DATASCI 120

SLIPPERY SNAILS: SEASONAL INDICES FOR PREDICTING SCHISTOSOMIASIS SPREAD IN WEST AFRICA

June 10, 2024

Syed Ashal Ali
Student ID: 06685228
Stanford University

Slippery Snails: Seasonal Indices for Predicting Schistosomiasis Spread in West Africa

Contents

Summary

More than **200,000** deaths are caused in Africa every year because of schistosomiasis [1].

Schistosomiasis, commonly known as snail fever, is the second most devastating parasitic disease after malaria, with a significant global health impact. Annually, it claims approximately 200,000 lives in Africa alone. This disease is primarily caused by three species of the Schistosoma parasite and is transmitted through contact with contaminated freshwater. The disease is most prevalent in tropical and subtropical regions, particularly in areas with inadequate sanitation and limited access to clean water. The disease causes a range of symptoms from urinary issues to liver and intestinal damage, and even severe long-term complications such as bladder cancer and liver cirrhosis.

This paper aims to explore a dataset encompassing temperature, precipitation, and terrain-related features to develop predictive indices for schistosomiasis prevalence caused by S. haematobium and S. mansoni. The goal is to provide actionable insights that can assist governmental bodies and health departments in Africa to formulate effective prevention strategies.

Based on the exploration presented in this paper, we present the following key findings:

- Geographical and Environmental Impact: The spread of schistosomiasis is heavily influenced by geographical distribution and environmental factors conducive to the lifecycle of the host snails.
- Importance of Seasonal Indices: Seasonal variations critically affect the prevalence of schistosomiasis by influencing the population dynamics of the host snails and the parasites. Identifying these indices is vital for timing prevention efforts optimally.

Based on the analysis presented in this paper, we propose the following key recommendations:

- Targeted Interventions: Utilize the developed seasonal indices and models presented in this paper to implement targeted interventions during specific times of the year, potentially reducing the transmission significantly.
- Enhanced Public Health Education: Increase awareness about the disease transmission and prevention methods, particularly in rural and impoverished areas.
- Improvement in Water Management and Sanitation: Develop infrastructure projects that enhance water quality and sanitation facilities to disrupt the lifecycle of the parasite.

[1] Ahmed, Shadab. Schistosomiasis (Bilharzia), Medscape, 23 Mar. 2023, emedicine.medscape.com/article/228392-overview?form=fpf

The seasonal indices and predictive models proposed in this study offer a cost-effective solution for mitigating the impact of schistosomiasis in Africa. By integrating these models with existing health systems, African countries can enhance their disease prevention strategies and significantly lower the incidence of this life-threatening disease. This approach not only aligns with the resource constraints of developing nations but also paves the way for sustainable health improvements across affected regions.

Continued research and data collection will enable the refinement of the predictive models discussed in this paper. By adapting strategies in response to environmental or social changes that affect disease patterns, the effectiveness of these approaches can be enhanced. Lastly, collaborating with international health organizations will further amplify the impact of the proposed strategies.

Introduction

According to the CDC, schistosomiasis is a disease caused by parasitic worms in humans, which is second only to malaria in terms of its devastating effect [2]. The three main species infecting humans are Schistosoma haematobium, Schistosoma japonicum, and Schistosoma mansoni. The vector for the spread of schistosomiasis are types of freshwater snails. Parasitic worm eggs in the environment due to feces or urine form miracadia in fresh water, which infects snails. From this point, it is very easy for the snails to transmit the disease to humans through various routes. With the role of the vector identified, regions with poor sanitation and a large number of freshwater bodies are most vulnerable to schistosomiasis.

Due to the devastating impacts of this disease, understanding the environments that help it spread can contribute to a more focused and effective prevention and mitigation strategy. Currently, the literature in the space consists of a landscape with several studies that hope to inform disease spread prevention. For instance, there have been geostatistical studies to understand schistosomiasis prevalence [3], as well as spatial mapping of Schistosoma infection risk in east Africa [4]. However, there is a gap in the existing literature when it comes to the seasonal indices that influence the spread of the disease. My project aims to address this by examining the relationship of biological climate and environmental indicators.

