roosts were observed, as colonies were detected in the modern- style buildings and only isolated bats were observed in the traditional cabins (López-Baucells *et al.* 2017). Building design varied depending on the location where mosquitoes were found (Figure 14). In remote locations, building design was usually a hut built using wattle and daub or a traditional straw hut. Modern building design mainly occurred close to larger villages.

Building design data were obtained by digitizing the houses in mosquito collection sites in QGIS and by classifying houses by observation in the field and by Google Earth in QGIS. The basic information of the houses in the collection location was filled in the collection sheet. Additionally, we used Google satellite map in QGIS and checked each building where mosquitoes were collected. Afterwards, building design data were inserted from QGIS into ArcMap.

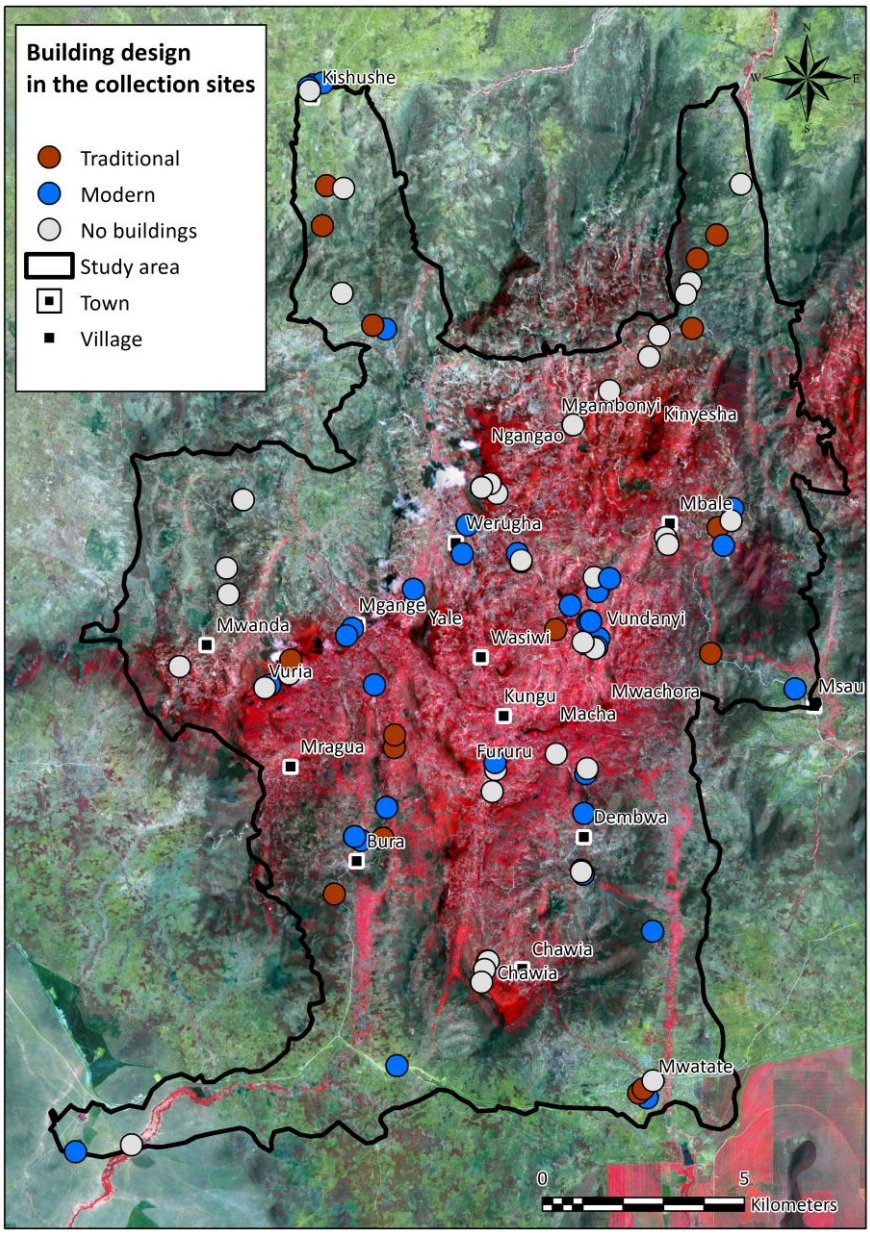


Figure 14. Houses built by modern design in the collection locations, were mainly located close to villages, and traditional huts were primarily located in remote areas.

These 12 introduced variables are representing the factors which may influence the presence and absence of mosquitoes. In R-statistical computing software, the biomod2-package enables prediction with a maximum of 6 to 8 explanatory variables. Multicollinearity of the predictor variables was calculated in order to eliminate correlations between variables and other predictor variables. The final predictor (explanatory) variables which are taken into account in the final distribution model will be introduced later, in the Results chapter.

6. Analysis and modeling methods

Themes such as data preparation, model creation, selection and evaluation were important parts of model process. It was evident to understand the characteristics of different statistical models and to learn differences in the use of them. Model evaluation tested the reliability and validity of the models, and therefore, resulted in ideas how good estimations the models produced. These points, among others, will be introduced more accurately in the next sections.

6.1 Modeling process

The SDM is used here to analyze and describe the dependence of selected variables on the spread of *Culex* and *Stegomyia* and to estimate their distribution to new areas by interpolation methods (Franklin & Miller 2010). In this study, the BIOMOD computation framework (Thuiller 2003) was used for fitting species distribution models. Fitting was implemented in the biomod2 package (Thuiller *et al.* 2016) in R (Version 3.0.3) which enables improved treatment of uncertainties in the models (Thuiller *et al.* 2016). Models were processed mainly using the default settings of biomod2 for the models in the R environment. The aim of biomod2 is to maximize the predictive accuracy of current species distributions and the reliability of future potential distributions using different types of statistical modeling methods (Thuiller 2003). Biomod2 is able to model species distribution using several techniques, to test a model with several approaches, and to project species occurrences into different environmental conditions and into dispersal functions (Thuiller *et al.* 2016). Biomod2 is also useful when assessing species temporal turnover, plotting species response curves, and testing the strength of species interactions with predictor variables, among others (Thuiller *et al.* 2016).

The process involved data preparation and compilation, model creation and calibration, validation, and evaluation. In the next sections, the modeling process of mosquito species is described stepwise. First, data preparation and compilation are introduced and different models are explained. Then, spatial autocorrelation, model calibration, validation and selection are the followed themes. At the end of this chapter, we focus on spatial prediction and model evaluation, which represent the final parts of SDM process.

6.2 Data preparation and compilation

When preparing the data, the quality and accuracy of data was important due to "the garbage in, garbage out rule", which refers to a computers` ability to process whatever type of data provided regardless of its quality or suitability (Drew *et al.* 2011). Therefore, it was necessary to pay attention to data quality from the very beginning of the modeling process. In this section, the data preparation and compilation processes and multicollinearity of predictor variables are described. Mosquito data were first modified in Excel and then inserted into ArcMap, as environmental and other explanatory variables of each collection point were added to the data matrix. When the dataset was finalized, it was saved as a .csv file, ready for insertion into the R- software.

The process began by transferring the GPS data for each sampling location from the GPS-receiver and entering it into Excel. Excel data were then inserted into ArcMap software, where environmental and other attributes of each sampling point were extracted after adding raster datasets and clipping all the predictor layers to the same spatial extent, geographic coordinate system and cell size or resolutions as those of the study area layer.

Next, geographic coordinate system, resolution and study extent were defined to the study area. Geographic coordinate system was defined to WGS 1984 UTM Zone 37S, which is commonly used in Kenya and Eastern Africa. Resolution for each environmental, anthropogenic and distance dataset was set to 20 meters x 20 meters. The extent of the study area defined the extent values of environmental and other layers. Clipping these aspects for each attribute was implemented by Extract by Mask-tool in ArcMap.

The compiled data consisted of species occurrence or non-occurrence and environmental and other data of each sampling point. This enabled starting the modeling process in R, as the basic settings was same in each layer. Occurrence data were saved as a .csv file in Excel and each raster dataset representing an explanatory variable was changed to an ASCII file in ArcMap, as biomod2 requires raster files to be in as .asc- format before inserting explanatory variables to the program. In R, biomod2 and raster packages were installed and all the required libraries were loaded so that inserting data were enabled. Then, data were assessed

relating to the multicollinearity of variables, so that the predictor variables used in the final model were confirmed. After that, the study proceeded to model creation.

6.2.1 Multicollinearity

In order for the model to produce as reliable and accurate results as possible, there exist a few methods to reduce errors and misinterpretation. In multiple regression, with the presence of correlated explanatory variables, a phenomenon called multicollinearity should be taken into account (Franklin & Miller 2010). Multicollinearity can render variables that may be causal or proximal to be insignificant or reverse the sign of the coefficient (Franklin & Miller 2010).

Multicollinearity was implemented in R with the command which seeks correlations between predictor variables using the Pearson correlation coefficient. Multicollinearity was investigated between predictor variables, resulting in the observation that many predictor variables were correlated with each other. Original environmental and anthropogenic data consisted of 12 variables, but only 8 were used after elimination of highly correlated variables. Highly correlated variables ($r \geq 0.9$) were mean precipitation, mean relative humidity, mean radiation, elevation and mean temperature. Only one of these variables was selected for the final modeling process. The others were removed. On the contrary, variables which have independent relevance, have to be retained in the model (Dormann *et al.* 2013). Thus, the distance to houses, distance to roads, slope, elevation, population density, land cover, building design and NDVI -variables were not highly correlated and were retained in the model process.

6.2.2 Spatial autocorrelation (SAC)

Autocorrelation relates to data or model residuals which are correlated with each other rather being independent (Drew *et al.* 2011). In order for appropriate results to be achieved, it was obvious to pay attention to this phenomenon on response variables. Spatial autocorrelation (hereafter SAC) refers to the covariation of properties with space, so that values of a variable are positively or negatively related as a function of proximity or distance (Cressie 1991; Legendre 1993; Anselin *et al.* 2004). SAC arises either from endogenous or exogenous processes; endogenous processes refer to the processes generated by the response variable of interest such as dispersal or competition, and exogenous processes refer to the processes

which occur independently of the variable of interest, such as disturbance or historical barriers to dispersal. Spatially correlated patterns in the study data of this research is measured by Moran´s Index (Moran 1950).

If spatial autocorrelation is not noticed before model creation, it may cause challenges in the development and implementation of appropriate statistical analyses (Drew *et al.* 2011). Some consequences of spatial autocorrelation are e.g. decreased precision of coefficients, shift towards models with more predictors, and variable selection predisposed towards more strongly autocorrelated predictors (Drew *et al.* 2011; Lennon 2002; Diniz- Filho *et al.* 2003). Solutions to the problems of spatial autocorrelation could be implemented e.g. by removing the spatial structure from the data by first mapping residuals for interpreting potential causes of spatially structured errors, or by accepting coefficients as significant even if they might not be. In this study, SAC of predictor variables of *Culex* and *Stegomyia* were tested.

6.3 Model fitting

Model fitting was implemented in biomod2 package after data were prepared and compiled. The biomod2 package uses regression techniques and modern machine learning methods, which have greater performance for nonlinear relationships (Elith 2009). These methods aim at predictive species distribution modeling. Predictive statistical models can be categorized as statistical, machine learning, classification and distance methods (Franklin & Miller 2010). The biomod2 package enables model comparison by accuracy and reliability assessment (Thuiller *et al.* 2016).

In biomod2, eight of its total ten available statistical models were used to estimate species distribution in this study. Generalized linear model (GLM), generalized additive model (GAM) and generalized boosted model (GBM) are introduced as statistical methods, classification tree analysis (CTA) represents the classification methods, and artificial neural networks (ANN), multivariate adaptive regression splines (MARS), random forest (RF) and maximum entropy modeling (Maxent) includes in machine-learning methods. These all methods were used in this study. In the following sections, models are explained in detail.

6.3.1 Generalized linear model (GLM)

Generalized linear model (GLM) is a traditional statistical modeling method which is an extension of the classical linear regression method, where a dependent variable is transformed based on the so-called link-function, describing mathematically the way the dependent variable is transformed relative to its mean value (Kienast *et al.* 2012). GLM can be used for binary data such as presence-absence data, and for being coping with non-normal distribution of response variable (Drew *et al.* 2011; Franklin & Miller 2010). Flexibility in GLM can be achieved by including additional transformations of predictors such as polynomial terms, as well as by including link-functions in the commands in R (Franklin & Miller 2010; Kienast *et al.* 2012). GLM is a limited model but has clear parametric shapes, contrary to generalized additive model which will be introduced next (Kienast *et al.* 2012).

6.3.2 Generalized additive model (GAM)

Generalized additive model (GAM) is a non-parametric extension of GLMs, and is a flexible and automated approach to identify and describe non-linear relationships between predictors and response (Yee *et al.* 1991). GAM lets the data find the best solution to the shape by adding a selection of local smoothing functions along the gradients of predictor variables, which affects flexibility (Kienast *et al.* 2012). The degree of freedom of a smooth fit is usually a real number, quite often either 3 or 4 (Drew *et al.* 2011). The fit of GAM is evaluated by testing non-linearity of the predictor variable by comparing a model with a linear fit for that predictor against non-parametric fit (Drew *et al.* 2011). According to Kienast *et al.* (2012), GAM models may result in more accurate predictions if calibration data is used to test the data.

6.3.3 Classification tree analysis (CTA)

CTA results in highly independent data, and being an iterative optimization algorithm, it seeks to optimize a dichotomous decision key in order to explain the dependent variable from a set of independent predictors (Kienast *et al.* 2012). With CTA, it is easy to interpret and address data interacting in a non-linear or hierarchical manner (Breiman *et al.* 1983; De'ath 2000; Rogan *et al.* 2008). In CTA, additional predictors can be virtually searched, until each data point is split and explained in an enormous tree (Kienast *et al.* 2012). Even if CTA can bring about almost perfect prediction results, it is prone to overfitting and a goal must be

simplifying these overfittings (Kienast *et al.* 2012). This can be done e.g. through cross-validation, where full data sets are split to classes and re-fit to the tree (Kienast *et al.* 2012).

6.3.4 Artificial neural networks (ANN)

Artificial neural networks (ANNs) are machine-learning methods which first have been developed for the human brain (Benediktsson *et al.* 1993). The basic step in ANNs is realized in deriving features that are linear combinations of predictors, then modeling the output (response) as a non-linear function of those features (Drew *et al.* 2011). In ANNs, overfitting training data using a large network is avoided by limiting the number of iterations of estimation, using the same method as in CTA, cross-validation. Thus, there exist large group of models classifying classes where there are units in the output layer (Moisen *et al.* 2006).

6.3.5 Multivariate adaptive regression splines (MARS)

Multivariate adaptive regression splines (MARS) is a form of stepwise linear regression which produces continuous models with continuous derivatives, and has more power and flexibility to model relationships which are nearly additive or involving interactions (Friedman 1991). MARS is able to model non- linearity, and introduce itself as a flexible regression modeling of high-dimensional data, taking the form of an expansion in product spline basis functions (Friedman 1991). The number of basic functions and parameters associated with each one are automatically determined by the data (Friedman 1991).

6.3.6 Generalized boosted model (GBM)

GBM is an extension of MARS which accepts non-linear transformations and interactions and treats missing values without filling in values or removing observations (Heikkinen *et al.* 2006; Ridgeway 2007). It is a more flexible regression model but can also result in over-fitted data. Boosting estimates classifiers iteratively by using a base learning algorithm, e.g. a decision tree, which systematically varies the training data with the aim of improved prediction accuracy (Heikkinen *et al.* 2006). The final prediction is based on an accuracy-weighted vote across the estimated classifiers (Heikkinen *et al.* 2006).

6.3.7 Random forest (RF)

Machine-learning methods also include tree-based methods, which include classification and regression trees, and so-called decision trees (DT) e.g. Random forest (RF), whose goal is to partition data into subgroups that are homogenous (Drew *et al.* 2011). A single split results in a tree with nested binary decision thresholds, dichotomous key (Drew *et al.* 2011). Partitioning is stopped when a resulting split doesn´t achieve some defined level of homogeneity, or if resulting subsets would have less than some minimum number of members (Drew *et al.* 2011). Advantages of these decision trees are that they can also handle categorical predictors and characterize hierarchical interactions (Drew *et al.* 2011). Moreover, threshold responses are presented and resulting outputs are informative and can also handle missing data and classify new data (Drew *et al.* 2011). Disadvantages include the fact that decision trees or RF can be unstable and over-fitted, as well as data-driven (Drew *et al.* 2011).

6.3.8 Maximum entropy model (Maxent)

Other machine-learning methods include Maxent which is a general-purpose machine-learning model (Phillips *et al.* 2013). The idea of Maxent lies in the fact that it estimates a target probability distribution by seeking the probability distribution of maximum entropy (Phillips *et al.* 2013). Previous discussions have defined Maxent as evaluating a distribution across the geographical space, but according to Elith *et al.* (2011), there exists a new pattern of Maxent which compares probability densities in covariate space through a correlate range of variables. The importance of Maxent is based on the fact that it minimizes the relative entropy between two probability densities defined in the feature space (Elith *et al.* 2011). However, it was argued that maximum entropy modeling has the problems of developing unrealistic and misleading habitat suitability maps when predicting species distributions in both contemporary environments and making future projections (Barnhart *et al.* 2014). It has a simple and precise mathematical formulation and is well-suited to model species distributions using *presence-only* data, as absence data is not always available for all species (Phillips *et al.* 2013). Some drawbacks of Maxent are that it is not as mature a statistical method as GLM or GAM, and there exist less established guidelines for use (Phillips *et al.* 2013). Also, Maxent uses an exponential model for probabilities, which gives very large predicted values for environmental conditions outside the range present in the study area (Phillips *et al.* 2013).

6.4 Model selection and prediction

In the model process, an obvious step was to pay attention to the calibration of the model. When evaluating data, it is necessary to divide data into one portion to use to calibrate the model, so-called training data, and another portion to validate the predictions, so-called testing data (Franklin & Miller 2010; Fielding & Bell 1997). Thus, we split our data into training data which consisted of 70% of the total samples, and into test data which included 30% of total samples.

Moreover, it was important to determine explanatory factors which had the strongest influence on the dependent variable. Before continuing to the analysis of variable importance, it was important to ensure that predictor variables were not correlated. This was important in order that interpretation of variable importance would be clear and straight forward (Murray *et al.* 2009). There exist many indices which measure variable importance in statistical models including zero-order correlations, partial correlations, semipartial correlations, standardized regression coefficients, Akaike weights and independent effects (Murray *et al.* 2009). All of these mentioned indices are useful when assessing the relative importance of variables. A relative importance of the predictor variables was assessed in this study and resulted in the factors which were most influential in estimating the mosquito occurrence. The variable contributions will be introduced in the chapter Results.

Data validation represents an important part of modeling process. Validation refers to the fact that a model must meet specified performance requirements to be accepted for its intended use (Rykiel 1996). Validating data prevents prediction errors which distort the results. Prediction errors may result from data errors or errors in model specification (Barry & Elith 2006). In data outputs, prediction errors were measured by R^2, proportion of variance explained, by RMSE (root mean square error) or by D^2, explained deviance (Drew *et al.* 2011). In calibration, the model is verified if predictions of 0.6 have a 60% chance of being occupied, and if these predictions are twice as likely to be occupied or as suitable as predictions of 0.3 (Vaughan & Ormerod 2005).

All models aim for high predictive accuracy and precision; and currently, there is a higher need for robust accuracy assessment of all models (Drew *et al.* 2011). There exist several methods for selecting the best models to the prediction. The Akaike Information Criterion

(AIC) presents a statistic used in an information –theoretic approach to model selection, and is the measure of goodness of fit that pays attention to the number of parameters (Burnham *et al*. 1998). AIC was used here in order to measure goodness of our resulted model fits.

Each of the earlier introduced 8 models was run for every mosquito species. These models were compared regarding the goodness and reliability of fit, and spatial predictions were obtained on the occurrence data. Predictive modeling of species distributions has two main objectives. The first objective of prediction is called model-based interpolation to unsampled locations, where new sites in the sampled area are sampled by the training data within the same general time frame as the species which have been sampled (Elith & Leathwick 2009). The second objective of prediction is extrapolation, which means that species occurrences are predicted to new and unsampled areas outside the study area, or to future climates (Elith & Leathwick 2009). In our study, we concentrated on the first objective of predictive modeling; to interpolate unsampled locations inside the study area in order to produce estimations for areas of potential distribution of *Culex* and *Stegomyia*.

6.5 Model evaluation

We evaluated the predictive power of models using the area under the receiver operating characteristic (ROC) curve (AUC), Cohen´s kappa (κ) (Cohen 1960) and the true skill statistic (TSS) (Peirce 1884). The classifications of each are shown in Table 2. However, we mainly focused on evaluating models by AUC (Mason and Graham 2002), which is a measure of rank-collection describing how well the model explains the correlation of selected predictor and explanatory variables (Hijmans & Elith 2016). Those models which had highest AUC-values ($0.7<AUC>1.0$) with significant statistical rates ($p\leq0.05$), were taken into account in the prediction.

Table 2. The classification of AUC, Kappa and TSS (Mason & Graham 2002; Cohen 1960; Peirce 1884).

AUC	Kappa	TSS
0.9-1.0 Excellent	0.7-1.0 Excellent	0.7-1.0 Excellent
0.8-0.9 Good	0.4-0.7 Good-moderate	0.4-0.7 Good-moderate
0.7-0.8 Moderate	0.0-0.4 Poor	0.0-0.4 Poor
0.6-0.7 Poor		
0.5-0.6 Failed		

In AUC, ROC-curve indicates sensitivity plotted against [1- specificity] where sensitivity refers to the proportion of true positives and specificity to the proportion of false negatives (Drew *et al.* 2011). In unbiased data, a high AUC value indicates that the prediction is accurate and that the produced predictive species map has sites with high predicted suitability values, which tend to be areas of known presence; locations with lower model prediction values tend to be areas where the species is not known to be present (Hijmans & Elith 2016).

