Application of Spatial Generalized Linear Models in Predicting Visceral Leshmaniasis causing vector : a case of Turkana County

Ngala, Hassan Samini

Submitted in partial fulfilment of the requirements for the degree of Master of Science in Statistical Science of Strathmore University

Strathmore Institute of Mathematical Sciences

Strathmore University

Nairobi, Kenya

June 2025

Declaration

I declare that this work has not been previously submitted and approved for award of a degree by this or any other University. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Name: **Hassan Samini Ngala**

Signature: .....

Date: April 4, 2025

Approval

The thesis of Hassan Samini Ngala was reviewed and approved by the following:

Dr. Evans Otieno Omondi

Principal supervisor,

Institute of Mathematical Sciences, Strathmore University.

April 4, 2025

Dr. Kennedy Senagi

Co-supervisor,

Institute of Mathematical Sciences, Strathmore University.

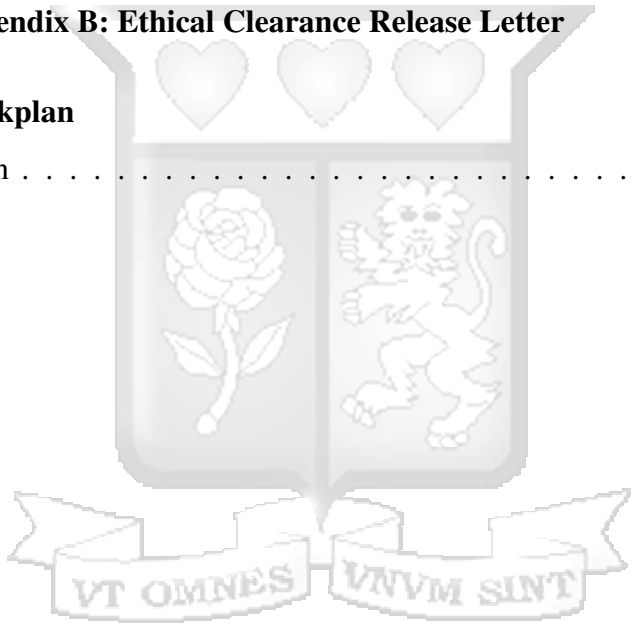
4th April 2025

Abstract

Visceral leishmaniasis (VL) remains a public health concern in Kenya, especially in Turkana County, where control strategies often lack a multidisciplinary One Health approach integrating human, animal, and environmental health. This study investigates how environmental (NDVI, brightness), weather (temperature, evapotranspiration, wetness), and spatial factors (latitude, longitude) influence vector abundance.We applied the Spatial Autoregressive Moving Average (SARMA) model alongside machine learning algorithms XGBoost, Spatial Random Forest (RF), and k-Nearest Neighbors (KNN) to assess predictive accuracy. Model performance was evaluated using the Akaike Information Criterion (AIC), residual variance, spatial dependence tests, and classification metrics such as sensitivity, specificity, and Area Under the Curve (AUC). XGBoost outperformed SARMA, achieving the highest AUC (0.8997) and sensitivity (99.63%), and ranked Latitude (43.38%) and Longitude (29.12%) as the most influential variables, followed by NDVI (8.95%), Indoors Sticks (4.23%), Wetness (4.19%), and Humidity (2.13%). SARMA identified NDVI ($\beta = 0.3638, p = 0.0079$), Wetness ($\beta = 0.4836, p = 0.0185$), and ET ($\beta = -0.0068, p = 0.0051$) as significant, with strong spatial effects observed ($\rho = 0.5440, \lambda = -0.7760$).These findings underscore the value of spatial modeling in vector surveillance; while SARMA effectively captures spatial dependencies, XGBoost offers superior predictive power. The results support the development of data-driven, location-specific interventions for VL control in Turkana County, enhancing the precision and impact of public health efforts.

Keywords: Sandflies, Leishmaniasis, OneHealth, Social determinant, Transmission, Ecology

Table of contents



List of tables

List of abbreviations

VL	Visceral Leshmaniasis	PKDL	Post-kala-azar dermal leishmaniasis

Acknowledgement

First and above all, I am grateful to the Almighty God for His provision, grace, and for granting me good health throughout the study period.

I am grateful to my wife, Mercy Marende for her unwavering support and cheering me on throughout the study period.

I am grateful to my brothers and sisters for their immeasurable support throughout the study period.

I am indebted to my supervisors, Dr. Evans Otieno Omondi, Dr.Kennedy Senagi (ICIPE), Dr. George Simons (IDEMS International), Dr. David Stern (IDEMS International) for their availability, kind guidance, and support without which this thesis would not have been a success.

I am grateful to my classmates for their support and encouragement.

Chapter 1

Introduction

1.1 Background to the study

Sandflies are tiny blood-sucking flies that belong to the group of insects known as the family Psychodidae and occur in tropical and subtropical regions. They are minute (1.5–3 mm long) with hairy, fringed wings that give them a distinctive "sand-like" appearance in flight. Despite their small size, sandflies are considered important public health pests in endemic regions due to their role as vectors of several diseases, including leishmaniasis.

One type of blood-feeding insect with a particularly notorious reputation as a vector of parasitic diseases infecting humans and animals alike is the sandfly. They tend to bite at night in dark, protected environments like caves, forests, or places with thick vegetation. Their bites can hurt and can lead to swelling, skin lesions, or infections. Leishmaniases Visceral, Mucocutaneous, and PKDL Leishmaniases are all three sandfly vector-borne diseases.

A study by Van Dijk and Schallig (2023) posits that visceral leishmaniasis (VL), a fatal parasitic infection due to the parasite Leishmania donovani or Leishmania infantum, has as its transmitting vector the female sandfly.

Further Phlebotomus orientalis and Phlebotomus martini are known as key vectors in East africa, while the principal vector in Kenya is Phlebotomus martini.

1.1.1 Physiology of Leishmaniasis

Leishmaniasis is a disease caused by protozoan parasites belonging to the genus Leishmania. As a zoonotic infection, it persists in animal reservoirs, making complete eradication

challenging. These parasites are spread to humans and animals through bites from infected female sandflies. Once inside the host, *Leishmania* parasites live within the host's cells as intracellular pathogens.

The lifecycle of the parasite consists of two distinct stages:

- **Promastigote Form:** Within the sandfly, the parasite exists in a flagellated state known as the *promastigote*. When the sandfly feeds on blood, it ingests these promastigotes, which then multiply in its midgut before being transmitted to a new host.
- **Amastigote Form:** Once inside a human or animal host, the promastigotes transform into the non-flagellated *amastigote* form. These amastigotes invade macrophages, a type of immune cell responsible for engulfing foreign pathogens.

Leishmaniasis primarily affects the skin, spleen, liver, and bone marrow. However, its clinical presentation varies significantly depending on the *Leishmania* species, the host's immune response, and the type of infection, which can be either cutaneous or visceral.

- **Cutaneous Leishmaniasis:** This is the most prevalent form of the disease, characterized by the development of ulcers or skin lesions that may result in permanent scarring and disfigurement.
- **Visceral Leishmaniasis:** Often called *kala-azar*, this term originates from *kala*, meaning black, referring to the characteristic darkening of exposed skin. This is the most severe form of the disease, as it can affect critical internal organs, including the liver, spleen, muscles, and bone marrow. If left untreated, visceral leishmaniasis can be fatal.

1.1.2 Optimal conditions for transmission

For sandflies to thrive and effectively transmit *Leishmania*, specific ecological conditions must be met.

1.1.3 Temperature

Temperature has a crucial role in sandfly survival and reproduction. Sandflies prefer warm environments, with optimal temperatures ranging between 20°C and 27°C, which support both their breeding and the survival of *Leishmania* parasites. When temperatures fall below 20°C or exceed 27°C, sandfly populations are negatively affected, reducing their ability to sustain transmission (El Omari et al., 2023).

1.1.4 Humidity

Humidity and rainfall also influence sandfly survival. According to El Omari et al. (2023), sandflies are highly sensitive to moisture, and their abundance shows a negative correlation with both rainfall ($r = -0.8$) and humidity ($r = -0.76$). This suggests that excessive moisture can hinder their proliferation. However, observations by Marangu et al. (2024) indicate a more dynamic pattern, where sandfly populations peak during the dry season and towards the end of the rainy season. Interestingly, a high prevalence has also been reported at the peak of the rainy season, highlighting potential variations in local environmental conditions and sandfly adaptability.

1.1.5 Breeding Sites

Breeding sites are another critical factor for sandfly persistence. Sandflies typically breed in ecotopes rich in decaying organic matter, such as animal dung, leaf litter, and other dark, humid environments. Poor sanitation and human settlements near these breeding grounds can elevate the risk of transmission.

1.1.6 Vegetation

Vegetation also plays a significant role, as dense vegetation provides a suitable microhabitat for sandflies. It offers shelter and a stable environment for both larvae and adult sandflies, further supporting their life cycle.

The combination of these factors temperature, humidity, breeding sites, and vegetation creates optimal conditions for sandflies and *Leishmania* parasites to persist, ultimately sustaining transmission and increasing infection rates in affected regions.

1.1.7 Life cycle of leishmania parasite

The life cycle of the *Leishmania* parasite consists of two stages, involving both a sandfly vector and a mammalian host. Initially, a female sandfly becomes infected when it feeds on an infected human or animal. Inside the insect's gut, the parasite exists in its promastigote form.

Inside the sandfly, these promastigotes replicate and develop into their motile form. During the next blood meal, the sandfly injects the parasite into a new host's skin, initiating the second stage of the cycle. Within the host, immune cells called macrophages engulf the parasites, where they transition into their amastigote form and multiply. This intracellular replication allows the infection to spread to other tissues, potentially leading to skin lesions or damage to internal organs. The cycle persists as sandflies continue to feed on infected hosts, sustaining the transmission as long as environmental conditions support their survival.

1.1.8 Social and economic impact

Recent research suggests that Eastern Africa bears the highest burden of visceral leishmaniasis (VL), accounting for nearly two-thirds of global cases. Within this region, a significant proportion of those affected are children under the age of 15, while women represent approximately one-third of reported infections. Meanwhile, the Indian subcontinent contributes

around 12% of the global VL burden (Alvar et al., 2023).Additionally, van Dijk et al. (2024) further highlights that Kenya is the second-largest VL focus globally. A meta-analysis by Geto et al. (2024) further supports this claim, showing that Eastern Africa contributes to approximately 73% of global VL cases.

VL has significant social and economic consequences at multiple levels, from global to local communities. A systematic review by Grifferty et al. (2021a) indicates that VL substantially reduces patients' quality of life. Furthermore, post-kala-azar dermal leishmaniasis (PKDL) often leads to severe skin disfigurement, which is associated with poor mental health and diminished quality of life due to social stigma and exclusion. As a result, stigma can negatively impact health-seeking behaviors, reducing treatment adherence and prolonging disease prognosis. Moreover, VL has recently emerged as a major opportunistic infection in HIV and malaria patients and accelerates HIV to full-blown AIDS, and if left untreated, VL has a mortality rate exceeding 95% (Andre´s F. Henao-Mart´ınez, 2022; Tadesse Hailu, 2016; Tarnas et al., 2024).

1.1.9 Global public health impact

Leishmaniasis represents a major global public health concern, affecting over 12 million individuals across 98 countries or prefectures. The disease has an annual incidence rate of 0.9 to 2 million new cases, with mortality rates ranging between 20,000 and 50,000 deaths annually (Trájer and Kniha, 2025).

The World Health Organization (WHO) estimates that leishmaniasis remains a significant public health concern, affecting millions globally. Around 350 million individuals live in regions where the disease is prevalent, with an estimated 2 million new cases annually. Of these, approximately 500,000 are classified as visceral leishmaniasis, while 1.5 million cases involve the cutaneous form.

Visceral leishmaniasis is responsible for more than 50,000 deaths each year, making it one of the deadliest parasitic diseases, second only to malaria. The disease also results in the loss

of approximately 2,357,000 disability-adjusted life years (DALYs), ranking it ninth among infectious diseases globally (Organization, 2010).

According to Schoenheit et al. (2018) the global burden of leishmaniasis, estimated to cost billions annually, is compounded by the rising incidence in non-endemic regions, particularly due to climate change and migration patterns.

The economic burden in the endemic areas is primarily driven by agricultural losses and healthcare costs, which are exacerbated by limited access to medical resources and poor sanitation.

A study conducted by Grifferty et al. (2021b) reports that the median cost to the health provider per VL patient in 7 health facilities surveyed in Morocco, Brazil, India, and Sudan was about $520 per patient, with the cost reducing on average from $636 to about $306 when care was on an outpatient basis.

In Sudan, a retrospective review of 250 medical records found that in 2008, three public hospitals showed that depending on the hospital, the medical cost per patient varied between $117 and $366 (Meheus et al., 2013).

Post-Kalaazar dermal leishmaniasis (a skin rash experienced by recovered VL patients) has also been known to cause social stigma and psychological distress, hence reducing quality of life (QoL)(Grace Grifferty and Wamai, 2021).

1.1.10 Regional economic and social impact

At the regional level, the World Health Organization (WHO) reported in 2022 that approximately 85% of global visceral leishmaniasis (VL) cases originated from seven countries: Brazil, Ethiopia, India, Kenya, Somalia, and South Sudan.

Leishmaniasis imposes significant social and economic burdens, particularly in endemic regions such as South Asia, Sub-Saharan Africa, and parts of Latin America. The financial impact of the disease is substantial, with direct costs including healthcare expenses that can

reach up to 300 USD, while the available resources for affected communities may be as low as 5 USD (van Dijk et al., 2024).

A systematic review by Grifferty et al. (2021b) examined the financial burden of visceral leishmaniasis, classifying costs into two primary categories: direct medical expenses and indirect costs. The study reported variations in treatment costs across different regions, with estimated per-patient expenses of approximately $760 in Sudan, $128 in Nepal, $197 in India, and $220 in Bangladesh. Additionally, the review noted substantial indirect costs due to lost productivity, with patients and caregivers in Nepal losing 57 days of work, economically active individuals in India losing 120 days, and patients in Sudan losing 51 days of productivity.

A study conducted by Schoenheit et al. (2018) highlighted that the cost of medical treatment, consultations, and hospital admissions significantly strains family resources to the extent of adopting unhealthy coping strategies such as seeking loans from loan sharks/informal outfits attracting huge interests Okwor and Uzonna (2016) and selling assets such as livestock and land, especially when treatment is prolonged.

The above discourse is also supported by Anoopa Sharma et al. (2006) who found out that patients requiring treatment for visceral leishmaniasis made a median of 7 visits to different healthcare service providers before starting treatment, with 79% reporting informal payments to access providers.

The financial burden on affected households was significant, with a median expenditure of approximately **US$87** per patient for direct medical costs and an estimated **US$40** in lost income. To manage these expenses, many households relied on different coping mechanisms. The most common strategies included selling or renting out assets, reported by about **62%** of households, and securing loans, which **64%** of families resorted to as a financial buffer.

Additionally, vector control programs, including insecticide spraying and the distribution of bed nets, contribute to the financial burden, costing governments and NGOs millions annually (Desjeux, 2004).

Indirect costs are perhaps even more substantial. The disease disrupts livelihoods, particularly in agricultural and pastoral communities, where workers experience prolonged illness, leading to a loss of productivity and income.

In rural areas, where farming is the primary source of income, workers suffering from visceral leishmaniasis may experience reduced crop yields and agricultural productivity, particularly in regions like Turkana County in Kenya (our study area), which relies heavily on farming and livestock. The disease could lead to economic instability as families are forced to divert resources toward healthcare, leaving them vulnerable to poverty.Below is a table showing the costs associated with VL.

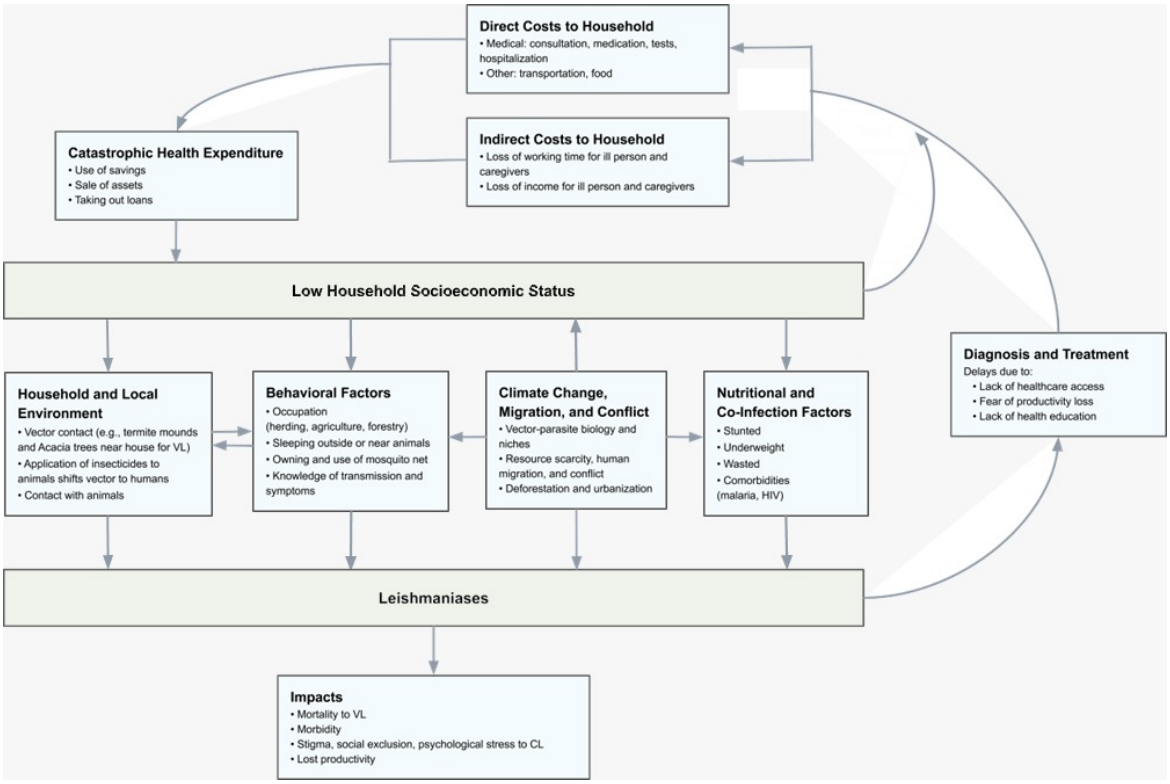


Figure 1.1: Social and Economic costs

1.1.11 Livestock losses and economic impact

In addition to human health, leishmaniasis also poses a significant threat to livestock populations, particularly in pastoral communities. Canine leishmaniasis, where dogs act as

reservoirs for the disease, can lead to reduced livestock productivity. Infected animals, such as dogs, often experience poor health, weight loss, and diminished strength, which directly impacts meat and milk production.

