Exploring Anopheles species dynamics: a retrospective study of malaria vectors

EEB313: Group F
Zhiqi Tang
Sarah Dickson
Fariha Razzak
Meichen Zhou

Abstract

Malaria is a mosquito-borne infectious disease that is widespread in tropical and subtropical regions. In 2020, there were more than 200 million cases of malaria worldwide resulting in an estimated 627,000 deaths. Approximately 95% of these cases and 97% of deaths occurred in sub-Saharan Africa[1]. In this project, we utilized information acquired from online open-source databases to investigate how geographical distribution and species richness of malaria vector mosquitoes in the *Anopheles* subgenus changes over time, and how they interact with the spread of malaria in Africa. Our results suggest that latitudinal ranges of most species in the subgenus have increased over the last two decades, and mosquito species richness has significantly decreased over the years. However, no statistically significant relationship was found between species richness and malaria infection rate in the African countries covered by our data source. Our findings have relevance in guiding future research concerning humans' ecological impact on the spread of malaria.

Introduction

Malaria is a disease causing significant economic and public health burden globally, especially in tropical and subtropical countries. The WHO estimates there were 619, 000 deaths from malaria in 2021, with African regions accounting for 95% of malaria cases and 97% of malaria deaths[1]. An individual develops malaria when they are bitten by an infectious female Anopheles mosquito carrying parasitic eggs[2]. Once infected, individuals often experience high fever, chills and extreme flu-like symptoms which can be cured if they receive appropriate medical attention. The disease is also preventable; some common approaches aim to prevent individuals from being bitten in the first place using repellents or bed nets. While the physical steps taken to prevent and treat malaria may prove effective in individual cases, they are not able to address the larger issue: the fact that malaria cases are still increasing and causing a significant number of deaths in Africa. This phenomenon may in large part be attributed to climate change, which supports the spread of the malaria vectors through an increasing range of favourable climate and rainfall patterns[3]. Furthermore, climate change is prompting patterns of anthropogenic human activity, such as trapping water through dams, clearing farmlands and forcing individuals to live in subpar housing conditions[4]. These activities can enable disease vectors, such as Anopheline mosquitoes to expand their scope and cause a greater number of infections[4].

This project aims to investigate trends in the Anopheline mosquito genus in African regions to understand the impact on malaria vector spread and disease cases by considering three related hypotheses. The first hypothesis will study the change in latitudinal range of species between 1989-2016. In this case, we are predicting a net expansion in latitudinal range from the equator. Secondly, we aim to investigate the change in species richness across African regions from 2007-2016. We predict that selection pressure will cause a decrease in species richness over time. Lastly, we are considering a potential correlation between regional species richness and the reported number of malaria cases. We predict the presence of a dominant sub-species contributes to the rise in malaria case number.

Methods

Data description

This study analyzed data from two datasets. The first was the Malarial Mosquito Dataset, acquired from Kaggle. This dataset is a compilation of malaria vector literature published from 1898 to 2016, which had 13,464 observations of 14 variables[5]. Each observation was a published study that examined the distribution of species in the Anopheles genus of malaria vector mosquitoes. The columns provide detailed information regarding the studies, including detailed geographic information, study observation period, the Anopheles complexes observed, species and subspecies identified, life cycle status, sampling method employed, laboratory methods used to identify the species, and study title. A detailed explanation of how the *Anopheles* genus is organized into species, complexes and subspecies is available in Appendix 1. For the analyses performed in our study, we modified the original dataset by removing variables unnecessary for our analysis and reformatting. For the first dataset, we only retained information regarding the location, year of observation, sampling methods, species, complex, and subspecies identified. We restructured the data from a long to a wide format according to

the taxonomy of the Anopheles genus. The final dataset had 24239 observations of 12 variables, and a detailed data dictionary can be found in Appendix 2.

