Centre for Geo-Information

Thesis Report GIRS-2014-24

The Spatial Distribution of Mosquitoes Related To the Environment on Rusinga Island, Western Kenya.

A Case Study

Corné Vreugdenhil

26 June 2014

WAGENINGEN UNIVERSITY
WAGENINGEN UR

The Spatial Distribution of Mosquitoes Related To the Environment on Rusinga Island, Western Kenya.

A Case Study

Corné Vreugdenhil

Registration number 891009-910-050

Supervisors:

Ron van Lammeren

Tobias Homan

A thesis submitted in partial fulfilment of the degree of Master of Science

at Wageningen University and Research Centre,

The Netherlands.

26 June 2014,

Wageningen, The Netherlands

Thesis code number: GRS-80436
Thesis Report: GIRS-2014-24
Wageningen University and Research Centre
Laboratory of Geo-Information Science and Remote Sensing

Acknowledgements

I would like to thank all the people who supported me during my thesis project. Special thanks to my supervisors, Ron van Lammeren and Tobias Homan, who supported me during the whole project and especially guided me in the right way during brainstorm sessions and the revisions of my report. Also special thanks to Alexandra Hiscox, who was involved in almost all stages of this project and supported me with ideas and knowledge from an entomologist point of view. Thanks to Richard Mukabana and Willem Takken, project leaders of the SolarMal project, for their support to my research in Kenya. I also want to thank the people from the SolarMal project who assisted me in Kenya during the fieldwork: Ibrahim Kiche who provided me the tablets, David Owaga who guided me in the field, Jackton Arija and Charles Wambua for the transportations, and in general all the members of the SolarMal project for the company and conversations which I appreciated a lot.

Thanks to my family, my wonderful wife Willeke and lovely kids Boas and Vera. They gave me the enthusiasm to work on this project and supported me at home through all the stages of the thesis. Most thanks to God and Jesus Christ for giving me the opportunity and the pleasure of exploring His amazing creation.

Abstract

The SolarMal project on Rusinga Island, Western Kenya, aims to eliminate malaria by reducing the mosquito population, using a new developed mosquito trap which will be installed at all households on the island. The effect of the installed traps is monitored by sampling mosquitoes from randomly selected selected households on the island during the whole duration of the project. This thesis project aims to support the SolarMal project by performing spatial analysis to the distribution of mosquitoes in relation to the environment on Rusinga Island. Environmental variables were searched that are, according to literature, potential determinants of mosquito presence. Second, potential determinants were validated against the mosquito distribution on Rusinga Island. Third, the fitness of the available spatial data, both environmental and mosquito catches, to relate with each other, was examined.

In order to study the environment of the island, an elevation map (30 m. ASTER) and a satellite image (2,4 m QuickBird) were available for the creation (in ArcGIS 10.2) of 14 environmental variables, for example the slope and topographical wetness index of the area. These environmental variables were related (within the R environment) to a spatial dataset consisting one year of mosquito catches.

This study shows that there are only weak correlations (R^2 < 0,11) between the studied environmental variables and the mosquito catches. The spatial distribution of malaria vector mosquitoes is highly varying over time and shows small preferences for specific areas on the island.

It is concluded that the available spatial data was not suitable for explaining the spatial distribution of adult mosquitoes on Rusinga Island. An explanation for this is found in the small scale breeding sites that are commonly found on Rusinga Island like tire tracks, footprints of cattle in drenched grass, or dumped waste in bushes, which were not detectable. The spatial data for the creation of environmental variables are limited in the spatial resolution to detect the small scale breeding sites (< 0,5 m) and are especially in the temporal resolution too limited for detection of temporal breeding sites. Continuations of this study are highly recommended to increase the spatial resolution of the data in combination with the inclusion of other environmental variables that are indicators for the chance on the mentioned small scale breeding sites. In order to increase the temporal resolution, inclusion of precipitation data is recommended.

Table of Contents

1 Introduction

1.1 Malaria

One of the major health threats worldwide is the malaria disease. In literature, from 1960 on, an increasing number of studies have been carried to research malaria. Despite worldwide activities to eliminate it, malaria is still one of the most lethal diseases in a large part of the world, and therefore a widely discussed topic in science. The World Health Organization estimated that worldwide 3.3 billion people were at risk of malaria in 2011. In the same year, 26 million infections of malaria were reported and 106.820 deaths that are caused by malaria were officially reported (WHO 2013). Humans can acquire malaria by infection of the *Plasmodium* parasite of which *Plasmodium falciparum* is the most lethal one. The *Plasmodium* parasite is transmitted from host to host by mosquitoes and is injected into human blood during blood meals of female mosquitoes. Malaria predominantly occurs in sub-Saharan Africa (WHO 2012b). In Kenya, 36% of the population is at high risk of malaria, of which almost all cases are due to the *Plasmodium falciparum* parasite. In Kenya, the parasite's vector species are mainly the mosquitoes *Anopheles gambiae sensu stricto, An. arabiensis, An. funestus and An. merus* (WHO 2012a).

So far, society tried to control malaria by the use of insecticides to eliminate mosquitoes, and drugs to clear parasites within infected people. Due to natural selection, these measures are proven not to be successful on the long term, since mosquitoes and *Plasmodium falciparum* are getting resistant against insecticides and drugs (Wernsdorfer 1994, Hiscox et al. 2012, Melmane et al. 2014). Next to the existing malaria control measures, an increasing number of projects and national programs try to decrease malaria infections by controlling the mosquito populations in order to limit the transmission risk of malaria. In literature, from around the year 1980 onwards, an increasing interest is found for the behavior of mosquitoes. In addition, from around the year 1990 on, due to technical developments in computer sciences, an increasing amount of research is dealing with the spatial distribution of mosquitoes in relation to environmental properties of the surrounding landscape. Controlling the malaria disease is nowadays also focusing on controlling the parasites' vector, the mosquito populations and especially their habitats. Knowing more about the type and locations of mosquito habitats, malaria could be restricted by elimination of its vector species. The use of GIS by these studies is of increasing interest since the last 15 years.

1.2 SolarMal Project

The SolarMal project aims to eliminate malaria on Rusinga Island, Western Kenya (Hiscox et al. 2012). Next to the existing nationwide strategy of case management and bed net use, it emphasizes substantially reducing vector abundance by mass trapping of vectors with mosquito traps. They use a newly developed trap that attracts mosquitoes, with an odor more attractive for mosquitoes than human odorants (Mukabana et al. 2012). These mosquito traps will be installed outside each household on Rusinga Island. The hypothesis is that the use of these traps will be an effective method of reducing mosquito populations to reduce malaria transmission eventually leading to malaria elimination (Hiscox et al. 2012). To monitor changes, they perform each round of 6 weeks an extensive entomology survey at 80 households, where they catch mosquitoes for further analysis. The setup of the SolarMal project included a broad survey on Rusinga Island. For all households on the island, information is gathered about the household members, its construction type, kind of neighborhood, etc. This information is used to see whether there is any relationship between the occurrence of mosquitoes and human activity. Next to this broad household data, more inventories are done over the whole island.

1.3 Problem Definition

A Spatial Decision Support System (SDSS) is an interactive system designed to support decision making with a spatial component in the information. A SDSS can for example be used during an epidemiologic disaster where fast decision making is necessary. The SDSS can support the decision makers by providing, preferably real-time, maps of the situation or even predictions of what is going to happen. Kelly et al. show that the use of a SDSS, based on a Geographical Information System (GIS), can contribute to an effective and efficient way of eliminating and controlling malaria (Kelly et al. 2013). Such a system could even be improved by implementation of a model that detects areas with high risk for malaria vector species. Therefore, more knowledge about the relationship between mosquito occurrence and environmental determinants is desirable.

As stated before, an increasing amount of research is now performed to the spatial distribution of mosquitoes in relation to environmental determinants. However, more research to these relationships is still needed. Where many studies show that a certain environmental variable correlates with the occurrence of mosquitoes, no study has succeed to develop a model that predicts the spatial distribution of mosquitoes and is valid anywhere else on earth.

This type of deficit may be explained by the limiting spatial resolution of the input data. Many mosquito populations arise close to small water bodies which often have a size of less than 1 m^2 (Clements et al. 2013). Especially tire tracks that are temporary flooded can be an interesting breeding site for mosquitoes (Mutuku et al. 2006). If one wants to capture these small-scale features, very high spatial resolution data is needed, or other data that are indicators for such small-scale features, to fully understand and predict the spatial distribution of mosquitoes. Another explanation may be the lack of ground truth data, since extensive mosquito monitoring is quite rare. For modeling the spatial distribution of mosquitoes, mosquito occurrence data is needed for calibration and validation of the model. Some studies suggest a promising model, however were not able to calibrate and validate their model, simply due to a lack of ground truth data; Bulsink, for example, modeled the probability of mosquito larvae, however could not validate his model due to a lack of mosquito larval data (Bulsink 2007). Next to a lack of ground truth data or a limiting spatial resolution of the data, a limiting set of input data also limits the completeness of a model.

Moreover, if a study would prove that a certain model is able to predict the spatial distribution of mosquitoes, it would still be debatable whether this model would also fit in another environment. It is more realistic to focus on a model that would be valid within a specific environment and for specific mosquito species within that environment. However, studies come with frameworks that could be used to build a SDSS that serves the targeted elimination of malaria, like is done on the Solomon Islands and Vanuatu (Kelly et al. 2013). Clements et al. suggest also the implementation of environmental data in such a SDSS (Clements et al. 2013). A SDSS can be set up for each individual malaria elimination program. A study to the local spatial distribution of the mosquitoes and its environmental determinants would be needed to implement the environmental information into the SDSS of a malaria elimination program.

The innovative SolarMal project made Rusinga Island an outdoor laboratory for research to malaria. Since the SolarMal project is monitoring and studying the mosquitoes on Rusinga Island, more inside is needed to the spatial distribution of mosquitoes and the cause of this distribution. The broad survey over the whole island provides a lot of malaria related information about the area, which is critical for such a study. Rusinga Island is therefore a perfect location for a case study to the relationship between mosquito spatial distribution and environmental determinants.

Recently a paper has been published, about similar research that was done in 2006 on Rusinga Island by the Nagasaki University (Nmor et al. 2013). Nmor et al. studied the relationship between DEM derivatives and mosquito breeding sites and showed the predictive power of elevation maps in this. Their ground truth data was collected during a survey in April 2006, which covered whole of Rusinga Island. Nmor et al. is mainly focusing on the distribution of breeding sites where larvae develop. The remaining challenge is to study the spatial distribution of adult mosquitoes.

The ground truth data that is used in this thesis concerns adult mosquitoes that were trapped close to households. This thesis focusses more on predicting the risk for malaria vectors near households, considering their surrounding environment. Next to a difference in ground truth data, this thesis has access to and apply more environmental data than the study of Nmor et al. Nmor et al. only used two resolutions of a DEM (30 and 90 m) for prediction of mosquito breeding sites. This thesis use, next to a DEM of 30 m resolution, a high resolution (2.4 m) multispectral image for prediction of adult mosquito occurrence. Despite the fact that a study is already carried out on mosquito larvae on Rusinga Island, the research done in this thesis contribute to more knowledge since other ground truth data and more environmental data will is used.

1.4 Objective and Derived Research Questions

The main objective of this thesis was to study the spatial distribution of mosquitoes on Rusinga Island and to predict the spatial distribution of mosquitoes using environmental variables. These environmental variables are all derived from only a DEM and a multispectral image. The focus here was on all mosquito types that occur on Rusinga Island as well as vector species for malaria. This thesis aimed to answer three specific research questions:

RQ1. Which environmental variables relate to the occurrence of mosquitoes, according to literature?

RQ2. Which environmental variables on Rusinga Island are related to the occurrence of mosquitoes?

RQ3. Is the available mosquito dataset suitable to relate with environmental variables on Rusinga Island?

The aim was to answer research question 1 by performing a literature study, searching for the environmental variables that probably relate somehow to the occurrence of mosquitoes. From literature, only the environmental variables are selected that could be created with the available data. The second research question was dealt with by performing data analysis. These data analysis include correlating environmental variables with the available mosquito data. Answering research question 3 was based on the results of the data analysis and on the fieldwork done in Kenya.

1.5 Reading Guide and Definitions

A description of the study area, Rusinga Island, and the available data for this study are given in chapter 2. Chapter 3 describes the methodology and results of the literature study that is done to the environmental variables that are possibly related to the occurrence of mosquitoes. This chapter is related to research question 1. Chapter 4 is about the derivation of environmental variables out of the available data. Chapter 5 describes the methodology and results of the correlation analysis and is related to research question 2. Chapter 6 describes in detail the available mosquito data with respect to the suitability of relating this data with the environmental variables, and is related to research question 3. Chapter 7 describes the work that is done on Rusinga Island for the SolarMal project, which involves a study to the accuracy of position measurements of houses on the island. Chapter 7 is related to research question 3 since the described work resulted in more knowledge of the study area and the accuracy of the spatial data used. In chapter 8 the conclusions of this study can be found inclusive the discussion associated with it. Also in chapter 8 the recommendations can be found for continuations of this study. This report contains some terminology and abbreviations that probably needs some extra explanation. The definition of these words are given in the terminology list.

2 Study Area and Available Data

The available data and the study area for this study were already fixed, since the study is part of the SolarMal project. This chapter briefly describes which data exactly was available and summarizes the characteristics of the study area.

2.1 Study Area

Rusinga Island is in Western Kenya, just in a corner of Lake Victoria. The island connects to the mainland by a causeway close to the town Mbita (Figure 1). The study area covers whole Rusinga Island, around 47,6 km^2. A small island, called Ngodhe Rao Island, is in the north of Rusinga Island. The maps in this report cover this island and a piece of mainland, however, Ngodhe Rao Island and the main land of Kenya are not part of the study area.

Lake Victoria lies on 1125 meters above sea level and is one of the biggest fresh water bodies in the world. Fresh water surrounds Rusinga Island and people therefore live close to the water. People on Rusinga Island mainly live from fishery and small-scale farming. A hill in the middle characterizes the island, which is up to a height of 300 meters above the water level of Lake Victoria. Around 25.000 people, living in approximately 4.500 households distributed over around 7.000 houses, populate the island. Almost all the houses are close to the water and not one on the hill in the middle of the island. Two rainy seasons, a long one from March until June and a short one from October until December characterize the climate on Rusinga Island.

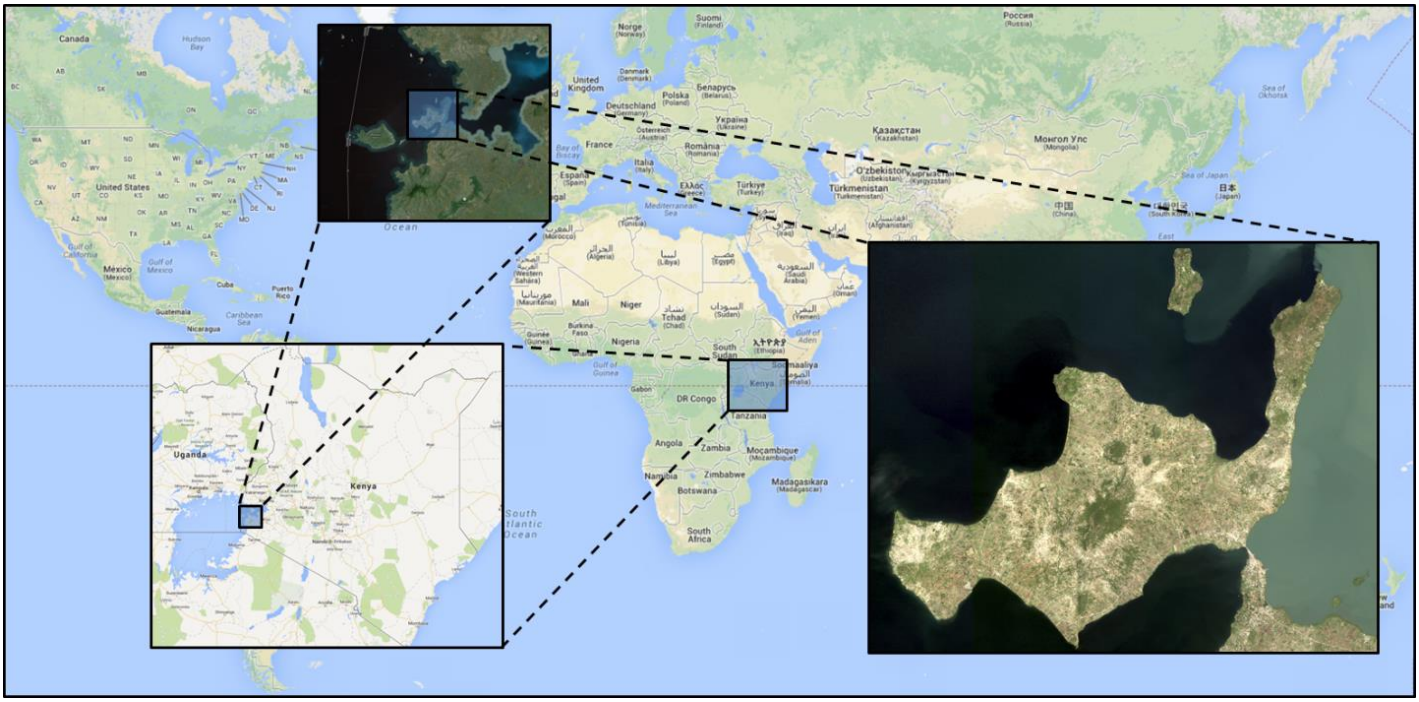


Figure 1: Rusinga Island, Western Kenya.

2.2 Data

2.2.1 DEM

The DEM that was available for this study is an ASTER GDEM 2. This Global Digital Elevation Model is generated using measurements of the Advanced Spaceborn Thermal Emission and Reflection Radiometer (ASTER 2011). The geographical coordinate system is in geographic latitude and longitude, and this DEM is referenced to the 1984 World Geodetic System (WGS84). The grid size of the DEM is one arc-second, which is approximately 30 meters on Rusinga Island. One GDEM was available for this study that covers Rusinga Island completely.

2.2.2 Multispectral Image

A QuickBird multispectral image was available for this study. This multispectral image includes the bands blue (450-520 nm), green (520-600 nm), red (630-690 nm) and NIR (760-900 nm) (DigitalGlobe 2006). The spatial resolution of the image is 2,4 m. Next to these four bands, a panchromatic band (450-900 nm) with a spatial resolution of 0,6 m was available. This panchromatic band can be used for pan sharpening the multispectral image to a spatial resolution of 0,6 m. The QuickBird image is an OR2a product, so the image is geo-referenced, radio-metrically corrected, corrected for sensor and platform-induced distortions, and is ready for orthorectification (DigitalGlobe 2005). Orthorectification can be done using the available DEM. Like the DEM, the QuickBird image also covers Rusinga Island completely. This image was acquired on March 17 2010, just in the beginning of the long rainy season.

2.2.3 Mosquito abundance data

During the SolarMal project, 80 households are randomly selected for mosquito monitoring each sampling round of 6 weeks, from 2012 until 2015. At these households, a mosquito trap with known GPS locations, is used to collect trap mosquitoes for two nights. The trap is placed one night outside and one night inside the house. For this study, the data of the first 7 sampling rounds, or 12 months is available, and consists of 1192 sample points. The following information is, amongst others, stored for these measurements:

- Date
- GPS location (Longitude & Latitude)
- Inside / Outside location
- Some identification codes for the house(hold), village and the location within the SolarMal project.
- Mosquito specie (*An. gambiae sensu stricto, An. funestus, Culex, Mansonia, Aedes* or ‘others’)
- Female / male mosquito
- Abdominal status of the female mosquito (Unfed, bloodfed or gravid)
- Room of the house where the trap was installed
- The number of people sleeping in the house during the measurement

This one year of mosquito data is the so-called baseline year of monitoring data, performed before the roll-out mosquito traps. This means that the mosquito data includes no influences of the mosquito traps yet, and therefore it is ensured that the mosquito distribution of the island is as usual.

2.2.4 Other data

Demographic data is collected on Rusinga Island by the SolarMal project, with a health and demographic surveillance system. This demographic data collected provides lots of information about especially humans. All households are labelled and surveyed 3 times a year, have a unique code within the SolarMal project and the GPS locations are known. Many characteristics of each household, both about the building itself and the people living in it, are compiled in a database. Due to the known GPS location of these households, all the household information can be useful within this study.

2.2.5 Software

The analysis and work is mainly done in two programs. First the program R, a language and environment for statistical computing and graphics. Due to the use of R-scripts, analysis in R are easily reproducible. ArcGIS is used for the creation of the environmental variables and for quick visualizing purposes since visualizing of maps in R is less straightforward.

3 Literature study

3.1 Methodology

In literature, more and more studies are found to malaria, also in combination with GIS analysis. The amount of literature is too extensive to start with reading all of these. Therefore the papers that are recently published were selected, assuming that these studies are built on the findings of previous studies. If a paper was relevant for this study, the references were checked for more papers that also could be of interest for this study. In such a way the more elementary papers could be found that are cornerstones for studies to malaria in relation to the environment.

The first research question covers a literature search to potential environmental determinants for the occurrence of mosquitoes and a first study to the needed spatial data to generate these environmental variables. This chapter describes the found environmental variables from the literature. See Table 1 for a quick overview and for all literature.

3.2 Results

3.2.1 DEM-derivatives

In the literature, many studies do point at some relationship between the occurrence of mosquitoes and environmental variables. Clements et al. summarizes some of these studies in a small list of potential environmental variables (Clements et al. 2013). With only a DEM, the elevation, slope and (distance to) water networks can be derived. The final DEM derivatives like water networks, seems almost logical to correlate with the occurrence of mosquitoes, since water means humid circumstances, which is essential for mosquito breeding activities. While it is less straightforward, the first derivatives like elevation and slope are also suggested by other studies (Myers et al. 2009, Moss et al. 2011, Kasasa et al. 2013, Obsomer et al. 2013). Moss et al. is even mentioning the aspect of the land surface. Where Clements et al. stopped with deriving information from a DEM, Moss et al. continues with the 'Index of Topographic Wetness' (ITW) and the 'Topographic Position Index' (TPI) (Moss et al. 2011). The ITW is an indicator of potential moisture, depending on the ratio between upslope area and local slope. The TPI is a classifier based on the slope and landform type of the center cell compared with its neighborhood. Both the ITW and TPI can be generated using only a DEM.

3.2.2 Land use and cover

Some studies suggests the land cover as potential environmental determinant (Bøgh et al. 2007, Clements et al. 2013, Obsomer et al. 2013). Clements et al. also call vegetation cover and the land use, as potential variables that can explain the occurrence of mosquitoes (Clements et al. 2013). A specific type of land use or land cover can be for instance an ideal habitat for mosquitoes. Ideally, maps of land use and land cover are kept track by (local) authorities; however, one can also derive these maps from classifications of multispectral remote sensing images.

3.2.3 Meteorology

Clements et al., Dom et al. and Kasasa et al. point all at meteorological variables, like temperature and rainfall (Clements et al. 2013, Dom et al. 2013, Kasasa et al. 2013). Ignoring spatial variation in soil characteristics etcetera, rainfall patterns can be measures for the wetness of an area. Quite a lot of studies do mention this meteorological variable where it is used for the temporal variation, using meteorological data from a weather station, or for the spatial variation, using data from more than one weather station (Mbogo et al. 2003, Coulibaly et al. 2013, Dom et al. 2013, Kasasa et al. 2013, Obsomer et al. 2013, Wee et al. 2013). Dom et al. also suggest the potential of relative humidity of air, which is more focusing on moisture conditions of the land, including its atmosphere (Dom et al. 2013).

3.2.4 Indices

Some studies use indices as measures for something else (Gaudart et al. 2009, Clements et al. 2013, Kasasa et al. 2013, Obsomer et al. 2013). The Normalized Difference Vegetation Index (NDVI) for example is a measure for greenness of the surface, a measure for living vegetation on the surface. The more there is living vegetation on the surface, the more humid circumstances can be expected since living vegetation needs enough water. Obsomer et al. suggest the NDVI, Normalized Difference Wetness Index (NDWI) and Enhanced Vegetation Index (EVI) (Obsomer et al. 2013). According to Bulsink (Bulsink 2007), the Normalized Difference Pond Index (NDPI) or Normalized Humidity Index (NHI) is a valuable measure for the wetness of an area. Bulsink and Kasasa et al. suggest the distance to open water as environmental variable (Bulsink 2007, Kasasa et al. 2013). A combination of indices like NDVI, NDWI and NDPI can be

used for detection of open water bodies. The NDWI and NDPI, or NHI, are more direct measures for the wetness of an area, while NDVI and EVI are measurements that are more indirect.

3.2.5 Other/Special Variables

Bulsink suggests among others four environmental variables, special for detection of mosquitoes habitats (Bulsink 2007). First Bulsink introduces the risk for landslides. A landslide occurs normally when the slope is steeper than it actually can be what depends on the soil characteristics. The more stable the soil, due to soil structure, water content etcetera, the steeper the slope can be. Normally a heavy rainfall event creates a moment of soil weakness and causes the landslide. Using a DEM for calculation of the slope and the EVI, as proxy for the rooting depth, a risk for landslides can be calculated. Second, Bulsink suggests the density of land use patches. The denser the area is with smaller patches, the higher the risk for mosquito habitats is assumed. Third, Bulsink suggests the detection of dirty roads or tire tracks, using a DEM and soil type. Small roads through fine textured soils are sensitive for tire tracks, which easily remain wet after a rainfall. Fourth, Bulsink suggests the population density or human settlements density. This can be produced using the geographical locations of households. Myers et al., suggests a variation to Bulsink's population density, the distance to administrative centers (Myers et al. 2009).

Table 1: Environmental variables suggested by literature and the needed data for these

Potential Environmental Determinant	Suggested in literature by among other	Needed spatial input data
Elevation	Clements et al., 2013 - Myers et al., 2009 - Moss et al., 2011 - Obsomer et al., 2013 - Kasasa et al., 2013	DEM
Slope	Clements et al., 2013 - Moss et al., 2011 - Obsomer et al., 2013	DEM
Aspect	Moss et al., 2011	DEM
(Distance to) Water Networks	Bulsink, 2008 - Clements et al., 2013 - Obsomer et al., 2013	DEM
(Distance to) Open Water Bodies	Bulsink, 2008 - Kasasa et al., 2013	NDVI, NDWI, NDPI
Landslide Risk	Bulsink, 2008	DEM, EVI
Topographic Wetness	Moss et al., 2011	DEM
Topographic Position Index	Moss et al., 2011 - Obsomer et al., 2013	DEM
Land Cover	Bogh et al., 2007 - Clements et al., 2013 - Obsomer et al., 2013	multispectral imagery
Vegetation Cover	Clements et al., 2013	multispectral imagery
Landuse	Clements et al., 2013	multispectral imagery
Population or Settlements Density	Bulsink, 2008	Household/Population informations
Patch Density	Bulsink, 2008	landuse/cover
Distance to Administrative Centres	Myers et al., 2009	Locations of administrative centres
Mud Roads	Bulsink, 2008	road map, DEM and soiltype
Temperature	Clements et al., 2013 - Dom et al., 2013 - Kasasa et al., 2013	Weather station measurements
Rainfall	Clements et al., 2013 - Mbogo et al., 2003 - Wee et al., 2013 - Dom et al., 2013 - Kasasa et al., 2013 - Coulibaly et al., 2013	Weather station measurements
Relative Humidity	Dom et al., 2013	Weather station measurements
NDVI	Clements et al., 2013 - Gaudart et al., 2009 - Kasasa et al., 2013 - Obsomer et al., 2013	RS(NIR, red)
NDWI	Obsomer et al., 2013	RS(SWIR, NIR) or RS(NIR, G)
NHI/NDPI	bulsink, 2008	RS(SWIR, green)
EVI	Kasasa et al., 2013 - Obsomer et al., 2013	RS(NIR, red, blue)

4 Environmental Variables

This chapter summarizes the environmental variables that were studied within this project, based on the potential environmental determinants from literature. This chapter will briefly describe why the environmental variables were expected to relate to the occurrence of mosquitoes, how they were generated and which data was required. The creation of these environmental variables is done in ArcGIS.

4.1 Methodology

4.1.1 Elevation

The first and most simple environmental variable is the elevation. Since the available DEM is a model of the elevation, no processing was needed for generation of an elevation map. The ASTER GDEM 2 has a spatial resolution of 30 meters and so the elevation map. The general assumption is that the higher the elevation, the more dry the environment since water will flow downwards, i.e. the less attractive for mosquitoes. The elevation is expressed in meters above sea level.

4.1.2 Slope

The slope of a surface is the first derivative of the elevation. The slope is expressed in degrees, where 0 degrees means a flat surface, 90 degrees a vertical slope. The assumption for this environmental variable is that a flat surface is more attractive for mosquitoes since water can easily stay on or in the ground surface, creating a suitable breeding site for mosquitoes.

4.1.3 Aspect

The aspect of a slope is the orientation of the earth's surface. This orientation is defined as the direction of maximum slope. The aspect is normally expressed in degrees from 0 to 360. The problem with such an expression is that an aspect of 5 degrees is completely different to an aspect of 355 degrees looking to the absolute numbers; however, both are in reality a North-facing slope. For this study, the cosines and sinus of the aspect will be used as environmental variables. The sinus of the aspect results in values from -1 (West) to +1 (East), the cosines from -1 (South) to +1 (North). The cosines and sinus of the aspect can be used for linear regression with the mosquito data. Combining the cosines and the sinus of the aspect, a clear distinction can be made between North or South (using the cosinus values) and East or West (using the sines values). The aspect of a slope is studied since a specific aspect could be more attractive for mosquitoes, generally due to more sunlight or more rainfall (orographic rainfall effect). However, due to the location of Rusinga Island, which is close to the equator, differences in the amount of sunlight over the aspects is not expected.

4.1.4 Flow Accumulation

Using the DEM that is available, a flow direction map is produced. This flow direction map represents for each cell the direction of maximum slope gradient. The flow direction map is actually a simulation of how water would flow over the surface. Using this flow direction map, a flow accumulation map is calculated. This flow accumulation map gives for each cell an indication of the watershed above the cell. It counts how many cells uphill are connected to the destination cell. The larger the watershed above the cell, the more water accumulates in this cell. The expectation is that the more water can accumulate in a cell, the wetter the area can be, so the more attractive this area will be for mosquitoes. The flow accumulation is expressed in the amount of cells. These cells are 30 by 30 meters.

4.1.5 Distance to River Networks

Using the flow accumulation map derived from the DEM, a river network map is generated. A threshold of 50 cells watershed was used to determine which cells are parts of a river network or not. This is equal to an area of (50cells x 900m^2/cell) 45000 m^2. The threshold for this selection is chosen such that the selected rivers were also visible in reality. To know where rivers are in reality, the multispectral QuickBird was used for a visual check. However, on the QuickBird image and on online maps, real rivers or streambeds were hardly recognizable. Therefore a line in the forest or a couple of trees in a row was used as indicator for a river. Real rivers appear during heavy rainfall, during the rain seasons. 50 cells as threshold seem to result in a realistic river network that could be reality during a rain season. Having a network of rivers, a map is generated which gives for each cell the Euclidian distance to the closest river, expressed in meters. The area around rivers is generally wet, so it is expected that mosquitoes occur more close to and around rivers.

4.1.6 Distance to Open Water Bodies

Like the distance to river networks, the distance to open water bodies is also of interest. First, one large open water body, Lake Victoria, surrounds Rusinga Island. Next to this large lake, other smaller ponds, artificial or not, can be attractive areas for mosquitoes. Using the multispectral QuickBird image, detection of open water is possible. McFeeters used NDWI for detection of swimming pools for abatement of mosquitoes (McFeeters 2013). According to McFeeters (McFeeters 1996), the NDWI can be calculated, using the Green (G) and Near-InfraRed (NIR) bands, as follows:

$$NDWI = \frac{(G - NIR)}{(G + NIR)}$$

McFeeters asserts that NDWI values larger than zero indicates a water surface, values less than or equal to zero indicates a non-water surface. Having a map of NDWI values for each cell, areas with open water can easily be selected using zero as threshold. Like the distance to river networks, the Euclidian distance to open water is calculated and expressed in meters.

4.1.7 Topographic Wetness Index

The Topographic Wetness Index (TWI) as defined by Moss et al. uses only a DEM as input (Moss et al. 2011). TWI is a combination of slope and flow accumulation and is calculated as follows:

$$TWI = Ln(\frac{AreaUpslope}{\tan(slope)})$$

Where the slope should be measured in degrees and AreaUpslope is the watershed above the cell to be calculated, i.e. the flow accumulation in that cell. The smaller the slope (flatter area) and the larger the watershed, the more wet the cell probably is, and the more attractive it could be for mosquitoes.

4.1.8 Topographic Position Index

Like the TWI, the Topographic Position Index (TPI) uses only a DEM as input (Moss et al. 2011). For calculation of the TPI, Moss et al. used an ArcView extension made by Jenness and Engelman (Jenness and Engelman 2013). This extension includes a function for calculation of the TPI and the classification of it (Jenness 2006) according to the ideas of Weiss (Weiss 2001). The ArcView extension of Jenness and Engelman will also be used in this study for calculation of the TPI values. Positive TPI values represent areas that are locally or overall higher than its surroundings, probably dry areas. Negative TPI values means the opposite, areas that is lower than its surroundings, probably more wet areas. TPI values close to zero represent areas that are equal to the surrounding. It is expected that the more negative the TPI values are, the wetter the area probably is, i.e. the more attractive for mosquitoes.

4.1.9 Population and Settlement Density

The broad survey with detailed household compositions, exact locations of the houses, etcetera, is used for generating a population and settlement density. The population density is defined as the number of people living per hectare. The settlement density is defined as the number of households per hectare. A circular neighborhood with a radius of 250 meters is used for the derivation of these densities. This distance is assumed to be the maximum distance mosquitoes will fly on Rusinga Island, if all the mosquitoes' necessities are within this distance. It is expected that the more people live in a certain area, the more man made water bodies are present like water reservoirs for consumption. These artificial water bodies can be extremely attractive for some mosquitoes, since humans and water are close to each other; humans for the blood meals, water for breeding.

4.1.10 Spectral Indices: NDVI, NDWI, EVI

The Normalized Difference Vegetation Index and the Enhanced Vegetation Index are commonly used indices for the greenness of an area. Both indices have higher values if more living vegetation is present on the earth surface. It is assumed that the more living vegetation is present, the more water is probably available to serve the needs of the vegetation. According to Lillesand et al., (Lillesand et al. 2004), the NDVI is using the Red (R) and Near-InfraRed (NIR) bands. According to Huete et al. (Huete et al. 2002), the EVI uses the Blue (B), Red (R) and Near-InfraRed (NIR) bands. The NDVI and EVI are calculated as follows:

$$NDVI = \frac{(NIR - R)}{(NIR + R)}$$

$$EVI = G * \frac{(NIR - R)}{(NIR + C_1 * R - C_2 * B + L)}$$

Where G is the gain factor, C_1 and C_2 coefficients for correction of the aerosol resistance term, and L the canopy background adjustment. Standard values for these four coefficients are given by Huete et al. as L = 1, C_1 = 6, C_2 = 7.5 and G = 2.5 (Huete et al. 2002).

The NDWI is used for detection of open water bodies, as described above. However, this NDWI value on its own could already be correlated with the mosquito occurrences (Dambach et al. 2012), since it is a measure for the area wetness at the moment the satellite image was taken (beginning of a raining season). The NDVI, EVI and NDWI are all three spectral indices that are generated within this study. It is expected that the higher the values of these indices, the more water is available, so the more attractive for mosquitoes.

4.1.11 Potential Environmental Determinants Which Will Not Be Dealt With

Some potential environmental determinants named in paragraph 3.2 are not studied. First, the land use or land cover would be possible to generate, since a multispectral image is available. However, Mulder does this already for the same SolarMal project (Mulder 2013), so this study left this up to Mulder. The risk for landslides mentioned by Bulsink (Bulsink 2007), is a potential determinant for the larval breeding sites, however, landslides do not seem to be an issue on Rusinga Island. The patch density as a determinant for the occurrence of mosquitoes (Bulsink 2007) is not possible since no land use information is available. Land use classification is done by Mulder (Mulder 2013), so the patch density is also outside the focus of this study. The distance to administrative centers, as mentioned by Myers et al. (Myers et al. 2009), will not be dealt with since no extremely large administrative centers are present on Rusinga Island. The meteorological variables are not part of this study. Meteorological spatial data was not available for this study, since the focus was on using only a DEM and a multispectral image. Last, the generation of mud roads, as suggested by Bulsink (Bulsink 2007), will be skipped since no complete roadmap is available and a soil type map not at all.

4.2 Results

The 14 environmental variables were generated according to the methodology described in paragraph 4.1. The resulting maps can be found in Appendix A. Figure 2 and Figure 3 show for example respectively the elevation and slope of the island. There is a hill in the middle of the island, one small hill in the northwest corner, another one in the southwest corner and an high area in the northeast part.

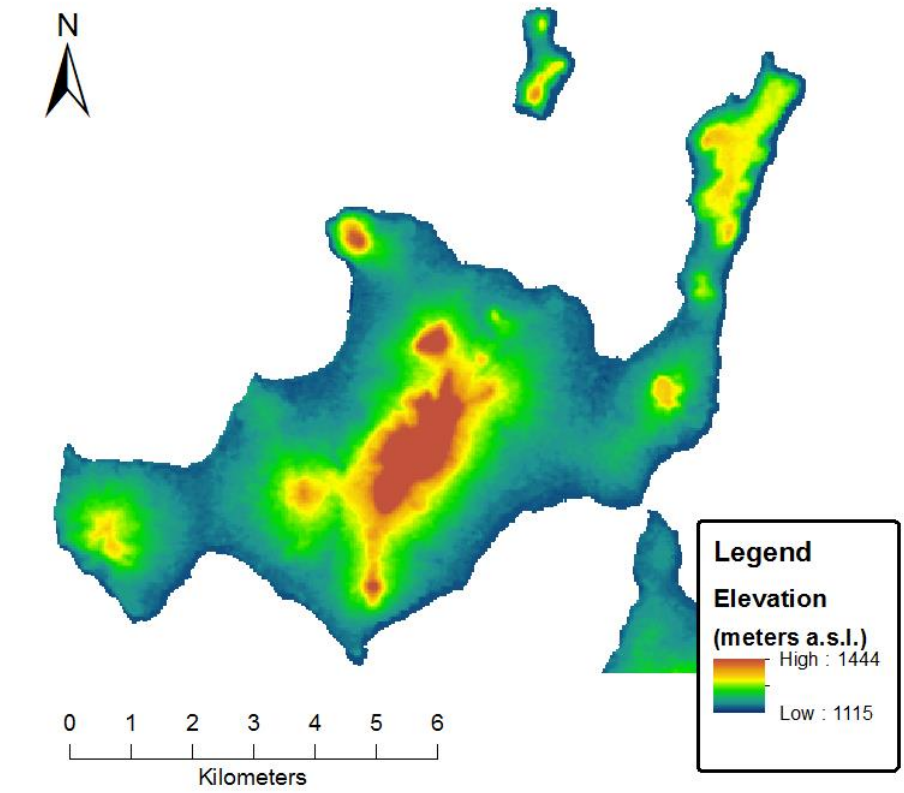


Figure 2: Created environmental variable: Elevation

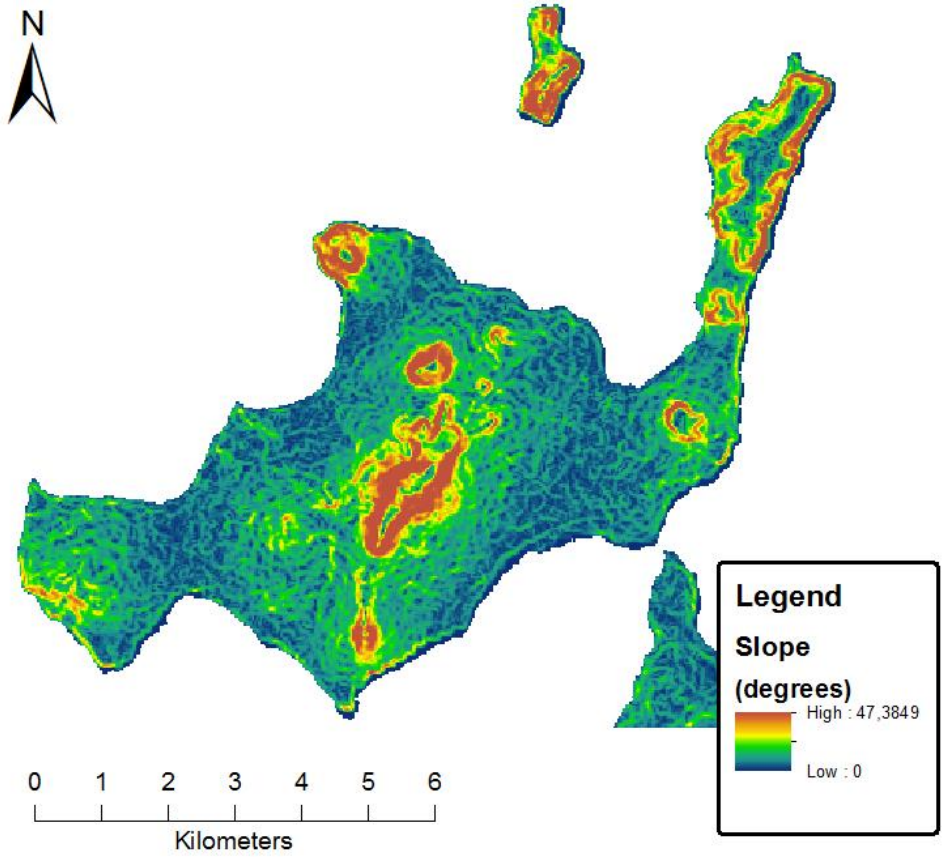


Figure 3: Created environmental variable: Slope

5 Correlations

5.1 Methodology

The correlation between catch sizes of mosquitoes, from the mosquito dataset (see paragraph 2.2.3 for a description of this mosquito data), and the environmental variables is assessed point based at each mosquito sample location. Amongst others, the correlation is expressed in Pearson's correlation coefficient, r, and its squared version, R^2. The 'r-squared' is a common way of expressing a correlation in sciences (Shieh 2010). The correlation of quantitative environmental variables, like slope or elevation, can easily be assessed with the R^2. There are no qualitative environmental variables, since these were all transformed into a quantitative version. The locations of open water bodies are for example not quantitative, however the distance to these open water bodies is a quantitative variable.

