

---

# **INTEGRATING MACHINE LEARNING WITH GEOSTATISTICAL TECHNIQUES IN MODELING THE PREVALENCE OF PLASMODIUM FALCIPARUM**

**by**

**NTURIBI GACHERI LYDIAH**

*Submitted on 1/17/2022 in partial fulfilment of the requirements for the award of the  
degree of Bachelor of Science in Geomatic Engineering.*



**Department of Geomatic  
Engineering and Geospatial  
Information Systems (GEGIS)**

---

---

## DECLARATION

*I declare that this project is my own work and has not been submitted by anybody else in any other university for the award of any degree to the best of my knowledge.*

Sign.....

Date 12<sup>th</sup> January 2022

Student name                NTURIBI GACHERI LYDIAH

Registration number      ENC221-0296/2016

Department of Geomatic Engineering and Geospatial Information Systems (GEGIS)

Jomo Kenyatta University of Agriculture and Technology

## CERTIFICATION

This project has been submitted for examination with my approval as the candidate's supervisor.

Sign.....

Date 11<sup>th</sup> January 2022

Dr -Ing. MISK Benson Kenduiywo

Lecturer

GEGIS Department

©GEGIS 2022

---

## **Acknowledgements**

First, I would like to thank God, for letting me through all the difficulties. I have experienced your guidance day by day. You are the one who let me finish my degree. I will keep on trusting you for my future.

I would like to acknowledge and give my warmest thanks to my supervisor Dr. - Ing. MISK Benson Kenduiywo who made this work possible. His guidance, criticism and advice carried me through all the stages of writing my project. He recognized and tapped my potential, challenging me to achieve more. To him, I am forever indebted.

I would also like to give special thanks to my parents and my family (Nturibis) for their continuous support and encouragement while undertaking my research and writing my project. Your prayers for me was what sustained me this far. Dad, mum, we made it to the end. Special thanks to my friend Rik for his input in this project. You all are my heroes.



---

## Abstract

Malaria is primarily caused by the *Plasmodium falciparum* parasite transmitted by the female anopheles mosquito. Many control measures that have been put in place to fight malaria have resulted in its significant decrease. Despite the decrease, malaria remains one of the leading causes of death in children below the age of five. The efforts to eradicate malaria continue to bear little success, hence the need to adopt new methods for identifying areas prone to malaria to form a basis for predicting future trends based on climatic, socio-economic and environmental factors. Existing studies employ either Machine Learning or Geostatistical Methods for modelling and predicting malaria prevalence. However, since more studies continue to show the success of integrating both Machine Learning and Geostatistics, this research focuses on developing hybrid models for modelling and predicting malaria prevalence in Kenya. It also seeks to identify the driving factors for *Plasmodium falciparum*. The Random Forest and Extreme Gradient Boosting Machine Learning algorithms and their regression kriging hybrids were developed. The hybrids performed better than the Machine Learning or geostatistics models alone, drawing the conclusion that integrated Machine Learning and Geostatistics can be used in modelling and predicting *Plasmodium falciparum* prevalence given a set of climatic, environmental and socio-economic factors.



---

## Table of contents

Acknowledgements.....	III
Abstract.....	V
Table of contents .....	VII
List of figures .....	XI
List of tables.....	XII
CHAPTER ONE .....	- 1 -
1 Introduction .....	- 1 -
1.1 Background .....	- 1 -
1.2 Motivation and problem statement .....	- 3 -
1.3 Research identification .....	- 4 -
1.3.1 Research objectives.....	- 4 -
1.3.2 Research questions .....	- 5 -
1.4 Significance of the Research.....	- 5 -
1.5 Study outline .....	- 6 -
CHAPTER TWO .....	- 9 -
2 Literature review .....	- 9 -
2.1 Malaria Background .....	- 9 -
2.2 Test for Correlation and Statistical Significance.....	- 12 -
2.3 Machine Learning Models .....	- 13 -
2.3.1 Random Forest .....	- 14 -
2.3.2 Extreme Gradient Boosting (XGBoost) .....	- 16 -
2.3.3 Multiple Linear Regression .....	- 17 -
2.4 Geostatistical Methods.....	- 17 -
2.4.1 Regression Kriging .....	- 18 -
2.5 Validation and Comparison of the Models.....	- 19 -

2.6	Gaps in the research .....	- 20 -
CHAPTER THREE .....		23
3	Materials and methods.....	23
3.1	Study area .....	23
3.2	Data.....	24
3.2.1	Malaria Prevalence Point Data .....	25
3.2.2	Rainfall Data .....	- 28 -
3.2.3	Land Surface Temperature.....	- 28 -
3.2.4	Enhanced Vegetation Index .....	- 28 -
3.2.5	Proximity to Water Bodies .....	- 29 -
3.2.6	Population Density .....	- 29 -
3.2.7	Global Human Footprint (GHF) .....	- 29 -
3.2.8	Digital Elevation Model (DEM) .....	- 29 -
3.3	Methods.....	- 29 -
3.3.1	Introduction .....	- 29 -
3.3.2	Flow Diagram.....	- 32 -
3.3.3	Correlation of Independent Variables.....	- 32 -
3.3.4	Identification of Statistically Significant Factors.....	- 34 -
3.3.5	Machine Learning Models.....	- 35 -
3.3.6	Geostatistical Methods .....	- 35 -
3.3.7	Model Validation .....	- 36 -
CHAPTER FOUR.....		39
4	Results.....	39
4.1	Exploratory Data Analysis .....	39
4.2	Correlation and Statistically Significant Factors.....	39
4.3	Distribution of PfPR Prevalence Points .....	- 49 -



---

4.4	Prediction of PfPR Prevalence.....	- 51 -
4.4.1	Using Machine Learning Models .....	- 51 -
4.5	Models Validation and Comparison .....	- 58 -
CHAPTER FIVE .....		- 59 -
5	Discussion .....	- 59 -
CHAPTER SIX .....		- 61 -
6	Conclusion and outlook .....	- 61 -
6.1	Conclusion.....	- 61 -
6.2	Recommendations .....	- 61 -
References .....		- 63 -
Appendix .....		A



---

## List of figures

Figure 1: Study Area Map .....	23
Figure 2: KDHS Sample Points.....	- 27 -
Figure 3: Flow Diagram of the research.....	- 32 -
Figure 4: Correlation matrix showing correlation of independent variables .....	40
Figure 5: EVI of Kenya (2015).....	42
Figure 6: Global Human Footprint (2015) .....	43
Figure 7: ITN Coverage for Kenya (2015).....	44
Figure 8: Kenya's Mean Temperature (2015) .....	45
Figure 9: Proximity to Water Bodies .....	46
Figure 10: Rainfall (2015).....	47
Figure 11: DEM.....	48
Figure 12: Distribution of 2015 PfPR Data. Class 1 0.008 – 0.079, Class 2 0.079 – 0.147, Class 3 0.147 – 0.234, Class 4 0.234 – 0.477 .....	- 50 -
Figure 13: Random Forest PfPR Prediction .....	- 52 -
Figure 14: XGBoost PfPR Prediction .....	- 53 -
Figure 15: Regression Kriging PfPR Predictions.....	- 54 -
Figure 16: RFRK PfPR Predictions .....	- 56 -
Figure 17: XGBoostRK PfPR Predictions .....	- 57 -

---

## **List of tables**

Table 1: Data, sources and date of access .....	24
Table 2: Summary Statistics for the Dependent Variable.....	39
Table 3: Statistically Significant Factors .....	41
Table 4: PfPR Classification .....	- 49 -
Table 5: Variogram Parameters .....	- 55 -
Table 6: Models Validation .....	- 58 -

---

## **CHAPTER ONE**

### **1 Introduction**

#### **1.1 Background**

*Plasmodium falciparum* (PfPR) malaria remains a global burden that affects thousands of people on a worldwide scale. Malaria is primarily caused by the *Plasmodium falciparum* parasite transmitted by the female anopheles mosquito (Centers for Disease Control and Prevention, 2020). PfPR malaria has declined over the last years. The significant decline is attributed to improved healthcare, malaria control interventions and increased research studies on the parasite (Weiss, et al., 2019). Despite a steady increase in population, malaria prevalence decreases due to the measures formulated by major global organisations to control the disease. The decrease in the number of infections, although significant, still leaves thousands of people vulnerable to the disease. As a result, PfPR malaria still poses a health risk, especially in areas prone to the disease.

Malaria is responsible for thousands of deaths each year in Kenya. According to the Centre for Diseases and Control (CDC), there are about 3.5 million cases of malaria reported each year in Kenya, with more than 10 000 deaths. It is also one of the leading causes of mortality in children under the age of five. The disease is more prevalent in western Kenya, whose climatic and geographic characteristics make it more prone to the PfPR parasite (Centers for Disease Control and Prevention, 2018). Like most sub-Saharan African countries, Kenya is characterised by a tropical climate with warm temperatures, high rainfall, and humid conditions (World Health Organization, 2021). Countries within the tropics are more likely to harbour the PfPR

---

due to the favourable conditions. Although Kenya does not have the highest number of deaths from malaria globally, the number of deaths experienced are still high enough to warrant measures for preventing the prevalence of the disease.

The CDC's collaboration with the Kenya Medical Research Institute (KEMRI) aims to combat malaria in Kenya. The significant operations put in place to curb the spread of the disease include "capacity building and technical support, surveillance, monitoring and evaluation, prevention, case management, transmission reduction research and laboratory (Centers for Disease Control and Prevention, 2018)". Established over thirty years ago, the collaboration between the two organisations was aimed at prevention other than treatment. The activities and operations of the collaboration have borne fruits characterised by a reduction in the number of malaria cases from 11% in 2010 to 8% in 2015 (Bashir, Nyakoe, & Sande, 2019). Despite the progress made in eliminating malaria, high endemic zones such as Western Kenya continue to register a high incidence rate, with malaria accounting for 25 to 35% of outpatient consultations, 20 to 45% of hospital admissions and 15 to 35% of hospital deaths (Kapesa, et al., 2018). Langat (2019) confirms the challenge at hand through the evidence of stagnation in the fight against malaria.

The recent rise in malaria cases throughout the country is attributed to mosquitoes' resistance to pyrethroid-based insecticides and relaxed prevention measures due to reduced funding from global organisations. The fight against malaria has also faced numerous challenges that have resulted in slow progress in fighting the disease. Some of the challenges include weak health systems and the inability to respond to conditions that favour the breeding of the parasites. There are concerns that malaria could rebound as a fatal drug-resistant disease that would require even greater

---

efforts to eradicate (Langat, 2019). There is a need to identify areas prone to malaria using predictive models to formulate preventative and control measures.

## **1.2 Motivation and problem statement**

The challenges faced in eradicating malaria within the country jeopardise the global efforts to eliminate malaria entirely by 2030. This is in accordance with the third Sustainable Developmental Goal (SDG), whose subsections two and three intend to end infectious diseases that result in premature mortalities. SDG 3.2 states, *"By 2030, end preventable deaths of new-borns and children under five years of age, with all countries aiming to reduce neonatal mortality to at least as low as 12 per 1,000 live births and under-5 mortality to at least as low as 25 per 1,000 live births"*. The 3.3 SDG states, *"By 2030, end the epidemics of AIDS, tuberculosis, malaria and neglected tropical diseases and combat hepatitis, water-borne diseases and other communicable diseases"* (United Nations, p. 3). The decline in the success in fighting malaria is enough to warrant further studies for identifying areas prone to malaria and form a basis for predicting future trends based on the existing climatic and geographic factors.

Numerous studies to estimate the prevalence of PfPR have been conducted (Kapesa, et al., 2018; Nkiruka, Prasad, & Clement, 2021; Macharia, et al., 2018). These studies are all aimed at identifying high-risk areas for planning and prevention. Most of them, conducted in Kenya and sub-Saharan Africa, focus on the whole country or specific regions. They employ either Machine Learning (ML) or Geostatistical (GS) Models for tackling the problem at hand. Macharia et al. (2018) used a geostatistical model to predict annual malaria risk for children from 1990 to 2015. Nkiruka, Prasad, & Clement (2021)'s study employs machine learning models to predict malaria

---

outbreaks based on climatic factors. The Machine Learning models used in the studies are XGBoost, Support Vector Machine (SVM), Naïve Bayes and Linear Regression (LR) (Nkiruka, Prasad, & Clement, 2021). The researchers perform a comparative analysis to determine the best-performing model for predicting malaria prevalence.

ML and GS models all have high accuracy levels, as presented in Macharia et al. (2018) and Nkiruka, Prasad, & Clement, (2021) papers. With such levels of accuracy, this sparks interest in conducting a comparative study to identify the performance of both techniques in a common study and integrate both methods for predicting PfPR prevalence in Kenya. Instead of focusing on a specific area, there is a need to conduct the study for the whole country since climatic and geographic data is available and also because of biases that result in negative results due to lack of resources in specific areas for testing and reporting PfPR cases (Langat, 2019). The study will help highlight malaria-prone regions in Kenya and shed more light on the performance of ML and GS models in predicting malaria prevalence.

### **1.3 Research identification**

The main aim of this research is to integrate Random Forest (RF) and Extreme Gradient Boosting (XGBoost) machine learning with Regression Kriging (RK) geostatistical algorithms in modelling and predicting the prevalence of Plasmodium falciparum in children aged 2-10 years from 2005 to 2015 in Kenya.

#### **1.3.1 Research objectives**

- i. To correlate and identify statistically significant driving factors influencing the prevalence of plasmodium falciparum.



- 
- ii. To develop and compare RF, XGBoost and RK models for predicting the prevalence of plasmodium falciparum.
  - iii. To integrate the ML models with OK to generate their RK hybrids.
  - iv. To validate and compare the models.

### **1.3.2 Research questions**

The limitations highlighted in the problem statement section prompted the following questions, which were the basis for the formation of the objectives listed above:

- i. What are the driving factors that influence the prevalence of PfPR? What is the significance of each driving factor?
- ii. What machine learning and geostatistical algorithms can be used in predicting the prevalence of Plasmodium falciparum?
- iii. What is the possible outcome of integrating the best performing Machine Learning Model with a Geostatistical Model in carrying out the prediction?
- iv. Which model is the overall best in predicting PfPR prevalence?

## **1.4 Significance of the Research**

Understanding the prevalence of malaria will form a basis for formulating better preventative policies for mitigating the disease. Environmental, climatic and socioeconomic factors are significant drivers that influence the prevalence of the disease. Understanding the significance of these factors will inform the training data to be used in training the models, but it will also help identify the impact of each factor in influencing malaria. This alone will allow policymakers and the national government to improve the identified factors in an informed manner. It will also allow


---

these parties to make informed decisions on matters relating to malaria eradication in Kenya.

Integrating ML and GS models in predicting malaria will help account for spatial autocorrelation and spatial heterogeneity in spatial data. Spatial autocorrelation, according to Zhang, Ma, & Guo (2009), "represents the correlations between the values of a random variable at "neighbouring" locations." In simpler terms, this means that close things are closely related, and vice versa. Spatial heterogeneity is "structural instability in the form of systematically varying model parameters or different response functions. (Zhang, Ma, & Guo, 2009, p. 533)" This basically refers to the variation of the level of influence of independent variables on the independent variable. Machine Learning caters for spatial autocorrelation effects while geostatistical models cater for the spatial heterogeneity effect. Therefore, integrating both models will help remove these spatial effects, resulting in better-fitted data and a more accurate prediction of the dependent variable.

## **1.5 Study outline**

This thesis is organised into 8 chapters. Chapter 1, the introduction, provides the background of the study, the motivation and problem statement, the research objectives and the research questions. Chapter 2 is the literature review which reviews the findings from previous literature. Chapter 3 is the materials and methods section. This chapter discusses the datasets and methodology adopted in realising the research objectives. Chapter 4 discusses the results obtained while comparing them with those obtained in other similar studies. Chapter 5 comprises the discussion of the results obtained in Chapter 4. Chapter 6 focuses on the conclusions drawn



---

from the study in reference to the results obtained. It also provides recommendations for future studies. Chapter 7 is the reference list used in the study.



---

## **CHAPTER TWO**

### **2 Literature review**

#### **2.1 Malaria Background**

Malaria is a communicable disease transmitted to people by female Anopheles mosquitoes. It is caused by Plasmodium parasites, which are spread to people through bites. The two parasite species that cause malaria are *P. falciparum* and *P. vivax*. *P. falciparum* accounts for the most malaria cases globally, with 99.7% of malaria cases in the African Regions in 2018 caused by the parasite. According to the WHO, malaria's symptoms include headaches, fevers and chills, whose onset is usually about 10 to 15 days after being bitten by an infected anopheles mosquito. In children, the symptoms are fatal. They include severe anaemia and respiratory distress and usually leads to cerebral malaria. In adults, the more severe symptom is a multiple-organ failure, resulting in death (World Health Organization, 2021). These symptoms show the fatality and danger that malaria poses to both children and adults.

Some people are more likely to contract malaria more than others. The high-risk groups are infants and children under five, pregnant women, patients living with HIV/AIDS, non-immune immigrants and mobile populations. Most malaria programmes within different countries target these particular groups, with policies to protect them from malaria infections. Globally, there has been a decline in the number of malaria infections, especially in recent years. According to the world malaria report, released in 2020, there were 229 million malaria cases in 2019, compared to 2018's 228 million cases. Additionally, the number of deaths in 2019 was 409000 in 2019 compared to 411000 deaths in 2018. Of these deaths, about

---

50% of them can be traced to the WHO African region, which is also home to more than 90% of all malaria cases at the global level. Children are more likely to die more than any other groups from the vulnerable groups discussed here. (World Health Organization, 2021) With the high death rates, it is imperative to perceive not only malaria as a fatal disease but also one that is more likely to frustrate the third SDG that aims at promoting well-being for all at all ages.

Owing to the large number of deaths from malaria experienced in the country, many measures and policies have since been put in place by international bodies to prevent the spread of the disease. More than 10,000 deaths occur each year in Kenya alone due to malaria (Centers for Disease Control and Prevention, 2018). The collaboration between CDC prevents malaria by providing resources and knowledge on tackling the endemic. This ranges from providing technical assistance to training scientists and clinicians on how to handle the disease. In addition, CDC and KEMRI conduct health facilities surveillance where the reported cases are recorded for monitoring purposes. Surveillance also involves testing community members in high-risk areas that are unlikely to visit hospitals due to the absence of symptoms (Centers for Disease Control and Prevention, 2018). The activities allow Kenya's Ministry of Health (MoH) to monitor the progress of the disease and gauge the success of the prevention and control measures put in place to control malaria. Other preventative measures include providing WHO with insecticide-treated mosquito nets (ITNs) and antimalarial drugs (World Health Organization, 2021).