The second dataset analyzed in the study was created by merging the aforementioned Malarial Mosquito Dataset with information from the Malaria in Africa dataset, retrieved from the World Bank. This dataset contains information regarding annual malaria occurrences in all African countries from 2007 to 2017, along with data about preventive measures taken in each country[6]. It has 595 observations, and reports country name, year of observation, malaria incidence measured as cases per 1,000 population at risk, total reported cases, as well as many demographic factors about each country and the malaria prevention measures in place. For the second dataset, we excluded columns with preventive measurements information due to a high number of missing observations. We added species richness data calculated from the previous dataset and merged it into this dataset by matching country and year of observation. The final dataset had 301 observations of 8 variables, and a detailed data dictionary can be found in Appendix 2.

Data Analysis

All analyses for this report were performed using R version 4.0.2, with the uploaded packages *tidyverse, readr, knitr, ggpubr, PairedData, lmer, lmerTest, sjmisc, reglass, ggalt, car* and *MuMIn.*

Hypothesis 1: Net latitudinal range was captured for every observed species in Africa in the first and last five-year period data was available. The five year period was established to account for sporadic observations in the dataset, as almost half of species only had one record for their latitude in the first year they were observed. Averaging a five year window allowed a more representative range estimate to be calculated. Specifically, for each subspecies, net latitudinal range at each time point was calculated by subtracting the minimum observed latitude in each respective window from the maximum observed latitude within the window. This hypothesis was tested using a paired t-test at a 95% confidence level, which examines whether the mean difference between pairs of measurements is significantly different ($p < 0.05$ indicates significant difference). The net latitudinal ranges at the earliest and the latest time points for each subspecies were compared. The paired t-test was conducted using the function t.test() from R package “stats”. For the assumptions of the test, normality was checked with the Shapiro-Wilk normality test, using the function shapiro.test(), also from R package “stats”. Equality of variance was checked with the F Test to Compare Two Variances, using the var.test() function from the R package “stats”.

Hypothesis 2: For the second hypothesis, our approach to investigate this question involves finding a “best” mixed effects model and utilizing it to understand the effects of three factors: species richness, country, and time. Since we are considering only a few parameters, a total of three models were compared. Models were fitted using the lmer() function from R package “lme4” (version 1.1-31). The assumptions for linear mixed models were verified, which include a linear relationship between predictors and the response variable, independence of observations, independence of errors, normality of errors, homoscedasticity of errors. The first model

compares species richness against country, and REML is set to true because the random effects are important. Species richness is the response variable, time is a fixed effect, and country is a random effect. The second model only looks at the random effect of a country on species richness. Whereas, the third model compares the effect of time and country on species richness. Upon creating these three separate models to account for different potential scenarios affecting species richness, the AICc method was used to determine model fit fairly without penalizing a model for its level of complexity. Model output (correlation value between time and species richness, significance level) of this best mixed effects model was used to determine if there is statistical evidence suggesting species richness decline over the years.

Hypothesis 3: To test hypothesis 3 we first evaluated a number of different models that are available by our dataset, to identify a best model. In this analysis, the best model was defined as the model with the lowest AIC value. Generalized linear models were fitted using glm() function and created using manual backward selection based on AIC values. The assumptions for linear mixed models were verified, which include a linear relationship between predictors and the response variable, independence of observations, independence of errors, normality of errors, homoscedasticity of errors. The starting model had malaria incidence as the response variable, and included all potential variables of interest as predictors, including species richness, time, country and rural population growth. The backwards selection process continued until the removal of a predictor no longer resulted in a lower AIC value, and this model was determined to be the best generalized linear model. Mixed effects models were fitted also by lmer() function and evaluated in a similar manner. Country was always included as a random effect and two reduced models were compared to the full model, one excluding time and another excluding rural population growth as predictors. Finally, the best generalized linear model was compared to the best mixed model, using AICc as selection criteria. Multicollinearity was checked using variance inflation factor values acquired via the vif() function from “regclass” package (version 1.6). Model output was used to examine whether any predictors had a significant relationship with the outcome, malaria incidence, and the strength and direction of these relationships. Of specific relevance to hypothesis 3 was the nature of the relationship between species richness and malaria incidence.