Analyzing the relationship between mosquito data and the environmental data involved first a study to inner-correlations in the independent variables, so the correlations between the environmental variables. Using amongst others Principal Component Analysis (PCA), redundancy in the environmental variables dataset is detected. PCA is a statistical method to describe a large amount of data with a limited subset of the same data, i.e. finding principal components (which consist of several original components) that explain most of the variation that occurs. The second step was assessing the direct relationship of one environmental variable to the mosquito dataset. This is done using single linear regression. Using Multiple linear regression analysis the relation is analyzed of more than one environmental variables to the mosquito dataset. All these correlation analysis are performed in R.

5.2 Results

5.2.1 Removal of two points

Having a first look to the data, it appeared that two points were located outside the study area. These points contained zero values for the environmental variables, which can be seen in Figure 4. Figure 4 shows that these two points lay close to each other south of Rusinga Island in the water. These two points were removed from the dataset and were not taken into account in any of the analysis.

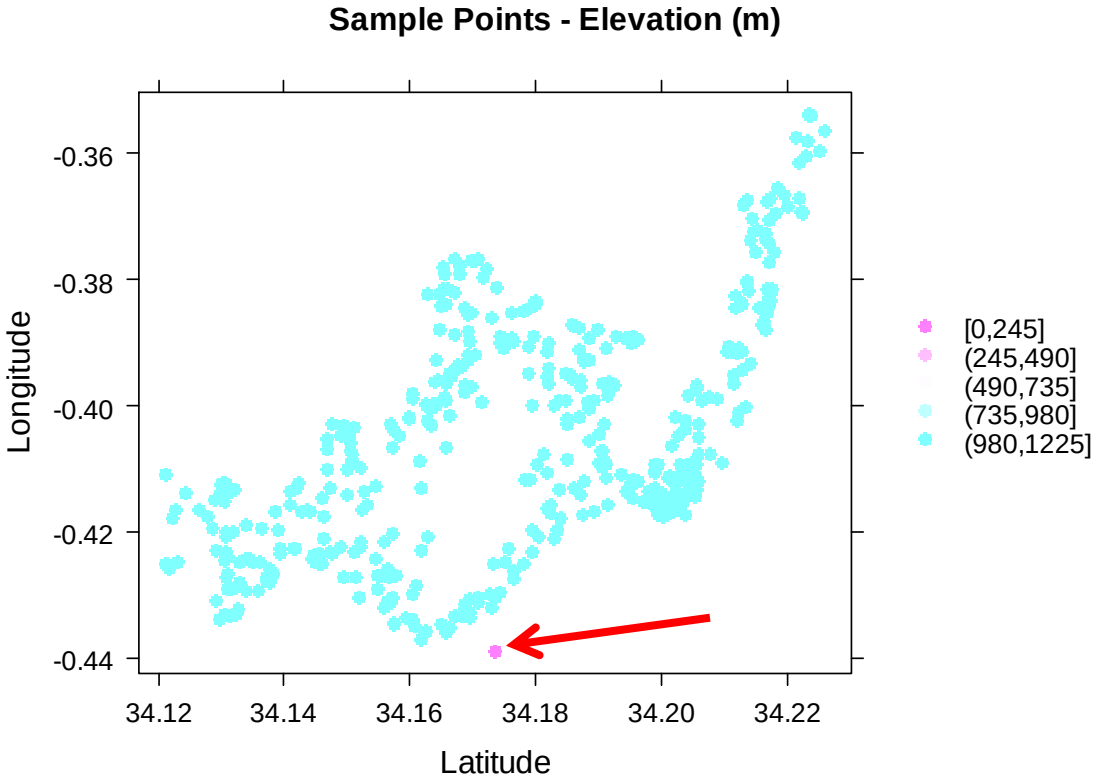


Figure 4: Elevation values at the sample locations

5.2.2 Inner-correlations in datasets

Before looking at the correlations between the dependent and the independent variables, the inner-correlations of the (in)dependent variables were assessed. Table 3 summarizes the correlations of the dependent variables, the mosquito groups, with each other. Table 4 and Table 5 represents the same as Table 3, however with subsequently only females selected and only males selected. Table 6 does this for the independent variables, the environmental variables. The linear correlations are expressed in Pearsons' coefficient. For the dependent variables, it appears that the *Culex* is highly correlated with the total group (AllMosq). This could be expected since the Culex mosquito is caught the most (see Table 2). Discarding the AllMosq group, the mosquito groups seems not to be correlated with each other. This means that the final model must have to deal with different prediction models for the different mosquito species.

Table 2: Number of trapped mosquitoes per mosquito group

Mosquito group	AllMosq		VectMosq	An.gambia		An.funestus		Culex		Mansonia		Aedes		Others
nr. of mosquitoes trapped	3805		474	120		402		2840		379		43		21
percentage of AllMosq	100%		12%	3%		11%		75%		10%		1%		1%
subdivison on gender (f/m)	f	m	f	f	m	f	m	f	m	f	m	f	m	-
nr. of mosquitoes trapped	3273	511	474	106	14	368	34	2394	446	363	16	42	1	-
percentage of AllMosq	86%	13%	12%	3%	0%	10%	1%	63%	12%	10%	0%	1%	0%	-
percentage of mosquito group	86%	13%	100%	88%	12%	92%	8%	84%	16%	96%	4%	98%	2%	-

For the independent variables, some high inner-correlation is visible. The population density and the density of households are highly correlated with a correlation coefficient of 0,99. It is logical that the more households there are the more people there live. This redundancy of data could be dealt with by removing one of the two variables out of the dataset. The three spectral indices (EVI, NDVI and NDWI) are also highly correlated with each other; values for correlation coefficient from 0,96 till 0,99. For these three variables, it is an option to select only one variable for regression analysis. The TWI is moderately correlated with the slope (0,42) and with the flow accumulation (0,54), which is due to the fact that the TWI is calculated based on these two variables. The TWI can be seen as a combination of the slope and the flow accumulation and will be kept in this study, since there is no real redundancy here. Figure 5 shows these inner-correlations in one graph. This figure is a biplot, produced using principal component analysis. The X-axis is the first principal component, which is a combination of variables and does explain the variation as far as possible with one component. The Y-axis is the second principal component. Two vectors representing the household- and population density show high correlation since their direction and length of the vectors are almost the same. The same counts for the spectral indices. Mind that NDWI is only negative correlated with the other spectral indices, however is of almost the same length and (negative) direction. The TWI here is visible as a vector on its own, the vectors of slope and flow accumulation are clearly of another length and direction.

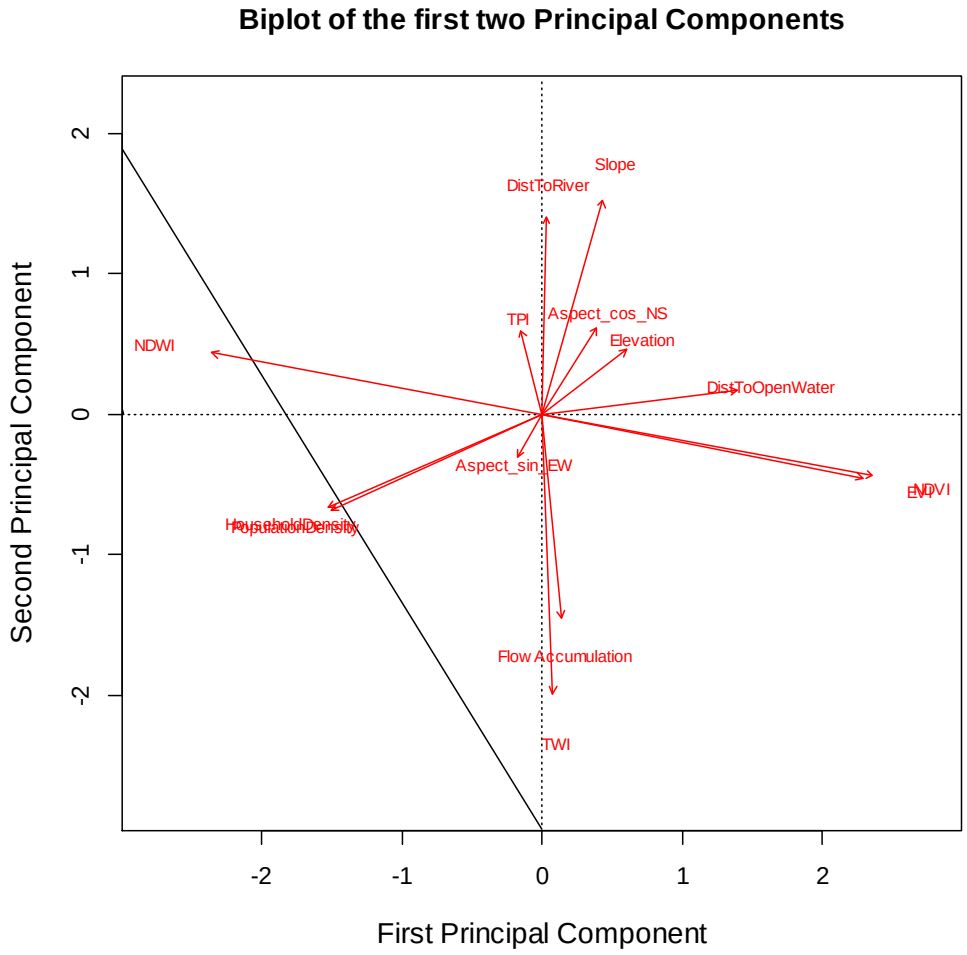


Figure 5: Biplot of the first two Principal Components

Table 3: Correlation coefficients within the mosquito groups (both female and male)

Inner-correlations: Mosquito groups (female & male)	AllMosq	VectMosq	An. gambia	An. funestus	Culex	Mansonia	Aedes
AllMosq	1	0,23	0,35	0,91	0,40	0,11	0,11
VectMosq	0,23	1	0,07	0,12	0,04	0,13	0,08
An.gambia	0,35	0,07	1	0,04	0,08	0,01	0,09
An.funestus	0,91	0,12	0,04	1	0,16	0,04	0,03
Culex	0,40	0,04	0,08	0,16	1	0,04	0,09
Mansonia	0,11	0,13	0,01	0,04	0,04	1	0,03
Aedes	0,11	0,08	0,09	0,03	0,09	0,03	1

Table 4: Correlation coefficients within the mosquito groups (only female)

Inner-correlations: Mosquito groups (only female)	AllMosq	VectMosq	An. gambia	An. funestus	Culex	Mansonia	Aedes
AllMosq	1	0,42	0,24	0,38	0,89	0,44	0,12
VectMosq	0,42	1	0,34	0,96	0,07	0,09	0,04
An.gambia	0,24	0,34	1	0,08	0,14	0,03	0,11
An.funestus	0,38	0,96	0,08	1	0,03	0,08	0,01
Culex	0,89	0,07	0,14	0,03	1	0,18	0,04
Mansonia	0,44	0,09	0,03	0,08	0,18	1	0,04
Aedes	0,12	0,04	0,11	0,01	0,04	0,04	1

Table 5: Correlation coefficients within the mosquito groups (only male)

Inner-correlations: Mosquito groups (only male)	AllMosq	VectMosq	An. gambia	An. funestus	Culex	Mansonia	Aedes
AllMosq	1	-	0,16	0,28	0,97	0,11	0,01
VectMosq	-	-	-	-	-	-	-
An.gambia	0,16	-	1	-0,01	0,07	-0,01	-0,00
An.funestus	0,28	-	-0,01	1	0,10	-0,01	-0,00
Culex	0,97	-	0,07	0,10	1	-0,01	-0,01
Mansonia	0,11	-	-0,01	-0,01	-0,01	1	-0,00
Aedes	0,01	-	-0,00	-0,00	-0,01	-0,00	1

Table 6: Correlation coefficients within the environmental (independent) variables

Inner-correlations: Environmental Variables	TPI	Aspect_cos_NS	Aspect_sin_EW	DistToOpenWater	DistToRiver	Elevation	FlowAccumulation	TWI	Slope	PopulationDensity	HouseholdDensity	EVI	NDWI	NDVI
TPI	1	-0,08	-0,01	0,04	0,08	0,35	-0,06	-0,33	-0,17	0,08	0,07	-0,03	0,02	-0,02
Aspect_cos_NS	-0,08	1	-0,09	0,13	0,10	0,06	-0,06	0,02	0,04	-0,26	-0,22	0,00	-0,01	0,01
Aspect_sin_EW	-0,01	-0,09	1	0,00	-0,01	0,00	0,06	0,01	0,05	0,14	0,13	-0,02	0,00	0,00
DistToOpenWater	0,04	0,13	0,00	1	-0,04	0,22	-0,02	-0,01	0,10	-0,26	-0,27	0,34	-0,39	0,37
DistToRiver	0,08	0,10	-0,01	-0,04	1	0,25	-0,18	-0,31	0,24	0,04	0,06	0,03	0,01	0,01
Elevation	0,35	0,06	0,00	0,22	0,25	1	-0,07	-0,24	0,27	-0,26	-0,27	0,07	-0,12	0,10
FlowAccumulation	-0,06	-0,06	0,06	-0,02	-0,18	-0,07	1	0,46	-0,09	-0,06	-0,06	0,03	-0,04	0,03
TWI	-0,33	0,02	0,01	-0,01	-0,31	-0,24	0,46	1	-0,41	-0,02	-0,02	0,00	-0,01	0,01
Slope	-0,17	0,04	0,05	0,10	0,24	0,27	-0,09	-0,41	1	-0,17	-0,18	0,04	-0,05	0,04
PopulationDensity	0,08	-0,26	0,14	-0,26	0,04	-0,26	-0,06	-0,02	-0,17	1	0,99	-0,22	0,27	-0,26
HouseholdDensity	0,07	-0,22	0,13	-0,27	0,06	-0,27	-0,06	-0,02	-0,18	0,99	1	-0,24	0,28	-0,27
EVI	-0,03	0,00	-0,02	0,34	0,03	0,07	0,03	0,00	0,04	-0,22	-0,24	1	-0,96	0,98
NDWI	0,02	-0,01	0,00	-0,39	0,01	-0,12	-0,04	-0,01	-0,05	0,27	0,28	-0,96	1	-0,99
NDVI	-0,02	0,01	0,00	0,37	0,01	0,10	0,03	0,01	0,04	-0,26	-0,27	0,98	-0,99	1

5.2.3 Single Linear Regression Analysis

Table 7 gives an overview of Pearsons' coefficient for each independent variable correlated with each dependent variable. The resulting correlation coefficients are low, with an highest coefficient of 10,5% found for the slope relating to all mosquitoes grouped together. This means that only 10% is correlated with each other. Looking at the predictive power of variable Slope, a R^2 value of 0,011 is found in relation to all mosquitoes grouped together. The R^2 value can be seen as a value for the predictability of the dependent variable using only the independent variable. This means that only 1,1% of the total variation in the dependent variable can be predicted by the independent variable, using a linear relationship. The R^2 values for all the other combinations of variables were all in the order of 0,01 – 0,1%.

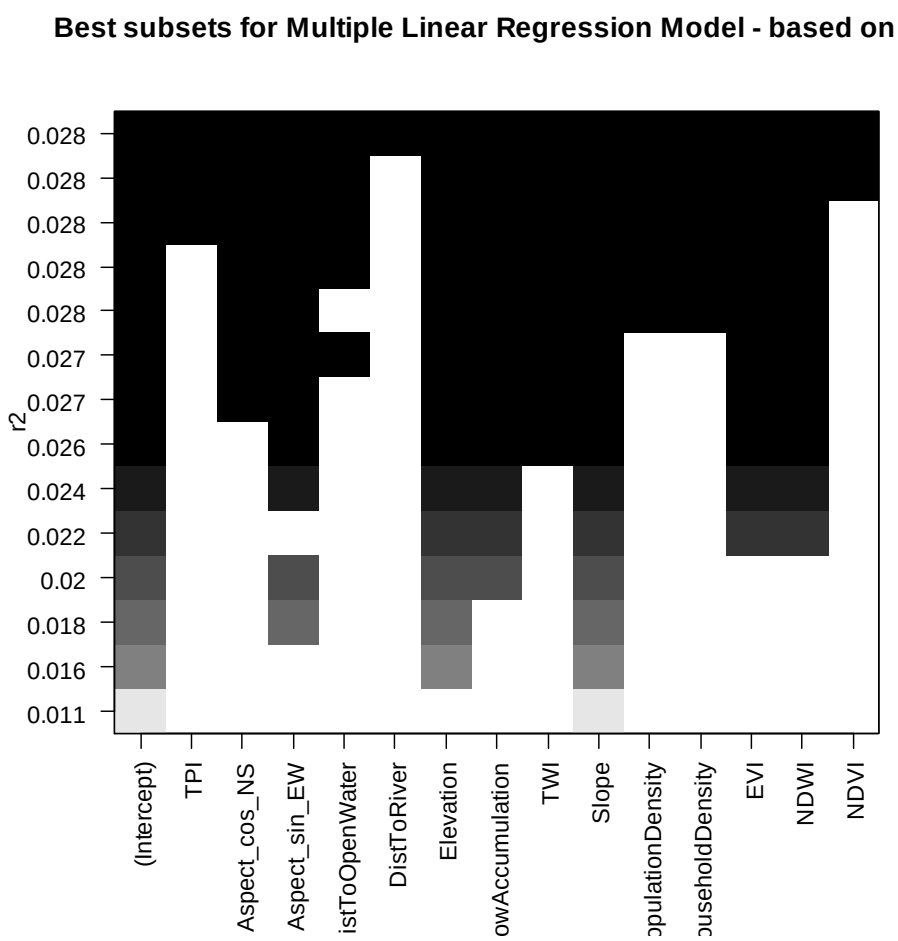
Table 7: Correlation coefficients between one dependent and one independent variable

Linear correlations (only female)	AllMosq	VectMosq	An. gambia	An. funestus	Culex	Mansonia	Aedes
TPI	-0,010	0,049	0,035	0,042	-0,037	0,014	0,002
Aspect_cos_NS	0,024	0,074	0,009	0,076	0,008	-0,021	-0,035
Aspect_sin_EW	-0,048	-0,039	0,028	-0,049	-0,043	0,015	-0,068
DistToOpenWater	-0,050	-0,021	-0,001	-0,021	-0,061	0,027	-0,002
DistToRiver	-0,036	-0,021	-0,070	-0,003	-0,043	0,031	-0,022
Elevation	-0,098	-0,058	-0,010	-0,059	-0,085	-0,025	-0,031
FlowAccumulation	0,056	0,007	-0,022	0,013	0,074	-0,028	0,004
TWI	0,041	-0,037	0,009	-0,041	0,071	-0,023	0,019
Slope	-0,105	-0,068	-0,057	-0,056	-0,089	-0,025	-0,046
PopulationDensity	0,040	-0,063	-0,034	-0,057	0,064	0,042	-0,035
HouseholdDensity	0,041	-0,052	-0,026	-0,048	0,062	0,037	-0,031
EVI	-0,009	-0,051	0,044	-0,066	-0,003	0,039	-0,005
NDWI	0,030	0,072	-0,037	0,087	0,021	-0,045	0,004
NDVI	-0,022	-0,058	0,035	-0,071	-0,017	0,045	-0,007

5.2.4 Multiple Linear Regression Analysis

Multiple regression modeling is the use of more than one independent variables to predict one dependent variable. Using all available independent variables for the prediction of one dependent variable might result in the highest possible R^2. However, due to redundancy in the independent variables, the model can be too complex and can contain an unnecessary number of variables. The adjusted R^2 takes care of redundancy and indicates whether an extra variable is still useful in a model.

The leaps package provides a method within R to find the best subset of independent variables with a limited number of variables. It simply goes through all possible subsets of variables and selects the best subset based on a given measure (Miller 2002). Figure 6 shows the best subsets for prediction of all mosquitoes together, expressed in R^2 and Figure 7 does this for the R^2 adjusted value. It can be concluded that selecting, based on the highest adjusted R^2 value, only the aspect (twice: NS- and EW-component), elevation, flow accumulation, TWI, slope, EVI and NDWI as independent variables results in a better (and simpler model) than selecting all the independent variables in one prediction model.



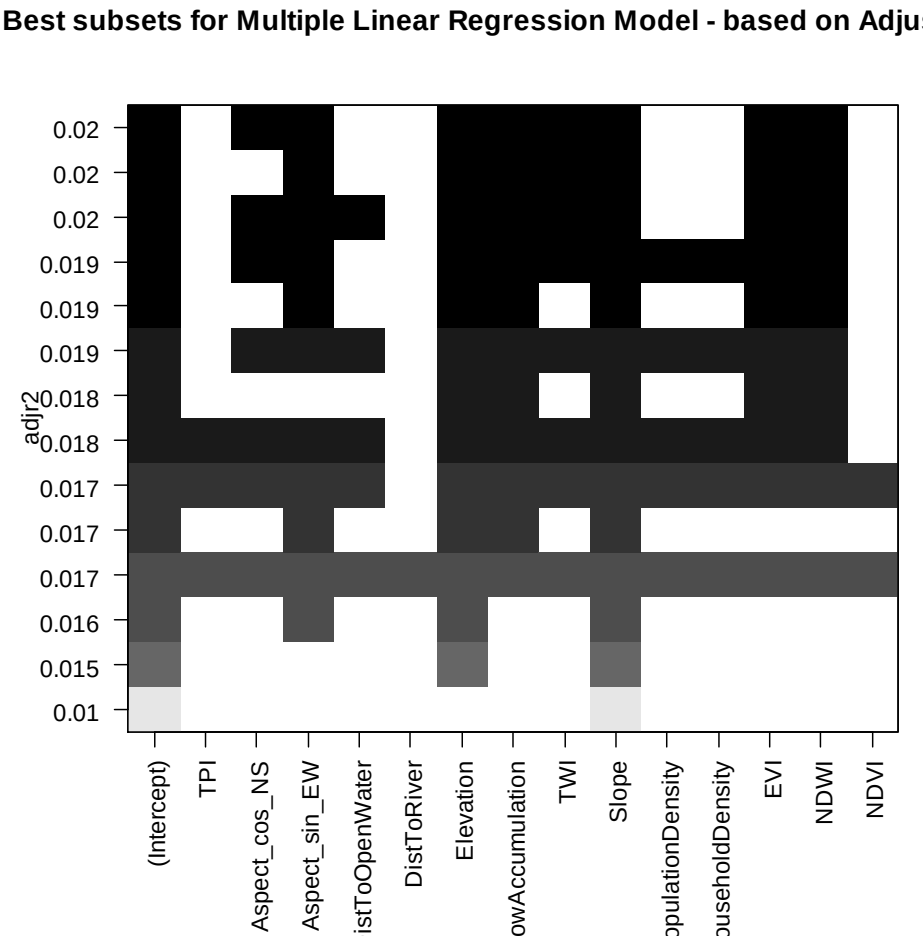


Figure 6: (Left) Selecting best subsets for predicting AllMosq (only female) based on R2 value, using leaps package

Figure 7: (Right) Selecting best subsets for predicting AllMosq (only female) based on adjusted R2 value, using leaps package

Table 8 summarizes the best subsets (selection based on R^2 adjusted value) for each mosquito group and gives the resulting R^2 and R^2 adjusted values. The highest values found are with the prediction of mosquito group *An. funestus*, with a R^2 adjusted value of only 3,83%.

Table 8: Overview of best subset for prediction of mosquito groups, based on R^2 adjusted value

Best subset for Multiple Linear Regression Model (based on R2adj)																	
Mosquito groups (only female)	Results			Environmental. Variables													
	R2	R2adj	AIC	TPI	Aspect_cos_NS	Aspect_sin_EW	DistToOpenWater	DistToRiver	Elevation	FlowAccumulation	TWI	Slope	PopulationDensity	HouseholdDensity	EVI	NDWI	NDVI
AllMosq	2,70%	2,04%	6.593,7		x	x			x	x	x	x			x	x	
VectMosq	4,44%	3,63%	4.006,5	x	x				x	x	x	x	x	x		x	x
Gambia	2,22%	1,48%	880,2	x		x		x		x		x	x	x	x		x
Funestus	4,55%	3,83%	3.864,9	x	x				x	x	x	x	x			x	x
Culex	2,70%	1,88%	6.131,5	x	x	x	x			x		x	x	x	x		x
Mansonia	0,97%	0,46%	3.440,3					x	x				x	x	x		x
Aedes	1,00%	0,66%	-336,3		x	x						x	x				

The MASS package provides a similar method as the leaps package, however is searching to the best subset, based on the AIC value. It starts with all variables into one model and removes step by step the variable that has least value (Venables and Ripley 2000). Table 9 summarizes the best subsets (selection based on AIC value) for each mosquito group and gives the resulting AIC and R^2 adjusted values. The resulting R^2 adjusted values look quite similar to the ones of Table 8, however small differences can be seen. Overall the R^2 adjusted values are again very low.

Table 9: Overview of best subset for prediction of each mosquito group, based on the AIC value

Best subset for Multiple Linear Regression Model (based on AIC)																	
Mosquito groups (only female)	Results			Environmental. Variables													
	R2	R2adj	AIC	TPI	Aspect_cos_NS	Aspect_sin_EW	DistToOpenWater	DistToRiver	Elevation	FlowAccumulation	TWI	Slope	PopulationDensity	HouseholdDensity	EVI	NDWI	NDVI
AllMosq	2,43%	1,94%	6.592,9		x				x	x		x			x	x	
VectMosq	4,17%	3,53%	4.005,8		x				x	x	x	x	x			x	x
Gambia	1,94%	1,36%	879,6	x		x		x					x	x	x		x
Funestus	4,46%	3,82%	3.864,1		x				x	x	x	x	x			x	x
Culex	2,42%	1,85%	6.128,9	x		x				x		x	x		x		x
Mansonia	0,51%	0,34%	3.437,8										x				x
Aedes	1,00%	0,66%	-336,3		x	x						x	x				

5.2.5 Non-Linear Regression Analysis

Analyzing whether there is a non-linear relationship between some variables needs some first ideas about the expected relationship. So far, linear relationships were mainly expected between the variables, however this does not appear to be the situation. Looking at the scatterplots of each variable against all other variables it does not point to a certain direction to think about. Figure 8 for example shows a scatterplot with the Slope against the mosquito group AllMosq (=all mosquito types grouped together), where a kind of cloud is visible, meaning that there is probably not a clear relationship between the two variables. See Appendix B for more scatterplots like Figure 8.

There is one more issue to point at in Figure 8, namely that there seems to be a relationship between the slope and the maximum possible occurrence of mosquitoes. The top of the points cloud looks like a linear or exponential relationship. The direction of this relationship is indicating that the higher the slope of an area, the lower the maximal number of mosquitoes can be expected. This sounds as expected (the higher the slope, the easier drainage of water, the dryer the area, the less water available for mosquitoes, the less mosquitoes) however, a relationship with the maximal number of mosquitoes that can occur, including some accuracy, is not really what was looked for. If there is a place with a high maximal number of mosquitoes expected, the real number of mosquitoes is more likely to be the average expected number of mosquitoes. However, the slope provides some information regarding the possibility of finding adult mosquitoes at a house, albeit with a wide range of predicted values/low predictive accuracy.

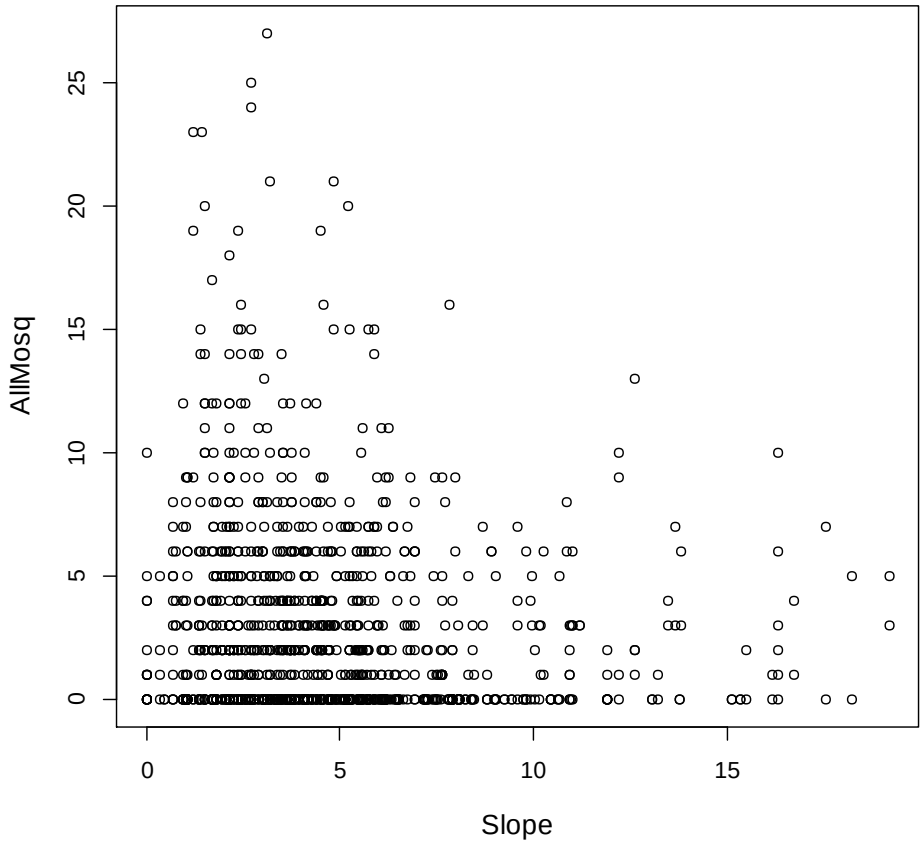


Figure 8: Scatterplot with the independent variable Slope on the X-axis, the dependent variable AllMosq (only female) on the Y-axis

6 Mosquito data description

6.1 Methodology

To study whether the mosquito dataset is representative for whole Rusinga Island, it was checked how well the sample points are distributed over the island. This was checked in the spatial and temporal dimension. To do this, first general statistics on the whole mosquito dataset were performed. Secondly, the temporal aspect was studied by dividing the data in months and sampling rounds and the average catch sizes were plotted per month and per sampling round in graphs. Thirdly, the spatial coverage of the mosquito dataset was assessed using histograms and violin plots. A violin plot is a combination of a boxplot and a Kernel density plot (Hintze and Nelson 1998). Finally, the spatial and temporal aspect were combined into distribution maps of the mosquito data per month. These maps show the sample points within one month or one sampling round, the catch sizes at these points and an interpolation of it covering the whole island.

6.2 Results

The mosquito dataset so far is gathered in the period from September 2012 until August 2013, so one year of data, forming a baseline year of mosquito data (see also the data description in 2.2.3). During these twelve months, a total of 3805 mosquitoes were caught in 1190 traps, at approximately 600 locations. All the caught mosquitoes were analyzed for their species type. The species *An. gambiae* and *An. funestus* are malaria vector species, the others not. For the analysis, the malaria vector species were grouped into one group called 'VectMosq', containing only the female mosquitoes of *An. gambiae* and *An. funestus*. The group 'AllMosq' is just all mosquito types grouped together, no matter they are vector of the malaria disease, however only female mosquitoes. Only female mosquitoes are dealt with since the traps are designed for attracting female mosquitoes that are hunting for a blood meal. Any male mosquito that is caught by a trap happened by accident, the reason why only 13% of the caught mosquitoes are male. Table 10 shows the numbers of mosquitoes caught for each mosquito type and group. Only 12% of the caught mosquitoes were vectors of *Plasmodium*. A large proportion of the caught mosquitoes are of the Culex type.

Table 10: Number of trapped mosquitoes per mosquito group

Mosquito group	AllMosq		VectMosq	An.gambia		An.funestus		Culex		Mansonia		Aedes		Others
nr. of mosquitoes trapped	3805		474	120		402		2840		379		43		21
percentage of AllMosq	100%		12%	3%		11%		75%		10%		1%		1%
subdivison on gender (f/m)	f	m	f	f	m	f	m	f	m	f	m	f	m	-
nr. of mosquitoes trapped	3273	511	474	106	14	368	34	2394	446	363	16	42	1	-
percentage of AllMosq	86%	13%	12%	3%	0%	10%	1%	63%	12%	10%	0%	1%	0%	-
percentage of mosquito group	86%	13%	100%	88%	12%	92%	8%	84%	16%	96%	4%	98%	2%	-

6.2.1 Temporal description

Table 11 shows the mosquito data per month, and shows that the traps were not equally distributed over the months. On average, a number of 99 traps were placed each month, probably at around 50 locations, since there was one outdoor and one indoor measurement per location. This is because the traps were placed during the sampling rounds of 6 weeks. Therefor the data is also summarized per sampling round in Table 12. Figure 9 provides a quick visual interpretation of what is given in Table 11, Figure 10 the visual interpretation of Table 12. There is a remarkable variation in catch size over time. The proportion of vector mosquitoes out of all the mosquitoes (red line in Figure 9 and Figure 10) is extremely varying during the year from 5% to 47%.

Table 11: Number of trapped mosquitoes (only female) per month

Month	Total	Sep12	Oct12	Nov12	Dec12	Jan13	Feb13	Mar13	Apr13	May13	Jun13	Jul13	Aug13
Number of traps	1.190	60	176	120	40	116	60	100	80	118	100	150	70
All Mosquitoes	3.273	103	185	590	163	356	116	353	173	620	358	190	66
All Mosquitoes per trap	2,8	1,7	1,1	4,9	4,1	3,1	1,9	3,5	2,2	5,3	3,6	1,3	0,9
Vector Mosquitoes	474	48	46	127	33	46	13	28	42	39	19	20	13
Vector Mosquitoes per trap	0,4	0,8	0,3	1,1	0,8	0,4	0,2	0,3	0,5	0,3	0,2	0,1	0,2
% Vector Mosquitoes	14%	47%	25%	22%	20%	13%	11%	8%	24%	6%	5%	11%	20%

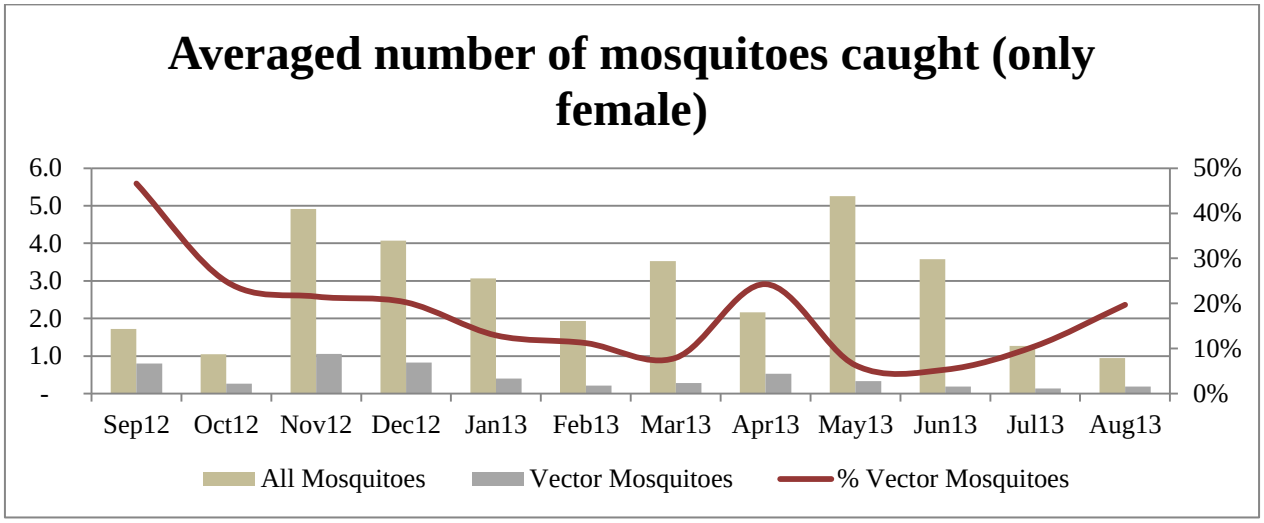


Figure 9: Trapped mosquitoes in time by month

Table 12: Number of trapped mosquitoes (only female) per sampling round

Round	Total	1	2	3	4	5	6	7
Number of traps	1.190	236	160	156	160	158	160	160
All Mosquitoes	3.273	288	753	423	488	707	475	139
All Mosquitoes per trap	2,8	1,2	4,7	2,7	3,1	4,5	3,0	0,9
Vector Mosquitoes	474	94	160	58	60	50	31	21
Vector Mosquitoes per trap	0,4	0,4	1,0	0,4	0,4	0,3	0,2	0,1
% Vector Mosquitoes	14%	33%	21%	14%	12%	7%	7%	15%

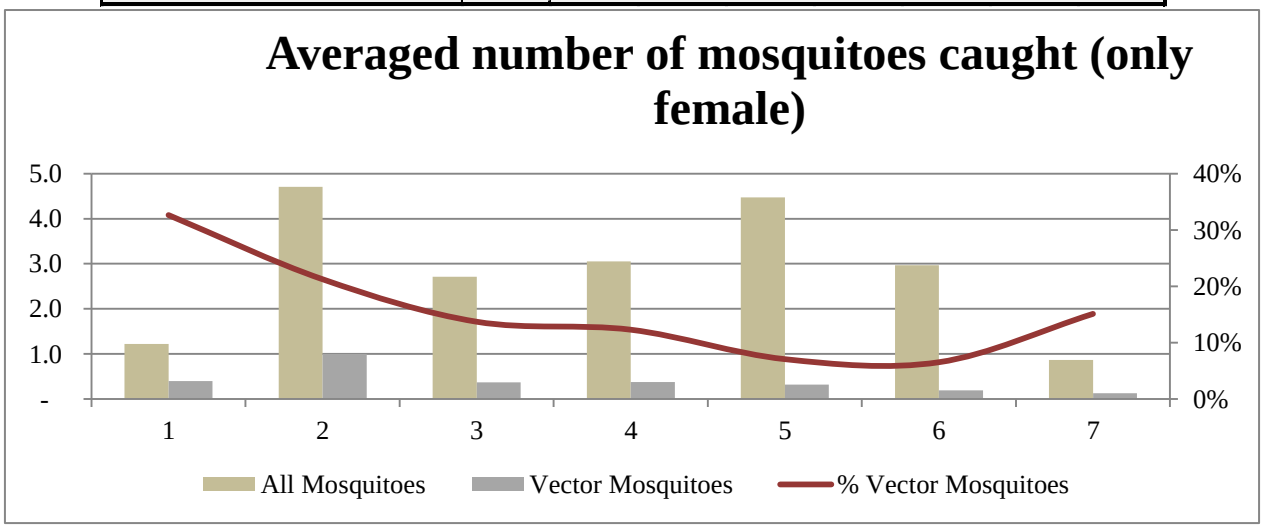


Figure 10: Trapped mosquitoes in time by sampling round

6.2.2 Spatial description: Distribution maps

The spatial distribution of the mosquito dataset is represented in Figure 11 (only malaria vector mosquitoes) and Figure 12 (all mosquitoes). In these figures the points are the locations of the traps, their size indicate the catch size and the colors in between an indication of the spatial interpolation of these data using IDW technique. The spatial distribution of all the mosquitoes looks quite randomly over the island. However, not for the spatial distribution of only the vector mosquitoes, were in the northern and eastern part many points are with no mosquitoes caught (black dots) and here the interpolated mosquito occurrence is quite low (dark green). The other side of the island shows a mix of black dots and high catch sizes. The interpolated areas with high mosquito occurrences are found in the western and northwest side of the island. The middle of the island is not sampled, due to the lack of houses over there.