Many researchers share concerns on malaria, especially in Kenya, where the infection rates in the Western and Coastal Regions remain high despite the many measures put in place by both the Kenyan Government and International organisations to

---

prevent and control the disease. Were et al. (2019) and Bashir, Nyakoe, & Sande (2019) highlight the high prevalence of malaria in Western Kenya, which can be attributed to proximity to Lake Victoria and the hot and wet climate experienced in the region. Additionally, this high prevalence and mortality rates can be attributed to socioeconomic factors, where people from underprivileged societies cannot afford healthcare, thus increasing their risk of succumbing to the disease. The lack of information on preventative measures also arrested the progress made in preventing and controlling the spread of malaria, especially in high-risk areas. Langat also highlights the burden of malaria in Baringo and Mombasa Counties, whose climatic, environmental and socioeconomic factors make them hotbeds for malaria (Langat, 2019). Going with the information presented in all these articles, it is without a doubt that malaria is a national burden that needs to be handled by identifying the areas likely to have the most favourable conditions that support the breeding of the anopheles mosquitoes, which are the primary carriers of the PfPR parasite.

Malaria prevalence can be linked to various predictors. The factors linked to PfPR prevalence in Weiss et al., (2019) article include geospatial environmental and socioeconomic data characterising mosquitoes and human habitats. Others highlighted in the article are coverage of ITNs, indoor spraying and the use of antimalarial drugs. Malaria prevalence can also be attributed to rainfall, temperature, presence or absence of vegetation, distances to waterbodies, health facilities, urban centres, elevation, land cover, humidity, demographic factors, time of the year, and malaria interventions in an area. The metrics that are used to measure these co-variates include but are not limited to rainfall/precipitation, Land Surface Temperature (LST), Weekly temperatures, Normalized Difference Vegetation Index

---

(NDVI), Enhanced Vegetation Index (EVI), relative humidity, wetness index and slope (Odhiambo, Kalinda, Macharia, Snow, & Sartorius, 2020). These indicators can be grouped into geographical, environmental, climatic, demographic/socioeconomic factors and are extensively used in building ML and GS models for predicting PfPR prevalence.

## **2.2 Test for Correlation and Statistical Significance**

Studies that use many co-variates in forecasting use different techniques to correlate the factors, such as Pearson's and Spearman's correlation analysis and Principal Component Analysis (PCA). Highly correlated factors are usually ignored because of their low impact to contribute to any significant changes in the models. For instance, in the (Nkiruka, Prasad, & Clement, 2021) study, Pearson's correlation analysis was conducted to determine the relationship between the independent and dependent variable, with the correlation coefficient ranging from -1 to +1. Negative numbers show a negative correlation, while positive numbers reveal a positive correlation between the feature and target variables. Usually, the researcher can choose to keep or discard some co-variates based on information acquired from secondary sources and the nature of their study, even when these drivers fail the collinearity test.

The measure of association is used to identify the statistically significant driving factors based on the p-values. This is usually achieved through linear regressions done at a defined confidence interval. All the factors that fall within the defined p-values are viewed as statistically significant, thus allowing the rejection of the null hypothesis (Kattan & Gerds, 2020). Factors that fall outside the defined p-value represent insignificant factors.



---

## 2.3 Machine Learning Models

Machine learning models are preferred for predictive studies because they can extract knowledge from data and identify patterns using both classification and regression. Various studies have employed different ML methods in the predictive analysis of PfPR, with the methodology based on the nature of data and the primary intent of the research. K-means clustering and Extreme Gradient Boosting (XGBoost) are the main ML models used in Nkiruka, Prasad, & Clement (2021) paper. The researcher cites the reasons for using K-means as its ability to identify outliers within the data. XGBoost is preferred due to its ability to speed up the classification process with high accuracy levels. The paper's main aim is to identify the most suitable ML model for the binary prediction of malaria outbreaks based on climatic factors. The other models considered in the paper are the auto-regressive integrated moving average (ARIMA) and seasonal ARIMA (SARIMA), both of which perform optimally with time-series data to predict the trend of malaria. Other models include VECTRI, SLIM, SINTEX-F2, SVM, Ensemble Learning and Artificial Neural Networks (ANN) (Nkiruka, Prasad, & Clement, 2021). Despite the success in most of these models, the researchers prefer using XGBoost as it had not been used in other studies and to test its performance against the other existing models.

Odhiambo et al.'s (2020) paper conducted a systematic review of papers, some of which use various ML models to forecast malaria risk at different temporal resolutions. Some studies employ the ARIMA model, others the Bayesian conditional autoregressive models, while others used the first-order autoregressive (AR) models. Linear models such as Poisson and logistic regression, ANNs and boosted regression trees were also used to analyse incidence patterns and examine malaria prevalence.

---

Similarly, (Harvey, Valkenburg, & Amara, 2021) study employs ML models such as naïve Bayes (Gaussian) and Random Forest (RF) regressor in predicting malaria epidemics in Burkina Faso. These models are selected due to their strength, making it possible to analyse data and make plausible conclusions based on the patterns and trends within the data.

The R-score, F-score, Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE) and Akaike Information Centre (AIC) statistics are used to validate the results from the models. The choice of performance metrics is dependent on the choice of ML model used (Steurer, Hill, & Pfeifer, 2021); hence there is a need to study the best performance metrics for use in different scenarios. ML models can comfortably be used for non-linear data, and it has capabilities to uncover patterns that would have otherwise been difficult to uncover through regular statistical analysis.

### **2.3.1 Random Forest**

Random forest is a machine learning technique that is used to solve both classification and regression problems. It utilises ensemble learning, a technique that combines many classifiers to provide solutions to complex problems. Random forest algorithms consist of many decision trees. The forest generated by the random forest algorithm is trained through bagging or bootstrap aggregating (Mbaabu, 2020). The limitations of decision tree algorithms are eradicated in Random Forest because it reduces the overfitting of datasets while increasing the precision of prediction. The algorithm generates predictions without requiring many configurations.

---

Each decision tree is constructed using a different bootstrap sample from the original data. About one-third of the cases are left out of the bootstrap sample and not used in the construction of the  $k^{\text{th}}$  tree. A test classification is obtained for each case in about one-third of the trees; hence there is no need for cross-validation or running separate tests to get unbiased estimates (Breiman & Cutler, n.d.).

Random Forest regression uses the mean squared error (MSE) to measure how the data branches from each node (Schott, 2019). The equation is expressed as:

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (f_i - y_i)^2 \quad (1)$$

Where  $N$  is the number of data points,  $f_i$  is the value returned by the model and  $y_i$  is the actual value for data point  $i$ .

The features of Random Forests, according to Breiman & Cutler (n.d.), are:

- RF algorithm is very accurate
- It can handle large datasets
- The algorithm can handle thousands of input variables without variable deletion.
- It gives estimates of what variables are important in the classification
- RF has an effective method for estimating missing data and maintains accuracy when a large proportion of data is missing.
- It has methods for balancing errors in class population unbalances datasets.
- The generated forests can be saved for use on other data.
- Prototypes are computed that give information about the relation between the variables and the classification.

- It computes proximities between pairs of cases that can be used in clustering, locating outliers, or giving interesting views of the data.
- RF offers an experimental method for detecting variable interactions.

### 2.3.2 Extreme Gradient Boosting (XGBoost)

Boosting simply means converting weak learners into strong learners. Gradient Boosting, just like Random Forest, uses decision trees where the machine learning model is developed upon repetitively asking questions to partition data and reach a solution. Gradient boosting involves subsampling the training dataset and training individual learners on random samples created by the subsampling. This action reduces the correlation between the results from individual learners and combines results with low correlation hence providing better results. The Gradient boosting algorithm finds optimal solutions to problems by finding the near-optimal solutions through the stochastic nature of random sampling (Gaurav, 2021). XGBoost is preferred by data scientists due to its high execution speed out of core computation (Osman, Ahmed, Chow, Huang, & El-Shafie, 2021). The goal of XGBoost is to find the optimised output value for the leaf to minimise the whole equation. It uses the loss function to build trees by minimising the following value:

$$\mathcal{L}(\Phi) = \sum_i l(y_i, A_i) + \sum_k \Omega(f_k) \quad (2)$$

Where  $l$  represents the loss function,  $\Omega$  is a measure of how complex the model is to assist in avoiding over-fitting of the model. It is calculated using the formula:

$$\Omega(f_k) = \gamma T + \frac{1}{2} \lambda \|w\|^2 \quad (3)$$

---

Where  $T$  represents the number of leaves of the tree, and  $w$  is the weight of each leaf.

### 2.3.3 Multiple Linear Regression

For multiple linear regression models, the response, or the independent variable, is influenced by more than one predictor variable. The independent variables are controlled by the researcher, although uncontrollable variables also play a key role in impacting the response variable. The linear model that relates the response,  $y$ , to several independent variables takes the form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon \quad (4)$$

Where parameters  $\beta_0, \beta_1, \dots, \beta_k$  are regression coefficients, and  $\varepsilon$  is a constant that provides for random variation in  $y$  that cannot be explained by the  $x$  variables. The random variation is partly due to other variables that affect  $y$  that are not known or remain unobserved (Rencher & Schaalje, 2008).

Similarly, according to Rencher and Schaalje (2008), multiple linear regression models can take the form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 \sin x_2 + \varepsilon \quad (5)$$

Where the model is linear in the  $\beta$  parameters and not necessarily linear in the  $x$  variables, multiple linear models provide a theoretical framework for understanding the  $y$  variable.

## 2.4 Geostatistical Methods

These provide an essential framework for performing interpolations and estimating spatial dependence from a given set of data (Odhiambo, Kalinda, Macharia, Snow, & Sartorius, 2020). In malaria studies, geostatistical methods are popular because they

account for data continuity uncertainties and provide high accuracy estimates and forecasts. The study by (Macharia, et al., 2018) uses Model-Based Geostatistics to conduct a spatiotemporal analysis of PfPR in Kenya. To test the validation of the results, mean error and mean absolute error (MAE) are used.

Bayesian Kriging and Geographical Weighted Regression (GWR) are some of the geostatistical models considered in the spatiotemporal modelling of malaria risk (Odhiambo, Kalinda, Macharia, Snow, & Sartorius, 2020) study. These models, just like in ML, help identify the patterns in malaria risks while catering for spatial heterogeneity. Geostatistical models such as regression kriging allow for the construction of prediction intervals and the mapping of prediction errors.

#### **2.4.1 Regression Kriging**

Regression kriging combines both non-geostatistical and geostatistical methods through ordinary kriging of residuals. The regression models can either be linear or non-linear. Linear models used include multiple linear regression and ordinary least squares, while the non-linear models are primarily products of Machine Learning, such as decision tree algorithms like Random Forest, Support Vector Regression, and XGBoost Regression. Regression kriging accounts for spatial autocorrelation by geostatistically modelling the spatial correlation of residuals. According to (Hengl, Heuvelink, & Rossiter, 2007), the regression kriging equations are given by:

$$\hat{y}(u_i v_i) = m_{mlr}(u_i v_i) + \varepsilon'_{ok}(u_i v_i) + \varepsilon''(u_i v_i) \quad (6)$$

Where  $\hat{y}(u_i v_i)$  is the target variable at location  $(u_i v_i)$ ,  $(u_i v_i)$  is the coordinates of the  $i^{th}$  location,  $m_{mlr}(u_i v_i)$  is the deterministic component,  $\varepsilon'_{ok}(u_i v_i)$  is the spatially correlated random component, and  $\varepsilon''(u_i v_i)$  is the spatially independent residuals error.

---

## 2.5 Validation and Comparison of the Models

Validation essentially entails finding the error rate of Machine Learning or Geostatistical algorithms. According to Kumar (2018), validation is critical as it allows researchers to understand their models and estimate an unbiased generalisation performance (Kumar, 2018). Grootendorst further reviews different validation techniques. These are discussed below:

i. Splitting data

Splitting data involves splitting the dataset into two different datasets; the training and the testing datasets. Splitting data works by ensuring that the model performs well when encountered by previously unseen data. It can take the form of train/test split or holdout set. The latter entails creating an additional holdout set of data which is data not previously used in any processing or validation steps. The holdout method is deployed after validating the model using the train/test split (Grootendorst, 2019).

ii. K-Fold Cross-Validation (k-Fold CV)

K-Fold Cross-Validation involves splitting the data into additional splits other than the split/test cross folds. This technique splits the data into k-folds, then trains the data on k-1 folds and tests on the one that was left out. It does this for all the combinations and finds the average results for each instance (Kumar, 2018).

iii. Leave-one-out Cross-Validation (LOOCV)

LOOCV is a variant of the k-Fold CV, which uses each sample in the data as a separate test set while all the remaining samples are from the training set. It

---

is identical to k-fold CV when k is equal to the number of observations (Grootendorst, 2019).

iv. Nested Cross-Validation

Nested Cross-Validation allows the researcher to separate the hyperparameters tuning from the error estimation step. It loops through the inner loop for hyperparameter tuning and the outer loop for estimating the accuracy of the model (Grootendorst, 2019).

v. Comparing Models

To compare the performance of different models. Various performance metrics such as R-squared, RMSE, MAE, and MAPE are used. High R-squared scores, lower RMSE, MAE and MAPE scores indicate better model performance.

## **2.6 Gaps in the research**

Malaria studies in Kenya focus on both ML and GS modelling, which, although successful, do not cater for both spatial autocorrelation and heterogeneity. Geostatistical modelling independently allows manipulation of ground-based data, thus supporting prediction and other operations. ML on its own offers little to no spatial support. Integrating both techniques would yield better results with ML's benefits within geostatistical data, which would form an even better baseline for predictive studies (McKinley & Atkinson, 2020). According to a study, hybrid approaches (integrating ML and GS models) bore better results marked by low RMSE, low mean error and a high R score. The hybrid approach outperformed GWR and ANN. These results were attributed to the model's capability of addressing spatial dependency in Soil Organic Carbon, which was the focus of the study. (Chen, et al.,



---

2019) Using this approach in malaria studies is expected to have a similar effect as presented in Chen et al.'s study.

This study will focus on integrating the best performing ML and GS models to develop a hybrid of both. This approach will allow malaria studies to benefit from the spatial capabilities of GS models while also reaping the full benefits of ML methods, whose robustness to noise and ability to handle large non-linear correlated datasets for predictive modelling is laudable.



---

## CHAPTER THREE

### 3 Materials and methods

#### 3.1 Study area

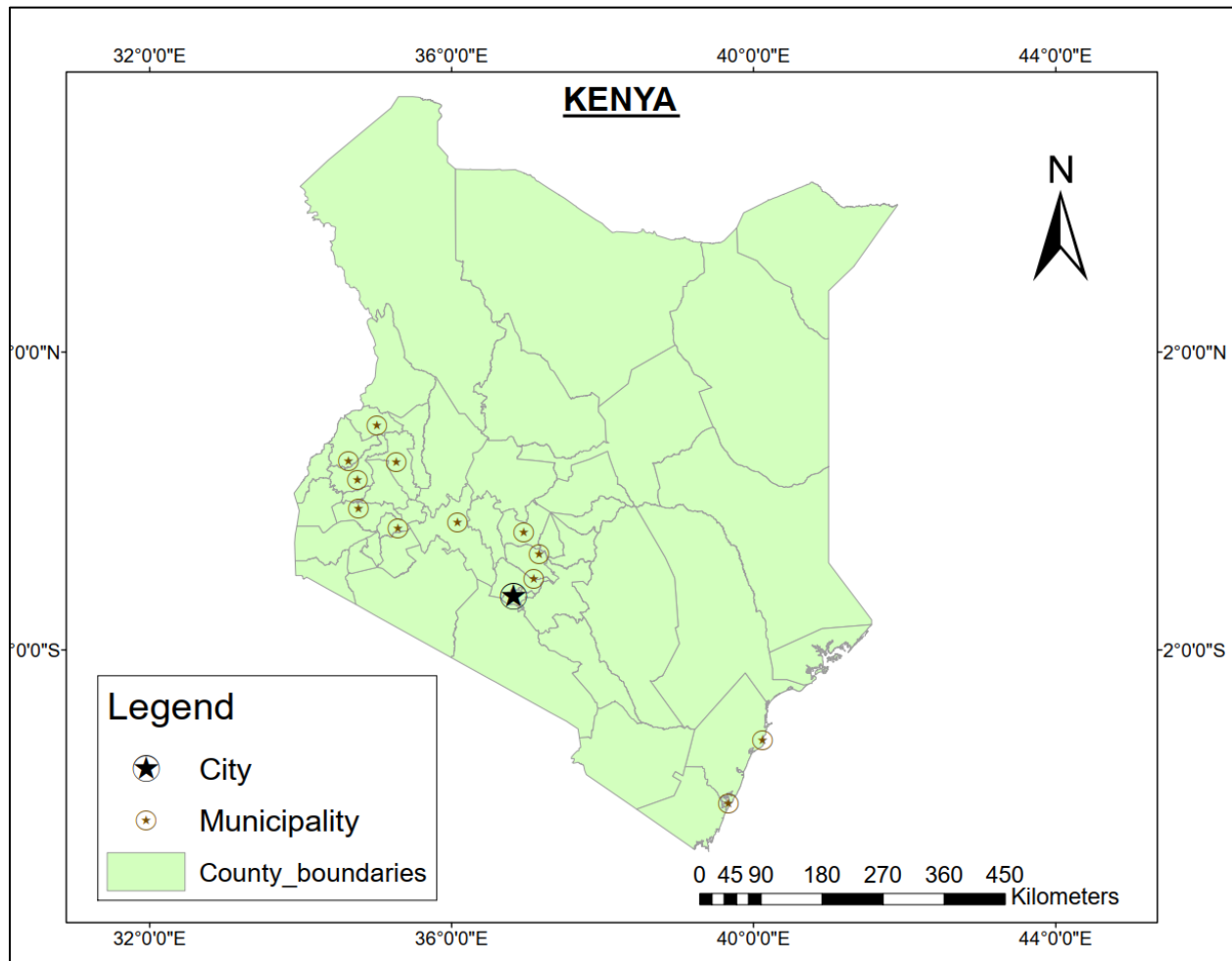


Figure 1: Study Area Map

The country has an approximate land mass of 582,646 square Km, with a 536 Km long coastline. Kenya lies between 0.0236° S, 37.9062° E (World Population Review, 2021). Kenya's capital is Nairobi City located in the heart of Nairobi County. Kenya is one of the 54 African countries. The East African Country borders Uganda to the East, Tanzania to the North, Ethiopia to the South, South Sudan to the South East, Somalia

---

to the West, and the Indian Ocean to the North West. The country is divided into 47 counties.