Results

Hypothesis 1

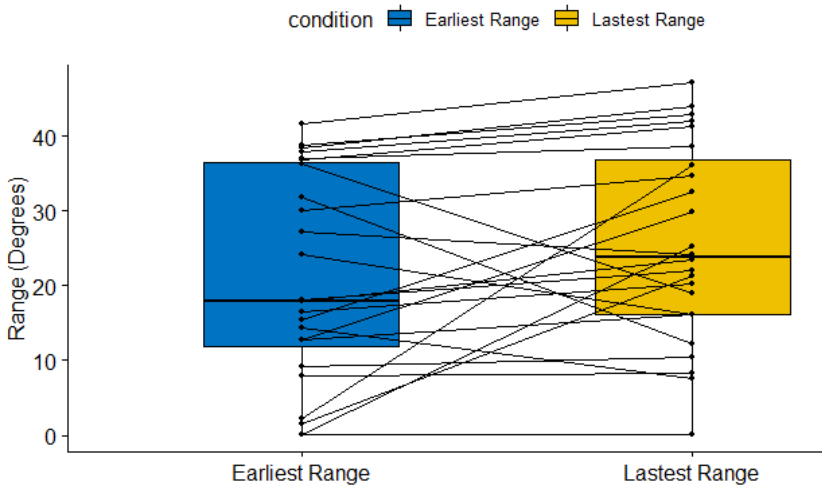
The goal for the first hypothesis is to examine whether the net latitudinal range of African *Anopheles* subgenus species has an increasing trend. The net latitudinal range is defined as the total range on either side of the equator where a species has been observed. Data was structured to show the maximal range in the first and last five-year period a species was observed during (Table 1). The assumptions of normality and equality of variance were tested before the t-test was conducted. The distribution of species’ latest observed ranges was confirmed to be normal using the Shapiro-Wilk normality test. The p-value was 0.5235, indicating that the data’s distribution is not significantly different from a normal distribution. The Shapiro-Wilk test for the distribution of species earliest ranges returned a p-value of 0.049, indicating that it differs significantly from a normal distribution. However, the p-value is very close to 0.05, allowing the assumption of normality to generally apply to this data. The

assumption of equal variance was assessed using an F test to compare two variances, where a significant p-value (<0.05) indicates groups of data have significantly different variances. The resulting p-value was 0.80, meaning a failure to reject the null hypothesis that the group's variances are equal, and validating the assumptions. The paired t-test returned a p-value greater than 0.05, meaning that the mean of two ranges are not significantly different at a 95% confidence level. A paired data plot was used to further examine the changing trend of the net latitudinal range (Figure 1). Based on the paired data plot, 20 out of 24 species underwent an increase in their net latitudinal range, over 84%. Given this finding, we have rejected the null hypothesis that the net latitudinal range of species in the Anopheles subgenus has not significantly increased from 1898 to 2016 in Africa.

Table 1: Earliest and Latest Ranges of Anopheles Sub Species

Species	Earliest Range	Latest Range
An gambiae ss (M Form)	17.8653	23.4342
An gambiae ss (S Form)	12.7940	29.8505
An gambiae ss (Unspecified Sub Species)	30.0190	34.6468
An..melas	9.1527	10.3694
An..merus	18.1333	22.0145
An.arabiensis	41.6093	47.0995
An.bwambae	0.0679	0.1393
An.paludis	36.2319	18.9251
An.ziemanni	31.7430	12.3036
Unspecified Species	37.8950	42.0119
An.funestus.s.s. . . specified.	2.2814	36.0857
An.leesoni	15.4076	32.5046
An.parensis	16.5249	20.1667
An.rivulorum	0.1287	25.1945
An.vaneedeni	1.6002	21.2458
Unspecified Species	38.4936	43.9550
Unspecified Species	27.1468	24.1990
An.hancocki	12.8120	16.1719
An.mascarensis	7.9861	8.3188
An gambiae ss (Unspecified Sub Species)	14.3340	7.5629
An.pharoensis	38.7132	42.7940
An gambiae ss (Unspecified Sub Species)	37.0380	38.6725
An.squamous	36.8140	41.3568
An gambiae ss (Unspecified Sub Species)	24.2398	16.0549

Figure 1: Comparison of Anopheles Species Range during First and Last Observed Period