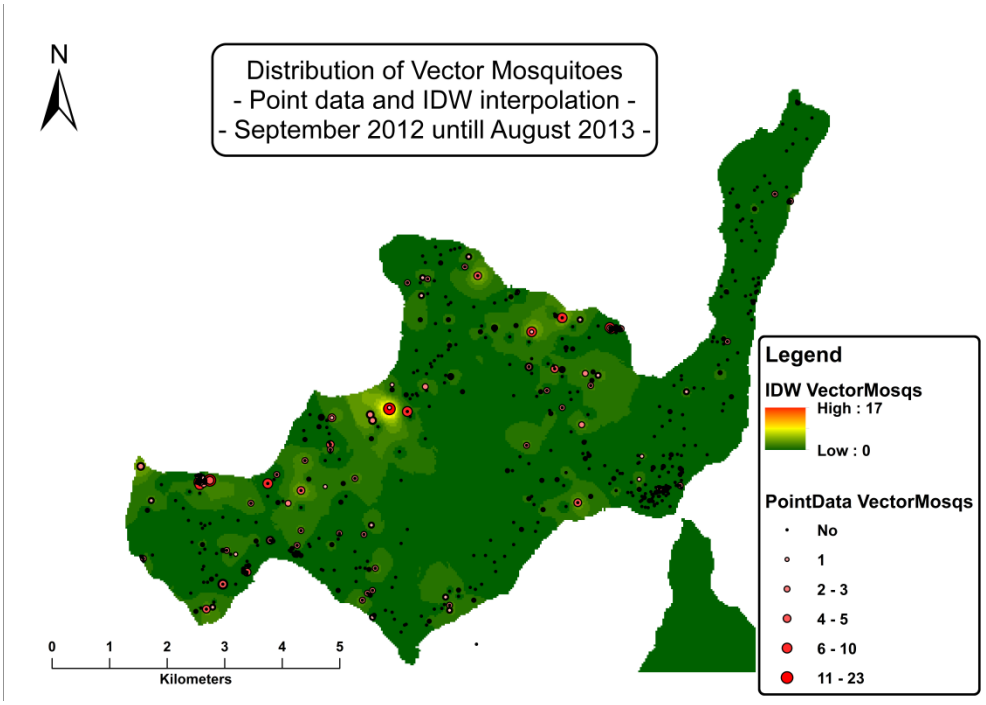


Figure 11: Spatial distribution of vector mosquito data (only female) with IDW interpolation

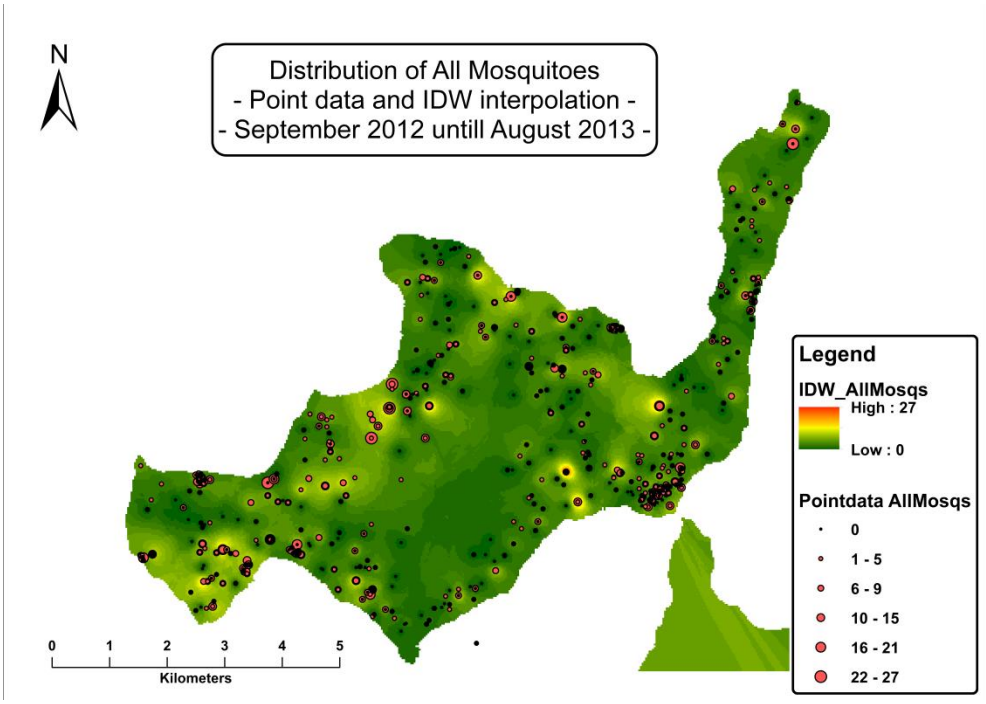


Figure 12: Spatial distribution of all mosquito data (only female) with IDW interpolation

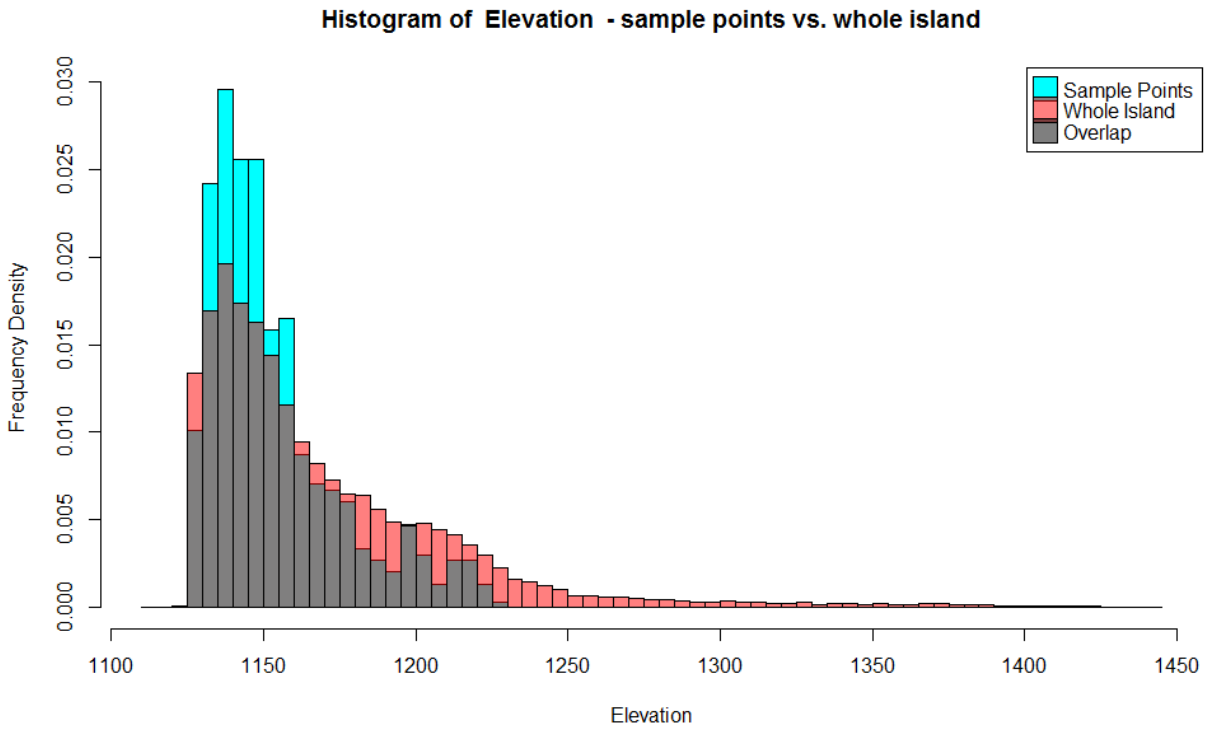


Figure 13: Combined histogram of Elevation

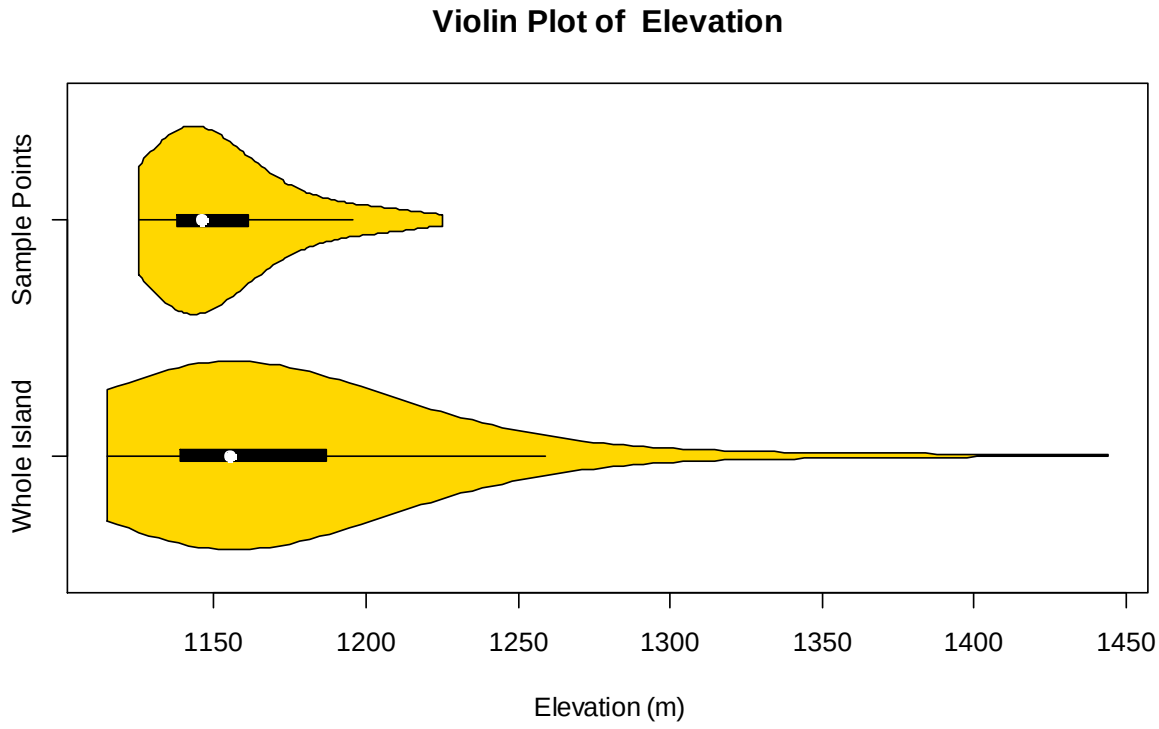


Figure 14: Violin plot of Elevation

6.2.3 Spatial description: Histograms & Violin Plots

Figure 11 and Figure 12 show that not whole Rusinga Island is covered by the mosquito trap measurements. Especially in the middle of the island, were no houses are located, there is also no mosquito data available. It is therefore good to have a look at the spatial distribution of the trap locations compared to the whole island. The environmental variables that were created are now used to describe the positions of the trap locations, compared with the characteristics of the whole island. Figure 13 shows a combination of two histograms, one for the sample points (the locations of the traps) and one for the whole island, both for the environmental variable Elevation. The grey area in this figure represents the overlap between the two separate histograms. The sample points cover heights from 1125 (level of Lake Victoria) until 1225, while the island is ranging until 1425. The blue area is peaking above the grey area, indicating that there are relatively more areas with elevations around 1150 compared with the elevation statistics of the whole island, indicating a bias in the dataset. The mosquito dataset is not covering the whole variation that is present on the island and is not a perfect subset of the whole island. This is in relation to the SolarMal project objectives not a big issue since it is focusing on the public health, so around the houses. For correlation purposes and finding relations with the environment on Rusinga Island this does matter since there is an interested in the whole island. In Appendix C, the combined histograms for all the environmental variables are given. Overall, the mosquito dataset is lacking the extremes of the island.

Instead of looking at a histogram, Figure 14 shows a Violin plot, which is just another representation of the two datasets and their variation. A violin plot is a combination of a boxplot and a Kernel density plot (Hintze and Nelson 1998). For the variation of elevation in the mosquito dataset, it appeared that the median is lower than the median of the whole island. Compared to the variation on the island, the mosquito dataset contains less variation (closer quartiles) and is restricted in extreme heights.

In Appendix C next to the histograms, also the violin plots are given for each environmental variable. Both violin plot and histogram are given since the violin provides a clear visualization of the relative difference in distributions, while the histogram adds the frequency values. Most of the violin plots show the same pattern as here for the elevation; extremes are missing, a shift in the median of the values and therefor a bias of the mosquito dataset. Two violin plots are that interesting that they deserve some special attention here. The left violin plot in Figure 15 visualizes the variation in household densities. Where the variation overall the island is limited to a median around 2 houses per hectare and some extreme values around 20 houses per hectare, the variation in the mosquito dataset shows almost two groups of most occurring values. The household density is mainly low, around 3 households per hectare (mind that this is already one more than the median of the whole island!), however there are relatively many locations that are within high dense areas. The right violin plot in Figure 15 visualizes the variation in distance to open water (Lake Victoria for example). The average mosquito trap was much more located close to open water than an average point on the island. This can be explained by the fact that almost all the houses are located close to Lake Victoria.

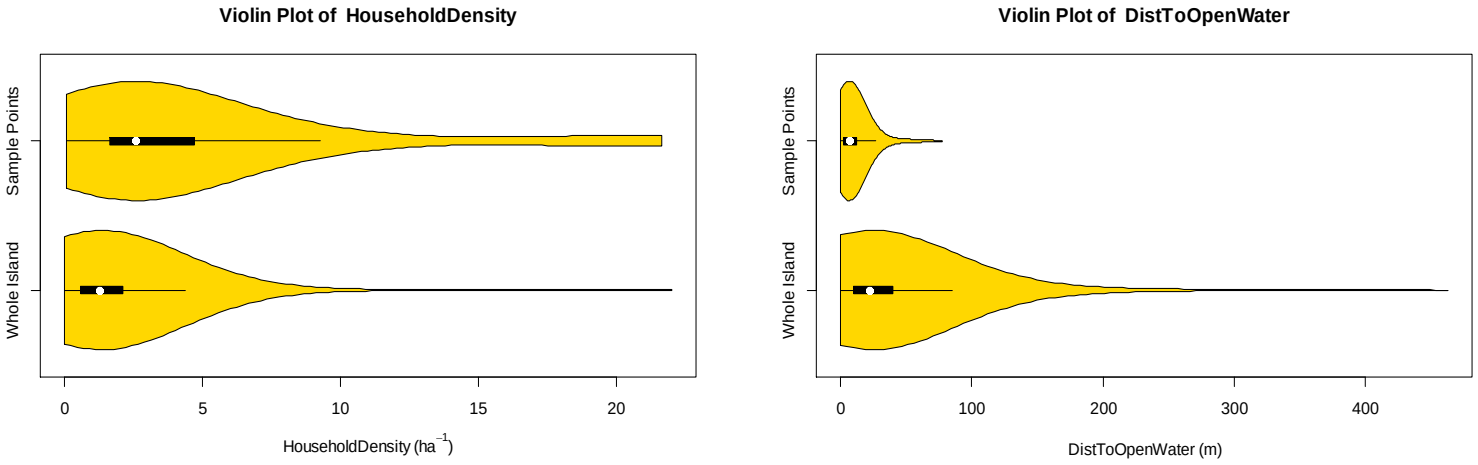


Figure 15: Violin plots of the Household Density and the Distance to Open Water

6.2.4 Spatial-Temporal description: Distribution Maps per month

In Appendix D the spatial distribution of the mosquito dataset is presented like is done in Figure 11 and Figure 12, however now for each month separately. Looking at the 12 months that are available, it becomes clear that the spatial distribution of the mosquitoes is varying over time. For some months, it appears that there is quite a limited number of sample points available like for September, December and February. This has to do that within each sampling round of 6 weeks, on average the last two weeks were used for processing of the data. Appendix E therefor shows the spatial distribution of the mosquito dataset per sampling round, ensuring a minimum amount of samples on each map. Overall, it appears that there is a small preference of the vector mosquitoes for the northwestern side of the island. The distribution of all mosquitoes together shows, like the vector mosquitoes, a high variation during the year. A preference for a side of the island is not clear for the distribution of all mosquitoes together.

7 Study to the accuracy of house positions measured

As supplementary study for the SolarMal project and as study to the usefulness of the spatial data used in this thesis, the accuracy of the measured house positions were validated. Since low correlations were found in chapter 5, it was desired to check how accurate the spatial input data was. The geographical positions of the mosquito dataset were measured with tablets, which could involve a large spatial inaccuracy. Large inaccuracies in this input dataset could be the cause for the low correlations found. This chapter first summarizes briefly the study and the outcomes of it. Secondly, this chapter discusses the accuracy of the data used in this thesis project, based on the findings of the extra study. The complete report on the study performed, can be found in Appendix F.

7.1 Background and objectives

7.1.1 Introduction

The dataset containing the coordinates of residential structures on Rusinga Island has a distribution of spatial accuracy of which the magnitude was unknown. Fieldworkers of the SolarMal project reported that some houses could not be located using the associated GPS coordinates, causing them to request help of local people and find houses using the names of household members. The inaccuracy of the geographic positions for some households has such a magnitude that the positions are not suitable for navigation purposes. Inaccuracy also affects the spatial analysis of the data that belongs to the houses, and most importantly influences the spatial analysis of the effect of the intervention, the mosquito trap, on malaria transmission. The mosquito abundance analysis included in this thesis project, is also relying on the house locations. Inaccuracy in the precision of GPS locations leads to inaccuracy of the data analysis, leading to incorrect results and conclusions.

7.1.2 Objective

The objective of this study is to validate the accuracy of the house positions on Rusinga Island and to assess whether this accuracy can be improved. This study aims to answer the following questions:

RQ1. How accurate were the positions of the houses on Rusinga Island measured?

RQ2. Which method of measuring positions has the best accuracy and is feasible within the SolarMal project?

7.2 Experimental methods

For this study three GPS devices were available. First of all a Garmin eTrex 30 (Garmin Device), provided by the GIS department of the Wageningen University. The second device is a black Samsung tablet, model GT-P7510 (Black Tablet). The third device is a white Samsung tablet, a newer model, the SM-T210 (White Tablet). Both Samsung tablets are equipped with a WIFI and GPS receiver and were both provided by the SolarMal project.

To answer the two research questions, three main experiments are performed. The first experiment aims to investigate the accuracy of a GPS device in the spatial and temporal dimension. The Garmin Device was set up for 24 hours on a fixed place, measuring its position every 10 seconds. The resulting cloud of points was analyzed for their spatial distribution. The time dependency of the accuracy and precision was studied by analyzing the cloud of points per hour. The average distance to the real position of the experiment was used as indication for the accuracy of the measurement. The standard distance in the spatial distribution of the cloud of point was used as indication for the precision of the measurement.

The second experiment aimed to validate several methods for measuring the positions of houses on Rusinga Island. Amongst others, the three devices were used to measure two waypoints at each house (one inside the house and one outside the house). A waypoint is a set of coordinates in the two- or three-dimensional space which actually are the distances to a certain reference position. The outside measurements were corrected afterwards to the middle of the house by the recorded direction and distance from the middle of the house. The measured positions were validated by calculating the distance to the ‘true’ positions of the houses. A dataset of ‘true’ positions was created by using the QuickBird image

available for this project. The sampled houses were searched for on this map and manually the positions were drawn in to it. To confirm which house exactly was sampled, pictures were made in the field of the sampled houses so these could be relocated on the QuickBird image. In the data analysis, also the original measured house positions were included in order to validate their accuracy.

The third experiment followed on from the second experiment, where the most feasible and best options for this project were validated again, but for a larger set (100) of houses. Four methods were chosen to be the most appropriate and feasible for this project, based on the results of the second experiment. These four methods were the use of the two available tablets, measuring a waypoint both inside and outside each house.

7.3 Results and conclusions from experiments

From the first experiment it became clear that the best accuracy possible, using the Garmin device, can be less than 10 meters after one measurement; however it is at least 6 meters. 3 meter inaccuracy is explained by the spatial accuracy, which is the variation in the spatial dimension, the other 3 meter inaccuracy is caused by a temporal variation of the spatial accuracy (see Figure 16 for a visualization of this), making the total accuracy 6 meters. When measuring once, a precision of around 4 meters has to be taken into account, resulting in a total uncertainty of the position of 10 meters. If one would measure several times and average the position, the precision can be ignored and a total uncertainty of approximately 6 meters remains.

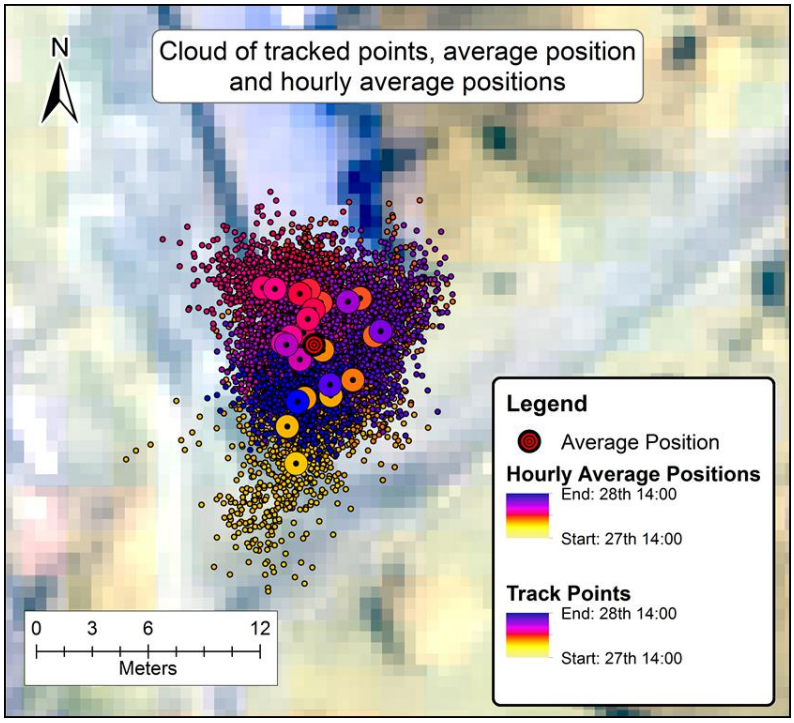


Figure 16: Cloud of points and hourly average positions, colored per hour

From the second and third experiments it became clear that the original dataset of house locations on average has an accuracy of 5-8 meters. The standard variation in this is 6-8 meters, so 97,5% of all positions have an accuracy less than 22 meters. One of the most suitable methods validated during the third experiment was measuring outside the house and subsequently correct for the distance and direction to the middle of the house afterwards. On average this method resulted in an accuracy of 6-7 meters, with a standard deviation of 4-5 meters. This means that 97,5% of all positions measured in this way have an accuracy less than 15 meters. A slightly less accuracy was found in the peri-urban area compared with the rural area. Houses in the peri-urban are built closer to each other, creating more reflection of satellite signals, resulting in a larger error. Figure 17 shows on the left an example of rural area, on the right on example of peri-urban area.

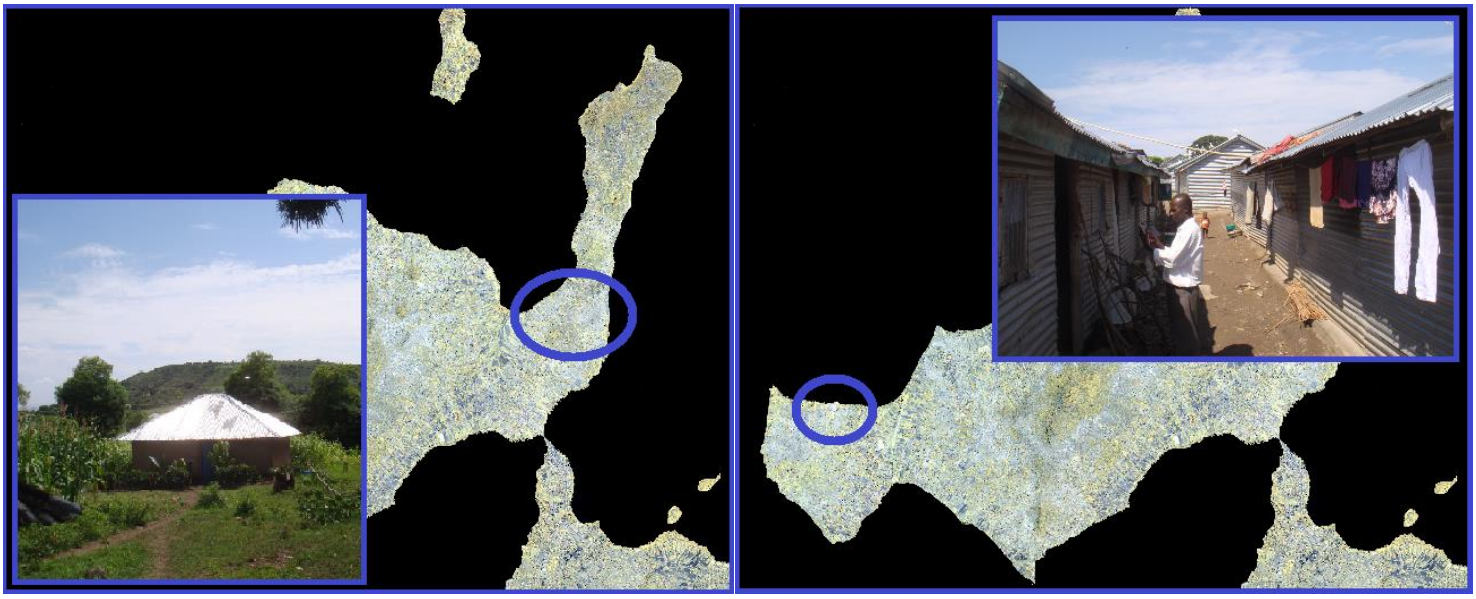


Figure 17: Areas of fieldwork for experiment 2 and two types of houses

7.4 Conclusion and discussion of the study in relation to the thesis project

An average accuracy of 6-7 meters, with a variation in this of another 4-5 meters would be fine for the rural areas, where the distance between two individual houses is at least 10 meters. However, the same accuracy is not sufficient enough for the peri-urban area, where within a distance of 15 meters more than 5 houses could be present. The question is whether it is even possible to get a high enough accuracy for the peri-urban area. An accuracy of maximal 15 meters, and an average of 6-7 meters, would at least limit the possible houses, since the coordinates will navigate to the right neighborhood.

For the spatial analysis conducted in this thesis, a point accuracy of maximal 22 meters seems large, while it is probably acceptable since the DEM has an resolution of 30 meters. However, in relation to the high-resolution QuickBird image (2,4 meters) it is fairly inaccurate. According to Wing et al. point accuracies of 5 meters can normally be expected using consumer-grade, top quality, GPS devices under clear-sky conditions and 10 meters under a closed canopy (Wing et al. 2005). An accuracy of less than 22 meters seems realistic from this since these measurements were performed using a simple tablet. The question arises, what accuracy is actually required. During this study, a great deal became clear about the way the population live on the island and how the environment in reality looks like. Guided by a fieldworker from the SolarMal project, typical breeding sites were detected on the island. Figure 18 shows examples for some of these breeding sites. It appeared that a breeding site is rarely a natural dip in the terrain where some water accumulates to a pool. Breeding sites are found in tire tracks (left picture in Figure 18), which can have a size of half a meter to a few meters. Sometimes an artificial pool was found near recently built houses and were full of larvae. These pools with a diameter of 1-2 meter are used as water reservoir during the construction of the house (lower right picture in Figure 18). After the construction, such a pool could be destroyed or filled with ground and sand, however some pools stay intact. Interesting are breeding sites that are found in the footprints of cattle in drenched grass (upper right picture in Figure 18), which have a size of only 10-20 cm. Additionally, it is striking that breeding sites are found in in waste along bushes and roads. Broken bottles or plastic bags can easily hold some water for a few days, enough for mosquitoes to be a breeding site. Such breeding sites can vary in size from a few centimeters to a few decimeters. Important to realize is that such breeding sites can be everywhere. If just somebody is throwing his waste in some bushes, a potential breeding site does appear. This all make it realizing that predicting the presence of breeding sites is probably not so much related to the natural environment (except for natural dips of course), rather to the activities of the human society. It can therefore be concluded that studies to the spatial distribution of mosquitoes have to take into account the way local people behave and what their living circumstances are.

Figure 18: Examples of typical breeding sites on Rusinga Island

High-resolution analysis in these kind of areas are not relevant any more, since the resolution is still not high enough. It is possible to search for breeding sites on the larger scale, with a focus on breeding sites presence in natural pools. It would also be possible to search for the typical small breeding sites like that shown in Figure 18 by searching for variables that can act as indicators of such breeding sites. The change for tire tracks for example could be indicated using a road map. For such purposes a DEM with a resolution of 30 meters would probably already be sufficient. However, a 30 m resolution DEM or a 2,4 m resolution satellite image is not suitable for direct detection of very small breeding sites like tire tracks, footprints of cows in drenched grass and waste in bushes. These breeding sites are not only of a very small size, they are also time-dependent. Waste dumped in bushes is an highly varying process which cannot be tracked by only one satellite image. For detection of these specific breeding sites, the use of satellite images and other remote acquired spatial data is nowadays too limited in the spatial and temporal space.

8 Conclusion, discussion and recommendation

This chapter presents the conclusions and related discussion per research question and for the overall objective. Afterwards recommendations considering data sampling, data and data analysis are given.

8.1 Conclusion and discussion

The first research question concerns which environmental variables could be determinants of the mosquito distribution according to literature. In order to answer this first research question, 22 environmental variables were found from literature that are somehow related to the occurrences of mosquitoes. These 22 environmental variables were expected to be the most logical and straight forward variables, since these were easily found in literature. Not all literature could be screened, so there could be more environmental variables to be found. The environmental variables found are summarized in Table 2. Based on the available data, a number of 14 environmental variables were selected to include in the spatial analysis of this research.

The second research question was meant to find the environmental variables that are related to the spatial distribution of mosquitoes on Rusinga Island. To answer the second research question data analysis were performed, an attempt was made to find any correlation between the mosquito dataset and 14 environmental variables. The resulting correlation coefficients were quite low. First, it can be concluded from the analysis that using single linear regression or multiple linear regression is not suitable for the prediction of mosquito occurrences. Performing non-linear analysis are difficult here since there is no idea what type of non-linear relation to think of. However, there seems to be a kind of relationship between the maximum catch size and the slope of a point (Figure 8), where the maximum catch size is decreasing with increasing slope. A relation with the maximum catch size was not exactly what was looked for, rather for a relation with just the catch size, however it gives the indication that slope is somehow related to the occurrence of mosquitoes. That no strong correlation is found within this study does not mean that there is no relation between the occurrences of adult mosquitoes and these 14 environmental variables. There is no relation found between these environmental variables and the specific mosquito dataset from the SolarMal project.

The main conclusion in respect to the objective of this study is that there is no strong relationship between the available adult mosquito dataset and the created environmental variables which can be used in prediction models for mosquito occurrence. It was expected that there would be a relationship, however the found relationships were weak. A possible explanation for this could be that the adult mosquito dataset is acquired in and outside houses. It could be that human related factors (like the house construction type, animal ownership or house occupancy) influences the catch sizes of the mosquitoes that much, that any relation with the environment is weakened. It could also be that the positions where the mosquito data was collected is not well distributed over the island and was not representative for the whole island.

The third research question was focused on the suitability of the mosquito dataset to relate it with environmental variables on Rusinga Island. For that reason, there was a critical look to the mosquito dataset in chapter 6. From the spatial analysis (histograms and violin plots) it can be concluded that the sample points are not completely covering all the characteristics of the island. Average values are relatively more present in the mosquito dataset and extreme characteristics are mostly missing. From the timeline analysis it can be concluded that there is a seasonal variation to deal with. Not only the catch sizes are varying over the year, but also the proportion of vector mosquitoes is varying over the year.

From the study conducted to the accuracy of the house locations measured, several conclusions can be made. First, it can be concluded that the accuracy of the positions is on average sufficient to relate with the larger scale breeding sites, like natural dips in the elevation or a year-round stable water body. For the smaller scale breeding sites, like footprints of cattle in drenched grass and dumped waste in bushes, the accuracy is limiting. However, for these small scale breeding sites, also the accuracy of the environmental variables is too limited. With spatial resolutions of 30 meter (DEM) and 2,4 meter (satellite image), the environmental variables are too limited in spatial accuracy to detect breeding sites that can be smaller than 0,5 meter. Since the datasets for the environmental variables have only one timestamp, especially the satellite image, they are also too limited in the temporal accuracy to detect breeding sites that can occur and disappear again within a few weeks. The dataset of house positions is on average acceptable for the larger scale breeding sites, however some extreme errors in the dataset should not be there. Some house positions are completely wrong, representing a location more than 100 meter further away than it should be. Such errors should be selected out of the dataset, since these errors can be partly the cause of

the random noise in the spatial analysis. However, since the found correlations are very low, it is not expected that the removal of this noise only, immediately would lead to strong correlations. The weak correlations found are rather caused by a combination of factors.

8.2 Recommendations

For this thesis project an attempt was made to relate the occurrence of adult mosquitoes with the environment. There is probably a relationship to be found according to literature, however it was not found within this study. This has probably to do with the fact that the mosquito occurrence data was gathered only at houses, so where people are present. Any relationship between the adult mosquito distribution and the spatial distribution of human beings cannot be studied since there is no adult mosquito occurrence data available for places where no humans live. It would however still be possible to include non-spatial information about the presence of human beings in prediction models. Since the mosquito data is only collected at houses, field data of mosquito occurrences is missing where no houses are present. If adult mosquito data was available independently from the houses, it would be possible to study the relation between adult mosquito occurrences and the environment without the human influence. It would even be possible to study the human influence on this relation if both types of mosquito data were available. It is therefore recommended that a continuation of this study would also include mosquito data that is not located inside and in the direct environment of houses.

Focus in this thesis project was to predict the spatial distribution of adult mosquitoes, using mainly environmental variables that indicate areas where larvae of mosquitoes can develop. It is highly recommended to extent this study with a dataset of larvae breeding sites, since these are more direct related to the environmental variables.

By the time of this writing, a paper was recently published by McCann et al., who did a comparable study in Western Kenya, however to the occurrences of larvae instead of adult mosquitoes (McCann et al. 2014). McCann et al. show that including precipitation data in malaria vector studies improves the accuracy of prediction models. Precipitation data was in their study especially improving the accuracy in the temporal dimension. This thesis project done on Rusinga Island was, in respect to the environmental variables, limited to a dataset with only one timestamp. This could be one of the reasons why the resulting correlations are low. The same paper states that the use of a so-called Random Forest Model (Breiman 2001), was more accurate in explaining the occurrences of larvae than a logistic regression model. The implementation of such a Random Forest Model could be an improvement to this study. It is recommended that a continuation of this thesis project would try to include precipitation data in the analysis and also assess the use of a Random Forest Model in this specific study.

For a continuation of this thesis project, it is recommended to have a critical look again at the environmental variables. In this thesis project, the environmental variables were handled as continuous variables. It could be an option to categorize (some) of these environmental variables. The slope could for instance be divided into a flat, medium and steep slope. The same counts for the mosquito dataset, where the occurrence of mosquitoes could also be expressed in categories instead of catch sizes. An option is to divide into measurements were no mosquitoes at all were caught and into measurements where one or more mosquitoes were caught.

It is also recommended, if it would be possible, to improve the house positions dataset. A second house position dataset for example would provide the chance to detect the errors and correct house positions if necessary. Performing the spatial analysis with a corrected house position dataset would probably lead to less noise in the results.

It is recommended to search for more environmental variables in a follow up of this thesis, with a focus on indirect indications for the small scale breeding sites. The chance for footprints of cattle in drenched grass could for instance be indicated by a map representing the depth of the groundwater table. The closer the groundwater table to the ground surface, the less stable the ground will be and the more chance there is on footprints (and tire tracks).

In order to increase the temporal accuracy of the analysis, the use of more satellite images is recommended. Acquiring more high-resolution images like the QuickBird image in this study is probably too expensive for the SolarMal project. However, it could be an option to ensure the temporal accuracy by the use of more satellite images with a lower resolution and accept a coarser spatial resolution.

9 Bibliography

ASTER. 2011. ASTER GDEM 2 Readme File. ASTER JAPAN/USA.

Bøgh, C., S. W. Lindsay, S. E. Clarke, A. Dean, M. Jawara, M. Pinder, and C. J. Thomas. 2007. High spatial resolution mapping of malaria transmission risk in The Gambia, West Africa, using landsat TM satellite imagery. American Journal of Tropical Medicine and Hygiene **76**:875-881.

Breiman, L. 2001. Random forests. Machine learning **45**:5-32.

Bulsink. 2007. Rapid identification of mosquito larval habitats: A remote sensing and GIS based quick-scan. Thesis Report. Wageningen University and Research Centre, Wageningen.

Clements, A. C. A., H. L. Reid, G. C. Kelly, and S. I. Hay. 2013. Further shrinking the malaria map: How can geospatial science help to achieve malaria elimination? The Lancet Infectious Diseases **13**:709-718.

Coulibaly, D., S. Rebaudet, M. Travassos, Y. Tolo, M. Laurens, A. K. Kone, K. Traore, A. Guindo, I. Diarra, A. Niangaly, M. Daou, A. Dembele, M. Sissoko, B. Kouriba, N. Dessay, J. Gaudart, R. Piarroux, M. A. Thera, C. V. Plowe, and O. K. Doumbo. 2013. Spatio-temporal analysis of malaria within a transmission season in Bandiagara, Mali. Malar J **12**.

Dambach, P., V. Machault, J. P. Lacaux, C. Vignolles, A. Sié, and R. Sauerborn. 2012. Utilization of combined remote sensing techniques to detect environmental variables influencing malaria vector densities in rural West Africa. International Journal of Health Geographics **11**.

DigitalGlobe. 2005. QuickBird Imagery Products FAQ. Digital Globe.

DigitalGlobe. 2006. QuickBird Imagery Products - Product Guide. Digital Globe.

Dom, N. C., A. H. Ahmad, Z. A. Latif, R. Ismail, and B. Pradhan. 2013. Coupling of remote sensing data and environmental-related parameters for dengue transmission risk assessment in Subang Jaya, Malaysia. Geocarto International **28**:258-272.

Garmin. 2009. Waypoint Averaging.

Garmin. 2011. Garmin eTrex and GLONASS: A powerful combination.

Garmin. 2014. What is WAAS?

Gaudart, J., O. Touré, N. Dessay, A. L. Dicko, S. Ranque, L. Forest, J. Demongeot, and O. K. Doumbo. 2009. Modelling malaria incidence with environmental dependency in a locality of Sudanese savannah area, Mali. Malar J **8**.

Hintze, J. L., and R. D. Nelson. 1998. Violin Plots: A Box Plot-Density Trace Synergism. American Statistician **52**:181-184.

Hiscox, A., N. Maire, I. Kiche, M. Silkey, T. Homan, P. Oria, C. Mweresa, B. Otieno, M. Ayugi, T. Bousema, P. Sawa, J. Alaii, T. Smith, C. Leeuwis, W. R. Mukabana, and W. Takken. 2012. The SolarMal Project: innovative mosquito trapping technology for malaria control. Malar J **11**:O45.

Huete, A., K. Didan, T. Miura, E. P. Rodriguez, X. Gao, and L. G. Ferreira. 2002. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. Remote Sensing of Environment **83**:195-213.

Jenness, J. 2006. Topographic Position Index extension for ArcView 3.x, v. 1.3a. Jenness Enterprises.

Jenness, J., and L. Engelman. 2013. Jenness Enterprises. Flagstaff, Arizona, USA.

Kasasa, S., V. Asoala, L. Gosoniu, F. Anto, M. Adjuik, C. Tindana, T. Smith, S. Owusu-Agyei, and P. Vounatsou. 2013. Spatio-temporal malaria transmission patterns in Navrongo demographic surveillance site, northern Ghana. Malar J **12**.

Kelly, G. C., E. Hale, W. Donald, W. Batarii, H. Bugoro, J. Nausien, J. Smale, K. Palmer, A. Bobogare, G. Taleo, A. Vallely, M. Tanner, L. S. Vestergaard, and A. C. Clements. 2013. A high-resolution geospatial surveillance-response system for malaria elimination in Solomon Islands and Vanuatu. Malar J **12**.

Lillesand, T. M., R. W. Kiefer, and J. W. Chipman. 2004. Remote sensing and image interpretation. John Wiley & Sons Ltd.

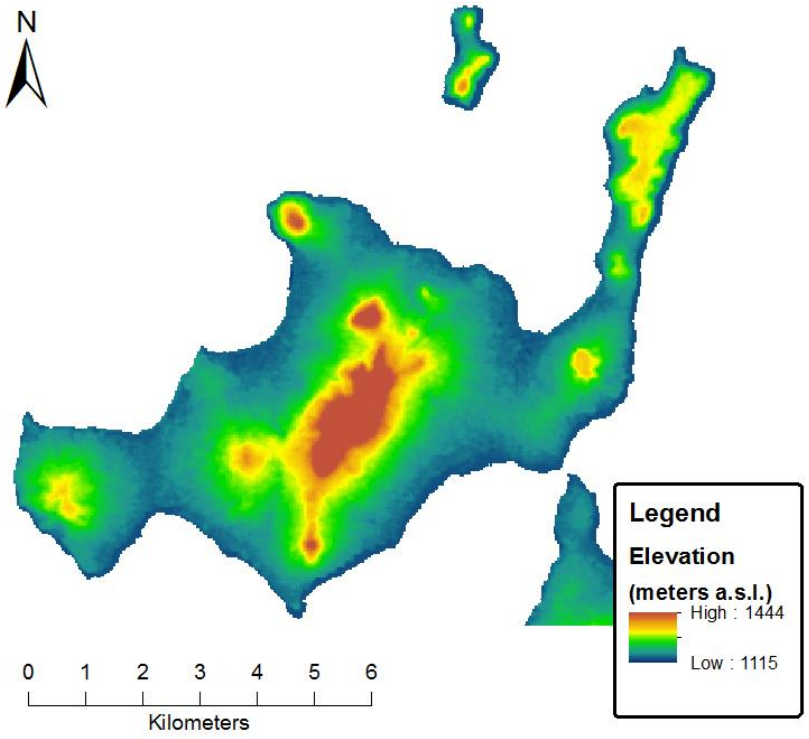
Mbogo, C. M., J. M. Mwangangi, J. G. Nzovu, W. Gu, G. Yan, J. T. Gunter, C. Swalm, J. Keating, J. L. Regens, J. I. Shililu, J. I. Githure, and J. C. Beier. 2003. Spatial and temporal heterogeneity of Anopheles mosquitoes and Plasmodium falciparum transmission along the Kenyan coast. American Journal of Tropical Medicine and Hygiene **68**:734-742.