In areas where livestock is essential for economic survival, the loss of animals to leishmaniasis can be devastating, especially for pastoralist communities who rely on dogs for security.

A study by Elnaiem et al. (2006) highlighted that livestock losses in areas with prevalent canine leishmaniasis reduce food security and hinder economic growth.

In Kenya, pastoralists in regions like northern Kenya suffer from decreased income and food security due to the loss of animals to diseases transmitted by sandflies, further complicating the local economic landscape (Kama, 2016).

1.1.12 Social consequences: stigma and exclusion

The social impact of leishmaniasis is profound, particularly for those suffering from PKDL forms of the disease, which leave visible scars and lesions. The visible nature of the disease often leads to stigmatization, psychological distress, and social exclusion. According to individuals with disfiguring lesions, they face anxiety, depression, and social isolation, especially in communities where physical appearance is strongly tied to social status (Alvar et al., 2006). This stigma can affect employment opportunities, education, and social participation.

A study by Moore (2017) showed that children with leishmaniasis, for example, are at higher risk of being ostracized at school, resulting in lower educational attainment and fewer opportunities for future employment.

The social burden also falls disproportionately on women, who are often expected to care for sick family members. In societies where women are the primary caregivers, the financial and emotional toll of caring for a loved one with leishmaniasis can lead to lost income-generating opportunities, putting women at greater risk of financial hardship and social marginalization (Ali et al., 2020).

1.1.13 Local impact: Kenya and East Africa

At the local level, the social and economic consequences of leishmaniasis are particularly severe in regions like Kenya, where the disease is endemic in areas such as Turkana County. In Kenya, leishmaniasis disproportionately affects rural and peri-urban communities, where poor housing conditions and inadequate sanitation provide ideal breeding grounds for sandflies. A few studies in Kenya on VL have been keen on collecting the costs (either direct or indirect costs);

For instance, a study by Kasili and Okindo (2016) found out that the direct cost on medical care in Baringo County for VL-affected patients was $259.83 per household, with around 50.26% expending at least US$ 200. In terms of productivity, an average of 178 lost days, was reported.

In Ethiopia, research on visceral leishmaniasis (VL) treatment expenses indicates that non-medical costs tend to be significantly higher than direct medical expenses. A study by Tessema et al. (2024) estimated that the average non-medical expenditure per patient was approximately USD 101, whereas medical costs amounted to around USD 30. Furthermore, findings revealed that a substantial proportion (81%) of VL patients experienced catastrophic health expenditures, with out-of-pocket payments exceeding 10% of their annual household income.

1.1.14 Challenges in mitigating Leishmaniasis economic and social impact

Efforts to mitigate the economic and social impact of leishmaniasis face significant challenges, especially in low-resource settings.

Public health initiatives, such as vector control programs and environmental management, have shown some success, but they remain costly and often unsustainable without adequate funding. In Kenya, for instance, local governments face challenges relating to weak surveillance systems, lack of a dedicated budget in endemic counties for VL control, diagnosis and

treatment, stockouts of medicine in the supply chain, and a lack of personnel to carry out diagnostic/screening tests (Ministry of Health, 2021).

At the stakeholder level, knowledge gaps manifested by limited community awareness of the roles different individuals play in Kala-azar prevention, control, and elimination.

In addition, there is limited research to understand the plant-feeding behaviors of Afrotropical species of sandflies to help inform sustainable and less costly vector control programs.

Hassaballa et al. (2021) posits that resource constraints and lack of comprehensive approaches in coordination and in implementing control measures.

Furthermore, climate change and urbanization complicate efforts to control the disease, as these factors alter the habitat and distribution patterns of sandflies, leading to increased transmission in new areas. The emergence of insecticide resistance in sandfly populations further hinders vector control efforts, making it essential to find new, more sustainable strategies for disease prevention and management (Alvar et al., 2006).

1.1.15 Global,regional and local successes

Over the years, a number of success stories have been recorded in Asia, where Bangladesh was able to eradicate leishmaniasis. Also in 2021, India's National Kala-azar Elimination Programme (NKAEP) reported having reduced VL incidence to below one case per 10,000 population per year in 625 of 633 endemic blocks across four states (Subramanian et al., 2024).

Kenya has made significant strides in combating visceral leishmaniasis (VL) with the implementation of its strategic plan for the control of VL from 2021 to 2025. The plan aims to reduce VL-related morbidity by 60% and decrease the frequency of leishmaniasis outbreaks in the country by 50% by the year 2025.

1.2 Statement of the problem

Kala-azar (visceral leishmaniasis) is a zoonotic and anthroponotic disease caused by sandflies, with about 50,000–90,000 new cases annually. In Kenya, 4,000 cases occur yearly, affecting 5 million people. Eastern Africa bears 66% of global VL cases, with 53% in children under 15 and one-third in women (Alvar et al., 2023).

Despite regional elimination efforts, including the WHO's 2020 East African strategy Alvar et al. (2021), most studies focus on isolated factors like community awareness and environmental influences, often neglecting the complex interactions of these factors.Environmental and climatic factors, such as seasonal rainfall and temperature, significantly influence the transmission of visceral leishmaniasis (VL) (Talbi et al., 2022).

Current knowledge on the distribution of VL and its vectors in Kenya is primarily derived from cross-sectional surveys and sandfly collections in accessible areas, with limited predictive insights on potential transmission hotspots.

However, the combined effects of comprehensive ecotypes, vegetation health (e.g., normalized difference vegetation index), geospatial factors (latitude, longitude, altitude), and weather conditions/meteorological factors (humidity, temperature, and rainfall) on the presence of VL-causing sandflies remain poorly understood. This study aims to examine the interactions among these factors and their influence on sandfly distribution.

1.3 Research objectives

1.3.1 General objective

The objective of this study is to assess the performance of spatial regression and machine learning models in the prediction of Phlebotomine sandflies in Turkana County, Kenya.

1.3.2 Specific objectives

1. To evaluate the relationship between phlebotomine sandfly presence and average Evapotranspiration.
2. To evaluate the relationship between phlebotomine sandfly presence and Wetness.
3. To evaluate the relationship between phlebotomine sandfly presence and average annual temperature.
4. To evaluate the relationship between phlebotomine sandfly presence and Normalized difference vegetation index (NDVI).
5. To evaluate the relationship between phlebotomine sandfly presence and average annual humidity.
6. To evaluate the relationship between phlebotomine sandfly presence and ecotypes.
7. To evaluate the relationship between phlebotomine sandfly presence and spatial effects.

1.4 Justification of the Study

This study highlights the significance of environmental characteristics, climatic conditions, geographical factors, and vegetation health (NDVI) in determining the presence of VL-causing vectors in Turkana County by identifying and understanding these influencing factors are essential for designing and implementing effective control, prevention, and eradication strategies.

Conventional disease surveillance and control methods are often reactive, with interventions typically initiated only after an outbreak has already spread. The Kenyan Strategic Plan for Leishmaniasis Control (2021–2025) underscores the importance of enhancing surveillance, monitoring and evaluation, and operational research to support evidence-based decision-making. Achieving these goals requires the development of advanced predictive models

capable of forecasting high-risk areas and future trends using readily available environmental and climatic data.

Adopting a proactive modeling approach offers a cost-effective alternative to relying solely on epidemiological data, which is both expensive to collect and often obtained only after disease occurrence. By integrating predictive modeling techniques, interventions can be more timely and precisely targeted, enhancing their overall effectiveness.

Spatial regression models and machine learning techniques are widely used frameworks for modeling infectious diseases, each with its own strengths and limitations. These modeling approaches incorporate a comprehensive understanding of environmental, demographic, and climatic variables that influence the presence of *Phlebotomine* sandflies. Previous studies have primarily relied on statistical association tests, such as chi-square, and descriptive statistics, including frequency counts, to explore the relationship between VL risk factors. For example, Stauch et al. (2011), in a study titled *Visceral Leishmaniasis in the Indian Subcontinent: Modelling Epidemiology and Control*, used ordinary differential equations to estimate sandfly density. However, this model lacked a comprehensive perspective, as it did not account for climatic and environmental variables or the migration patterns of hosts, both of which are critical in determining VL vector density (Joshi et al., 2009).

This study addresses these gaps by conducting a causal analysis of how climatic factors influence the presence of VL vectors in Turkana County. By applying spatial and machine learning models to real world vector abundance data, this research assesses which modeling approaches and factors are most relevant for predicting the presence of VL causing vectors. Such insights are particularly valuable for public health officials and policymakers in Kenya, where leishmaniasis remains a significant health challenge (Pigott et al., 2014). The findings of this study hold practical implications for disease control efforts in Turkana County and similar endemic regions.

Accurately predicting the presence and absence of VL vectors is crucial for optimizing resource allocation in prevention and control programs, enabling strategic intervention planning and the efficient deployment of healthcare resources. Additionally, this study

contributes to the broader field of vector modeling by offering insights into the trade-offs between model complexity and predictive accuracy in vector-borne disease dynamics.

In conclusion, this study is justified not only by the pressing need to combat a critical public health challenge in Kenya but also by its potential to advance the field of vector-borne disease modeling, ultimately aiding in the development of more effective disease control strategies.

1.5 Significance of the study

This study is significant in advancing the discourse on the use of Spatial econometric modelling and machine learning applications on disease vector modeling and the understanding of the dynamics of the VL-causing sandfly population in Turkana County and offers valuable information on which modeling techniques might be effective in predicting the prevalence of VL-causing vectors given a set of conditions.

This research will be relevant to multiple stakeholders who are key in the operationalization of the promising OneHealth approach, such as public health authorities, ecological researchers, demographic experts, and policymakers who seek to mitigate the spread of diseases like leishmaniasis, which are transmitted by sandflies.

Furthermore, the findings will inform the design of more targeted and effective prevention and control strategies in areas where sandfly populations pose a significant public health risk.

Chapter 2

Literature review

2.1 Introduction

Leishmaniasis is a vector-borne disease caused by *Leishmania* parasites, which are mainly transmitted through the bites of sandflies belonging to the Phlebotominae subfamily. It remains a significant public health concern in tropical and subtropical regions across the globe (Alvar et al., 2020).

Despite significant advances in understanding its pathogenesis and epidemiology, the persistence of leishmaniasis in endemic areas compounded by socio-economic, ecological, and biological factors demands the adoption of multidimensional theoretical frameworks for comprehensive disease control.

In this review, we examine four theoretical frameworks, the epidemiological triad, One Health, social determinants of health (SDH), and the social-ecological model (SEM) detailing their strengths, weaknesses, assumptions, and applications in leishmaniasis control.

Additionally, we explore statistical methods that can predict the abundance of VL-causing vectors based on geographical, climatic, and environmental characteristics, thereby strengthening surveillance efforts and enabling more targeted interventions.

2.2 Theoretical review

Generalized linear models are a framework or a family of statistical distributions that extend the linear regression to allow for modeling of response variables that do not follow a normal distribution or with error distribution models that are non-normal. Generalized linear models

can not only allow inference of responses that follow a normal distribution but also those that follow Poisson, gamma, binomial, or at least those that belong to the exponential family of distributions (McCullagh, 1984).

This is made possible through what we call a link function, which is a transformation applied to the mean of the response variable.

The choice of the link function is dependent on the nature of the response variable. The generalized linear model takes the below form

$$Y \sim f(\mu) = g^{-1}(\eta)$$

. In this model, x is regarded as fixed and the primary objective is to investigate the relationship between

$$E(Y) = X\beta = \mu = \eta$$

For instance, if the response variable is continuous, then the link function is identity. Moreover for a binary response, the link function is a logit, on the other hand if the response is categorical the link function is

2.2.1 Poisson regression

Poisson regression, a type of generalized linear model, is commonly used for modeling count data by establishing a relationship between a count-based response variable and one or more independent variables. The log link function serves as the canonical link for Poisson regression and is represented as:

$$\log(\lambda) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k,$$

where λ represents the expected count or rate (i.e., $E(Y)$), and $x_1, x_2, \ldots, x_k$ are the predictor variables. This formulation allows us to model the expected number of events (such as the

number of VL cases) by transforming the linear predictor through the exponential function:

$$\lambda = \exp\left(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k\right).$$

This model has been used, for instance, to predict the number of visceral leishmaniasis (VL) cases in China following the rollout of vector control initiatives (Hao et al., 2024).

2.2.2 Multinomial regression

Multinomial regression is a generalized linear model (GLM) method used to model a categorical response variable with more than two categories. For example, it can be applied to model the distribution of various VL-causing species as a function of a set of explanatory variables. The model uses the softmax link function, which converts a set of linear predictors into probabilities that sum to one. The softmax function is presented as follows:

$$P(Y = j \mid x) = \frac{\exp\left(\beta_{j0} + \beta_{j1} x_1 + \cdots + \beta_{jk} x_k\right)}{\sum_{l=1}^{J} \exp\left(\beta_{l0} + \beta_{l1} x_1 + \cdots + \beta_{lk} x_k\right)}, \quad j = 1, 2, \ldots, J,$$

where J is the number of outcome categories. This formulation estimates the probability of each category, allowing researchers to understand how explanatory variables influence the likelihood of observing each VL-causing species.

2.2.3 Logistic regression

Logistic regression is a type of generalized linear model (GLM) commonly used for predicting binary outcomes, such as success or failure, or yes and no responses (McCullagh, 1984). It estimates the probability of an event occurring by employing the logit function to establish a relationship between the linear predictor and the probability. This method has been widely applied in various studies, including malaria transmission modeling (Ayele et al., 2024) and exploring environmental and socioeconomic factors influencing visceral and cutaneous leishmaniasis (Valero et al., 2021).

In the GLM framework, the response variable Y is assumed to follow a binomial distribution, and its probability $p = P(Y = 1)$ is linked to a linear combination of predictors through the logit function:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon_k.$$

This formulation models the probability that an observation falls into one of the two categories by transforming the linear predictor into a probability using the inverse logit function. Logistic regression has been effectively applied in various fields; for instance, it has been used by Ayele et al. (2024) in malaria transmission modeling and by Valero et al. (2021) in investigating environmental and socioeconomic risk factors for visceral and cutaneous leishmaniasis.

2.2.4 The Epidemiological triad

The epidemiological triad is a foundational model in public health that conceptualizes disease transmission as an interaction between three critical components: the host, the agent, and the environment. Each element plays a vital role in the persistence and spread of diseases like leishmaniasis. This model is particularly effective in identifying intervention points, such as controlling sandfly vectors, improving host immunity, or modifying environmental conditions. Notably, the model highlights the role of environmental changes, such as deforestation, in disease spread (Tesh and colleagues, 2019).

Despite its usefulness, the epidemiological triad has limitations. It oversimplifies disease dynamics by adopting a linear approach that ignores the complex feedback loops between socio-economic status, behavioral practices, and environmental shifts. Additionally, it does not account for human behavior, socio-cultural factors, or healthcare access, all of which are crucial in mitigating disease spread (Ranjan et al., 2020).

The model assumes that host, agent, and environment can be manipulated independently to influence transmission, often overlooking their intricate interconnections.

Nevertheless, this model has been instrumental in guiding vector control programs. Efforts such as insecticide use, bed nets, and environmental management have yielded positive results in reducing leishmaniasis transmission (Killick-Kendrick, 2000).

However, the model's primary shortcoming lies in its inability to address the social and economic factors fueling transmission. Controlling sandfly populations alone proves insufficient when poverty, poor sanitation, and limited healthcare access continue to contribute to disease persistence (Murray et al., 2015).

2.2.5 One health approach

The One Health approach provides an integrated framework that highlights the linkages between human, animal, and environmental health. It recognizes that the transmission dynamics of leishmaniasis are influenced by the interactions between humans, domestic animals (such as dogs), and sandfly vectors. By encouraging collaboration across various disciplines, this approach fosters the development of holistic strategies for disease control, incorporating insights from veterinary, medical, and environmental sciences (Tesh and colleagues, 2019).