Data Set, Research Question, and Hypotheses

The dataset that I will be using contains information about the source of schistosomiasis parasite prevalence and location data from West

[2] CDC. (2021, January 13). CDC - Schistosomiasis. Www.cdc.gov. https://www.cdc.gov/parasites/schistosomiasis/index.html#:~:text=Schistosomiasis%2C%20also%20known%2

[3] Schur, N. et al. (2011). Geostatistical model-based estimates of Schistosomiasis prevalence among individuals aged 20 years in West Africa. PLoS neglected tropical diseases, 5(6), e1194. https://doi.org/10.1371/journal.pntd.0001194

[4] Schur, N. et al. (2013). Spatially explicit Schistosoma infection risk in eastern Africa using Bayesian geostatistical modelling. Acta tropica, 128(2), 365–377. https://doi.org/10.1016/j.actatropica.2011.10.006

Africa. This dataset is sourced from the Global Neglected Tropical Disease database (GNTD). This dataset was accessed from GTND in 2016, and spans the years 1934-2010, with the majority of the data post-1980. The geophysical and climate variables were sampled at the point locations using Google Earth Engine. The data explores the relationship between two different parasites that cause schistosomiasis in humans: Schistosoma mansoni and Schistosoma haematobium, through spatial data, time, 19 climate predictor variables, and 9 terrain predictor variables.

More information about the data such as the data dictionary can be found in the appendix.

Based on this data, the primary research question that I hope to address is "what environmental and terrain features predict the presence of Schistoma mansoni and Schistosoma haematobium most effectively?" In this question, environmental features can be futher divided into temperature related and precipitation related variables. Terrain features include variables that describe the soil and land features (eg. pH of soil, slope of land etc.) Further description of these variables can be found by referring to the data dictionary linked right above.

I have three hypotheses that I hope to prove to accompany my research question investigation:

1. Schistoma mansoni and Schistosoma haematobium have different sets of indices that predict them best.
2. Since schistomiasis is known to spread through water, there is a higher risk of schistomiasis parasite presence closer to pH = 7.
3. Following the same logic as Hypothesis 2, there is a higher risk of schistosomiasis presence in the periods with higher precipitation.

Quality Control

The first step of quality control is imputing the NaN values as per the summary table provided at the start of the exploratory data analysis.

pH, clay, and sand have 169 missing values. Since these factors are so heavily dependent on location, I am going to fill in these missing values with the average of the 10 nearest neighbors of those observations in terms of longitude and latitude.

The biological climate variables, "bio01" to "bio19", have 31 missing values. Upon investigation, I saw that these missing values were all in the same observation. Because of this, I decided to remove those 31 observations entirely as none of them would have recorded data for those 19 variables. My dataset size was reduced from 11,330 observations to 11,299 observations.

"hnd" and "upa" both had 25 missing values. Since the hydrological metrics of a particular region would be similar to the geography

around it, I handled these values through the 10 nearest neighbors average method based on longitude and latitude.

"Slope," "elevation," and "aspect" all had 24 missing values. Since the landscape metrics of a particular region would be similar to the geography around it, I handled these values through the 10 nearest neighbors average method based on longitude and latitude.

Lastly, "gHM" had 55 missing values. Since the human modification and terrain metrics of a particular region would be similar to the locations close to it, I handled these values through the 10 nearest neighbors average method based on longitude and latitude.

This concludes the section of quality control dealing with missing values as there are no more missing values left in my dataset.

There were no obvious outliers in the dataset. I would also include them in my dataset because they could be good examples to uncover some unique characteristic behind the spread of schistosomiasis and parasite prevalence.

Unscrambling the Data: Exploratory Data Analysis

The target variable for my investigation is "percent_pos," which is defined as the percentage that represents the schistomiasis presence at each study site. It is a quantitative variable whose values range from 0 to 100 with units in percentage. The distribution of the values of the variable can be plotted for greater insight as done in Figure 1.

I am interested in exploring the indices that contribute to a higher percentage presence of a particular schistomiasis causing parasite. At this point, the distribution of the two different parasites(S.mansoni and S.haematobium) can provide greater insight into which parasite is more commonly recorded in this dataset. This information is stored as the "parasite_s" variable in my dataset.