The AUC value can vary between 0 and 1, and a value of 0.5 implies that model is as good as a random prediction (Heikkinen *et al.* 2006). Values below 0.5 indicate systematically wrong estimations (Kienast *et al.* 2012). In ecological modeling, 0.7 is minimum AUC value for reliable estimations and a score of 1 indicates a perfect prediction (Drew *et al.* 2011). In order to receive the highest possible prediction accuracy values, we compared the different sets and orders of predictor variables.

Depending on the strength of relationship to response, we selected the best combination of predictors to include in a model. We also compared the variable contributions in the models in order to define the most powerful environmental, anthropogenic or distance factors, and their relative magnitude and order in the share. In this study, we concentrated on analyzing the models with best accuracy which obtained AUC value of 0.7 at minimum.

7. Results

Results from the mosquito identifications provided an answer to our first study question; namely; which mosquito genera are present around the Taita Hills and how they are distributed. The modeling outputs responded to two other study questions; which environmental, anthropogenic and distance factors influenced in *Stegomyia* and *Culex* presence most and whether any of the models can reliably estimate the distribution of *Culex* and *Stegomyia* mosquitoes. More than seven mosquito genera were collected from the Taita Hills, of which only two genera, *Culex* and *Stegomyia*, obtained a sufficient number of observations for modeling (approx. 30 locations of presence).

The models predicted an increasing probability of *Culex* in the locations with high human population densities and with moderate NDVI values at minimum. The probabilities and most influencing variables of estimations for *Stegomyia* were quite different from those of *Culex.* The locations with high solar radiation values and high temperature were favorable for *Stegomyia* presence, and highly-populated locations were revealed to be an unsuitable condition. However, anthropogenic factors had high influence in both *Culex* and *Stegomyia* presence, and of the environmental drivers, topographic and temperature factors had the greatest impact. In this study, all statistical significance values (p-values), are marked as follows: statistically highly significant *** = <0.001, statistically significant ** = <0.01, statistically significant *= <0.05 and ns = not statistically significant.

7.1 Observed mosquito genera and their distribution in the Taita Hills

The seven named mosquito genera were collected in the Taita Hills during the fieldtrip, included *Culex, Stegomyia, Uranotaenia, Aedimorphus, Eretmapodites, Anopheles* and *Lutzia.* The largest collections were of *Culex* (73 locations) with 2600 mosquitoes and *Stegomyia* (28 locations) with 180 mosquitoes. Other genera *Uranotaenia, Aedimorphus, Eretmapodites, Anopheles* and *Lutzia tigripes* were collected from fewer than 10 locations and each collection consisted of fewer than 40 mosquitoes; for this reason, they were not used in the modeling process. Furthermore, altogether 300 mosquitoes in 19 locations were not recognized within the timeframe for inclusion, so; were not included in the model process.

Culex mosquitoes, potential WNV vectors, were collected from 73 of the 122 locations, from over 2600 individuals, which suggests that in a majority of collection locations, *Culex* mosquitoes were present. This is not an uncommon outcome, as it is one of the largest mosquito genera in the world (Figure 15) and at least three subgenera populated the collections; Culex (Culex), Cx. (Culiciomyia) and Cx. (Eumelanomyia). Large collections were made around the Taita Hills, and they were widely distributed across the region. The largest number of *Culex* mosquitoes was collected from primary and secondary schools and from septic tanks. They were found in the villages as well as in forests and national parks. High elevation was not an obstacle for *Culex* occurrence, as they were collected both on the plains at an elevation of 600 m and in the mountains at an elevation of 2000 m.

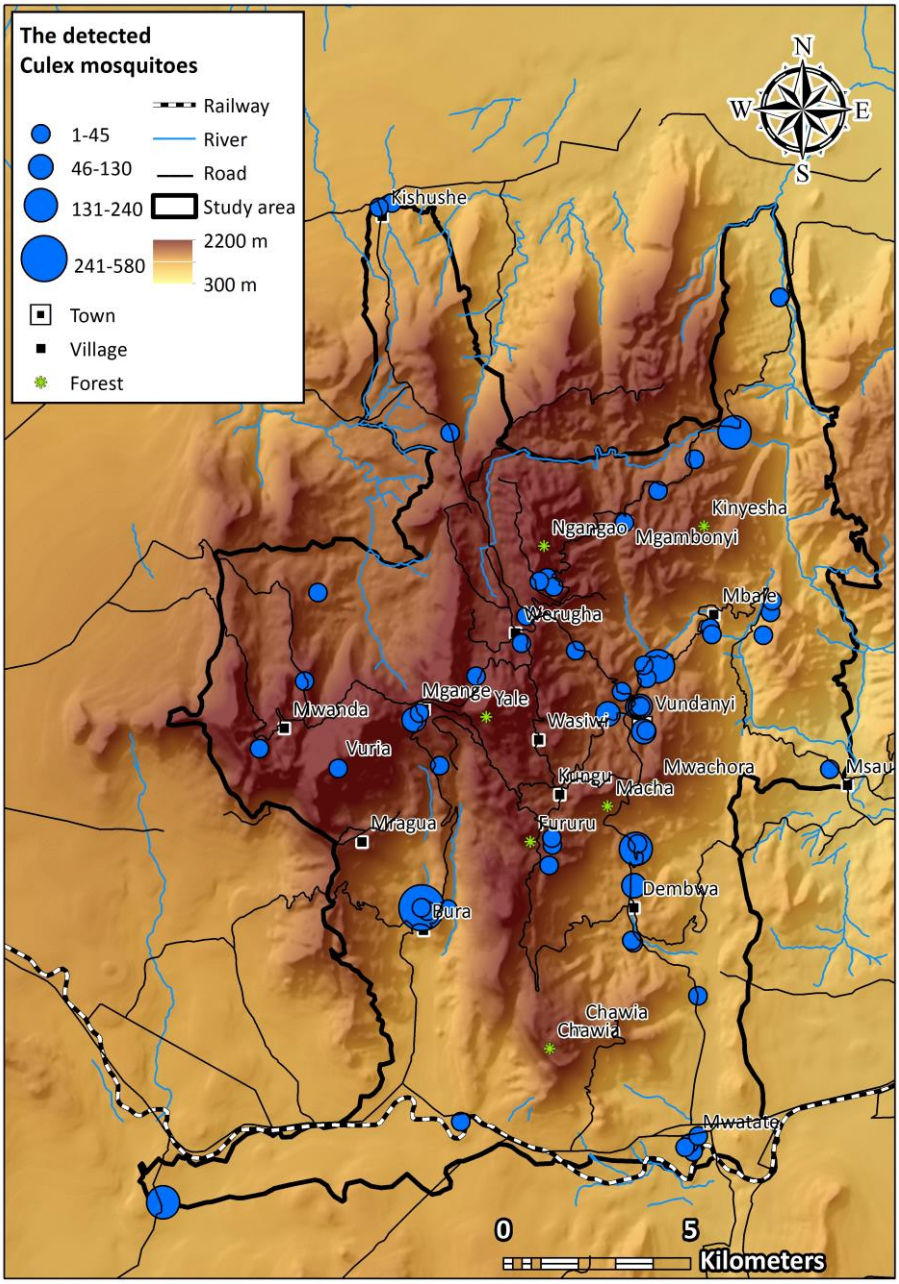


Figure 15. *Culex* mosquitoes were collected at varying altitudes both on the plains and in the mountainous area.

The second most abundant genus in the region was *Stegomyia* with 180 mosquitoes (Figure 16). Like *Culex*, *Stegomyia* was also observed across the study area, including high-elevation areas (≥1500 m). Both adults and larvae of *Stegomyia* were collected close to human settlements. Larvae were found in water tanks and small ponds, and adults were collected mainly biting in roadside homes or homes in the villages.

Stegomyia aegypti was recognized in the collections without the need for molecular identification. *St. aegypti* has been labeled as the most dangerous mosquito species in the

world, in addition to *St. albopicta*, as they both are capable of transmitting over 22 different viruses (WHO 2017c). A large amount of *St. aegypti* larvae were found in water tanks in Paranga Primary School, and also in water tanks in Makyambu and Kishushe in northern Taita.

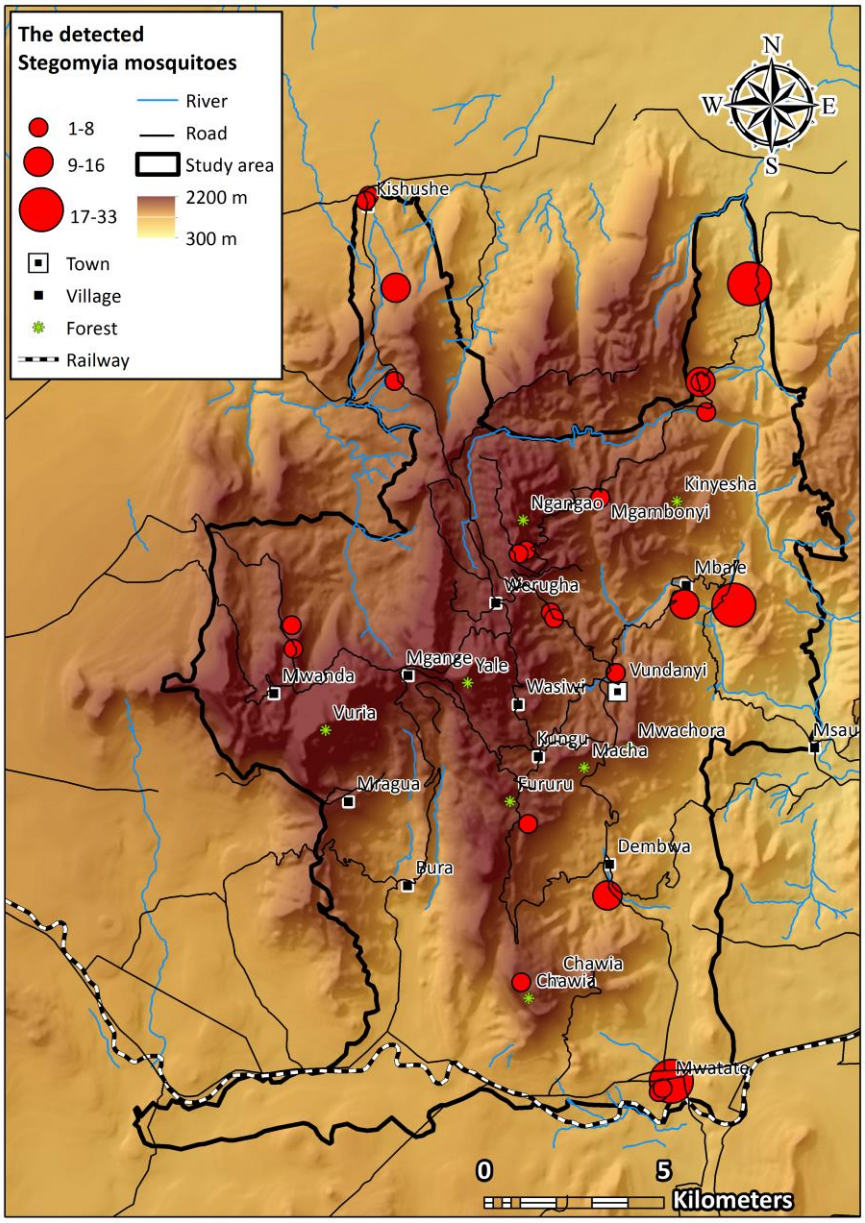


Figure 16. *Stegomyia* were collected from a variety of habitats all over the Taita Hills.

Another great collection of *St. aegypti* larvae were found in car tyres in Mwatate village. All of these collection sites were located at an elevation below 900 m. Nevertheless, *St. aegypti* adults were also collected whilst biting at Taita research station at an elevation of 1360 m above sea level. Furthermore, many as yet unidentified *Stegomyia* individuals were found at a variety of elevations in the Taita Hills, also in Ngangao montane forest, which is located at

over 1800 meters elevation. These mosquitoes will be identified in due course. Originally members of the genus *Stegomyia* were forest species, but it´ s suitability to inhabit a variety of habitats has driven this genus to new areas especially, where the land use is changing (MTI 2017). This was confirmed in the Taita collections since *Stegomyia* mosquitoes were mainly collected close to humans, but were also collected in Ngangao montane forest.

In addition to *Culex* and *Stegomyia*, several other genera were observed in the study region with much lower collection rates. The distributions of these genera, *Uranotaenia, Aedimorphus, Eretmapodites*, *Anopheles* and *Lutzia tigripes* are given in (Figure 17). *Uranotaenia* was collected as adults in the dry river close to research station in Wundanyi. *Uranotaenia* mosquitoes were resting in banana plants and were collected at night.

Only a few *Eretmapodites* mosquitoes were collected from the Taita Hills. These collections occurred in the mountain areas both at the Taita research station and in the Mlondo market in the village of Werugha. *Eretmapodites* was collected as larvae from a turkey baster at the station, and as adults resting in vegetation close to the Mlondo market. As with *Eretmapodites* and *Uranotaenia, Aedimorphus* also had a very low collection rate in the Taita Hills. *Aedimorphus* was only found in the Chawia montane forest. They were collected as larvae in a rut filled with rain water.

Lutzia tigripes had the third highest collection rate after *Culex* and *Stegomyia*, as it was collected at 9 variable locations across the Taita Hills from 40 mosquitoes. Immature staged were found in the water tanks in roadside homes in Mgambonyi and Paranga villages, and from the waste water well at the Taita research station in Wundanyi, as well as from the puddle in Chawia montane forest. As adults, *Lutzia tigripes* was collected both resting on water surface of a water tank. They were also collected resting in a well in downtown Mwatate. Even though *Lt. tigripes* were aggressive biters, they have not been linked to transmission of pathogens (MTI 2017).

A surprising outcome was, that the major malaria vector in Africa, *Anopheles*, was only collected at four locations with a total of 25 mosquitoes in the collections which were mainly found as larvae. In the mountainous area, *Anopheles* was collected as immature from a ground

pool in a cultivated field close to the research station in Wundanyi and also from a ditch in a cultivated field in Josa. Other locations for occurrence situated on the plateau, which is not astonishing, as *Anopheles* prefer plain areas for survival. *Anopheles* larvae were found at an elevation of 300 m in a dam in savanna located at Lumo national park. The only adult *Anopheles* was found resting in a toilet in Timbila Primary Board close to the Voi-Taveta main road between Mwatate and Bura villages. Low collection rates of *Anopheles* in the Taita region was an interesting observation considering that high densities are found in other regions of Kenya.

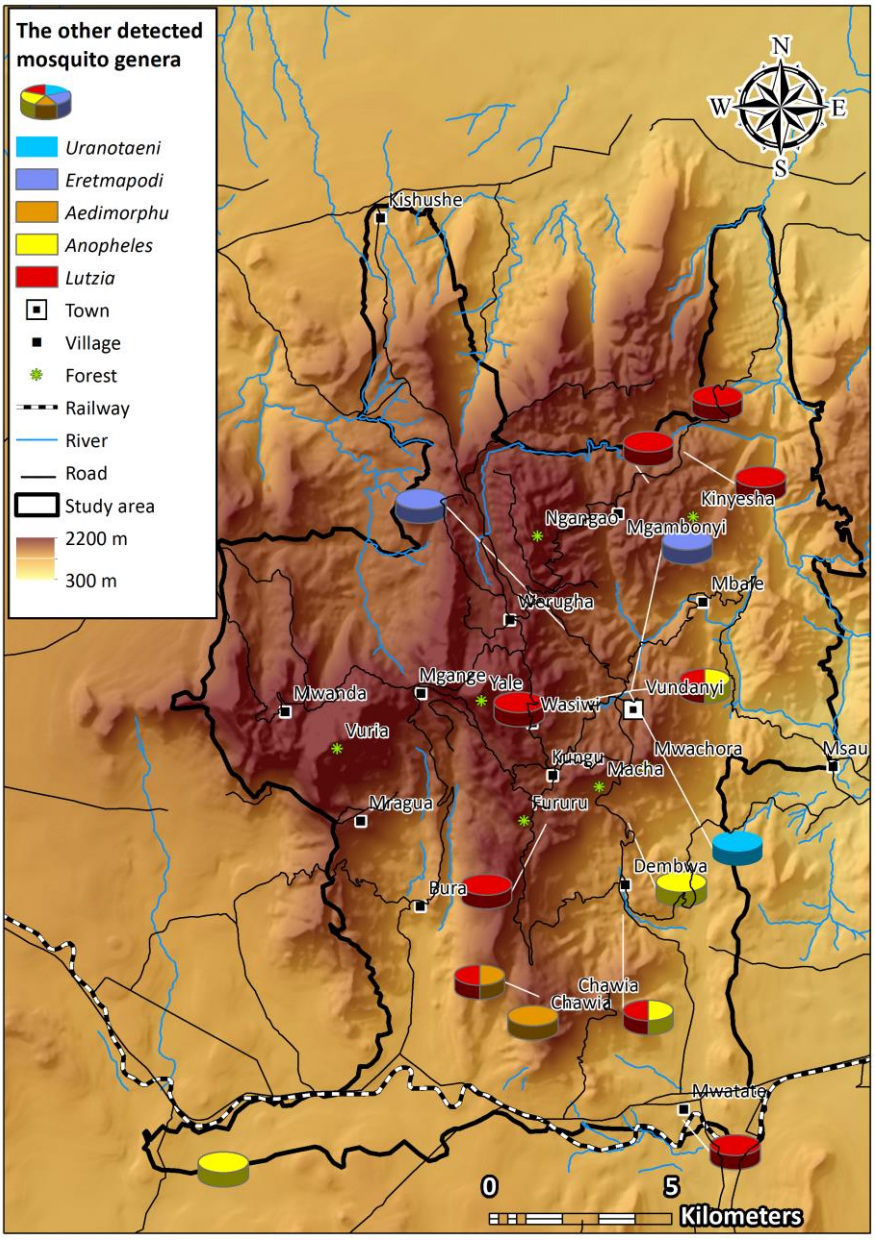


Figure 17. Five genera other than *Culex* or *Stegomyia* were found in the Taita Hills. *Lutzia* was found across Taita. *Anopheles* was found only in a few locations. *Aedimorphus* was collected from Chawia montane forest.

7.2 The multicollinearity of environmental, anthropogenic and distance variables

Regarding the multicollinearity of predictor variables, correlations between variables were run by the Pearson correlation coefficient. The variables NDVI, one of either variables of elevation, mean precipitation, mean temperature or mean relative humidity, either variable of slope or mean radiation, building design, land cover, distance to roads, distance to houses, and human population density were not highly correlated and were retained in the model process. Elevation, mean precipitation, mean temperature and mean relative humidity were highly positively or negatively correlated ($r \geq 0.9$ or $r \geq -0.9$); thus only one of these variables was retained in the model in order to avoid bringing errors or misinterpretation in the study results. Slope and mean radiation were also highly negatively correlated ($r \geq -0.75$), and thus, only the other one was retained in the model. Other environmental, anthropogenic and distance variables did not result in high correlations.

The best (resulting in highest evaluation values) combination of non-correlating variables was run in order to obtain the models with high evaluation value and reliable estimations for *Culex* and *Stegomyia* mosquitoes. Environmental, anthropogenic and distance factors which best estimated the presence and absence of *Culex* were slope, population density, NDVI, distance to roads and elevation. These factors resulted in the best set of predictors to produce reliable models to estimate *Culex* distributions (Table 3). The correlation graphs for *Culex* and *Stegomyia* predictors are shown in Appendices 1 to 10.

Table 3. Selected *Culex* predictors were mainly not highly correlated. Correlation (r) between NDVI and elevation was higher than 0.5 but NDVI was retained in the model.

Variable	Slope	NDVI	Population density	Distance to roads	Elevation
Slope		0.115ns	-0.294**	0.172ns	0.279**
NDVI	0.115ns		-0.226*	0.391***	0.557***
Population density	-0.294**	-0.226*		-0.26**	-0.313***
Distance to roads	0.172ns	0.391***	-0.26**		0.143ns
Elevation	0.279**	0.557***	-0.313***	0.143ns	

The predictors which were not highly correlated and resulted in the best models for estimating the distributions of *Stegomyia* were mean radiation, NDVI, human population density, distance to roads and mean temperature. The highest correlation with significant p-value occurred between mean temperature and NDVI (r = -0.54), but that did not hinder the model process, as biomod2 can manage with intermediate correlations (Table 4). All other predictors resulted low correlations (r ≤0.4) with significant p-values being suitable for estimations.

Table 4. Selected *Stegomyia* predictors were not highly correlated apart from NDVI and mean temperature which obtained r > -0.5. Nevertheless, they were included the model.

Variable	Mean radiation	NDVI	Population density	Distance to roads	Mean temperature
Mean radiation		-0.04ns	0.312***	-0.12ns	0.10ns
NDVI	-0.04ns		-0.226*	0.391***	-0.543***
Population density	0.313***	-0.226*		-0.26**	0.286**
Distance to roads	-0.12ns	0.391***	-0.26**		-0.12ns
Mean temperature	0.10ns	-0.543***	0.286**	-0.12ns	

7.3 Variable contributions in estimations of *Culex* and *Stegomyia* distributions

The influence of each variable in all eight models was consistent for variables that were very strong or moderate predictors. Human population density had the greatest effect on the distributions of both *Culex* and *Stegomyia* in all models, but its importance in each model varied. Moderately explaining variables, NDVI, slope, distance to roads and elevation were not ranked consistently in the models. The influential variables for both *Culex* and *Stegomyia* mosquitoes were NDVI and distance to roads, these resulted in moderate influences. Elevation was a surprisingly poor predictor in all models for *Culex*, and it was not included in the set of predictors for *Stegomyia* either.