One of the major strengths of the One Health approach is its ability to address multiple transmission pathways simultaneously. Unlike the Epidemiological triad, it considers human-animal-environment interactions and emphasizes collaborative disease surveillance and intervention. However, its implementation can be challenging due to the need for coordinated efforts across various sectors. Moreover, funding and policy integration remain key obstacles in achieving optimal outcomes (Ranjan et al., 2020).

2.2.6 Social determinants of health (SDH)

The Social Determinants of Health framework focuses on the socio-economic and environmental factors that influence disease outcomes. In the context of leishmaniasis, SDH highlights how poverty, education levels, housing conditions, and access to healthcare affect

disease vulnerability. It shifts the focus from biological and environmental factors to broader systemic issues (Murray et al., 2015).

A significant strength of this framework is its emphasis on health equity and addressing root causes of disease transmission. Unlike the epidemiological Triad, which mainly emphasizes agent-host-environment interactions, SDH provides a deeper understanding of why certain populations are disproportionately affected. However, its primary limitation is the difficulty in implementing policy changes that address systemic socio-economic disparities. Interventions based on SDH require long-term governmental and institutional commitments (Alvar et al., 2020).

2.2.7 Social-ecological model (SEM)

The Social-Ecological Model provides a comprehensive perspective by examining individual, interpersonal, community, and societal influences on disease transmission. SEM is particularly useful in understanding how community behaviors, urbanization, and environmental modifications contribute to the spread of leishmaniasis (Tesh and colleagues, 2019).

This model's advantage lies in its multi-level approach, which facilitates targeted interventions at various levels, including individual health behaviors, community practices, and national policies. However, its complexity makes implementation challenging, as it requires integrated efforts across multiple domains (Ranjan et al., 2020).

2.3 Conceptual Framework

A conceptual framework integrates the various theoretical perspectives discussed, illustrating the interconnected factors influencing Leishmaniasis incidence. This framework encapsulates human activities, socio-economic conditions, environmental changes, and climatic factors as key determinants of disease transmission.

Figure 2.1 presents the conceptual framework for Leishmaniasis transmission, highlighting the interactions between climate, vectors, reservoirs, and human activities.

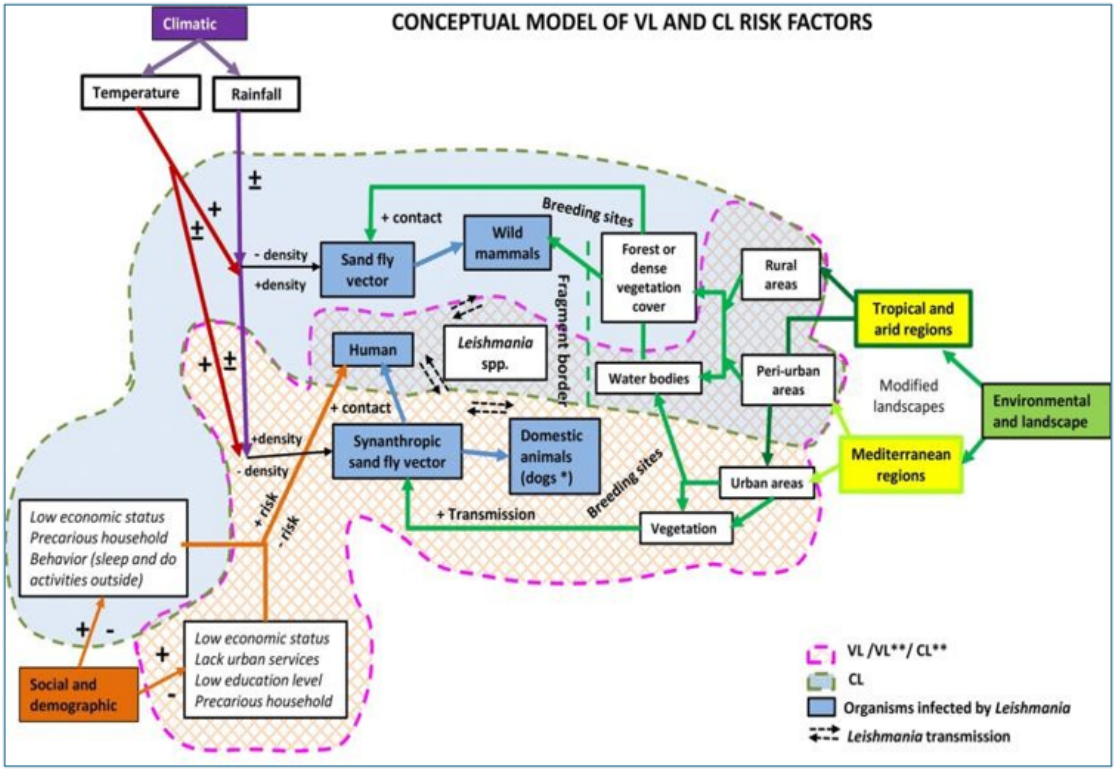


Figure 2.1: Conceptual framework for leishmaniasis transmission.

This framework underscores the need for integrative and multi-sectoral strategies to mitigate disease burden. By considering biological, environmental, and socio-economic factors together, it provides a holistic approach to disease control.

2.4 Empirical review relevant to the study

Understanding the environmental preferences and adaptive behaviors of sandflies is crucial for effective vector control and the elimination of leishmaniasis. A cross-sectional study by Gebre-Michael et al. (2004) identified key ecological determinants for *Phlebotomus martini* and *Phlebotomus orientalis*.

Logistic regression analysis revealed that land surface temperature (LST) during the dry season, altitude, mean annual temperature, and soil moisture were significant predictors of *P. martini* distribution, while *P. orientalis* was primarily influenced by LST annual composites. Spearman's rank correlation reinforced the importance of these factors, but soil type was not a determinant in East Africa, unlike in Sudan, where *P. orientalis* is associated with eutric vertisol (black cotton clay soil). The study also highlighted acacia and termite hills as critical habitats for Leishmania-transmitting sandflies.

Similarly, a study by Fernández et al. (2012) examined the seasonal distribution of sandflies in recently deforested areas of Argentina, emphasizing the role of climate and habitat in vector abundance. The research found that *Nyssomyia whitmani* and *Migonemyia migonei*, key vectors of American Tegumentary Leishmaniasis (ATL), were present year-round, with their populations strongly influenced by temperature and precipitation. Notably, sandfly abundance was significantly higher in pigsties compared to houses, indicating a high-risk environment for disease transmission.

Further, research by Philipo Gwandi et al. (2022) highlighted the socioeconomic and environmental factors that exacerbate leishmaniasis transmission. The study found that low levels of education, poverty, temporary housing, and proximity to termite hills and soil cracks increased the risk of infection. It recommended government collaboration with development partners to improve livelihoods and urged communities to eliminate dormant termite hills near residential areas as a preventive measure. This study corroborated the findings of Ferreira et al. (2022) who found out that precipitation and socio-economic proxies as important risk factors behind VL incidence.

A study conducted by Risueño et al. (2024) a multivariate logistic regression revealed that sand fly presence was independently and negatively related to both latitude and altitude, with the predicted probability decreasing as these factors increased.A bivariate analysis of the data revealed a significant inverse relationship between sandfly presence and latitude, suggesting that higher latitudes correspond to lower sandfly populations. However, no statistically significant association was observed between sandfly presence and altitude.

Geto et al. (2024) conducted a meta-analysis on human visceral leishmaniasis (VL) in Eastern Africa, estimating a pooled prevalence of 26.16% and highlighting various risk factors. These factors include demographic characteristics such as gender and age, as well as environmental and behavioral determinants like household size, proximity to termite mounds, presence of livestock, outdoor sleeping habits, and close contact with infected individuals. Additionally, living near water sources or frequently used pathways was associated with increased risk.

The study underscored the necessity of multifaceted intervention strategies in the region. These include strengthening public health education to enhance awareness about VL transmission, vectors, and breeding sites. Effective collaboration between health organizations, including the WHO, government bodies, non-governmental organizations, and local healthcare providers, is crucial for implementing sustainable control measures.

Furthermore, promoting early healthcare seeking behavior and adopting preventive measures such as minimizing outdoor exposure at peak sandfly activity times, wearing protective clothing, and utilizing insect repellents could play a significant role in reducing VL incidence in Eastern Africa.

A study by Valero et al. (2021) in Brazil revealed that municipalities with high vegetation cover tend to report a greater occurrence and number of CL cases, a pattern not observed for VL. This is attributed to the reliance of CL vectors and reservoirs on natural vegetation.

The studies discussed collectively highlight the complex and multifaceted nature of leishmaniasis transmission. Gebre-Michael et al. (2004) examined the role of land and temperature variables in East Africa while Fernández et al. (2012) emphasized the influence of seasonal and habitat factors in South America.

Additionally, Philipo Gwandi et al. (2022) also shed light on socioeconomic and structural risk factors on VL endemicity. Together, these findings underscore the critical need for integrated vector control strategies that account for climatic, ecological, socioeconomic, and public health determinants.A summary of these studies, along with identified research gaps, is presented in Table 2.1.

Table 2.1: Summary of studies on leishmaniasis research.

Focus of Study	**Findings**	**Research Gap**	**Focus of Current Study**
Risk factors for leishmaniasis in rural Marigat Sub-County, Kenya(Philipo Gwandi et al., 2022)	Presence of termite hills and soil cracks increased infection odds (multivariate regression).	Limited to Baringo County; excluded climatic factors and ecotypes.	Expanding to Turkana County, Kenya, with climatic/ecological analysis.
Climate/land use impact on sand fly density in Sri Lanka(Senanayake et al., 2024a)	Identified seasonal and land-use influences on vector populations.	Geographically restricted to Sri Lanka; longitudinal design.	Focus on Turkana County, Kenya, with cross-sectional spatial analysis.
Sandfly distribution along altitudinal gradients in Ethiopia(Alemayehu et al., 2024)	Altitude strongly correlated with *P. pedifer* density (highest in hyrax dwellings).	No focus on microclimatic or human behavioral factors.	Incorporating microclimate and human activity data in Turkana County.
Environmental drivers of sand fly density in Central Spain(McCullagh, 1984)	Higher densities linked to rural habitats, livestock presence, and unpaved roads.	Limited to Spain; excluded ecological interactions.	Kenyan focus with integrated environmental-ecological analysis.

Continued on next page

Table 2.1 – *Continued from previous page*

Focus of Study	**Findings**	**Research Gap**	**Focus of Current Study**
Ecological niche modeling of *Lutzomyia longipalpis* in Brazil(Fonseca et al., 2021)	Precipitation, temperature seasonality, and altitude predicted vector distribution.	Relied on remote sensing (NDVI/LST) without ground validation.	Ground-truthed remote sensing data in Turkana County, Kenya.
Socio-economic and Environmental determinants of leishmaniasis in Brazil(Valero et al., 2021)	Municipalities with high vegetation cover reported higher occurrence and number of CL cases, while VL was not significantly affected.	Limited investigation on VL; the dependency of CL and VL vectors on natural vegetation was not extended to other influencing eco-types/factors.	Integrating additional ecological variables in leishmaniasis research, with an emphasis on Kenyan regions.

Continued on next page

Table 2.1 – *Continued from previous page*

Focus of Study	**Findings**	**Research Gap**	**Focus of Current Study**
Environmental and geographic factors influencing sand fly presence using logistic regression(Risueño et al., 2024)	Bivariate analysis indicated a significant negative association with increasing latitude only, while multivariate analysis revealed that sand fly presence was independently and negatively associated with both latitude and altitude.	The discrepancy between bivariate and multivariate findings highlights the complexity of geographic influences on sand fly distribution, suggesting potential confounding factors that require further investigation.	Expanding environmental and geographic analyses to additional regions to better understand the determinants of sand fly presence.

2.4.1 Current Research

This study aims to model the presence of female sandflies in Turkana County, Kenya, by comparing spatial models and machine learning techniques. By identifying the most effective approach, the research addresses a gap in the literature on using mixed methods to analyze the dynamics of the leishmaniasis vector in low-resource settings. The findings will provide essential insights for predicting disease outbreaks, assessing risk, and optimizing resource allocation for public health interventions in Turkana County.

Chapter 3

Methodology

3.1 Research methodology

3.1.1 Introduction

The chapter presents the research methodology that will be adopted in the study. It covers the research design, population of the study, sampling design and technique, data collection method, operationalization of study variables, data analysis and ethical considerations. The research methodologies that will be used to collect and analyse data to answer the research questions and meet the objectives earlier outlined

3.1.2 Research design

This study will utilize a cross-sectional design to collect data on the various species of sandflies in Turkana County, Kenya, as set out in the ICIPE study design protocol for the End fund project. A cross-sectional study is a research design that involves the collection of data at a single point in time, with the objective of estimating specific parameters of interest within a defined population. This study will provide a snapshot of the infection rate of leishmaniasis from sandfly populations in Turkana County, allowing for an assessment of the factors influencing their distribution, abundance, and diversity at a particular moment.

The cross-sectional design is particularly useful for estimating the prevalence of a phenomenon (such as the population density of sandflies) within a defined time frame (Babbie, 2016). In the case of vector-borne diseases like leishmaniasis, understanding the current status of vector populations can provide valuable insights into disease risk and potential

control strategies (Alvar et al., 2012). The cross-sectional nature of this study allows for the identification of temporal patterns in sandfly populations, as well as correlations between environmental variables and vector presence.

In cross-sectional studies, the sample is typically drawn from a larger population, with random sampling techniques ensuring that each unit in the population has an equal chance of being selected (Gerritsen and Clements, 2014). This is critical in achieving generalizable results. For the proposed study, random sampling will be employed to select appropriate sampling locations across different ecological zones in Turkana County, ensuring a comprehensive representation of the area. The study assumes that the sampled units are independent, meaning that the data collected from one location does not influence the data from another. This assumption is important for the integrity of the statistical analyses, as violations could lead to biased results (Cohen, 2003).

Furthermore, a cross-sectional design can provide useful estimates of how environmental factors, such as temperature, humidity, and vegetation type, influence sandfly populations at the time of data collection. This approach aligns with previous research on vector ecology, which has demonstrated that sandfly populations fluctuate in response to environmental and climatic variables (Desjeux, 2004). By employing a well-structured cross-sectional design, the study will contribute to the growing body of knowledge on sandfly ecology and its implications for vector control programs in endemic regions (Khalil and El-Beshbishy, 2018).

3.1.3 Population of the study

A population refers to the entire group or set of individuals, items, or elements that a researcher intends to study or from which a sample will be drawn (Andrade, 2021). In research, the population represents the complete set of subjects or units that meet specific criteria within a given boundary or context at a particular point in time (Majid, 2018). The target population for this study consists of the sandfly populations in Turkana County, Kenya, which serve as the focus of the investigation.

For this study, the sandfly population in Turkana County is defined as the total group of sandfly species present within the geographic boundaries of the county. This encompasses all ecological zones of Turkana, including both urban and rural areas, as well as various environmental settings such as agricultural land, forested regions, and human settlements.

Turkana County, located in central Kenya, is known for its diverse ecological zones, which provide varied habitats for sandfly species. The sandfly population in this area plays a crucial role in the transmission dynamics of vector-borne diseases, particularly leishmaniasis, and understanding its distribution and density is critical for disease control and prevention efforts (Gingrich and Pugh, 2017).

The county is characterized by diverse climatic and environmental conditions that can influence the presence and abundance of sandflies. The county is warm and hot, characterized by temperatures ranging from 20-30.5°C (average annual temperatures) and 200 mm annual rainfall, with the rains mostly distributed in the east-west gradient, with the west experiencing relatively high rainfall amounts.As a result, the study aims to assess sandfly populations across different environmental gradients, including areas with varying levels of vegetation, proximity to water sources, and human activity.This broad definition of the target population allows for the identification of key environmental factors that may affect sandfly distribution and provides a comprehensive understanding of the dynamics of vector and infection rates in the region.

3.1.4 Sample size calculation

The sample size determination for this study on sandflies in Turkana County, Kenya, will be that set out and elaborated by ICIPE End Fund project study design protocols.

3.1.5 Data collection methods

As per the ICIPE study design protocol, the sandflies will be collected from caves using CDC light traps and preserved in 70% ethanol. The specimens will be dissected, with the

head and terminal segments used for morphological identification, while the remainder of the specimens will be stored for molecular analysis. DNA will be extracted from the ethanol-fixed specimens and subjected to PCR amplification for further analysis.

To confirm species identification, the mitochondrial cytochrome c oxidase subunit 1 (COI) gene will be amplified, purified, and sequenced. Phylogenetic analysis will be performed using sequence alignment and maximum-likelihood methods to classify the species.

For detecting Leishmania parasites, sandfly DNA will be screened using a nested kDNA PCR assay targeting kinetoplast DNA. If initial results are negative, high-sensitivity qPCR-HRM will be employed for confirmation. Individual sandflies suspected of carrying the parasite will be tested using PCR amplification of the internal transcribed spacer 1 (ITS1) region to detect Leishmania infection.

Species identification will be confirmed through phylogenetic analysis of the COI sequences, and Leishmania infections will be determined using molecular techniques.

Data on the infected sandflies will be recorded over time. Environmental data, including temperature, humidity, and precipitation, will also be mined from satellite data to assess potential environmental influences on sandfly populations.