Figure 2 highlights that S.haematobium is more commonly recorded in the dataset than S. mansoni. The entire dataset consists of observations from Africa. The distribution of the geographical origins of each observation can be visualized for a better understanding of the spread of the data.

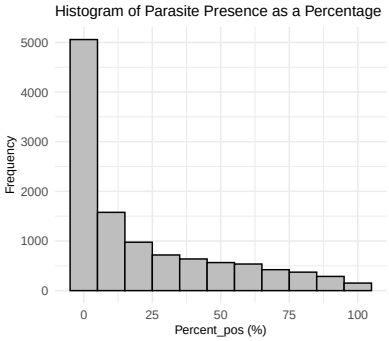

Figure 1: Distribution of the recorded positive test (%) for the prevalence of schistosomiasis prevalence. There is a clear heavier focus on the lower percentages than the higher ones, which communicates that most of the studies had lower percentage positivity of schistosomiasis.

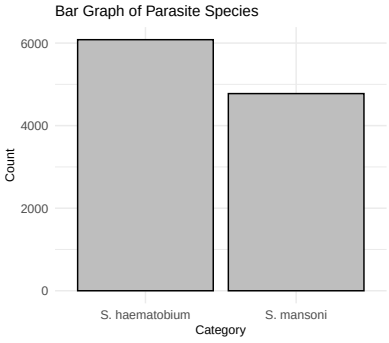

Figure 2: Distribution of the two different schistosomiasis parasites present in the dataset with more observations for S. haematobium than S. mansoni.

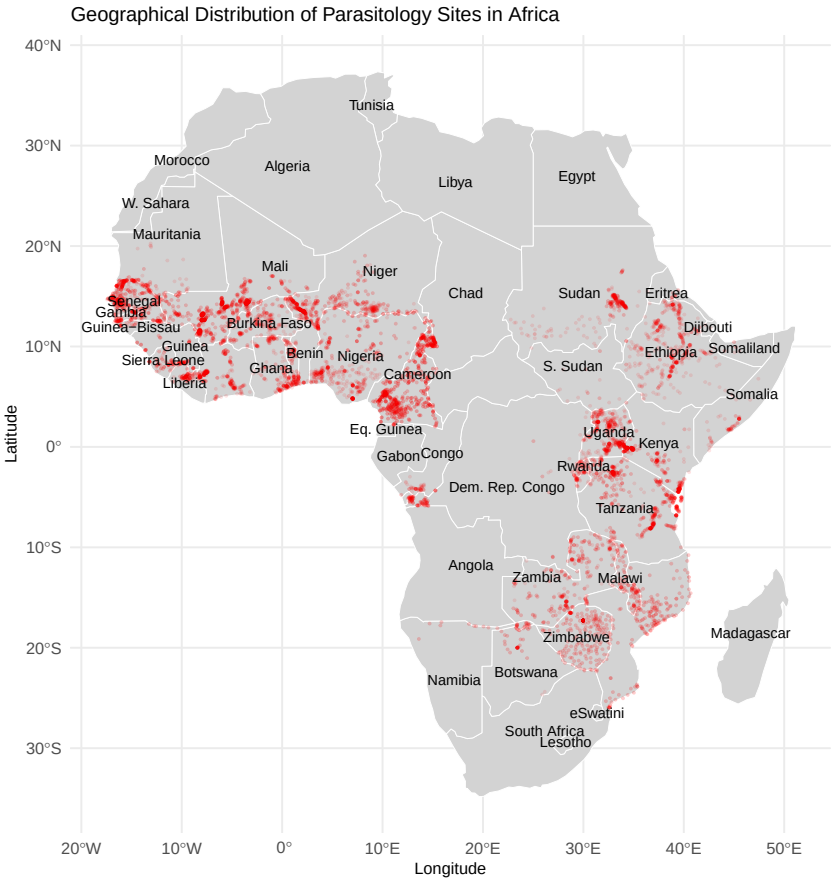

Figure 3: The geographical distribution of parasitology sites in Africa

Figure 3 communicates where the parasitology studies are conducted, with a high concentration being in the South, East and West of the country.