Kenya experiences tropical climate with relatively high temperatures throughout the year. However, these temperatures vary throughout the country, with highlands experiencing cooler temperatures than the lowland and coastal regions. According to the World Bank Group (2021), the average annual precipitation for Kenya is about 680mm. The northern Arid and Semi-Arid Lands receive less than 250mm of rainfall, while the western region of Kenya receives up to 2000mm of rainfall yearly. Based on data for the last 30 years, Kenya's mean annual temperature is 25.09°C and the mean annual precipitation is 699.10mm (World Bank Group, 2021).

### 3.2 Data

The datasets used in this research are tabulated below:

Table 1: Data, sources and date of access

<b>Data</b>	<b>Source</b>	<b>Date of access</b>
Malaria Prevalence Point Data	Demographic and Health Surveys (DHS)	25/07/2021
Rainfall Data	CHIRPS	26/07/2021
Land Surface Temperature(LST)	MODIS	25/07/2021
Enhanced Vegetation Index (EVI)	MODIS	04/04/2021

Proximity to Water Bodies	World Resources Institute	26/07/2021
Population Density	Earth Data	04/08/2021
Global Human Footprint	Earth Data	04/08/2021
DEM	SRTM	26/07/2021

All the datasets were resampled to 1km by 1km resolution through the nearest neighbors algorithm. They were then reprojected to the World Geographic System (WGS) – WGS84 coordinate reference system (CRS) and projected to Universal Transverse Mercator (UTM) zone.

### 3.2.1 Malaria Prevalence Point Data

This dataset represents the average parasite rate of plasmodium falciparum (PfPR) in children between the ages of 2 and 10 years old within the cluster locations. The points depict cluster locations which are within the correct administrative units. This data was downloaded from the KDHS website. It contained vector data for the years 2000, 2005, 2010 and 2015. All the datasets (in vector format) contained in this dataset are:

- |                          |                                 |                            |
|--------------------------|---------------------------------|----------------------------|
| i. All Population Count  | vi. Diurnal Temperature Range   | x. Global Human Footprint  |
| ii. Annual Precipitation | vii. Drought Episodes           | xi. Gross Cell Production  |
| iii. Aridity             | viii. Enhanced Vegetation Index | xii. Growing Season Length |
| iv. BUILT Population     | ix. Frost Days                  | xiii. Irrigation           |
| v. Day Land Surface Temp |                                 |                            |

xiv.	Insecticide Treated Net (ITN) Coverage	xxiii.	Nightlights Composite
xv.	Land Surface Temperature	xxiv.	Potential Evapotranspiration (PET)
xvi.	Livestock Cattle, Livestock Chickens, Livestock Goats, Livestock Pigs, Livestock Sheep	xxv.	Proximity to National Borders
xvii.	Malaria Incidence	xxvi.	Proximity to Protected Areas
xviii.	Malaria Prevalence	xxvii.	Proximity to Water
xix.	Maximum Temperature	xxviii.	Rainfall
xx.	Mean Temperature	xxix.	SMOD Population
xxi.	Minimum Temperature	xxx.	Temperature
xxii.	Night Land Surface Temp	xxxi.	Travel Times
		xxxii.	UN Population Count
		xxxiii.	UN Population Density
		xxxiv.	UN Population Density

To preserve confidentiality of the respondents, the locations are displaced. Urban clusters are displaced from the actual location by up to 2 Km, while rural clusters experience a displacement of up to 10 Km (Mayala, Fish, Eitelberg, & Dontamsetti, 2018). The sample points are as depicted in the image below:

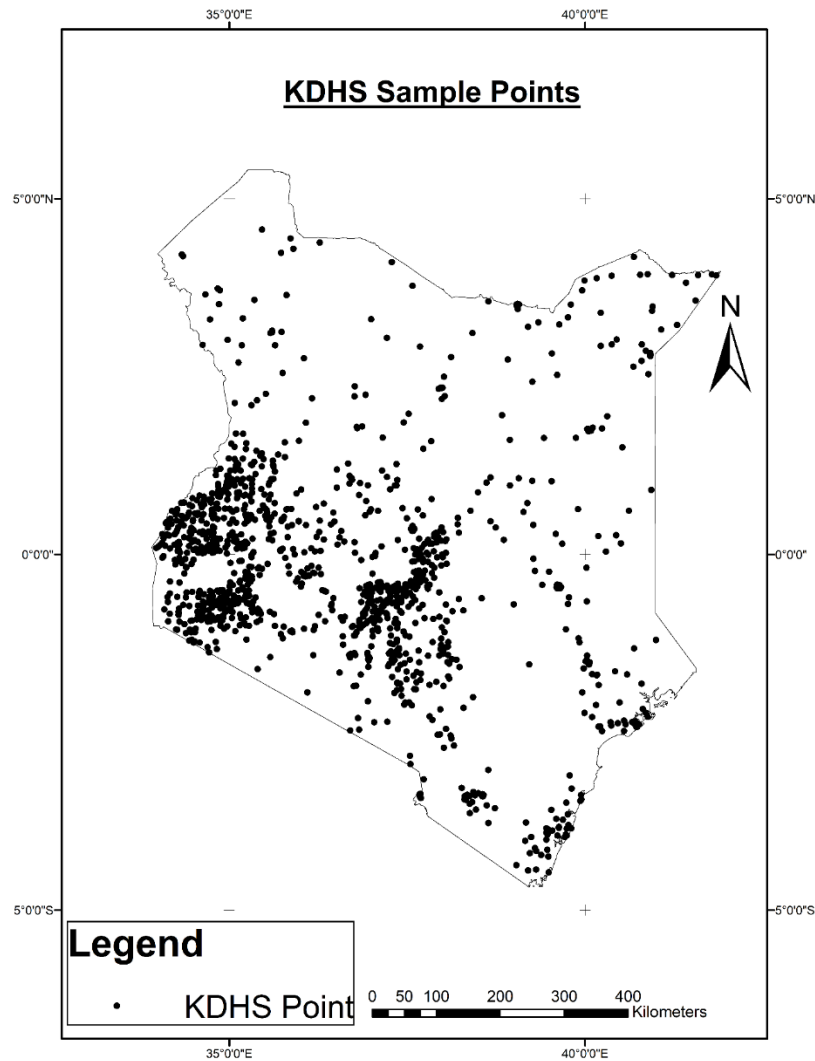


Figure 2: KDHS Sample Points

---

### 3.2.2 Rainfall Data

The rainfall data is the average rainfall within the study area. This data was obtained from Climate Hazards Group InfraRed Precipitation with Stations (CHIRPS). The mean summary statistic was performed on the yearly data obtained to obtain the average rainfall for 2015. CHIRPS data was selected due to its high temporal and spatial resolution, as well as the fact that it is based on different data sources.

### 3.2.3 Land Surface Temperature

This refers to the radiative skin temperature of the land surface derived from solar radiation (Copernicus Global Land Service, 2021). The average annual land surface temperature, obtained from the MYD11C3: MODIS satellites. To obtain the yearly average, the mean statistics was used.

### 3.2.4 Enhanced Vegetation Index

EVI is an index similar to Normalized Difference Vegetation Index (NDVI) that is used to quantify the greenness of vegetation while correcting for some atmospheric conditions and canopy background noise. The formula for EVI is given as:

$$EVI = G \times \left[ \frac{NIR - R}{NIR + C1 \cdot R - C2 \cdot B + L} \right] \quad (7)$$

Where NIR the near infrared, R the red, B the blue bands, and C1, C2 and L are coefficients, with constants, 6, 7.5, and 6 respectively. G is the gain factor, given as 2.5 (Index Database, 2021). The EVI products were downloaded from the MODIS sensor on the Terra satellite (MOD13A3). Average EVI was computed by finding the mean.



---

### **3.2.5 Proximity to Water Bodies**

The lakes dataset was considered in finding this dataset, which was obtained by computing the Euclidean distances to the lakes. This data was obtained from the World Resources Institute.

### **3.2.6 Population Density**

This is the total number of people living within a square kilometer of land. The gridded population of the world (GPW) from Earth Data was used to obtain the dataset.

### **3.2.7 Global Human Footprint (GHF)**

GHF is the Human Influence Index (HII) normalized by biome and realm, expressed in percentage form. This ranges from 0, denoting extremely rural to 100 representing extreme urban (Mayala, Fish, Eitelberg, & Dontamsetti, 2018). This data was obtained from Earth Data.

### **3.2.8 Digital Elevation Model (DEM)**

DEM represents the topography and terrain of the bare ground in digital format. The DEM product was obtained from the Shuttle Radar Topography Mission (SRTM) satellite in the World Resources Institute.

## **3.3 Methods**

### **3.3.1 Introduction**

This section highlights the methodology that was used in the research, detailing the data analysis process.

The Insecticide Treated Net Coverage data was obtained by ordinary kriging of the ITN Coverage data by KDHS to obtain a continuous surface. Clustering and Outlier Analysis (Anselin Local Moran's I) was performed on the ITN coverage data from

KDHS to check for spatial autocorrelation. The equations for the Local Moran's I are given as:

$$I_i = \frac{x_i - \bar{X}}{S_i^2} \sum_{j=1, j \neq i}^n w_{i,j} (x_j - \bar{X}) \quad (8)$$

Where  $x_i$  is an attribute for feature  $i$ ,  $\bar{X}$  is the mean of the corresponding attribute,  $w_{i,j}$  is the spatial weight between feature  $i$  and  $j$ , and;

$$S_i^2 = \frac{\sum_{j=1, j \neq i}^n (x_j - \bar{X})^2}{n - 1} \quad (9)$$

With  $n$  representing the number of features.

The  $z_I$  score for the statistics are computed as:

$$z_{I_i} = \frac{I_i - E[I_i]}{\sqrt{V[I_i]}} \quad (10)$$

Where:

$$E[I_i] = - \frac{\sum_{j=1, j \neq i}^n w_{ij}}{n - 1} \quad (11)$$

$$V[I_i] = E[I_i^2] - E[I_i]^2 \quad (12)$$

A positive  $i$  value depicts high spatial autocorrelation while a negative  $i$  value indicates low spatial autocorrelation. In both instances, the  $p$ -value for the feature needs to be small enough for the cluster or outlier to be considered statistically significant (ESRI, 2018).

Ordinary kriging was then selected due to its ability to cater for spatial autocorrelation. It is also a spatial interpolation tool that is used in the estimation of

---

a parameter at an unsampled location. The formula for point kriging used takes the form:

$$Z^*(u) = \sum_{i=1}^n \lambda_i Z(u_i) \quad (13)$$

Where  $\lambda_i$  is the weights and  $u_i$  is the unsampled location (Bárdossy, 2008).

Raster images for the rest of the datasets were then obtained, clipped to the study area and resampled to 1km by 1km resolution. The main reason for resampling to this grid size is due to the large study area. This resolution would clearly depict the changes within the clusters clearly while utilizing the processing resources available without strain to the computer. These datasets are:

- i. Land Surface Temperature
- ii. Day Land Surface Temperature
- iii. Enhanced Vegetation Index
- iv. Global Human Footprint
- v. Insecticide Treated Net (ITN) Coverage
- vi. Mean Temperature
- vii. Proximity to Water Body
- viii. Rainfall
- ix. Population Density
- x. Digital elevation Model (DEM)

▪ **Extraction of the Independent Variables Data**

The next step in data preparation was to extract the independent variable data for each KDHS sample point, such that for each cluster point, there was a set of

independent variables associated to it. A total of 1296 predictor and labels data were obtained for each of the years, 2000, 2005, 2010, and 2015. This data was then exported in CSV format in preparation for data analysis.

### 3.3.2 Flow Diagram

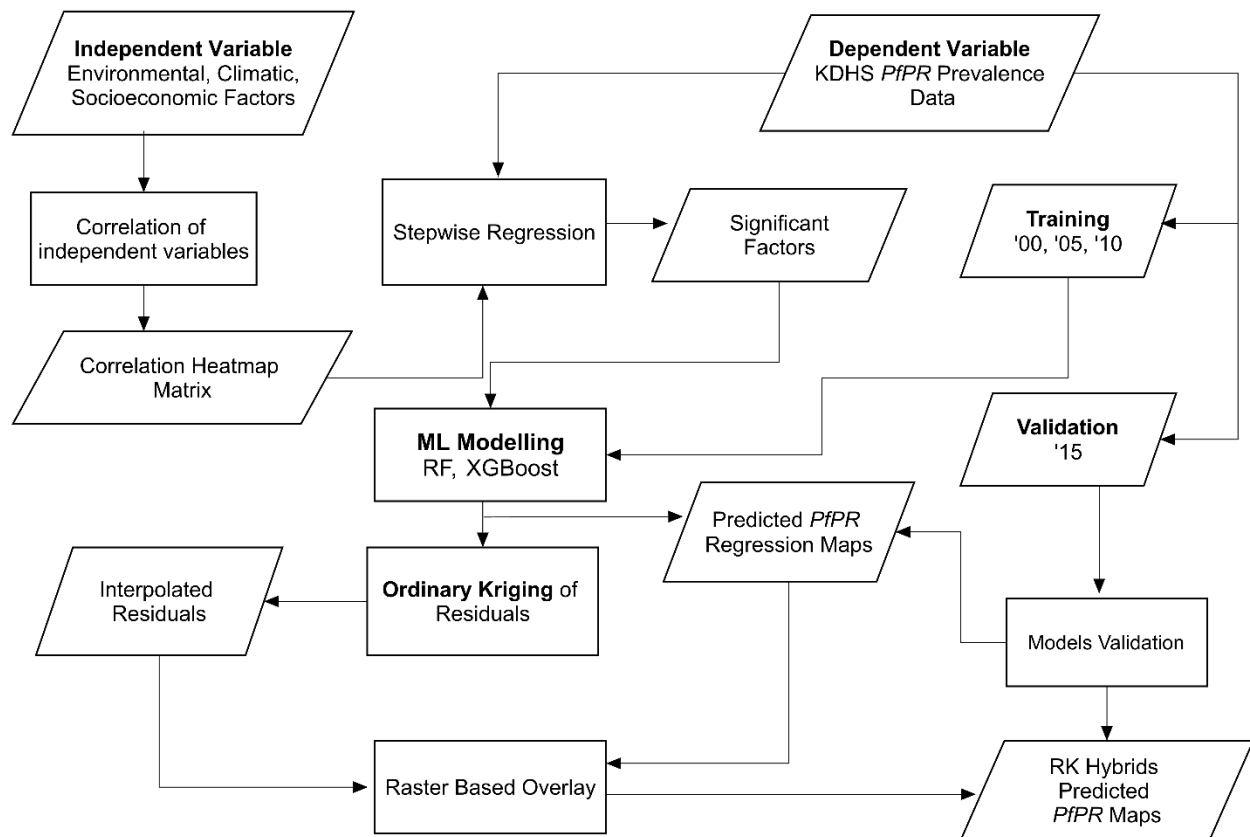


Figure 3: Flow Diagram of the research

### 3.3.3 Correlation of Independent Variables

The bivariate Pearson Correlation was used to determine the correlation between the independent variables listed in section 3.3.1 above. According to Kent State University (2019) Pearson Correlation evaluates whether there exists a linear relationship between the independent variables. It shows whether a statistically significant linear relationship exists between the independent variables while showing

---

both the strength and direction of the linear relationship. The strength of the linear relationship is evaluated by the magnitude of the relationship, while the direction is measured by the positivity or negativity of the relationship. The correlation coefficient between variables are represented by:

$$r_{xy} = \frac{\text{cov}(x,y)}{\sqrt{\text{var}(x)} \cdot \sqrt{\text{var}(y)}} \quad (14)$$

where  $\text{cov}(x,y)$  is the sample covariance of  $x$  and  $y$ ;  $\text{var}(x)$  is the sample variance of  $x$ ; and  $\text{var}(y)$  is the sample variance of  $y$  (Kent State University, 2019). Related variables tending to  $+1$  and  $-1$ , that is  $>+0.8$  and  $<-0.8$  were eliminated to improve the models' performance.

The primary reason for correlating the independent variables is to get rid of highly correlated variables. Highly correlated variables have little to no impact to Machine Learning models. They do not influence the predictions, precision of the predictions nor the goodness-of-fit statistics. The test for multicollinearity was done using the Variance Inflation Factor (VIF) which is given as:

$$\text{VIF}(\beta_i) = \frac{1}{1 - R_i^2} \quad (15)$$

where  $R_i^2$  is the coefficient of determination of  $i^{\text{th}}$  predictor.

Two highly correlated independent variables will lead to high variance in predictions, even if both variables are relevant for prediction. In small samples, therefore, it may be beneficial to omit one of the pair in order to decrease the variance (Gregorich, Strohmaier, Dunkler, & Heinze, 2021). On the contrary Frost (2017) advises researchers to leave the highly correlated variables in the model, if the main intention

---

is predictive analysis, and not to study the impact of each independent variable on the model.

### **3.3.4 Identification of Statistically Significant Factors**

To identify the statistically significant factors, the independent variables that passed the Pearson's correlation test were passed through stepwise regression and tested under a 95% confidence level and a p value of 0.5%. The backward stepwise regression was used to further check the independent variables for collinearity and to ensure they fell within the stipulated criteria. The R-squared, F-statistics and Akaike's information criterion (AIC) metrics were used to check the model's performance with different independent variables. The variables that passed the criteria were then categorized as statistically significant factors that would be used in the Machine Learning and Geostatistical models.

- **Creating the Vector data for the Independent Variables**

To create the surface for prediction, the independent variable rasters were converted to point data in ArcMap. All the covariates identified as statistically significant were loaded onto the ArcMap platform. After ensuring that they were resampled to 1km by 1km cells size and in the WGS84 geographic coordinate system, one of the rasters was converted to point data using the raster to point conversion tool. The rest of the rasters were then converted using the Multivalue to point conversion tool, and then the geometry for the raster points was computed. This created a vector dataset of the independent variables with known geometry, which would form the prediction surface for the PfPR prevalence.