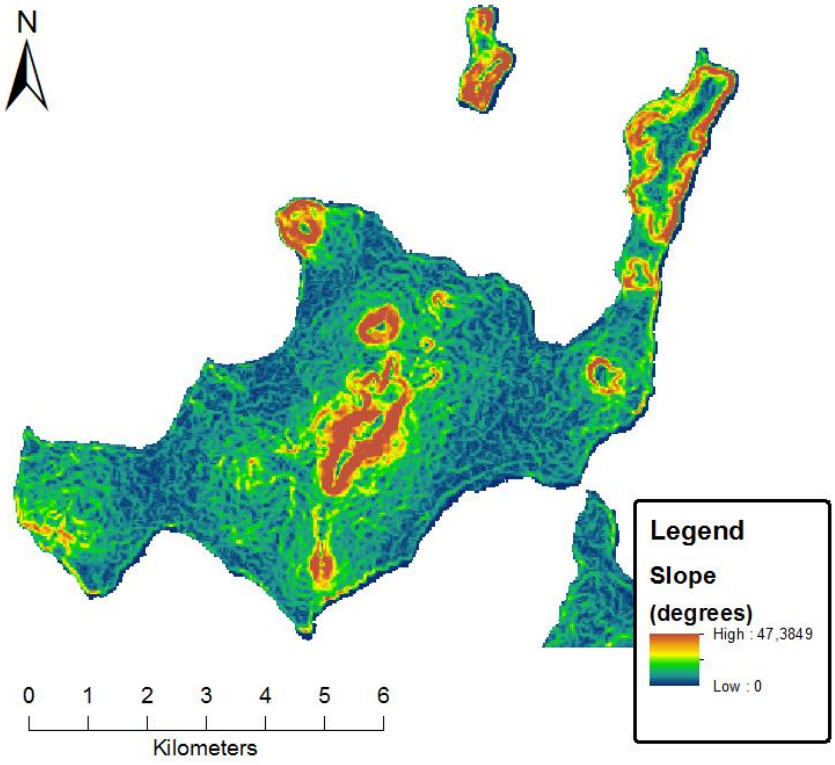
McCann, R. S., J. P. Messina, D. W. MacFarlane, M. N. Bayoh, J. M. Vulule, J. E. Gimnig, and E. D. Walker. 2014. Modeling larval malaria vector habitat locations using landscape features and cumulative precipitation measures. Int J Health Geogr **13**:17.

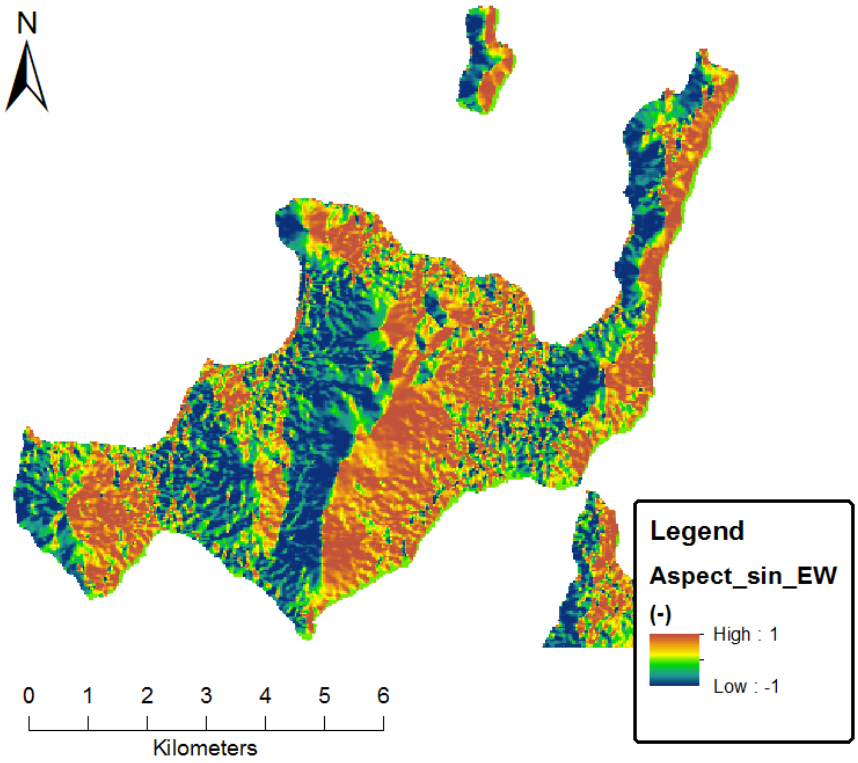
McFeeters, S. K. 1996. The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. International Journal of Remote Sensing **17**:1425-1432.
McFeeters, S. K. 2013. Using the normalized difference water index (ndwi) within a geographic information system to detect swimming pools for mosquito abatement: A practical approach. Remote Sensing **5**:3544-3561.
Melmane, P., S. Shetty, and D. Gulati. 2014. A study of drug resistance in malaria. Journal, Indian Academy of Clinical Medicine **15**.
Miller, A. 2002. Subset selection in regression. CRC Press.
Moss, W. J., H. Hamapumbu, T. Kobayashi, T. Shields, A. Kamanga, J. Clennon, S. Mharakurwa, P. E. Thuma, and G. Glass. 2011. Use of remote sensing to identify spatial risk factors for malaria in a region of declining transmission: A cross-sectional and longitudinal community survey. Malar J **10**.
Mukabana, W. R., C. K. Mweresa, B. Otieno, P. Omusula, R. C. Smallegange, J. J. A. van Loon, and W. Takken. 2012. A Novel Synthetic Odorant Blend for Trapping of Malaria and Other African Mosquito Species. J Chem Ecol **38**:235-244.
Mulder, A. 2013. Thesis Proposal: Malaria, Mosquitoes and Agricultural Land Use Patterns. Wageningen University.
Mutuku, F. M., J. A. Alaii, M. N. Bayoh, J. E. Gimnig, J. M. Vulule, E. D. Walker, E. Kabiru, and W. A. Hawley. 2006. Distribution, description, and local knowledge of larval habitats of Anopheles gambiae s.l. in a village in western Kenya. American Journal of Tropical Medicine and Hygiene **74**:44-53.
Myers, W. P., A. P. Myers, J. Cox-Singh, H. C. Lau, B. Mokuai, and R. Malley. 2009. Micro-geographic risk factors for malarial infection. Malar J **8**.
Nmor, J. C., T. Sunahara, K. Goto, K. Futami, G. Sonye, P. Akweywa, G. Dida, and N. Minakawa. 2013. Topographic models for predicting malaria vector breeding habitats: Potential tools for vector control managers. Parasites and Vectors **6**.
Obsomer, V., M. Dufrene, P. Defourny, and M. Coosemans. 2013. Anopheles species associations in Southeast Asia: Indicator species and environmental influences. Parasites and Vectors **6**.
Olynik, M., M. Petovello, M. Cannon, and G. Lachapelle. 2002. Temporal Variability of GPS Error Sources and Their Effect on Relative Positioning Accuracy. Proceedings of the Institute of Navigation NTM 2002.
Shieh, G. 2010. Estimation of the simple correlation coefficient. Behav Res Methods **42**:906-917.
Venables, W., and B. D. Ripley. 2000. S programming. Springer.
Wee, L. K., S. N. Weng, N. Raduan, S. K. Wah, W. H. Ming, C. H. Shi, F. Rambli, C. J. Ahok, S. Marlina, N. W. Ahmad, A. McKemy, S. S. Vasan, and L. H. Lim. 2013. Relationship between rainfall and Aedes larval population at two insular sites in Pulau Ketam, Selangor, Malaysia. Southeast Asian Journal of Tropical Medicine and Public Health **44**:157.
Weiss, A. 2001. Topographic Position and Landform Analysis. Poster presentation. ESRI User Conference, San Diege, CA.
Wernsdorfer, W. H. 1994. Epidemiology of drug resistance in malaria. Acta Tropica **56**:143-156.
WHO. 2012a. Country Profile Kenya 2012. Page 142 World Malaria Report 2012. World Health Organization, Geneva.
WHO. 2012b. World Malaria Report 2012. World Health Organization, Geneva.
WHO. 2013. World Health Organisation.
Wing, M. G., A. Eklund, and L. D. Kellogg. 2005. Consumer-grade global positioning system (GPS) accuracy and reliability. **103**:169-173.

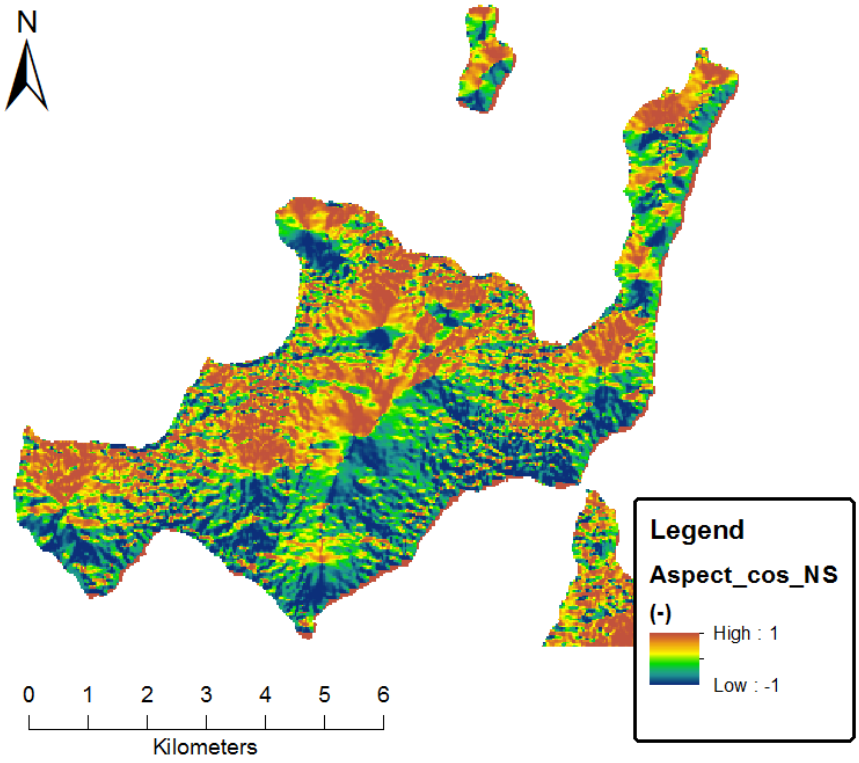
Appendix A

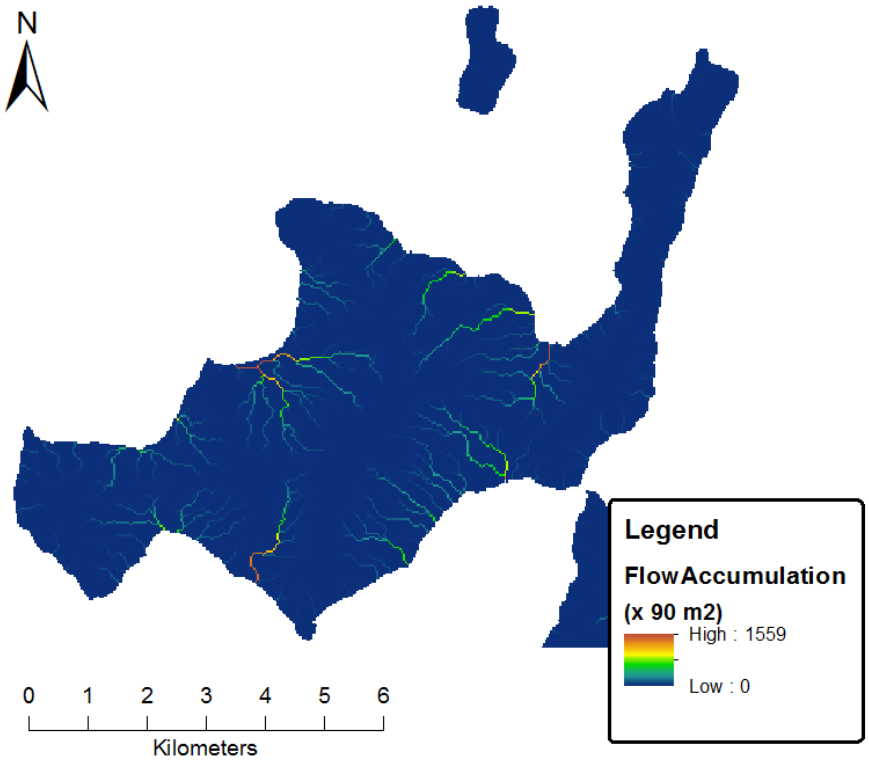
Environmental variables created for Rusinga Island

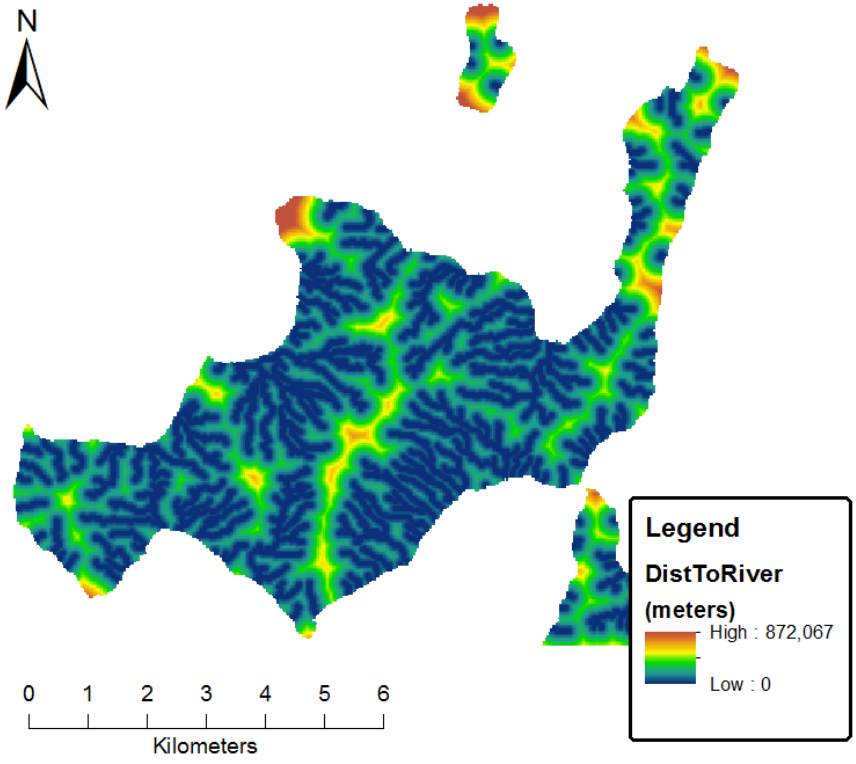


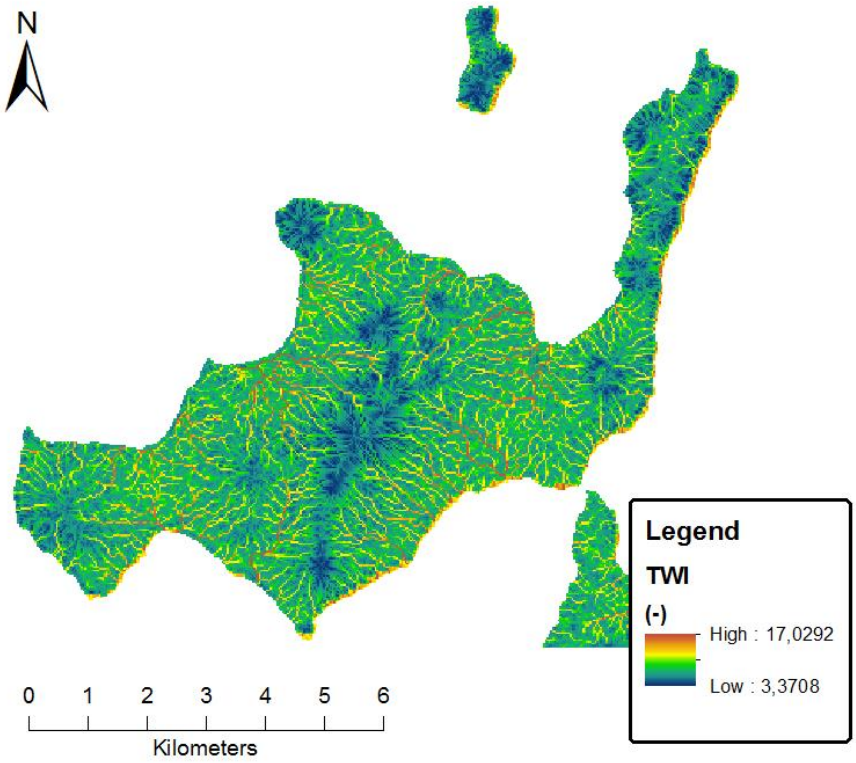


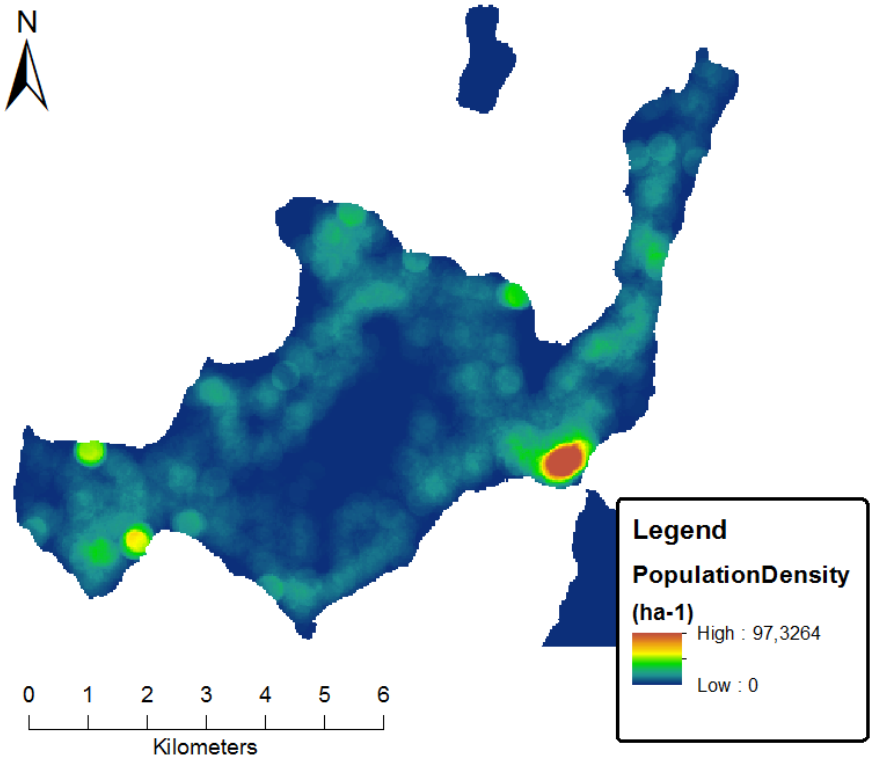

N
0 1 2 3 4 5 6
Kilometers
Legend
Aspect_sin_EW
(-)
High : 1
Low : -1

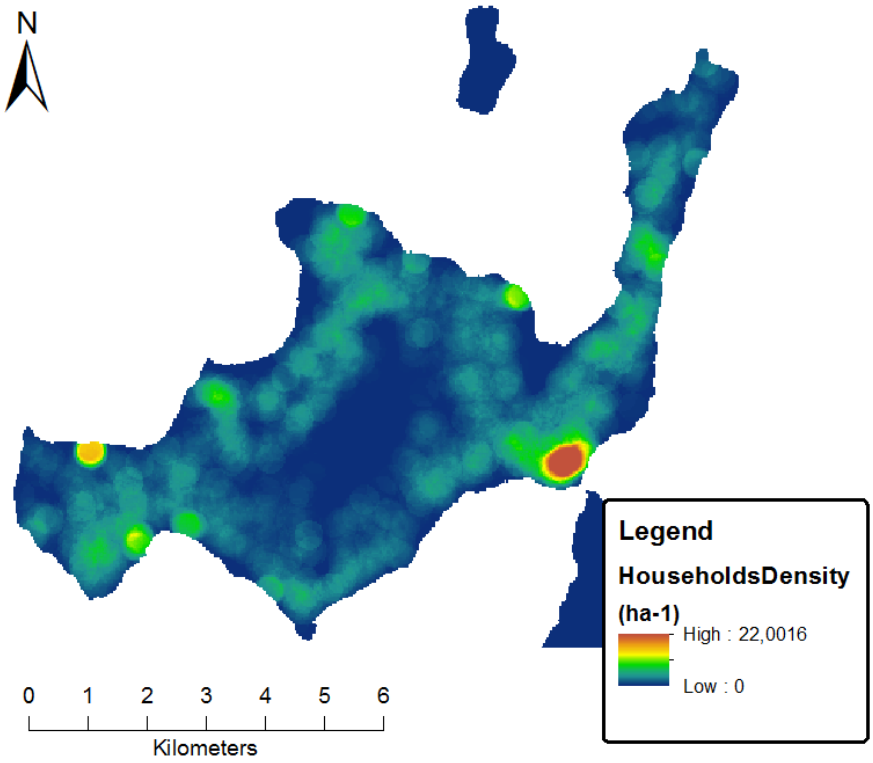

N
0 1 2 3 4 5 6
Kilometers
Legend
Aspect_cos_NS
(-)
High : 1
Low : -1

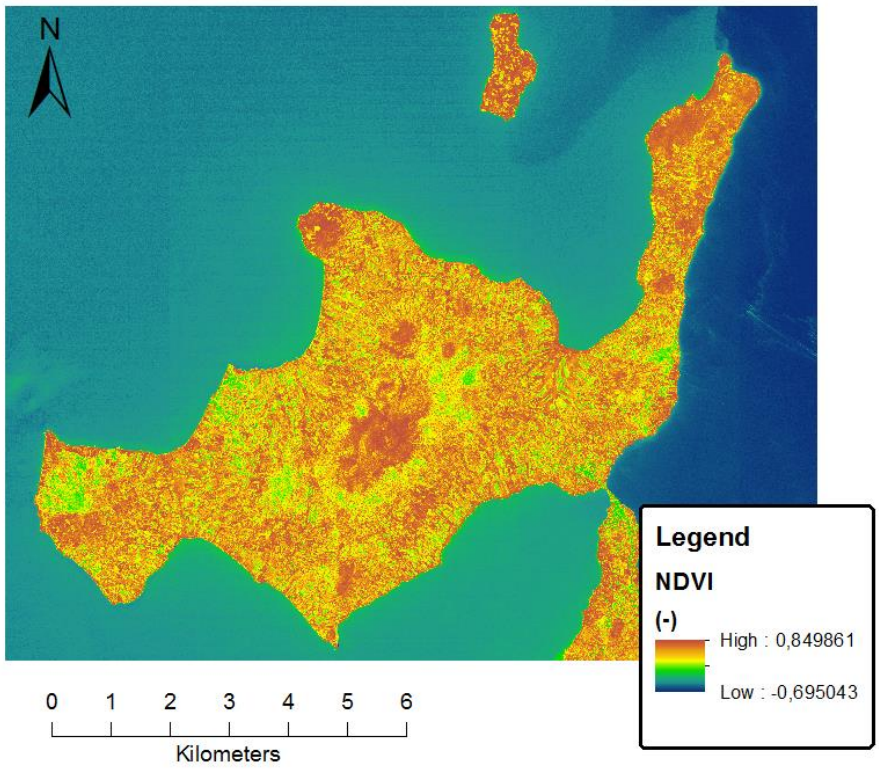

N
0 1 2 3 4 5 6
Kilometers
Legend
FlowAccumulation
(x 90 m2)
High : 1559
Low : 0

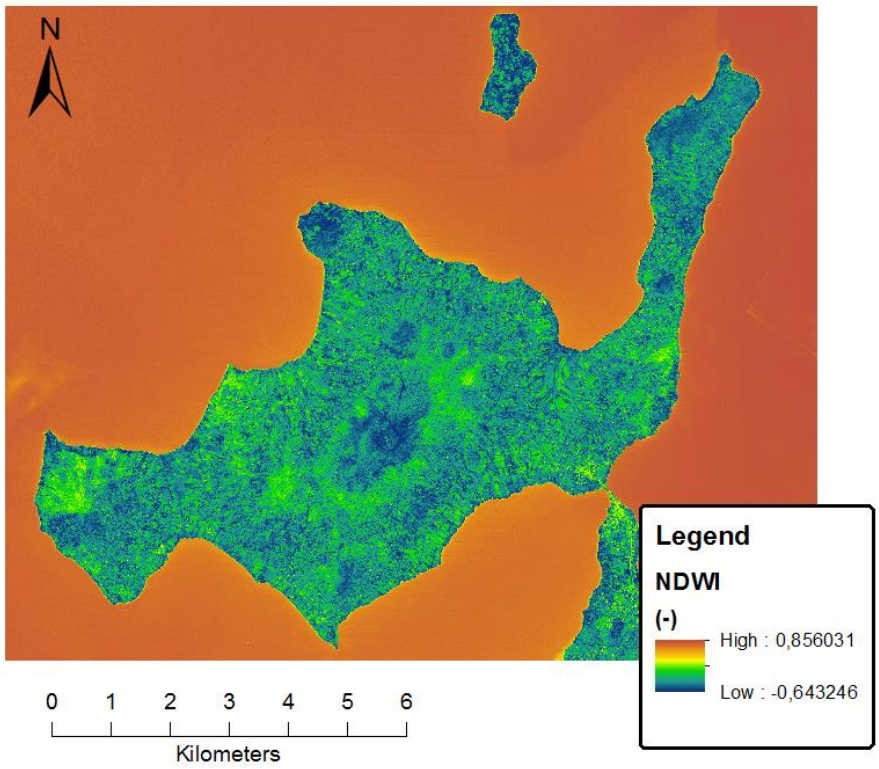

N
0 1 2 3 4 5 6
Kilometers
Legend
DistToRiver
(meters)
High : 872,067
Low : 0

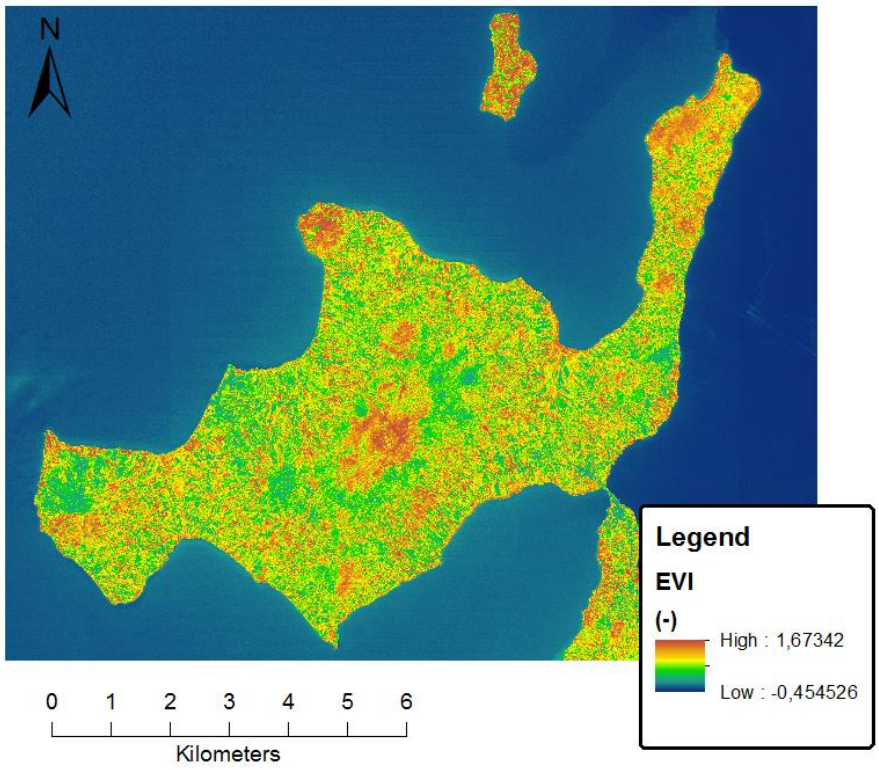

N
0 1 2 3 4 5 6
Kilometers
Legend
TWI
(-)
High : 17,0292
Low : 3,3708

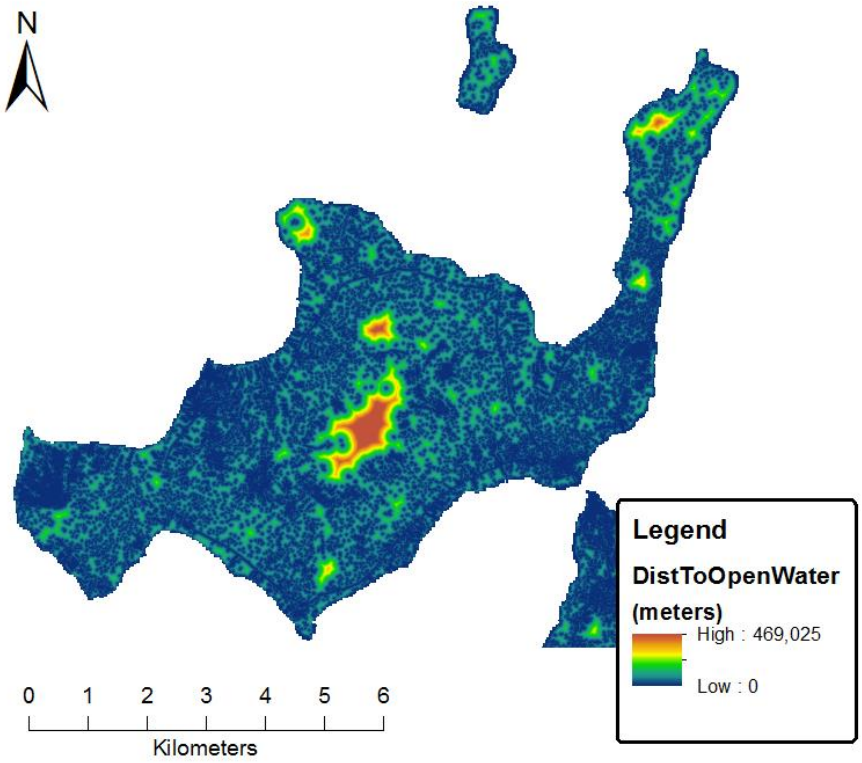

N
0 1 2 3 4 5 6
Kilometers
Legend
PopulationDensity
(ha-1)
High : 97,3264
Low : 0

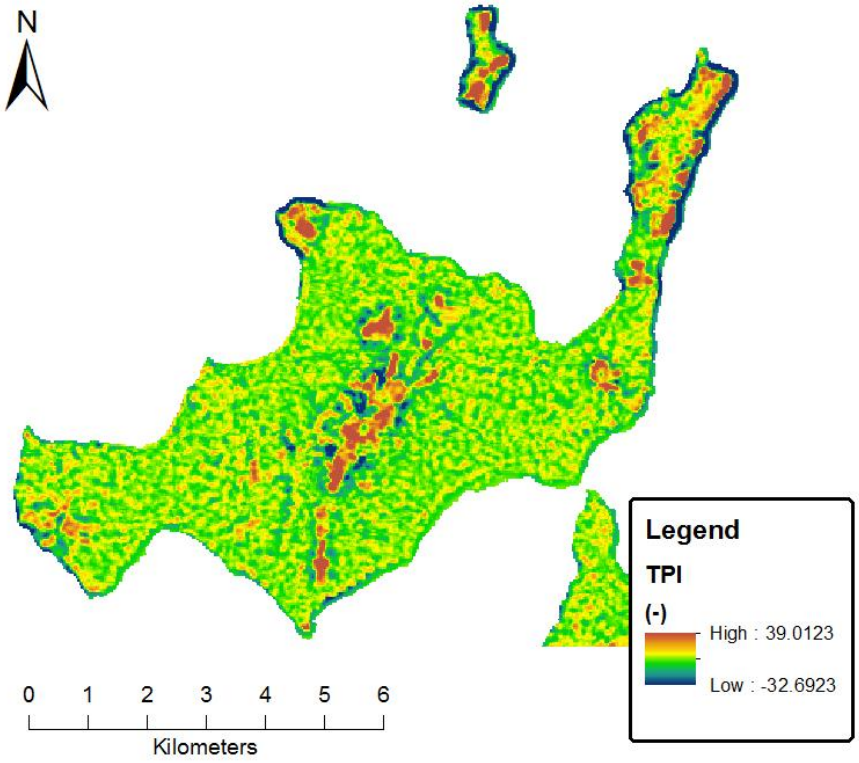

N
0 1 2 3 4 5 6
Kilometers
Legend
HouseholdsDensity
(ha-1)
High : 22,0016
Low : 0


N
0 1 2 3 4 5 6
Kilometers
Legend
NDVI
(-)
High : 0,849861
Low : -0,695043


N
Legend
NDWI
(-)
High : 0,856031
Low : -0,643246
0 1 2 3 4 5 6
Kilometers

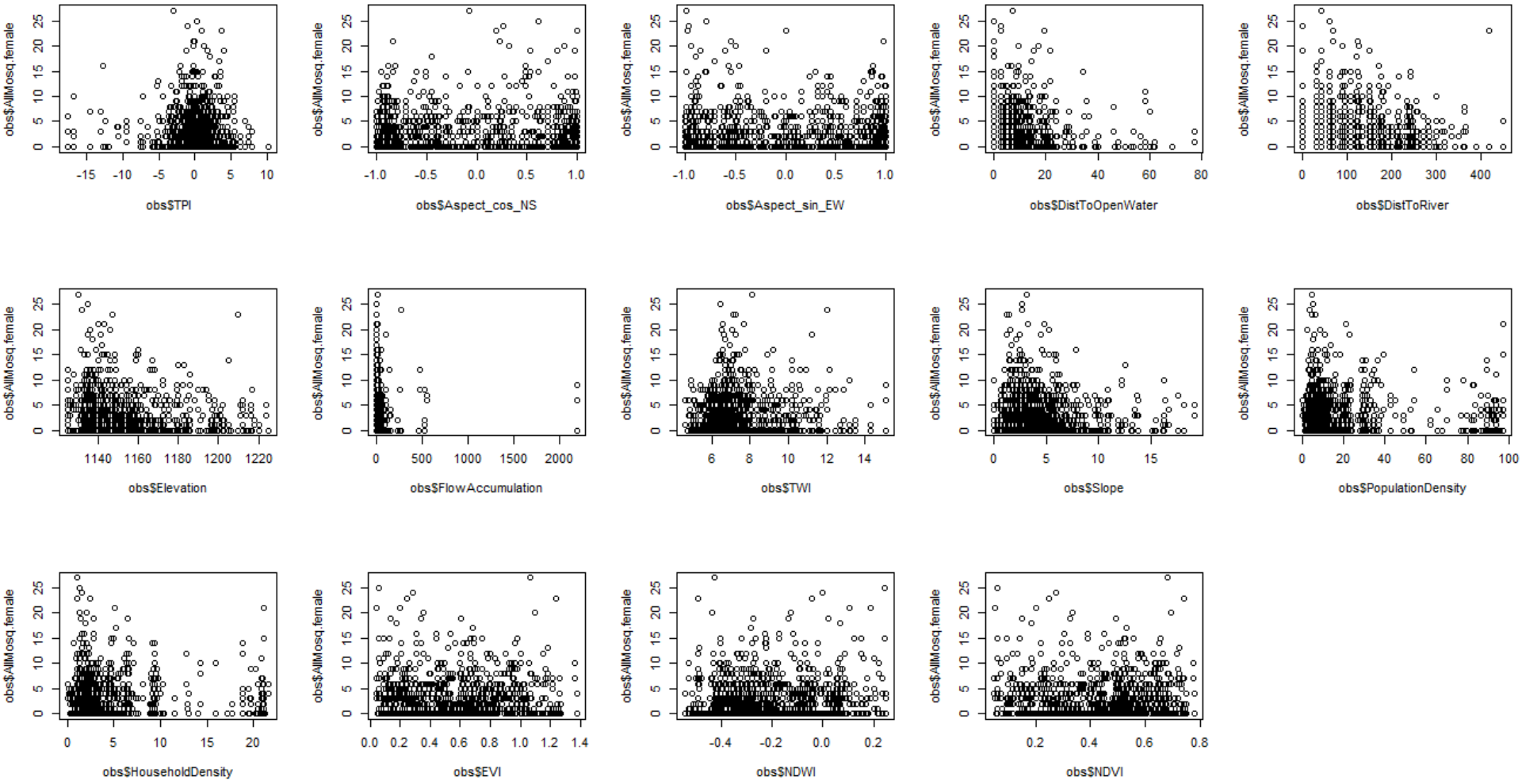

N
Legend
EVI
(-)
High : 1,67342
Low : -0,454526
0 1 2 3 4 5 6
Kilometers


N
0 1 2 3 4 5 6
Kilometers
Legend
DistToOpenWater
(meters)
High : 469,025
Low : 0


N
0 1 2 3 4 5 6
Kilometers
Legend
TPI
(-)
High : 39.0123
Low : -32.6923

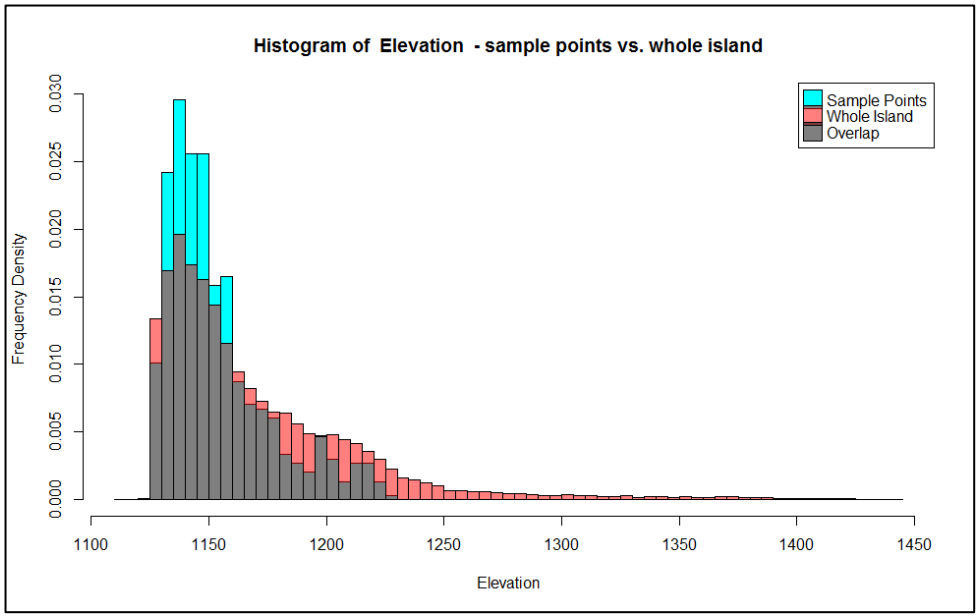
Appendix B

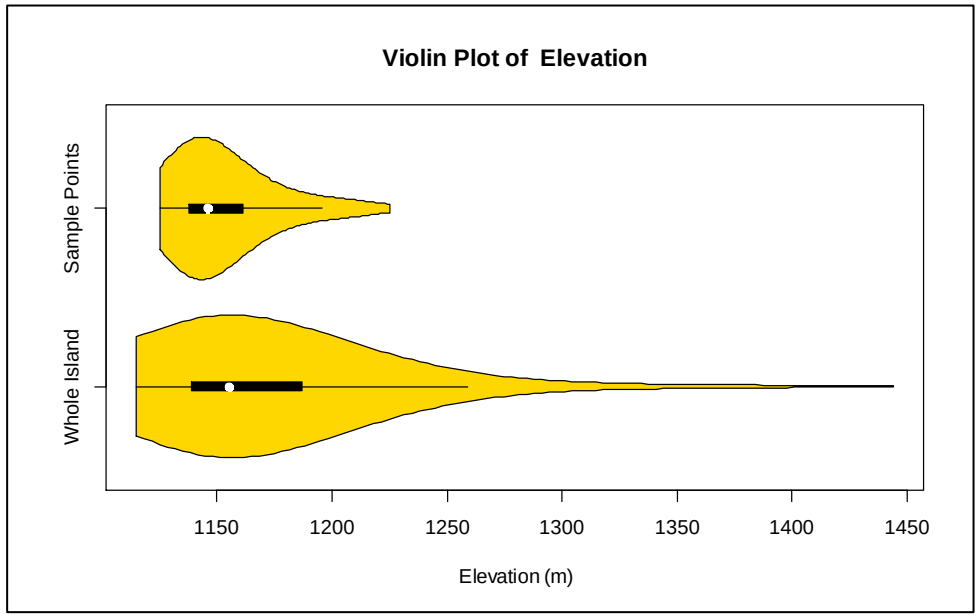
Scatterplots of all independent variables against dependent variable AllMosq (=All mosquitoes grouped together) (only female selected)