Anthropogenic and distance variables generally were good predictors compared to environmental drivers. From the environmental variables, only mean temperature, mean radiation, elevation, slope and NDVI were influential factors in the models. Most of them

indirectly affect the distributions of *Culex* and *Stegomyia*; thus, they had little effect on model performance. Only mean temperature and NDVI are direct factors affecting the distributions of *Culex* and *Stegomyia*. At first, we concentrate on studying the variable contributions of *Culex* estimations, followed by the variable contributions of predictors in *Stegomyia* estimations. We begin with an overall review of variable contributions in each model. Afterwards, we focus on considering variable contributions in GLM and GAM models by response curves in *Culex* estimations; later, we examine variable contributions in GBM and RF models estimating *Stegomyia* distributions. Finally, we consider the response curves of each mentioned models. Overall, human population density was a highly important variable explaining the *Culex* estimations in each model (Table 5). The contribution of population density varied from 43% up to 100% in the models. Other variables were slightly significant. NDVI, slope and elevation were intermediate predictors with contribution from 4% to 45%. Elevation had a surprisingly low effect on *Culex* presence as it obtained low contributions (<10%) in all models apart from Maxent (44%).

Table 5. Variable importance presented in each model for *Culex* estimations. Overall, human population density was the most influential predictor. Elevation had surprisingly little effect on *Culex* distributions.

Variable	GLM	GAM	GBM	CTA	ANN	MARS	RF	Maxent
Slope	0	0.147	0.078	0	0.147	0.385	0.200	0.432
Population density	0.658	0.764	0.678	1	0.773	0.702	0.429	0.513
NDVI	0.261	0.165	0.088	0	0	0	0.114	0.382
Distance to roads	0	0.179	0.039	0	0.092	0	0.064	0.453
Elevation	0	0.028	0.006	0	0.097	0	0.030	0.436

The most important predictor for *Culex* in the GLM model was human population density, which had a relative influence of 66% (Figure 18). In the GLM model, only human population density and NDVI (26%) were influential when estimating *Culex* distributions. Slope, distance to roads and elevation did not have any impact in GLM.

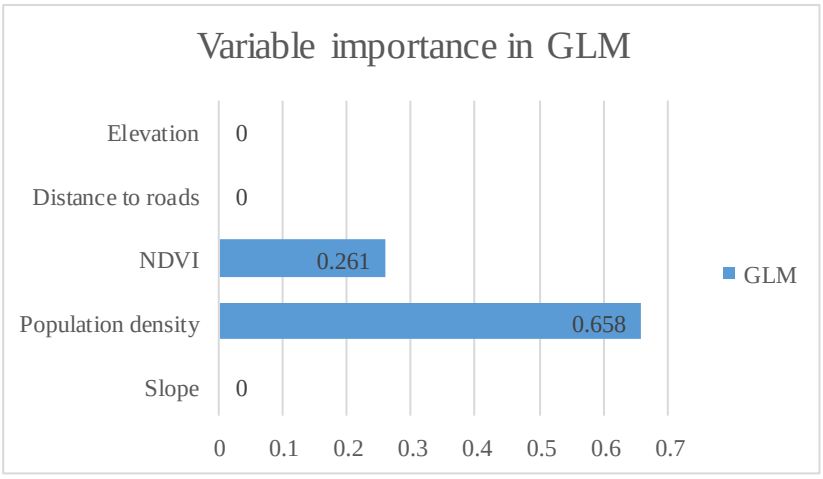


Figure 18. Population density and NDVI were influential factors to *Culex* estimations in GLM model.

According to the GLM model, it is rarer to find *Culex* in the locations with low human population density than in the locations with higher population densities (Figure 19). In the locations, where human population density is lower than 1500 people per km², the probability of *Culex* presence is less than 50%. When the amount of people increased to 2000 persons per square kilometer, the probability of observing *Culex* rose to 80%. Thus, villages and towns are more favorable areas for *Culex* than remote areas.

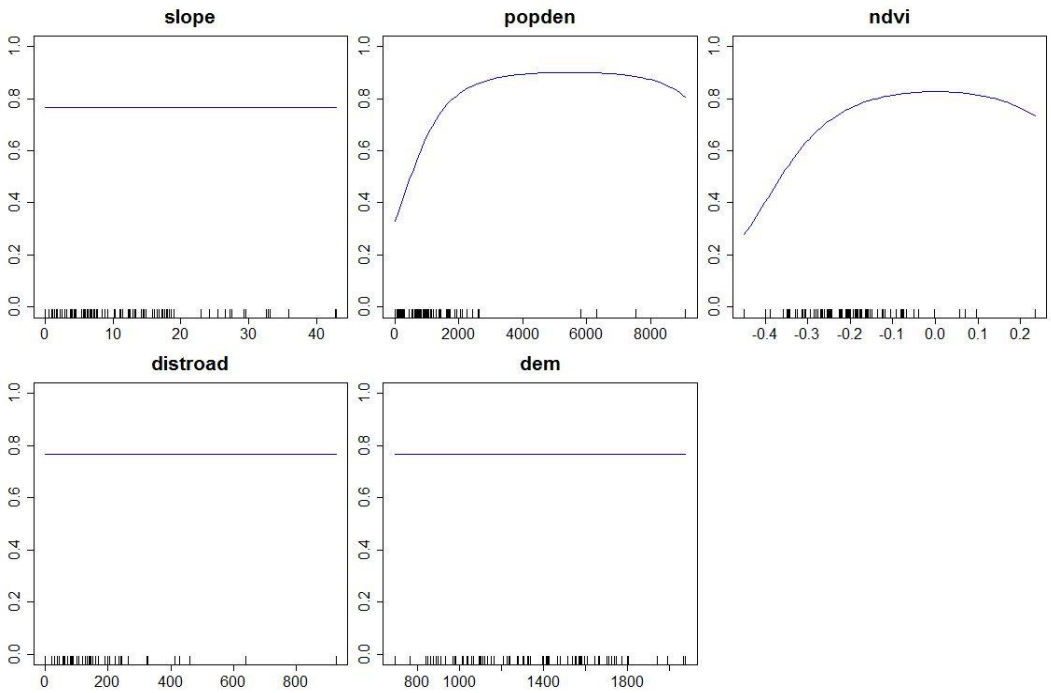


Figure 19. The response curves for *Culex* estimations by GLM model. We can notice that only population density and NDVI influence the probability of presence values in the GLM. The tick marks on the x-axis imply observations.

The trend is quite similar with the NDVI variable in the GLM model. *Culex* receives lower presence estimations (less than 25%) for NDVI with a value of -0.4 or less. The areas with moderate NDVI values ($-0.2 \leq NDVI \geq 0.2$) prove to be the most favorable areas for *Culex*, according to GLM estimation. When NDVI values rise up to 0.2, it seems that the probability of finding *Culex* decreases. This assumption would strengthen the link between *Culex* presence and population density and imply that *Culex* mosquitoes thrive in the vicinity of people. Hence, moderately green areas are more suitable habitat for *Culex* than areas with poor vegetation or extremely green areas.

GAM estimations for *Culex* presence differed fairly from those of GLM. The variable contributions varied more in GAM (Figure 20). As well as in GLM, human population density responded well to the presence of *Culex*. Even if human population density had high contribution (76%) in GAM, other predictor variables were also influential. Slope (15%), elevation (3%), distance to roads (18%) and NDVI (17%) had moderate or low, but significant, contributions when assessing *Culex* distributions.

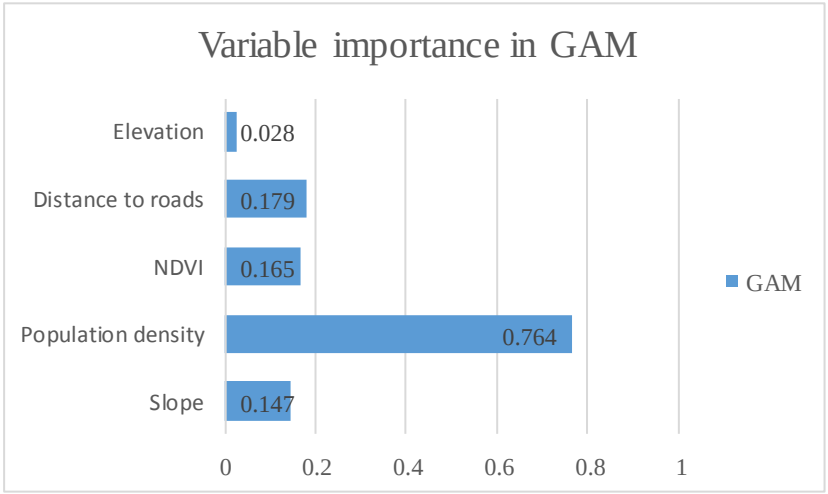


Figure 20. Population density was a major factor also in GAM model, but other predictors were also influential.

Each of these five predictors responded quite well to the estimations of the presence of *Culex* mosquitoes (Figure 21). With moderate slope angles (0°–35°), the estimations were really high (80%) for *Culex* presence. When slope angle obtained values greater than 35°, the likelihood of observing *Culex* quickly decreased. Thus, steep locations were not favorable for *Culex* mosquitoes.

Human population density in GAM differed from the contribution in GLM. Suitable locations for finding *Culex* (≥80%) were locations with population density between 500–2000 persons per km² and locations with population density of ≥6000 persons per km². However, the locations with low population density (≤ 500 persons/km²) were not the most favorable areas to *Culex* occurrence (≤ 40%). A surprising gap occurs in the locations with the population density between 2500 persons/ km² and 5000 persons/km². According to GAM, there is not a huge variability for different NDVI values between -0.4 and 0.2. The probability to locate *Culex* is just slightly lower (≤80%) in the locations with NDVI ≤ -0.4 than in the locations with -0.4≤ NDVI ≥0.2 (≥80%).

The probability for *Culex* to be present is high (≥80%) when distance to roads is 500 meters or less. This also affirms the presumption that several *Culex* species are dependent on humans and their activity in the vicinity. When distance to roads from the location increases (≥500 m), the probability of observing *Culex* decreases quickly to less than 20%. According to GAM, elevation responds quite similarly to the presence of *Culex* between elevations of 800 and 2000 meters (≥80%).

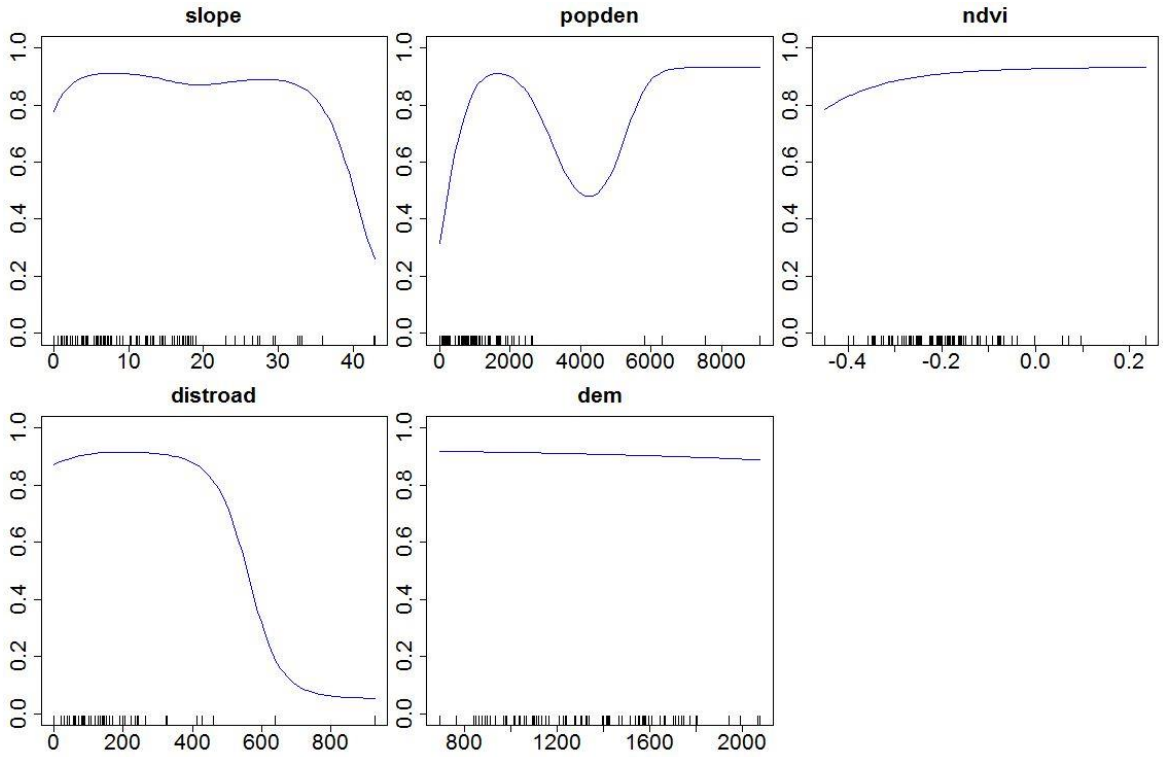


Figure 21. Response curves of predictors for *Culex* estimations by GAM model. Each predictor variable responds to the probability of presence of *Culex* mosquitoes. The black tick marks on the x-axis mean observations.

Stegomyia estimations differed greatly from the *Culex* estimations. Variable importance of *Stegomyia* predictors in each model ranged considerably (Table 6). As in the estimations of *Culex* distributions, population density was the most influential factor in each model. All the other predictors were less than 40% influential in the models. Only the generalized boosted model and random forest provided high enough AUC- values for reliable estimations. Next, we focus on considering more variable contributions in both GBM and RF models.

Table 6. Variable contributions of *Stegomyia* predictors are introduced by each model. Overall, population density was the most influential predictor, but other predictors were also important.

Variable	GLM	GAM	GBM	CTA	ANN	MARS	RF	Maxent
Mean radiation	-	0.041	0	0	0.289	-	0.059	0.176
Population density	-	0.575	0.680	0.966	1	-	0.206	0.357
NDVI	-	0.388	0	0	0	-	0.210	0.291
Distance to roads	-	0.337	0	0	0.215	-	0.087	0.028
Mean temperature	-	0	0.358	0	0.081	-	0.089	0.026

When further studying GBM, we can note that only two predictors were influential, when modeling *Culex* by GLM (Figure 22). Population density (68%) and mean temperature (36%) were the most influential factors in *Stegomyia* estimations. Mean radiation, distance to roads and NDVI had no contribution at all when estimating *Stegomyia* presence by GBM.

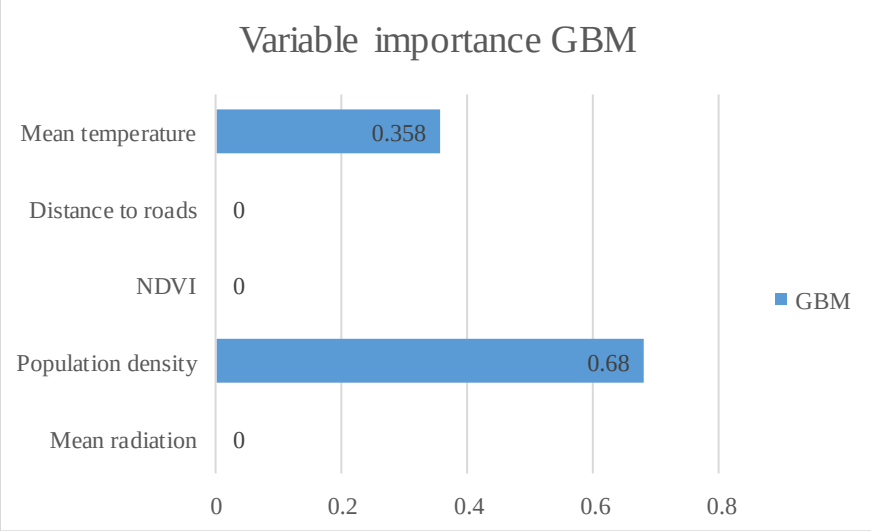


Figure 22. Population density and temperature were the only influential factors in the GBM model when distance to roads, NDVI and mean radiation were not important.

Compared to GBM, random forest resulted in more contributions for each predictor (Figure 23). NDVI (21%) obtained even higher contributions than human population density (20%) and was an even more important factor in random forest. Mean temperature, distance to roads and mean radiation, also have an effect on *Stegomyia* presence, but they obtain only low contributions (≥10%).

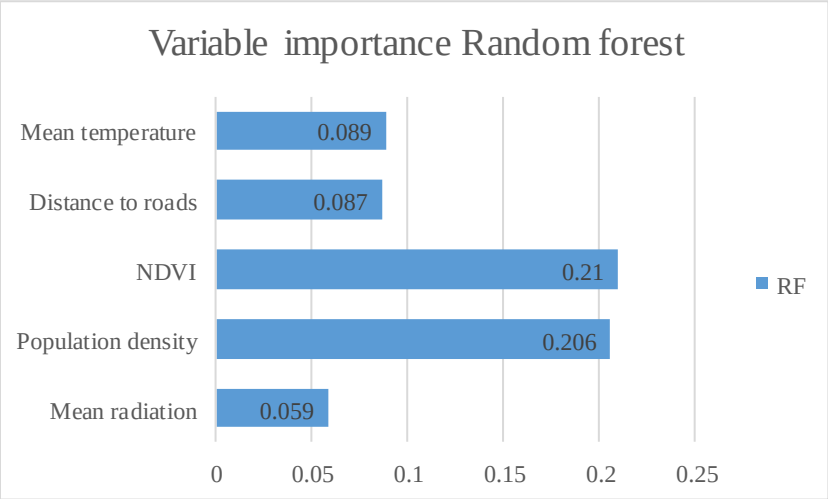


Figure 23. Population density and NDVI were major factors also in the random forest model, but other predictors were also influential.

When focusing on assessing the GBM model and the response curves which show how predictor variables respond to *Stegomyia* estimations, we can make a few observations (Figure 24). Mean radiation, NDVI and distance to roads were not responding to *Stegomyia* presence, as they were not influential variables in GBM. A surprising observation was that when population density was between 0 and 1000 persons per km², the probability of observing *Stegomyia* was higher (40–65%). However, as human population density increased over 1000 persons per km², the likelihood of observing *Stegomyia* decreased to less than 40%. According to this observation, high population density would not be a prerequisite for the presence of *Stegomyia* species.

Stegomyia is also affected by changes in mean temperature in GBM model. The locations with temperatures between 20º and 23 Cº were not suitable for *Stegomyia*, as the probability of their presence decreases (less than 40%). Nevertheless, when mean temperature in the locations remains between 15º and 20 Cº, or if it rises (≥ 23 Cº), the conditions are more favorable to *Stegomyia* and the probability of observing them increases to over 40%.

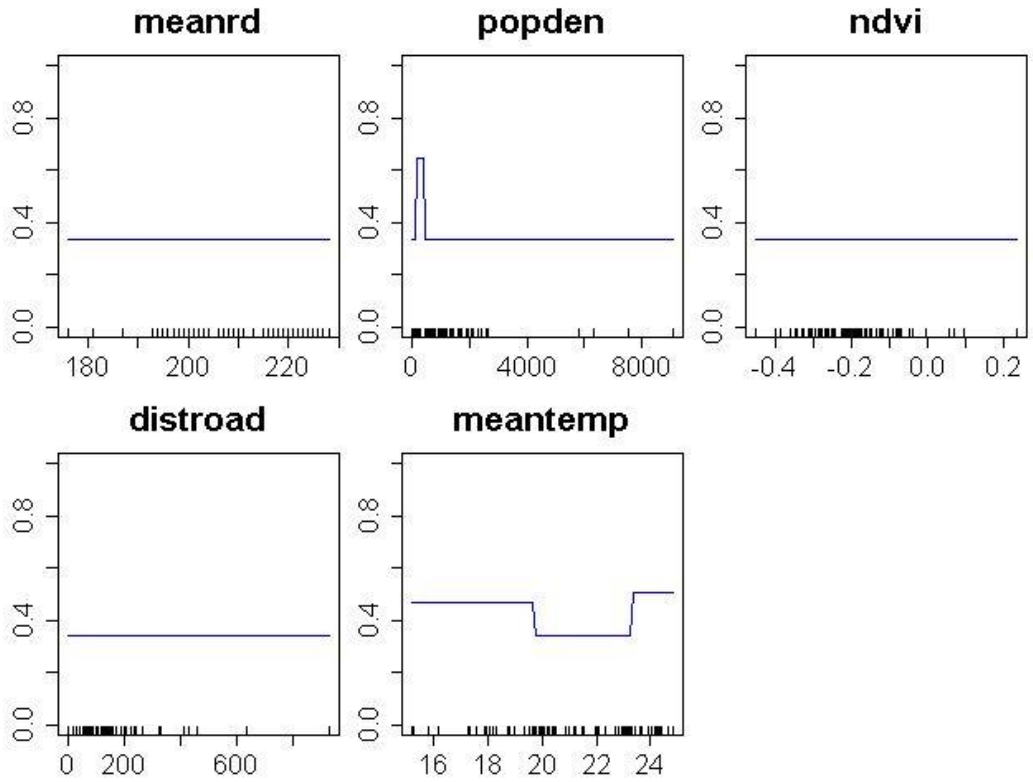


Figure 24. The response curves of predictors for *Stegomyia* estimations in generalized boosted model. Only human population density and mean temperature responded to the probability of presence for *Stegomyia*. The black tick marks on the x-axis imply observations.

In the random forest model, all predictors responded varyingly to the presence of *Stegomyia* (Figure 25). When mean radiation is at value of 220 Wh/m² or less, the probability of observing *Stegomyia* is low (≥20%). A surprising remark is when the rate of solar radiation rises (≥230 Wh/m²), the probability of observing *Stegomyia* increases, even up to 80%. This finding may imply something about the species and its capability to adapt to warm and dry conditions.

Similar findings to GBM are observed regarding the human population density factor. The locations with less than 1000 persons per km² are more favorable locations for *Stegomyia* (40-80%) than the locations with more than 1000 persons/km² (40%). *Stegomyia* is originally a forest species, but has later spread to human settlements. *Stegomyia* is more likely (80%) to be found in locations which are relatively poor in vegetation (NDVI ≤ -0.2) or with moderate vegetation (NDVI ≥0.1). In the locations where NDVI is between -0.2 and 0.1, *Stegomyia* is hardly detected (20%).