---

### **3.3.5 Machine Learning Models**

Both the Random Forest and XGBoost models were deployed and used to predict PfPR prevalence for the year 2015. In both models, the 2000, 2005 and 2010 data was used in training and fitting the models. The Randomized Search CV and Grid Search CV were used to find the optimal parameters for the models. The optimal parameters were then passed and used to build and train the models. With the models built, they were then used to predict PfPR prevalence for the year 2015. The independent variables vector created in section 3.3.3 above was the prediction surface. Upon completing prediction, the models were validated using Mean Absolute Error (MAE), R-squared( $R^2$ ), Root Mean Squared Error(RMSE) and Mean Absolute Percentage Error (MAPE). The Sklearn metrics library was used to calculate the performance metrics. The predicted Comma Separated Values (CSV) file generated after predicting PfPR was exported and imported into ArcMap to show the predictions. The points were converted back to raster using ArcMap's conversion tools. This resulted in raster surfaces, showing the predicted PfPR prevalence for the year 2015 for each of the Machine Learning Models.

### **3.3.6 Geostatistical Methods**

The geostatistical model used in to predict PfPR prevalence was the regression kriging, which was based on multilinear regression. The regression estimates obtained were assessed to check for spatial autocorrelation through ArcMap's Anselin Local Moran's I clustering tool. The presence of clustering is an indication of spatial autocorrelation, hence justifying the need for kriging. The residuals were therefore interpolated and added to the multilinear regression model to come up with a better PfPR surface catering for spatial autocorrelation.

## ▪ Regression Kriging

To obtain the hybrids of the Machine Learning Models, ordinary kriging was performed on the residuals from each of the models to obtain their residual kriging equivalents.

Residuals are given by:

$$\text{Residual} = [Y_{\text{actual}} - Y_{\text{observed}}] \quad (16)$$

The purpose for regression kriging is to account for spatial autocorrelation, especially in geographic data. Hence, the interpolation of the observed residuals allowed the generation of a more approximate estimate of PfPR predicted surface. The results of kriging were added to the Machine Learning regression surfaces through ArcMap's raster based overlay to obtain more accurate surfaces catering for spatial autocorrelation. The regression kriging models were then validated using MAE, MAPE, R-Squared and RMSE.

### 3.3.7 Model Validation

The Sklearn metrics library was used in model validation, where R-squared, RMSE, MAE, and MAPE were computed. They are given as follows:

#### 1) Mean Absolute Error

MAE represents the average of the absolute difference between the actual and predicted values in the dataset. It measures the average of the residuals in the dataset (Chugh, 2020), and is given as:

$$\text{MAE} = \frac{\sum_{i=1}^n |y_i - x_i|}{n} \quad (17)$$

Where, MAE is the mean absolute error,  $y_i$  the prediction,  $x_i$  the true value and  $n$  the total number of data points.



## 2) Mean Squared Error

MSE represents the average of the squared difference between the original and predicted values in the dataset by measuring the variance of the residuals (Chugh, 2020). It is given as:

$$MSE = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (18)$$

Where MSE is mean squared error, n the number of data points  $Y_i$  the observed values and  $\hat{Y}_i$  the predicted values.

## 3) Root Mean Squared Error

This is the square root of the MSE, and measures the standard deviation of the residuals (Chugh, 2020). It is given as;

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (Obs_i - Pred_i)^2} \quad (19)$$

where n is the number of validation samples, and  $Obs_i$  and  $Pred_i$  denote the corresponding observed and predicted values at  $i^{th}$  locations.

## 4) R-Squared

This represents the proportion of the variance in the dependent variable which is explained by the regression model (Chugh, 2020). It is a scale-free score, and is given as;

$$R^2 = \left( \frac{\sum_{i=1}^n (\widehat{Pred}_i - \overline{Obs_i})^2}{\sum_{i=1}^n (Pred_i - \overline{Obs_i})^2} \right) \quad (20)$$

---

where  $n$  is the number of validation samples, and  $Obs_i$  and  $Pred_i$  denote the corresponding observed and predicted values at  $i^{th}$  locations.

### **5) Mean Absolute Percentage Error**

MAPE is a measure of how accurate a prediction system is measured as a percentage (Chugh, 2020). It is given as:

$$M = \frac{1}{n} \sum_{t=1}^n \left| \frac{A_t - F_t}{A_t} \right| \quad (21)$$

Where  $M$  is the mean absolute percentage error,  $n$ , the number of times the summation iteration happens,  $A_t$  the actual value, and  $F_t$ , the forecast value.

---

## CHAPTER FOUR

### 4 Results

#### 4.1 Exploratory Data Analysis

There was a need to perform exploratory data analysis to check the summary statistics as well as the appearance of the data. The table below shows the summary statistics for the dependent variable:

Table 2: Summary Statistics for the Dependent Variable

PfPR Prevalence	Count	Mean	Std	Min	Max
	1296	0.116941	0.080848	0.008223	0.476596

#### 4.2 Correlation and Statistically Significant Factors

A correlation matrix showing the correlation of the various independent variables considered was generated after performing the Pearson's correlation of the independent variables. The results obtained after performing the correlation test are as depicted below:

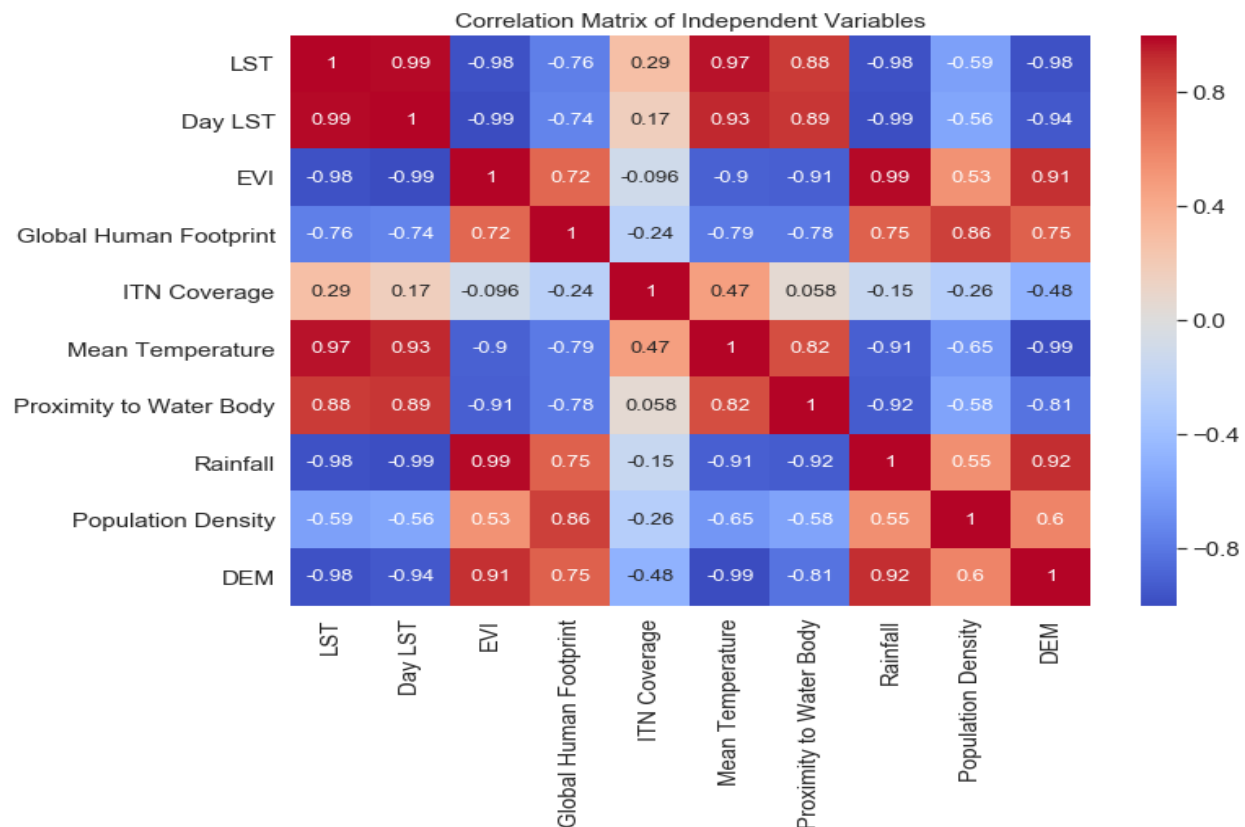


Figure 4: Correlation matrix showing correlation of independent variables

From the results obtained, several independent variable pairs are highly correlated. Values tending to +1 and -1 indicate a high positive and negative correlation respectively. Factors such as Day LST and LST were eliminated due to their high correlation with mean temperature and other factors. Given the high correlation of the independent variables, it is impossible to eliminate most of the highly correlated variables as the reduction of the dataset would severely impact the model and affect the accuracy predictions (Frost, 2017). Therefore, based on the problem, which is to study the performance of different Machine Learning and Geostatistics models, these factors were preserved. The other reason for keeping these highly correlated variables was because past studies used these environmental, climatic and

socioeconomic variables to predict PfPR prevalence (Bashir, Nyakoe, & Sande, 2019; Macharia, et al., 2018; Nkiruka, Prasad, & Clement, 2021; Were, et al., 2019).

When the remaining factors; EVI, Global Human Footprint, ITN coverage, Mean Temperature, Proximity to water Bodies, Rainfall, population density and DEM were passed through backward stepwise regression at 95% confidence level and 5% p-value, the following results were obtained. The lowest AIC value was 403 at different confidence levels and with different p-values, hence the reason the 5% confidence interval and 5% p-value was selected. At the stipulated confidence level and p-value, only one factor, that is, population density did not pass the stepwise regression test, and was therefore a potential candidate for elimination.

Table 3: Statistically Significant Factors

<b>Coefficients</b>	<b>Estimate</b>	<b>Std. Error</b>	<b>t value</b>	<b>P-value</b>
Enhanced Vegetation Index	0.001	0.000	-4.023	0.001
Global Human Footprint	0.001	0.001	3.432	0.001
ITN Coverage	0.131	0.027	4.861	0.000
Mean Temperature	-0.023	0.001	-3.318	0.001
Proximity to Water Body	0.009	0.000	-4.709	0.000
Rainfall	0.117	0.000	22.17	0.000
<b>Population Density</b>	<b>0.001</b>	<b>0.000</b>	<b>-1.695</b>	<b>0.090</b>
DEM	0.012	0.000	-6.285	0.000
Significant at 5%				
Multiple R-Squared		0.873	AIC	403.000
F-statistic		3.240		

---

The figures below shows the rasters for the significant independent variables identified from stepwise regression. They were used in PfPR prevalence.

Below is the raster for Enhanced Vegetation Index.

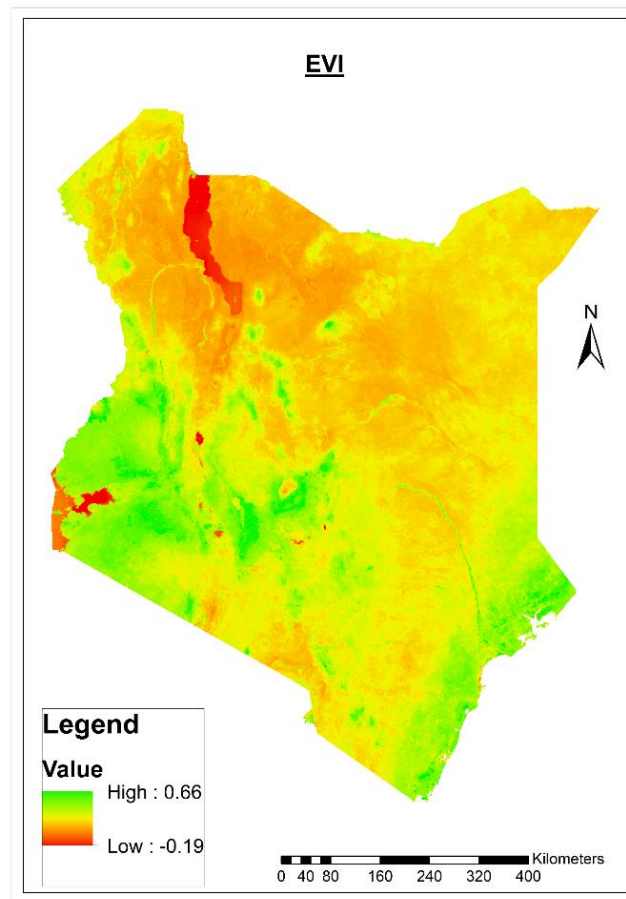


Figure 5: EVI of Kenya (2015)

The areas represented by colour green have higher EVI values than the red areas. The western, central and coastal areas have higher EVI than the rest of the country.

Below is the raster for Global Human Footprint.

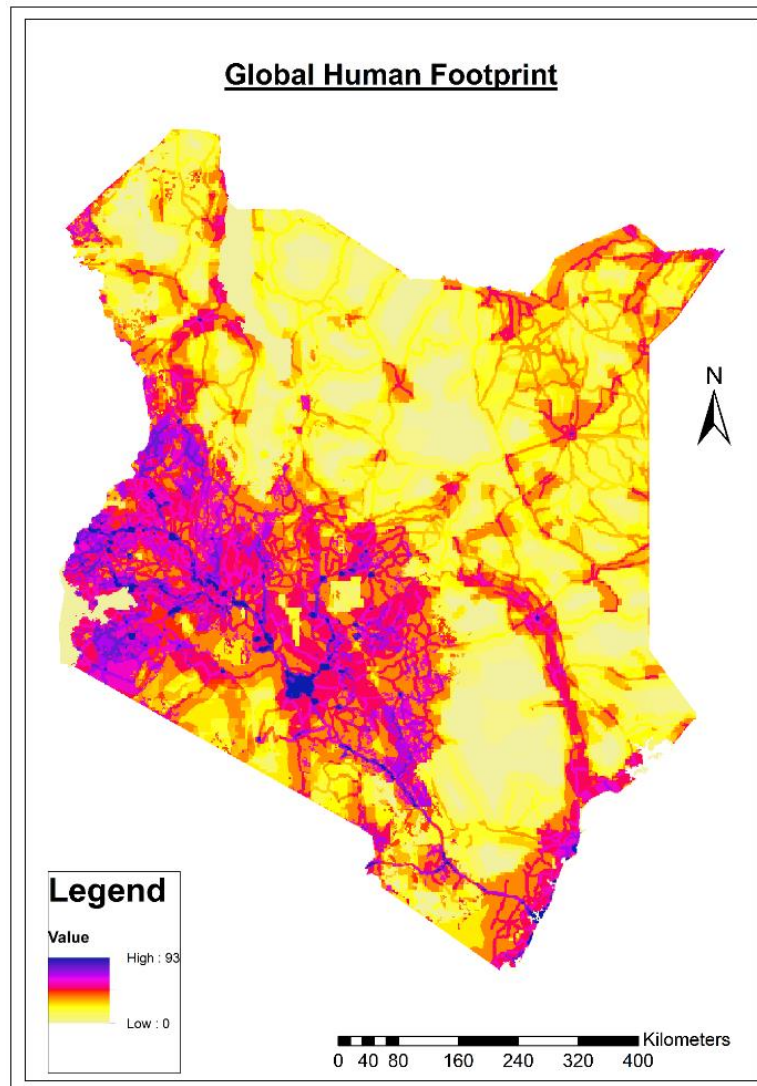


Figure 6: Global Human Footprint (2015)

Purple areas, majorly concentrated in the western and central regions of Kenya show higher GHF than the rest of Kenya. These areas represent those highly impacted by human activities.

Below is the raster for Insecticide Treated Nets, derived from ordinary kriging of the given KDHS data.

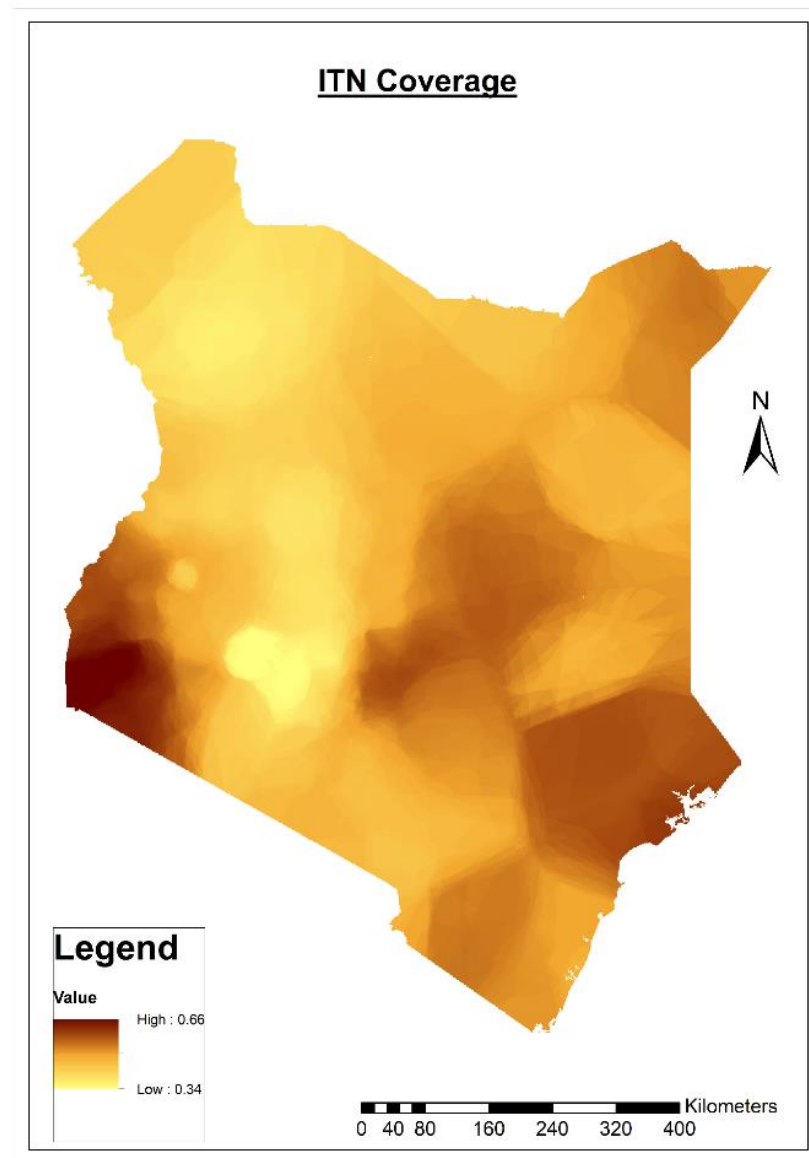


Figure 7: ITN Coverage for Kenya (2015)



According to the map, the areas with the highest ITN coverage are the western and coastal regions, with ITN values of 66%. The North Western region has low ITN coverage.