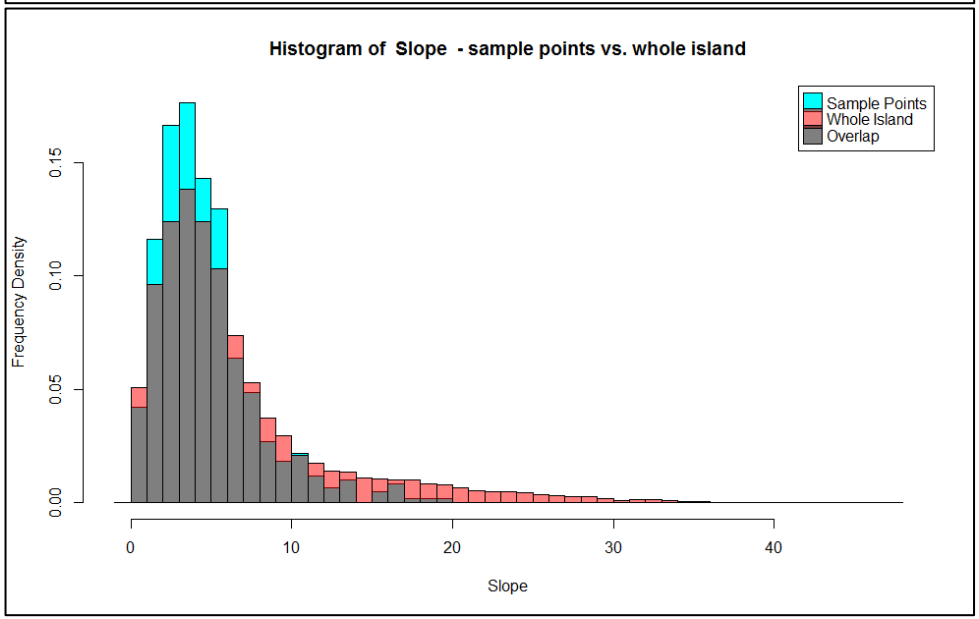


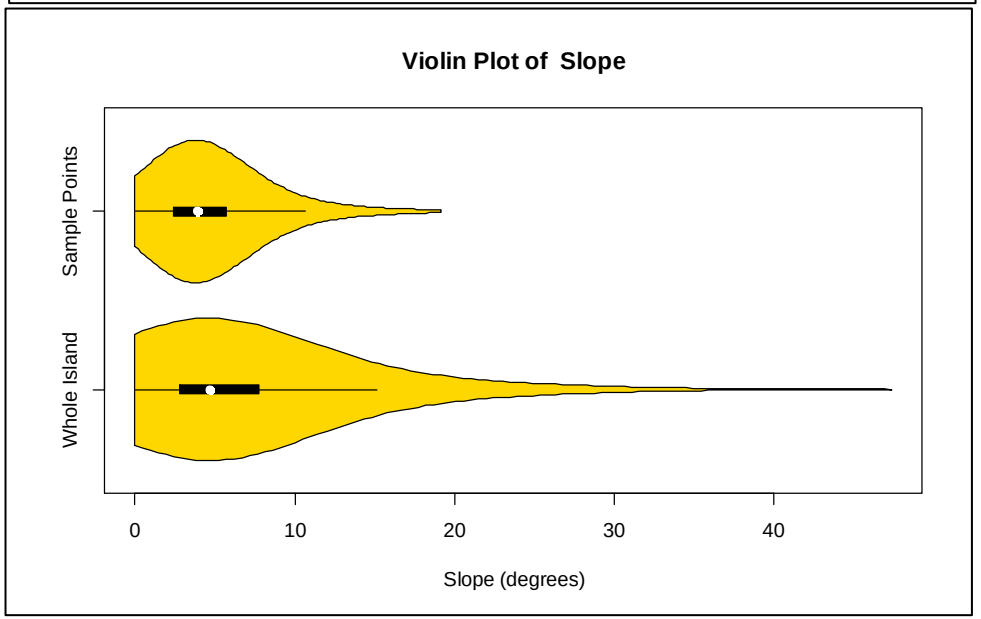
Appendix C

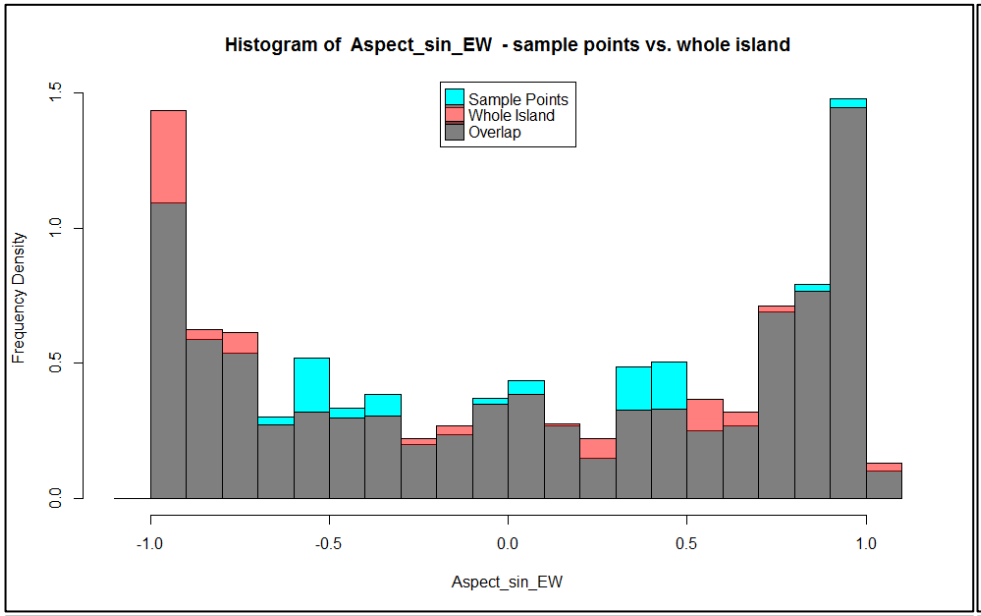
Histograms and Violin plots of the mosquito sample dataset and the whole island, for all environmental variables

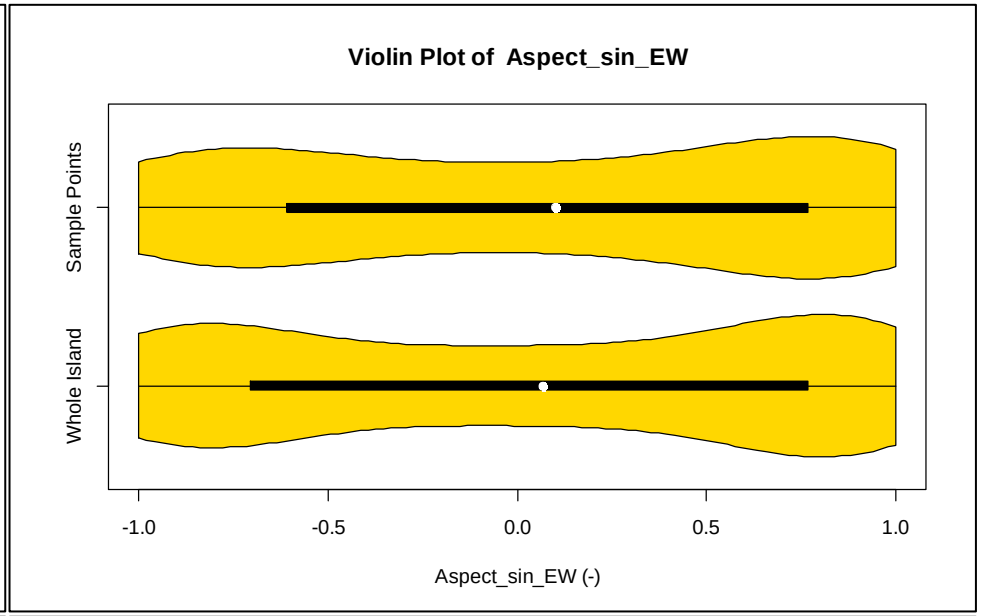


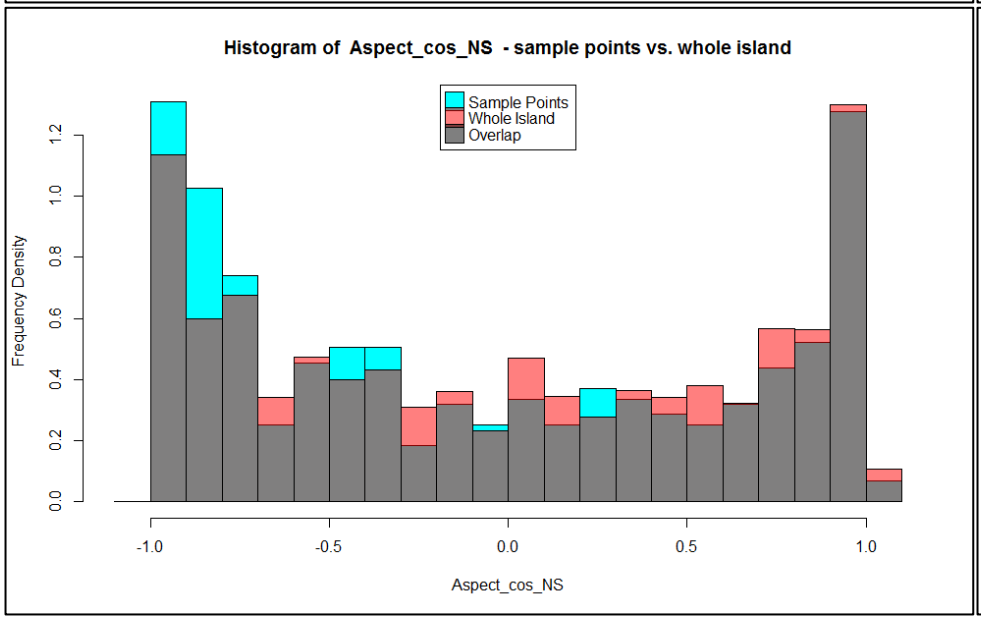


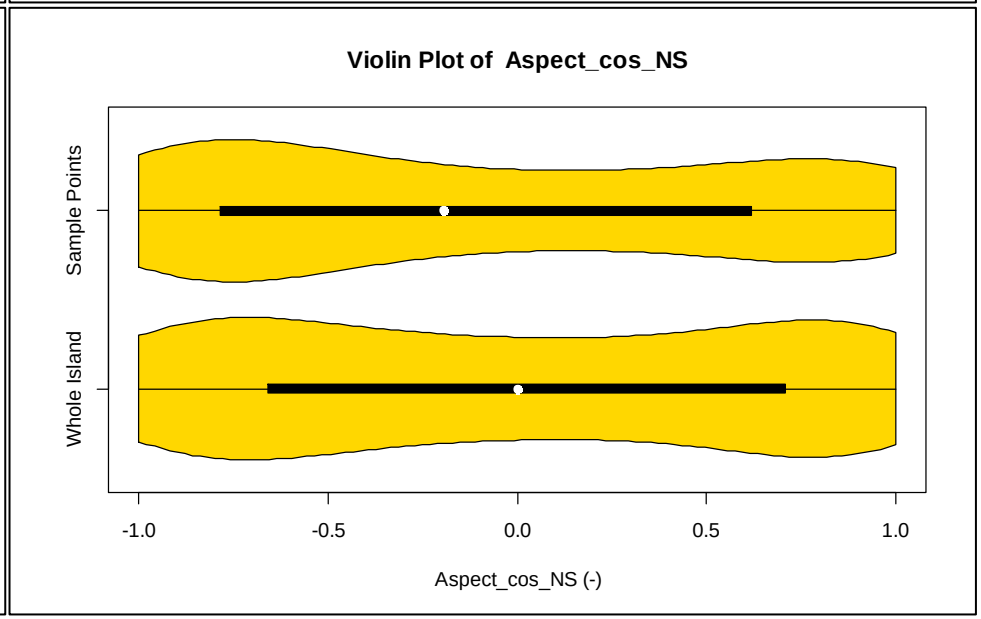


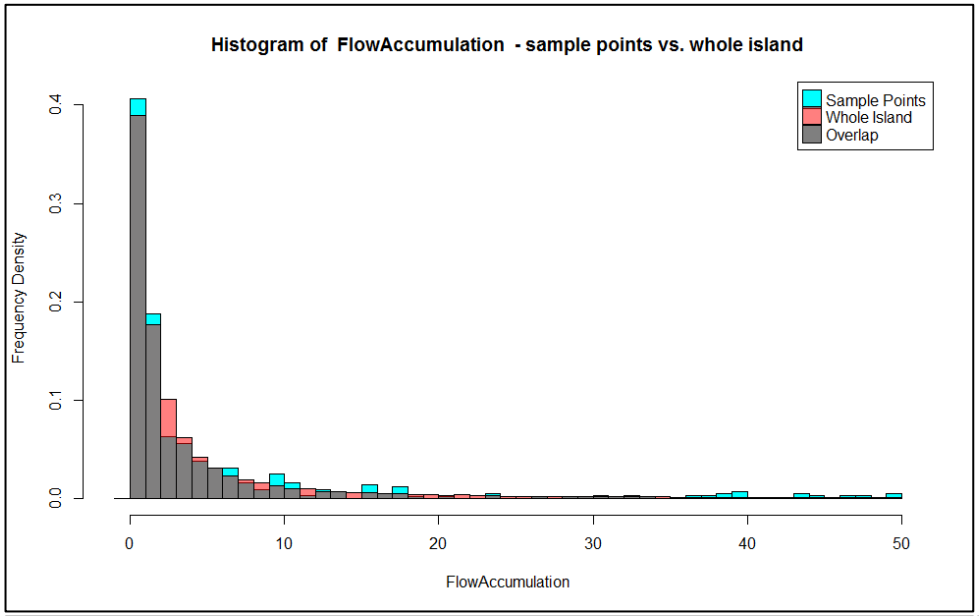


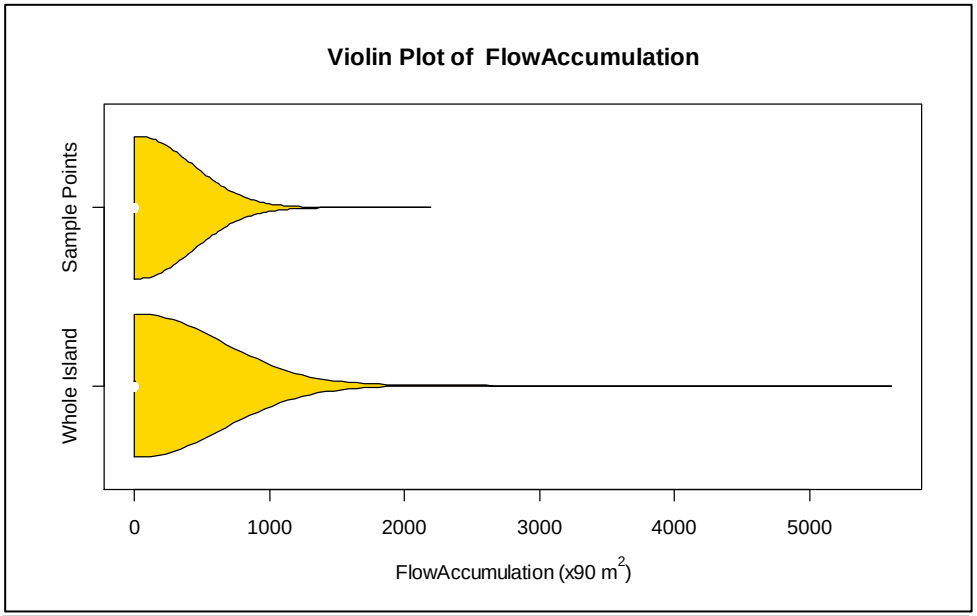

Histogram of Aspect_sin_EW - sample points vs. whole island
Sample Points
Whole Island
Overlap
Frequency Density
0.0
0.5
1.0
1.5
-1.0
-0.5
0.0
0.5
1.0
Aspect_sin_EW

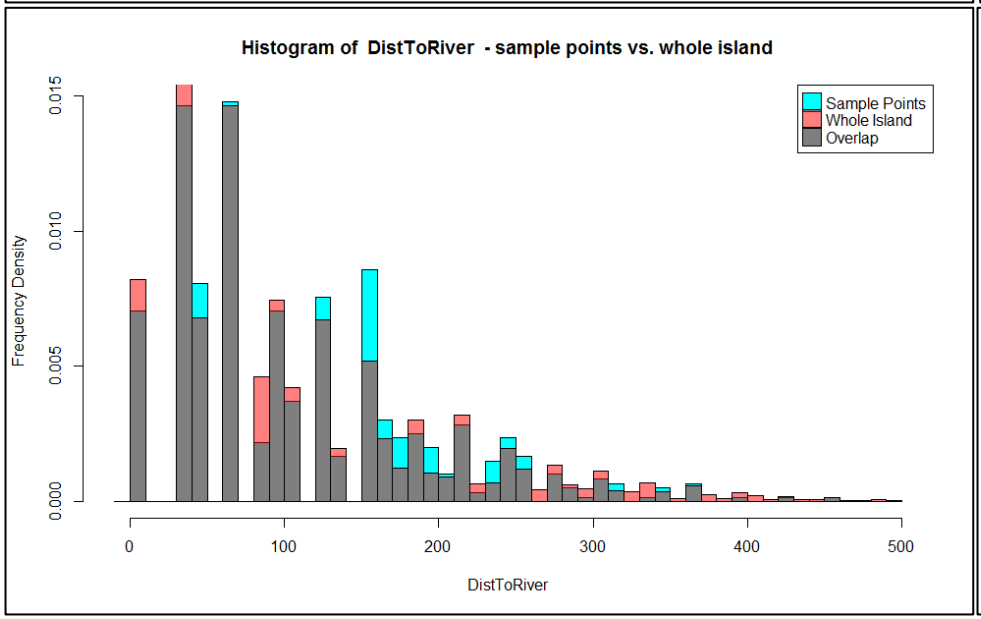

Violin Plot of Aspect_sin_EW
Sample Points
Whole Island
-1.0
-0.5
0.0
0.5
1.0
Aspect_sin_EW (-)

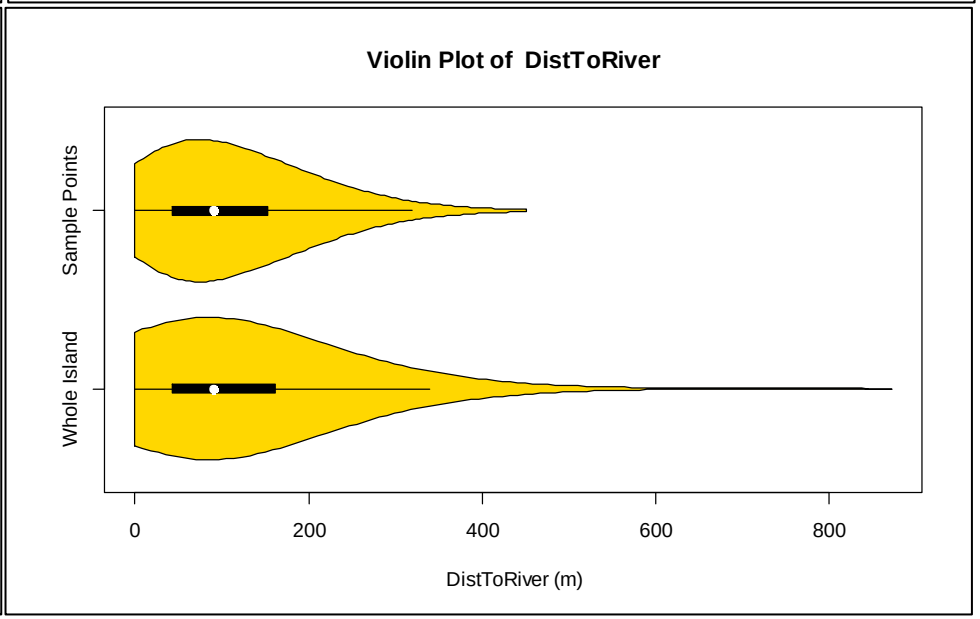

Histogram of Aspect_cos_NS - sample points vs. whole island
Sample Points
Whole Island
Overlap
Frequency Density
0.0
0.2
0.4
0.6
0.8
1.0
1.2
-1.0
-0.5
0.0
0.5
1.0
Aspect_cos_NS

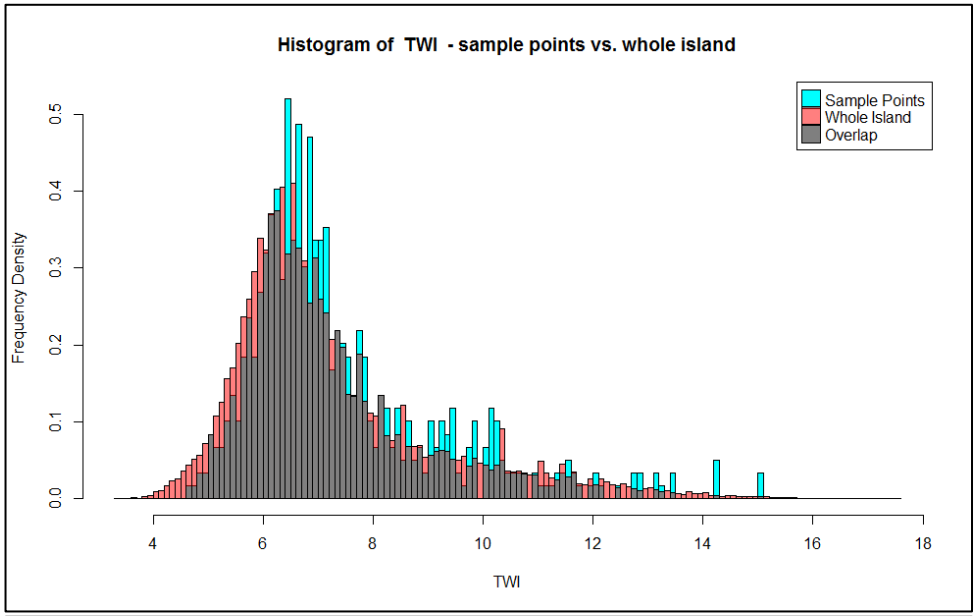

Violin Plot of Aspect_cos_NS
Sample Points
Whole Island
-1.0
-0.5
0.0
0.5
1.0
Aspect_cos_NS (-)

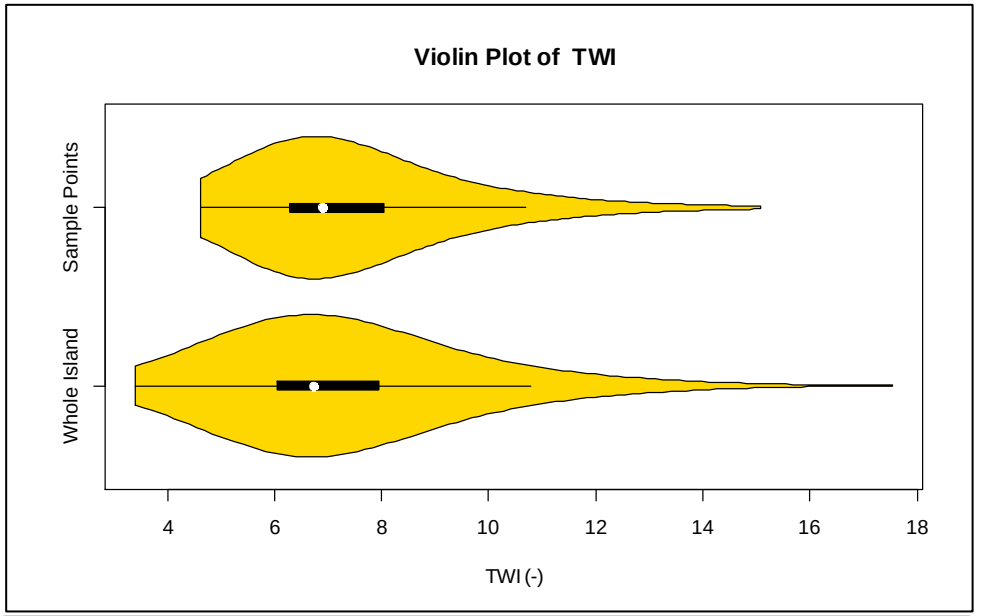

Histogram of FlowAccumulation - sample points vs. whole island
Sample Points
Whole Island
Overlap
Frequency Density
0.0
0.1
0.2
0.3
0.4
0
10
20
30
40
50
FlowAccumulation

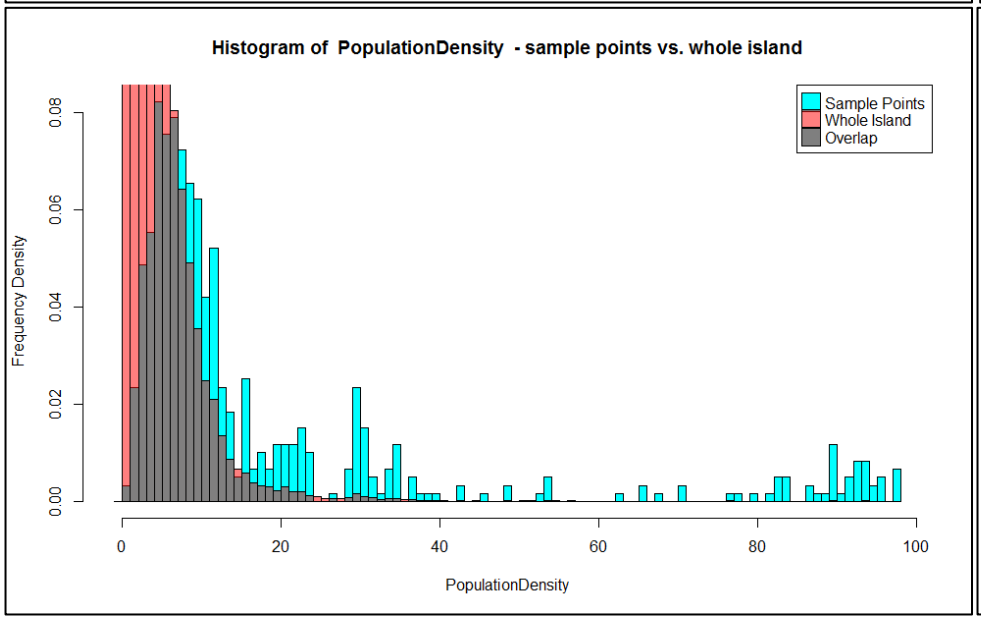

Violin Plot of FlowAccumulation
Sample Points
Whole Island
0
1000
2000
3000
4000
5000
FlowAccumulation (x90 m2)

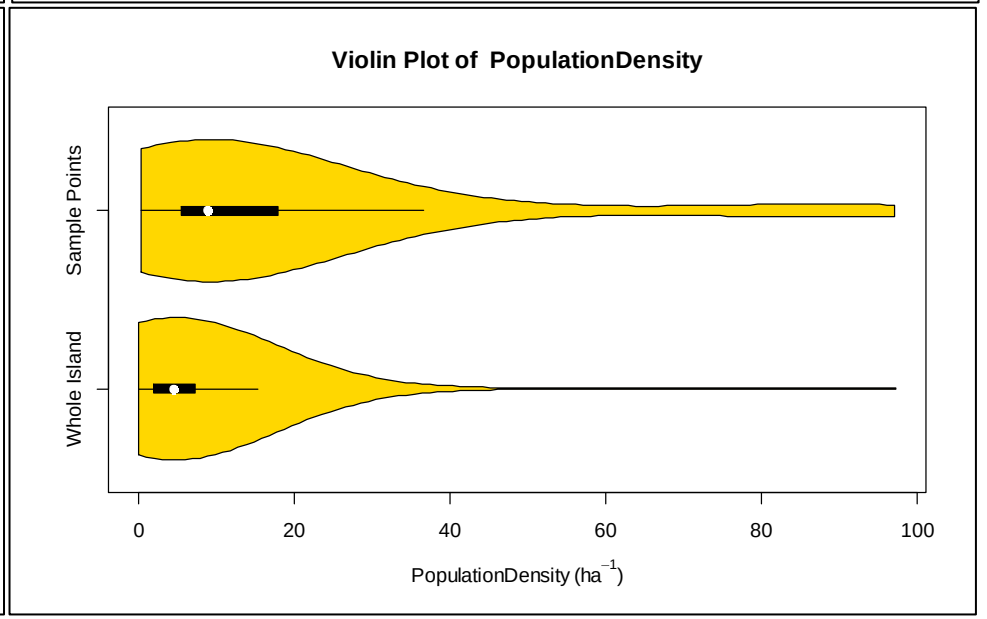

Histogram of DistToRiver - sample points vs. whole island
Sample Points
Whole Island
Overlap
Frequency Density
0.000
0.005
0.010
0.015
0
100
200
300
400
500
DistToRiver

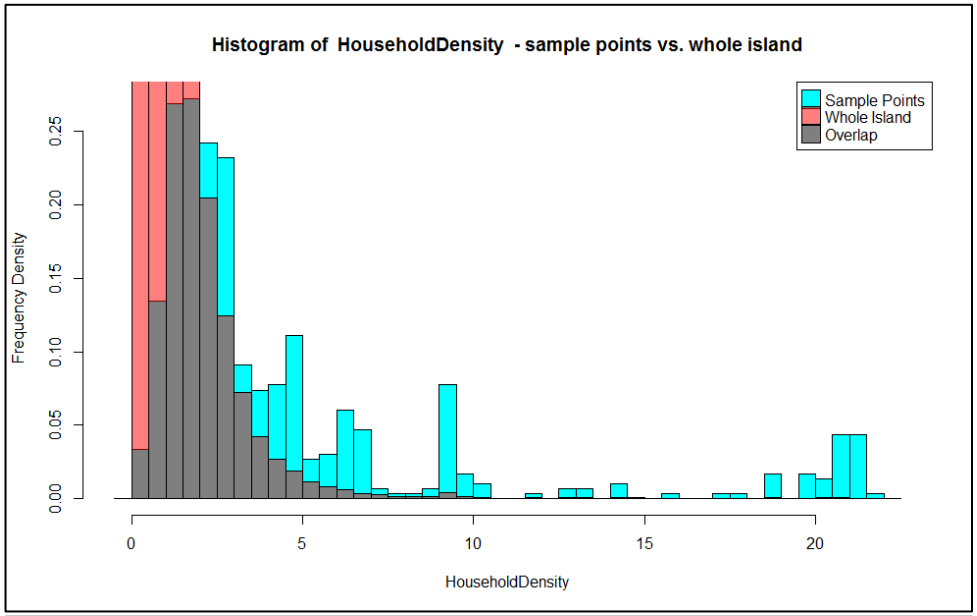

Violin Plot of DistToRiver
Sample Points
Whole Island
0
200
400
600
800
DistToRiver (m)

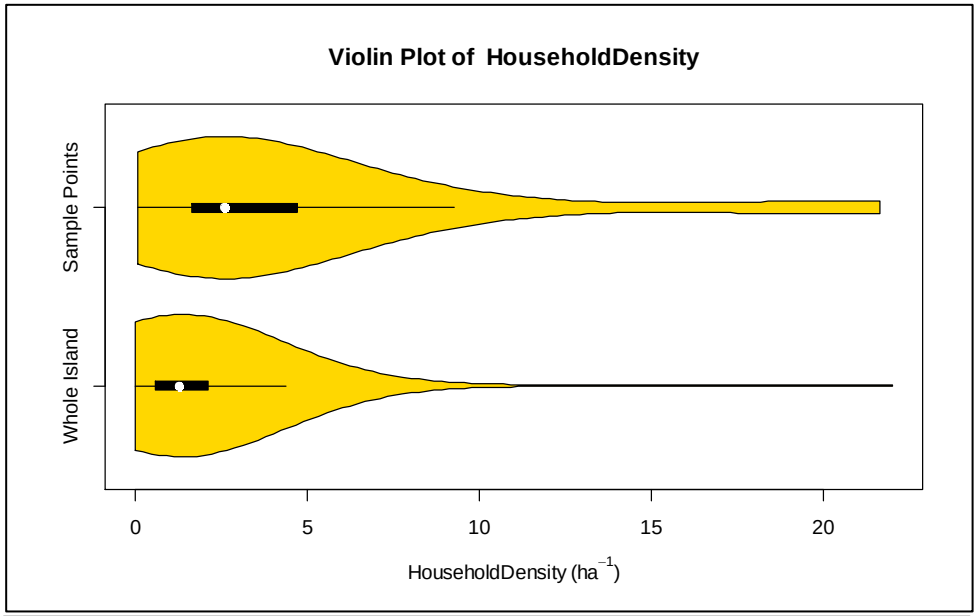

Histogram of TWI - sample points vs. whole island
Sample Points
Whole Island
Overlap
Frequency Density
0.0
0.1
0.2
0.3
0.4
0.5
4
6
8
10
12
14
16
18
TWI

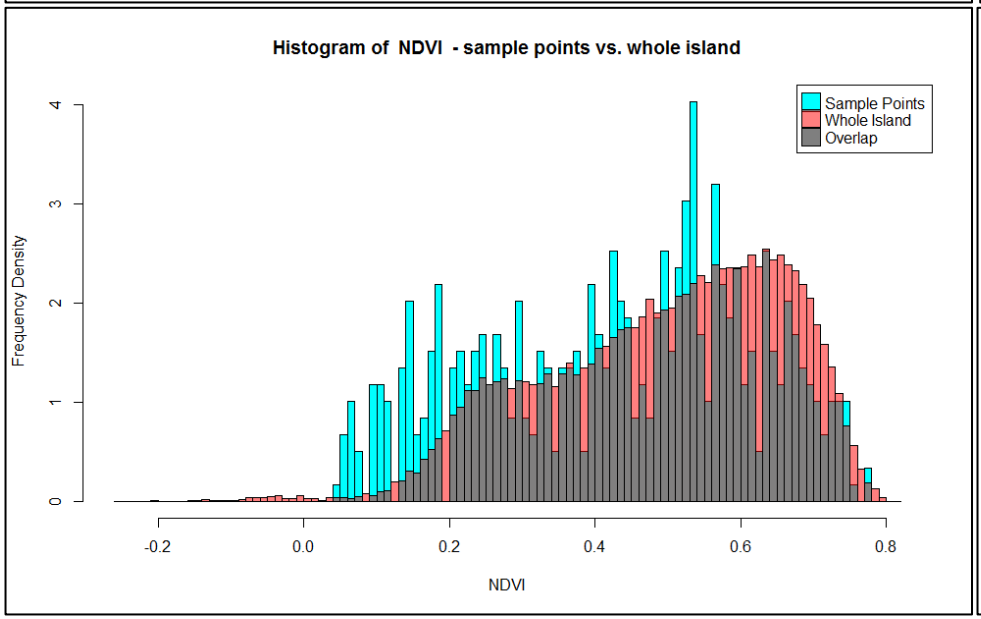

Violin Plot of TWI
Sample Points
Whole Island
4
6
8
10
12
14
16
18
TWI (-)

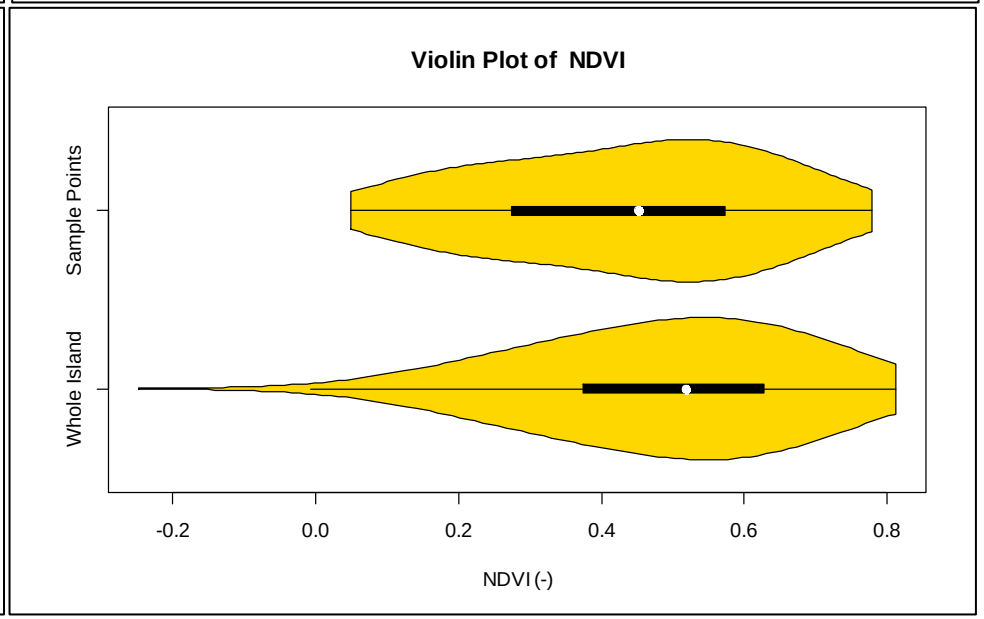

Histogram of PopulationDensity - sample points vs. whole island
Sample Points
Whole Island
Overlap
Frequency Density
0.00
0.02
0.04
0.06
0.08
0
20
40
60
80
100
PopulationDensity

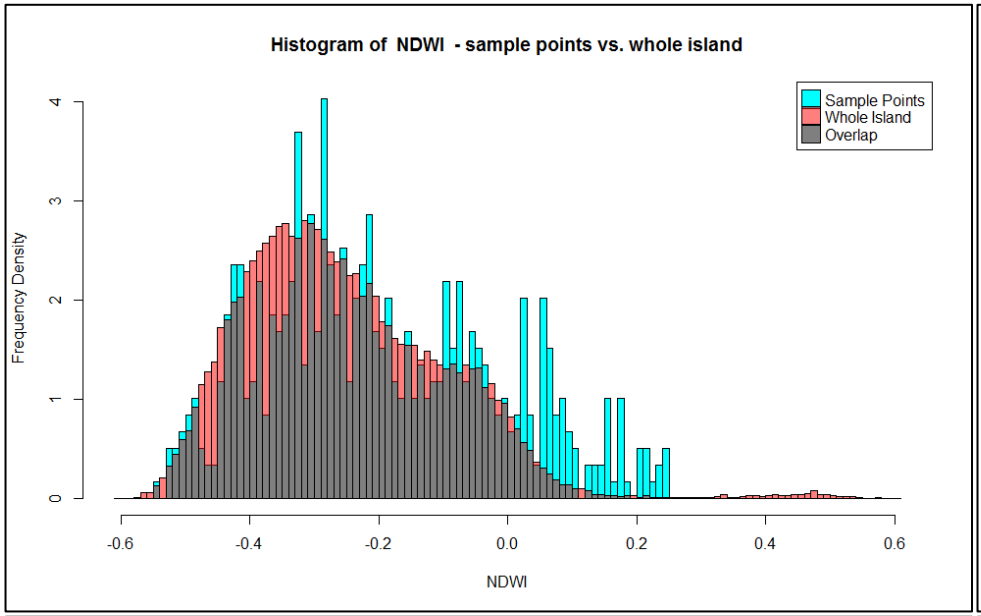

Violin Plot of PopulationDensity
Sample Points
Whole Island
0
20
40
60
80
100
PopulationDensity (ha^{-1})

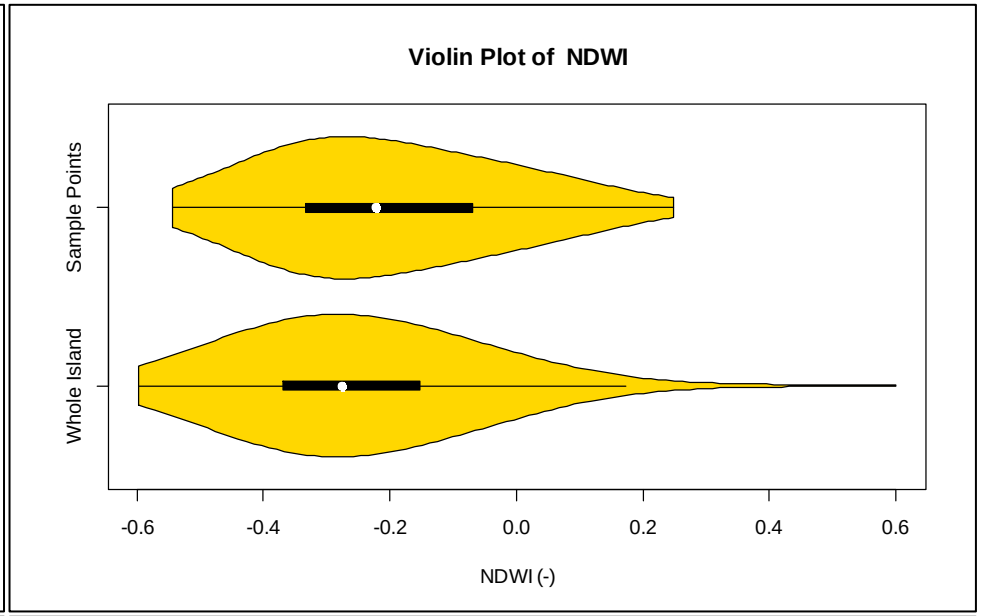

Histogram of HouseholdDensity - sample points vs. whole island
Sample Points
Whole Island
Overlap
Frequency Density
0.00
0.05
0.10
0.15
0.20
0.25
0
5
10
15
20
HouseholdDensity