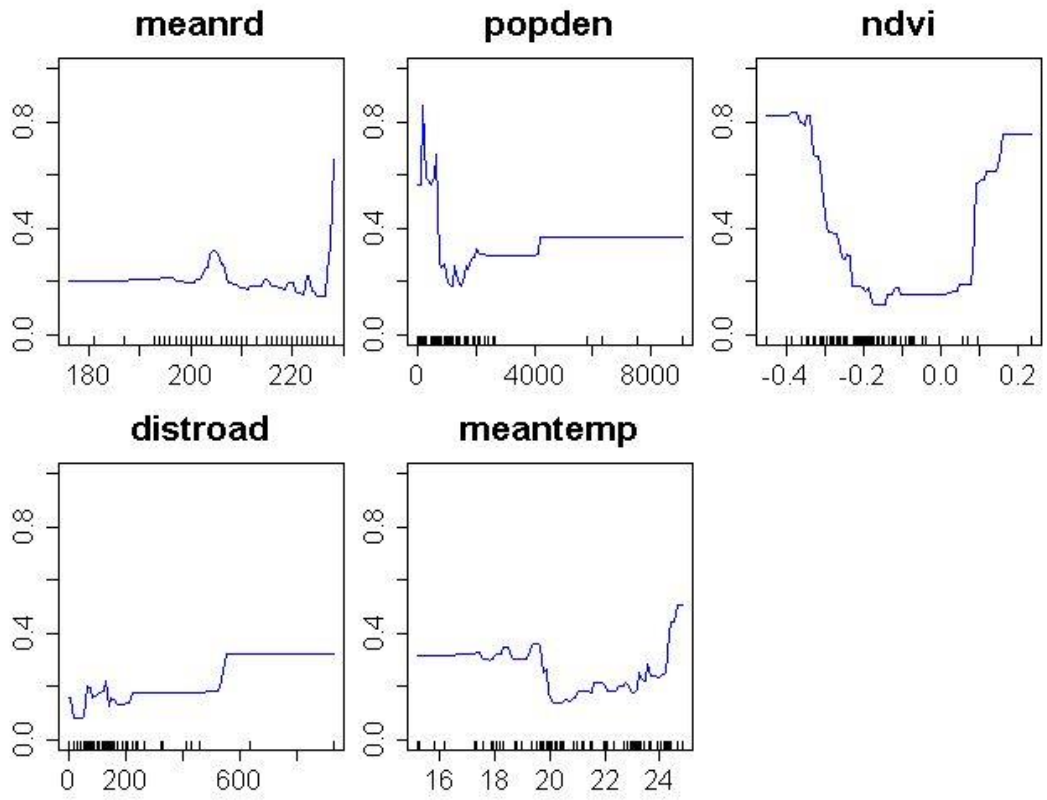


Figure 25. The response curves of predictors for *Stegomyia* estimations in random forest model. Each predictor variable responded to the probability of presence of *Stegomyia*. The black tick marks on the x-axis imply observations.

The variability of distance to roads had little influence on the probability of *Stegomyia*. In all distances between 0 and 600 meters from roads, *Stegomyia* could be detected with less than 30% probability. The changes in mean temperature slightly affect the probability of *Stegomyia* presence. Similarly to GBM, the most favorable locations for *Stegomyia* were those with temperature between 15Cº and 20 Cº, and over 23 Cº (≥ 30%). The areas with temperatures between 20 Cº and 23 Cº are not suitable locations for *Stegomyia*, as the probability of their presence decreases to less than 30%.

7.4 Evaluating the best model to estimate *Culex* and *Stegomyia* distributions

The aim of this study was to find the best model to estimate *Culex* and *Stegomyia* presence. First we will focus on evaluating the models estimating *Culex* distributions run by GLM and GAM models. Later, we concentrate on evaluating the GBM and RF models, which were best at estimating *Stegomyia* distributions.

In *Culex* estimations, the statistical models with highest AUC-value with significant p-values were GAM (AUC=0.791) and MARS (AUC=0.806) (Table 7). Altogether six of the total eight models proved reliable estimations (AUC≥0.7). Only CTA and Maxent resulted in models with poor AUC values (AUC≤ 0.7), and are as effective at estimating the presence of *Culex* as at random. In this case, we will focus on investigating GLM (AUC=0.730) and GAM models, as the former represents a traditional statistical model and the latter is a model utilizing more smoothing functions.

Table 7. AUC-, Kappa- and TSS values of all resulted models for *Culex* are shown below.

Model	AUC	Kappa	TSS
GLM	0.730	0.455	0.479
GAM	0.791	0.570	0.594
GBM	0.750	0.363	0.370
CTA	0.620	0.247	0.239
ANN	0.764	0.462	0.482
MARS	0.806	0.539	0.591
RF	0.729	0.363	0.352
Maxent	0.585	0.168	0.170

In *Stegomyia* estimations, only two models of eight resulted in a higher evaluation value than random (AUC ≥0.7). These were GBM and RF models, which are highly over-fitting models using smoothing functions (Table 8). A traditional GLM did not produce estimations at all, so they cannot be compared. GBM and RF both resulted in an AUC- value of 0.708, which is sufficient to produce reliable estimations for *Stegomyia* presence. Other models resulted in an AUC-value less than 0.7, indicating that they were assessing *Stegomyia* presence as effectively as at random.

Table 8. AUC-, Kappa- and TSS values of all resulted models for *Stegomyia* are shown below.

Model	AUC	Kappa	TSS
GLM	-	-	-
GAM	0.643	0.239	0.304
GBM	0.708	0.362	0.411
CTA	0.616	0.217	0.179
ANN	0.612	0.165	0.250
MARS	-	-	-
RF	0.708	0.409	0.500
Maxent	0.690	0.362	0.411

Furthermore, the most important parameters for the models were assessed. The GLM model for *Culex* predictions resulted in a deviance explained value (D^2) of 11%, which is not very high. GLM resulted in an Akaike information criterion (AIC) - value of 110.16. The GAM model resulted in D^2- value of 31%, which is more than twice higher than to that of GLM. A maximum degree of freedom value for the GAM model was -1, which indicates that fit was highly traditional, and not affected by overfitting. The important parameters for GLM, GAM and for RF models are more accurately introduced in Appendices (see Appendix 11, 12 &13).

7.5 The predictive maps of potential *Culex* and *Stegomyia* distributions

By utilizing the models described earlier, predictive maps for both *Culex* and *Stegomyia* were created. The predictive maps estimate the probability of presence of these genera over the study area. We first introduce predictive maps for *Culex* created by a traditional GLM model and GAM, which uses smoothing functions. A predictive map for *Culex* created by MARS model, which produced the highest AUC-value, is introduced in Appendix 14. Here, we also introduce predictive maps produced by GBM and RF models for *Stegomyia* distributions.

In the GLM model, the values of slope, population density, NDVI, distance to roads and elevation in the observed presence and absence points were interpolated to the entyre study area (Figure 26). In GLM estimation overall, the probabilities of detecting *Culex* in the Taita Hills were higher in mountainous areas than on the surrounding plateaus.

The probability of *Culex* presence were highest (80–100%) in the central Taita Hills and lowest in the surrounding plains (0–60%). The influence of two important predictors, human population density and NDVI, can be recognized on the map. The probability for *Culex* detection was higher (60–100%) close to the towns and villages of Mgange, Mwatate and Wundanyi. Contrarily, remote northern and western areas characterized by national parks and cropland were not the most favorable locations for *Culex*, according to the GLM model. In these locations, the probability of detecting *Culex* was less than 20%.

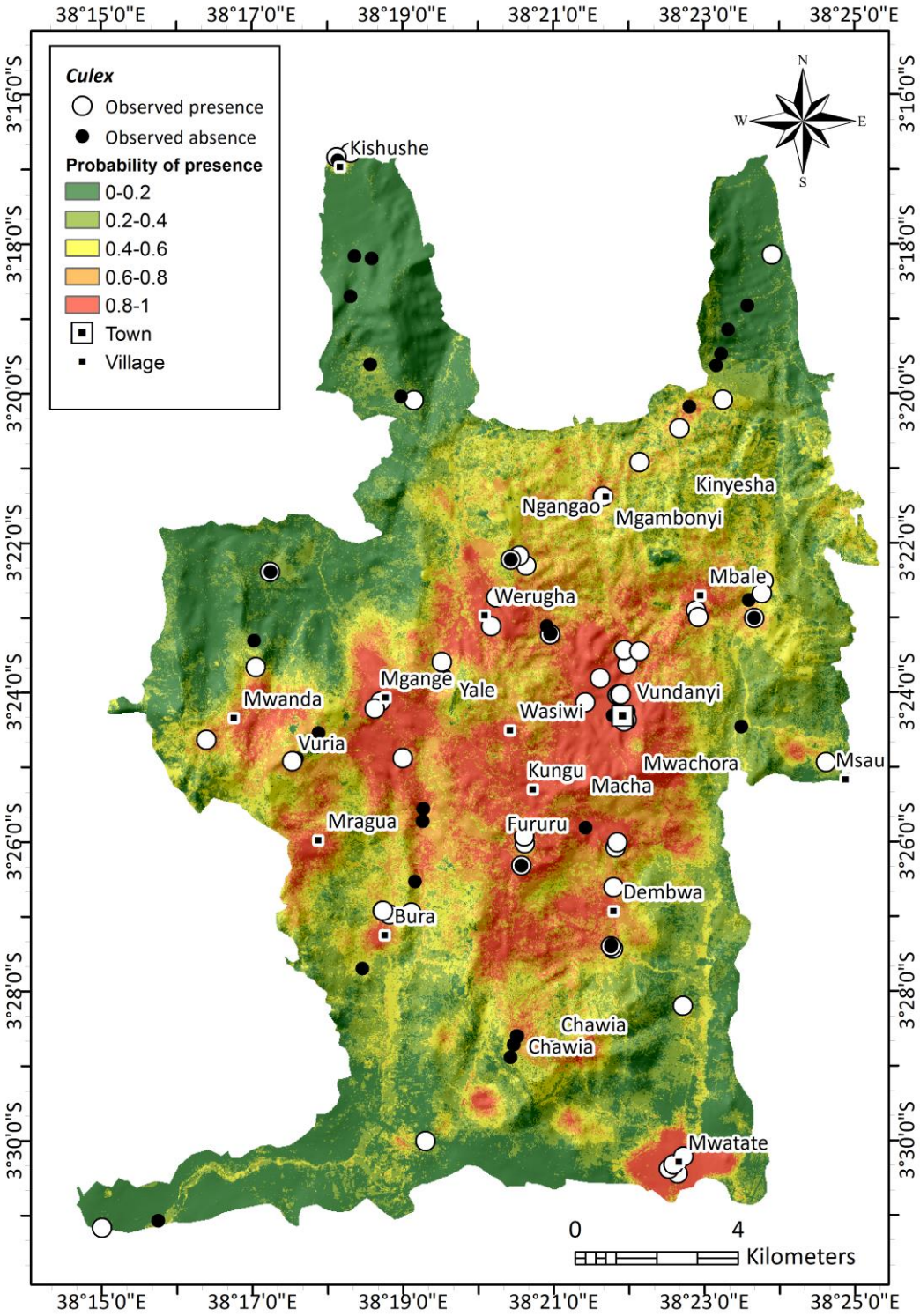


Figure 26. An influence of NDVI and human population density factors can be recognized in the GLM-based prediction map.

A GAM model produced a similar predictive map for *Culex* presence as GLM, but with a slightly different appearance (Figure 27). The gaps in the area probabilities were stronger in GAM. The village areas, forestall regions and the central mountainous Taita were shown as the areas with high probability rates for *Culex* detection (80–100%). In the Taita Hills, the

high-elevation locations also, were suitable habitats for *Culex*. On the plateaus apart from Mwatate village in the South, the probability of detecting *Culex* was only 0–20%. The moderate probabilities (20–60%) for *Culex* presence occurred in the remote areas close to the roads in both mountainous areas and on the plateau.

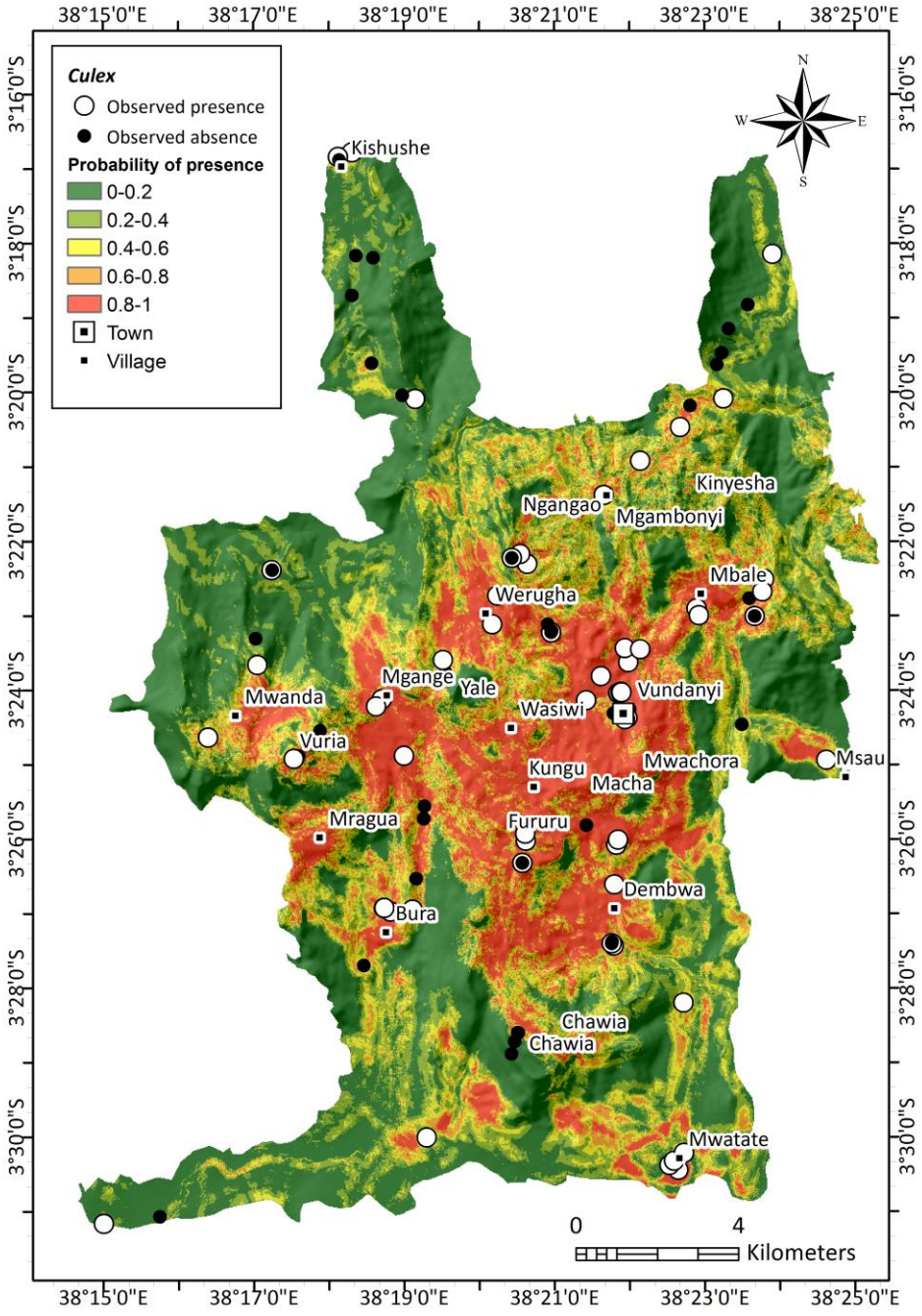


Figure 27. A GAM model estimated well the presence of *Culex*. The probability of *Culex* presence was highest (80–100%) in the central and southern Taita Hills. The lowest likelihoods for presence occurred on the surrounding plateaus.

From now on, we concentrate on analyzing the estimations for *Stegomyia* distributions in the Taita Hills. The estimations for *Stegomyia* presence in Taita differed widely from the areas

with high probability of *Culex* presence. A GBM model produced a predictive map which shows that the probability of detecting *Stegomyia* varied greatly in the study area (Figure 28).

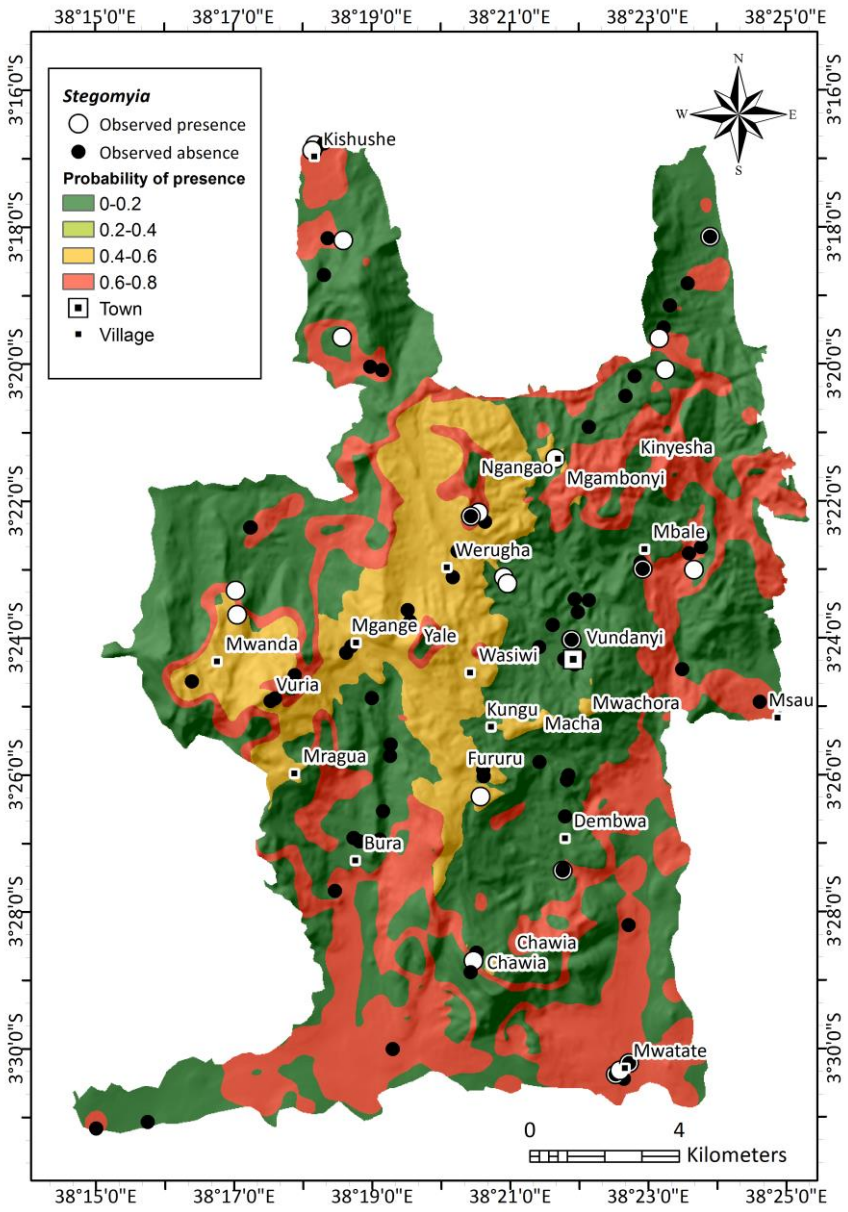


Figure 28. A GBM model estimated the presence of *Stegomyia*. The probability of presence was highest (60–80%) on the plateau. The lowest likelihoods for presence (0–20%) occurred at the high elevations.

On the plateau and close to main roads or railway, the probability of detecting *Stegomyia* was highest (60–80%). On the contrary, the lowest probability (≤20%) of finding *Stegomyia* occurred in deep forestal areas, such as in Ngangao and Chawia montane forests, and in Tsavo West national park. Furthermore, the probability of finding *Stegomyia* at high elevations (≥1800 m) was lower (0–60%) than on the plateau even though they were still favorable areas for occurrence.