3.2 Overview of methods

This study integrates spatial econometric models and machine learning techniques to examine the distribution patterns of visceral leishmaniasis (VL) vectors. The spatial econometric models applied include the **Spatial Error Model (SEM)**, **Spatial Lag Model (SLM)**, and **Spatial Autoregressive Moving Average (SARMA)** model, which are tailored to capture varying spatial dependencies.

To complement these models, machine learning techniques such as **XGBoost**, **k-nearest neighbors (KNN)**, and **spatial random forest** are utilized. Additionally, **logistic regression** is included as a reference classifier. By leveraging these methodologies, the study aims to

enhance predictive accuracy and provide a deeper understanding of the spatial distribution of VL vectors.

3.2.1 Spatial error model (SEM)

The spatial error model (SEM) accounts for spatially correlated errors, which often arise due to unobserved factors influencing vector distribution. The SEM is specified as follows:

$$y = X\beta + u, \quad u = \lambda Wu + \varepsilon, \tag{3.1}$$

where y represents the dependent variable, X is the matrix of covariates, β is the vector of regression coefficients, W is the spatial weight matrix, λ is the spatial error parameter, and $\varepsilon \sim N(0, \sigma^2 I)$ represents the error term. The model is estimated using Maximum Likelihood Estimation (MLE), ensuring that spatially structured errors are appropriately accounted for.

The SEM has been effectively utilized in vector-borne disease studies where unmeasured environmental or socio-economic factors influence disease distribution. For instance, da Paixão Sevá et al. (2023) applied SEM to analyze VL and tegumentary VL incidence in Brazil, revealing that Brazilian Savannah was negatively associated with VL.

3.2.2 Spatial lag model (SLM)

The Spatial Lag Model (SLM) explicitly incorporates spatial dependence in the dependent variable, making it suitable for understanding the propagation of disease across adjacent regions. The model is expressed as:

$$y = \rho Wy + X\beta + \varepsilon, \tag{3.2}$$

where ρ is the spatial lag parameter, measuring the extent to which the disease burden in a given location is influenced by neighboring regions. The SLM is typically estimated using MLE to account for the spatial structure in the response variable.

This model has been particularly useful in identifying disease hotspots and their spatial spillover effects. For example, Rotejanaprasert et al. (2021) employed the SLM to examine Malaria incidence in Greater Mekong Subregion uncovering significant spatial clustering of cases. Their study demonstrated that villages with high malaria incidence were surrounded by other high-incidence villages, reinforcing the need for spatially targeted vector control interventions.

3.2.3 Spatial autoregressive moving average (SARMA) model

The Spatial Autoregressive Moving Average (SARMA) model extends traditional spatial autoregressive models by incorporating both spatial lags and moving average components. The model is specified as:

$$y = \rho W y + \theta W \varepsilon + X\beta + \varepsilon, \tag{3.3}$$

where θ represents the spatial moving average parameter. The SARMA model is particularly useful in capturing both spatial and temporal dependencies in disease data. Estimation techniques for SARMA often include Bayesian inference or the Generalized method of moments (GMM).

SARMA has been employed in disease surveillance systems due to its ability to provide early warning signals for outbreaks. Patel et al. (2023) utilized SARMA to assess malaria risk in India, demonstrating its applicability to leishmaniasis control. By integrating environmental and epidemiological data, their approach improved the accuracy of early warning systems, allowing for timely intervention measures.

3.3 Machine learning methods

To enhance predictive performance, this study incorporates machine learning algorithms capable of capturing complex spatial relationships in vector distribution.

3.3.1 XGBoost

Extreme Gradient Boosting (XGBoost) is an advanced ensemble learning method that enhances decision trees through sequential improvements. It systematically reduces prediction errors by optimizing a predefined loss function at each iteration. The model effectively balances bias and variance, making it well-suited for handling structured datasets and complex relationships.

$$L(\Theta) = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i) + \sum_{k} \Omega(f_k), \tag{3.4}$$

where ℓ represents the error term, and $\Omega(f_k)$ is a regularization function to control model complexity and prevent overfitting.

XGBoost has been extensively used in vector surveillance studies due to its superior predictive capabilities. In a study on malaria risk assessment, Johansson et al. (2021) demonstrated that XGBoost outperformed traditional regression models in predicting disease hotspots based on climatic and ecological variables.

3.3.2 k-nearest neighbors (KNN)

The k-Nearest Neighbors (KNN) algorithm is a non-parametric classifier that assigns class labels based on the majority vote of the k closest neighbors in the feature space. The prediction is computed as:

$$\hat{y} = \arg\max_{c} \sum_{i \in N_k} \mathbb{I}(y_i = c), \tag{3.5}$$

where N_k represents the set of the k nearest neighbors, and $\mathbb{I}$ is an indicator function.

KNN has been widely applied in epidemiology for classification tasks, particularly in predicting disease presence based on spatial covariates. Smith and Brown (2020) applied KNN to predict dengue transmission zones, demonstrating its utility in vector control strategies.

3.3.3 Spatial random forest

The Spatial Random Forest model extends traditional Random Forest by incorporating location-based features to enhance prediction accuracy. The model aggregates the outputs of multiple decision trees:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(X, W), \tag{3.6}$$

where h_t represents individual trees trained on bootstrapped samples of X, while W introduces spatial proximity features.

Random Forest models have been successfully applied in vector-borne disease research. Rivera et al. (2022) used spatially aware Random Forest models to analyze the distribution of cutaneous leishmaniasis in the Amazon region, incorporating remote sensing data to improve model accuracy.

3.3.4 Logistic regression

Logistic regression serves as a baseline classifier, estimating the probability of vector presence as:

$$P(y = 1|X) = \frac{1}{1 + e^{-(X\beta)}}. \tag{3.7}$$

The coefficients β are estimated using MLE, with predictor selection guided by the Akaike Information Criterion (AIC). While simpler than other models, logistic regression remains a valuable benchmark for assessing the performance of more complex algorithms.

3.4 Derivation of spatial models in vector control

3.4.1 Spatial error model (SEM)

The SEM accounts for spatially correlated errors in regression models, which is crucial when unobserved factors influencing vector control outcomes exhibit spatial dependence.

Model specification

The Spatial Error Model (SEM) accounts for spatial dependence in the error term rather than the dependent variable. The general form of the model is given by:

$$y = X\beta + \lambda W\varepsilon + \varepsilon \tag{3.8}$$

where:

- y is the dependent variable (e.g., vector density or disease incidence),
- X is the matrix of independent variables,
- β represents the regression coefficients,
- W is the spatial weight matrix capturing spatial relationships,
- λ is the spatial error parameter,
- $\varepsilon \sim N(0, \sigma^2 I)$ is the error term.

Rewriting the error term:

$$\varepsilon = (I - \lambda W)^{-1} u \tag{3.9}$$

where $u \sim N(0, \sigma^2 I)$. Substituting this into the main equation:

$$y = X\beta + (I - \lambda W)^{-1} u. \tag{3.10}$$

Thus, the variance-covariance structure of y becomes:

$$\text{Var}(y) = \sigma^2 (I - \lambda W)^{-1} (I - \lambda W)^{-T}. \tag{3.11}$$

Since y follows a multivariate normal distribution, the probability density function is:

$$f(y|\beta, \lambda, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} |I - \lambda W| \exp\left(-\frac{1}{2\sigma^2} (y - X\beta)^T (I - \lambda W)^T (I - \lambda W)(y - X\beta) \right). \tag{3.12}$$

Taking the log-likelihood function:

$$\log L(\beta, \lambda, \sigma^2) = -\frac{N}{2} \log(2\pi\sigma^2) + \log|I - \lambda W| - \frac{1}{2\sigma^2} (y - X\beta)^T (I - \lambda W)^T (I - \lambda W)(y - X\beta). \tag{3.13}$$

The parameters β, λ, and σ^2 are estimated as follows:

1. **Estimate** λ by maximizing the log-likelihood function:

$$\hat{\lambda} = \arg\max_{\lambda} \left(\log|I - \lambda W| - \frac{1}{2\sigma^2} (y - X\beta)^T (I - \lambda W)^T (I - \lambda W)(y - X\beta) \right). \tag{3.14}$$

   This requires numerical optimization as the determinant term $\log|I - \lambda W|$ is computationally complex.

2. **Estimate** β by solving:

$$\hat{\beta} = (X^T(I-\lambda W)^T(I-\lambda W)X)^{-1}X^T(I-\lambda W)^T(I-\lambda W)y. \tag{3.15}$$

3. **Estimate** σ^2 using:

$$\hat{\sigma}^2 = \frac{(y-X\hat{\beta})^T(I-\lambda W)^T(I-\lambda W)(y-X\hat{\beta})}{N}. \tag{3.16}$$

3.4.2 Spatial lag model (SLM)

The estimation of parameters in the Spatial Lag Model (SLM) is essential, as ignoring spatial dependence results in biased and inconsistent estimators. The model is given by:

$$y = \rho Wy + X\beta + \varepsilon, \tag{3.17}$$

where y is an $N \times 1$ vector of observations, X is an $N \times k$ matrix of explanatory variables, β is a $k \times 1$ vector of coefficients, W is an $N \times N$ spatial weights matrix, ρ is the spatial autoregressive parameter, and $\varepsilon \sim N(0, \sigma^2 I)$ is an error term with constant variance.

Rewriting the equation:

$$(I-\rho W)y = X\beta + \varepsilon. \tag{3.18}$$

Taking expectations:

$$E[y] = (I-\rho W)^{-1}X\beta. \tag{3.19}$$

Since y appears on both sides, applying Ordinary Least Squares (OLS) directly leads to inconsistent estimates. To address this, Maximum Likelihood Estimation (MLE) is used.

Rewriting the error term:

$$\varepsilon = (I - \rho W)y - X\beta. \tag{3.20}$$

Since $\varepsilon \sim N(0, \sigma^2 I)$, y follows:

$$y \sim N((I - \rho W)^{-1} X\beta, \sigma^2 (I - \rho W)^{-1} (I - \rho W)^{-T}). \tag{3.21}$$

The log-likelihood function is:

$$\log L(\rho, \beta, \sigma^2 | y) = -\frac{N}{2} \log(2\pi\sigma^2) + \log|I - \rho W| - \frac{1}{2\sigma^2} (y - X\beta)^T (I - \rho W)^T (I - \rho W)(y - X\beta). \tag{3.22}$$

The MLE estimates are obtained as follows:

1. Estimate β:

$$\hat{\beta} = (X^T (I - \rho W)^T (I - \rho W) X)^{-1} X^T (I - \rho W)^T (I - \rho W) y. \tag{3.23}$$

2. Estimate ρ by maximizing the log-likelihood:

$$\hat{\rho} = \arg\max_{\rho} \left(\log|I - \rho W| - \frac{1}{2\sigma^2} (y - X\hat{\beta})^T (I - \rho W)^T (I - \rho W)(y - X\hat{\beta}) \right). \tag{3.24}$$

3. Estimate σ^2:

$$\hat{\sigma}^2 = \frac{(y - X\hat{\beta})^T (I - \rho W)^T (I - \rho W)(y - X\hat{\beta})}{N}. \tag{3.25}$$

The term $\log|I - \rho W|$ requires numerical optimization techniques such as eigenvalue decomposition or Taylor series expansion.

3.4.3 Spatial autoregressive moving average (SARMA) model

The SARMA model extends traditional autoregressive models by incorporating both spatial lags and moving average components. This allows for capturing spatial dependence in both the dependent variable and the error structure.

Model specification

The general SARMA model is given by:

$$y = \rho W y + \theta W \varepsilon + X\beta + \varepsilon \tag{3.26}$$

where:

- y is the $N \times 1$ vector of observations,
- X is the $N \times k$ matrix of independent variables,
- β is the $k \times 1$ vector of regression coefficients,
- W is the $N \times N$ spatial weight matrix,
- ρ is the spatial autoregressive parameter,
- θ is the spatial moving average parameter,
- $\varepsilon \sim N(0, \sigma^2 I)$ is the error term.

Rewriting the model by isolating y:

$$(I - \rho W)y = X\beta + \varepsilon + \theta W \varepsilon. \tag{3.27}$$

Substituting ε:

$$(I - \rho W)y = X\beta + (I + \theta W)\varepsilon. \tag{3.28}$$

The variance structure of y follows:

$$\text{Var}(y) = \sigma^2(I-\rho W)^{-1}(I+\theta W)(I+\theta W)^T(I-\rho W)^{-T}. \tag{3.29}$$

Since y follows a multivariate normal distribution, the probability density function is:

$$f(y|\rho,\theta,\beta,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}}|I-\rho W|\exp\left(-\frac{1}{2\sigma^2}(y-X\beta)^T\Omega^{-1}(y-X\beta)\right), \tag{3.30}$$

where the spatial covariance matrix Ω is:

$$\Omega = (I-\rho W)^{-1}(I+\theta W)(I+\theta W)^T(I-\rho W)^{-T}. \tag{3.31}$$

Taking the log-likelihood function:

$$\log L(\rho,\theta,\beta,\sigma^2) = -\frac{N}{2}\log(2\pi\sigma^2) + \log|I-\rho W| - \frac{1}{2\sigma^2}(y-X\beta)^T\Omega^{-1}(y-X\beta). \tag{3.32}$$

The parameters ρ, θ, β, and σ^2 are estimated as follows:

1. **Estimate ρ and θ** by maximizing the log-likelihood function:

$$\hat{\rho},\hat{\theta} = \arg\max_{\rho,\theta}\left(\log|I-\rho W| - \frac{1}{2\sigma^2}(y-X\beta)^T\Omega^{-1}(y-X\beta)\right). \tag{3.33}$$

   This requires numerical optimization as ρ and θ appear in both the determinant and the covariance matrix.

2. **Estimate β** by solving:

$$\hat{\beta} = (X^T\Omega^{-1}X)^{-1}X^T\Omega^{-1}y. \tag{3.34}$$

3. **Estimate** σ^2 using:

$$\hat{\sigma}^2 = \frac{(y - X\hat{\beta})^T \Omega^{-1} (y - X\hat{\beta})}{N}. \tag{3.35}$$

3.4.4 Exploratory data analysis (EDA)

Exploratory data analysis (EDA) is crucial for gaining insights into the structure of data before conducting more advanced analyses. It utilizes a range of statistical and graphical methods to examine data distribution, trends, and relationships.

A key aim of EDA is to uncover patterns that can inform future modeling and hypothesis testing (Tukey, 1977). It also helps identify anomalies, such as outliers or data entry errors, which could skew the study's results.

To visualize variable distributions, various graphical techniques will be applied, including boxplots, histograms, bar charts, and pie charts. Boxplots offer a visual summary of data spread, central tendency, and outliers, making them ideal for comparing distributions across different groups (McGill et al., 1978).

Histograms will help illustrate the frequency distribution of continuous variables, enabling the detection of skewness or multimodality (Friendly, 2007).

Bar charts and pie charts will present categorical data, highlighting the proportion of each category in an easily interpretable format.

Beyond distribution analysis, EDA plays a vital role in detecting outliers—data points that significantly differ from others. These outliers might result from data entry errors or could indicate rare yet important occurrences (Barnett and Lewis, 1994).

Properly identifying and managing outliers is crucial, as they can disproportionately affect statistical analyses and lead to inaccurate conclusions (Rousseeuw and Leroy, 1987).

In addition to traditional EDA, spatial descriptive analysis will be performed to explore the spatial characteristics of the dataset. Spatial descriptive analysis aims to identify patterns in

spatial distributions of VL vectors, detect spatial clustering or dispersion, and analyze the spatial relationships between data points.

This is particularly important when the data has a geographical or spatial component, as it helps uncover underlying spatial structures that could influence subsequent modeling.

Spatial autocorrelation will be assessed using two key techniques: the Global Moran's I statistic and the Getis-Ord Gi statistic.

Using ArcGis,the Global Moran's I statistic will provide an overall measure of spatial correlation, indicating whether data points tend to be clustered or dispersed across the study area (Blazquez et al., 2018).

Positive values suggest clustering, while negative values indicate dispersion. A value near zero suggests no spatial correlation.

The Getis-Ord G_i statistic will be used to identify spatial clusters of high or low VL vector values (hotspots and cold spots) by considering neighboring data points.

This statistic will help pinpoint areas with significant spatial patterns, indicating potential areas of interest for further investigation (Blazquez et al., 2018). Z-scores and P-values will be used to determine the statistical significance of the hotspots and cold spots.

These spatial descriptive analyses will provide additional insights into the spatial structure of the data, revealing how variables are distributed across geographical space and identifying regions that exhibit significant clustering. Understanding the spatial relationships in the data is essential for building accurate models and making informed conclusions, especially when spatial dependencies exist.

Through these EDA techniques, including both traditional and spatial analysis, we can gain a deeper understanding of the data, assess its quality, and make informed decisions about the appropriate statistical models and analyses to apply.

Following the insights from the EDA, the suitable model for fitting the data will be identified.

Next, the data will be split into training and testing sets, with the training set used to train the model.

After training, the model will be evaluated on the testing set to assess its prediction performance using statistical measures such as RMSE, which will help determine how accurately the model predicts the data (Burnham and Anderson, 2002).