Hypothesis 2

From the model selection outputs it appears that mixed_model_sr_v_c comparing the effects of species richness by country is the best fit with an AICc value of 1080.13 and four degrees of freedom. In comparison, mixed_model_c had an AICc of 1151.69 and used three degrees of freedom, whereas mixed_model_c_a_y had an AICc value of 931.33 and used 81 degrees of freedom. The dredge function from the package MuMin was also used as a second measure to confirm the results from the AICc, and this function actually ranked mixed_model_sr_v_c second to mixed_model_c. The model we selected also meets the correct assumptions according to the QQ plot for residuals showing an approximately linear trend, suggesting normality of model error. Plotting the residuals showed no discernible pattern, which confirms the requirement of independence of model errors. However, a plot of residuals versus fitted values showed a pattern of diagonal lines, which violates the constant variance of residuals requirement. Considering the individual output from mixed_model_c, it shows that a few African Countries exhibit a significant response. However, both models consistently show that time is the most significant consideration. These results suggest that external geographical or behavioral influences may be important considerations to determine which model is truly most representative of the biological relationship. Overall, the best model shows that species richness is decreasing with time, contributing to a significant response (p-value < 0.05). As such, we are able to reject the null hypothesis.

Table 2: Reporting Mixed Model Statistics

Candidate Models	DF	R^2	AIC_c
Year + (1|Country)	4	0.8029250	1080.13
(1|Country)	3	0.7221861	1151.69
Year * Country	81	0.8654433	931.33

Hypothesis 3

Several general linear and mixed effects models were created using backwards selection, with the goal of describing the relationship between malaria incidence and species richness. AICc was used to select the best fit model at each stage of backwards selection. The most well fitting general linear model had species richness, country of observation and year of observation as predictors. Table 3 describes the model diagnostics for the general linear models. The most well fitting mixed model contained richness, rural population growth and year of observation as fixed effects, and country of observation as a random effect. Table 4 describes the model diagnostics for the mixed effect models. Based on AICc, the best mixed effect model was better at describing variation observed in the data than the best general linear model, so it was selected as the final model. The variance inflation factor for this model was under 5 for all predictor variables, indicating that they were not correlated. A QQ plot for model residuals showed an approximately linear trend, indicating approximate normality of model error. A plot of ordered residuals showed no discernible pattern, confirming the requirement of independence of model errors. A plot of residuals versus fitted values showed a slight fanning pattern, indicating a slight violation of the constant variance of residuals requirement. This should be noted as a limitation. According to the output from this model, year of observation was the only significant predictor of malaria incidence ($p < 0.001$). The coefficient for this relationship was a large negative number, suggesting a strong negative relationship ($\beta_{Year} = 5 \times 10^8$). The variable of interest, species richness, had a p-value of 0.438, which does not allow the rejection of the null hypothesis that there is a significant relationship between species richness and malaria incidence. When explored graphically, the relationship between species richness, year and malaria incidence was not consistent between countries. The seven African countries with the largest population were highlighted for this graphical analysis (Figure 2). Several countries, including the Democratic Republic of Congo, did in fact show the hypothesized relationship between richness and incidence. However, countries such as Tanzania showed the opposite trend. This suggests that a predictor other than species richness should be used to explain trends in malaria incidence data.

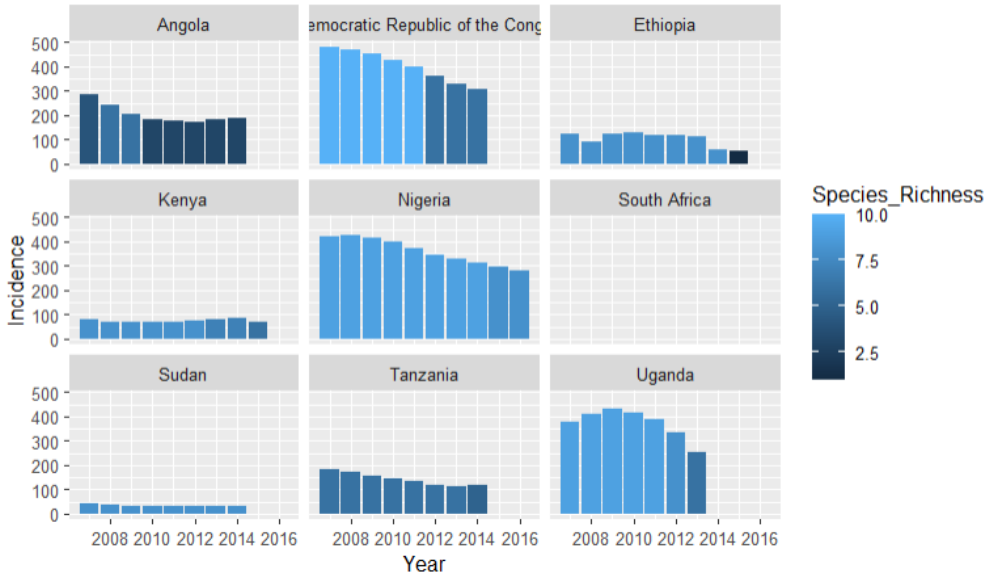
Table 3: Reporting General Linear Model Statistics