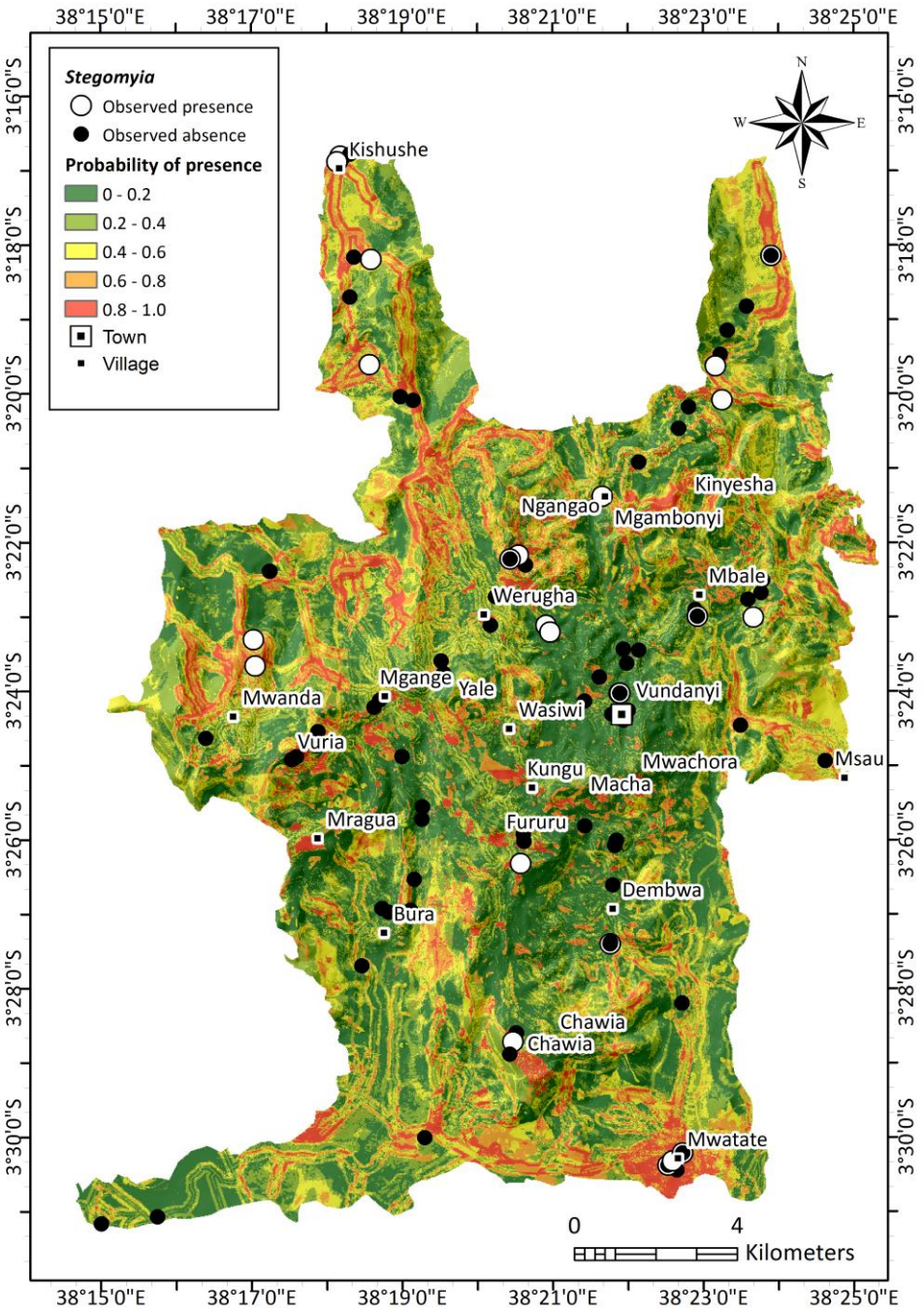


Figure 29. The random forest model estimated the presence of *Stegomyia*. The probability of presence was highest (80–100%) in many fragmented locations. This phenomenon verifies the *Stegomyia´*s ability to adapt to new habitats.

The predictive map produced by the random forest model differed considerably from GBM-based prediction. The locations with different probabilities vary highly (Figure 29). The locations where the probability of detecting *Stegomyia* is 80–100% mainly occurred in the villages on the plateau and close to the roads. The areas with moderate probability (40–80%) were situated sporadically. Wundanyi town center and national parks were the areas where the probability of *Stegomyia* presence was minimal (0–20%).

8. Discussion and conclusions

It was a great opportunity to model mosquito distributions over the Taita Hills, as it is a new location for such research. As a recognized hotspot of biodiversity in Kenya, it is also a potential source of unrecognized mosquito species. Furthermore, the outbreaks of mosquito-borne infections, mainly transmitted by *Culex*, *Stegomyia* and *Anopheles* mosquitoes, are currently topical in Kenya and in southeastern Africa in general (WHO 2016). The results of studying *Culex* and *Stegomyia* distribution gave us a greater understanding of the potential suitable habitats for these, and ideas how they react to the environment. This study could produce basic information that health officials can utilize in order to prevent diseases in the Taita Hills. In this chapter, we discuss the uncertainties relating to the study process and the possible explanations for the results. We also discuss the study achievements of this research and the possibilities for future studies on this subject.

8.1 Differences in the use of *presence-only* and *presence-absence* data

In our study, we used *presence-absence* data for statistical modeling. Advantages and disadvantages exist both in the use of *presence-only* and *presence-absence* data. *Presence-only* data is commonly used in models estimating species distributions. Absences are usually strong signs of biotic interactions, dispersal constraints and disturbances in order to preclude modeling of potential distributions (Svenning *et al.* 2004). The presence points often indicate many of the factors affecting absences, that are revealed e.g. in the situation when a species is absent from an environmentally suitable habitat because of past disturbances (Elith *et al.* 2011). Also, species prevalence can´ t be identified from *presence-only* data (Ward *et al.* 2009). In addition, sample selection bias has a stronger effect on *presence-only* models than on *presence-absence* models, when some areas are sampled more intensively than others (Elith *et al.* 2011). Due to these drawbacks of *presence-only* data, *presence-absence* data were used in this study. Recording true absence events adds to the depth and the breadth of insights gained from predictive modeling efforts (Drew *et al.* 2011). However, data on true absences is difficult to obtain, because surveys must be closed to individual movement and conducted such a way that no individual escapes detection (Drew *et al.* 2011). For example, wind conditions and temperature during collection may have affected in the situation of an absence event in this study.

8.2 Uncertainties relating to the collections and model process

Data uncertainties should be always incorporated in model predictions as environmental covariates affect the results and species distribution data proposes special challenges to all model types using observed data (Beale *et al.* 2011). There may be weather and other conditions affecting the results in the field, as well as uncertainties in the study process.

8.2.1 Conditions in fieldwork

Weather or climatic conditions are usually the major reasons for data biases in the field. Early on in the data collection, there was some uncertainty regarding the reasoning for the absence data. In some locations, mosquitoes were not collected, even after hours of searching for them. The reasons for their absence may include the wrong time of the day due to the heat, or temporary weather conditions such as a strong wind, localized differences in rain fall or humidity which may affect larval habitats, or the time of the year. In addition to this, collections was begun daily at 1 pm and generally finished after darkness at 7 pm. The afternoons were characterized by hot sunshine, which was a reason for not collecting adults during the first few hours. Nevertheless, it was possible to collect larvae were despite the heat. As the collection day progressed, more adult mosquitoes were observed. Despite there being more mosquitoes in the evening, we had to stop the collections for practical reasons as it was not appropriate to visit homes after darkness without arousing suspicions.

Seasonal dynamics of vector populations and the frequency of blood feeding are also dependent on temperature and precipitation (Drew *et al.* 2011). Two rainy seasons and two dry seasons occur annually in Kenya. Long rains occur from March until July and short rains from October to December. According to the residents in Taita, mosquitoes are prevalent almost everywhere during the rainy season. Mosquitoes were collected from late January until mid-March (the dry season), with only a few rainy days. Thus, an earlier timeframe could have enabled a much larger collection, which may have positively affected the sample size. If collections would have been implemented for example over a five year period twice a year during both rainy and dry seasons, the results may be more reliable, and therefore, more significant for public health decisions.

8.2.2 Spatial autocorrelation (SAC) of predictor variables

Spatial autocorrelation (hereafter SAC) was tested through predictor variables of *Culex* and *Stegomyia*. Moran´s Index was used to verify the potential spatial autocorrelation. Among both *Culex* and *Stegomyia* predictors, environmental, anthropogenic and distance variables were autocorrelated to some extent. Fifteen spatial distance classes were selected in order to reveal distances where variables are autocorrelated spatially.

At first, the SAC of *Culex* predictors was tested. This proved that human population density and elevation were highly autocorrelated (Moran´s I ≥ 0.8) with significant p-values (<0.05) for short distances (Figure 30). For longer distances, population density and elevation were no longer autocorrelated. NDVI and distance to roads were moderately autocorrelated (Moran´s I ≤0.6) for short distances from the collection sites. Slope was the only predictor of *Culex*, which resulted in only slight autocorrelation in space.

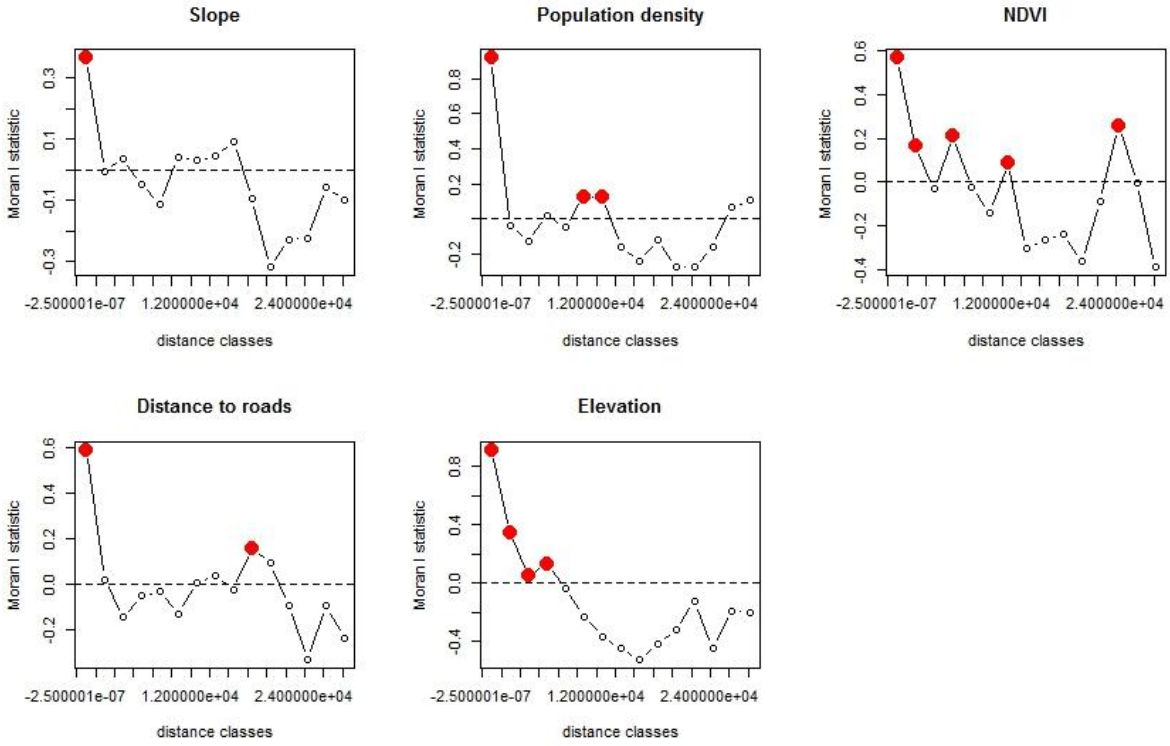


Figure 30. Spatial autocorrelation of slope, population density, NDVI, distance to roads and elevation. Population density and elevation were highly autocorrelated variables in the short distances but not in the longer distances. Slope was slightly autocorrelated for short distances as well as NDVI and distance to roads. Red rounds indicated the significant p-value (p <0.05) and were located at distances where variable was autocorrelated.

Secondly, we focused on analyzing the potential spatial autocorrelation of *Stegomyia* predictor variables (Figure 31). As mentioned earlier, human population density was highly autocorrelated in short distances but not in longer distances. In addition, mean temperature resulted in high Moran`s I values (≥0.8) in short distances, but autocorrelation totally vanished when distance was on the increase. NDVI and distance to roads were moderately autocorrelated (Moran´s I = 0.6) in short distances, but spatial autocorrelation decreased when distance increased. Mean radiation was hardly autocorrelated (Moran´s I ≤ 0.25) for both short and long distances from the occurrence location.

According to these findings, we can state that all other predictor variables of *Culex* and *Stegomyia* resulted in either moderate or high positive autocorrelation values at very short distances, apart from slope and mean radiation variables. Only these two variables indicated that they were not spatially dependent on each other either at short distances or long distances from the collection locations.

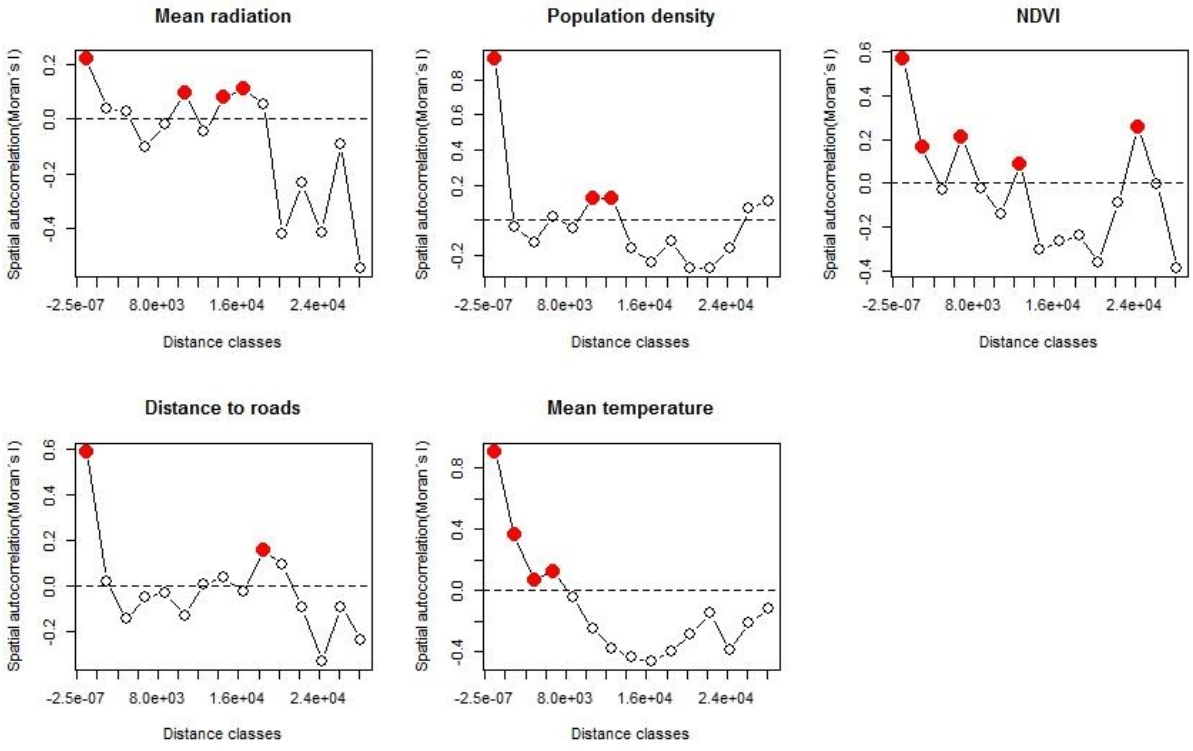


Figure 31. Spatial autocorrelation of mean radiation, population density, NDVI, distance to roads and mean temperature. Population density and temperature were highly autocorrelated in the short distances but not for the longer distances. Mean radiation was very little autocorrelated for short distances. Distance to roads and NDVI were slightly autocorrelated for short distances. Red rounds indicated the significant p-value ($p<0.05$) and were located at distances where variable is autocorrelated.

Nearly all predictor variables, excluding mean radiation and slope variables, were spatially autocorrelated for short distances but not for longer distances. This may have brought about bias or distortion in the results. In this study, bias could mainly be found in the estimations of neighboring areas.

8.2.3 Other uncertainties in model process

Uncertainties also exist regarding the modeling framework. Simple models usually ignore historical events and temporal processes, despite their impact on the structure and the function of present day landscapes (Bürgi *et al.* 2004; Rhemtulla *et al.* 2007; Gillson 2009). Furthermore, many models ignore species interactions due to their dynamic nature or poor representation in mapped form, which excludes the complexity of potential community level efforts such as competition on the distribution and abundance patterns of the species (Drew *et al.* 2011). Another limitation to empirical ecological modeling is that correlation doesn´t always imply causation (Drew *et al.* 2011). The models also assume that species that are modeled are in equilibrium with their environments, and that models are static and are not able to account for dispersal (Drew *et al.* 2011). This uncertainty relates to the accuracy of environmental and other data used as explanatory variables. Mosquito habitats vary due to climate change, and this is why environmental data must be updated often enough. Some of the satellite imagery data were from 2011 but some, e.g. NDVI and vegetation maps, were updated last year.

We can also question the reliability of our modeling results for a few reasons. In the beginning, we had twelve anthropogenic, distance, and environmental factors determining the distributions of *Culex* and *Stegomyia*. Later on, we selected five not-highly-correlated factors for the final model. This means that a majority of factors were excluded from the model. Thus, there certainly exist other predictors, also outside the twelve, which affect the species distributions of these two mosquito genera.

Furthermore, select environmental and other variables accurately predicted the distribution of *Culex* (AUC $\geq$0.7) in a majority of models, including more traditional models that do not use smoothing functions such as GLM. Also, machine-learning models and the models using smoothing functions obtained accurate results for *Culex*. Thus, we can argue that predictive

maps give quite reliable estimations for *Culex* distributions. For *Stegomyia* however, often only highly overfitting models (GBM and RF) resulted in AUC scores high enough for reliable prediction. GLM didn´t obtain reliable AUC values for *Stegomyia,* so we could not compare the model outputs.

In addition to the potential uncertainties stated earlier, there also exist questions regarding the use of results as a straight linkage to the estimations of MBD distributions. Regarding modeling for virus patterns by the distributions of virus vectors *Culex* and *Stegomyia,* we cannot make very reliable predictions for dengue, West Nile virus or chikungunya distributions in the Taita Hills for several reasons. *Culex* and *Stegomyia* include hundreds of subgenera and dozens of species of which only some are vectors of mosquito-borne infections. Furthermore, even if the estimations for *Culex* and *Stegomyia* distributions are reliable, not all mosquitoes of those vector species are carriers of viruses.

8.3 Notes about the mosquito genera of the Taita Hills

We confirmed seven mosquito genera from the Taita Hills, including three main MBD vectors *Culex, Stegomyia* and *Anopheles* in addition to other genera, *Uranotaenia, Eretmapodites, Lutzia* and *Aedimorphus*. Only *Culex* and *Stegomyia* resulted in large enough collections for modeling purposes. *Culex* was collected from a variety of environments, which strengthens its´ characteristics as a widely distributed mosquito genus, in general.

Stegomyia is linked to transmission of a large number of mosquito-borne viruses. In the Taita Hills, a great amount of *St. aegypti* larvae were observed in high numbers in car tyres in Mwatate village. This is an interesting observation, because car tyres are one of the main methods of *Stegomyia* dispersal worldwide. They mainly spread from Asia in used car tyres via international transportation, and are capable of withstanding extremely dry and warm conditions for years (WHO 2017c). *Stegomyia* is not often found at high-elevations as it has temperature-based limits to survival. In temperatures below 14°C, *St. aegypti* suffers from reduced mobility and capability to suck blood (Brady *et al.* 2013). In the Taita Hills, *Stegomyia* was collected at surprisingly high-elevations, above 1700 meters. Even though *Stegomyia* originated in tropical forests, we found both *Stegomyia* adults and larvae in the

Ngangao montane forest at an elevation of 1800 meters. These mosquitoes found in Ngangao, have not yet been identified to species, but will be in due course.

Lutzia tigripes is not known to act as a vector of known parasites or viruses, thus is considered to be harmless to humans. On the contrary, *Aedimorphus, Eretmapodites* and *Uranotaenia* are potential vectors transmitting Rift Valley fever virus, West Nile virus or yellow fever virus. *Aedimorphus* and *Eretmapodites* did not result great collection size during the fieldtrip; thus we can note that they may not be widely spread in the Taita region.

In addition to these observations mentioned earlier, we can note a few findings of *Anopheles*. *Anopheles,* although it is a genus containing the only human malaria vector, resulted in four occurrence locations from our total 122 collection sites in the Taita Hills. This is an extremely low number of observations, and was a rather unexpected outcome. The majority of *Anopheles* collections occurred in the plateau areas which strengthens the assumption that highlands with low temperatures are not favorable habitats for their eggs to survive (Afrane *et al.* 2012). Nevertheless, a few small collections were found in the mountain area in Wundanyi. This outcome may confirm why health officials have not diagnosed malaria cases; instead, they have diagnosed positive samples of chikungunya and dengue viruses from the antibodies of patients in Taita, as currently on-going research will explore.

Based on the collections, *Uranotaenia*, *Eretmapodites* and *Aedimorphus* were not widely distributed in the Taita Hills; thus these are not the highest threat regarding the transmission of MBDs in the area. The greatest risk of spreading MBDs is caused by *Culex* and *Stegomyia* genera in the Taita Hills. The 300 unrecognized mosquito specimens from 19 locations are of continuing interest to other researchers, since these specimens may result in new mosquito distributions to the area, and there is always a chance to find species as yet unknown to science from remote areas which have previously not been studied.

8.4 Influential factors for *Culex* and *Stegomyia*

Anthropogenic and distance factors appeared to be even more important factors than environmental drivers for *Culex* and *Stegomyia* distributions in a majority of models. This strengthens the fact that they are capable of adapting to new habitats modified by humans and

human activity, such as land use changes and urbanization. *Culex* mosquitoes prefer locations with high population densities located at close to roads and the locations with moderate NDVI values and low slope angles. The distribution of *Culex* was not dependent on elevation, which is an interesting outcome and may explain part of its widespread distribution. In contrast, *Stegomyia* prefer sites with lower human population densities and higher distances to roads, as well as high temperatures and solar radiation, in addition to either very poor or rich vegetation. These outcomes tell something of its origin as a forest species, and this finding may also strengthen the assumption of its amazing capability to stand hot and dry weather conditions for survival (Brady *et al.* 2013).

Human population density, mean temperature and NDVI were direct factors affecting *Culex* and *Stegomyia* presence. Slope, distance to roads, solar radiation and elevation presented as indirect factors affecting *Culex* and *Stegomyia* distributions. For example, in locations with high slope angles (º), weather conditions are usually windier than in locations with low slope angles. In this case, wind is the main factor affecting the presence and absence of mosquitoes. Also, elevation is indirect factor, as at high elevations, factors of precipitation and temperature are the main influential drivers in mosquito presence and absence.

Environmental factors, along with the variable of human population density, have widely been used as predictors in the studies related to mosquito distribution modeling (Ibañez-Justicia *et al.* 2015; Sallam *et al.*2016; Fatima *et al.*2016) We included additional variables, distance to houses and distance to roads in the models; of these, distance to roads was influential factor almost in all models in both *Culex* and *Stegomyia* estimations. This is an important driver to pay attention to in potential further research, as the distribution of *Culex* and *Stegomyia* are strongly affected by human mobility.