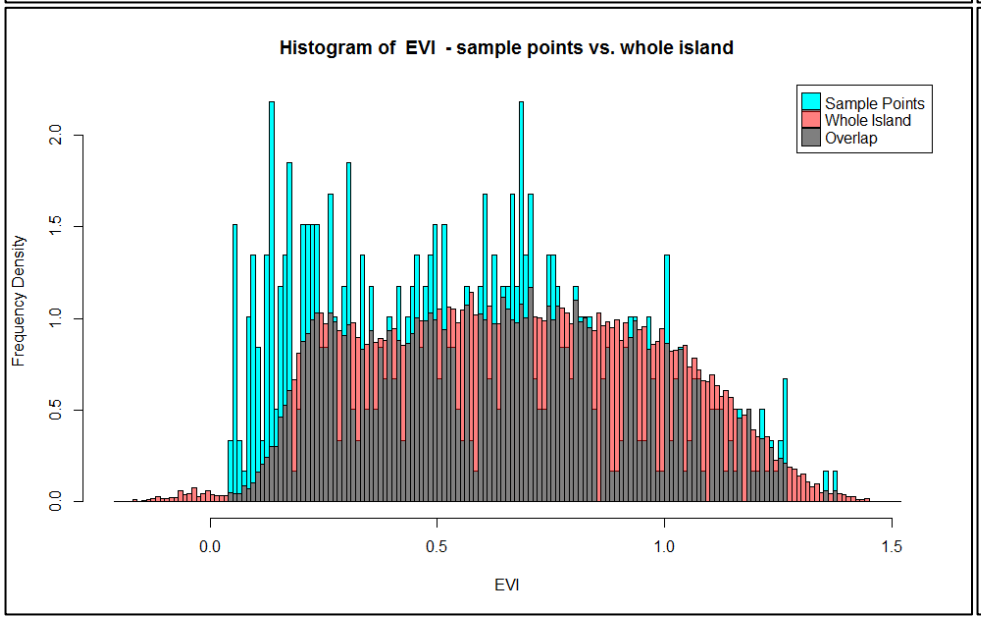

Violin Plot of HouseholdDensity
Sample Points
Whole Island
0
5
10
15
20
HouseholdDensity (ha^{-1})

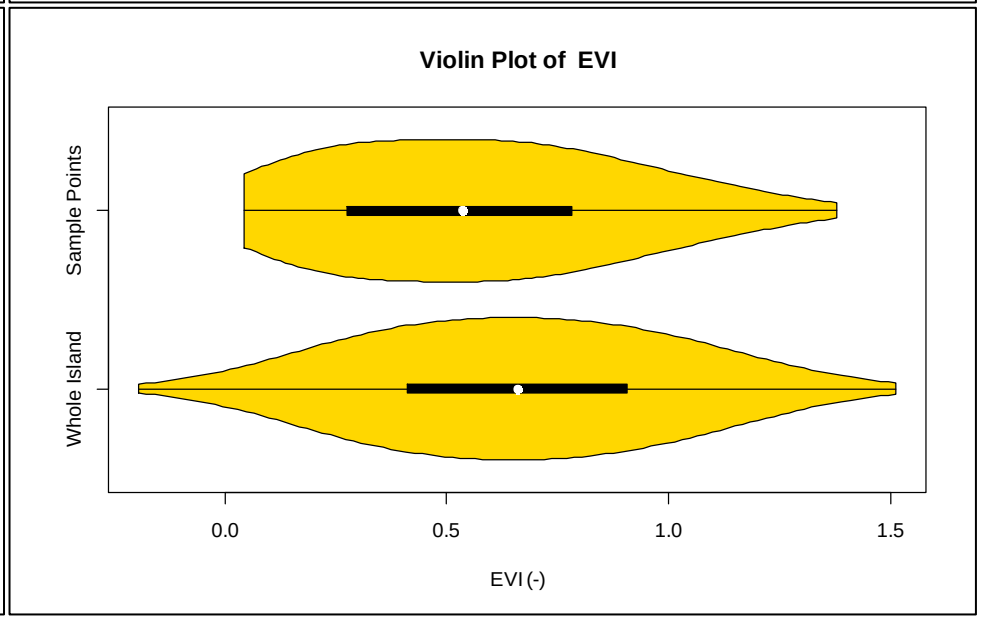

Histogram of NDVI - sample points vs. whole island
Sample Points
Whole Island
Overlap
Frequency Density
0
1
2
3
4
-0.2
0.0
0.2
0.4
0.6
0.8
NDVI

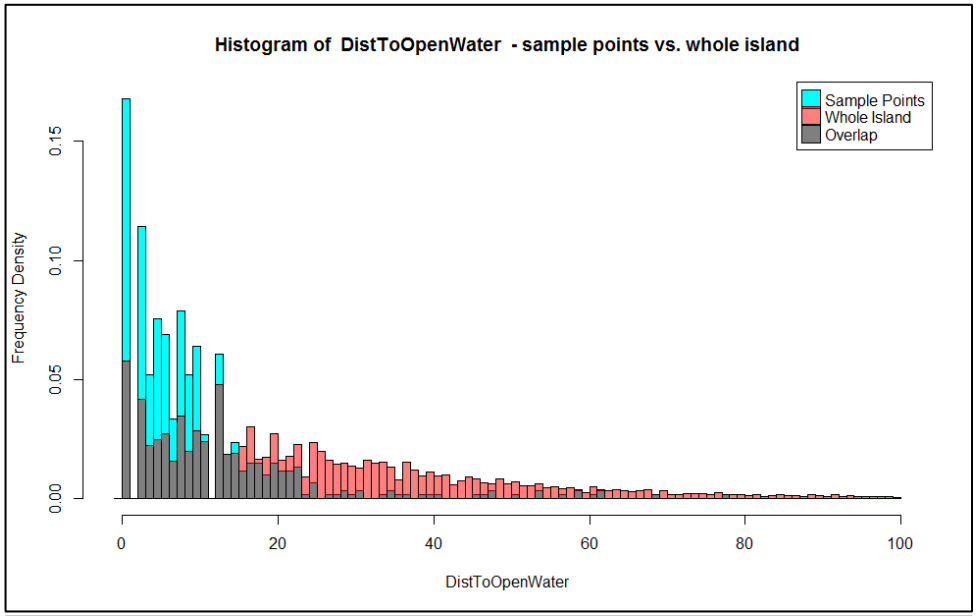

Violin Plot of NDVI
Sample Points
Whole Island
-0.2
0.0
0.2
0.4
0.6
0.8
NDVI (-)

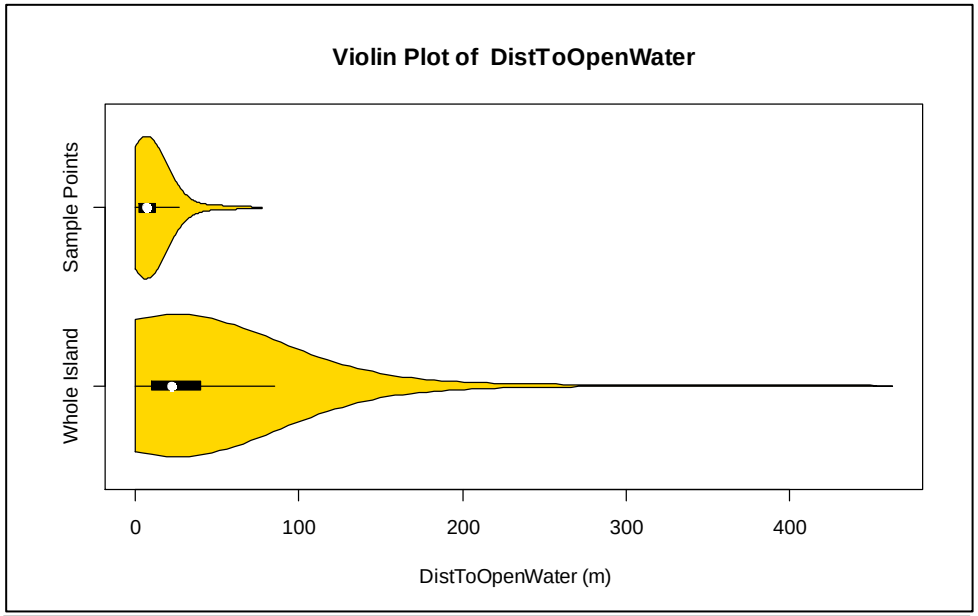

Histogram of NDWI - sample points vs. whole island
Sample Points
Whole Island
Overlap
Frequency Density
0
1
2
3
4
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
NDWI

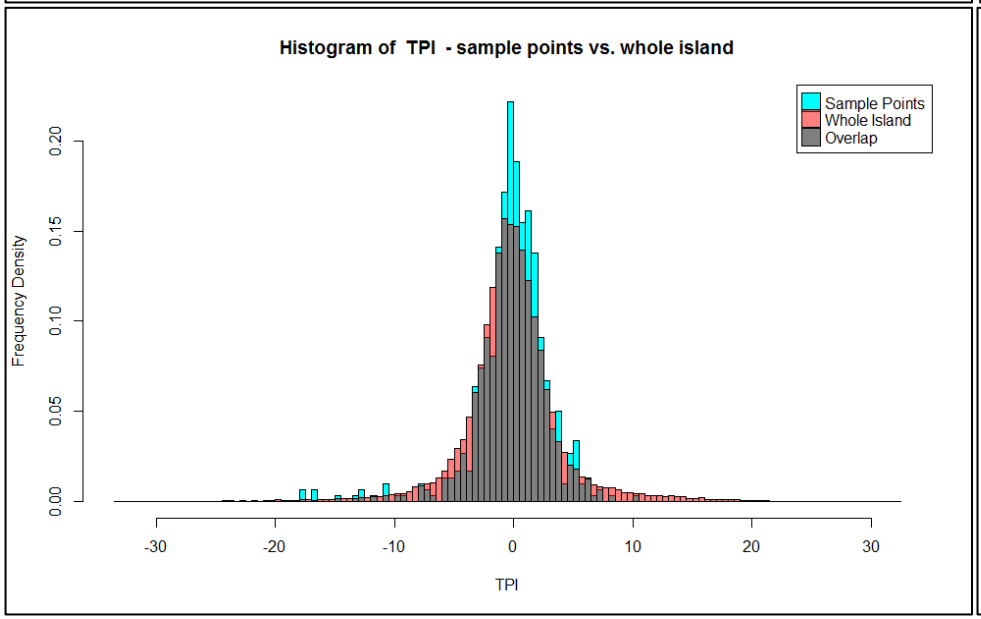

Violin Plot of NDWI
Sample Points
Whole Island
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
NDWI (-)

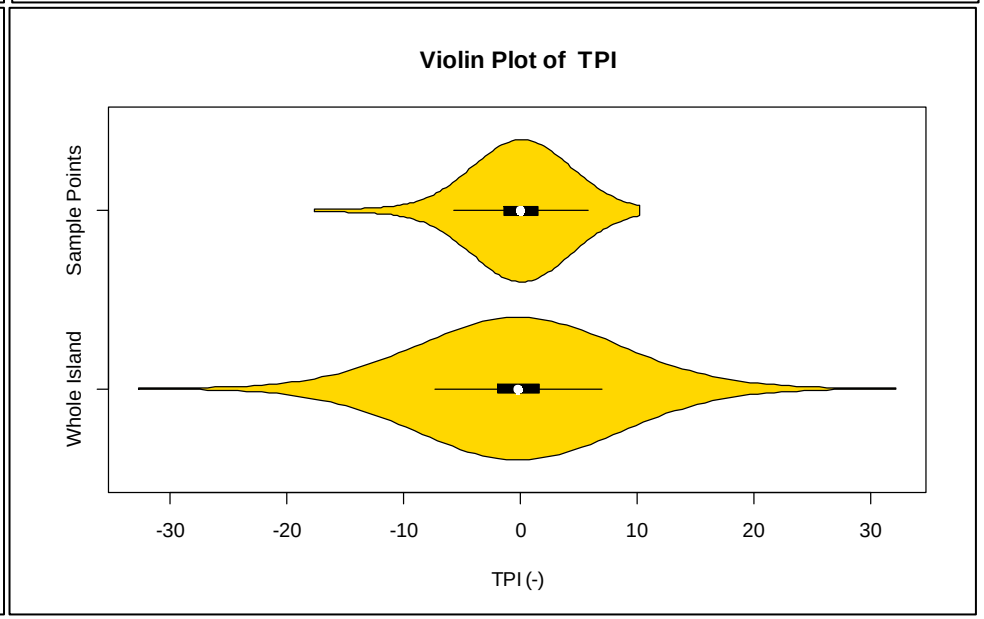

Histogram of EVI - sample points vs. whole island
Sample Points
Whole Island
Overlap
Frequency Density
0.0
0.5
1.0
1.5
2.0
0.0
0.5
1.0
1.5
EVI


Violin Plot of EVI
Sample Points
Whole Island
0.0
0.5
1.0
1.5
EVI (-)


Histogram of DistToOpenWater - sample points vs. whole island
Sample Points
Whole Island
Overlap
Frequency Density
0.00
0.05
0.10
0.15
0
20
40
60
80
100
DistToOpenWater

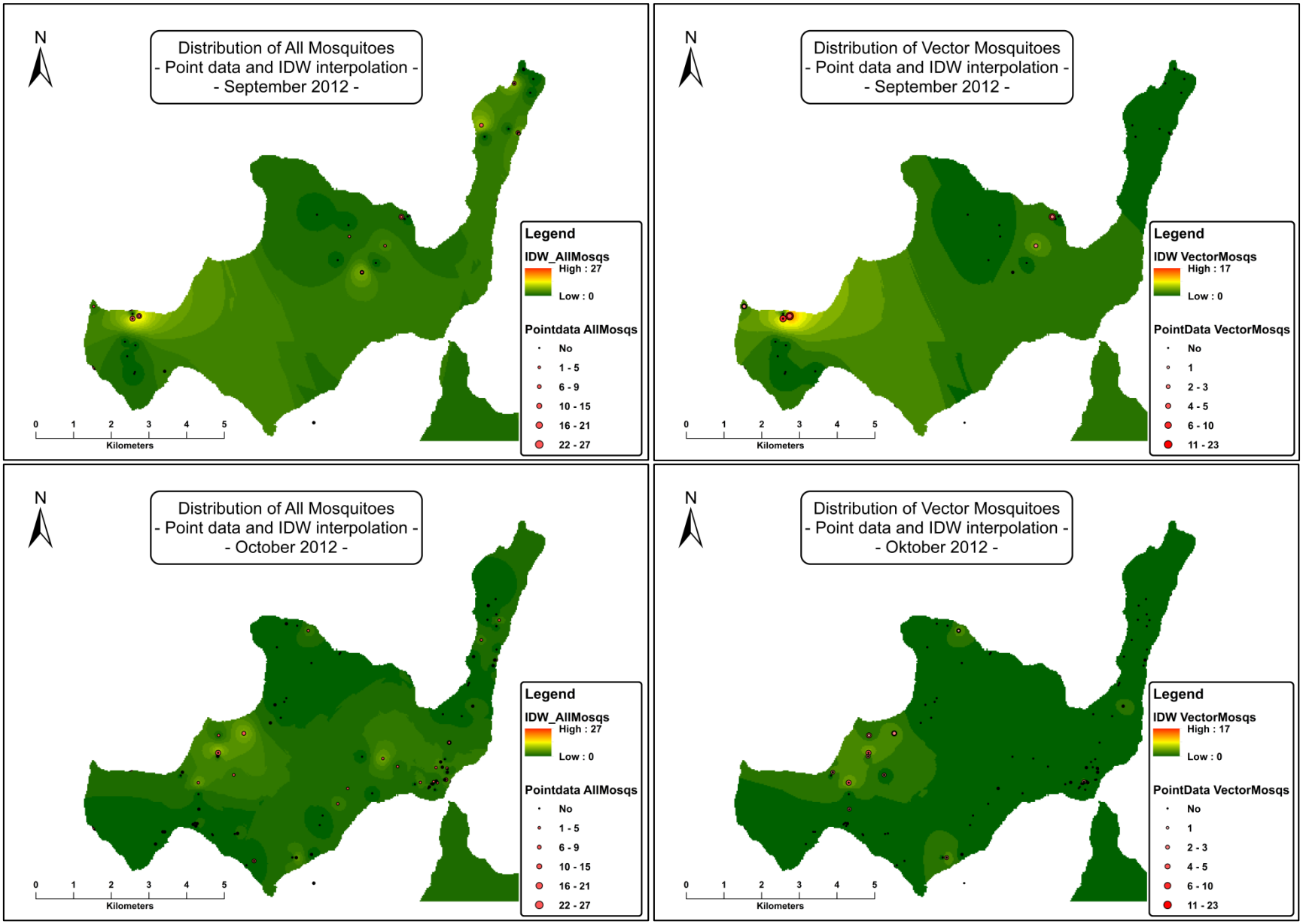

Violin Plot of DistToOpenWater
Sample Points
Whole Island
0
100
200
300
400
DistToOpenWater (m)

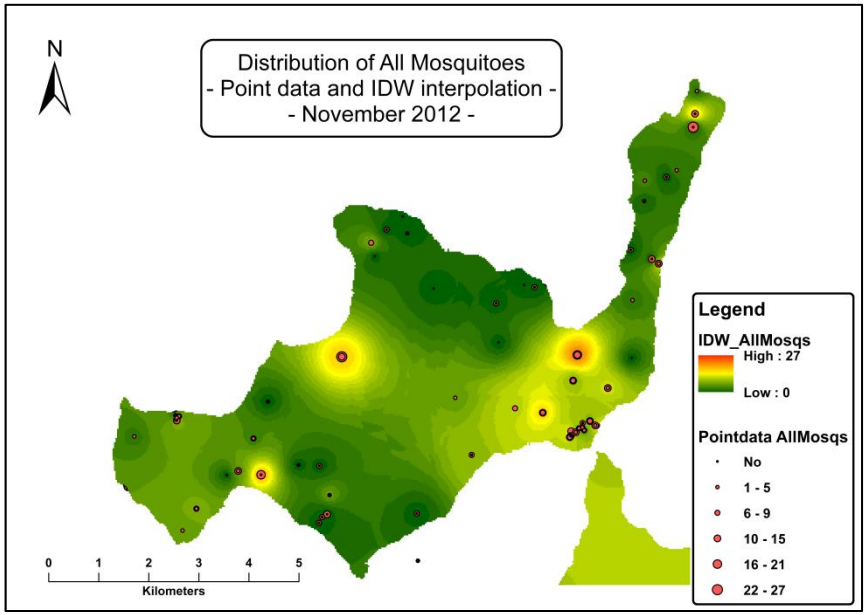

Histogram of TPI - sample points vs. whole island
Sample Points
Whole Island
Overlap
Frequency Density
0.00
0.05
0.10
0.15
0.20
-30
-20
-10
0
10
20
30
TPI

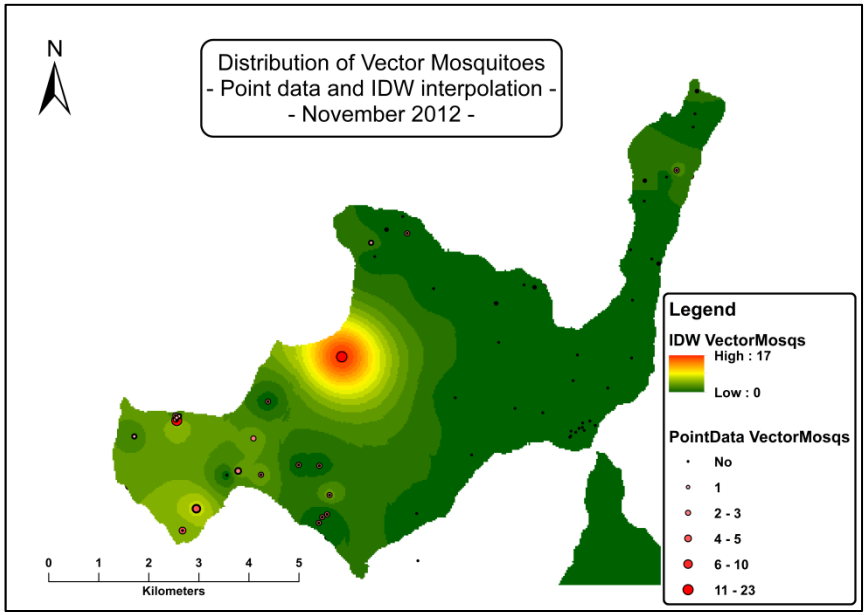

Violin Plot of TPI
Sample Points
Whole Island
-30
-20
-10
0
10
20
30
TPI (-)

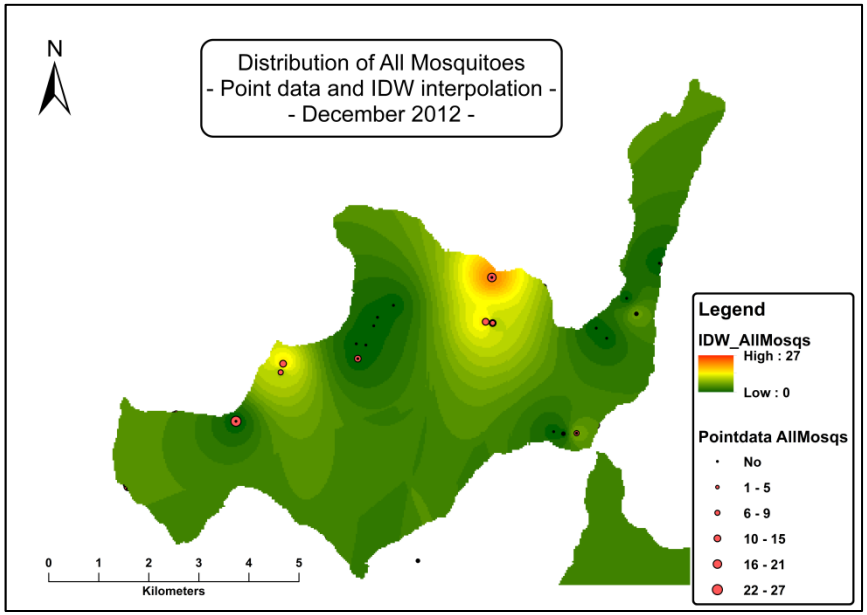
Appendix D

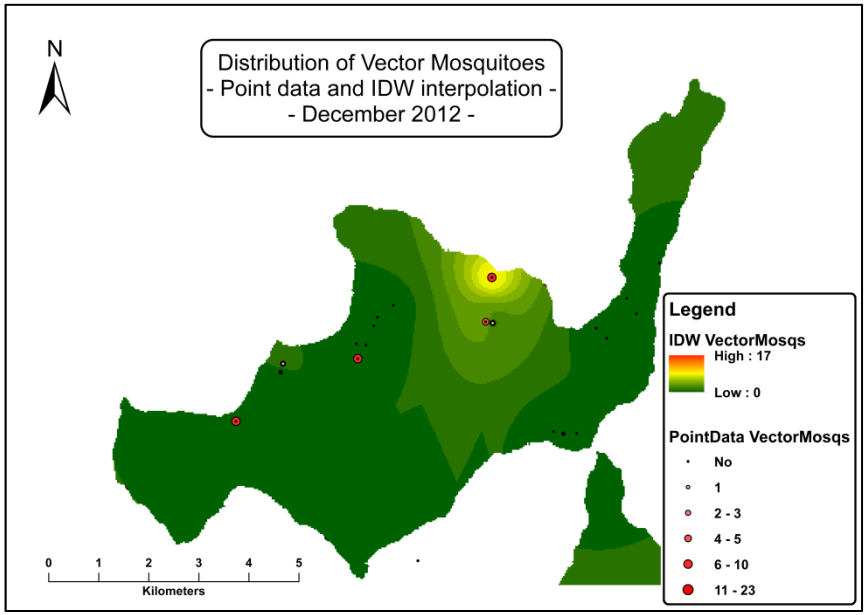
Spatial distribution of mosquitoes per month, for vector mosquitoes and all mosquitoes together (only female).

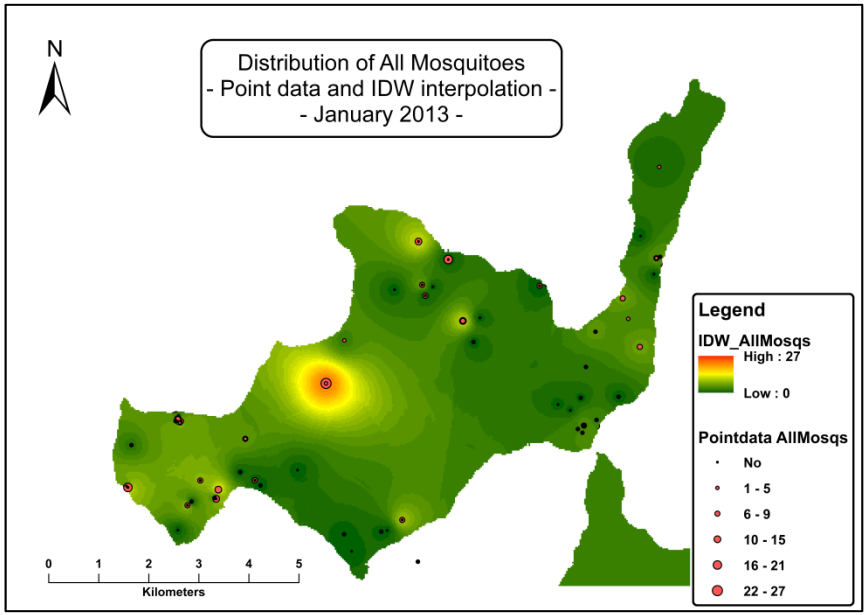


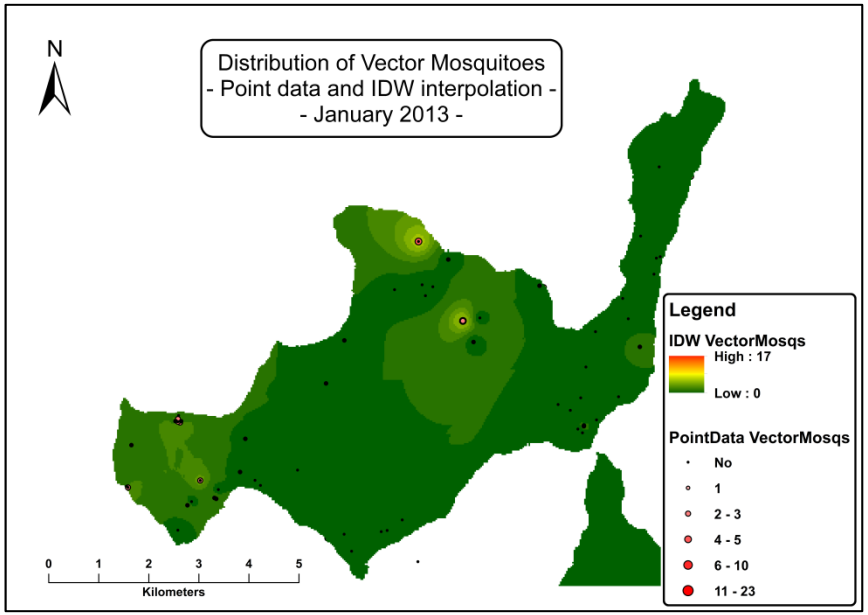

N
Distribution of All Mosquitoes
- Point data and IDW interpolation -
- November 2012 -
Legend
IDW_AllMosqs
High : 27
Low : 0
Pointdata AllMosqs
No
1 - 5
6 - 9
10 - 15
16 - 21
22 - 27
0 1 2 3 4 5
Kilometers

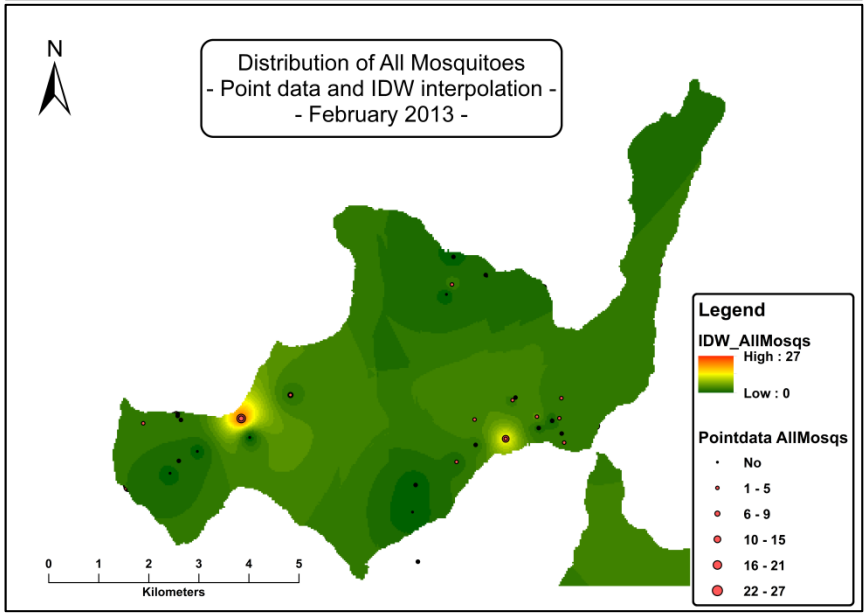

N
Distribution of Vector Mosquitoes
- Point data and IDW interpolation -
- November 2012 -
Legend
IDW VectorMosqs
High : 17
Low : 0
PointData VectorMosqs
No
1
2 - 3
4 - 5
6 - 10
11 - 23
0 1 2 3 4 5
Kilometers

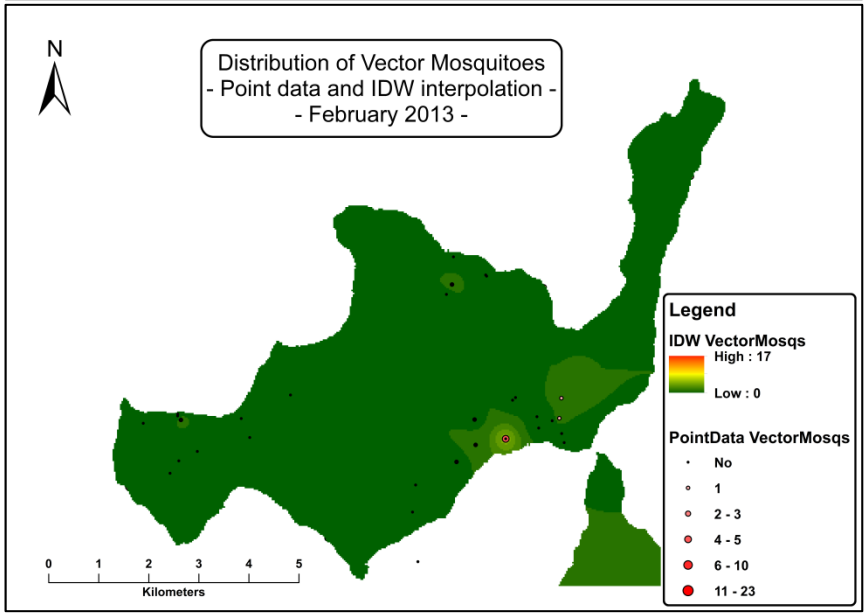

N
Distribution of All Mosquitoes
- Point data and IDW interpolation -
- December 2012 -
Legend
IDW_AllMosqs
High : 27
Low : 0
Pointdata AllMosqs
No
1 - 5
6 - 9
10 - 15
16 - 21
22 - 27
0 1 2 3 4 5
Kilometers

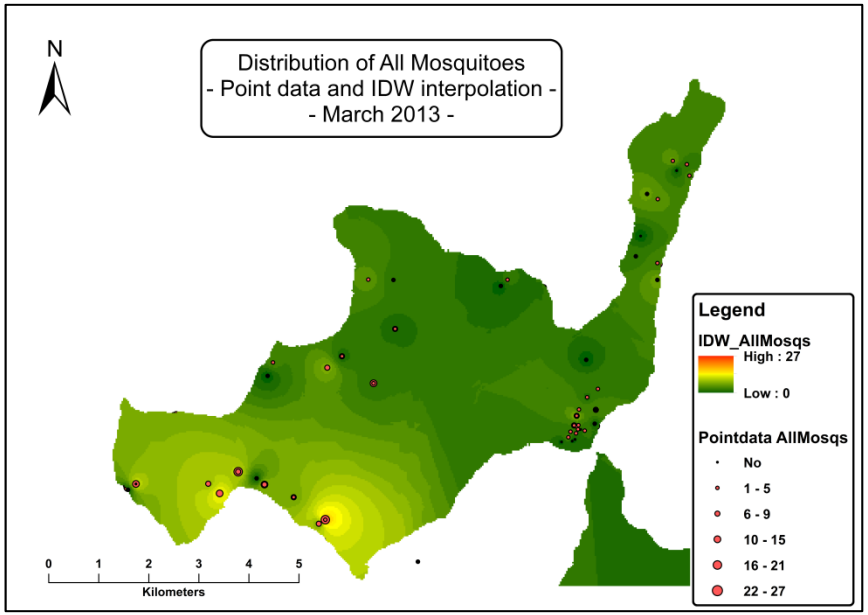

N
Distribution of Vector Mosquitoes
- Point data and IDW interpolation -
- December 2012 -
Legend
IDW VectorMosqs
High : 17
Low : 0
PointData VectorMosqs
No
1
2 - 3
4 - 5
6 - 10
11 - 23
0 1 2 3 4 5
Kilometers

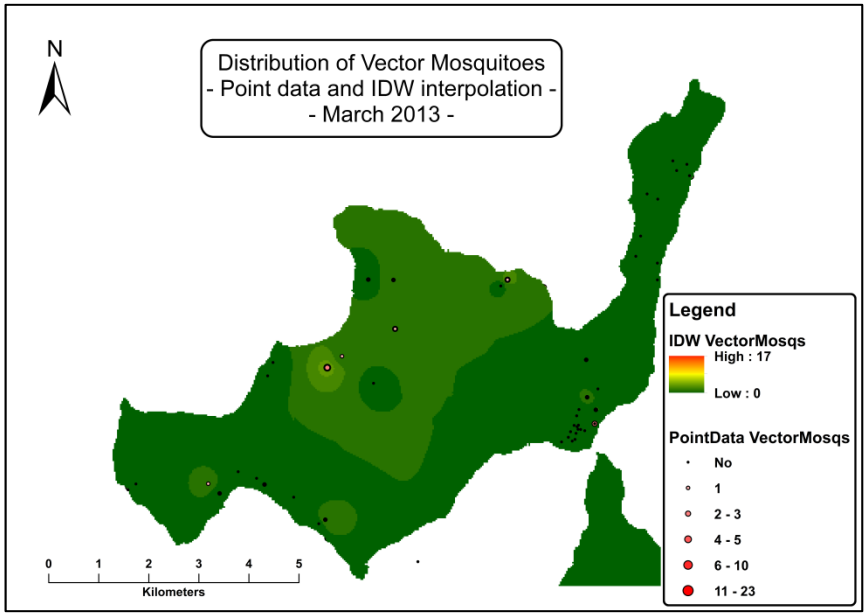

N
Distribution of All Mosquitoes
- Point data and IDW interpolation -
- January 2013 -
Legend
IDW_AllMosqs
High : 27
Low : 0
Pointdata AllMosqs
No
1 - 5
6 - 9
10 - 15
16 - 21
22 - 27
0 1 2 3 4 5
Kilometers

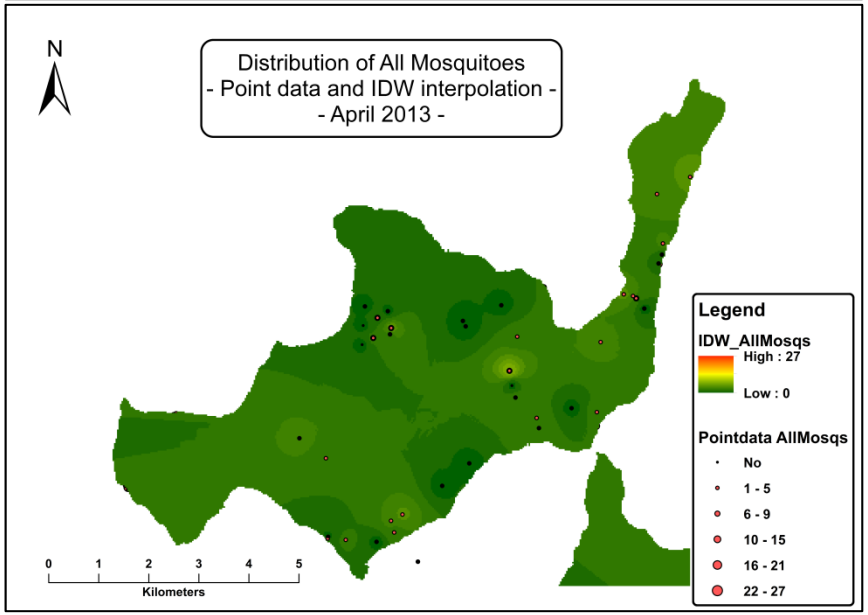

N
Distribution of Vector Mosquitoes
- Point data and IDW interpolation -
- January 2013 -
Legend
IDW VectorMosqs
High : 17
Low : 0
PointData VectorMosqs
No
1
2 - 3
4 - 5
6 - 10
11 - 23
0 1 2 3 4 5
Kilometers

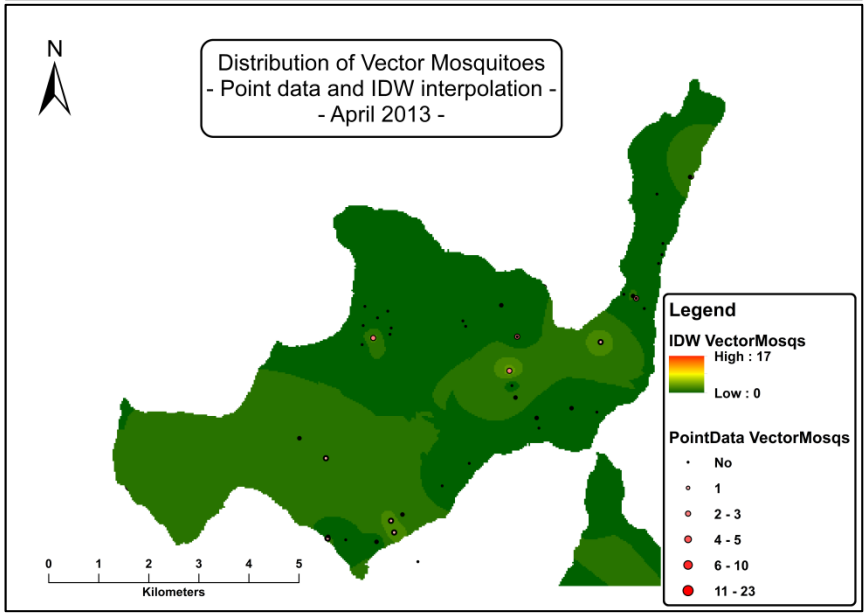

N
Distribution of All Mosquitoes
- Point data and IDW interpolation -
- February 2013 -
Legend
IDW_AllMosqs
High : 27
Low : 0
Pointdata AllMosqs
No
1 - 5
6 - 9
10 - 15
16 - 21
22 - 27
0 1 2 3 4 5
Kilometers

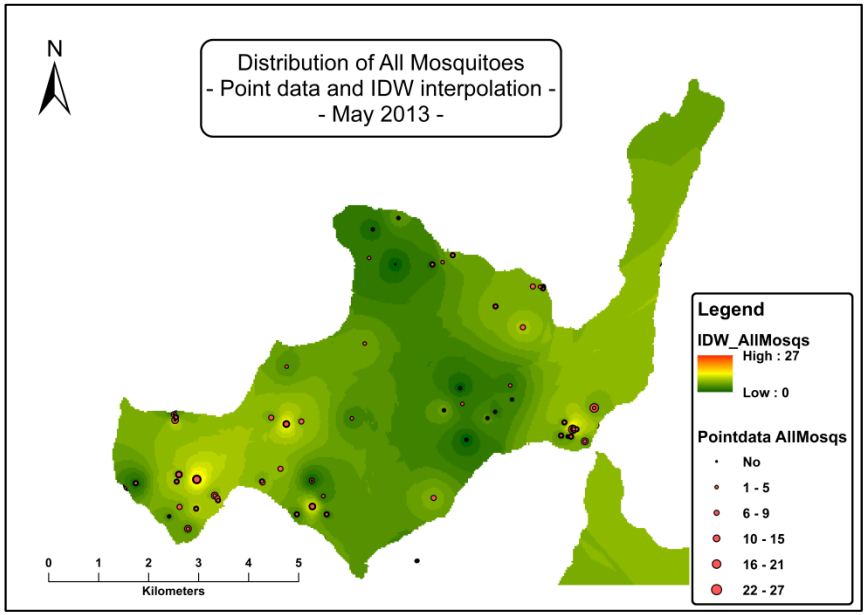

N
Distribution of Vector Mosquitoes
- Point data and IDW interpolation -
- February 2013 -
Legend
IDW VectorMosqs
High : 17
Low : 0
PointData VectorMosqs
No
1
2 - 3
4 - 5
6 - 10
11 - 23
0 1 2 3 4 5
Kilometers