This is the raster for mean temperature used in the research.

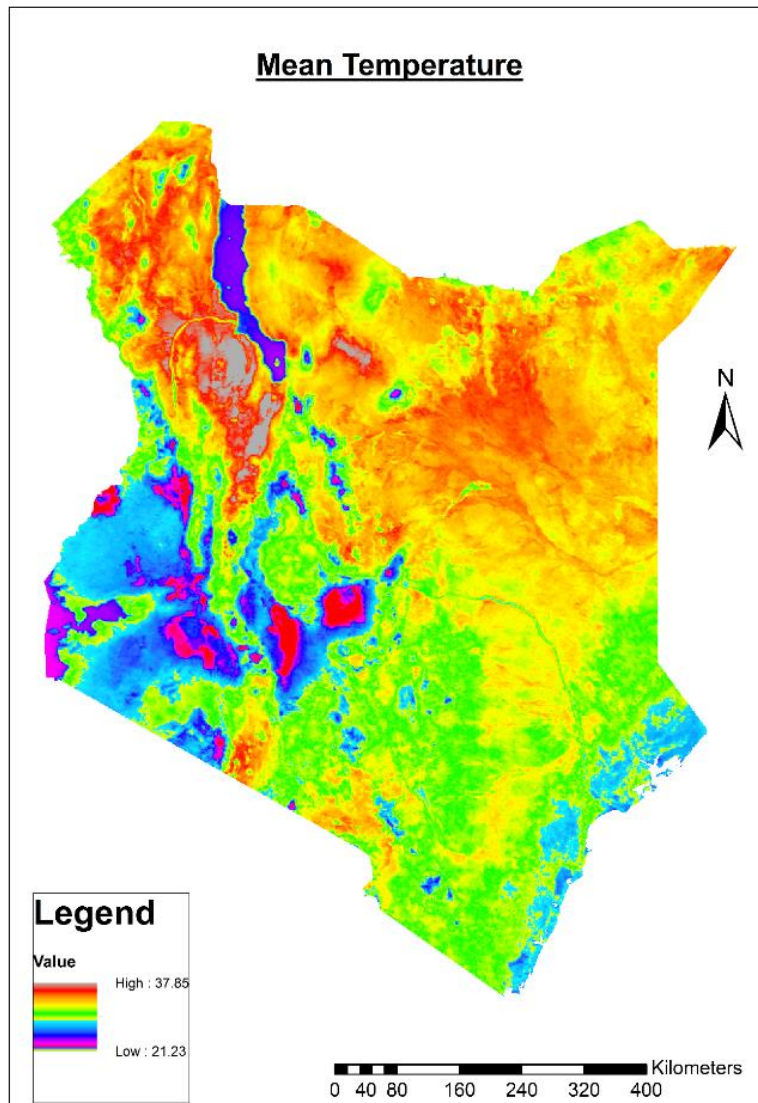


Figure 8: Kenya's Mean Temperature (2015)

High temperatures, represented by red and orange are experienced in the northern parts of Kenya, with water bodies experiencing low mean temperatures.

Below is the raster for proximity to water derived from the calculation of Euclidean distances to lakes and rivers.

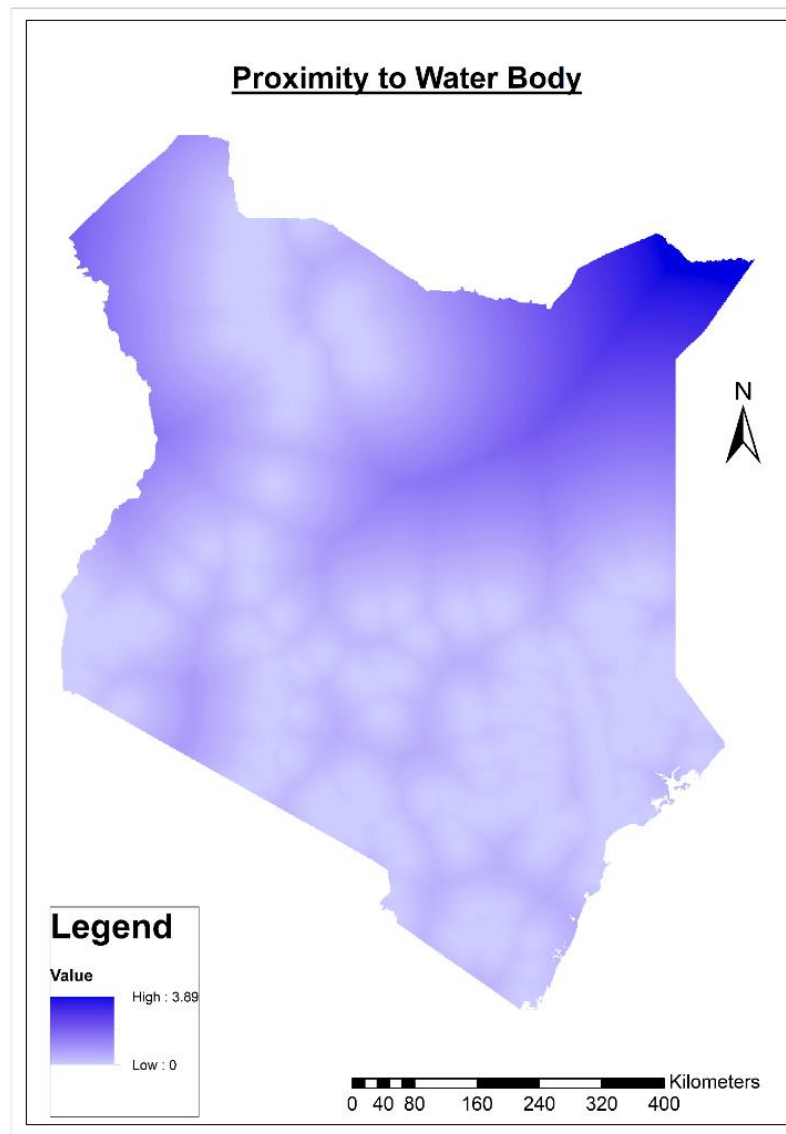


Figure 9: Proximity to Water Bodies

The areas with lighter shades of blue are much closer to water bodies than the areas with deeper blue shades. The North Eastern region is farthest from water bodies as shown by the map.

This is the raster for rainfall data obtained from CHIRPS.

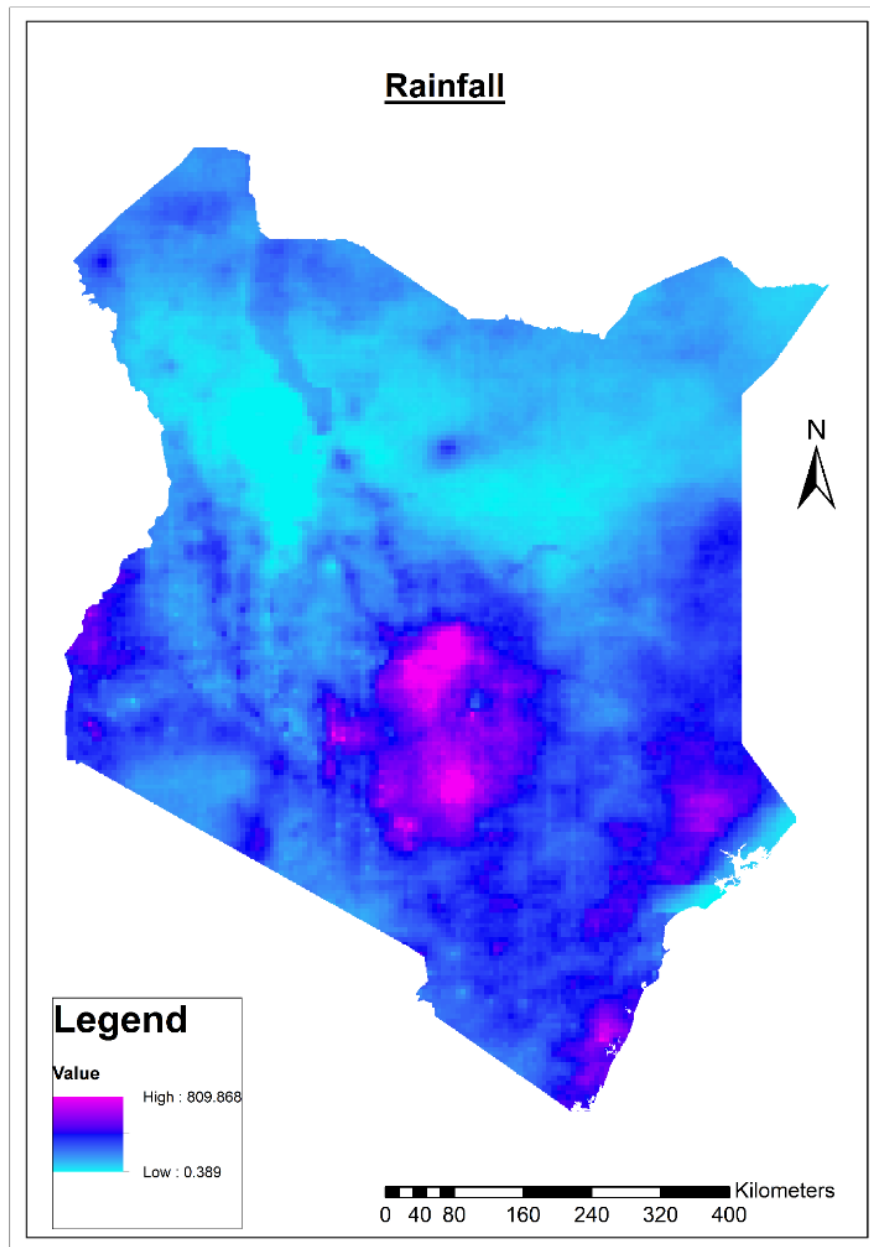


Figure 10: Rainfall (2015)

Central Kenya has the highest rainfall as represented by the purple colour. The northern Kenya region has the lowest rainfall.

This is the DEM raster for Kenya obtained from SRTM.

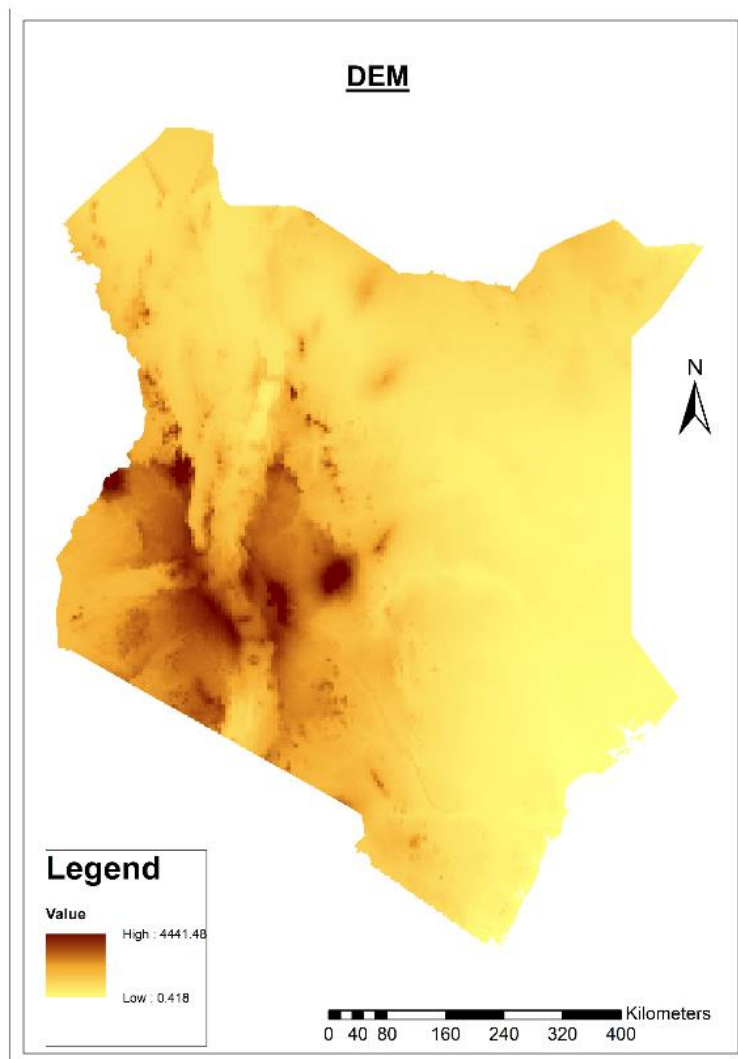


Figure 11: DEM

The highest elevations are concentrated on areas bordering the rift valley as well as some part of western and central Kenya. The eastern and coastal regions have low elevations.

---

### 4.3 Distribution of PfPR Prevalence Points

There is a need to show the actual 2015 data which would serve as the control experiment for which to compare the predicted 2015 PfPR values. The actual points were categorized into 4 classes that showed high and low prevalence areas. The classes were defined as represented in this table below:

Table 4: PfPR Classification

<b>PfPR Prevalence Rate (%)</b>	<b>Class</b>
0.008 – 0.079	1
0.079 – 0.147	2
0.147 – 0.234	3
0.234 – 0.477	4

The following map was generated based on the classification data:

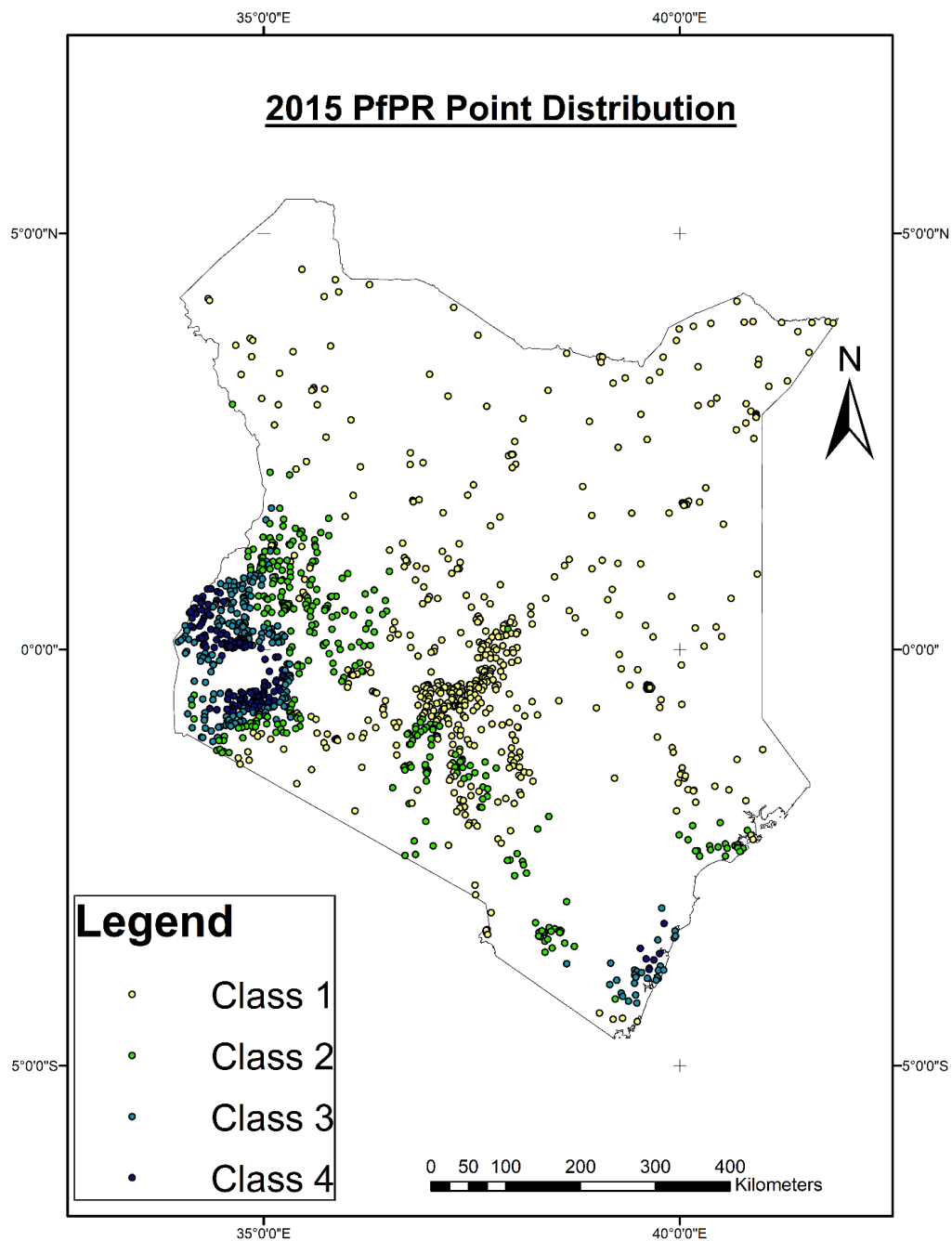


Figure 12: Distribution of 2015 PfPR Data. Class 1 0.008 – 0.079, Class 2 0.079 – 0.147, Class 3 0.147 – 0.234, Class 4 0.234 – 0.477

From the map, the highest PfPR values, depicted by navy blue dots are found in the western and coastal regions. Northern Kenya has low PfPR prevalence. It is important

---

to note that the data points are just a sample of the actual situation and therefore not all regions are covered by the KDHS data.