8.5 Model validity or incompetence

Our results affirm the utility and reliability of the use of the biomod2 package in R as a valid modeling method for species distributions. Several models accurately estimated (AUC ≥0.7) the distribution of both *Culex* and *Stegomyia* (Figure 32).

An inclusion of GLM, GAM, GBM and RF models gave different perspectives on modeling and predictive map performances. Our results reject the null hypothesis regarding the unusability of these algorithms to model mosquito distributions, since the models with these four techniques had very good accuracy on their AUC scores. Different environmental and anthropogenic variable contributions easily resulted high AUC scores for *Culex*; however, only a model with contributions of mean radiation, NDVI, distance to roads, human population density and mean temperature resulted AUC scores high enough to produce accurrate estimations for *Stegomyia*. The reason for this may also be a different number of presence and absence points in the occurrence data as the *Culex* model was run with 73 observed presence and 49 true absences, while the *Stegomyia* model was run with 28 observed presence and 42 true absence points.

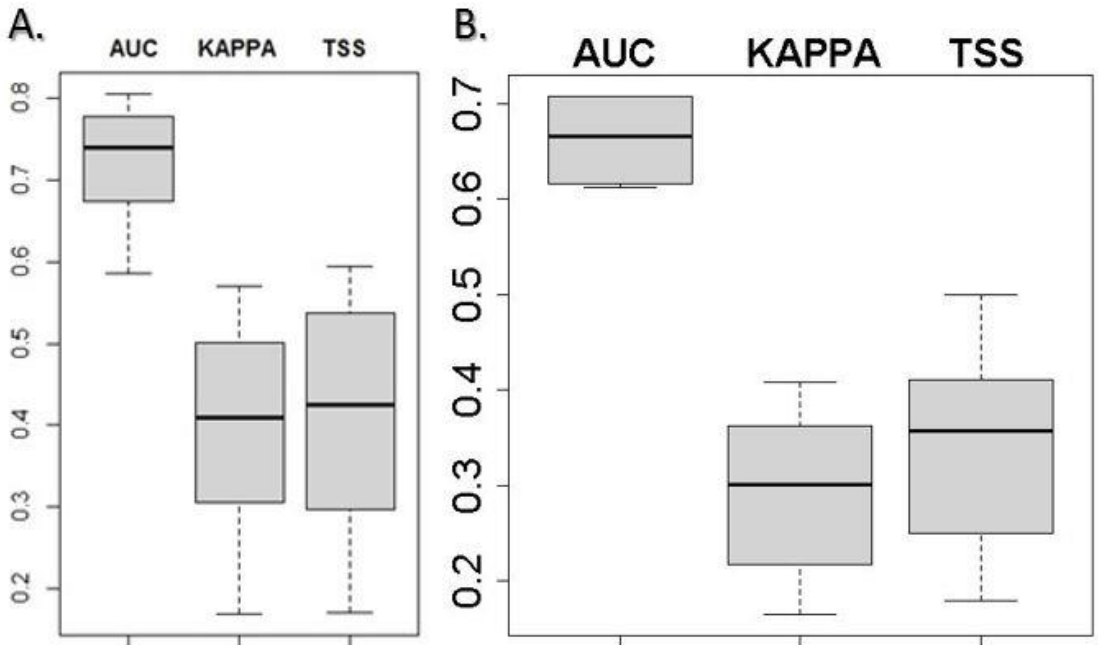


Figure 32. A. The distribution of prediction accuracy for *Culex*. A majority of the models accurately estimated (AUC ≥0.7 or κ ≥0.4 or TSS ≥0.4) the distribution of *Culex* apart from few residuals. B. The division of prediction accuracy for *Stegomyia* differs from the left Figure, as a majority of models didn´t estimate *Stegomyia* accurately (AUC ≤0.7) apart from generalized boosted regression model and random forest model.

8.6 A potential new predictor for mosquito distribution modeling

We recognized a potential new predictor variable influencing the presence of mosquitoes. A distinction occurred in the abundance of mosquitoes in different building designs (Figure 33). A majority of large mosquito collections were found in schools or in other buildings of modern design. In general, mosquitoes were rarely collected in houses built by traditional

methods. Modern, airtight buildings appear to be favorable breeding sites for mosquitoes, as the humidity remains inside the buildings where no appropriate ventilation system exists. On the contrary, in the buildings built by traditional methods, air gaps exist from which moisture can evaporate away.

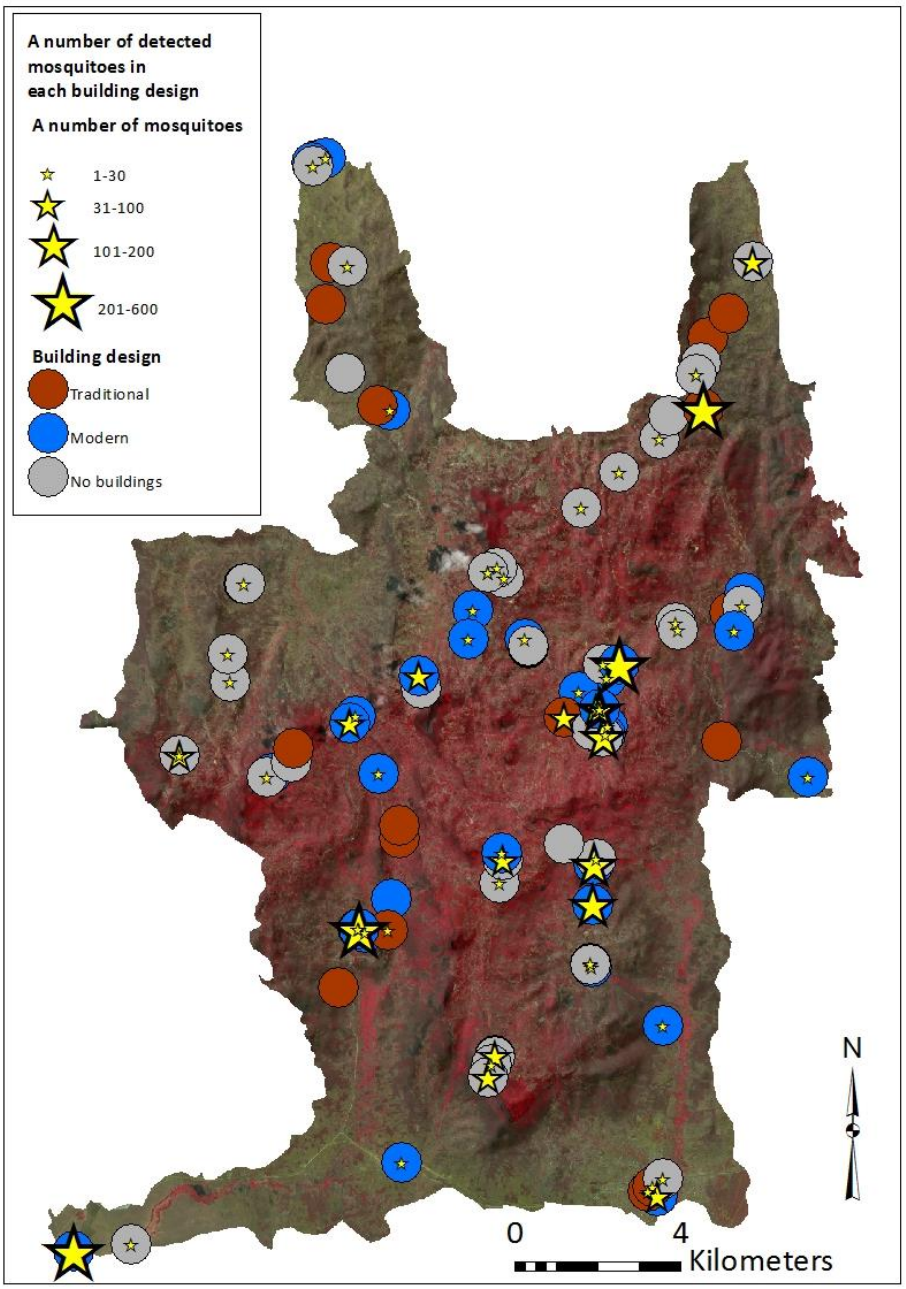


Figure 33. The number of observed mosquitoes in each building design. The majority of large collections were implemented in modern buildings. Buildings with traditional design were not favorable occurrence sites for mosquitoes.

Mosquitoes and their association with building design has not yet been investigated, but a study on a similar theme was conducted in southwestern USA. A contrary argument to our

finding exists concerning the detection of *Stegomyia aegypti* eggs in the old houses. Higher numbers of *St. aegypti* eggs were observed in older homes than in modern houses, according to Walker *et al.* (2011). Thus, the older homes were asserted to be associated with *St. aegypti* abundance. Even if the main study perspective is similar, we cannot compare their study results to ours for several reasons. First, climate in southeastern Africa differs from the climate in southwestern USA. Also, the classification of older homes and modern houses is also totally different than that in Kenya. Furthermore, even though the houses mentioned in the study are old, there may have been modern air conditioning systems, which was not the case in Kenya.

Because the Taita Hills region is characterized by a humid and warm climate, it provides suitable conditions for mosquitoes to breed inside modern buildings, where humidity lingers due to the lack of air gaps or the absence of ventilation system. This fact should be taken into account in the prevention of mosquito-borne infections.

8.7 Potential distributions of *Culex* and *Stegomyia* in the Taita Hills

In the Taita Hills, the distribution of the West Nile virus vector *Culex* is widespread, excluding the surrounding savannas located at the national parks and the croplands in north. *Culex* were collected at all elevations in Taita, and particularly in the Wundanyi town and Mwatate village, which are the largest population centers in the region. This confirms that humans are important hosts for *Culex* species, and explains why the probabilities to find them in national parks are lower than in villages.

The distribution of the dengue virus and chikungunya virus vector *Stegomyia* is fragmented in the Taita Hills. Especially the main roads of Taita and Mwatate region are possible locations for observing *Stegomyia*. The northern and southern plateau areas are also suitable locations for *Stegomyia*, with high temperatures and solar radiation values as well as intermediate population densities. The Wundanyi town area is not a suitable area for *Stegomyia* distribution, and neither are the areas in Taita which are located at high elevations, unlike with *Culex*.

8.8 Conclusion

The importance of studying distribution modeling of mosquitoes lies in the fact that more outbreaks of mosquito-borne infections have been revealed around in the world, including in Africa. Our results proved that mosquito distributions can reliably be modeled by the biomod2 package in R, resulting in an insight into vector-ecological interactions on local, regional and global scale. By defining the suitable habitat and potential distributions of vectors of mosquito-borne diseases, brings more ideas of where and how to concentrate on the disease intervention strategies. An important finding was the link between building design and abundance of mosquitoes. This argument still requires further study, but this consideration could be utilized in the prevention of mosquito-borne infections, already in the construction phase of buildings.

With our study results, general assumptions can be made about the distribution of main West Nile virus, dengue virus, Zika virus and chikungunya vectors *Culex* and *Stegomyia* in the Taita Hills. After molecular identification is completed, thus exact species will be recognized, and virus isolations will be completed, it is possible to more accurately model the distribution of virus vectors over the Taita Hills. Additional interest lies among unrecognized mosquito collections, which may result in new mosquito species being discovered. These issues, among others, may bring new opportunities for future research.

Acknowledgements

I gratefully thank everybody who participated to this extremely interesting study process. I express my gratitude to Mika Siljander, who was a great supervisor, always ready to help and advice in all the issues regarding the thesis. I thank prof. Petri Pellikka, who opened up this opportunity for me to participate in the project. I am thankful for the other supervisor; Lorna Culverwell, from the Department of Virology, who taught me so much more about mosquitoes and other virus vectors, their morphology, habitats and the public health concern. I also thank Kristian Forbes for being my supervisor; encouraging and supporting me, particularly, with the public health concern. I am thankful to everybody involved in *the Wildlife screening*-project in Kenya and Finland; Essi Korhonen, Kristian Forbes, Lorna Culverwell, Joni Uusitalo, Olli Vapalahti, Masika Moses and Eili Huhtamo, who taught me a plenty of new matters about rodents, mosquitoes and bats, and about their biology and connections to virus transmission. I really loved to work with them. Also, I would like to thank prof. Miska Luoto, who organized the course regarding the spatial analysis and modeling in R, which brought me the skills to model mosquito distributions using R. I am thankful to Juha Aalto, who gave me final advice for the use of biomod2 methods. Moreover, I would like to thank Sakari Keipi who helped me with the language, and Sakari Äärilä and Ninna Malinen who were my mental and technical support at the university. I am so grateful to them for their company and all the advice they gave me. Finally, I also would like to thank my family for all the support and encouragement they gave me during the study process. All of them, had a significant impact on this process.

References

Afrane1, Y., Githeko1, A. & Y. Guiyun (2012). The Ecology of Anopheles Mosquitoes under Climate Change: Case Studies from the Effects of Environmental Changes in East Africa Highlands. *Annals of the New York Academy of Sciences* 1249: 204–210.

AMCA (2014). American Mosquito Control Association. Mosquito info. 15.4.2016. <http://www.mosquito.org/mosquito-info>

Anselin, L., Florax, R., & S. Rey (2004). Advances in Spatial Econometrics: Methodology, Tools and Applications.

Ara´ujo, M. & M. New (2007). Ensemble forecasting of species distributions. *Trends in Ecology & Evolution* 22: 42–47.

Attaway, D., Jacobsen, K., Falconer, A., Manca, G., Bennett, L. & N. Waters (2014). Mosquito habitat and dengue risk potential in Kenya: alternative methods to traditional risk mapping techniques. *Geospatial health* 9(1): 119-130.

Austin, M. (1971). Role of regression analysis in plant ecology. *The Proceedings of the Ecological Society of Australia* 6: 63-75.

Austin, M. (1980). Searching for a model for use in vegetation analysis. *Vegetatio* 42: 11–21.

Austin, M. (2002). Spatial prediction of species distribution: an interface between ecological theory and statistical modeling. *Ecological Modeling* 157: 101–118.

Barnhart, P. & E.Gillam (2014). The impact of sampling method on maximum entropy species distribution modeling for bats. BioOne. *Acta Chiropterologica* 16(1): 241-248.

Barry, S. & J. Elith (2006). Error and uncertainty in habitat models. *Journal of Applied Ecology* 43: 413–423.

Beale, M. & J., Lennon (2011). Incorporating uncertainty in predictive species distribution modeling. *Philosophical Transactions of the Royal Society* 367: 247–258.

Becker, N., Petric, D., Zgomba, M., Boase, C., Madon, M., Dahl, C. & A. Kaiser (2010). Mosquitoes and Their Control. The biology of mosquitoes. 2nd edition, 565 p.

Benediktsson, J., Swain, P. & O. Ersoy (1993). Conjugate-gradient neural networks in classification of multisource and very-high-dimensional remote sensing data. *International Journal of Remote Sensing* 14: 2883–2903.

Bhatt, S., Gething, P., Brady, O., Messina, J., Farlow, A. & C. Moyes (2013). The global distribution and burden of dengue. *Nature* 496: 504-507.

Brady, O., Johansson, M., Guerra, C., Bhatt, S., Golding, N., Pigott, D., Delatte, H., Grech, M., Leisnham, P., Maciel-de-Freitas, R., Styer, L. & D. Smith (2013). *Parasites & Vectors* 6: 351.

Breiman, L., Friedman, J., Olshen, R., & C. Stone (1983). *Classification and regression trees.* Wadsworth, London.

Broberg, A., Salminen, H., Tolvanen, R. & J.Ylhäisi (2004). Werugha - village in the heart of the Taita Hills. p.108-113. In: Pellikka, P., J. Ylhäisi & B. Clark (eds.) Taita Hills and Kenya, 2004 –seminar, reports and journal of a field excursion to Kenya. *Expedition reports of the Department of Geography*. University of Helsinki. 148 pp.

Brown. J., Evans, B., Zheng, W., Obas, V., Barrera- Martinez, L., Egizi, A., Zhao, H., Caccone, A. & J. Powell (2014). Human impacts have shaped historical and recent evolution in *Stegomyia aegypti*, the dengue and yellow fever mosquito. *Evolution* 68(2): 514–525.

Burnham, K. & D. Anderson (1998). Model Selection and Multimodel Inference: A Practical Information – Theoretic Approach. New York, USA.

Bürgi, M., Hersperger, M., & N. Schneeberger (2004). Driving forces of landscape change – current and new directions. *Landscape Ecology* 19: 857–868.

CIA (2016). Central Intelligence Agency. Kenya. The World Factbook October 2016. 20.4.2016.<https://www.cia.gov/library/publications/the-world-factbook/geos/ke.html>

CDC (2012). Centers for Disease Control and Prevention. Dengue and Climate. 27.9.2016. <//www.cdc.gov/dengue/entomologyEcology/climate.html>

Cohen, Jacob (1960). A coefficient of agreement for nominal scales. *Educational and Physiological Measurement* XX: 1.

Clark, B. & P. Pellikka (2005). The development of a land use change detection methodology for mapping the Taita Hills, South-East Kenya. Proceedings of the 31st *International Symposium of Remote Sensing of the Environment*. 20-24 June 2005. St. Petersburg, Russia. CD-Rom publication.

Conley, A., Fuller, D., Haddad, N., Hassan, A., Gad, A. & J. Beier (2014). Modeling the distribution of the West Nile and Rift Valley Fever vector Culex pipiens in arid and semi-arid regions of the Middle East and North Africa. *Parasites & Vectors* 7: 289.

Cressie, N. (1991). Statistics for Spatial Data. New York: John Wiley and Sons. 920 pp.

Culverwell, Lorna (2016). Personal Communication. Mosquito Protocol.

De'ath, G. & K. Fabricius (2000). Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology* 81: 3178–3192.

Diniz- Filho, J., Bini, L. & B. Hawkins (2003). Spatial autocorrelation and red herrings in geographical ecology. *Global Ecology & Biogeography* 12: 53–64.

Dormann, C., Purschke, O., Garc´ıa-Marquez, J., Lautenbach, S. & B. Schroder (2008).Components of uncertainty in species distribution analysis: a case study of the Great Grey Shrike. *Ecology* 89: 3371–3386.

Drew, C., Wiersma, Y. & F. Huettmann (2011). Predictive Species and Habitat Modeling in Landscape Ecology. 319 p. The United States.

Elith, J. & J. Leathwick (2009). Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics* 40: 677–697.

Eisen, L., Bolling, B., Blair, C., Beaty, B. & C. Moore (2008). Mosquito Species Richness, Composition, and Abundance along Habitat-Climate-Elevation Gradients in the Northern Colorado Front Range. *Journal of Medical Entomology* 45(4):800-811.

Elith, J., Phillips, S., Hastie, T., Dudik, M., Chee, Y. & C. Yates (2011). A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions* 17: 43–57.

Environment Systems Research Institute (ESRI) (1991). ARC/INFO User's guide. Cellbased modelling with GRID. Analysis, Display and Management (Redlands, CA: ESRI).

Erdogan, E.H., Pellikka, P. & B. Clark (2011). Modelling the impact of land-cover change on potential soil loss in the Taita Hills, Kenya, between 1987 and 2003 using remote-sensing and geospatial data. *International Journal of Remote Sensing*, to be published.

Fatima, S., Atif, S., Rasheed, S., Zaidi, F. & E. Hussain (2016). Species distribution modeling of *Aedes aegypti* in two dengue-endemic regions of Pakistan. *Tropical Medicine and International Health* 21: 427-436.

Fielding, A. & J. Bell (1997). A review of methods for the assessment of prediction errors in conservation presence/ absence models. *Environmental Conservation* 24: 38–49.

Franklin, J. & J. Miller (2010). Statistical methods – modern regression. In: Franklin, J.(ed.) Mapping species distribution: spatial inference and prediction. Cambridge: Cambridge University Press. 340 p.

Friedman, J (1991). Invited paper: Multivariate adaptive regression splines. *The Annals of Statistics* 19(1):1-141.

Gamaa, M., Crespo, D., Dolbeth, M. & P. Anastácioa (2016). Predicting global habitat suitability for *Corbicula fluminea* using species distribution models: The importance of different environmental datasets. *Ecological Modeling* 319: 163–169.

Gillson, L. (2009). Landscapes in time and space. *Landscape Ecology* 24:149–155.

Guisan, A., Lehmann, A., Ferrier, S., Austin, M., Overton, J., Aspinall, R. & T. Hastie (2006). Making better biogeographical predictions of species. *Journal of Applied Ecology* 43: 386-392.

Guisan, A. & W. Thuiller (2005). Predicting species distribution: offering more than simple habitat models. *Ecology Letters* 8: 993-1009.

Heikinheimo,Vuokko (2015). Impact of land change on aboveground carbon stocks in the Taita Hills. Master thesis. Geography. Geoinformatics. University of Helsinki.

Heikkinen, R., Luoto,M., Araújo, M., Virkkala,R., Thuiller, W. & M.Sykes (2006). Methods and uncertainties in bioclimatic envelope modeling under climate change. *Progress in Physical Geography* 30 (6): 751–777.

Heiskanen, J., Korhonen,L., Hietanen, J., Heikinheimo, V., Schöfer, E. & P. Pellikka (2015). Comparison of field and airborne laser scanning based crown cover estimates across land cover types in Kenya. The International Archives of the Photogrammetry. *Remote Sensing and Spatial Information Sciences*: XL-7/W3.

Hijmans, R. & J. Elith (2016). Species Distribution Modeling with R. January 2016. <https://cran.r-project.org/web/packages/dismo/vignettes/sdm.pdf>

Hijmans, R. & J. Elith (2016). Species distribution modeling with R. 79 p.