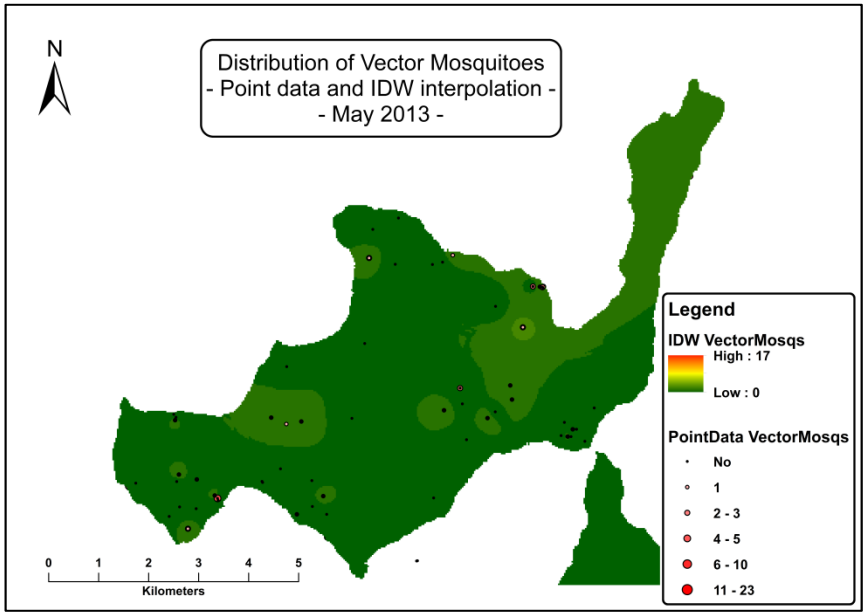

N
Distribution of All Mosquitoes
- Point data and IDW interpolation -
- March 2013 -
Legend
IDW_AllMosqs
High : 27
Low : 0
Pointdata AllMosqs
No
1 - 5
6 - 9
10 - 15
16 - 21
22 - 27
0 1 2 3 4 5
Kilometers

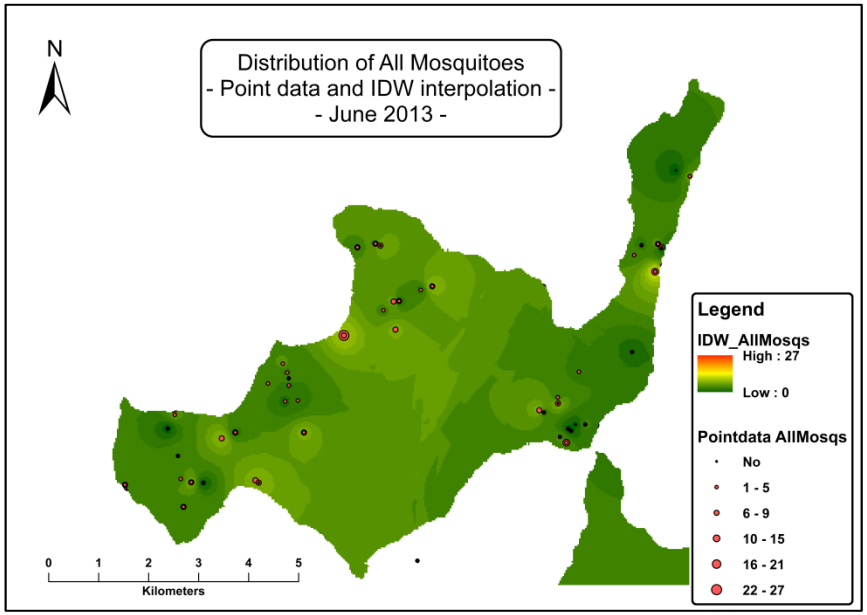

N
Distribution of Vector Mosquitoes
- Point data and IDW interpolation -
- March 2013 -
Legend
IDW VectorMosqs
High : 17
Low : 0
PointData VectorMosqs
No
1
2 - 3
4 - 5
6 - 10
11 - 23
0 1 2 3 4 5
Kilometers

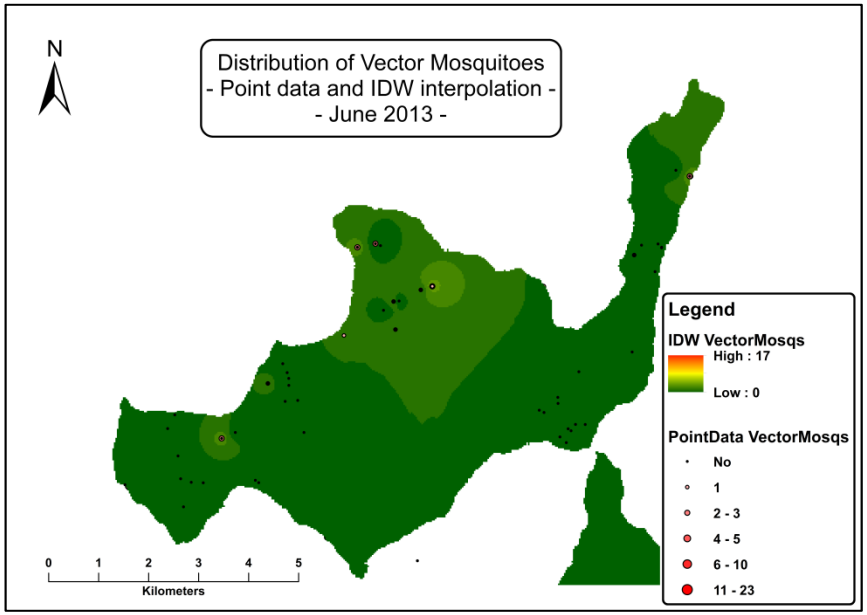

N
Distribution of All Mosquitoes
- Point data and IDW interpolation -
- April 2013 -
Legend
IDW_AllMosqs
High : 27
Low : 0
Pointdata AllMosqs
No
1 - 5
6 - 9
10 - 15
16 - 21
22 - 27
0 1 2 3 4 5
Kilometers

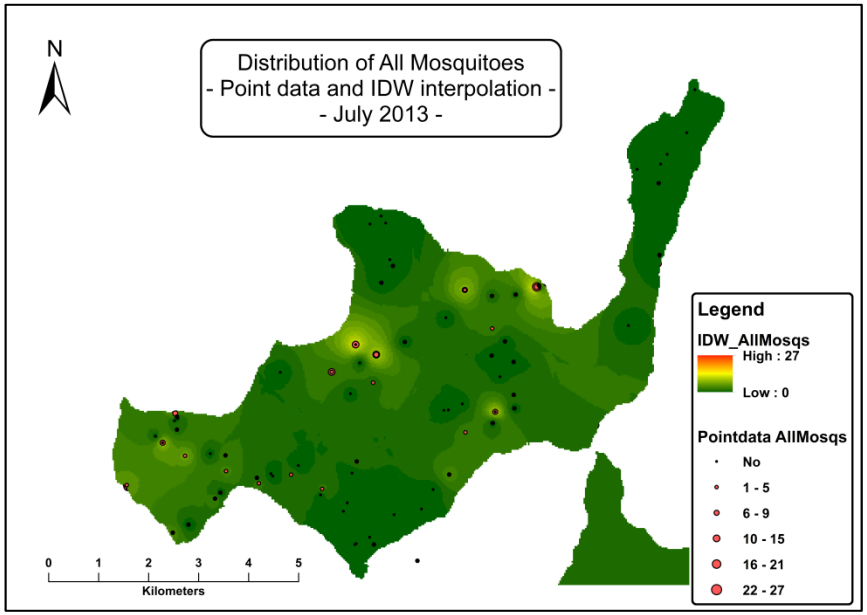

N
Distribution of Vector Mosquitoes
- Point data and IDW interpolation -
- April 2013 -
Legend
IDW VectorMosqs
High : 17
Low : 0
PointData VectorMosqs
No
1
2 - 3
4 - 5
6 - 10
11 - 23
0 1 2 3 4 5
Kilometers

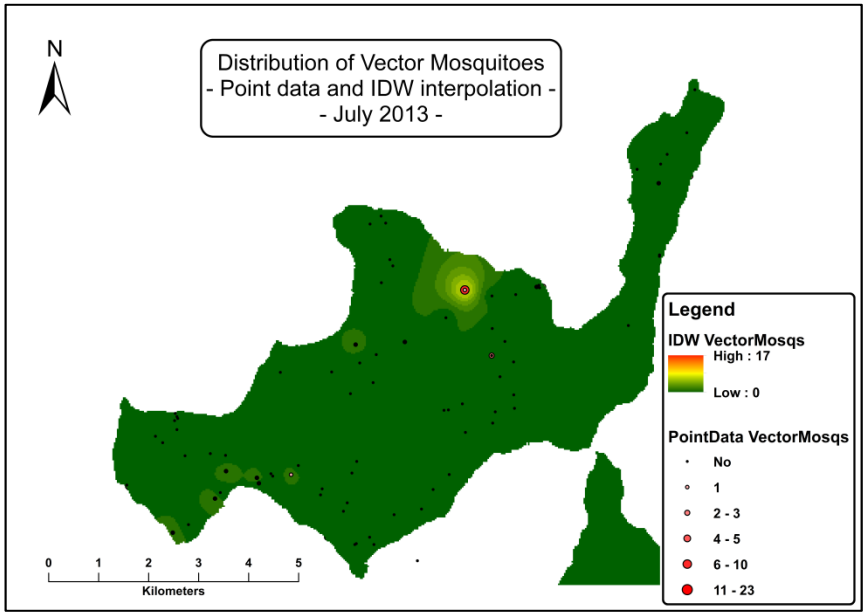

N
Distribution of All Mosquitoes
- Point data and IDW interpolation -
- May 2013 -
Legend
IDW_AllMosqs
High : 27
Low : 0
Pointdata AllMosqs
No
1 - 5
6 - 9
10 - 15
16 - 21
22 - 27
0 1 2 3 4 5
Kilometers

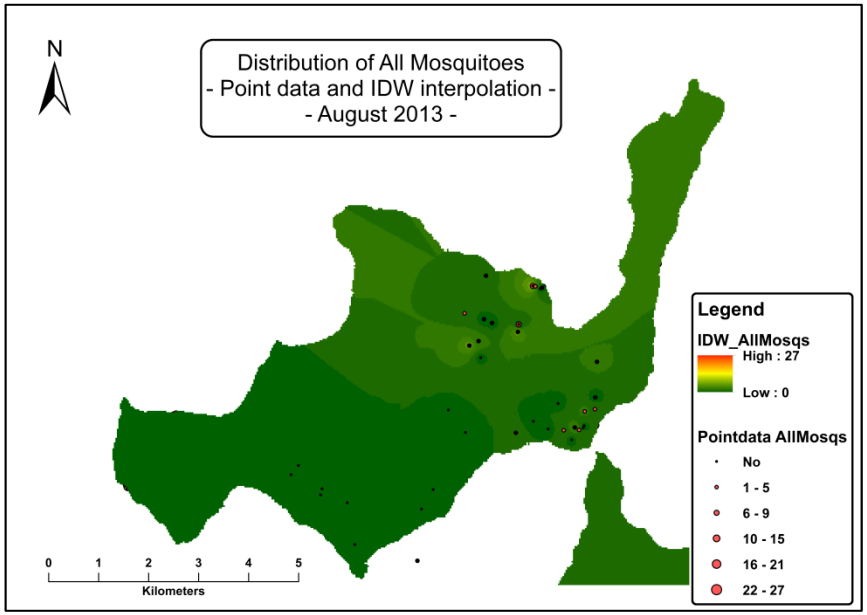

N
Distribution of Vector Mosquitoes
- Point data and IDW interpolation -
- May 2013 -
Legend
IDW VectorMosqs
High : 17
Low : 0
PointData VectorMosqs
No
1
2 - 3
4 - 5
6 - 10
11 - 23
0 1 2 3 4 5
Kilometers

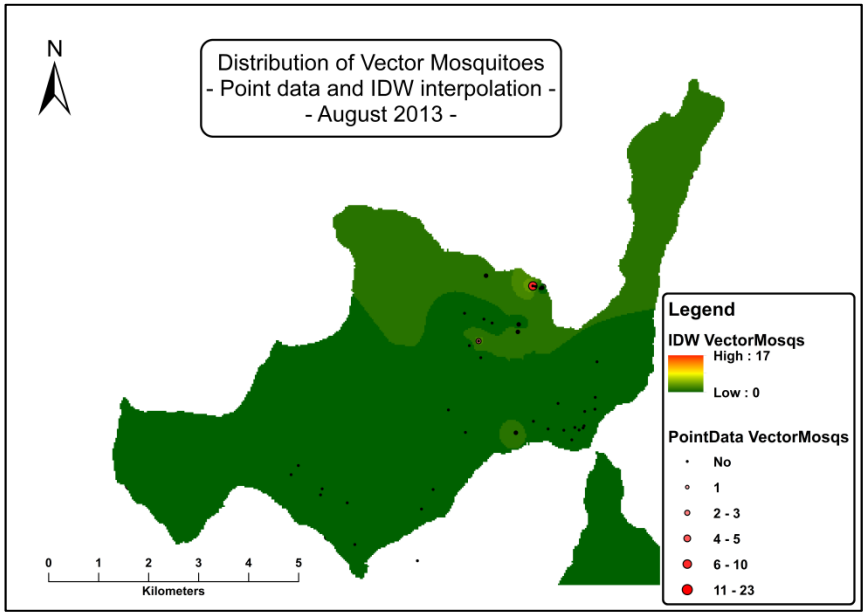

N
Distribution of All Mosquitoes
- Point data and IDW interpolation -
- June 2013 -
Legend
IDW_AllMosqs
High : 27
Low : 0
Pointdata AllMosqs
No
1 - 5
6 - 9
10 - 15
16 - 21
22 - 27
0 1 2 3 4 5
Kilometers


N
Distribution of Vector Mosquitoes
- Point data and IDW interpolation -
- June 2013 -
Legend
IDW VectorMosqs
High : 17
Low : 0
PointData VectorMosqs
No
1
2 - 3
4 - 5
6 - 10
11 - 23
0 1 2 3 4 5
Kilometers


N
Distribution of All Mosquitoes
- Point data and IDW interpolation -
- July 2013 -
Legend
IDW_AllMosqs
High : 27
Low : 0
Pointdata AllMosqs
No
1 - 5
6 - 9
10 - 15
16 - 21
22 - 27
0 1 2 3 4 5
Kilometers

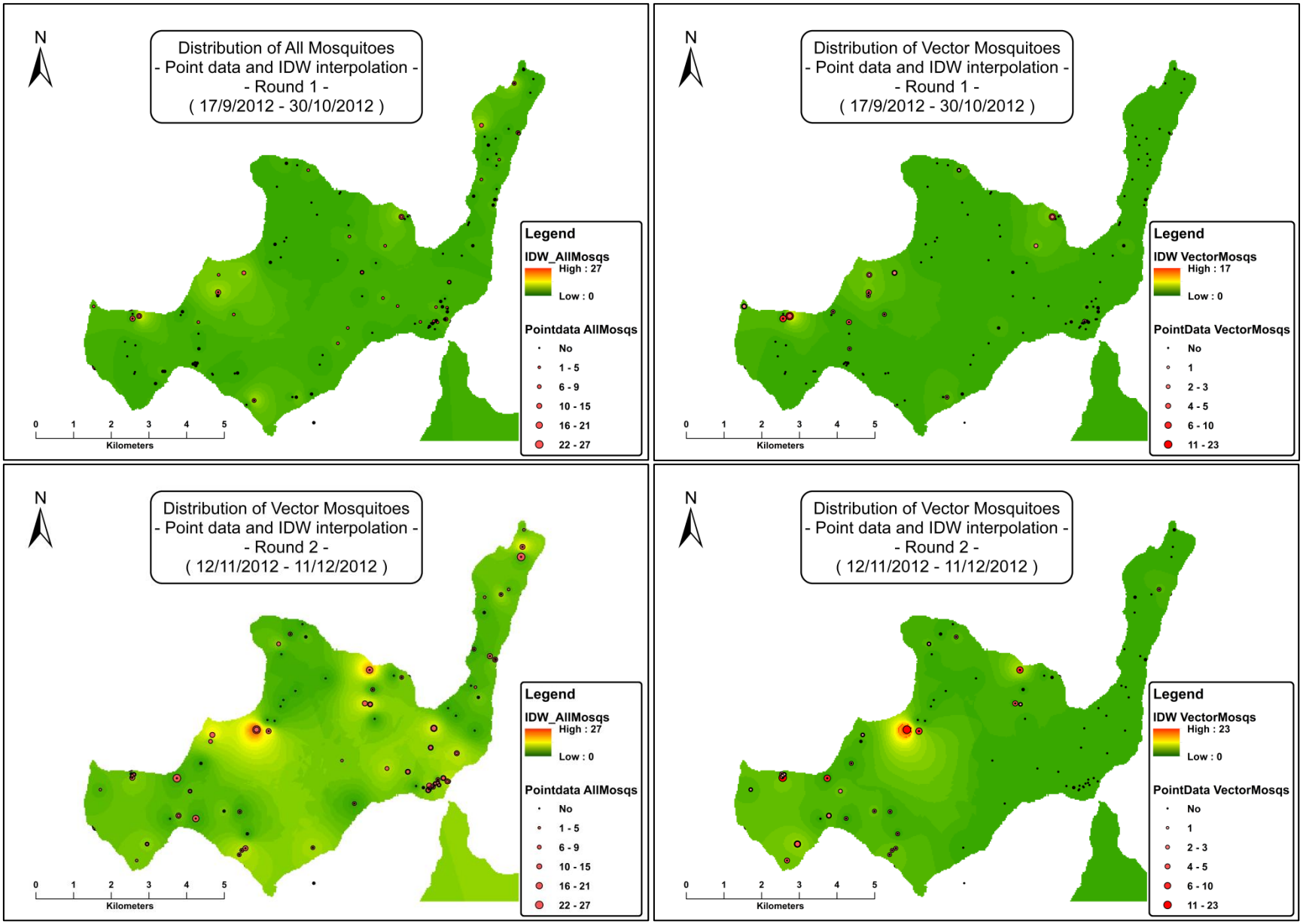

N
Distribution of Vector Mosquitoes
- Point data and IDW interpolation -
- July 2013 -
Legend
IDW VectorMosqs
High : 17
Low : 0
PointData VectorMosqs
No
1
2 - 3
4 - 5
6 - 10
11 - 23
0 1 2 3 4 5
Kilometers

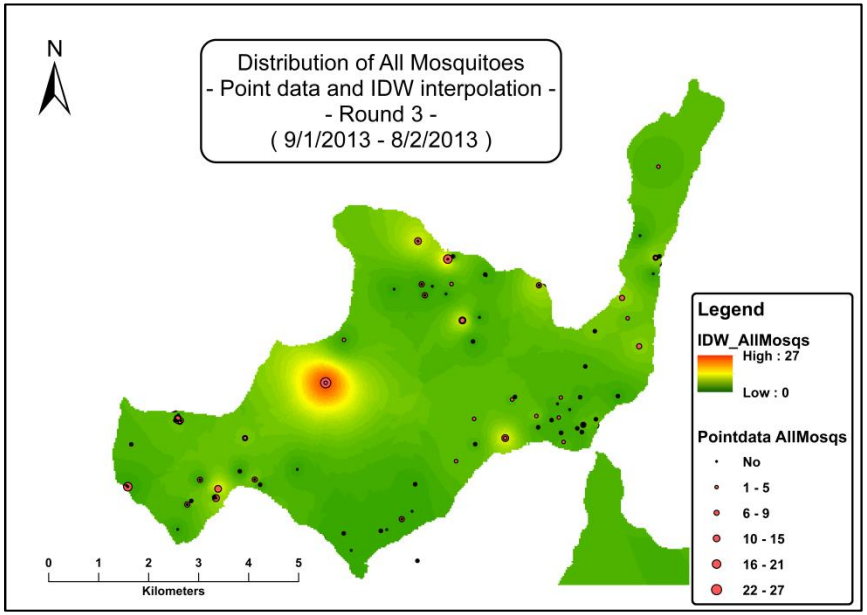

N
Distribution of All Mosquitoes
- Point data and IDW interpolation -
- August 2013 -
Legend
IDW_AllMosqs
High : 27
Low : 0
Pointdata AllMosqs
No
1 - 5
6 - 9
10 - 15
16 - 21
22 - 27
0 1 2 3 4 5
Kilometers

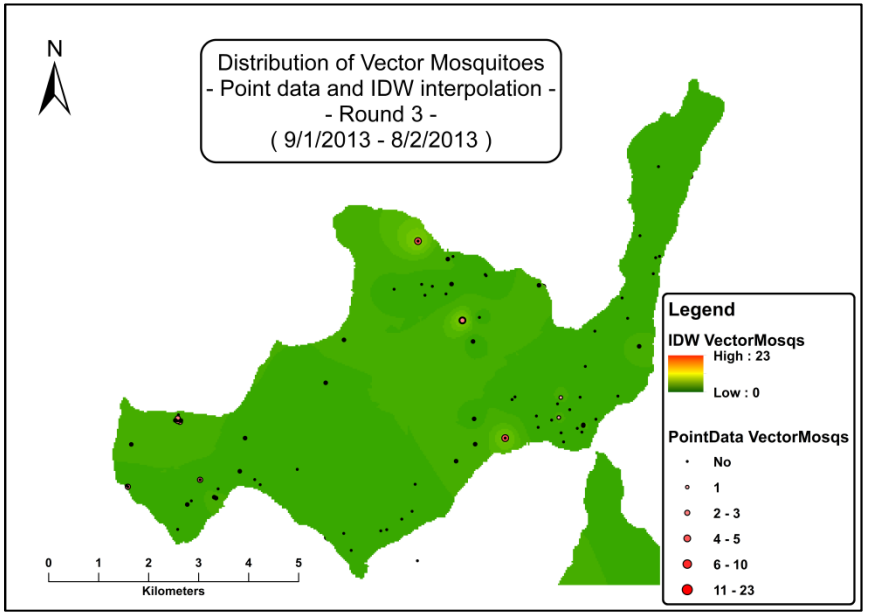

N
Distribution of Vector Mosquitoes
- Point data and IDW interpolation -
- August 2013 -
Legend
IDW VectorMosqs
High : 17
Low : 0
PointData VectorMosqs
No
1
2 - 3
4 - 5
6 - 10
11 - 23
0 1 2 3 4 5
Kilometers

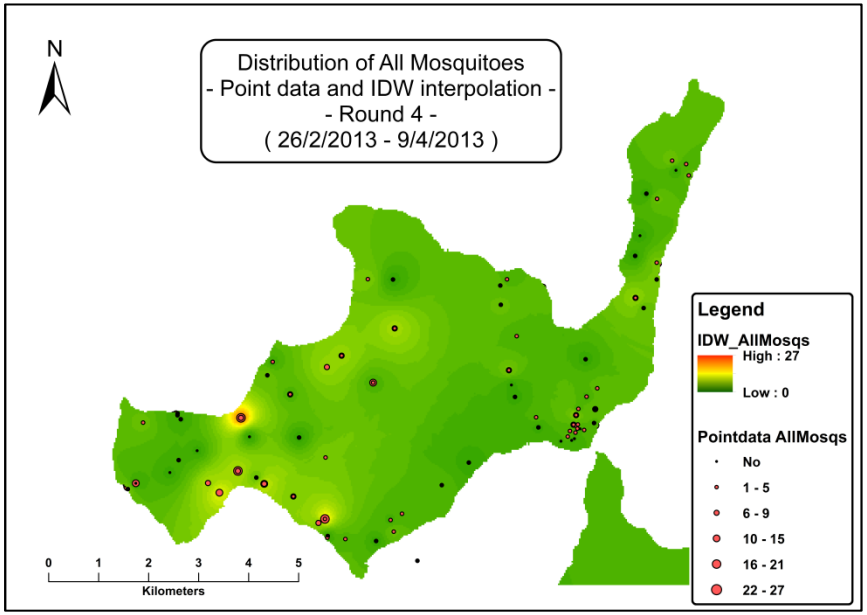
Appendix E

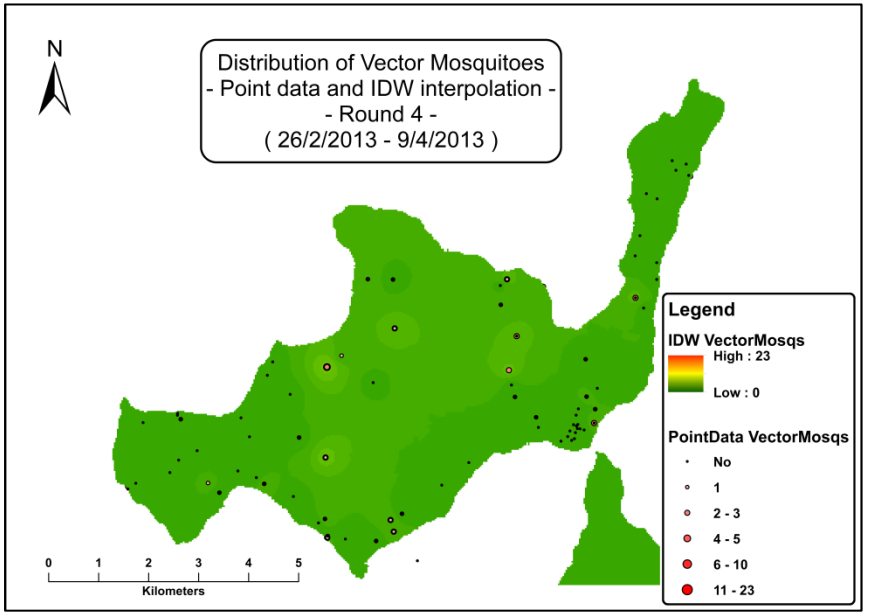
Spatial distribution of mosquitoes per sampling round, for vector mosquitoes and all mosquitoes together (only female).

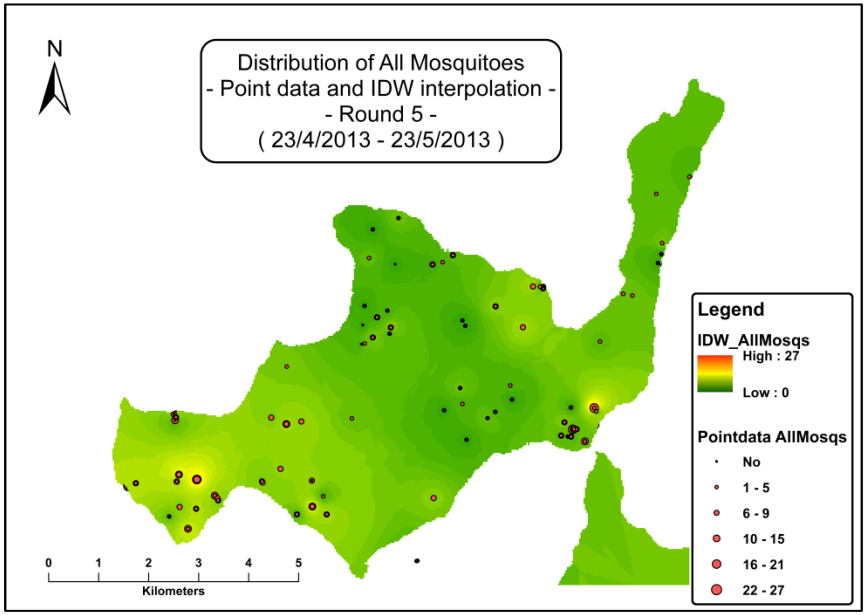


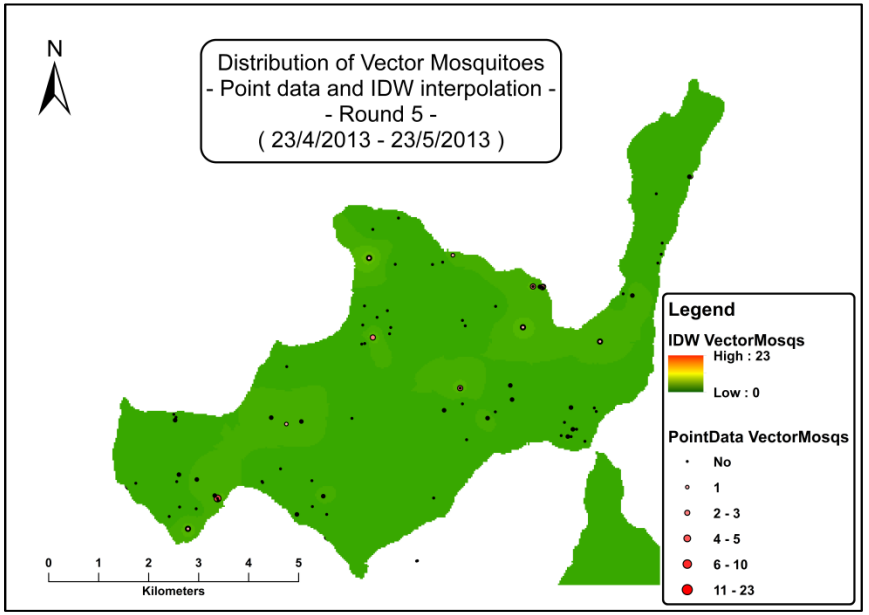

N
Distribution of All Mosquitoes
- Point data and IDW interpolation -
- Round 3 -
(9/1/2013 - 8/2/2013)
Legend
IDW_AllMosqs
High : 27
Low : 0
Pointdata AllMosqs
No
1 - 5
6 - 9
10 - 15
16 - 21
22 - 27
0 1 2 3 4 5
Kilometers

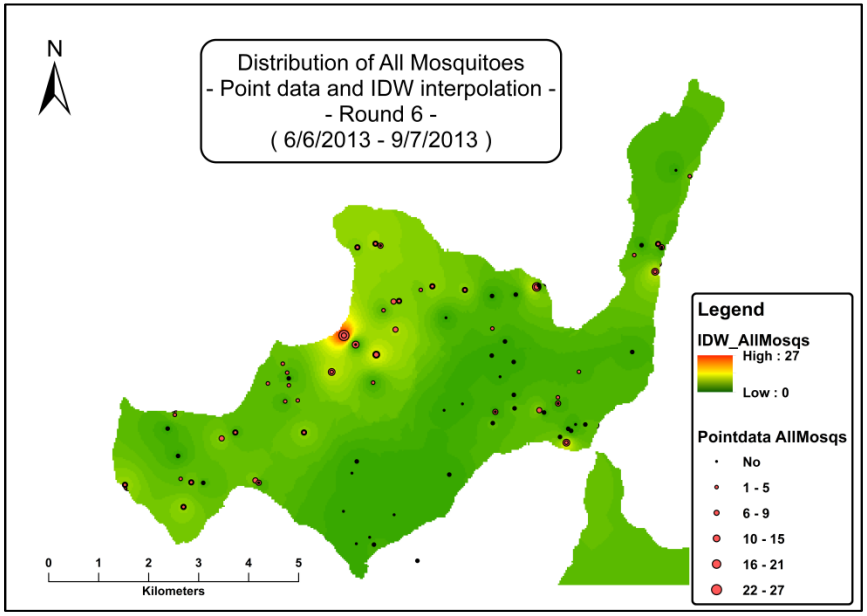

N
Distribution of Vector Mosquitoes
- Point data and IDW interpolation -
- Round 3 -
(9/1/2013 - 8/2/2013)
Legend
IDW VectorMosqs
High : 23
Low : 0
PointData VectorMosqs
No
1
2 - 3
4 - 5
6 - 10
11 - 23
0 1 2 3 4 5
Kilometers

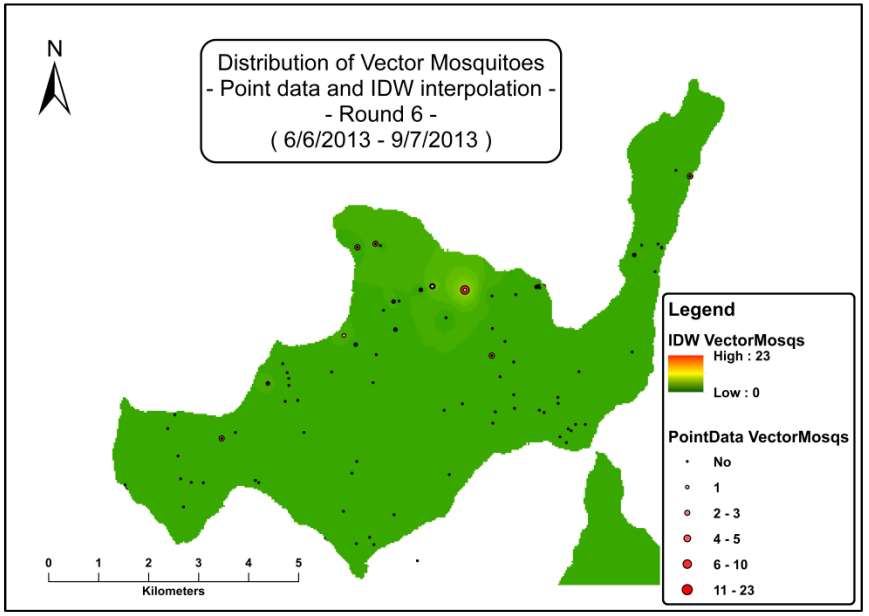

N
Distribution of All Mosquitoes
- Point data and IDW interpolation -
- Round 4 -
(26/2/2013 - 9/4/2013)
Legend
IDW_AllMosqs
High : 27
Low : 0
Pointdata AllMosqs
No
1 - 5
6 - 9
10 - 15
16 - 21
22 - 27
0 1 2 3 4 5
Kilometers

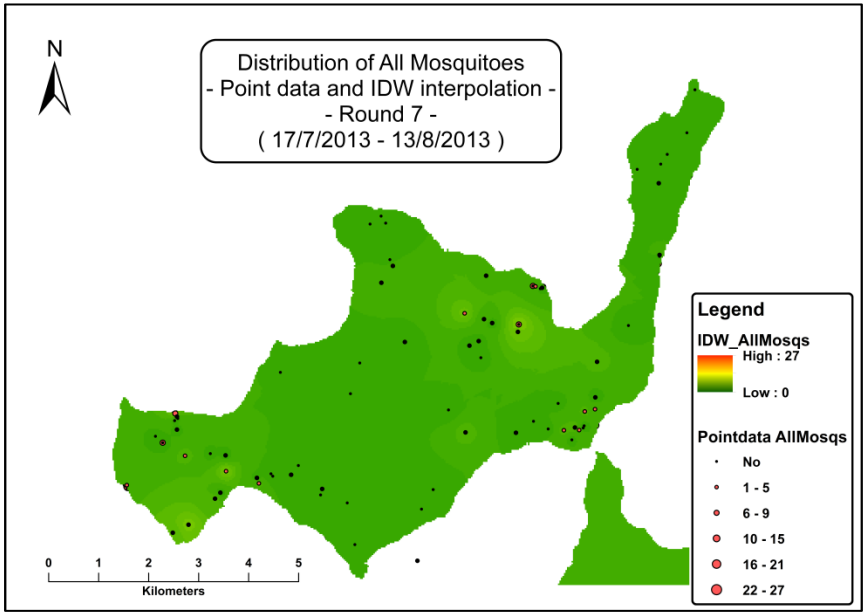

N
Distribution of Vector Mosquitoes
- Point data and IDW interpolation -
- Round 4 -
(26/2/2013 - 9/4/2013)
Legend
IDW VectorMosqs
High : 23
Low : 0
PointData VectorMosqs
No
1
2 - 3
4 - 5
6 - 10
11 - 23
0 1 2 3 4 5
Kilometers

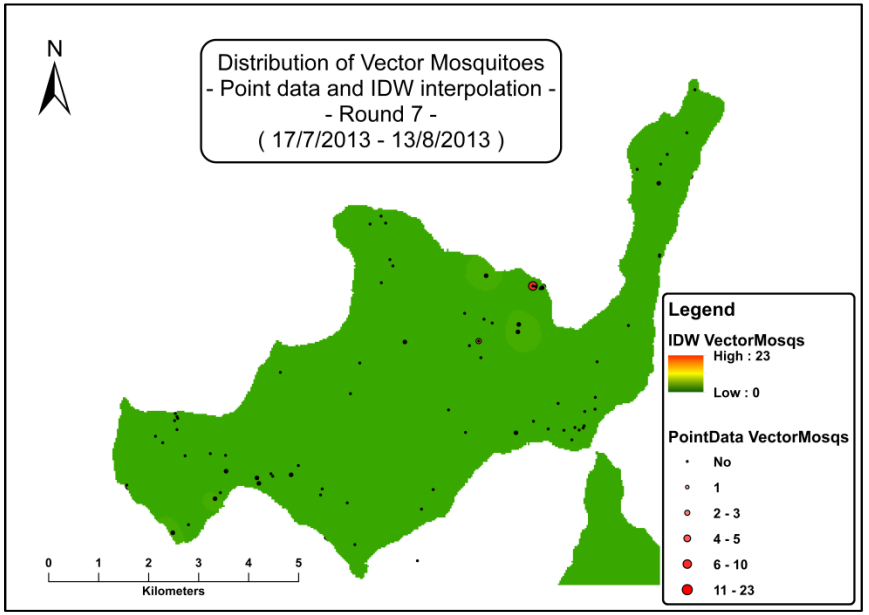

N
Distribution of All Mosquitoes
- Point data and IDW interpolation -
- Round 5 -
(23/4/2013 - 23/5/2013)
Legend
IDW_AllMosqs
High : 27
Low : 0
Pointdata AllMosqs
No
1 - 5
6 - 9
10 - 15
16 - 21
22 - 27
0 1 2 3 4 5
Kilometers


N
Distribution of Vector Mosquitoes
- Point data and IDW interpolation -
- Round 5 -
(23/4/2013 - 23/5/2013)
Legend
IDW VectorMosqs
High : 23
Low : 0
PointData VectorMosqs
No
1
2 - 3
4 - 5
6 - 10
11 - 23
0 1 2 3 4 5
Kilometers


N
Distribution of All Mosquitoes
- Point data and IDW interpolation -
- Round 6 -
(6/6/2013 - 9/7/2013)
Legend
IDW_AllMosqs
High : 27
Low : 0
Pointdata AllMosqs
No
1 - 5
6 - 9
10 - 15
16 - 21
22 - 27
0 1 2 3 4 5
Kilometers


N
Distribution of Vector Mosquitoes
- Point data and IDW interpolation -
- Round 6 -
(6/6/2013 - 9/7/2013)
Legend
IDW VectorMosqs
High : 23
Low : 0
PointData VectorMosqs
No
1
2 - 3
4 - 5
6 - 10
11 - 23
0 1 2 3 4 5
Kilometers


N
Distribution of All Mosquitoes
- Point data and IDW interpolation -
- Round 7 -
(17/7/2013 - 13/8/2013)
Legend
IDW_AllMosqs
High : 27
Low : 0
Pointdata AllMosqs
No
1 - 5
6 - 9
10 - 15
16 - 21
22 - 27
0 1 2 3 4 5
Kilometers


N
Distribution of Vector Mosquitoes
- Point data and IDW interpolation -
- Round 7 -
(17/7/2013 - 13/8/2013)
Legend
IDW VectorMosqs
High : 23
Low : 0
PointData VectorMosqs
No
1
2 - 3
4 - 5
6 - 10
11 - 23
0 1 2 3 4 5
Kilometers

Appendix F

Validation of the geographical positions of houses on Rusinga Island

1. Background and problem description

All houses on Rusinga Island are positioned by measuring the GPS coordinates. The resulting dataset is used in this study to know the locations where the mosquito abundance data was collected. Since the start of the SolarMal project it was known that there is an inaccuracy involved in these measured positions; however the magnitude of the accuracy was uncertain. Fieldworkers of the SolarMal project reported that some houses could not be found using the belonging GPS coordinates, causing them to request help of local people. Inaccuracy of the measured positions leads to problems in finding the houses. Inaccuracy also affects the spatial analysis of the data that belong to the houses. The mosquito abundance data for example, which is included in the spatial analysis of this thesis project, is also linked to the houses. Inaccuracy in this leads to inaccuracy of the data analysis, since mosquito data is not collected at the location they seem to be collected. Inaccuracy in the data can finally lead to incorrect results and conclusions. It is important for the SolarMal project to know what the accuracy of the spatial data is, so it can adapt on this if necessary. It is for this thesis project especially important to investigate what the accuracy is of the mosquito data, to investigate whether this influenced the spatial analysis.

2. Objective

The objective is to validate the accuracy of the house positions on Rusinga Island and to assess whether this accuracy could be improved. This study will tackle the following research questions:

RQ4. How accurate were the positions of the houses on Rusinga Island measured?

RQ5. Which method of measuring positions is feasible within the SolarMal project and leads to the best achievable accuracy?

3. Experiments

This study to the accuracy of house positions on Rusinga island was performed from the 22th of May till the 18th of June, 2014. This is just in the last part of the long rain season. All experiments done, analysis and results are discussed in this chapter. Three main experiments were performed. The first experiment aims to investigate the accuracy of a GPS device in the spatial and temporal dimension. The second experiment aims to validate several methods of measuring the positions of houses on Rusinga Island. The last experiment was a consecutive of the second experiment, where the most feasible and best options for this project were validated again for a larger set of houses.

For the experiments three devices were available. First of all a Garmin eTrex 30, which was provided by the GIS department of the Wageningen University. In this chapter this device will be called the Garmin Device. The second device is a Samsung tablet, model GT-P7510, black colored and is therefore called the Black Tablet. The third device is also a Samsung tablet, a newer model, the SM-T210, white colored and therefore called the White Tablet. Both Samsung tablets are equipped with a WIFI and GPS receiver and were both provided by the SolarMal project.

On the Garmin device, the WAAS/EGNOS option was turned on for better accuracy of the measured position. WAAS and EGNOS stand respectively for Wide Area Augmentation System and Euro Geostationary Navigation Overlay Service. Both correct for typical GPS signal errors like for example atmospheric disturbances of the GPS signal, where WAAS is an American system and EGNOS the European version (Garmin 2014). According to the Garmin company this option would only matter in North-America and Europe and not in Africa (Garmin 2014). However during a small experiment it appeared that the precision of the measurements was higher with the WAAS/EGNOS option on. See Appendix H for the details of this experiment.

The Garmin Device was also set to receive signals from both GPS and GLONASS satellites. Both are chosen since this only increases the amount of satellites available to receive information from. The only disadvantage is a reduction of the battery life (Garmin 2011).