3.4.5 Data quality

Data quality is a crucial aspect of ensuring the validity and reliability of research findings. The data collected for this study will undergo thorough checks for both completeness and validity. Completeness refers to the extent to which all expected data points have been recorded, while validity ensures that the data accurately reflects the real-world phenomena being measured. Missing or incomplete data can introduce bias and affect the robustness of the analysis, especially in studies involving ecological or epidemiological data where every data point may be critical to understanding disease transmission dynamics.

In the case of incomplete data, imputation techniques will be applied to replace missing values. Specifically, missing values will be imputed using the median of the respective variable. Median imputation is commonly employed in such studies because it is less sensitive to outliers than the mean, making it a reliable method for preserving the central tendency of the data without introducing significant skew (Allison, 2001). By imputing missing values instead of excluding cases with missing data, we minimize the risk of losing valuable observations, which could otherwise impact the statistical power of the analysis and reduce the generalizability of the results (Rubin, 1987). The imputed data will then be retained for analysis, and sensitivity analyses may be conducted to assess the potential impact of the imputation on the findings (Enders, 2010).

Chapter 4

Results and Interpretation

4.1 Introduction

This section presents the results of **parametric models** (SEM, SLM, and SARMA) used to examine the relationship between various climatic ,environmental and spatial factors influencing the occurrence of VL-causing vectors in Turkana County.

Similarly, non-parametric (machine learning) models, including **K-Nearest Neighbors (KNN), XGBoost**, and **Spatial Random Forest**, are also evaluated.

The best-performing model will be determined based on key evaluation metrics, including AIC,BIC,Log Likelihood and **Area Under the Curve (AUC), sensitivity**, and **specificity**.

4.2 Descriptive statistics of Turkana data collection sites,VL vector distribution across ecotypes

Data was collected from three sub-counties in Turkana, with Loima and Turkana south contributing **75.0% (83/110)** of the total VL-causing vectors .The distribution of the VL vectors was however proportional across the sub counties (*p-value = 0.291*).

The VL vector distribution across administrative village areas was also analysed and Moruege recorded the highest catch, accounting for **16.4% (18/110)**, followed by Nang'omo with **16.4% (18/110)**. Nakipi contributed **13.6% (15/110)**, while Kaidir and Lokwatuba each accounted for **9.1% (10/110)** of the total VL vector catch. The distribution of VL vectors across the sites was however significantly different (*p-value = 5.28716E-08****).

Further, **P. alexandri**, a vector for visceral leishmaniasis (VL), dominated the catch, accounting for **54.5% (60/110)**, while **P. orientalis** made up **32.7% (36/110)**. Other species were less common, with **P. duboscqi** accounting for **8.2% (9/110)**, **P. martini** contributing just **4.5% (5/110)**, and **P. saevus** accounting for only **3.6% (4/110)**. Overall, the distribution of VL-causing vectors was significantly non-uniform (*p-value = 9.999E-05****).

The VL causing vectors were analyzed based on the ecotypes from which they were collected. The largest proportion came from **termite mounds**, which accounted for **34.5% (38/110)**. This was followed by **river beds**, contributing **15.5% (17/110)**, and **animal sheds (goats)** with **13.6% (15/110)**. **Vegetation with palm** made up **10.0% (11/110)**, while **vegetation with acacia** contributed **6.4% (7/110)**. Other ecotypes included **vegetation with balanites** at **1.8% (2/110)**, **animal sheds (sheep)** at **1.8% (2/110)**, **indoors (sticks)** at **3.6% (4/110)**, **sleeping areas** at **2.7% (3/110)**, **rocks** at **0.9% (1/110)**, and **resting areas** at **3.6% (4/110)**. The presence of VL-causing vectors was significantly associated with ecotypes (*p-value = 0.00260***).

In terms of environmental variables, the relationship between **Mean NDVI** and VL causing vector was significant (**p-value = 0.00546**). Similarly, **Max average temperature** also exhibited a significant relationship (**p-value = 0.00546**). Additionally, **Wetness** showed a significant association (**p-value = 0.0202**). The above observations indicate that both maximum average temperature,Wetness,Mean NDVI influence the abundance of VL causing vectors.

Table 4.1: Comparison of Non-VL and VL causing vectors across different factors

Factors	**Non-VL causing vector (n%)**	**VL causing vector (n%)**	**P-value**
Soil Type			0.337
Sandy clay loam	445 (48.0%)	54 (5.8%)	
Sandy loam	372 (40.1%)	56 (6.0%)	

Continued on next page

Table 4.1 – *Continued from previous page*

Factors	**Non-VL causing vector (n%)**	**VL causing vector (n%)**	**P-value**
Species			0.000♯***
P. alexandri	36 (3.9%)	60 (6.5%)	
P. duboscqi	1 (0.1%)	9 (1.0%)	
P. martini	0 (0.0%)	5 (0.5%)	
P. orientalis	10 (1.1%)	36 (3.9%)	
P. saevus	4 (0.4%)	0 (0.0%)	
S. adleri	63 (6.8%)	0 (0.0%)	
S. affinis vorax	3 (0.3%)	0 (0.0%)	
S. africana	132 (14.2%)	0 (0.0%)	
S. antenata	42 (4.5%)	0 (0.0%)	
S. antenatus	16 (1.7%)	0 (0.0%)	
S. clydei	359 (38.7%)	0 (0.0%)	
S. schwetzi	139 (15.0%)	0 (0.0%)	
S. squamipleuris	10 (1.1%)	0 (0.0%)	
Sub-county			0.291
Loima	321 (34.6%)	44 (4.7%)	
Turkana Central	254 (27.4%)	27 (2.9%)	
Turkana South	242 (26.1%)	39 (4.2%)	
Site			0.000***
Kaidir	86 (9.3%)	10 (1.1%)	

Continued on next page

Table 4.1 – *Continued from previous page*

Factors	**Non-VL causing vector (n%)**	**VL causing vector (n%)**	**P-value**
Lokwatuba	45 (4.9%)	10 (1.1%)	
Moruege	87 (9.4%)	18 (1.9%)	
Nadoto	77 (8.3%)	14 (1.5%)	
Nakipi	26 (2.8%)	15 (1.6%)	
Nang'omo	88 (9.5%)	18 (1.9%)	
Nangitony	43 (4.6%)	8 (0.9%)	
Naoyaodome	103 (11.1%)	6 (0.6%)	
Natorube	48 (5.2%)	2 (0.2%)	
Nawoiyatira	134 (14.5%)	5 (0.5%)	
Nayanae Ng'ilimo	80 (8.6%)	4 (0.4%)	
Ecotype			0.003^{F}**
Animal shed (goats)	88 (9.5%)	15 (1.6%)	
Animal shed (sheep)	25 (2.7%)	2 (0.2%)	
Dry river	1 (0.1%)	0 (0.0%)	
Indoors (sticks)	124 (13.4%)	4 (0.4%)	
Resting area	19 (2.0%)	4 (0.4%)	
River bed	57 (6.1%)	17 (1.8%)	
Rocks	16 (1.7%)	1 (0.1%)	
Sleeping area	16 (1.7%)	3 (0.3%)	
Termite mound	288 (31.1%)	38 (4.1%)	

Continued on next page

Table 4.1 – *Continued from previous page*

Factors	**Non-VL causing vector (n%)**	**VL causing vector (n%)**	**P-value**
Vegetation	33 (3.6%)	6 (0.6%)	
Environmental Variables	-	-	
Latitude	-	-	0.665†
Longitude	-	-	0.550†
Average Humidity	-	-	0.697†
Elevation	-	-	0.682†
Total Precipitation	-	-	0.280†
Mean Temperature	-	-	0.351†
Mean NDVI	-	-	0.005†**
Max Avg Temperature	-	-	0.005†**
Min Avg Temperature	-	-	0.351†
Population Density	-	-	0.129†
Brightness	-	-	0.166†
Greenness	-	-	0.442†
Wetness	-	-	0.020†*
ET	-	-	0.534†
Tree Height (km)	-	-	0.875†

Note:

All p-values from Pearson Chi-square test unless otherwise stated.

* p-value < 0.05; ** p-value < 0.01;

Continued on next page

Table 4.1 – *Continued from previous page*

Factors	**Non-VL causing vector (n%)**	**VL causing vector (n%)**	**P-value**

♯ Fisher Exact Test; † Kruskal-Wallis Test.

4.3 Estimation of models parameters

4.3.1 Full Logistic regression model fitting (incl.correlated factors)

A logistic regression model was fitted to assess the variables influencing VL presence this was done using a combination of the significant predictors based on the initial bivariate associations on Table 4.1, as well as additional variables that were flagged as important features from the Ridge regression .

These predictors included *Ecotype*, *Mean NDVI*, *Population density*, and *Wetness*. Each of these variables was evaluated for its potential contribution to explaining the observed patterns in the data, ensuring a comprehensive analysis.

Although *Max Average Temperature* was identified as a significant predictor, it was flagged as an aliased variable, leading to its removal to enhance model stability and interpretability.

The final model results confirmed that *Ecotype*, *Mean NDVI*, and *Wetness* were significant predictors of VL-causing vector presence. Additionally, certain administrative factors, such as the data collection site and *Species* (previously established as an important factor), were also found to be significant. However, these variables were not included in the final model.

4.3.2 Multicollinearity and correlation analysis

The correlation analysis of numeric variables in Turkana Data reveals several significant relationships that provide insights into environmental and climatic patterns in the region.

Latitude exhibits a strong positive correlation with total precipitation ($r = 0.95$), indicating that higher latitudes receive more rainfall. Similarly, longitude is highly correlated with minimum average temperature ($r = 0.98$), suggesting that temperature patterns vary significantly across different longitudes.

Vegetation dynamics are strongly influenced by temperature, as evidenced by the perfect positive correlation between mean NDVI and maximum average temperature ($r = 1.00$). This implies that higher temperatures positively impact vegetation greenness. Additionally, elevation is negatively correlated with minimum average temperature ($r = -0.98$), confirming that higher altitudes experience lower temperatures, which aligns with expected climatic variations.

Mean temperature and longitude also share a strong positive relationship ($r = 0.98$), further supporting the idea of regional temperature fluctuations across different longitudes.

Population density is positively correlated with latitude ($r = 0.92$), suggesting that higher latitudes tend to have more populated areas, likely due to settlement patterns and resource distribution.

Brightness, which may indicate urbanization or land cover, is negatively correlated with average humidity ($r = -0.73$). This suggests that areas with higher brightness values tend to be drier, reflecting a decrease in moisture in more developed or exposed regions. Finally, evapotranspiration (ET) is positively correlated with elevation ($r = 0.74$), indicating that higher elevations experience increased evapotranspiration, likely due to variations in vegetation cover and atmospheric conditions.

To further assess the interdependence among predictor variables, multicollinearity was evaluated using the Generalized Variance Inflation Factor (GVIF). The GVIF values provide a measure of collinearity among the independent variables in the regression model. A $\text{GVIF}^{1/(2Df)}$ value greater than 10 indicates severe multicollinearity, suggesting that the corresponding predictor is highly correlated with other variables in the model.

Among the predictors, Max avg temperatures was flagged as an aliased variable and was removed,further upon re-running the VIF on the full logistic model comprising ten vari-

ables excluding Max average temperature the remaining variables did not exhibit high multicollineariy as indicated by Table 4.2 which presents the GVIF values, degrees of freedom (Df), and normalized GVIF values for each variable, while Figure 4.1 visualizes the correlation structure among the numeric variables.

Table 4.2: GVIF and Normalized GVIF Values for variables

Variable	**GVIF**	**Df**	**GVIF**$^{1/(2Df)}$
Mean NDVI	8.1373	1	2.8526
Wetness	12.8225	1	3.5809
Ecotype	10.3808	13	1.0942
Soil Type	3.4065	1	1.8457
Brightness	29.4819	1	5.4297
Pop Density	20.2194	1	4.4966
Total Precipitation	52.6499	1	7.2560
Tree Height (km)	12.4489	1	3.5283
Elevation	68.2950	1	8.2641
Mean Temp	45.4083	1	6.7386

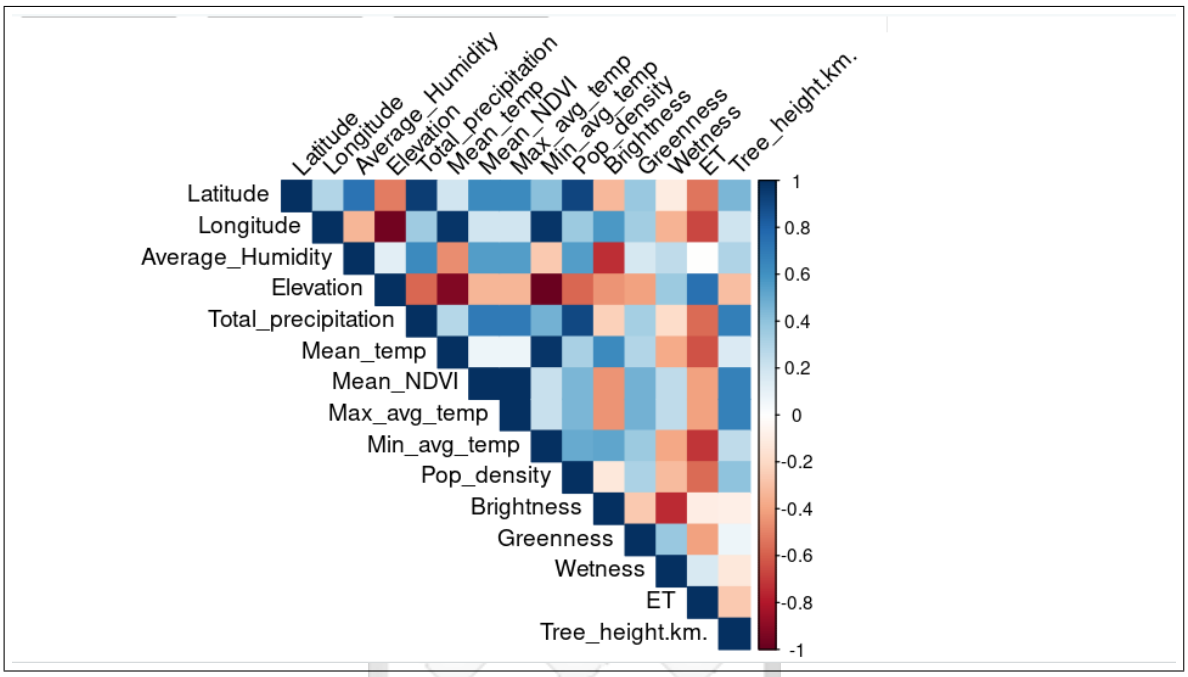


Figure 4.1: Correlation matrix

4.3.3 Logistic regression model results (excl.correlated factors)

After resolving multicollinearity, a simplified logistic regression model with fewer explanatory variables was fitted to analyze the relationship between the presence of VL-causing vectors and selected predictor variables that exhibited no multicollinearity. This model served as a baseline and was also used to assess the spatial dependence of error terms.

Table 4.3 presents the estimated coefficients, standard errors, z-values, and significance levels.

Only one Ecotype was identified as a significant predictor of the presence of VL-causing vectors this was the Indoor ecotype (made with sticks),similarly Sandy loam ($\beta = 1.031$, p-value = 0.0075) were also found to be significant predictors of VL-causing vectors.

Table 4.3: Logistic Regression Model Results

Variable	Estimate	Std. Error	p-value
Intercept	2.2978	17.9137	0.89793
Mean NDVI	9.3659	5.3728	0.08130 .
Wetness	-2.9801	9.2535	0.74741
Animal Shed Sheep	-0.7265	0.8206	0.37594
Dry River	-12.6778	882.7435	0.98854
Indoors Sticks	-1.5528	0.6027	0.00998 **
Resting Area	0.1668	0.6429	0.79529
River Bed	0.3818	0.4781	0.42456
Rocks	-1.9008	1.1417	0.09595 .
Sleeping Area	-0.2062	0.7249	0.77602
Termite Mound	0.1350	0.4416	0.75984
Vegetation	0.2309	0.6377	0.71725
Vegetation Acacia	-0.2821	0.5598	0.61431
Vegetation Balanites	-1.6019	0.8474	0.05872 .
Vegetation Palm	0.1554	0.4930	0.75265
Vegetation Prosopis	-13.1364	882.7435	0.98813
Sandy loam	1.0313	0.3855	0.00746 **
Brightness	-4.4072	5.5445	0.42668
Pop Density	-0.0231	0.0153	0.13048
Total Precipitation	0.0468	0.0987	0.63532
Tree Height (km)	-189.3890	246.4275	0.44217
Elevation	-0.0040	0.0057	0.48400
Mean Temp	-0.0138	0.5024	0.97815

4.3.4 Evaluation of the logistic regression model

The performance of the fitted logistic regression model was assessed using accuracy, sensitivity, specificity, and the Area Under the Curve (AUC), as presented in Table 4.4.

To further evaluate the model, it was tested on an independent dataset. Predictions were generated, and a confusion matrix was computed to assess classification performance. The model achieved an accuracy of 88.1% with a 95% confidence interval (CI): (85.9%, 90.2%). Additionally, the sensitivity was 100%, indicating a high ability to correctly identify positive cases, while the specificity was 0%, suggesting limited effectiveness in identifying negative cases.

The AUC of 0.72 for the Receiver Operating Characteristic (ROC) curve suggests that the model exhibits moderate predictive performance in distinguishing between positive and negative instances.