Huang, Yiau-Min (2001). A pictorial key for the identification of the subfamilies of Culicidae, genera of Culicinae, and subgenera of *Aedes* mosquitoes of the Afrotropical Region (Diptera: Culicidae). *Proceedings of the Entomological Society of Washington* 103: 1-53.

Huang, Yiau-Min (2004). The subgenus *Stegomyia* of *Aedes* in the Afrotropical Region with keys to the species (Diptera: Culicidae). *Zootaxa* 700(1): 1-120.

Hutchinson, M. (1995). Interpolating mean rainfall using thin plate smoothing splines. *International Journal of Geographical Information Science* 9: 305–403.

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

Junglen, S., Kopp, A., Kurth, A., Pauli, G., Ellerbrok, H., & F. Leendertz (2009). A New Flavivirus and a New Vector: Characterization of a Novel Flavivirus Isolated from *Uranotaenia* Mosquitoes from a Tropical Rain Forest. *Journal of Virology* 83(9): 4462–4468.

Kadmon, R., Farber, O. & A. Danin (2004). Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecological Applications* 14:401–413.

Kaplan, I., Dobert, M.K., Marvin, B.J., McLaughlin, J.L. & D.P. Whitaker (1976). Area Handbook for Kenya. 472 p. Foreign Area Studies (FAS).The American University, Washington D.C., U.S.A.

Kienast. F., Bolliger, J. & N.E. Zimmermann (2012). Species Distribution Modeling (SDM) with GLM, GAM and CART. *Advanced Landscape Ecology* 1-16. 31.12.2012.

Kumar, L., Skidmore, A.K. and Knowles, E., 1997. Modelling topographic variation in solar radiation in a GIS environment. *International Journal for Geographical Information Science* 11(5): 475-497.

Kwon, Y., Bae, M., Chung, N., Lee, Y., Hwang, S., Kim, S., Choi, Y. & Y. Park (2015). Modeling Occurrence of Urban Mosquitos Based on Land Use Types and Meteorological Factors in Korea. *International Journal of Environmental Research and Public Health 12*: 13131-13147.

Larson, S., DeGroote, J., Bartholomay, L. & R.Sugumaran (2009). Ecological niche modeling of potential West Nile virus vector mosquito species in Iowa. *Journal of Insect Science* 10:110.

Lautenbach (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* 36: 27–46.

Legendre, P. (1993). Spatial autocorrelation: problem or new paradigm. *Ecology* 74: 1659–1673.

Lennon, J. (2002). Red shifts and red herrings in geographical ecology. *Ecography* 23: 101–113.

López- Baucells, A., Rocha, R., Andriatafika,Z., Tojosoa, T., Kemp, J., Forbes, K.A. & M. Cabeza (2017). Roost selection by synanthropic bats in rural Madagascar: what makes non-traditional 2 structures so tempting? To be published.

Maeda, E. (2011). Agricultural expansion and climate change in the Taita Hills, Kenya: an assessment of potential environmental impact. Academic dissertation. University of Helsinki.

Mason, S & N. Graham (2002). Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Quarterly Journal of the Royal Meteorological Society* 128: 2145–2166.

Miller, J. (2010). Species Distribution Modeling. *Geography Compass* 4(6): 490-509.

Moisen, G., Freeman, E., Blackard, J., Frescino, T., Zimmermann, N. & T. Edwards Jr. (2006). Predicting tree species presence and basal area in Utah: a comparison of stochastic gradient boosting, generalized additive models and tree-based methods. *Ecological Modeling* 199: 176–187.

Moran, P (1950). Notes on continuous stochastic phenomena. *Biometrika* 37:17–23.

Mosquito World (2017a). Mosquito Habitats. 15.2.2017. <http://www.mosquitoworld.net/about-mosquitoes/habitats/>

Mosquito World (2017b). Mosquito Species. 15.2.2017.

<http://www.mosquitoworld.net/about-mosquitoes/habitats/>

Msagha, Johanna (2004). Population development in Taita Hills, Kenya. Published in In:

Pellikka, P., J. Ylhäisi & B. Clark (eds.) Taita Hills and Kenya, 2004 – seminar, reports and journal of a field excursion to Kenya. *Expedition reports of the Department of Geography*. University of Helsinki 40, 39-46. Helsinki 2004, ISBN 952- 10-2077-6, 148 pp.

MTI (2017). The mosquito taxonomic inventory. *Anopheles*; *Culex*; *Stegomyia*. 5.4.2017.

< http://mosquito-taxonomic-inventory.info/simpletaxonomy/term/8577>

Mughini-Gras, L., Mulatti, P., Severini, F., Boccolini,D., Romi, R., Bongiorno,G., Khoury,

C., Bianchi, R., Montarsi, F., Patregnani, T., Bonfanti, L., Rezza, G., Capelli, G. & L. Busani (2014). Ecological Niche Modeling of Potential West Nile Virus Vector Mosquito Species and Their Geographical Association with Equine Epizootics in Italy. *EcoHealth* 11: 120–132.

Murphy, C., Grenouillet, G. & E.García-Berthou (2015). Natural abiotic factors more than

anthropogenic perturbation shape the invasion of Eastern Mosquitofish (Gambusia holbrooki). *Freshwater Science* 34(3):965–974.

Murray, K. & M. Conner (2009). Methods to quantify variable importance: implications for

the analysis of noisy ecological data. *Ecology* 90(2): 348-355.

Naimi, B. & M. B. Araújo (2016). Sdm: a reproducible and extensible R platform for species

distribution modeling *Ecography* 39: 368–375.

Nduire, John (2016). Common types of building materials in Kenya. *Construction Business*

Review. 20.4.2017. < http://www.constructionkenya.com/1599/building-materials-in-kenya/>

Neteler, M., Roiz, D., Rocchini, D., Castellani, C. & A. Rizzoli (2011). Terra and Aqua

satellites track tiger mosquito invasion: modeling the potential distribution of *Stegomyia albopictus* in north-eastern Italy. *International Journal of Health Geographics* 10:49.

Pachler, K., Lebl, K., Berer, D., Rudolf, I., Hubalek, Z. & N. Nowotny (2014). Putative New West Nile Virus Lineage in *Uranotaenia unguiculata* Mosquitoes, Austria, 2013. *Emerging Infectious Diseases 20*(12): 2119-2122.

Peirce, C (1884). The Numerical Measure of the Success of Predictions. *Science* 4:93: 453-454.

Pellikka, P., Lötjönen, M., Siljander, M. & L. Lens (2009). Airborne remote sensing of spatio-temporal change (1955-2004) in indigenous and exotic forest cover in the Taita Hills, Kenya. *International Journal of Applied Earth Observations and Geoinformation* 11(4): 221-232.

Peyton, E. (1972). A subgeneric classification of the Genus *Uranotaenia Lynch Arribalzaga*, with a Historical Review and Notes on Other Categories. The mosquito taxonomic inventory.<http://mosquito-taxonomic-inventory.info/sites/mosquito-taxonomicinventory.info/files/Peyton%201972.pdf>

Phillips S., Anderson R. & R. Schapire (2013). Maximum entropy modeling of species geographic distribution. *Ecological Modeling* 190: 231-259.

Phillips, S (2004). A Maximum Entropy Approach to Species Distribution Modeling. Proceedings of the Twenty-First International Conference on Machine Learning. p. 655-662.

Platts P., Omeny P. & R. Marchant (2015). AFRICLIM: high-resolution climate projections for ecological applications in Africa. *African Journal of Ecology* 53: 103-108.

Rasheed, S., Boots, M., Frantz, A. & R. Butlin (2013). Population structure of the mosquito *Stegomyia aegypti (Stegomyia aegypti)* in Pakistan. *Medical and Veterinary Entomology* 27: 430-440.

Rhemtulla, J., Mladenoff, D. & M. Clayton (2007). Regional land-cover conversion in the U.S. upper Midwest: magnitude of change and limited recovery (1850–1935–1993). *Landscape Ecology* 22: 57–75.

Ridgeway, G (2007). Generalized Boosted Models: A guide to the gbm package. 1-12.

Rogan, J., Franklin, J., Stow, D., Miller, J., Woodcock, C. & D. Roberts (2008). Mapping landcover modification over large areas: a comparison of machine learning algorithms. *Remote Sensing of Environment* 112: 2272–2283.

Rykiel, E., Jr. (1996). Testing ecological models: the meaning of validation. *Ecological Modeling* 90: 229–244.

Sallam, M., Xue, R., Pereira, R. & P. Koehler (2016). Ecological niche modeling of mosquito vectors of West Nile virus in St. John's County, Florida, USA. *Parasites & Vectors* 9:371.

Salminen, H. (2004). A geographic overview of The Taita Hills, Kenya. *Expedition reports of the Department of Geography.* University of Helsinki 40, 31-38.

Schowalter, T., (2011). Insect Ecology: An Ecosystem Approach. 3rd edition, 633 p. Elsevier.

Sharp, T (2015). Unveiling the Burden of Dengue in Africa. Published in the blog of Centers for Disease Control and Prevention. 28.7.2015. <https://blogs.cdc.gov/publichealthmatters/2015/07/unveiling-the-burden-of-dengue-in-africa/>

Smith, A. (2012). An Introduction to Best Practices in SDM. Kansas State University. <http://www.earthskysea.org/!ecology/sdmShortCourseKState2012/sdmShortCourse_kState.pdf>

Service, M. (1991). Mosquito Ecology: Field sampling methods.2nd edition, 988 p. Elsevier.

Sharp, T (2015). Unveiling the burden of dengue in Africa. Centers for Disease Control and Prevention, CDC. 24.10.2016.

Siljander, M., Clark, B. & P. Pellikka (2011). A predictive modelling technique for human population distribution and abundance estimation using remotesensing and geospatial data in a rural mountainous area in Kenya. *International Journal of Remote Sensing*. iFirst: 1–27.

Soini, Eija (2005). Livelihood capital, strategies and outcomes in the Taita Hills of Kenya. ICRAF Working Paper no. 8. Nairobi, Kenya: World Agroforestry Centre.

Svenning, J. & F. Skov (2004). Limited filling of the potential range in European tree species. *Ecology Letters* 7: 565–573.

Swets, K. (1988). Measuring the accuracy of diagnostic systems. *Science* 240: 1285–1293.

Thuiller, W., Georges, D., Engler, R. & F. Breiner (2016). Package "biomod2". Ensemble platform for species distribution modeling. Version 3.3-7. <https://cran.r-project.org/web/packages/biomod2/biomod2.pdf>

Thuiller,W., Ara´ujo, M. & S. Lavorel (2003). Generalized models vs. classification tree analysis: Predicting spatial distributions of plant species at different scales. *Journal of Vegetation Science* 14: 669–680.

Thuiller, W., Lavorel, S., Sykes, M. & M. Ara´ujo (2006). Using niche-based modeling to assess the impact of climate change on tree functional diversity in Europe. *Diversity and Distributions* 12: 49–60.

Thuiller, Wilfried (2003). BIOMOD – optimizing predictions of species distributions and projecting potential future shifts under global change. *Global Change Biology* 9:1353–1362.

Tucker, Compton (1979). Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment* 8: 127–150.

University of Helsinki (2003-2009). The Taita Hills. 22.4.2016. <www.helsinki.fi/science/taita/taitahills.html>

Vaughan, I. & S. Ormerod (2005). The continuing challenge of testing species distribution models. *Journal of Applied Ecology* 42: 720–730.

Virtanen, Elina (2015). Fine-resolution climate grids for species studies in data-poor regions. Master´s Thesis. University of Helsinki.

Walker, K., Joy, T., Ellers-Kirk, C. & F. Ramberg (2011). Human and environmental factors affecting *Stegomyia aegypti* distribution in an arid urban environment. *Journal of the American Mosquito Control Association* 27(2): 135–141.

Ward, G., Hastie, T., Barry, S., Elith, J. & J. Leathwick (2009). Presence-only data and the EM algorithm. *Biometrics* 65: 554–563.

WHO (2016a). The World Health Organization. Vector-borne diseases. February 2016. Fact sheets. 20.10.2016.<http://www.who.int/mediacentre/factsheets/fs387/en/>

WHO (2016b). The World Health Organization. Chikungunya- Kenya. Disease outbreak news. August 2016. 20.4.2017. <http://www.who.int/csr/don/09-august-2016-chikungunya-kenya/en/>

WHO (2016c). World health Organization. Dengue and severe dengue. Fact sheets July 2016. 20.10.2016. <www.who.int/mediacentre/factsheets/fs117/en/>

WHO (2017a). The World Health Organization. Zika virus. 5.4.2017. <http://who.int/emergencies/zika-virus/mediacentre/en/>

WHO (2017b). The World Health Organization. Dengue control: The mosquito. 5.4.2017. <http://www.who.int/denguecontrol/mosquito/en/>

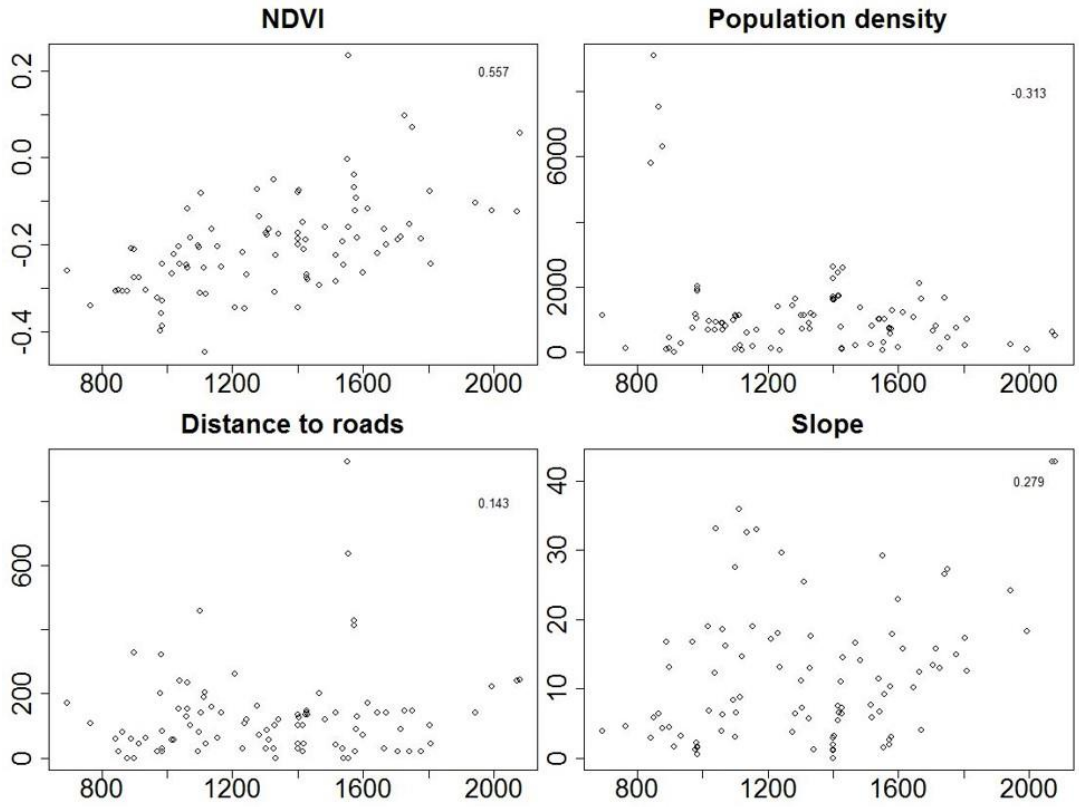
WHO (2017c). The World Health Organization. Neglected tropical diseases: Mosquito-borne diseases. 5.4.2017. <http://www.who.int/neglected_diseases/vector_ecology/mosquito-borne-diseases/en/>

Yee, T. & N. Mitchell (1991). Generalized additive models in plant ecology. *Journal of Vegetation Science* 2: 587–602.

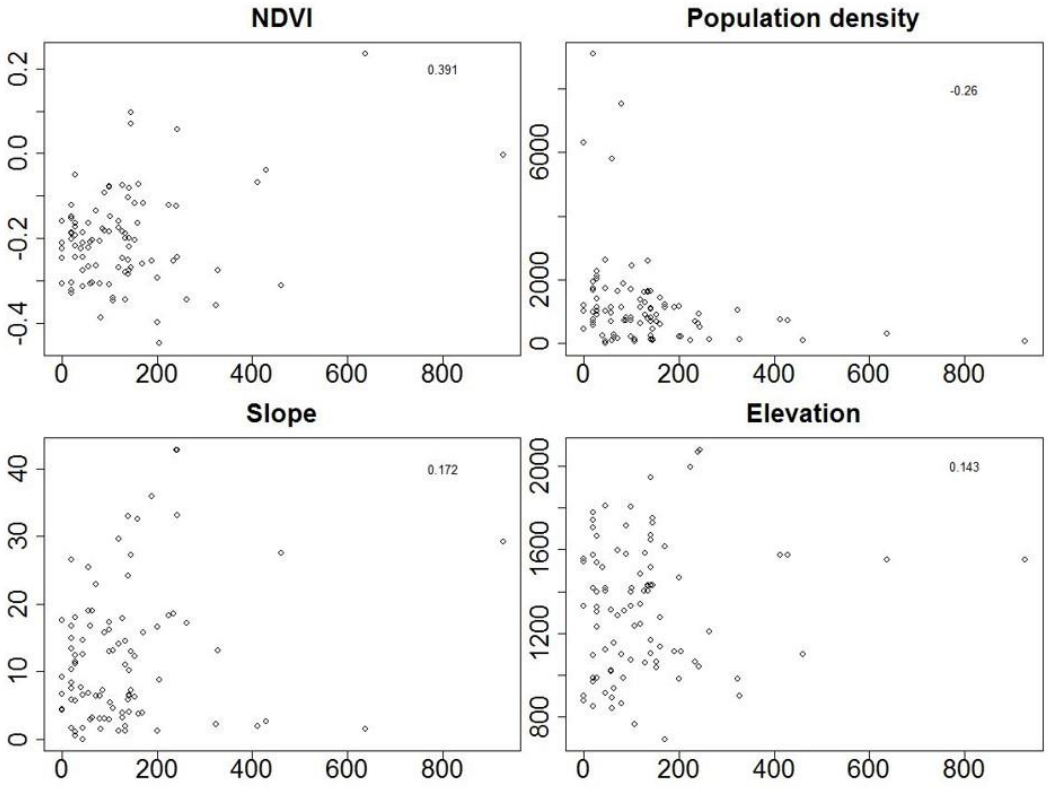
Zimmermann, N.E. (2000). Shortwavc.aml. Available online at: http://www.wsl.ch/staff/ niklaus.zimmermann/programs/aml1_1.html (accessed 01 January 2009).

Appendices

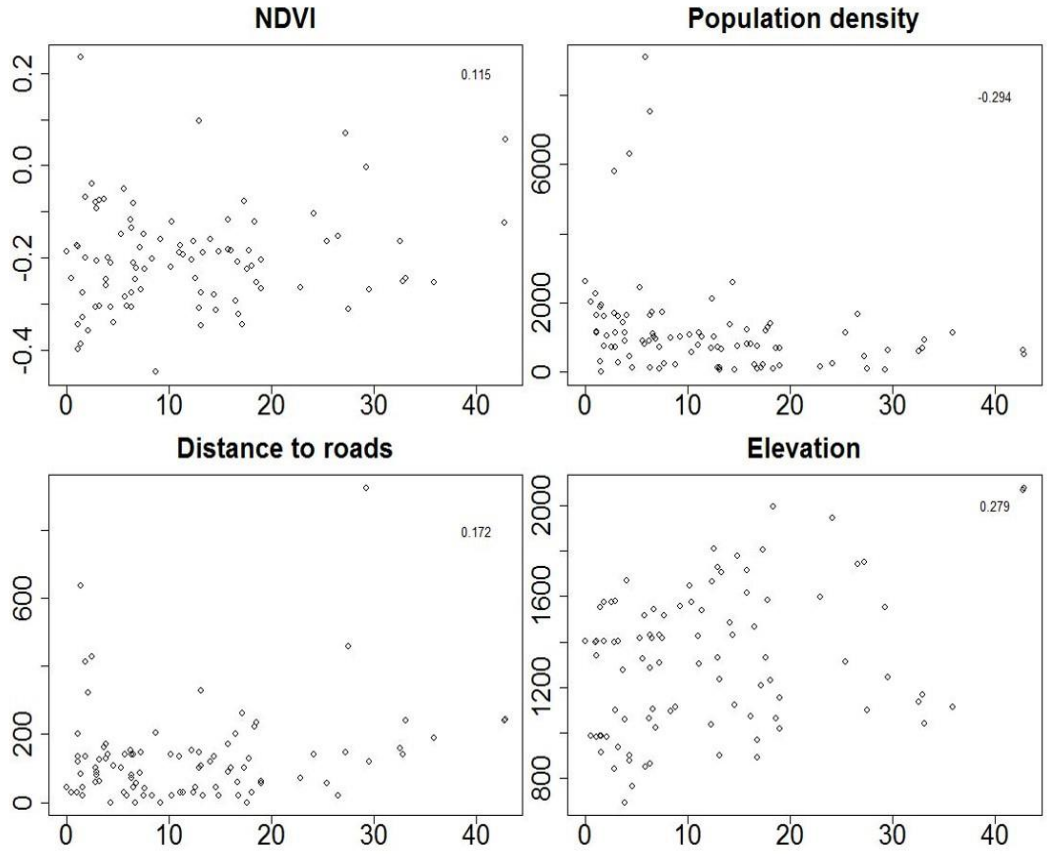
Appendix 1. Correlations of *Culex* predictors are introduced here; elevation and other variables.