## **4.4 Prediction of PfPR Prevalence**

### **4.4.1 Using Machine Learning Models**

Machine Learning regression prediction algorithms were done using the Python programming language. The results obtained from the predictive analysis through Random Forest and XGBoost algorithms are as displayed in the pages that follow:

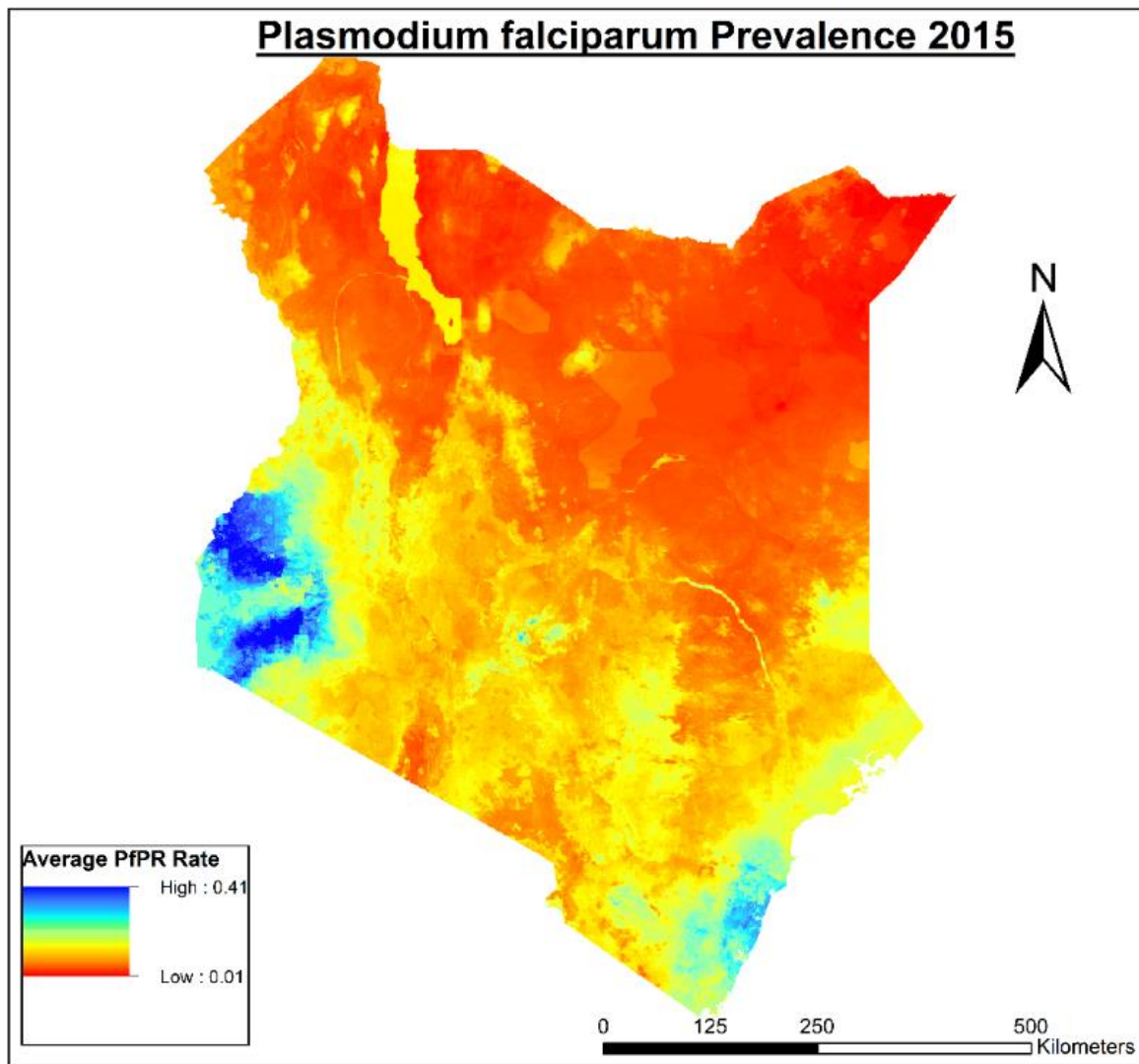


Figure 13: Random Forest PfPR Prediction



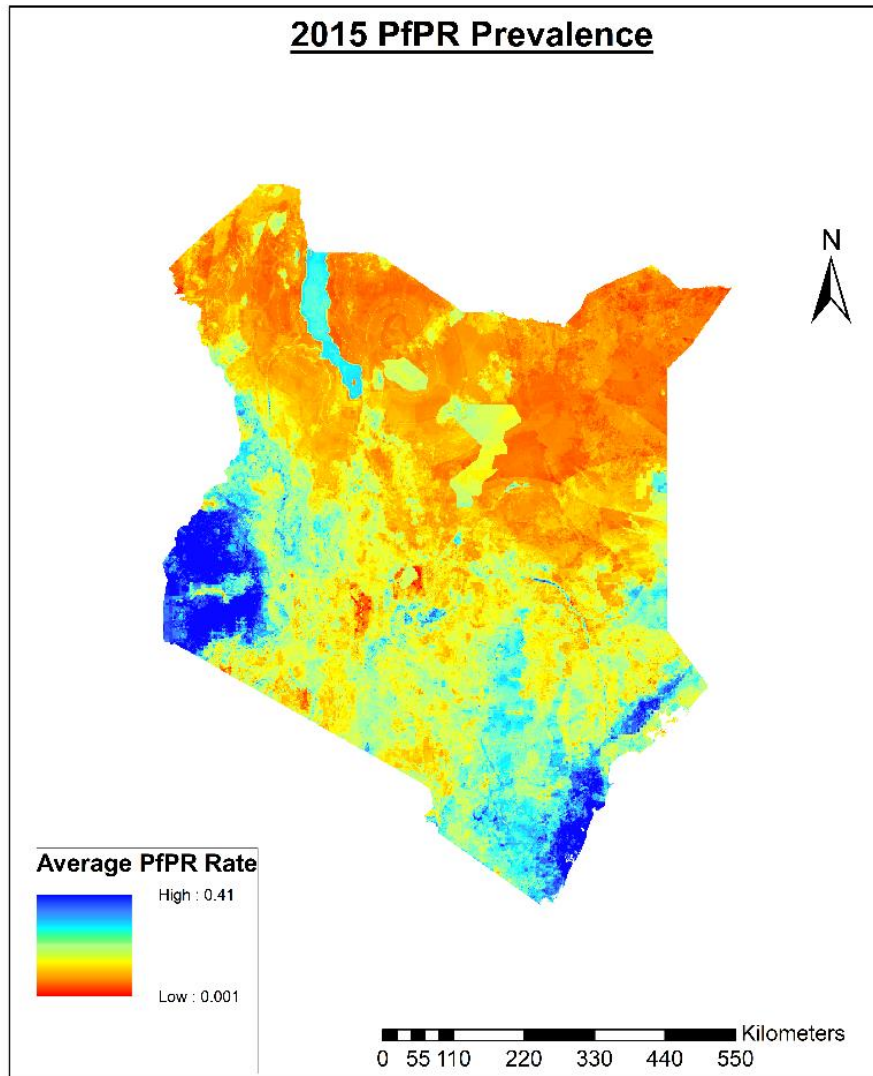


Figure 14: XGBoost PfPR Prediction

From the results displayed above, PfPR prevalence is higher in the Western Kenya, Central and Coastal regions of Kenya. The XGBoost model, however, shows more high areas of PfPR prevalence, as represented by the increased values in the Western, Central and Coastal regions of Kenya. The results depict an actual representation of the raw data, which follows a similar trend, as discussed in the section 4.3 above. Similarly, the data values are a true representation of the actual data, where the highest values at 0.40 and the lowest at 0.01. In the actual dataset, the extreme

---

classes reveal a similar pattern. These results show that both Machine Learning models predict PfPR prevalence with a high accuracy.

Multiple Linear Regression was performed in R and used to predict PfPR prevalence for the year 2015. The residuals from the model were obtained and kriged to obtain regression kriging results shown in the map below:

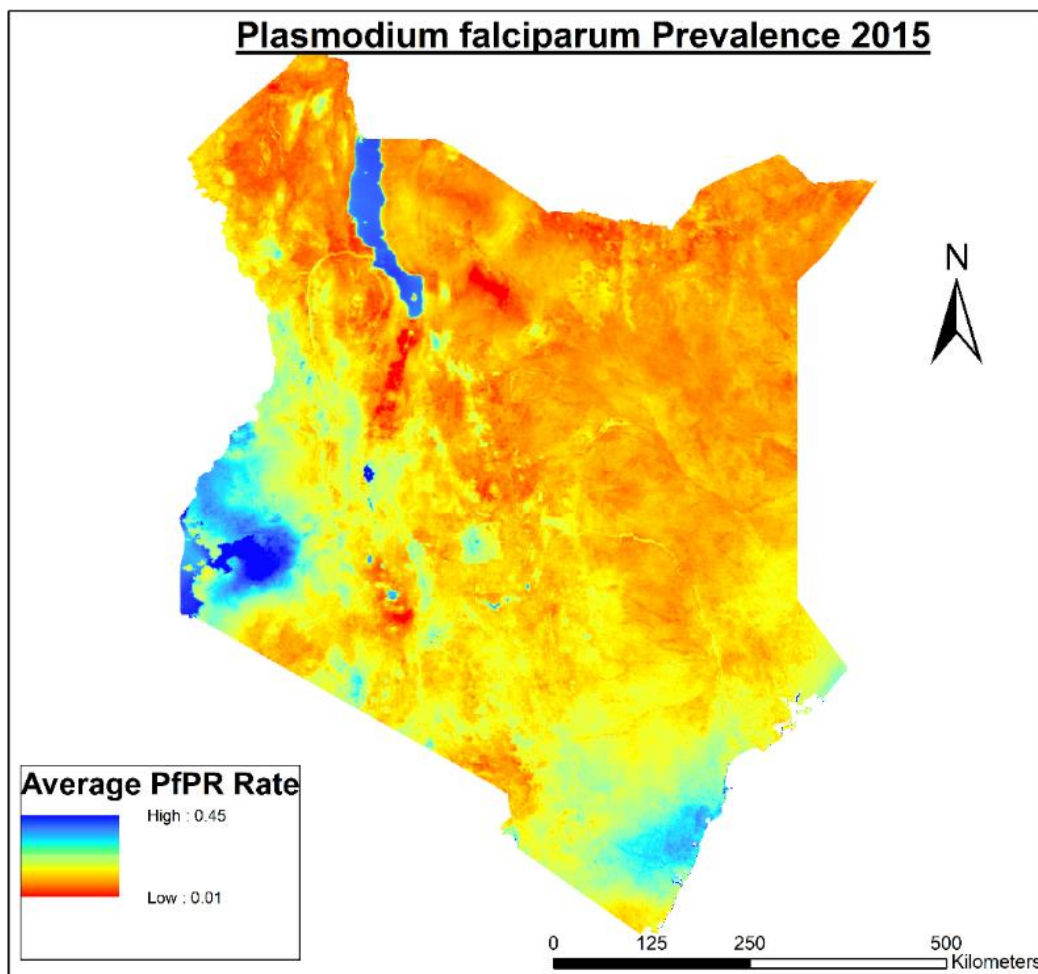


Figure 15: Regression Kriging PfPR Predictions

From the map above, the western region shows the highest PfPR prevalence with values of up to 45%. The majority of northern and eastern regions of Kenya have low PfPR prevalence.

---

The residuals from both the RF and XGBoost models were then analyzed to ascertain spatial autocorrelation. The residuals were analyzed for clustering through the Local Moran's I clustering test.

For regression kriging to proceed the following variogram parameters were adopted after fitting the data in the optimal variogram parameters:

Table 5: Variogram Parameters

<b>Nugget</b>	0.0003
<b>Sill</b>	0.00045
<b>Range</b>	200 kilometers
<b>Semi-variogram model</b>	Spherical

The kriged RF and XGBoost residuals were then added through raster based overlay to their corresponding Machine Learning models to obtain their regression kriging hybrid equivalents, that is Random Forest Regression Kriging (RFRK) and XGBoost Regression Kriging (XGBoostRK). These results are depicted in the pages that follow:

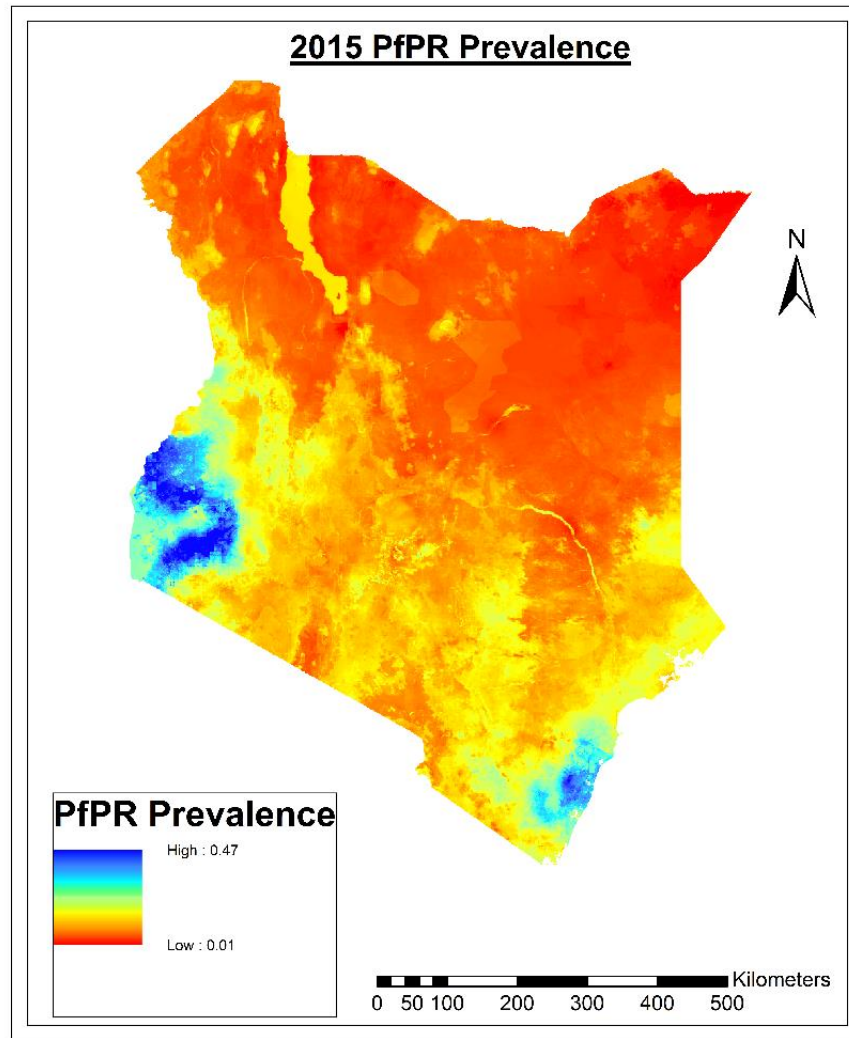


Figure 16: RFRK PfPR Predictions

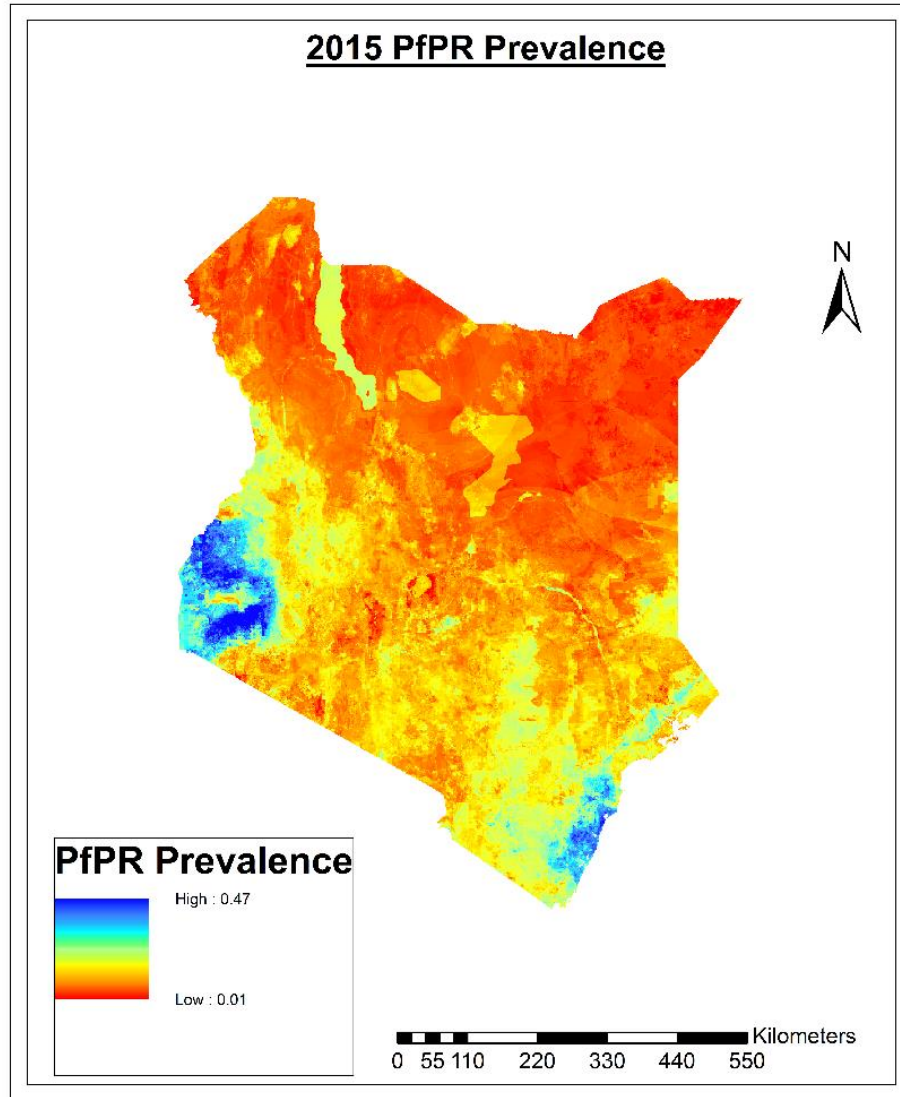


Figure 17: XGBoostRK PfPR Predictions

Based on these results, we can clearly see the corrections for spatial autocorrelation that happen in regression kriging. Both RFRK and XGBoost results look almost similar, with negligible differences, revealing their closeness to the actual data, which shows that spatial autocorrelation is catered for. Further comparisons can be done by analyzing the Machine Learning and Hybrid maps through the performance metrics.