3.1. Accuracy & precision, and the time dependency, of the Garmin device

The first experiment performed, aims to validate the accuracy and the precision, including the time dependency of these, of GPS positions measured by the Garmin device.

3.1.1. Methodology

To measure any variation of the Garmin device during the day, the Garmin device was placed on one fixed place. Figure 19 shows the position of the Garmin device on the *icipe* campus in Mbita, Western Kenya. It was installed on the roof of a car parking place at the campus. The Garmin device was collecting GPS data points with a time interval of 10 seconds, for a duration of 24 hours. It started tracking at 13:00 local time on the 27th of May till 14:30 on the 28th of May. Of this time series, 24 hours from 14:00 till 14:00 was selected for the analysis. The direct environment was not changed during the 24 hours. There was a typical clear-sky situation during daytime on the 27th, followed by overcasting and rain during the night. The morning of the 28th was more cloudy than usual for this area, however without rain. After 11:00 it became a typical clear-sky situation again. Overall, the weather circumstances during this experiment were typical for the local area during the rainy season.

The data from the Garmin device was imported in ArcGIS and plotted on the QuickBird image, resulting in a so-called cloud of points. During the data analysis the distribution of this cloud of points was studied in the spatial dimension by looking to the average distance and standard deviation in this to the average position of the point cloud. The time dependency was investigated by studying the spatial distribution of the point cloud per hour. These average distances gave an impression of the precision of the measurements. To investigate the accuracy of the measurements, the average position of the point cloud was compared with the 'true' position according to several map providers.

Figure 19: Location of the 24-hours experiment

3.1.2.Results

Figure 20 shows a map on the left side with the average position of all these points and the cloud of points itself. The right map in Figure 20 shows next to this the calculated standardized distances. The standard deviation of this cloud of points equals 3,68, meaning that 68% of the tracked points fall within 3,68 meters from the average position. Approximately 97.5% of the tracked points fall within 7,36 meters from the average position. According to the right map in Figure 20, one could state that the precision of the measured positions equals approximately 7,4 meter (with a 97,5% confidence interval).

The theoretical accuracy of the Garmin device given by the device itself was equal to 3 meter during the experiment. However this is a calculated value, not a measured accuracy. Figure 21 shows the position of the experiment according to several different map providers which is summarized in Table 13. If one compares this to the average position of the tracked points, the distances to this average point vary from 0,6 to 3,4 meter, with an average of 1,8 meter.

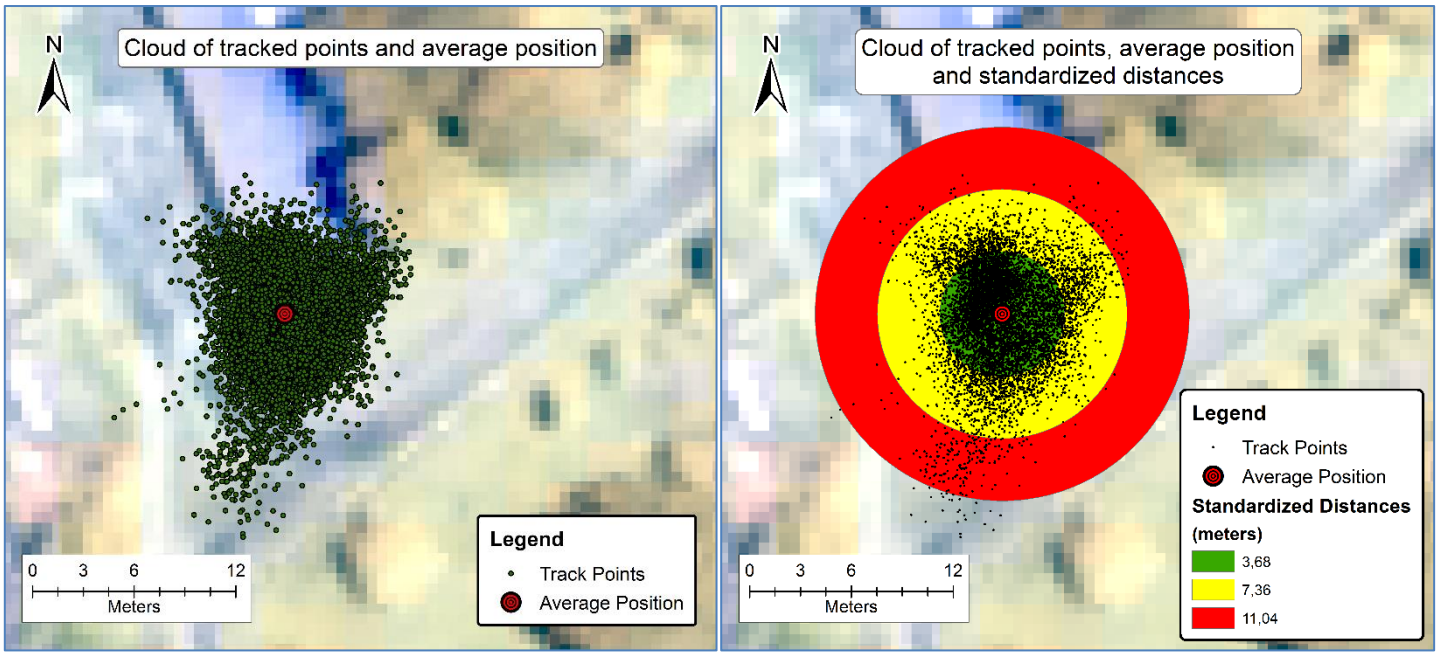


Figure 20: Cloud of tracked points and standardized distances of it (right)

Table 13: 'True positions' according to several map providers

'True Positions'			
Map Provider	Longitude	Latitude	Distance to Average Position (m)
QuickBird image	34,206157	-0,431877	0,6
maps.google.com	34,206167	-0,431881	0,7
GoogleEarth	34,206139	-0,431867	2,9
FlashEarth	34,206167	-0,431861	2,2
BingMaps	34,206165	-0,431850	3,4
ESRI basemap - World Imagery	34,206153	-0,431881	0,9
Average Position of points	34,206161	-0,431881	-

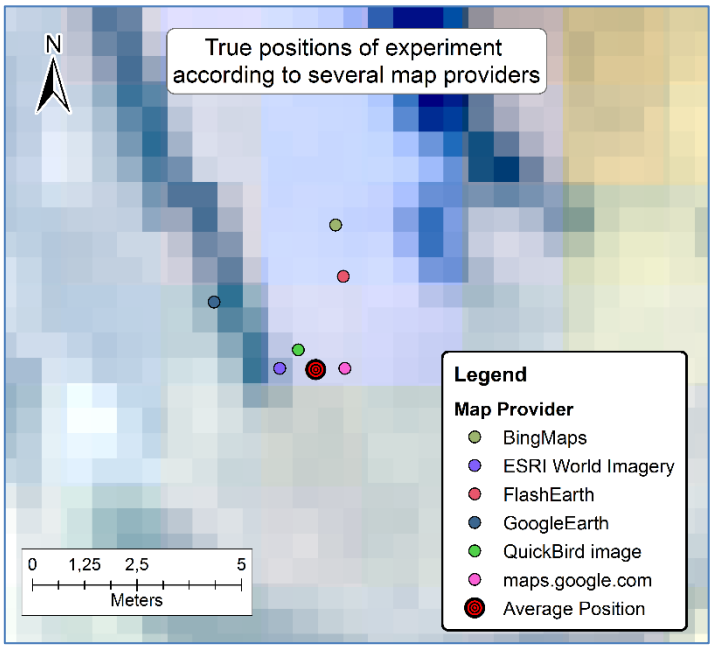


Figure 21: 'True positions' according to several providers plotted

The accuracy of the GPS measurements can vary in time, due to a temporal variation in the errors that cause the inaccuracy of GPS measurements (Olynik et al. 2002), wherefore it is necessary to look at the temporal variation in the GPS accuracy. Figure 22 therefor shows the cloud of points again, with for each hour another color. On top of the individual points, hourly average positions are plotted in the color of that hour. It appears that the points start more south than on average, move towards the north during the night hours and goes to the south again in the morning. Table 13 summarizes the hourly statistics like the standardized distance of the cloud of points on an hourly basis. The distances of the hourly averages to the total average position vary between 0,5 and 6,4 meter, with on average a distance of 2,7 meter. Depending on the moment of the day, the measured position has an extra accuracy factor on top of the already known accuracy. On average the displacement is 2,7 meters relative to the 24-hours measured position, which already had an accuracy of approximately 3 meters. This means that the total accuracy is less than 6 meters. The hourly based standardized distances, with a 95% confidence level, vary between 2,1 and 6,5 meters, with on average 4,2 meters. This means that the precision of measurements is varying over the day, with an average precision of around 4 meters.

Table 14: Hourly spatial statistics of tracked points

Hourly Statistics	average	min	max
Distance Hourly average position to total average position (m)	2,7	0,5	6,4
Standardized Distance (95%) of point cloud (m)	4,2	2,1	6,5

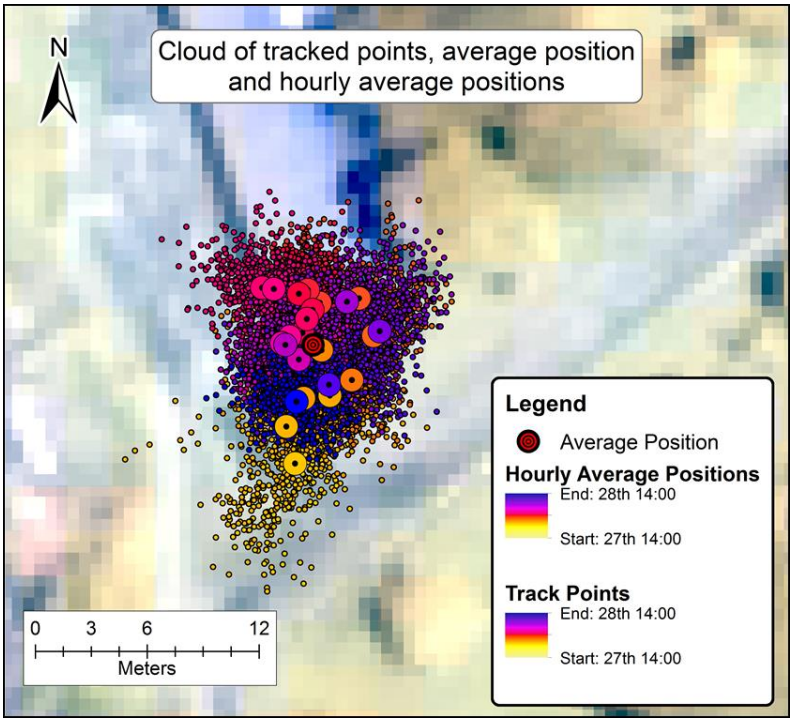


Figure 22: Cloud of points and hourly average positions, colored per hour

3.1.3.Conclusion and discussion

The distances from the average position of the point cloud and the 'true' locations according to the several map providers, varies from 0,6 to 3,4 meter. The question remains which map, if any, is giving the real true coordinates of the experiment. However, since the variation in possible true locations is within a distance of 3,4 meter, one could state that the accuracy provided by the Garmin device itself is realistic and for this project it can be assumed that the accuracy of the Garmin device is less than 3 meters.

Measuring the position at a certain moment in time involves a total accuracy of less than 6 meters. 3 meter inaccuracy is caused by the spatial accuracy, the other 3 meter inaccuracy is caused by a temporal variation of the spatial accuracy (Table 14). Measuring only one time, one should take into account a precision of around 4 meters (Table 14), resulting in a total uncertainty of the position of 10 meters. If one would measure several times and average the position, the precision can be ignored and a total uncertainty of approximately 6 meters remains.

3.2. Comparing and validation of methods for measuring positions of houses

The second experiment conducted concerns the comparison and validation of several methods for measuring the positions of houses on Rusinga Island.

3.2.1. Methodology

The position of houses was measured using different devices and with different methodology of measuring. Table 15 shows the combinations of devices, measures and location, which are described in detail below.

Different types of building structures and areas

First of all, measurements are done at two typical situations for Rusinga, namely rural and so-called 'peri-urban'. The first area is rural, in the North-Western part of the Island (see left picture in Figure 23). In that area the houses are mainly built separately from each other. The second area, peri-urban, was at a beach community, Southwest of the Island (see the right picture in Figure 23). Such a fishermen community is characterized by barracks positioned really close to each other. Figure 23 shows clearly these two types of situations.

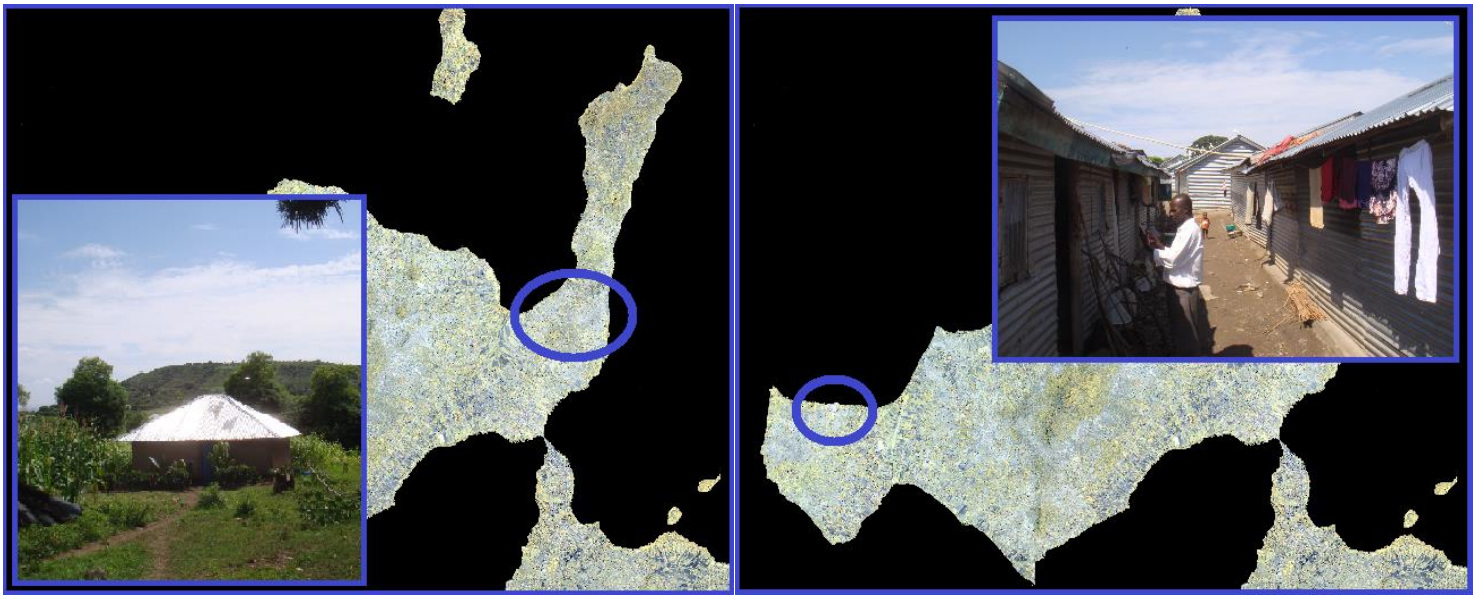


Figure 23: Areas of fieldwork for experiment 2 and two types of houses

Different devices

Three different devices were used, namely the Garmin device, the Black Tablet and the White Tablet. The Black Tablet is the tablet that is used by the SolarMal project for the initial positioning of the houses.

Different measures

All three devices were used to measure the position of houses by creating a so called Way Point (WP). Measuring a waypoint means that the device is deriving its current position from the available satellite signals and saves the coordinates that represent that position. Before measuring the position, the devices were hold stable at one place for a few seconds to ensure that the measured position would be as accurate as possible by fixing a WP. The Garmin device was used to measure next to a WP also a so called Average Way Point (AWP), where the Garmin device starts deriving its position for a duration of some minutes until its measured position reaches a stable average (Garmin 2009). During this measurement the Garmin device was put down on a fixed place until the device reported to be finished. The AWP is only measured in the rural area since it takes relatively a long time per measurement. In the peri-urban area this AWP measurement was skipped, so more measurements could be done on the second day in the same amount of time. The two tablets were also used, next to measuring a WP, to manually specify the position of a house. This was done by using satellite imagery provided by Google. Due to a human mistake, there was not a map available in the peri-urban area. It was therefore not possible to perform this manually method in the peri-urban area.

Different locations
The objective was to measure the coordinates that represent the middle of the houses. Measuring the position inside a house normally is less accurate due to the obstruction of the satellite signals by the roof of the house. All three devices were measuring the position while staying inside the house itself and outside. The outside measurements were taken 5 meters from the middle of the house in front of the main door. This normally resulted in a distance of 2 meters from the main door. For the outside measurements, the direction and distance to the middle of the house was recorded. During the preprocessing phase of the data analysis, the outside measurements were corrected using the recorded direction and distance to the middle of the house.

True locations
To validate the measured positions of houses, a reference dataset is needed which contains the real positions of the houses. Since there is only one dataset of this, and this is the one to be validated, another reference dataset is needed. A dataset of 'true' locations was produced for this experiment for all the houses that were sampled. This was done by using the QuickBird image available for this project. The sampled houses were searched for on this map and manually the positions were drawn on it. To know which house exactly was sampled, pictures were made in the field of the sampled houses so these could be relocated on the QuickBird image.

Original houses dataset
The dataset with the positions of all houses on Rusinga Island, that is being used by all disciplines in the SolarMal project was also included in this experiment to validate this dataset. This dataset was created by fieldworkers that visit all houses on Rusinga Island in the beginning of the project. The protocol was to stand as close as possible to the main door, while standing outside without a roof above you. These measurements done two years ago are called here Black Tablet Original.

Table 15: Combinations of house type, devices, measures and location for experiment 2

Experiment 2 - measurement combinations			
Housetype	Device	Measure	Location
Both	PC	manually	Inside
Rural	Black Tablet	manually	Inside
Rural	Black Tablet	WP	Inside
Rural	Black Tablet	WP	Outside
Rural	Black Tablet (original)	WP	front door
Rural	Garmin Device	AWP	Inside
Rural	Garmin Device	AWP	Outside
Rural	Garmin Device	WP	Inside
Rural	Garmin Device	WP	Outside
Rural	White Tablet	manually	Inside
Rural	White Tablet	WP	Inside
Rural	White Tablet	WP	Outside
Barrack	Black Tablet	WP	Inside
Barrack	Black Tablet	WP	Outside
Barrack	Black Tablet (original)	WP	front door
Barrack	Garmin Device	WP	Inside
Barrack	Garmin Device	WP	Outside
Barrack	White Tablet	WP	Inside
Barrack	White Tablet	WP	Outside

3.2.2.Results

In total 59 houses were sampled during this experiment. Out of these 59 houses, 12 houses were left out of the data analysis since they represented locations far away from the real position. The cause of these errors will be discussed in 3.2.3.

Table 16 shows the summarizing statistics of the data analysis. It shows for each combination of measuring the average distance to the 'true' locations and the standard deviation in this. What strikes first is that the smallest distance is found for manually drawing the position, while in the field on the black tablet. This is only found for the black tablet and not for the white tablet. Overall, the use of the Garmin device resulted in the smallest distances. It strikes that measuring an AWP is not explicitly better than measuring a simple WP. An AWP took several minutes per house, while a WP takes a few seconds to create. Looking to the difference between measuring outside or inside the houses, it appeared that overall the outside measurement have a smaller distance than the inside one. The largest distances are overall found for the Black Tablet Original measurements. The measurements in the rural area have overall a smaller distance than the ones in peri-urban area. Last, it strikes that in the rural area, the black tablet was more accurate than the white tablet, while this is the opposite in the peri-urban area.

Table 16: Summary statistics of error distances, experiment 2

Experiment 2					Distance to 'true' location (m)			
HouseType	Device	Measure	Location	Houses	mean	stdev	min	max
Both	PC	manually	Inside	47	0,0	0,0	0,0	0,0
Rural	Black Tablet	manually	Inside	17	5,1	8,5	0,0	27,7
Rural	Black Tablet	WP	Inside	17	6,7	7,9	1,1	27,1
Rural	Black Tablet	WP	Outside	17	6,2	7,4	0,9	27,4
Rural	Black Tablet (original)	WP	front door	17	10,1	6,6	3,7	27,9
Rural	Garmin Device	AWP	Inside	17	6,2	6,5	1,8	24,3
Rural	Garmin Device	AWP	Outside	17	5,6	8,2	0,8	28,2
Rural	Garmin Device	WP	Inside	17	6,4	8,1	0,8	26,9
Rural	Garmin Device	WP	Outside	17	5,5	7,8	1,4	28,0
Rural	White Tablet	manually	Inside	17	12,8	11,4	0,8	42,6
Rural	White Tablet	WP	Inside	17	8,6	9,0	1,1	32,6
Rural	White Tablet	WP	Outside	17	8,5	9,7	0,7	29,8
Barrack	Black Tablet	WP	Inside	30	8,3	5,7	1,9	21,6
Barrack	Black Tablet	WP	Outside	30	7,4	4,4	1,0	16,8
Barrack	Black Tablet (original)	WP	front door	30	11,8	6,2	2,3	26,9
Barrack	Garmin Device	WP	Inside	30	6,7	5,7	0,8	22,2
Barrack	Garmin Device	WP	Outside	30	6,2	5,2	0,3	19,5
Barrack	White Tablet	WP	Inside	30	7,3	5,4	1,2	21,5
Barrack	White Tablet	WP	Outside	30	7,5	4,5	1,0	19,7

To check whether the QuickBird image has a certain displacement relative to the measured positions, the direction of the error distances are studied. Figure 24 shows a plot of the error distances of all individual measurements and averaged error distances per measurement type. Around the middle of this figure, the zero-zero position, a cloud of points can be seen. It appeared that the average error distance in the x-direction is -0,56 meter and in the y-direction -0,96 meter.

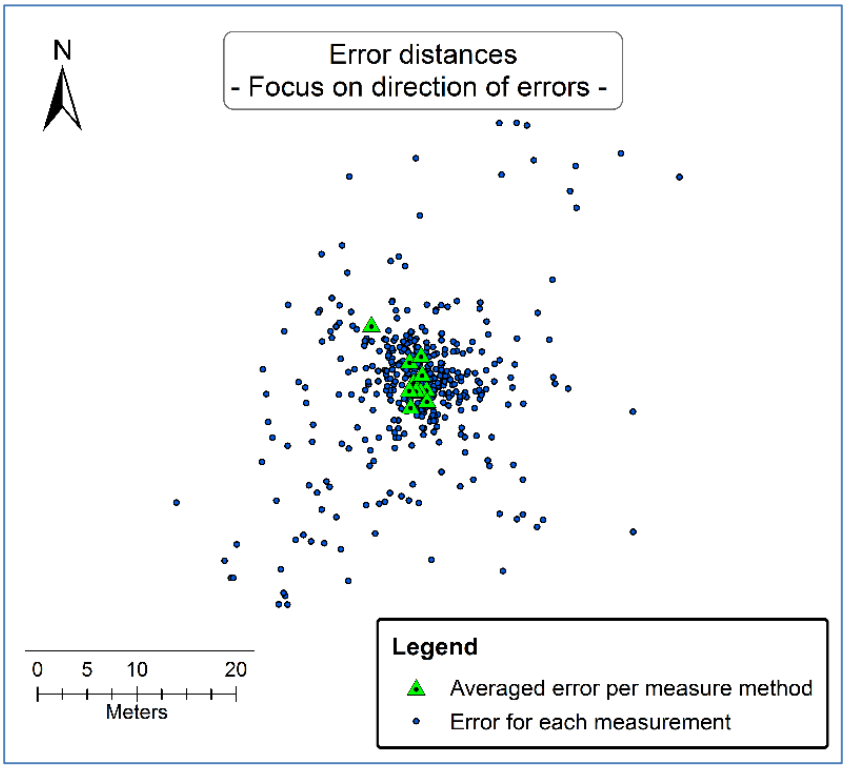


Figure 24: Error distances and their direction, experiment 2

Figure 25: Barrack area where satellite signals are easily blocked

3.2.3.Conclusion and discussion

Out of the 59 houses sampled, 12 had to be removed from the data analyses since three typical errors occurred. First of all, the tablets lost satellite signal while standing in a typical barrack area. Figure 25 shows the situation where actually 7 houses are present and were sampled. During the data analysis it appeared that the measured coordinates are representing the SolarMal office on the icipe terrain, more than eight kilometers away from the real position. It appeared that the tablets lost signal that moment and provided the position of probably its home or standard location. It is therefore important that if one would like to measure its position within such an area, that he ensures that the device has a contact with satellites. Secondly, it appeared that according to the Black Tablet Original dataset the positions of three separate houses were all three at the real position of the house in the middle. Probably a fieldworker forgot to measure the positions of these houses, did only go back to the middle house and measured three times the location as they were really there. In this case a human error is involved, which apparently also has to be taken seriously. Third, one of the Black Tablet Original positions was somewhere three kilometers away from the real position. The cause of this error is not really clear, but it could be an error in coding the houses. These three errors together lead to 12 houses that were finally not selected in the data analysis for the accuracy validation. The results of the data analysis tell something about the accuracy of a measured point when there was not an error involved of these kinds.

From this experiment it appeared that measuring a position by an AWP is not explicitly better compared to a WP. This strikes since it was expected to be more accurate. It could be that the AWP measurements were less accurate in this case, since they were lay down on the ground surface during the measurement. The other measurements were done at a height of 1,5-2 meters above the ground surface, while holding in a hand. Based on these results, a WP would have the preference over an AWP since measuring a simple WP cost less time.

The Garmin device had overall the smallest distances (+- 6 meters), however the distances of the two tablets are close to this (6-8 meters). For this project it would be sufficient to make use of the tablets and purchasing Garmin devices for positioning houses is not necessary. Especially for the rural area where the distance between two individual houses is at least 10 meters. For the peri-urban area the accuracy must be as high as possible to be able to locate each individual house, however accuracies of 6 meter or 8 meter are of the same order. Within a distance of 8 meter already 5 houses can be found, also within a distance of 6 meters the same 5 houses can be found.

The measurements in the rural area were more accurate compared with the measurements done in the peri-urban area. This has to do with a better satellite signal in the rural open field. It was expected that the accuracy would be less in the peri-urban environment since satellite signals are getting blocked by the density of metal roofs and objects.

The outside measurement including the correction to the middle of the house is overall the best measure in terms of accuracy. However, it only differs maximally 1 meter with the measurements done inside. Advantage of the inside measurement is that no correction is needed afterwards and less human mistakes can be expected. However, one has to be sure that there is a satellite signal while measuring inside houses, otherwise the measured positions will be completely wrong and has to be ignored like it was the situation in Figure 25. It was expected that the corrected outside measurements would be more accurate than the inside measurement, mainly due to an open sky situation for more accurately measuring the position. The direction and distance for correction also contains some inaccuracy, however it appeared that in total this inaccuracy is less than the inaccuracy caused by obstructing roofs.

It strikes that the manually method of positioning resulted in the highest accuracy with the Black Tablet and in the lowest accuracy with the White Tablet. This method has apparently a high potential looking to the results of the Black Tablet. However, there were probably some errors made by the fieldworker handling the White Table. This indicates that this manual method is also highly sensitive for mistakes by the fieldworker.

It seems that, for measuring waypoints, the Black Tablet is more accurate in the rural area than the White Tablet, while this is the opposite for the peri-urban area. It is not clear why this appears, however this must be studied again in the next experiment, where more houses will be sampled.

The measurements from the Black Tablet Original show overall the lowest accuracy. However there is a correction needed for this, since these measurements were done in front of the main door, which is different from this experiment were the focus was on the middle of the houses. The distance from the main door to the middle of the house will generally be around 3-5 meters, meaning that there is already a certain displacement between this dataset and the ‘true’ locations dataset. If this is taken into account, the distances of the original

dataset decreases to approximately 7 meters on average, which is equal to the measurements done by the two tablets.

For this experiment it was assumed that the dataset of drawn ‘true’ locations are the real positions of the sampled houses. Creating and using the drawn ‘true’ location dataset included some inaccuracy of the data analysis itself. First of all, the use is made of the QuickBird image, which already has some inaccuracy if it is compared to other map providers (see Figure 21). Secondly, some sampled houses were not visible on the QuickBird imagery, since these houses were recently build and the QuickBird image was taken before these houses were build. In such a case the location of the house was determined using surrounding objects that were recognizable. Third, within the barracks areas the houses are built so close to each other that it was tough to locate the right sampled house. However, the pictures taken in the field made it feasible to do this.

An average displacement of almost 1 meter in the y-direction would suggest to correct the QuickBird image for a standard displacement. However, looking to the well distributed point cloud in Figure 24, there seems not to be a clear direction of the displacement and the average displacement could just be the result of random error (Olynik et al. 2002). Based on Figure 24, there is no reason to correct the QuickBird image for a standard displacement.

3.3. Second validation of most suitable methods

The third experiment concerns validating the most suitable methods resulting from experiment 2. Based on field knowledge and the results of experiment 2, measuring waypoints inside and outside houses are the most feasible methods, using tablets. This experiment had the objective to validate the performance, in terms of accuracy, of these most feasible methods.

3.3.1. Methodology

The position of houses was measured by taking waypoints inside and outside houses with the two tablets. It was made sure that within the set of sampled houses both types of houses occurred, both in a rural area and a peri-urban area since these are the two typical situations on Rusinga Island. Like is done in experiment 2, a dataset of 'true' locations was created as reference positions. The Black Tablet Original dataset was again included to compare these results with the new measurements.

3.3.2. Results

A 100 houses were measured, of which 57 rural houses and 43 peri-urban houses. Unfortunately, 4 measurements had to be removed from the dataset since they represented locations far away from the real position. The cause of this will be discussed in 0.

Table 17 shows the summary statistics of the data analysis. What strikes first is that the smallest distances are found for the outside measurements, with distances of 6-7 meters. The inside measurements had on average a slightly larger distance than the outside measurements with distances of 7-14 meters. It appears again that the measurements done in the peri-urban area have larger distances than the ones in rural area. The Black Tablet Original dataset shows the same difference between the two types of houses. The Black Tablet Original shows overall the largest distances.

Looking to the size of the standard deviations in the error distances, the same structure is found as when looking to the average error distances. This means that the average accuracy of for instance the White Tablet outside, which is 6,2 meter, is not only on average so small. The accuracy of the White Tablet outside does not vary so much (3,6 meter), meaning that 97,5% of the measurements had an error distance less than 13,4 meter. Comparing this with the inside measurements of the same White Tablet, there is already an average distance of 13,9 meters, with a standard variation of another 15,3 meters. This finally results in that 97,5% of these measurements had an error distance smaller than 44,5 meter, which is more than three times the distance of the outside measurements.

To check again whether there is a specific directional displacement of the QuickBird image relative to the measured positions, the direction of the error distances was also studied for experiment 3. There was again a well distributed cloud of points around the center point. It appeared that for this dataset the average error distance in the x-direction was +1,45 meter and in the y-direction -0,15 meter.

Table 17: Summary statistics of error distances, experiment 3

Experiment 3					Distance to 'true' location (m)			
HouseType	Device	Measure	Location	Houses	mean	stdev	min	max
Both	PC	manually	Inside	96	0,0	0,0	0,0	0,0
Barrack	Black Tablet	WP	Inside	42	11,1	8,8	0,8	45,2
Barrack	Black Tablet	WP	Outside	42	7,1	5,3	0,6	24,6
Barrack	Black Tablet Original	WP	Front door	42	10,5	8,5	0,9	42,5
Barrack	White Tablet	WP	Inside	42	13,9	15,3	1,2	86,8
Barrack	White Tablet	WP	Outside	42	6,2	3,6	0,4	17,0
Rural	Black Tablet	WP	Inside	54	8,5	7,5	1,1	33,5
Rural	Black Tablet	WP	Outside	54	6,0	3,8	0,7	19,0
Rural	Black Tablet Original	WP	Front door	54	8,2	5,9	1,9	26,6
Rural	White Tablet	WP	Inside	54	7,1	5,1	1,4	27,1
Rural	White Tablet	WP	Outside	54	6,4	4,1	0,7	21,3

3.3.3.Conclusions and discussion

Out of the 100 houses, 4 had to be removed from the dataset since two typical errors occurred. First, like happened during experiment 2, two measured positions were done without satellite signal and the coordinates represent again the standard or home position in Mbita. The second error is that two positions from the Black Tablet Original dataset were so far away (> 300 meters) from the real position that there is probably again a problem with the coding of the houses or something else. These 4 houses are removed from the data analysis to be able to study the accuracy without such errors.

It appeared that outside measurements have the highest accuracy of 6 to 7 meters. The inside measurements do show a larger distance to this, which differs in this experiment more(1-6 meter) than in experiment 2 (1 meter maximal). Comparing the Black Tablet Original dataset with the outside measurements, there is a difference of 2-4 meter. This could be explained by the fact that the Black Tablet Original dataset was focusing on the front door and not at the middle of the house, so there is already a certain displacement of 3-5 meters. Taking this into account, one could conclude that the accuracy of the Black Tablet Original is of same accuracy as the outside measurements. This make sense since both methods measure the position outside, ensuring a good satellite signal. Overall it can be concluded that the accuracy of the Black Tablet Original is of equal accuracy of the measurements outside. The difference in accuracy between the White Tablet and Black Tablet does not seem to be clear, rather randomly.

The average directional displacements in this experiment (x- and y-direction respectively +1,45 and -0,15 meter) is not at all equal to the ones found in experiment 2 (x- and y-direction respectively -0,56 and -0,96 meter). It can therefore be concluded that there is no clear displacement of the QuickBird image relative to the measured positions.

4. Summary conclusion of experiments

Looking at the first experiment, it became clear that the best accuracy possible, when measuring just one time the position, can be less than 10 meters, however is at least 6 meters. According to the second and third experiment it became clear that the already existing dataset of houses has on average an accuracy of 8-11 meters. The standard variation in this is 6-8 meters, so 97,5% of all positions have an accuracy less than 25 meters. Accounting for the difference in measuring the location relative to the house, a correction is needed for this Black Tablet Original dataset to the middle of the houses. Accounting for this, 3-5 meters decrease in accuracy would be the result, ending in an average accuracy of 5-8 meters and with still an standard variation of 6-8 meters. 97,5% of all positions would then have an accuracy of less than 22 meters. One of the most suitable methods validated during the third experiment was measuring outside the house and correct for the distance and direction to the middle of the house afterwards. On average this method resulted in an accuracy of 6-7 meters, with a standard deviation of 4-5 meters. This means that 97,5% of all positions measured in this way have an accuracy less than 15 meters.

5. Recommendation for the SolarMal Project

Based on the experiments done and the knowledge acquired from the field it is recommended to redo exactly the same measurements as previous (the Black Tablet Original). While doing this, the fieldworkers have to ensure that they have an active satellite signal while taking the WP. With a second version of this dataset it would be possible to easily select measurements that need more attention by comparing one dataset with the other dataset. It is recommended to keep the methodology of standing in front of the main door, just outside of the roof so a good receiving capacity of satellite signals is available. If the same methodology would be followed as with the Black Tablet Original dataset, the two datasets would also be comparable with each other since both datasets represent the exact same location in relation to the house.

Another option is to go in the field with a laptop or one of the tablets and locate the houses manually on the map. If done properly, this will result in the highest accuracy possible. However this requires some skills in recognizing features from the environment on a map. If this would be the wished methodology, it is the best to train one or two fieldworkers in manually locating houses, so the least as possible human mistakes will occur. However, since this method is highly sensitive for human mistakes, it is rather recommended for this project to locate houses by measuring a simple WP using the tablets available for this project.

An advantage of the recommended method is that the two datasets could be used to check each other data. If for example one house has two measured positions which are more than 50 meters away from each other, then there is probable something wrong with one of the two measured positions. Having the two datasets it is possible to automatically select out the houses that need more attention. Such a way there can be a control system to ensure that no extreme errors appear in the dataset.

6. References

Garmin. 2009. Waypoint Averaging.

Garmin. 2011. Garmin eTrex and GLONASS: A powerful combination.

Garmin. 2014. What is WAAS?

Olynik, M., M. Petovello, M. Cannon, and G. Lachapelle. 2002. Temporal Variability of GPS Error Sources and Their Effect on Relative Positioning Accuracy. Proceedings of the Institute of Navigation NTM 2002.

Appendix H

Influence of the WAAS/EGNOS option (Garmin device) on the accuracy of measurements

This small experiment is done to check whether the WAAS/EGNOS option on the Garmin device is important in determining the accuracy of measurements.

Data Collection

The Garmin device was installed at a fixed place, which was equal to the location of the first experiment described in Appendix F. It was installed there on two consecutive days, measuring its position for both days at the same time. It measured its position every 10 seconds for a duration of 3 hours (13:00 – 16:00 local time, 24th and 26th of May, 2014). The first day the WAAS/EGNOS option on the Garmin device was turned ON. The second day the WAAS/EGNOS option on the Garmin device was turned OFF.

Data Analysis Methodology

The average point position from the cloud of points was calculated, after which an average position error for all points together relative to the average point position was calculated. This gives an impression of the precision of the measured points. The two average point positions from both data series were compared with each other, as to the 'true' positions according to several map providers. See Figure 21 and Table 13 for these positions. Comparing the average point positions with the 'true' locations gives an impression of the accuracy of the measurement.

Results and Conclusion

It appeared that the standard distance of the cloud of points, with the WAAS/EGNOS option on the Garmin device turned on was 1,72 meter. With a 95% confidence level this result in a precision of the measurement of 3,4 meter. With the WAAS/EGNOS option turned off the standard distance was 2,21 meter, resulting in a precision of the measurement of 4,4 meter (95% confidence). The average position of the two cloud of points were 4.79 meters away from each other. Why this differs almost 5 meters is not clear. Based on this experiment it can be concluded that using the WAAS/EGNOS option as turned ON results at least in a higher precision of the position.

Appendix G

Table Of Contents: Zip-file of all data, analysis and results

1. Report (both in Word and PDF format)

2. Presentations
   2.1. Midterm presentation
   2.2. Final presentation

3. Literature
   3.1. Excel file: Literature Search Review (Keywords and review results)
   3.2. EndNote file: All citations used in EndNote format
   3.3. Folder: PDF's of all papers referenced and used for inspiration

4. Data
   4.1. Folder: ASTER GDEM (contains a zip-file of the data as originally received)
   4.2. Folder: QuickBird Image (contains a zip-file of the data as originally received)
   4.3. Excel file: Baseline_individuals (The results of the broad survey over all the individuals)
   4.4. Excel file: entomology survey Sep 2012-Aug 2013 (the mosquito dataset)
   4.5. Excel file: Householdinfo_outcomes (the results of the broad survey over all the households)

5. Analysis
   5.1. R
      5.1.1. Folder: Scripts
      5.1.2. Folder: Input data used
      5.1.3. Folder: Output data and results
   5.2. ArcGIS
      5.2.1. Folder: GeoDataBases (just all data worked with within ArcGIS)
      5.2.2. Toolbox: MSc_CV_Rusinga (ArcGIS toolbox with models created for these analysis)
      5.2.3. Mxd-file: Rusinga

6. Results
   6.1. Folder: EnvVars (.png files of all the created environmental variables presented in Appendix)
   6.2. Folder: Histograms (.png files of the histograms presented in Appendix)
   6.3. Folder: ViolinPlots (.png files of the violin plots presented in Appendix)
   6.4. Folder: SpatialDistributionMaps (.png files of the spatial distribution maps presented in Appendix and Appendix)
   6.5. Excel file: SummaryResultsAnalysis (Tables and summaries used for reporting)

7. Extra Study in Kenya(Work folder of extra study done in Kenya to the accuracy of house positions, Geodatabases and mxd files can be found in 5.2)
   7.1. Folder: Experiments (input data, results etc per experiment done)
   7.2. Others: Some GPX files recorded during the stay in Kenya

Terminology list

List of used terminology

DEM	A Digital Elevation Model is a dataset or image with values of the elevation of a certain area.
Environmental variables	Anything that tells something about the environment, such as the height of an area, the slope, vegetation indices etc. In this study, also the population density and house hold density are seen in this thesis as environmental variables.
Multispectral image	An image that contains per pixel the intensity of several wavelengths, also called the bands. Commonly the blue, green and red light is combined to one image, a picture people are used to see it. A multispectral image provides the different wavelengths as different layers so researchers can use this.
nm	Abbreviation for nanometer, or 10^{-9} meter
Orthorectification	A satellite image of the earth is just a two-dimensional representation of what one sees from one point in time and space. Due to terrain characteristics like ridges and hills, this representation can be incorrect. A location on the image could be in reality be further from a ridge or something, which is called terrain displacement. Orthorectification involves the correction for these terrain displacements and corrects the image using an elevation model.
Panchromatic band	A panchromatic band is an image that represents a scene as it appears to the human eye. It combines the intensity of all wavelengths into one layer. This is also seen as the black-and-white representation of what human eye normally sees.
Pan sharpening	Panchromatic images generally have higher resolutions than multispectral images. Using a 1-meter resolution panchromatic image one can create a 1-meter resolution multispectral image, which originally had a lower resolution. This process of interpolating is called ‘pan sharpening’.
Spectral band	A spectral band is a range of wavelengths for which the intensity is measured, and saved in a multispectral image. A spectral image can contains more than one band.
Vector species	Mosquitoes are vectors of the malaria parasite since the mosquitoes are the transmitters of this parasite from human host to another.
Waypoint (WP)	A waypoint is a set of coordinates in the two- or three-dimensional space which actually are the distances to a certain reference position. A waypoint is normally used for navigation purposes.