Candidate Model	DF	R^2	AIC_c
Richness + Country + Year + Rural Growth	44	0.51	9300
Richness + Year + Rural Growth	5	0.24	9360
Richness + Country + Rural Growth	43	0.41	9354
Richness + Country + Year	43	0.51	9298
Country + Year	42	0.50	9301
Richness + Year	4	0.21	9368
Richness + Country	42	0.41	9354

Table 4: Reporting Mixed Effect Model Statistics

Candidate Model	DF	R^2	AIC_c
Richness + Rural Growth + Year + (1|Country)	6	0.46	9219
Richness + Year + (1|Country)	5	0.47	9244
Richness + Rural Growth + (1|Country)	5	0.46	9313

Figure 2: Bar Plots of Malarial Incidence over Time, colored by species richness



Discussion

Our research paper used *Anopheles* vector distribution and malaria case count data to retrospectively examine how patterns in mosquito subspecies have changed over time, and the health consequences. We found that in 20 of 24 subspecies examined, latitudinal range increased significantly over the examined time period. However, this pattern was not true of all species, and a paired t-test did not find a statistically significant increasing trend. Our report also explored changing regional species richness within the *Anopheles* subgenera. A mixed linear model showed a significant negative relationship between species richness and time, when the country of observation was considered as a random effect. Finally, we examined the relationship between *Anopheles* subgenera species richness and malaria incidence. In a mixed linear model, species richness had no significant impact on case count, even controlling for regional correlations. The analyses in our report have quantified trends in range shift of *Anopheles* malaria vectors, and explored how species richness interacts with time and ecology.

Our study used robust datasets that contained information from all countries in Africa which were partially sourced from World Bank data, a trustworthy source. Exact latitude and longitudes were included, making geographic data very specific. Merging several datasets allowed us to answer novel research questions that had not yet been explored. However, incomplete data was the biggest weakness of this project. While some countries had important

data about prevention measures in place annually, we could not use this variable as it was missing in many observations. Sensitivity could also have been vastly improved by the availability of more regionally specific demographic and health information to control for confounding factors.

There is a strong body of previous literature concerning how the changing range of vector species impacts malaria transmission[7]. This has specific relevance in the context of climate change, where global warming has increased Anopheles migration from regions where malaria is endemic to regions traditionally less affected[7]. Temperature increases allow mosquitoes to survive in new areas, and cause increased biting activity and malaria pathogen growth rate[3,7,8]. In our analysis, we found that the ranges of malarial vector species had significantly increased in the observed window. This finding, when taken into the context of the available literature, suggests that the theorized range change is in fact occurring. According to global health experts and epidemiologists, this is cause for concern. If new patterns in exposure to malaria begin to shape as a consequence of climate change, even more of the population will be exposed to the disease. This is also a trend that is unlikely to reverse in the coming decades, with the global temperature predicted to rise between 0.5 and 25 °C by 2500[9]. Our results are a preliminary step in the important investigation of how climate change and vector-borne diseases interact. Future studies should also examine how climate change interacts with altitude, and may cause previously unaffected mountain communities to experience malaria epidemics[10].