## 4.5 Models Validation and Comparison

MAE, RMSE, R-squared and MAPE were determined as shown in the table below:

Table 6: Models Validation

	<b>RF</b>	<b>XGBoost</b>	<b>RK</b>	<b>RFRK</b>	<b>XGBoostRK</b>
<b>MAE (%)</b>	1.33	2.22	0.92	0.14	1.20
<b>RMSE (%)</b>	1.75	0.11	1.67	0.27	0.07
<b>R-Squared (%)</b>	89.34	83.35	90.91	95.55	94.08
<b>MAPE (%)</b>	15.07	23.25	4.90	3.23	17.03

From the table above, RFRK has the best performance metrics with a R-squared value, MAE, RMSE and MAPE of 95.55%, 0.14%, 0.27%, 3.23% respectively. The high R-Squared value and low MAE, RMSE and MAPE values show that overall, this hybrid model posted the best performance as compared to the rest of the models. RFRK is followed closely by XGBoostRK that has the lowest RMSE value of 0.07%, which is almost negligible.

The poorest performing models were the RF and XGBoost ML models which have the lowest R-Squared values and highest MAPE scores. Despite their poor performance, R-Squared values of above 80% and MAPE values of below 25% are still classified as good performance as discussed by (Chicco, Warrens, & Jurman, 2021). All the models generally performed well in predicting PfPR prevalence in Kenya.

---

## CHAPTER FIVE

### 5 Discussion

The main finding from this study is that hybrid models perform better than Machine Learning or geostatistics alone. This is clearly represented by the increase in prediction accuracy of both the RF and XGBoost Models once regression kriging was done on the residuals. The hybrid models have a predictive accuracy of up to 95.55%, showing that they can be widely used in predictive PfPR prevalence studies. The best performing hybrid model is RFRK which reveals a better prediction accuracy as indicated by its lower MAPE, RMSE, MAE, and higher R-squared score. For spatial studies, hybrid models can be used to increase the predictive accuracy and to make sure that Tobler's laws of geography are catered for. These results can be attributed to their ability to cater for spatial autocorrelation for the residuals. Similar studies conducted by (Chen, et al., 2019; Al-Mudhafar, 2020) reveal that ML and geostatistics hybrids outperform Machine learning only and geostatistics only algorithms. Geostatistical models, however, tend to outperform Machine Learning models in PfPR prevalence prediction, and this again can be attributed to the latter's inability to cater for spatial autocorrelation.

The second finding of this study revealed that the Western, Central and Coastal regions in Kenya have higher PfPR prevalence as compared to the rest of the country. All the Machine Learning, geostatistics and integrated model revealed a similar pattern in PfPR prevalence, where high values (greater than 0.4) are registered in these areas. The findings mean that in areas with high PfPR prevalence, every 4 in 10 children is likely to contract malaria during their lifetime. Similarly, areas with low PfPR prevalence depict lower chances of contracting malaria in the identified regions,

---

with only 1 in 0 children likely to contract the disease in their lifetime. These findings are in tandem with (Macharia, et al., 2018; Nkiruka, Prasad, & Clement, 2021), whose findings show that the Western, Central and Coastal regions have a higher PfPR rate as compared to the rest of the country.

There are several contributing factors that can be attributed to the results represented in figures 13-17 above. These factors can be categorized into climatic, environmental and socioeconomic factors. The statistically significant factors identified in this study are Enhanced Vegetation Index, Global Human Footprint, ITN Coverage, Mean Temperature, Proximity to water bodies, Rainfall and Digital Elevation Models. These variables are used in predictive modelling in Nkiruka, Prasad, & Clement (2021) research. Of these identified factors, most of them are highly correlated but can be used in predictive modelling to generate accurate results, as depicted by the results of the models used in this study.



---

## **CHAPTER SIX**

### **6 Conclusion and outlook**


#### **6.1 Conclusion**

The aims of the study were fully realized. The findings of this study reveal that the hybrid of Machine Learning and Geostatistics Models is optimal for predicting the prevalence of PfPR in Kenya. This is because these hybrids cater for both spatial heterogeneity and spatial autocorrelation. The regression kriging hybrids used in the study, RFRK and XGBoost, accurately predict PfPR prevalence in Kenya with high accuracy values and MAPE, MAE and RMSE values suggesting that the results of these models are acceptable. Of the two hybrids, RFRK performs the best, with XGBoostRK showing the highest improvement when ordinary kriging is done on the residuals. Comparable outcomes are observed in (Chen, et al., 2019; Zhang, Ma, & Guo, 2009) studies that focus on integrating Machine Learning with Geostatistics in predictive studies.

High PfPR prevalence in the Western, Central and Coastal regions of Kenya can be attributed to various factors such Enhanced Vegetation Index, Global Human Footprint, ITN Coverage, Mean Temperature, Proximity to water bodies, Rainfall and Digital Elevation Models. The trends depicted in the findings reveal similar outcomes as those discussed in (Macharia, et al., 2018; Nkiruka, Prasad, & Clement, 2021). The areas with high PfPR prevalence conditions favour the presence of PfPR parasites, hence resulting in a high prevalence rate.

#### **6.2 Recommendations**

Predictive hybrid models should be embraced in malaria studies due to their high predictive accuracies. Such studies can be conducted on a large scale as these models



---

are robust, with their accuracies not largely affected even in performing predictions over larger areas. The hybrid approach utilizes the various climatic, socioeconomic and environmental covariates to establish the relationship between PfPR prevalence and these factors. The ability of these models to cater for spatial autocorrelation makes them excellent choices for use for spatial data.

Future studies should incorporate hybrids of deep learning models such as Convolutional Neural Networks and Artificial Neural Networks in PfPR prevalence prediction.

---

## References

- Alex Smith, S. S. (2017). Measuring sustainable intensification in smallholder agroecosystems: A review. *Global Food Security Volume 12*, 127-138.
- Al-Mudhafar, W. (2020). Integrating machine learning and data analytics for geostatistical characterization of clastic reservoirs. *Journal of Petroleum Science and Engineering 195(1)*, 1-14.
- Bárdossy, A. (2008). *Introduction to Geostatistics*. Institute of Hydraulic Engineering: University of Stuttgart.
- Bashir, I. M., Nyakoe, N., & Sande, M. v. (2019). Targeting remaining pockets of malaria transmission in Kenya to hasten progress towards national elimination goals: an assessment of prevalence and risk factors in children from the Lake endemic region. *Malaria Journal (18)*.
- Breiman, L., & Cutler, A. (n.d.). *Random Forests*. Retrieved from Berkley Edu: [https://www.stat.berkeley.edu/~breiman/RandomForests/cc\\_home.htm](https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm)
- C. Godfray, J. B. (2010). Food security: the challenge of feeding 9 billion people. *Science, 327*, 812-818.
- Centers for Disease Control and Prevention. (2018, July 23). *Malaria*. Retrieved from Centers for Disease Control and Prevention: [https://www.cdc.gov/malaria/malaria\\_worldwide/cdc\\_activities/kenya.html#:~:text=In%20Kenya%2C%20there%20are%20an,of%20Health%20to%20fight%20malaria.](https://www.cdc.gov/malaria/malaria_worldwide/cdc_activities/kenya.html#:~:text=In%20Kenya%2C%20there%20are%20an,of%20Health%20to%20fight%20malaria.)
- Centers for Disease Control and Prevention. (2020, October 6). *Malaria*. Retrieved from Centers for Disease Control and Prevention: <https://www.cdc.gov/dpdx/malaria/index.html>
- Chen, L., Ren, C., Li, L., Wang, Y., Zhang, B., Wang, Z., & Li, L. (2019). A Comparative Assessment of Geostatistical, Machine Learning, and Hybrid

- 
- Approaches for Mapping Topsoil Organic Carbon Content. *International Journal of Geo-Information* 8(174), DOI: doi:10.3390/ijgi8040174.
- Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, DOI: 10.7717/peerj-cs.623.
- Chugh, A. (2020, December 8). *MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared — Which Metric is Better?* Retrieved from Medium: <https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e>
- Ciro Gardhi, Y. Y. (2012). Continuous Mapping of Soil pH Using Digital Soil Mapping Approach in Europe. *Eurasian Journal of Soil Science*, 64-68.
- Copernicus Global Land Service. (2021, January 18). *Land Surface Temperature*. Retrieved from Copernicus Global Land Service: [https://land.copernicus.eu/global/products/lst#:~:text=The%20Land%20Surface%20Temperature%20\(LST,direction%20of%20the%20remote%20sensor.](https://land.copernicus.eu/global/products/lst#:~:text=The%20Land%20Surface%20Temperature%20(LST,direction%20of%20the%20remote%20sensor.)
- ESRI. (2018). *How Cluster and Outlier Analysis (Anselin Local Moran's I) works*. Retrieved from ArcGIS Desktop: <https://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-statistics-toolbox/h-how-cluster-and-outlier-analysis-anselin-local-m.htm>
- Frost, J. (2017, September). *Multicollinearity in Regression Analysis: Problems, Detection, and Solutions*. Retrieved from Statistics by Jim: <https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/>
- Gaurav. (2021, June 12). *An Introduction to Gradient Boosting Decision Trees*. Retrieved from Machine Learning Plus: <https://www.machinelearningplus.com/machine-learning/an-introduction-to-gradient-boosting-decision-trees/>

- 
- Gregorich, M., Strohmaier, S., Dunkler, D., & Heinze, G. (2021). Regression with Highly Correlated Predictors: Variable Omission Is Not the Solution. *International Journal of Environmental Research and Public Health*, 1-12.
- Grootendorst, M. (2019, September 26). *Validating your Machine Learning Model*. Retrieved from Towards Data Science: <https://towardsdatascience.com/validating-your-machine-learning-model-25b4c8643fb7>
- Hengl, T., Heuvelink, G. B., & Rossiter, D. G. (2007). About regression-kriging: From equations to case studies. *Journal of Computers & Geosciences*, 1301-1315.
- Index Database. (2021). *Index: Enhanced Vegetation Index*. Retrieved from Index Database: A database for remote sensing indices: <https://www.indexdatabase.de/db/i-single.php?id=16>
- Kapesa, A., Kweka, E. J., Atieli, H., Afrane, Y. A., Kamugisha, E., Lee, M.-C., . . . Yan, G. (2018). The current malaria morbidity and mortality in different transmission settings in Western Kenya. *PLOS One*, 1-19.
- Kattan, M. W., & Gerds, T. A. (2020). A Framework for the Evaluation of Statistical Prediction Models. *Chest Journal*, 529-538 DOI: <https://doi.org/10.1016/j.chest.2020.03.005>.
- Kent State University. (2019, November 19). *SPSS Tutorials: Pearson Correlation*. Retrieved from Kent State University: <https://libguides.library.kent.edu/spss/pearsoncorr>
- Kumar, A. (2018, February 12). *Machine Learning: Validation Techniques*. Retrieved from Dzone: <https://dzone.com/articles/machine-learning-validation-techniques>
- Langat, A. (2019, August 13). *In Kenya, a stagnating fight against malaria calls for new strategies*. Retrieved from The New Humanitarian: <https://www.thenewhumanitarian.org/analysis/2019/08/13/kenya-stagnating-fight-against-malaria-calls-new-strategies>

- 
- Macharia, P. M., Giorgi, E., Noor, A. M., Waqo, E., Kiptui, R., Okiro, E. A., & Snow, R. W. (2018). Spatio-temporal analysis of Plasmodium falciparum prevalence to understand the past and chart the future of malaria control in Kenya. *Malaria Journal* (17), DOI: <https://doi.org/10.1186/s12936-018-2489-9>.
- Mayala, B., Fish, T. D., Eitelberg, D., & Dontamsetti, T. (2018, September). *The DHS Program Geospatial Covariate Datasets Manual (Second Edition)*. Rockville, Maryland: USAID. Retrieved from The Demographic and Health Surveys Program: [www.DHSprogram.com](http://www.DHSprogram.com)
- Mbaabu, O. (2020, December 11). *Introduction to Random Forest in Machine Learning*. Retrieved from Section: <https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/>
- McKinley, J. M., & Atkinson, P. M. (2020). A Special Issue on the Importance of Geostatistics in the Era of Data Science. *Mathematical Geosciences*, 311-315 DOI: <https://doi.org/10.1007/s11004-020-09858-1>.
- Nkiruka, O., Prasad, R., & Clement, O. (2021). Prediction of malaria incidence using climate variability and machine learning. *Informatics in Medicine Unlocked*, DOI: <https://doi.org/10.1016/j.imu.2020.100508> .
- Odhiambo, J. N., Kalinda, C., Macharia, P. M., Snow, R. W., & Sartorius, B. (2020). Spatial and spatio-temporal methods for mapping malaria risk: a systematic review. *Biomedical Journal*.
- Osman, A. I., Ahmed, A. N., Chow, M. F., Huang, Y. F., & El-Shafie, A. (2021). Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia. *Ain Shams Engineering Journal*. Vol. 12, Issue 2, 1545-1556 <https://doi.org/10.1016/j.asej.2020.11.011>.
- Rencher, A. C., & Schaalje, G. B. (2008). *Linear Models in Statistics*. Hoboken, New Jersey: John Wiley & Sons.
- Ritchie, H. (2017, August 22). *Our World in Data*. Retrieved from <https://ourworldindata.org/yields-vs-land-use-how-has-the-world-produced-enough-food-for-a-growing-population/>

- 
- Schott, M. (2019, April 25). *Random Forest Algorithm for Machine Learning*. Retrieved from Medium: <https://medium.com/capital-one-tech/random-forest-algorithm-for-machine-learning-c4b2c8cc9feb>
- Steurer, M., Hill, R. J., & Pfeifer, N. (2021). Metrics for evaluating the performance of machine learning based automated valuation models. *Journal of Property Research* (38), 99-129 DOI: <https://doi.org/10.1080/09599916.2020.1858937>.
- United Nations. (n.d.). *Global indicator framework for the Sustainable Development Goals and targets of the 2030 Agenda for Sustainable Development*. Retrieved from UN Stats: [https://unstats.un.org/sdgs/indicators/Global%20Indicator%20Framework%20after%202021%20refinement\\_Eng.pdf](https://unstats.un.org/sdgs/indicators/Global%20Indicator%20Framework%20after%202021%20refinement_Eng.pdf)
- VAN ORSHOVEN Jos, T. J. (2012). *Updated common bio-physical criteria to define natural constraints for agriculture in Europe : Definition and scientific justification for the common biophysical criteria*. Europe: Publications Office of the European Union.
- Weiss, D. J., Lucas, T. C., Nguyen, M., Nandi, A. K., Bisanzio, D., Battle, K. E., . . . Collins, E. L. (2019). Mapping the global prevalence, incidence, and mortality of *Plasmodium falciparum*, 2000–17: a spatial and temporal modelling study. *The Lancet*, Volume 394, 322-331.
- Were, V., Buff, A. M., Desai, M., Kariuki, S., Samuels, A. M., Phillips-oward, P., . . . Niessen, L. W. (2019). Trends in malaria prevalence and health related socioeconomic inequality in rural western Kenya: results from repeated household malaria cross-sectional surveys from 2006 to 2013. *Biomedical Journal*.
- World Bank Group. (2021). *Kenya*. Retrieved from Climate Change Knowledge Portal: <https://climateknowledgeportal.worldbank.org/country/kenya/climate-data-historical>

---

World Health Organization. (2021, April 1). *Malaria*. Retrieved from World Health Organization: <https://www.who.int/news-room/fact-sheets/detail/malaria#:~:text=In%202019%2C%20nearly%20half%20of,Americas%20are%20also%20at%20risk.>

World Population Review. (2021). *Where is Kenya in the World?* Retrieved from World Population Review: <https://worldpopulationreview.com/country-locations/where-is-kenya>

Zhang, L., Ma, Z., & Guo, L. (2009). An Evaluation of Spatial Autocorrelation and Heterogeneity in the Residuals of Six Regression Models. *Journal of Forest Science*.



## Appendix

This appendix contains the codes used in the study.