We gain greater insight into the geographical distribution of the different parasite species through Figure 4. It seems that West Africa has a stronger detection of S. haematobium while the east has a stronger detection of S. mansoni. There is also a dominant purple color in the visualization which can be explained by the overlapping points of both studies. The overlapping points (purple color) is common in the West and South.

This insight can inform Hypothesis 1, as different geographical spread may indicate that there will be different factors that contribute to the presence of each specie. To explain this further, there must be some geographical variable (eg. precipitation) that varies from West to East Africa that may explain the difference in the presence of the red vs. blue dots, or the presence of S. haematobium vs S. mansoni.

The distribution of soil pH values across the dataset seems to be close to pH 7, or the neutral pH. This seems to correspond with infor-

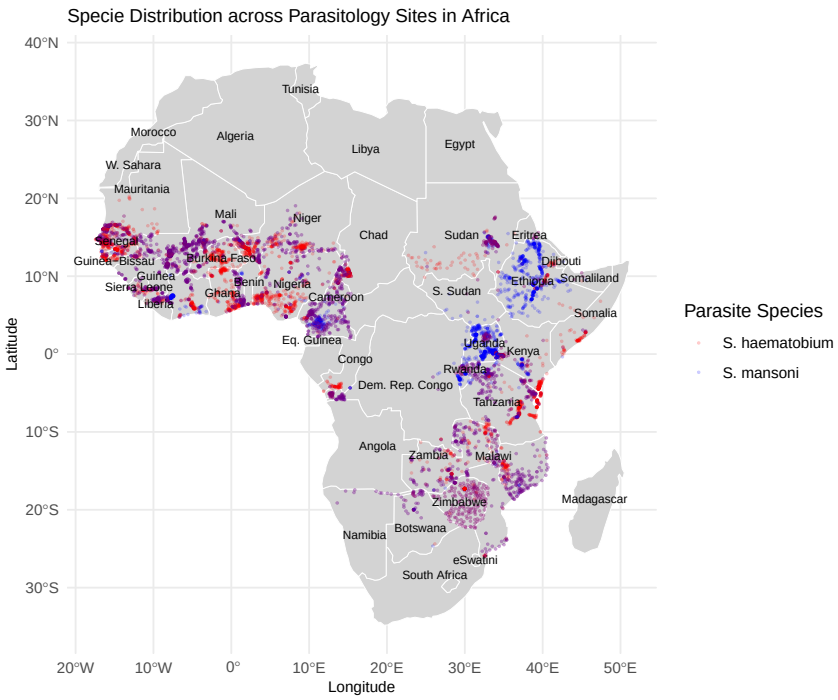

Figure 3: The distribution of different parisotology sites scattered across Africa. A red color indicates presence of S. haematobium while a blue color indicates presence of S. mansoni. The purple shades indicate locations where S. haematobium and S. mansoni are both present as it is produced when the red and blue points overlay each other.