Table 4.4: Logistic Regression Model Performance

Metric	**Value**
Accuracy	88.13%
95% CI	(85.88%, 90.15%)
Sensitivity	100.00%
Specificity	0.00%
AUC	70.2%

4.3.5 Comparison of performance of machine learning models (XGBoost, KNN, Spatial Random Forest)

The performance of the logistic regression model, along with three machine learning models (XGBoost, K-Nearest Neighbors (KNN), and Spatial Random Forest), was evaluated using key classification metrics. The comparison of the machine learning models is summarized in Table 4.5.

Table 4.5: Comparison of ML Models (XGBoost, KNN, Spatial Random Forest)

Metric	**KNN**	**Spatial RF**	**XGBoost**
Accuracy	89.19%	89.43%	89.97%
Kappa	0.0647	0.2176	0.2620
Sensitivity	98.80%	99.51%	99.76%
Specificity	5.26%	14.55%	17.27%
Balanced Accuracy	52.03%	57.03%	58.51%

From the above table, XGBoost, Spatial Random Forest, and KNN was assessed using multiple evaluation metrics, including accuracy, sensitivity, specificity, balanced accuracy, and the Kappa statistic. Among these models, XGBoost achieved the highest accuracy at 89.97%, followed closely by Spatial Random Forest with 89.43%. KNN, while slightly lower, still performed relatively well with an accuracy of 89.19%. This indicates that all models effectively classify instances, although XGBoost demonstrated the best overall predictive performance.

All models exhibited high sensitivity, meaning they were proficient in identifying positive cases. XGBoost had the highest sensitivity (99.76%), with Spatial Random Forest close behind (99.51%). KNN, while slightly lower at 98.80%, still performed well in detecting positive cases. However, specificity—the ability to correctly classify negative cases—was significantly lower across all models. XGBoost had the highest specificity at 17.27%, followed by Spatial Random Forest at 14.55%, while KNN struggled the most, with a specificity of only 5.26%. The low specificity suggests that all models tend to over-predict positive cases, likely due to class imbalance in the dataset.

Balanced accuracy, which accounts for both sensitivity and specificity, was highest for XGBoost (58.51%), followed by Spatial Random Forest (57.03%). KNN had the lowest balanced accuracy (52.03%), indicating that it struggled the most in correctly classifying negative cases. The Kappa statistic, which measures agreement between the predicted and actual classifications beyond chance, was highest for XGBoost (0.2620), followed by Spatial

Random Forest (0.2176). KNN had the lowest Kappa value (0.0647), indicating weaker agreement with the true class labels.

4.3.6 Non parametric model selection

Based on the foregoing results, XGBoost emerged as the best-performing model, achieving the highest accuracy and specificity while maintaining high sensitivity. Spatial Random Forest performed similarly in some aspects but demonstrated weaker specificity. KNN performed the worst, struggling with specificity and overall predictive agreement. Given these results, XGBoost is the most robust model for predicting VL-causing vectors, whereas Spatial Random Forest may still be useful in certain scenarios. KNN, however, is not recommended due to its poor specificity, Kappa statistic, and sensitivity.

An analysis of feature importance from the XGBoost model identified the key spatial, environmental and ecological factors influencing the distribution of VL-causing vectors. Among these, **Latitude** emerged as the most influential predictor, accounting for **43.38%** of the model's importance. This was followed by **Longitude (29.12%)**, **Mean NDVI (8.95%)**, and **Indoors Sticks (4.23%)**, which also played significant roles in shaping the spatial distribution of vectors.

Additional contributing factors included **wetness (4.19%)**, **average humidity (2.13%)**, **elevation (1.67%)**, and **greenness (1.49%)**. Furthermore, **brightness (0.82%)** and **population density (0.82%)** were identified as important predictors. Within the broader ecotype category, specific land features such as **ecotype (vegetation palm) (0.70%)**, **evapotranspiration (ET) (0.59%)**, **ecotype (animal shed for sheep) (0.55%)**, **tree height (0.36%)**, and **mean temperature (0.26%)** were also observed to influence vector distribution.

Other minor but noteworthy contributing factors included **soil type (0.25%)**, **vegetation (0.15%)**, **ecotype(vegetation acacia) (0.12%)**, **ecotype(termite mounds) (0.09%)**, **ecotype (resting areas) (0.09%)**, **ecotype (vegetation Balanites) (0.04%)**, and **ecotype (riverbeds) (0.02%)**.

These results highlight the complex interplay of geographical location, vegetation indices, climatic conditions, and land-use characteristics in determining VL vector habitats. The high importance of **Latitude and Longitude** suggests that spatial positioning is crucial, while factors such as **NDVI, brightness, and soil type** further refine the understanding of vector hotspots. These insights provide a robust foundation for developing targeted vector control strategies in Turkana County.

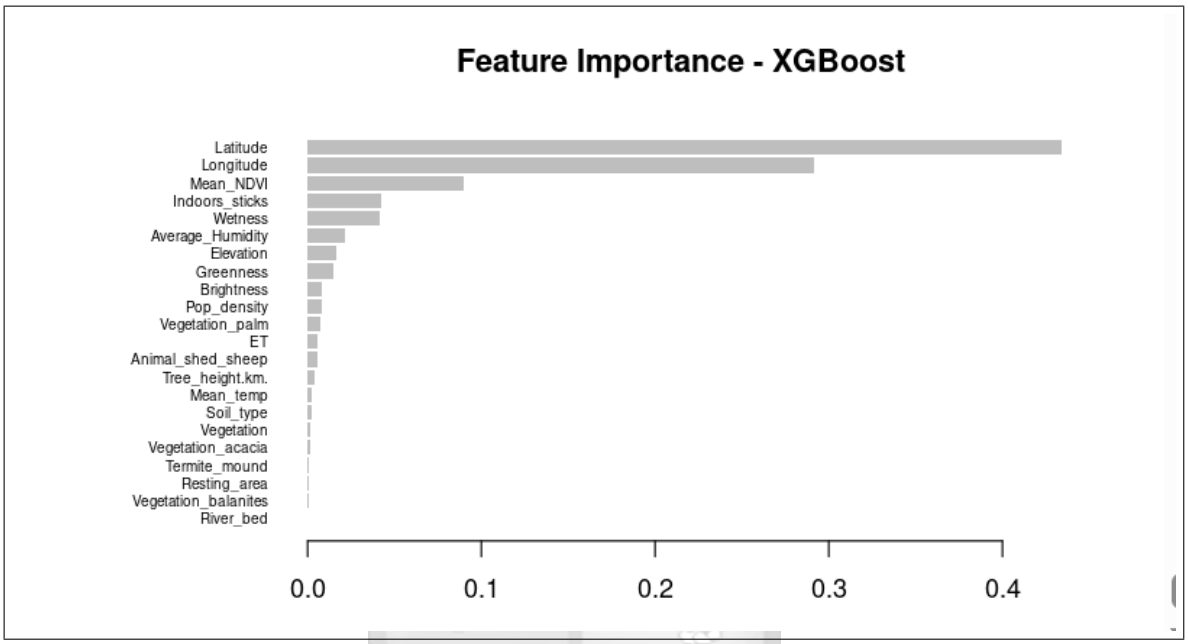


Figure 4.2: Feature Importance

4.3.7 Spatial dependence in regression residuals

The results of the Global Moran's I test indicate that the residuals of the ordinary least squares (OLS) logistic regression model exhibit significant spatial autocorrelation. With a Moran's I statistic of 2.339 and a p-value of 0.009674, we reject the null hypothesis of non-spatial randomness. The observed Moran's I value (0.0242) is greater than the expected value (-0.0153), suggesting that similar values cluster together spatially. This confirms that a standard Generalized logistic model is insufficient, and spatial regression techniques must be considered.

4.3.8 Lagrange multiplier (LM) tests for spatial dependence

To determine the most appropriate spatial regression model, we conducted Rao's Score (Lagrange Multiplier) tests to detect spatial lag and spatial error dependencies. The results are summarized in Table 4.6.

Table 4.6: Tests for Spatial Dependence in Models

Test	**Statistic**	**DF**	**p-value**
Moran's I (Residuals)	2.3388	-	0.009674
LM Test (Spatial Error)	1.8631	1	0.1723
LM Test (Spatial Lag)	3.1158	1	0.07753
Robust LM (Error)	7.0901	1	0.007751
Robust LM (Lag)	8.3429	1	0.003872
SARMA Test	10.206	2	0.006079

The results indicate that both spatial lag and spatial error dependencies are significant (from the robust LM tests), highlighting the need for an appropriate spatial regression model. The LM error test yielded a p-value of 0.1723; however, the robust LM error test returned a p-value of 0.007751, suggesting the presence of spatial error dependence, indicating that factors are unaccounted for by the model influence the VL vector abundance.

Similarly, the LM Lag test produced a p-value of 0.07753, while the Robust LM Lag test showed a p-value of 0.003872, indicating significant spatial dependence in the dependent variable. This suggests that the outcome variable is directly affected by neighboring values, making a spatial lag model (SLM) another viable option.

Furthermore, the SARMA test, which evaluates both spatial lag and spatial error dependencies simultaneously, reported a highly significant p-value of 0.006079. This finding indicates that both spatial processes are at play, reinforcing the necessity of a Spatial Autoregressive Moving Average (SARMA) model as the best approach. By incorporating both lag and error

dependence, the SARMA model offers a comprehensive solution for capturing the spatial relationships inherent in the data.

Given these results, it is evident that spatial dependence must be accounted for in the analysis, and the SARMA model presents the most robust framework for accurately modeling the distribution of VL-causing vectors in Turkana County.

Further, the comparison of spatial parametric models (SLM, SEM, and SARMA) suggests that the SARMA model provides the best fit, as it has the highest log-likelihood (-236.4), the lowest AIC (514.8), and the smallest ML residual variance (0.087). These indicators suggest that the SARMA model captures spatial dependencies more effectively than the SLM and SEM models.

Table 4.7: Comparison of spatial parametric models

Metric	**SLM**	**SEM**	**SARMA**
Log Likelihood	-241.50635	-242.03235	-236.42382
AIC	523.01270	524.06471	514.84765
Rho (SLM) / Lambda (SEM)	0.09888	0.08302	-0.77598
P-Value for Spatial Dependence	0.08679	0.17023	0.00143
ML Residual Variance	0.09846	0.09861	0.08727

4.3.9 Parametric model selection

Given that both the spatial lag and spatial error effects were significant, we conclude that a SARMA parametric model could be used. This model will account for both forms of spatial dependence and provide the most reliable estimates of the relationships between variables in the study. These results reinforce that spatial dependence must be accounted for when modeling the distribution of VL-causing vectors in Turkana County. The Spatial Autoregressive Moving Average (SARMA) model was selected as the best-fitting parametric model due to its superior performance in capturing spatial dependencies. Table 4.8 presents the estimated coefficients, standard errors, z-values, and p-values.

Table 4.8: SARMA Model Results

Variable	**Estimate**	**Std. Error**	**z-value**	**p-value**
Intercept	0.1342	0.0674	1.9924	0.0463 *
Mean NDVI	0.3315	0.1501	2.2086	0.0272 *
Animal_shed_sheep	-0.0598	0.0412	-1.4530	0.1462
Dry_river	-0.1473	0.2870	-0.5131	0.6079
Indoors_sticks	-0.0322	0.0254	-1.2663	0.2054
Resting_area	-0.0033	0.0419	-0.0799	0.9364
River_bed	0.0537	0.0299	1.7946	0.0727 .
Rocks	-0.0489	0.0467	-1.0468	0.2952
Sleeping_area	0.0164	0.0488	0.3362	0.7368
Termite_mound	0.0215	0.0239	0.9026	0.3668
Vegetation	0.0214	0.0356	0.6006	0.5481
Vegetation_acacia	-0.0551	0.0352	-1.5642	0.1178
Vegetation_balanites	-0.0128	0.0312	-0.4096	0.6821
Vegetation_palm	0.0045	0.0296	0.1535	0.8780
Vegetation_prosopis	-0.0289	0.2913	-0.0992	0.9210
Sandy loam	0.0666	0.0212	3.1423	0.0017 **
ET	-0.0068	0.0024	-2.7986	0.0051 **
Wetness	0.4836	0.2053	2.3561	0.0185 *
ρ (Spatial Lag)	0.5440	0.0689	7.8924	< 0.0001
λ (Spatial Error)	-0.7760	0.1686	-4.6013	< 0.0001
LR test value	13.098	-	-	0.0014 **
Log likelihood	-236.4238	-	-	-
ML residual variance	0.0873	-	-	-
AIC	514.85	-	-	-

Further, the results above in Table 4.8 indicate that Mean NDVI ($\beta = 0.3315$, $p = 0.0272$) is a significant predictor at the 5% level, suggesting that areas with higher Normalized Difference Vegetation Index (NDVI) are associated with an increased presence of VL-causing vectors.

Additionally, Sandy loam ($\beta = 0.0666$, $p = 0.0017$) is significant at the 1% level, indicating that areas with this soil type have a higher likelihood of VL-causing vectors. Similarly, wetness ($\beta = 0.4836$, $p = 0.0185$) is significant at the 5% level, showing a positive relationship between wetness and the presence of VL-causing vectors.

Evapotranspiration (ET) showed a negative correlation ($\beta = -0.0068$, $p = 0.0051$) with VL-causing vectors, indicating that an increase in evapotranspiration was associated with a decrease in the abundance of VL-causing vectors.

The results from the spatial analysis provide strong evidence for the presence of significant spatial dependencies in the distribution of VL-causing vectors. The spatial parameters ρ and λ were both found to be highly significant, with p-values less than 0.0001, indicating that spatial effects are crucial to understanding the patterns of vector distribution.

The spatial lag parameter ($\rho = 0.5440$) reflects the degree to which the presence of VL-causing vectors in one location is influenced by their presence in neighboring locations. The positive value for ρ suggests that there is a positive spatial autocorrelation, meaning that areas with a higher abundance of VL-causing vectors tend to have neighboring areas with similarly high abundance. This can be attributed to shared environmental or ecological factors that favor the proliferation of vectors in certain areas, creating clusters of high vector abundance across neighboring regions. For instance, regions with favorable environmental conditions for the vectors, such as the presence of breeding sites or particular vegetation types, will likely see these conditions extend to nearby areas, reinforcing the clustering of vector populations.

On the other hand, the spatial error parameter ($\lambda = -0.7760$) provides insight into the residual spatial correlation in the error term, capturing unobserved spatial factors that influence the vector distribution. A negative value for λ indicates negative spatial autocorrelation in the error term, meaning that areas with higher-than-expected vector populations might

be surrounded by areas with lower-than-expected vector populations, and vice versa. This suggests that there are additional, unaccounted-for spatial processes or factors that might be influencing the distribution of the vectors, such as microhabitat conditions or other local environmental factors that differ from those of the surrounding regions.

The SARMA model was chosen as the final parametric model due to its ability to account for both spatial lag and spatial error dependence. The significant spatial parameters ρ and λ confirm the importance of spatial effects. Among the predictor variables, ET, mean NDVI, sandy loam, and wetness were found to be significant factors influencing VL-causing vectors.

4.3.10 Comparison of SARMA and XGBoost

To evaluate the predictive performance of the models, the Area Under the Curve (AUC) was used as a comparative metric. The results indicate that the XGBoost model outperformed the SARMA model, achieving an AUC of 0.8997 compared to SARMA's AUC of 0.8878.

Based on this performance, XGBoost was selected as the most effective model for predicting the presence of VL-causing vectors in Turkana County. The comparison of the models' ROC curves is presented in Figure 4.3.

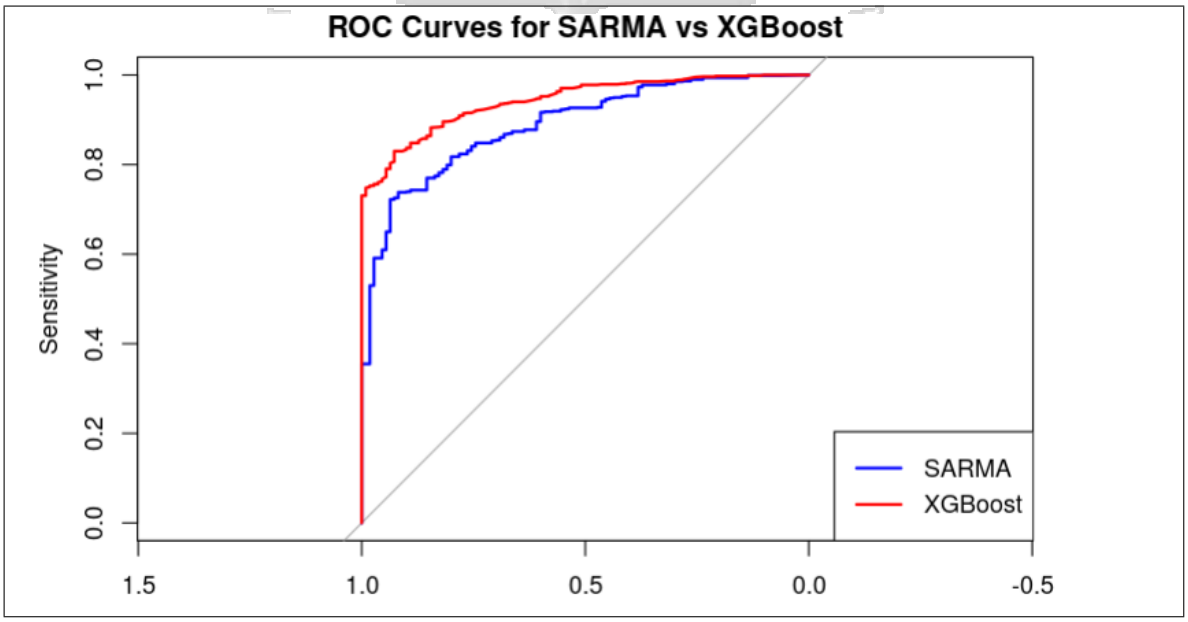


Figure 4.3: Comparison of Model Performance: SARMA vs. XGBoost

Given its superior classification performance, XGBoost was selected as the final predictive model for VL vector presence in this study. Its high sensitivity (**99.63%**) ensures accurate detection of vector-positive areas, making it the most reliable tool for vector surveillance and control planning in Turkana County.