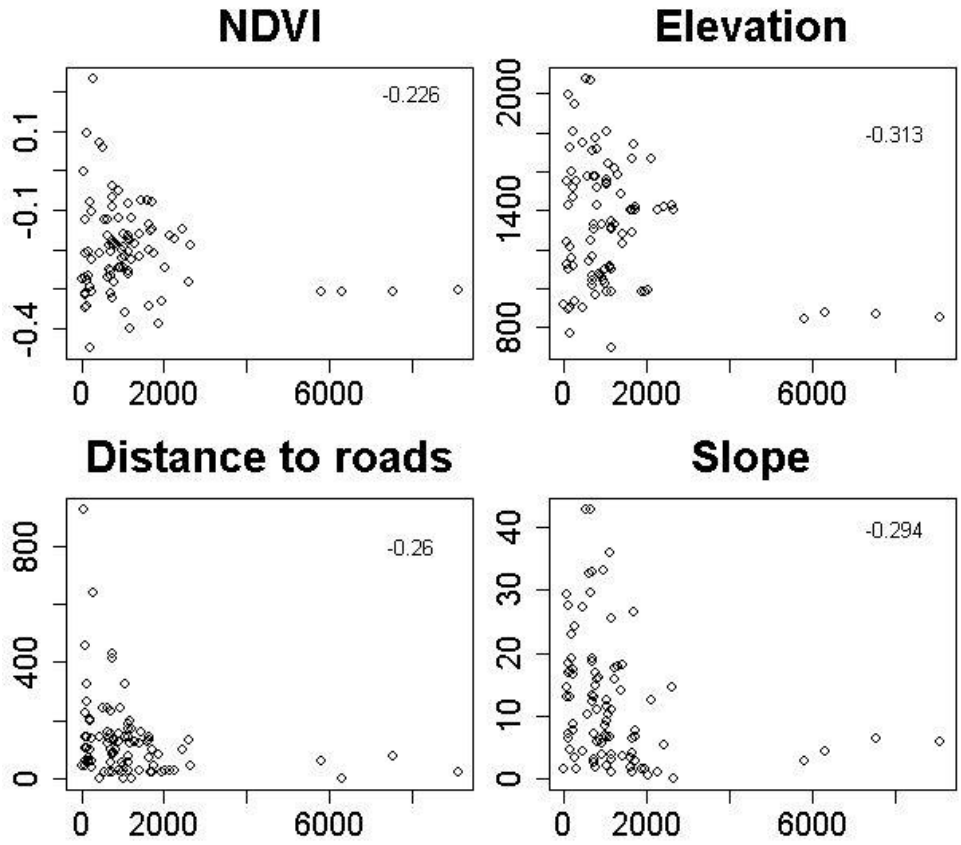
Appendix 2. Correlations of *Culex* predictors are introduced here; distance to roads and other variables.



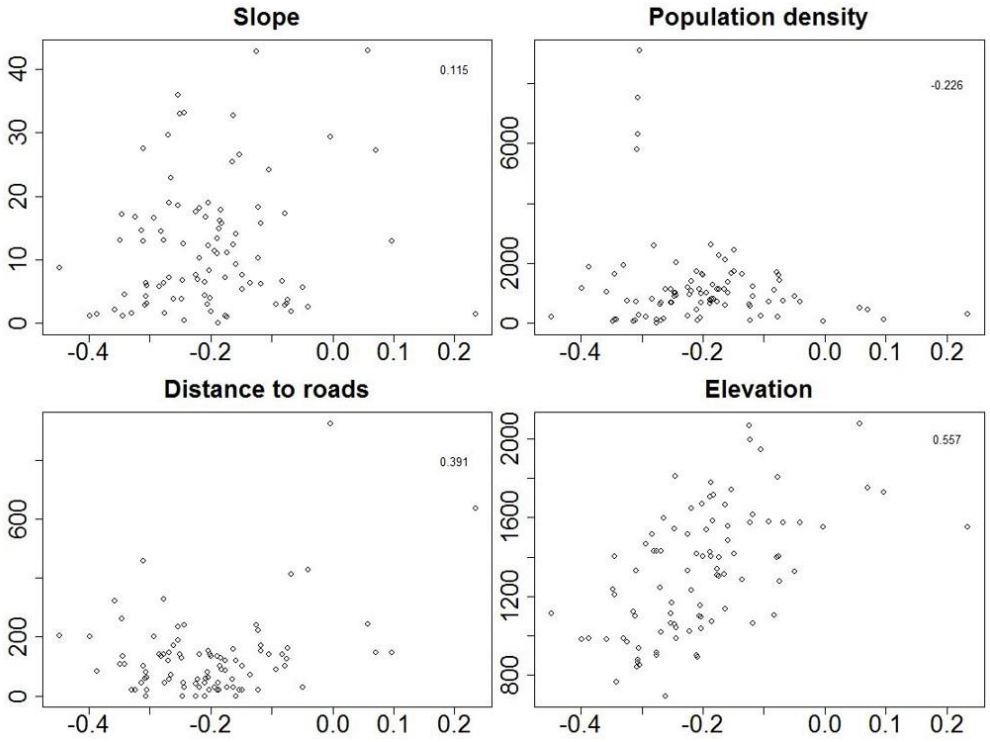
Appendix 3. Correlations of *Culex* predictors are introduced here; slope and other variables.



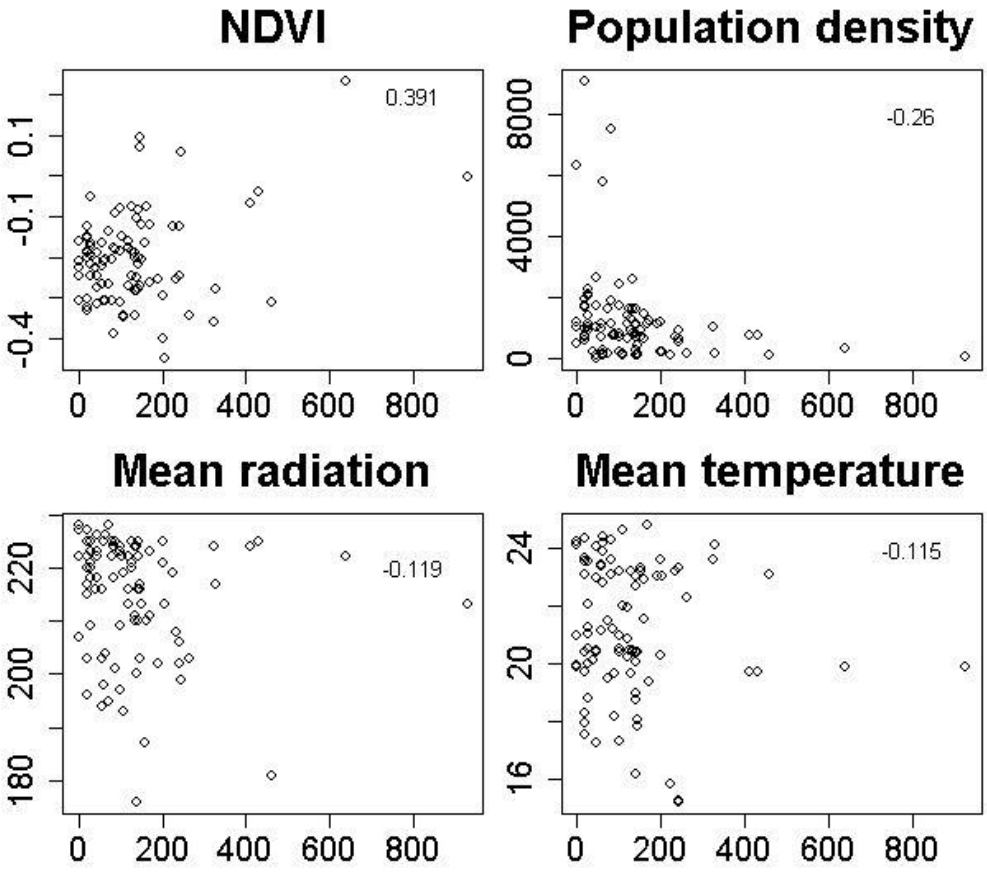
Appendix 4. Correlations of *Culex* predictors are introduced here; human population density and other variables.



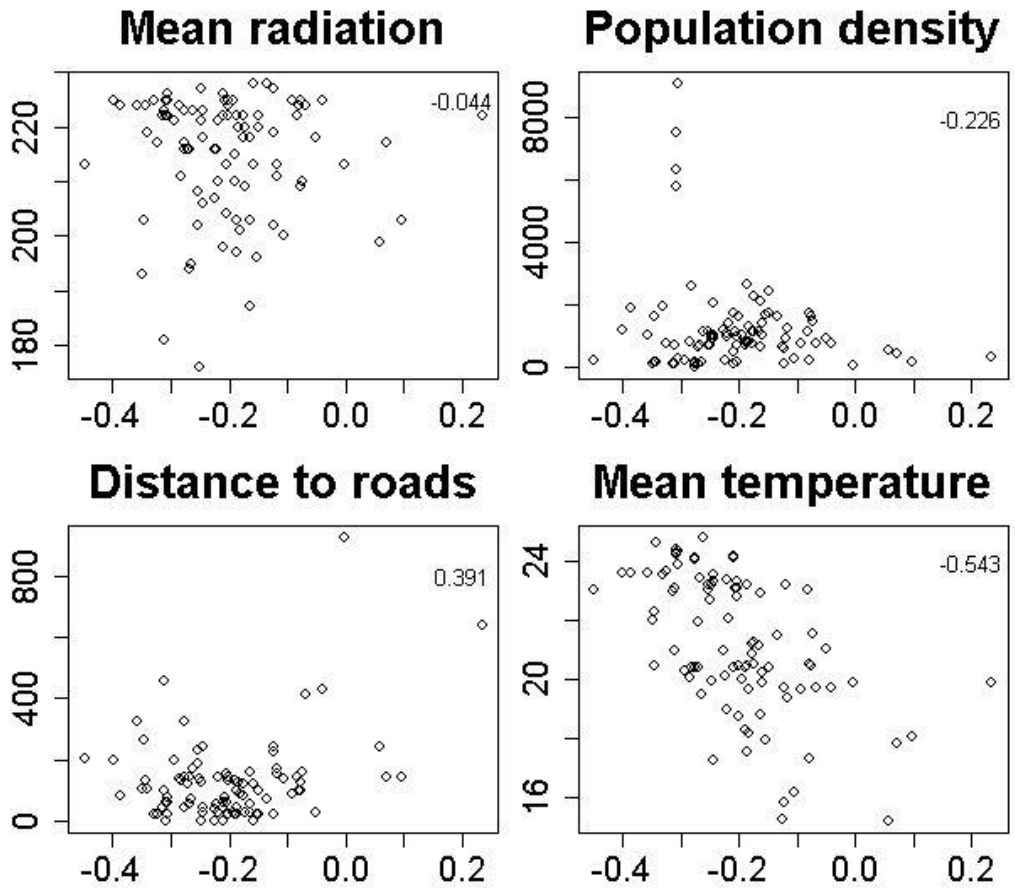
Appendix 5. Correlations of *Stegomyia* predictors are introduced here; distance to roads and other variables.



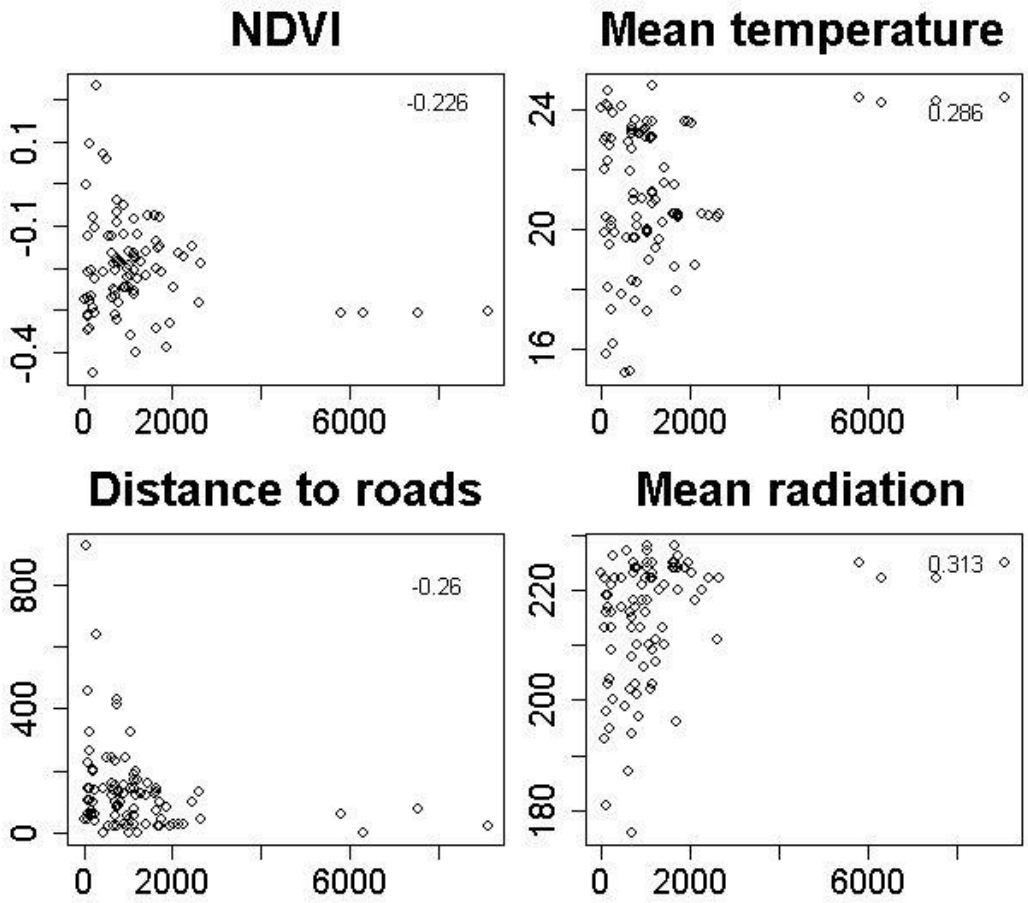
Appendix 6. Correlations of *Stegomyia* predictors are introduced here; distance to roads and other variables.



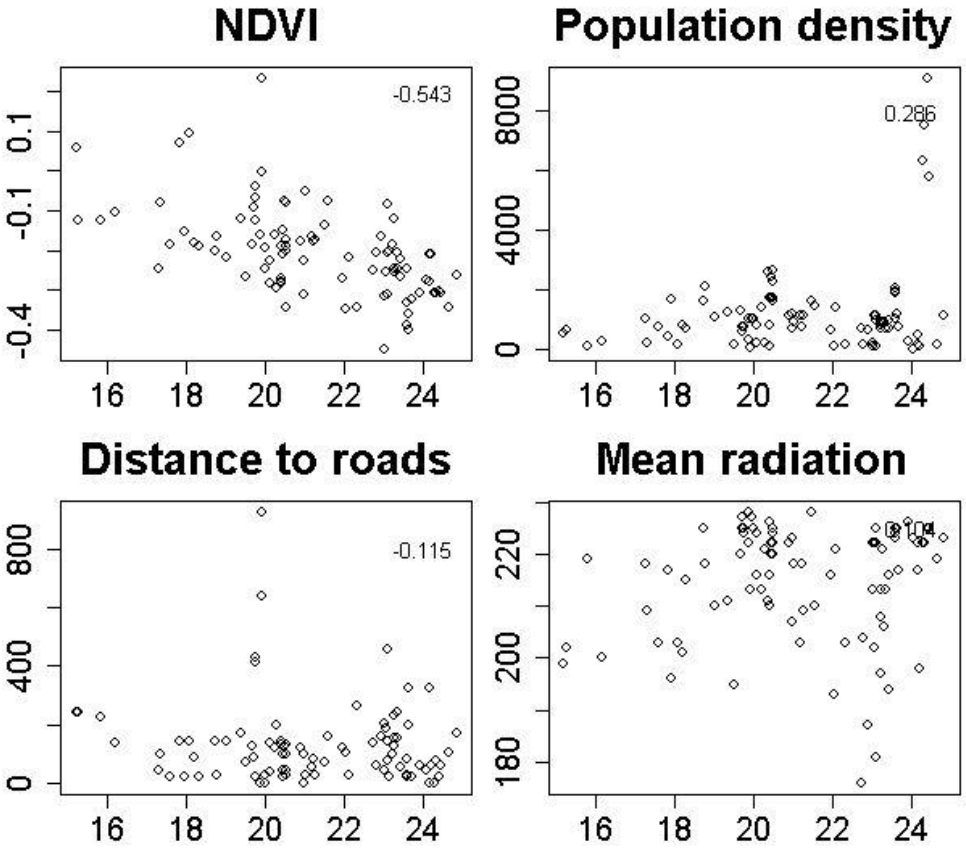
Appendix 7. Correlations of *Stegomyia* predictors are introduced here; NDVI and other variables.



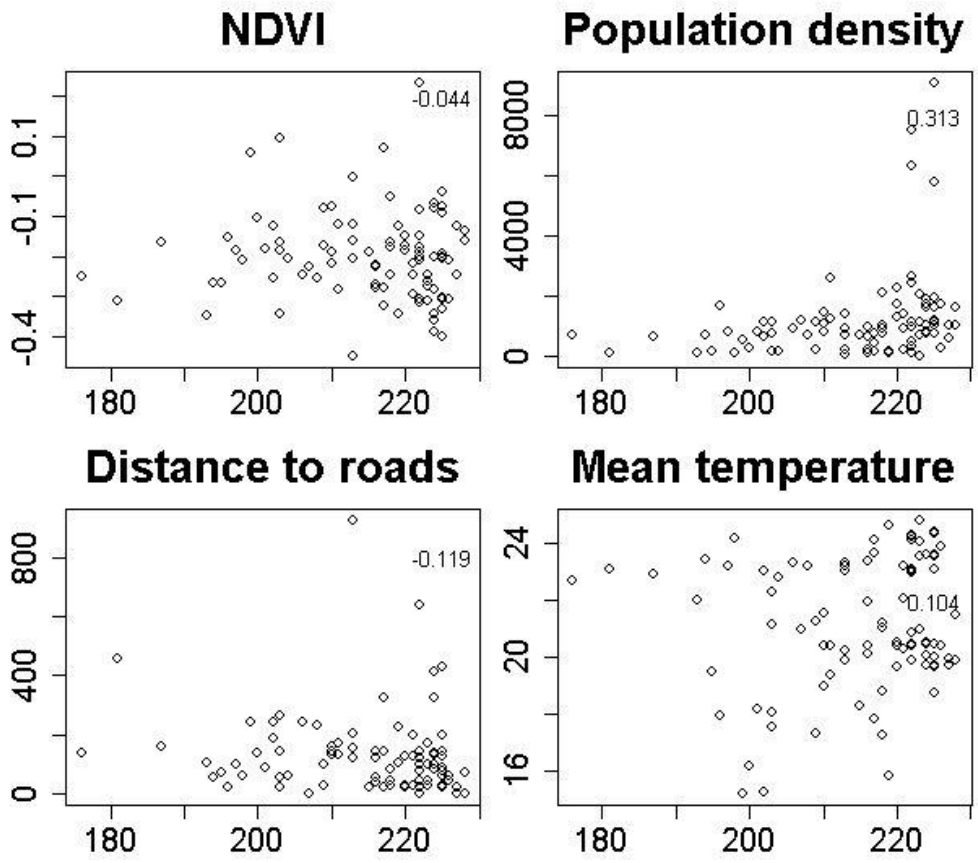
Appendix 8. Correlations of *Stegomyia* predictors are introduced here; human population density and other variables.



Appendix 9. Correlations of *Stegomyia* predictors are introduced here; mean temperature and other variables.



Appendix 10. Correlations of *Stegomyia* predictors are introduced here; mean radiation and other variables.



Appendix 11.The most important parameters of GLM model that predicted *Culex* distributions.

```
Call:
glm(formula = Culex ~ popden + I(ndvi^2) + I(popden^2), family = Options@GLM$fa
mily,
    data = Data[calibLines, , drop = FALSE], weights = Yweights[calibLines],
    mustart = rep(Options@GLM$mustart, sum(calibLines)), control = eval(Options
@GLM$control),
    model = TRUE)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1709  -1.0550   0.7155   0.9315   1.7140

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.081e-02  5.926e-01  -0.052   0.9585
popden       1.236e-03  5.068e-04   2.438   0.0147 *
I(ndvi^2)   -1.147e+01  7.299e+00  -1.572   0.1160
I(popden^2) -1.128e-07  6.330e-08  -1.782   0.0747 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 114.41  on 84  degrees of freedom
Residual deviance: 102.16  on 81  degrees of freedom
AIC: 110.16

Number of Fisher Scoring iterations: 4
```

Appendix 12. The most important parameters of the GAM model that predicted *Culex* distributions.

```
Family: binomial
Link function: logit

Formula:
Culex ~ 1 + s(slope, k = -1) + s(popden, k = -1) + s(ndvi, k = -1) +
    s(distroad, k = -1) + s(dem, k = -1)

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.6260     0.3505   1.786   0.0741 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
             edf Ref.df Chi.sq p-value
s(slope)   3.959  4.882  4.811  0.4328
s(popden)  3.977  4.825 11.440  0.0324 *
s(ndvi)    1.000  1.000  3.237  0.0720 .
s(distroad) 2.067  2.556  4.832  0.1769
s(dem)     1.000  1.000  0.480  0.4885
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.263   Deviance explained = 30.8%
UBRE = 0.2376  Scale est. = 1         n = 85
```

Appendix 13. The most important parameters of the random forest model that predicted *Culex* distributions.

	Length	Class	Mode
call	9	-none-	call
type	1	-none-	character
predicted	86	factor	numeric
err.rate	1500	-none-	numeric
confusion	6	-none-	numeric
votes	172	matrix	numeric
oob.times	86	-none-	numeric
classes	2	-none-	character
importance	5	-none-	numeric
importanceSD	0	-none-	NULL
localImportance	0	-none-	NULL
proximity	0	-none-	NULL
ntree	1	-none-	numeric
mtry	1	-none-	numeric
forest	14	-none-	list
y	86	factor	numeric
test	0	-none-	NULL
inbag	0	-none-	NULL
terms	3	terms	call

Appendix 14. A MARS model for *Culex* estimations resulted in AUC-value of 0.806.