In the course of our analysis, we also found that species richness in Anopheles subgenera observed in African regions was changing during the observed study period, from 2007 to 2016. The composition of Anopheles subspecies in a given region is influenced by environmental factors, such as vegetation, breeding site prevalence and food sources[11]. However, anthropomorphic factors are increasingly being acknowledged as modifiers of subspecies dynamics[11]. Despite this acknowledgement, there remains wide debate as to whether humans increase or decrease vector richness, and the longevity of these changes[11]. A 2017 study examined the effect of dam construction on species richness found a decrease in richness, and an increase in the prevalence of anthropophilic subspecies[11]. Conversely, an American study determined that mosquito richness increased by 10% in Connecticut from 2001 to 2019, including within the Anopheles genus[12]. The uncertainty in this field of research makes it difficult to contextualize our fundings. However, this reinforces the importance of conducting studies to determine how urbanization, human development and mass infrastructure are affecting the diversity of Anopheles subspecies. Future studies would benefit from long-term monitoring windows to ensure that trends detected have a broad perspective.

Understanding and controlling the reservoir species is an opportunity to prevent malaria infections and deaths. We need to understand how the direct and indirect effects of humans are changing how malaria is transmitted by Anopheles vectors. This study provided preliminary results demonstrating significant changes in malaria vector range and species richness over time. Further research needs to be done to analyze the health impacts of changing malaria vector ranges and diversity. This problem will only worsen in the future and presents an opportunity for high impact research.

References

1. World malaria report 2021. [accessed 2022 Dec 8]. https://www.who.int/teams/global-malaria-programme/reports/world-malaria-report-2021

2. Hemming-Schroeder E, Zhong D, Machani M, Nguyen H, Thong S, Kahindi S, Mbogo C, Atieli H, Githeko A, Lehmann T, et al. Ecological drivers of genetic connectivity for African malaria vectors Anopheles gambiae and An. arabiensis. Scientific Reports. 2020;10(1):19946. doi:10.1038/s41598-020-76248-2

3. Afrane YA, Githeko AK, Yan G. The ecology of Anopheles mosquitoes under climate change: case studies from the effects of deforestation in East African highlands. Annals of the New York Academy of Sciences. 2012;1249:204–210. doi:10.1111/j.1749-6632.2011.06432.x

4. Ryan SJ, Lippi CA, Zermoglio F. Shifting transmission risk for malaria in Africa with climate change: a framework for planning and intervention. Malaria Journal. 2020;19(1):170. doi:10.1186/s12936-020-03224-6

5. Boysen, Jacob. Malarial Mosquito Database. [accessed 2022 Dec 6]. https://www.kaggle.com/datasets/jboysen/malaria-mosquito

6. Lydia. Malaria in Africa. [accessed 2022 Dec 6]. https://www.kaggle.com/datasets/lydia70/malaria-in-africa

7. Rossati A, Bargiacchi O, Kroumova V, Zaramella M, Caputo A, Garavelli PL. Climate, environment and transmission of malaria. :12.

8. Afrane YA, Githeko AK, Yan G. The Ecology of Anopheles Mosquitoes under Climate Change: Case Studies from the Effects of Environmental Changes in East Africa Highlands. Annals of the New York Academy of Sciences. 2012;1249:204–210. doi:10.1111/j.1749-6632.2011.06432.x

9. Predictions of Future Global Climate | Center for Science Education. [accessed 2022 Dec 8]. https://scied.ucar.edu/learning-zone/climate-change-impacts/predictions-future-global-climate

10. Adugna T, Getu E, Yewhalaw D. Species diversity and distribution of Anopheles mosquitoes in Bure district, Northwestern Ethiopia. Heliyon. 2020;6(10):e05063. doi:10.1016/j.heliyon.2020.e05063

11. Rodrigures MS, Batista EP, Silva AA, Costa FM, Neto VAS, Gil LHS. Change in Anopheles richness and composition in response to artificial flooding during the creation of the Jirau hydroelectric dam in Porto Velho, Brazil. Malaria Journal. 2017;16(1):87. doi:10.1186/s12936-017-1738-7

12. Petruff TA, McMillan JR, Shepard JJ, Andreadis TG, Armstrong PM. Increased mosquito abundance and species richness in Connecticut, United States 2001–2019. Scientific Reports. 2020;10(1):19287. doi:10.1038/s41598-020-76231-x