### Insecticide Treated Nets Kriging

```
1 library(sp)
2 library(gstat)
3 library(automap)
4 library(maptools)
5 library(rgdal)
6 library(raster)
7 library(sf)
8
9 setwd("C:\\Users\\Lenovo X1\\Desktop\\JKUAT Project\\New Stuff\\
   Malaria Project\\R Scripts")
10 data=read.csv("bednets_15.csv", header = T)
11 head(data)
12 str(data)
13 plot(data$LONG,data$LAT)
14
15 # convert simple data frame into spatial point data frame
16 coordinates(data)= ~LONG+LAT
17
18 # Set coordinate system to WGS84
19 proj4string(data)=CRS("+init=epsg:4326")
20 class(data)
21
22 # Compute Experimental variogram as variogram object
23 vgm=variogram(ITN_Covera~1, data)
24 summary(vgm)
25 plot(vgm)
26
27 # estimates of variogram model parameters estimated visually
28 partial_sill = 0.006
29 range = 200
30 nugget = 0.0009
31
32 # creates a variogram model object
33 est <- vgm(0.006,"Sph",200,0.000)
34
35 #plot(est)
36
37 # fit model parameters by weighted least-squares
38 fitted <- fit.variogram(vgm, est)
39 plot(vgm, model=fitted)
40
41 # import study area raster boundary
42 bound=raster("Raster.tif")
43 class(bound)
44 str(bound)
45
46 #st_read()
47
48 # create spatial pixel data frame grid from imported raster
49 grd=rasterToPoints(bound,spatial=TRUE)
50 grd1=as(grd,"SpatialPixelsDataFrame")
51 #coordinates(grd1)=~x+y
52 #proj4string(grd1)=CRS("+init=epsg:21096")
53 class(grd1)
54
55 # interpolate using ordinary kriging method on unsampled grid
   points
```

---

```
56 k.o <- krige(ITN_Covera~1, data, grd1, fitted)
57 summary(k.o)
58
59 plot(k.o)
```

## Correlation and Stepwise Regression

```
1 #!/usr/bin/env python
2 # coding: utf-8
3
4 #importing all the libraries needed
5 import seaborn as sns
6 import numpy as np
7 import pandas as pd
8 import matplotlib.pyplot as plt
9
10 #Reading the data
11 dataset = pd.read_csv('new_correlation.csv')
12
13 #Setting display params for the correlation matrix
14 sns.set(font_scale=1)
15 corr_mat = sns.heatmap(dataset.corr(), annot = True, vmin=-1, vmax
    =1, center=0, cmap='coolwarm', annot_kws={"size":13})
16 plt.gcf().set_size_inches(11, 8)
17 corr_mat.set_xticklabels(corr_mat.get_xmajorticklabels(), fontsize
    = 15)
18 corr_mat.set_yticklabels(corr_mat.get_ymajorticklabels(), fontsize
    = 15)
19 plt.title('Correlation Matrix of Independent Variables', fontsize =
    15)
20 cbar = corr_mat.collections[0].colorbar
21 cbar.ax.tick_params(labelsize=15)
22
23 #Read the clean data
24 df = pd.read_csv("Clean_kdhs_data_1.csv")
25
26
27 #Getting the column names
28 x_columns = ['Land_Surface_Temperature_2015',
    'Day_Land_Surface_Temp_2015', 'Enhanced_Vegetation_Index_2015',
    'Global_Human_Footprint', 'ITN_Coverage_2015',
    'Mean_Temperature_2015', 'Proximity_to_Water', 'Rainfall_2015',
    'UN_Population_Density_2015', 'DEM']
29
30
31 y = df['Malaria_Prevalence_2015']
32
33
34 #Creating a function to get model statistics
35 import statsmodels.api as sm
36
37 def get_stats():
38     x = df[x_columns]
39     results = sm.OLS(y, x).fit()
40     print(results.summary())
41 get_stats()
42
43
44 #Remove the columns that do not meet specified criteria
45 x_columns.remove("Day_Land_Surface_Temp_2015")
46
47 #Check performance
48 get_stats()
49
```

## Random Forest Regression

```
1  #!/usr/bin/env python
2  # coding: utf-8
3
4  #import libraries
5  import sys
6  import scipy
7  import numpy as np
8  import matplotlib
9  import matplotlib.pyplot as plt
10 import pandas as pd
11 import sklearn
12 import seaborn as sns
13 from sklearn.model_selection import RandomizedSearchCV
14 from sklearn.ensemble import RandomForestRegressor
15
16 # load datasets
17 train = pd.read_csv('train_kdhs.csv')
18 train
19
20 # Checking for null values
21 train.isnull().sum()
22 train.info()
23
24 # Define X and y values
25 #Test data
26 maskE= train["Year"] == 2015
27 data_2015=train[maskE]
28 data_2015.head()
29
30 test_data = data_2015[['Enhanced_Vegetation_Index_2000', '
    Global_Human_Footprint', 'ITN_Coverage_2000', '
    Mean_Temperature_2000', 'Proximity_to_Water', 'Rainfall_2000', '
    DEM', 'Year']]
31 y_test = data_2015[['Malaria_Prevalence_2000']]
32
33 test_data
34
35 y_test.describe()
36
37 #Train Data
38 maskF= train["Year"] < 2015
39 data00_10=train[maskF]
40
41 data00_10
42
43 train_data = data00_10[['Enhanced_Vegetation_Index_2000', '
    Global_Human_Footprint', 'ITN_Coverage_2000', '
    Mean_Temperature_2000', 'Proximity_to_Water', 'Rainfall_2000', '
    DEM', 'Year']]
44 y_train = data00_10[['Malaria_Prevalence_2000']]
45 y_train
46
47 y_train.describe()
48 train_data
49
50 # Splitting the dataset for training and testing
51 X_train = train_data.values
```

```

52 X_train
53
54 # Number of trees in random forest
55 n_estimators = [int(x) for x in np.linspace(start = 200, stop =
    2000, num = 10)]
56 # Number of features to consider at every split
57 max_features = ['auto', 'sqrt']
58 # Maximum number of levels in tree
59 max_depth = [int(x) for x in np.linspace(10, 110, num = 11)]
60 max_depth.append(None)
61 # Minimum number of samples required to split a node
62 min_samples_split = [2, 5, 10]
63 # Minimum number of samples required at each leaf node
64 min_samples_leaf = [1, 2, 4]
65 # Method of selecting samples for training each tree
66 bootstrap = [True, False]
67 # Create the random grid
68 random_grid = {'n_estimators': n_estimators,
69                'max_features': max_features,
70                'max_depth': max_depth,
71                'min_samples_split': min_samples_split,
72                'min_samples_leaf': min_samples_leaf,
73                'bootstrap': bootstrap}
74
75
76 # Use the random grid to search for best hyperparameters
77 # First create the base model to tune
78 rf = RandomForestRegressor()
79 # Random search of parameters, using 3 fold cross validation,
80 # search across 100 different combinations, and use all available
    cores
81 rf_random = RandomizedSearchCV(estimator = rf, param_distributions
    = random_grid, n_iter = 100, cv = 3, verbose=2, random_state
    =42, n_jobs = -1)
82
83 # Fit the random search model
84 rf_random.fit(X_train, y_train)
85 rf_random.best_params_
86
87 # Make pipeline
88 from sklearn.pipeline import make_pipeline
89 from sklearn.preprocessing import StandardScaler
90 from sklearn.ensemble import RandomForestRegressor
91
92 model = make_pipeline(StandardScaler(), RandomForestRegressor(
    n_estimators=200, bootstrap=True, max_depth=50,
    min_samples_split=2, min_samples_leaf=1, max_features='sqrt',
    n_jobs=-1, random_state=42)).fit(X_train,y_train.ravel())
93 print(model.score(X_train, y_train))
94
95 X_test = test_data.values
96 X_test
97
98 # Make prediction on validation set
99 y_test
100
101 y_pred = model.predict(X_test)

```

```

102 print(model.score(X_test, y_test))
103
104 y_test, y_pred
105
106 from sklearn.metrics import mean_squared_error
107 from math import sqrt
108
109 rms = sqrt(mean_squared_error(y_test, y_predi))
110 print(rms)
111
112 #df_val=pd.DataFrame({'Actual': y_test, 'Predicted' : y_predi},
113                       index=[0])
114 #df_val
115 z= y_pred.tolist()
116 y_predi=pd.DataFrame(z)
117 #final_gpd=pd.concat([y_test,y_predi],axis=1)
118 #final_gpd
119
120 t = y_test.tolist()
121 y_testi=pd.DataFrame(t)
122 y_testi
123
124 final_pred=pd.concat([y_testi,y_predi],axis=1)
125 final_pred
126
127 #Use the inverse transform method to reverse the standardized
128 #variables back in their original form
129 #Importing the various accuracy measures
130 from sklearn import metrics
131 from sklearn.metrics import r2_score
132 from sklearn.metrics import mean_absolute_error
133 from sklearn.metrics import mean_squared_error
134
135 #Calculating the various accuracy measures
136 mae = metrics.mean_absolute_error(y_testi, y_predi)
137 mse = metrics.mean_squared_error(y_testi, y_predi)
138 rmse = np.sqrt(mse)
139 r2 = r2_score(y_testi, y_predi)
140 mape = np.mean(np.abs((y_testi - y_predi)/y_testi))*100
141
142 # #Calculating the various accuracy measures
143 mae = metrics.mean_absolute_error(y_test, prediction)
144 mse = metrics.mean_squared_error(y_test, prediction)
145 rmse = np.sqrt(mse)
146 r2 = r2_score(y_test, prediction)
147 mape = np.mean(np.abs((y_test - prediction)/y_test))*100
148
149 #printing the accuracy measures, rounded to 2 dp
150 print("Results of Random Forest Model:")
151 mae = round(mae, 4)
152 mse = round(mse, 4)
153 rmse = round(rmse, 4)
154 r2 = round(r2, 4)
155 mape = round(mape, 4)
156
157 print("MAE:", mae)
158 print("MSE:", mse)

```

```

157 print("RMSE:", rmse)
158 print("R-Squared:", r2)
159 print("MAPE:", mape)
160
161 from sklearn import model_selection
162 from sklearn.model_selection import train_test_split, KFold
163
164 model = make_pipeline(StandardScaler(), RandomForestRegressor(
165     n_estimators=200, bootstrap=True, max_depth=50,
166     min_samples_split=2, min_samples_leaf=1, max_features='sqrt',
167     n_jobs=-1, random_state=42)).fit(X_train, y_train.ravel()) #####
168     This is the important line
169 results = model_selection.cross_val_score(model, X_train, y_train,
170     cv=kfold)
171 print(results)
172 print ("Accuracy:", results.mean()*100)
173 print ("\n")
174 accu = results.mean()*100
175
176 #Reading the independent variables prepped from ArcMap
177 Ind_Vars15=pd.read_csv("Ind_vars_15.csv")
178 Ind_Vars15.dropna(inplace=True)
179 Ind_Vars15.isnull().sum()
180
181 #Reading the independent var. points, and coordinates. Converting
182     them to array
183 Predictors_grid = Ind_Vars15[['EVI_15', 'GHF', 'ITN_15', 'Mean_Temp', '
184     Prox_Water', 'Rainfall_15', 'DEM_Surface', 'Year']]
185 grid_coordinates=Ind_Vars15[['LONG', 'LAT']]
186
187 ##### Run RF model on the predictor raster values
188 rf_predictors= model.predict(Predictors_grid)
189 ##fn=rf_predictors.reshape(val1.shape[1], val1.shape[2])
190 predi_df=pd.DataFrame(rf_predictors)
191 final_gpd=pd.concat([grid_coordinates, predi_df], axis=1)
192 final_gpd.columns=['LONG', 'LAT', 'Predi_15']
193
194 ##print(final_gpd)
195 final_gpd
196 final_gpd.to_csv("RF_Malaria_pred_15.csv")

```



## XGBoost

```
1 #!/usr/bin/env python
2 # coding: utf-8
3
4 #importing all the libraries needed
5 import seaborn as sns
6 import numpy as np
7 import pandas as pd
8 import matplotlib.pyplot as plt
9
10 #Importing the various accuracy measures
11 from sklearn.metrics import r2_score
12 from sklearn.metrics import mean_absolute_error
13 from sklearn.metrics import mean_squared_error
14
15 # Importing the regressor
16 from sklearn.model_selection import RandomizedSearchCV
17 from xgboost import XGBRegressor
18
19 # load datasets
20 train = pd.read_csv('train_kdhs.csv')
21
22 # Define X and y values
23 maskE= train["Year"] == 2015
24 data_2015=train[maskE]
25 data_2015.head()
26
27 X_test = data_2015[['Enhanced_Vegetation_Index', '
    Global_Human_Footprint', 'ITN_Coverage', 'Mean_Temperature', '
    Proximity_to_Water', 'Rainfall', 'DEM', 'Year']]
28 y_test = data_2015[['Malaria_Prevalence']]
29
30 maskF= train["Year"] < 2015
31 data00_10=train[maskF]
32
33 X_train = data00_10[['Enhanced_Vegetation_Index', '
    Global_Human_Footprint', 'ITN_Coverage', 'Mean_Temperature', '
    Proximity_to_Water', 'Rainfall', 'DEM', 'Year']]
34 y_train = data00_10[['Malaria_Prevalence']]
35 y_train
36
37 X_train.shape, y_train.shape
38
39
40 # ### Second Model
41 RegModel=XGBRegressor(max_depth=3, learning_rate=0.1, n_estimators
    =500, objective='reg:linear', booster='gbtree')
42
43 #Printing all the parameters of XGBoost
44 print(RegModel)
45
46 #Creating the model on Training Data
47 XGB=RegModel.fit(X_train,y_train)
48 prediction=XGB.predict(X_test)
49
50 #Measuring Goodness of fit in Training data
51 from sklearn import metrics
52 print('R2 Value in training:',metrics.r2_score(y_train, XGB.predict
```



```

        (X_train)))
53
54 #Measuring accuracy on Testing Data
55 print('Accuracy',100- (np.mean(np.abs((y_test - prediction) /
    y_test)) * 100))
56
57 #Calculating the various accuracy measures
58 mae = metrics.mean_absolute_error(y_test, prediction)
59 mse = metrics.mean_squared_error(y_test, prediction)
60 rmse = np.sqrt(mse)
61 r2 = r2_score(y_test, prediction)
62 mape = np.mean(np.abs((y_test - prediction)/y_test))*100
63
64 #printing the accuracy measures, rounded to 4 dp
65 print("Results of sklearn.metrics:")
66 mae = round(mae, 4)
67 mse = round(mse, 4)
68 rmse = round(rmse, 4)
69 r2 = round(r2*100, 4)
70 mape = round(mape, 4)
71
72 print("MAE:",mae)
73 print("MSE:", mse)
74 print("RMSE:", rmse)
75 print("R-Squared:", r2)
76 print("MAPE:", mape)
77
78 #Reading the independent variables prepped from ArcMap
79 Ind_Vars15=pd.read_csv("New_Ind_Var15.csv")
80
81 #Checking if there are null values in the independent variables
82 Ind_Vars15.isnull().sum()
83
84 #Reading the independent var. points, and coordinates. Converting
    them to array
85 Predictors_grid = Ind_Vars15[['Rainfall','Mean_Temp','
    Water_Bodies_Surface','ITN_Surface','GPW_Pop','
    Global_Human_Footprint','DEM_Surface']]
86 Predictors_grid.columns=['Rainfall_2015','
    Land_Surface_Temperature_2015','Proximity_to_Water', '
    ITN_Coverage_2015','UN_Population_Density_2015','
    Global_Human_Footprint', 'DEM' ]
87 Predictors_grid.columns
88
89 grid_coordinates=Ind_Vars15[['LONG','LAT']]
90
91 ##### Run XGB model on the predictor raster values
92 XGB_predictors= XGB.predict(Predictors_grid)
93 ##fn=rf_predictors.reshape(val1.shape[1],val1.shape[2])
94 predi_df=pd.DataFrame(XGB_predictors)
95 final_gpd=pd.concat([grid_coordinates,predi_df],axis=1)
96 final_gpd.columns=['LONG','LAT','Predi_15']
97 ##print(final_gpd)
98
99 final_gpd.to_csv("XGBPredicted_Malaria_Prevalence.csv")

```

## Ordinary Kriging (Sample)

```
1 library(sp)
2 library(gstat)
3 library(automap)
4 library(maptools)
5 library(rgdal)
6 library(raster)
7 library(sf)
8 library(mgcv)
9
10 setwd("E:/Scripts_R_Python_store/Shee")
11 data=read.csv("Xgboost_kdhs_correlation.csv", header = T)
12 head(data)
13
14 # convert simple data frame into spatial point data frame
15 # Set coordinate system to GCS-WGS84
16 proj4string <- "+proj=longlat +datum=WGS84 +no_defs +ellps=WGS84 +
    towgs84=0,0,0"
17 statPoints <- SpatialPointsDataFrame(coords = data[,c("LONGNUM", "
    LATNUM")],
18   data      = data,
19   proj4string = CRS(proj4string))
20
21 ##remove duplicate locations call sp::zerodist within a bracket
    index
22 ##statPoints <- statPoints[-zerodist(statPoints)[,1],]
23
24 # Compute variogram parameters automatically that will be required
    later.
25 ok_param = autofitVariogram(Residuals~1, statPoints,
26   model = c("Sph", "Exp", "Gau"),
27   kappa = c(0.05, seq(0.2, 2, 0.1,0.5,0.001), 5,
    10,15,20,25))
28 summary(ok_param)
29 plot(ok_param)
30 str(ok_param)
31
32
33 ##Create experimental variogram object based on sample points
34 variog=variogram(Residuals~1, statPoints)
35 plot(vgm)
36
37 ## create variogram model by passing appropriate variogram
    parameters - Equivalent to model training
38 est <- vgm(0.0004,"EXP",130,0.0001677214)
39
40 ## fit Variogram model in to the variogram object
41 fitted <- fit.variogram(variog, est)
42 plot(variog, model=fitted)
43
44 #Importing the predicted grid data-unsampled locations(either csv
    or shp)
45 pred_grid=read.csv("Interpolation_points_2.csv")
46
47 # convert simple data frame into spatial point data frame
48 # Set coordinate system to GCS-WGS84
49 proj4string <- "+proj=longlat +datum=WGS84 +no_defs +ellps=WGS84 +
    towgs84=0,0,0"
```

```

50 predic_stat_grid <- SpatialPointsDataFrame(coords = pred_grid
      [,c("LONG","LAT")],
51 data = pred_grid,
52 proj4string = CRS(proj4string))
53
54 head(pred_grid)
55 # interpolate using ordinary kriging method on unsampled grid
    points
56 o.k <- krige(Residuals~1, statPoints, predic_stat_grid, fitted)
57 summary(o.k)
58
59 plot(o.k)
60
61 # write the regression results as csv table
62 write.csv(o.k, file = "Malaria_Xgboost_Residuals.csv")
63 interpolated_residuals = read.csv("Malaria_Xgboost_Residuals.csv",
      header = T)
64 print(interpolated_residuals)
65
66 ## Import raster grid that has the same setting for interpolation i
    .e Night light
67 night_light=raster("Random Forest_updated_updated.tif")
68
69 ext=extent(night_light)
70
71 ## create raster object using same raster paramaters as Night light
72 raster_object = raster(ncol=night_light@ncols, nrow=night_
    light@nrows, xmn=ext@xmin, xmx=ext@xmax, ymn=ext@ymin, ymx=
    ext@ymax)
73
74 # retrieve predicted and variance data
75 Residual_values_df=interpolated_residuals[,c("LONG","LAT","var1.
    pred")]
76 Variance_values_df=interpolated_residuals[,c("LONG","LAT","var1.var
    ")]
77 # will need to rename colnames for raster
78 colnames(Residual_values_df) <- c('x', 'y', 'vals')
79 colnames(Variance_values_df) <- c('x', 'y', 'vals')
80
81 # use rasterize to create Residual raster
82 Residual_rasterize <- rasterize(x=Residual_values_df[, 1:2], # lon-
    lat data y=raster_object, # raster object field=Residual_values
    _df[, 3], # vals to fill raster with
83 fun=mean) # aggregate function
84
85 # use rasterize to create Variance raster
86 Variance_rasterize <- rasterize(x=Variance_values_df[, 1:2], # lon-
    lat data
87 y=raster_object, # raster object
88 field=Variance_values_df[, 3], # vals to fill raster with
89 fun=mean) # aggregate function
90
91 # write the raster data as tiff
92 Residual_raster<-writeRaster(Residual_rasterize,'Malaria_XGBoost_
    Residuals.tiff',overwrite=TRUE)
93 Variance_raster<-writeRaster(Variance_rasterize,'Malaria_XGBoost_
    Variance.tiff',overwrite=TRUE)

```