The insights derived from XGBoost provide a strong foundation for developing targeted vector control strategies. By leveraging its predictive capabilities, public health initiatives can focus on high-risk regions, improving resource allocation and intervention effectiveness in controlling VL transmission.

Chapter 5

Discussions, Conclusions and Recommendations

5.1 Introduction

This chapter presents the findings from the spatial modeling of VL-causing vectors in Turkana County, Kenya. The study utilized both spatial parametric and machine learning models to explore the influence of environmental, spatial, and ecological factors on the distribution of these vectors. The Spatial Autoregressive Moving Average (SARMA) model, a parametric approach, was employed to account for spatial dependencies, while XGBoost, a non-parametric machine learning model, emerged as the best-performing model in terms of predictive accuracy and robustness. This section contextualizes the key results, comparing them with existing literature and highlighting their relevance for vector surveillance and control strategies.

5.2 Discussion

This study significantly enhances our understanding of the ecological and spatial determinants influencing the distribution of phlebotomine sandflies, the vectors responsible for leishmaniasis, in Turkana County, Kenya. The results highlight the multifaceted nature of vector habitats, where environmental, ecological, and spatial factors interact to shape the distribution of VL-causing vectors. These findings are critical not only for understanding the epidemiology of the disease but also for informing vector surveillance and control strategies in regions vulnerable to leishmaniasis.

A key finding of this study is the substantial contribution of geographic factors, specifically **Latitude** and **Longitude**, to the model's predictive power. The XGBoost model indicated that geographic location was the most influential factor in determining the presence of VL vectors, contributing 43.38% and 29.12% to the overall model feature importance, respectively. These results are in line with previous studies, which have emphasized the role of spatial positioning in the distribution of vector-borne diseases Rotejanaprasert et al. (2021)..

The SARMA model, which accounts for spatial dependencies in the data, corroborated the importance of geographic location by revealing significant spatial autocorrelation. This suggests that the distribution of VL vectors is not random but is influenced by neighboring areas. The presence of a positive spatial lag parameter in the SARMA model points to the clustering effect, where areas with high vector populations are likely to influence adjacent regions. Conversely, the negative spatial error parameter suggests that certain unobserved factors, potentially linked to local environmental conditions, could lead to negative spatial autocorrelation, where regions with high vector populations are surrounded by areas with unexpectedly low numbers of vectors. These findings underline the need for a spatially aware approach to vector control, where interventions in one area may have spillover effects in neighboring regions Bhunia and et al. (2018).

Furthermore, the study identified **Mean NDVI** (Normalized Difference Vegetation Index) as a significant predictor of vector distribution in both the XGBoost and SARMA models. The XGBoost model assigned an 8.95% importance to NDVI, while the SARMA model confirmed that regions with higher NDVI values were more likely to harbor VL vectors. The positive relationship between NDVI and vector distribution aligns with existing literature that points to NDVI as an important factor in influencing vector abundance (El Fattouhi et al., 2023). Vegetation provides essential microhabitats for sandflies by offering shade and moisture, which are critical for their survival and reproductive cycles. NDVI, a widely used satellite-derived index for assessing vegetation cover, proved to be an effective predictor of vector habitats in this study. These results highlight the potential of remote sensing data to support vector surveillance efforts, enabling the identification of areas with suitable

vegetation cover that could serve as breeding grounds for sandflies Almeida and et al. (2021); Bhunia and et al. (2018).

Another important environmental factor identified in this study is **wetness**, which emerged as a key determinant in both models. The SARMA model demonstrated a significant positive relationship between wetness and vector presence, while the XGBoost model attributed 4.19% importance to this variable. Wet environments are known to support the development of vector populations, particularly those that breed in moist conditions such as sandflies. The positive correlation between wetness and vector distribution may be explained by the availability of breeding sites in areas with high moisture content, which are ideal for larval development. This finding is consistent with previous research that has highlighted the association between sandfly abundance and wetland habitats (Hosseini et al., 2021). The identification of wetness as a significant predictor also suggests that changes in water availability due to seasonal rainfall patterns or land use changes could have a direct impact on the spread of VL vectors in Turkana County.

In addition to the environmental factors discussed above, the study also explored the role of ecological variables in shaping vector distribution. The SARMA model revealed that **sandy loam** and **evapotranspiration (ET)** were important predictors, while the XGBoost model highlighted these as important features as well; these results align with a longitudinal study by Senanayake et al. (2024b) who found out that an increase in evaporation reduced the density of phlebotomine sandflies.

The inclusion of these ecological variables underscores the complexity of sandfly habitats, which are not solely influenced by environmental factors but also by local land use practices and human settlements. For instance, the presence of **ecotype (indoors sticks)**, which are common in rural households in Turkana, could provide suitable resting or breeding sites for sandflies, a behavior that is well-documented in the literature (Philipo Gwandi et al., 2022). The use of natural materials for construction, such as palm leaves and sticks, may inadvertently create microhabitats that support vector populations. This finding suggests that vector control strategies should not only target natural habitats but also consider anthro-

pogenic factors that contribute to the proliferation of sandflies in human settlements (Baker and Others, 2015; Philipo Gwandi et al., 2022).

Additionally, ecological factors such as **Animal Shed for Sheep**, **River Beds**, and **Termite Mounds** were found to influence vector populations in areas where human activities intersect with vector ecology. Livestock shelters and riverbeds are known to attract sandflies due to the availability of food sources (such as mammals) and microhabitats that provide favorable conditions for breeding. Termite mounds, often found in arid regions like Turkana, have been identified as important ecological features that contribute to vector distribution (Baker and Others, 2015; Ready, 2013). These results emphasize the need for integrated vector control strategies that address both natural and human-induced habitats, promoting a holistic approach to managing sandfly populations.

The findings from the XGBoost and SARMA models have several implications for vector surveillance and control strategies. Identifying key predictors such as **Latitude**, **Longitude**, **Mean NDVI**, and **Wetness**, along with ecological features like **Indoors Sticks** and **Animal Shed for Sheep**, can help design targeted intervention programs. These programs could focus on modifying habitats, such as removing breeding sites in areas with high NDVI or wetness, and conducting spatially targeted insecticide applications in regions identified as high-risk based on the spatial models. The inclusion of spatial modeling in surveillance systems could also improve the detection of vector hotspots, facilitating timely interventions in areas with a high likelihood of vector presence (Bhunia and et al., 2018).

The spatial dependencies revealed in the SARMA model offer additional insights for vector control. Since neighboring regions with high vector populations tend to influence one another, interventions in one area could have a cascading effect, potentially reducing the spread of vectors across a larger region. This finding underscores the importance of implementing region-wide or multi-area control programs that consider the spatial relationships between areas. Furthermore, incorporating spatial autocorrelation into surveillance and control strategies can improve the detection of emerging vector hotspots, which may otherwise go unnoticed in traditional monitoring systems (Baker and Others, 2015; Bhunia and et al., 2018).

5.3 Strengths and limitations

This study integrates both parametric and non-parametric modeling techniques to provide a comprehensive understanding of VL vector distribution in Turkana County. While the use of both SARMA and XGBoost offers complementary insights, the limitations of the study design, such as the lack of temporal analysis, warrant further investigation. Additionally, the reliance on environmental and ecological predictors without incorporating socioeconomic variables may limit the broader applicability of the findings.

5.4 Recommendations

5.4.1 Recommendations for further studies

Future studies should address the seasonal variation in vector populations by incorporating temporal data into spatial models. Such studies would provide insights into how seasonal climate patterns, such as temperature and rainfall, influence vector distribution. Furthermore, socioeconomic factors, such as human mobility and population density, should be integrated into the models to assess their impact on vector ecology.

In addition, expanding the geographic scope to other regions with endemic VL transmission could validate the generalizability of the findings and help refine predictive models for different ecological settings. The integration of remote sensing data with ground-level ecological data could enhance the spatial accuracy of the models, providing more precise predictions for vector control efforts.

5.4.2 Policy recommendations

The findings from this study can inform policy development for vector control strategies in Turkana County and similar endemic regions. Key recommendations include the development of targeted surveillance programs that focus on high-risk areas identified by spatial modeling.

Additionally, policies that promote habitat modification, such as reducing breeding sites in areas with high NDVI or wetness, could help curb the spread of VL-causing vectors.

Public health interventions, such as the distribution of insecticide-treated nets or indoor residual spraying in high-risk zones, could also be prioritized based on the findings of this study. In addition, fostering community engagement in vector control activities can enhance the effectiveness of these interventions.

Similarly, the ministries of Environment and Livestock should monitor land use and anthropogenic activities, such as forest encroachment, that impact evapotranspiration and wetness. This could help mitigate human-vector interactions and prevent the creation of conditions that may promote an increase in vector populations.

5.5 Conclusion

In conclusion, this study provides valuable insights into the environmental and spatial factors influencing the distribution of VL-causing vectors in Turkana County. The results underscore the importance of geographic location, vegetation, and moisture in shaping vector habitats. The combination of SARMA and XGBoost models offers complementary perspectives on vector distribution, highlighting the need for integrated control strategies that consider both ecological and spatial factors. Further research incorporating temporal data and socioeconomic variables, along with expanded geographic coverage, will improve the accuracy and applicability of these findings for global vector control efforts.

References

Alemayehu, B., Koroto, N., Tomas, T., Matusala, T., Megaze, A., and Leirs, H. (2024). The abundance and distribution of sandflies (with emphasis on phlebotomus pedifer)(diptera: Psychodidae) along the altitudinal gradient in kindo didaye district, wolaita zone, south ethiopia. *Journal of Medical Entomology*, page tjae049.

Ali, H. et al. (2020). Modeling the spread of leishmaniasis in east africa. *Journal of Mathematical Biology*, 59(5):876–890.

Allison, P. D. (2001). *Missing Data*. Sage Publications.

Almeida, P. S. and et al. (2021). Land use and vegetation cover impact on sandfly distribution. *Parasites Vectors*, 14:1–14.

Alvar, J., Beca-Martínez, M. T., Argaw, D., Jain, S., and Aagaard-Hansen, J. (2023). Social determinants of visceral leishmaniasis elimination in eastern africa. *BMJ Global Health*, 8(6):e012638.

Alvar, J., den Boer, M., and Dagne, D. A. (2021). Towards the elimination of visceral leishmaniasis as a public health problem in east africa: reflections on an enhanced control strategy and a call for action. *The Lancet Global Health*, 9(12):e1763–e1769.

Alvar, J. et al. (2006). Epidemiology of leishmaniasis in south america. *The Lancet Infectious Diseases*, 6(11):718–725.

Alvar, J. et al. (2020). Leishmaniasis: Global burden and control strategies. *The Lancet Infectious Diseases*, 20(11):1185–1192.

Alvar, J., Velez, A. C., Bern, A., Barros, E. H., and Herrero, L. (2012). Leishmaniasis and sandflies: A global perspective. *The Lancet Infectious Diseases*, 12:553–563.

Andrade, J. (2021). *Research Design in Social Science*. Oxford University Press.

Andre´s F. Henao-Mart´ınez, University of Colorado Anschutz Medical Campus: University of Colorado Anschutz Medical Campus, U. S. (2022). Challenges of using modelling evidence in the visceral leishmaniasis elimination programme in india. *Research and reports in tropical medicine*, pages 1–2.

Anoopa Sharma, D., Bern, C., Varghese, B., Chowdhury, R., Haque, R., Ali, M., Amann, J., Ahluwalia, I. B., Wagatsuma, Y., Breiman, R. F., et al. (2006). The economic impact of visceral leishmaniasis on households in bangladesh. *Tropical Medicine & International Health*, 11(5):757–764.

Ayele, D. G., Mohammed, M. O. M., Abdallah, A. S. R., and Wacho, G. A. (2024). Assessment of malaria transmission in kenya using multilevel logistic regression. *Heliyon*, 10(21):e39835.

Babbie, E. (2016). *The Practice of Social Research*. Cengage Learning.

Baker, R. L. and Others (2015). Geospatial analysis of vector-borne diseases: Mapping the distribution of phlebotomine sandflies. *Journal of Vector Ecology*, 40(2):203–213.

Barnett, V. and Lewis, T. (1994). *Outliers in Statistical Data*. John Wiley & Sons, 3rd edition.

Bhunia, G. P. and et al. (2018). Prediction of visceral leishmaniasis using remote sensing techniques. *Acta Tropica*, 187:1–12.

Blazquez, C. A., Picarte, B., Calderón, J. F., and Losada, F. (2018). Spatial autocorrelation analysis of cargo trucks on highway crashes in chile. *Accident Analysis Prevention*, 120:195–210.

Burnham, K. and Anderson, D. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer, New York.

Cohen, J. (2003). *Statistical Power Analysis for the Behavioral Sciences*. Routledge.

da Paixão Sevá, A., Mao, L., Galvis-Ovallos, F., Oliveira, K. M. M., Oliveira, F. B. S., and Albuquerque, G. R. (2023). Spatio-temporal distribution and contributing factors of tegumentary and visceral leishmaniasis: A comparative study in bahia, brazil. *Spatial and Spatio-temporal Epidemiology*, 47:100615.

Desjeux, P. (2004). Leishmaniasis: Current situation and new perspectives. *The Lancet Infectious Diseases*, 4:563–568.

El Fattouhi, Y., Guemmouh, R., Idrissi, A. J., Darkaoui, N., Hakam, O., El Ouaryaghli, A., Talbi, F. Z., Benrezzouk, R., and Lalami, A. E. O. (2023). Spatiotemporal dynamics of phlebotomus perniciosus (diptera: Psychodidae) and characterization of its habitats using satellite images in fez city, north central morocco. *Tropical Journal of Natural Product Research*, 7(10):4133–4140.

El Omari, H., Chahlaoui, A., Talbi, F. Z., Chlouchi, A., El-Akhal, F., Lahouiti, K., Taroq, A., Mrani Alaoui, M., and El Ouali Lalami, A. (2023). Entomological survey and impact of climatic factors on the dynamics of sandflies in central morocco. *The Scientific World Journal*, 2023(1):6952992.

Elnaiem, D. et al. (2006). Seasonal and geographic distribution of leishmaniasis in the sudan. *American Journal of Tropical Medicine and Hygiene*, 75(2):254–260.

Enders, C. K. (2010). *Applied Missing Data Analysis*. The Guilford Press.

Fernández, M. S., Lestani, E. A., Cavia, R., and Salomón, O. D. (2012). Phlebotominae fauna in a recent deforested area with american tegumentary leishmaniasis transmission (puerto iguazú, misiones, argentina): Seasonal distribution in domestic and peridomestic environments. *Acta Tropica*, 122(1):16–23.

Ferreira, J., Silva, K., Cavalcanti, M., Ferreira-Júnior, G., Souza, E., Magalhães, P., Gomes, D., Fonseca, S., Maranhão, T., Rocha, M., et al. (2022). Spatio-temporal analysis of the visceral leishmaniasis in the state of alagoas, brazil. *Brazilian Journal of Biology*, 84:e253098.

Fonseca, E. D. S., Guimaraes, R. B., Prestes-Carneiro, L. E., Tolezano, J. E., Rodgers, M. D. S. M., Avery, R. H., and Malone, J. B. (2021). Predicted distribution of sand fly (diptera: Psychodidae) species involved in the transmission of leishmaniasis in são paulo state, brazil, utilizing maximum entropy ecological niche modeling. *Pathogens and global health*, 115(2):108–120.

Friendly, M. (2007). *A Brief History of Data Visualization*. Springer.

Gebre-Michael, T., Malone, J., Balkew, M., Ali, A., Berhe, N., Hailu, A., and Herzi, A. (2004). Mapping the potential distribution of phlebotomus martini and p. orientalis (diptera: Psychodidae), vectors of kala-azar in east africa by use of geographic information systems. *Acta Tropica*, 90(1):73–86.

Gerritsen, R. and Clements, M. J. (2014). Methods of random sampling in ecological research. *Ecology and Evolution*, 4:3300–3312.

Geto, A. K., Berihun, G., Berhanu, L., Desye, B., and Daba, C. (2024). Prevalence of human visceral leishmaniasis and its risk factors in eastern africa: a systematic review and meta-analysis. *Frontiers in Public Health*, 12.

Gingrich, P. and Pugh, T. (2017). *Vector-Borne Diseases in Africa: Challenges and Prevention*. Cambridge University Press.

Grace Grifferty, Hugh Shirley, J. M. J. K. A. O. and Wamai, R. (2021). Vulnerabilities to and the socioeconomic and psychosocial impacts of the leishmaniases: A review. *Research and Reports in Tropical Medicine*, 12:135–151.

Grifferty, G., Shirley, H., McGloin, J., Kahn, J., Orriols, A., and Wamai, R. (2021a). Vulnerabilities to and the socioeconomic and psychosocial impacts of the leishmaniases: A review. *Research and reports in tropical medicine*, pages 135–151.

Grifferty, G., Shirley, H., McGloin, J., Kahn, J., Orriols, A., and Wamai, R. (2021b). Vulnerabilities to and the socioeconomic and psychosocial impacts of the leishmaniases: A review. *Research and reports in tropical medicine*, pages 135–151.

Hao, C., Zhao, Z., Zhang, P., Wu, B., Ren, H., Wang, X., Qiao, Y., Cui, Y., and Qiu, L. (2024). Application of improved harmonic poisson segmented regression model in evaluating the effectiveness of kala-azar intervention in yangquan city, china. *Frontiers in Public Health*, 12:1326225.

Hassaballa, I. B., Sole, C. L., Cheseto, X., Torto, B., and Tchouassi, D. P. (2021). Afrotropical sand fly-host plant relationships in a leishmaniasis endemic area, kenya. *PLoS Neglected Tropical Diseases*, 15(2):e0009041.

Hosseini, S. H., Allah-Kalteh, E., and Sofizadeh, A. (2021). The effect of geographical and climatic factors on the distribution of phlebotomus papatasi (diptera: Psychodidae) in golestan province, an endemic focus of zoonotic cutaneous leishmaniasis in iran, 2014. *Journal of Arthropod-Borne Diseases*, 15(2):225–235. Published online 2021 Jun 30.

Johansson, A., Smith, L., and Patel, R. (2021). Machine learning approaches for predicting malaria risk using climatic and environmental variables. *International Journal of Epidemiology*, 50(4):987–1002.

Joshi, A. B., Das, M. L., Akhter, S., Chowdhury, R., Mondal, D., Kumar, V., Das, P., Kroeger, A., Boelaert, M., and Petzold, M. (2009). Chemical and environmental vector control as a contribution to the elimination of visceral leishmaniasis on the indian subcontinent: cluster randomized controlled trials in bangladesh, india and nepal. *BMC medicine*, 7:1–9.

Kama, A. (2016). Livestock losses due to sandfly-transmitted diseases in kenya. *East African Journal of Veterinary Studies*, 15(1):56–65.

Kasili, S. and Okindo, E. (2016). Socioeconomic impacts of leishmaniases on households of marigat sub county, baringo county of kenya. *Journal of Tropical Diseases*, 04.

Khalil, E. A. and El-Beshbishy, A. M. S. (2018). Vector ecology and control in leishmaniasis-endemic regions. *International Journal of Infectious Diseases*, 70:79–85.

Killick-Kendrick, R. (2000). Leishmaniasis: The vector. *The Lancet Infectious Diseases*, 1(3):168–174.

Majid, A. (2018). *The Art of Research Design*. Springer.

Marangu, V. M., Kei, R. M., and Kithinji, D. K. (2024). The one health approach: combating the persistent visceral leishmaniasis (kala-azar) outbreaks in isiolo county, kenya. *International Journal of Community Medicine and Public Health*, 11(5):1750.

McCullagh, P. (1984). Generalized linear models. *European Journal of Operational Research*, 16(3):285–292.

McGill, R., Tukey, J. W., and Larsen, W. A. (1978). Variations of box plots. *The American Statistician*, 32(1):12–16.

Meheus, F., Abuzaid, A. A., Baltussen, R., Younis, B. M., Balasegaram, M., Khalil, E. A. G., Boelaert, M., and Musa, A. M. (2013). The economic burden of visceral leishmaniasis in sudan: An assessment of provider and household costs. *The American Society of Tropical Medicine and Hygiene*, 89(6):1146 – 1153.

Ministry of Health, K. (2021). The strategic plan for control of leishmaniasis 2021-2025. Technical report, Ministry of Health, Kenya. Available at: https://www.health.go.ke/wp-content/uploads/2021/07/KSPC-OF-LEISHMANIASIS-STRATEGY-2021-2025.pdf.

Moore, R. (2017). Leishmaniasis and educational outcomes in affected regions. *Social Science and Medicine*, 189:86–95.

Murray, C. et al. (2015). Global burden of leishmaniasis: A systematic review of epidemiology and treatment. *PLOS Neglected Tropical Diseases*, 9(3):e0004102.

Okwor, I. and Uzonna, J. (2016). Social and economic burden of human leishmaniasis. *The American journal of tropical medicine and hygiene*, 94(3):489.

Organization, W. H. (2010). Who, "control of the leishmaniases," world health organ technical report series, vol. xii-xiii, pp. 1- 186, 2010. *WHO Fact Sheet*. Accessed: 2025-01-18.

Patel, A., Gupta, R., and Singh, P. (2023). Application of spatial autoregressive moving average models in malaria risk assessment. *International Journal of Epidemiology*, 52(5):1012–1025.

Philipo Gwandi, M., Odongo, A. O., Kirira, P. G., and Jeruto, E. (2022). Risk factors associated with leishmaniasis among residents of rural marigat sub-county, baringo county - kenya. *International Journal of TROPICAL DISEASE amp; Health*, 43(7):1–11.

Pigott, D. M., Bhatt, S., Golding, N., Duda, K. A., Battle, K. E., Brady, O. J., Messina, J. P., Balard, Y., Bastien, P., Pratlong, F., et al. (2014). Global distribution maps of the leishmaniases. *Elife*, 3:e02851.

Ranjan, S. et al. (2020). Visceral leishmaniasis: Epidemiology and treatment approaches. *Journal of Parasitology Research*, 45(2):56–64.

Ready, P. D. (2013). Biology of phlebotomine sand flies as vectors of disease agents. *Annual Review of Entomology*, 58:227–250.

Risueño, J., Bersihand, S., Bender, C., Cornen, T., De Boer, K., Ibáñez-Justicia, A., Rey, D., Rozier, Y., Schneider, A., Stroo, A., et al. (2024). A survey of phlebotomine sand flies across their northern distribution range limit in western europe. *Journal of the European Mosquito Control Association*, 1(aop):1–11.

Rivera, L., Gonzalez, R., and Santos, J. (2022). Remote sensing applications in cutaneous leishmaniasis control: A spatial-temporal approach using sarma models. *Journal of Vector Ecology*, 47(2):210–225.

Rotejanaprasert, C., Ekapirat, N., Sudathip, P., and Maude, R. J. (2021). Bayesian spatio-temporal distributed lag modeling for delayed climatic effects on sparse malaria incidence data. *BMC Medical Research Methodology*, 21(1):287.

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. John Wiley & Sons.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons.

Schoenheit, E. et al. (2018). Understanding the transmission dynamics of leishmaniasis. *Tropical Medicine and Infectious Disease*, 6(4):75–82.

Senanayake, S. C., Liyanage, P., Pathirage, D. R., Siraj, M. R., De Silva, B. N. K., and Karunaweera, N. D. (2024a). Impact of climate and land use on the temporal variability of sand fly density in sri lanka: A 2-year longitudinal study. *PLOS Neglected Tropical Diseases*, 18(11):e0012675.

Senanayake, S. C., Liyanage, P., Pathirage, D. R. K., Siraj, M. F. R., De Silva, B. G. D. N. K., and Karunaweera, N. D. (2024b). Impact of climate and land use on the temporal variability of sand fly density in sri lanka: A 2-year longitudinal study. *PLOS Neglected Tropical Diseases*, 18(11):1–23.

Smith, J. and Brown, E. (2020). Spatial modeling of infectious diseases: A case study on leishmaniasis. *Spatial Epidemiology Journal*, 12(3):215–230.

Stauch, A., Sarkar, R. R., Picado, A., Ostyn, B., Sundar, S., Rijal, S., Boelaert, M., Dujardin, J.-C., and Duerr, H.-P. (2011). Visceral leishmaniasis in the indian subcontinent: Modelling epidemiology and control. *PLOS Neglected Tropical Diseases*, 5(11):1–12.

Subramanian, S., Maheswari, R. U., Prabavathy, G., Khan, M. A., Brindha, B., Srividya, A., Kumar, A., Rahi, M., Nightingale, E. S., Medley, G. F., Cameron, M. M., Roy, N., and Jambulingam, P. (2024). Modelling spatiotemporal patterns of visceral leishmaniasis incidence in two endemic states in india using environment, bioclimatic and demographic data, 2013–2022. *PLOS Neglected Tropical Diseases*, 18(2):1–28.

Tadesse Hailu, Mulat Yimer, W. M. . B. A. (2016). Challenges in visceral leishmaniasis control and elimination in the developing countries: A review. *Research and reports in tropical medicine*, page 196.

Talbi, F. Z., Merabti, A., El-Akhal, F., Alami, A., El Omari, H., Najy, M., Amaiach, R., Alaoui, M. M., Taroq, A., El Khayyat, F., et al. (2022). Influence of basic environmental factors on seasonal variation and distribution of sand flies at ben slimane sites in fez city, morocco. *Tropical Journal of Natural Product Research*, 6(9).

Tarnas, M. C., Abbara, A., Desai, A. N., and Parker, D. M. (2024). Ecological study measuring the association between conflict, environmental factors, and annual global cutaneous and mucocutaneous leishmaniasis incidence (2005–2022). *PLOS Neglected Tropical Diseases*, 18(9):1–13.

Tesh, R. and colleagues (2019). Epidemiology and control of leishmaniasis in tropical regions. *International Journal of Tropical Medicine*, 25(4):231–239.

Tessema, S. B., Hagos, T., Kehasy, G., Paintain, L., Adera, C., Herrero, M., den Boer, M., Temesgen, H., Price, H., and Mulugeta, A. (2024). The economic burden of visceral leishmaniasis and barriers to accessing healthcare in tigray, north ethiopia: A field based study. *PLOS Neglected Tropical Diseases*, 18(10):1–19.

Trájer, A. J. and Kniha, E. (2025). Climatic and meteorological factors shaping the potential activity season of sand flies in southeast europe. *Acta Tropica*, 261:107486.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley.

Valero, N. N. H., Prist, P., and Uriarte, M. (2021). Environmental and socioeconomic risk factors for visceral and cutaneous leishmaniasis in são paulo, brazil. *Science of The Total Environment*, 797:148960.

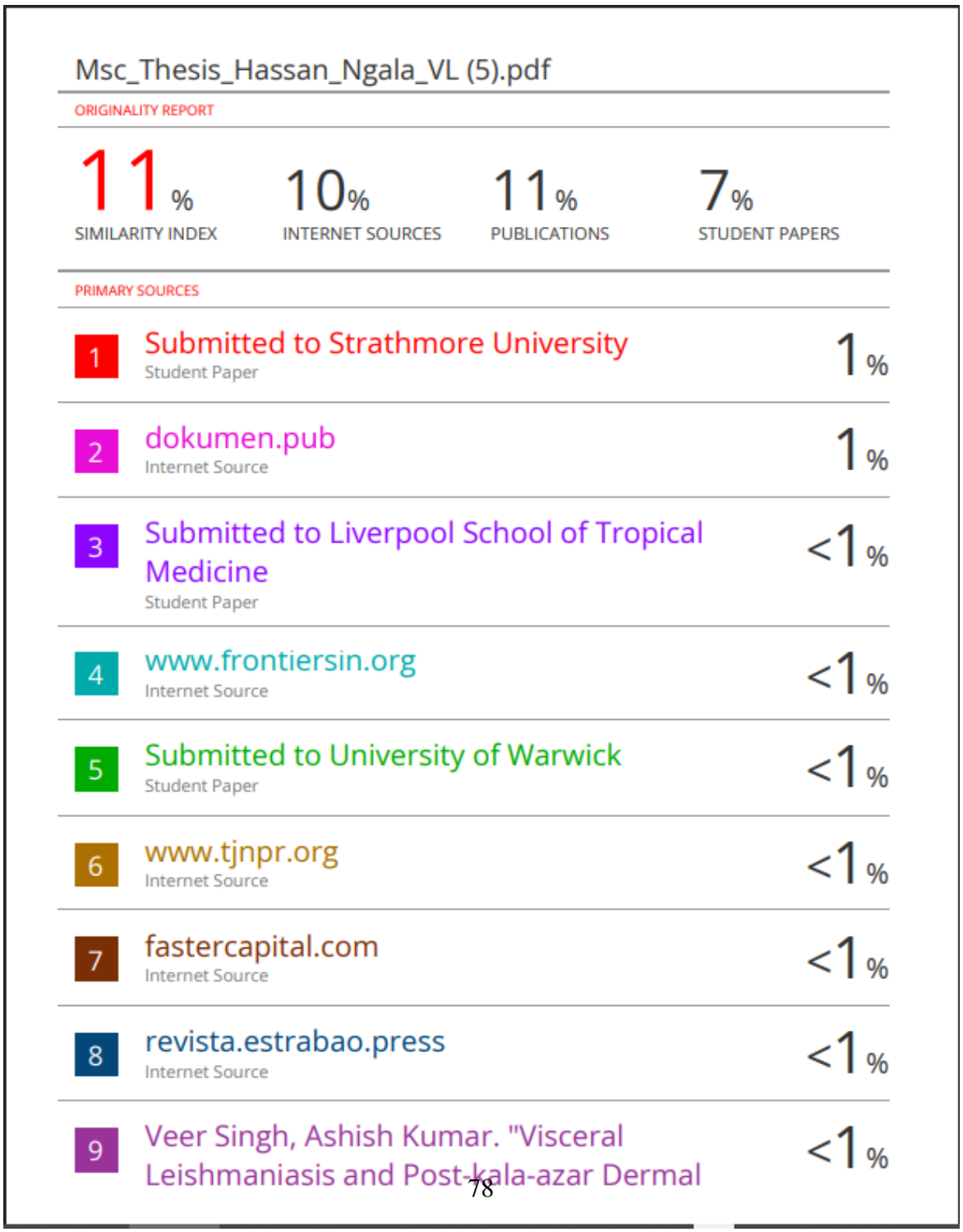
Van Dijk, Norbert, J. C. W. O. P. M. and Schallig, H. (2023). 'clinical features, immunological interactions and household determinants of visceral leishmaniasis and malaria coinfections in west pokot, kenya: Protocol for risk mapping of visceral leishmaniasis infections in west pokot, kenya: Characterisation of local environmental risk factors 36 an observational study'. *BMJ Open 13(4):e068679. doi: 10.1136/BMJOPEN-2022- 068679.*

van Dijk, N. J., Amer, S., Mwiti, D., Schallig, H. D., and Augustijn, E.-W. (2024). An epidemiological and spatiotemporal analysis of visceral leishmaniasis in west pokot, kenya, between 2018 and 2022. *BMC infectious diseases*, 24(1):1169.

VT OMNES
VNVM SINT

Appendix A

Appendix A: Similarity Report



Msc_Thesis_Hassan_Ngala_VL (5).pdf

ORIGINALITY REPORT

11%	10%	11%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Strathmore University Student Paper	1%
2	dokumen.pub Internet Source	1%
3	Submitted to Liverpool School of Tropical Medicine Student Paper	<1%
4	www.frontiersin.org Internet Source	<1%
5	Submitted to University of Warwick Student Paper	<1%
6	www.tjnpr.org Internet Source	<1%
7	fastercapital.com Internet Source	<1%
8	revista.estrabao.press Internet Source	<1%
9	Veer Singh, Ashish Kumar. "Visceral Leishmaniasis and Post-kala-azar Dermal	<1%

Figure A.1: Similarity Report - Page 1

VT OMNES
VNVM SINT

Appendix B

Appendix B: Ethical Clearance Release Letter

Strathmore UNIVERSITY

3rd April 2025

Mr Ngala Hassan,
hassan.ngala@strathmore.edu

Dear Mr Ngala,

RE: Application of Spatial Generalized Linear Models in Predicting Visceral Leishmaniasis-Causing Vector: A Case of Turkana County

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU-masters** proposal. Your application reference number is **SU-ISERC2773/25.** The approval period is from **3rd April 2025 to 2nd April 2026.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used.
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-ISERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-ISERC within 72 hours of notification.
iv. Any changes anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-ISERC within 72 hours.
v. Clearance for the export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to the expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days of completion of the study to SU-ISERC.

Before commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology, and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and obtain other clearances needed.

Figure B.1: Ethical clearance

Appendix C

Workplan

C.1 Study plan

Table C.1: Project Timeline and Phases

Phase	Timeline	Activities
Phase 1: Data Collection & Preliminary Analysis	**December 2024 - January 2025**	- Aggregate data collected on sandfly populations, climatic NDVI, geospatial and environmental factors, and Leishmaniasis cases. - Perform initial data cleaning and exploration.
Phase 2: Model Development & Calibration	**February - March 2025**	- Develop spatial and machine learning models to analyze VL-causing sandfly vectors. - Calibrate the models based on collected data and initial training.
Phase 3: Model Comparison & Performance Evaluation	**March 2025**	-Compare the performance of models using specificity, sensitivity, AUC, AIC, and other statistical tools. - Evaluate model accuracy in predicting sandfly population dynamics.
Phase 4: Final Analysis, Report Writing & thesis Submission	**April 2025**	- Conduct final analysis based on model outputs. - Write the final report, thesis and prepare for submission.