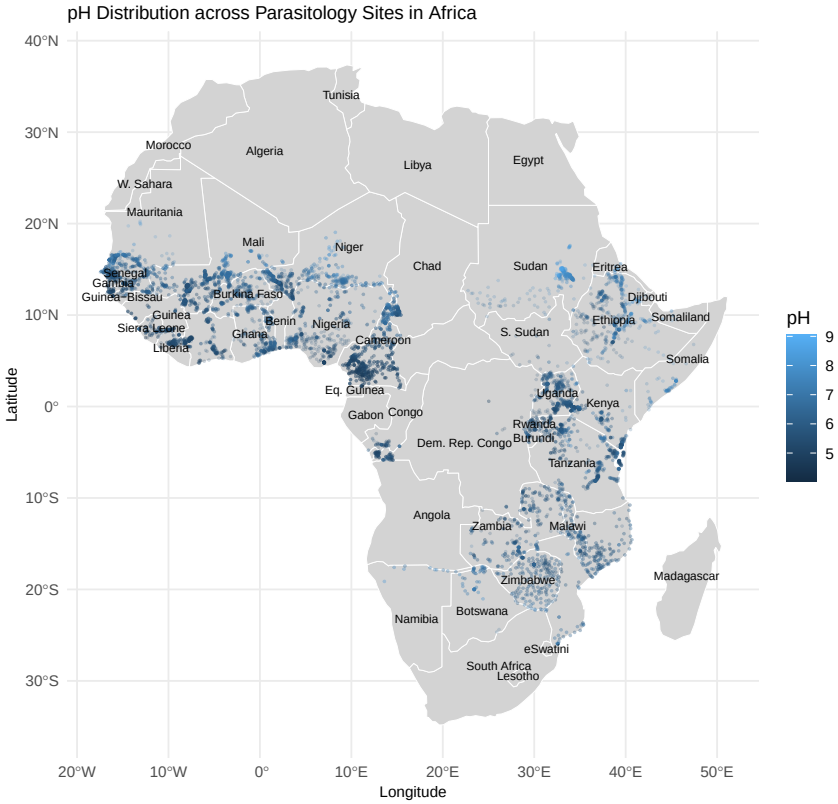

Figure 4: The distribution of different parisotology sites scattered across Africa. A darker blue color indicates presence of more acidic pH, while a lighter blue color indicates presence of more basic pH.

mation in the literature, which states that schistosomiasis prevalence is influenced by soil pH as snail species that act as vectors for the disease prefer a neutral to slightly alkaline pH: this corroborates Hypothesis 2. This preference can be attributed to several ecological and physiological factors. For instance, neutral pH levels typically support better calcium availability in water bodies, which is crucial for the snails' shell formation and overall survival. Additionally, neutral pH conditions are often associated with optimal microbial activity, providing a rich food source for the snails. Therefore, understanding the relationship between soil pH and snail habitats is crucial for predicting and managing schistosomiasis prevalence. By identifying areas with pH levels that support vector snail populations, public health officials can better target interventions such as snail control measures and environmental modifications. This proactive approach can significantly reduce the incidence of schistosomiasis and improve health outcomes in affected regions.

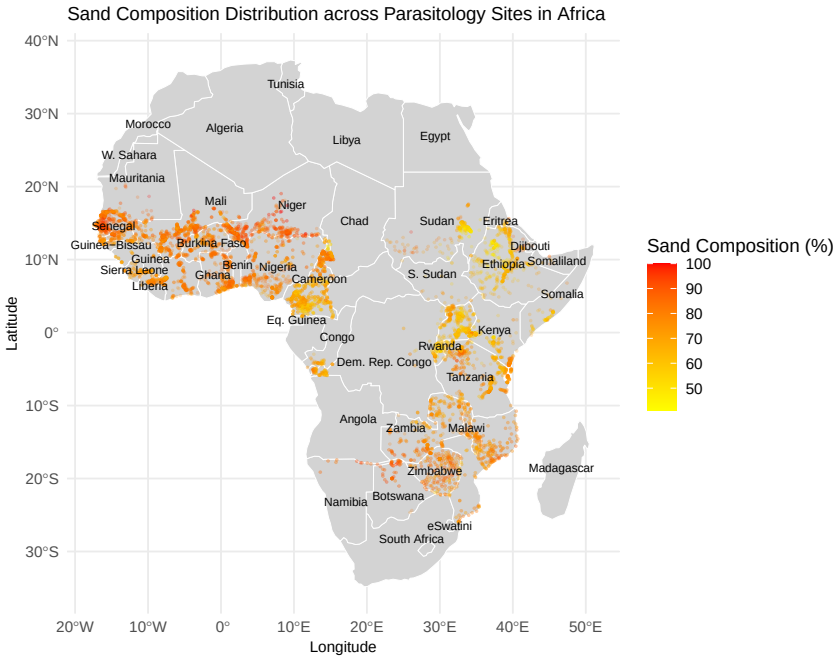

Figure 5: The distribution of different parisotology sites scattered across Africa. A yellow color indicates presence of soil composition containing less sand compared to clay, while a red color indicates presence of soil composition containing more sand compared to clay.

Based on the dataset, the presence of schistosomiasis-causing parasites in the soil prefer environments with greater sand than clay composition showing that the snails prefer sandy soil. Sandy soils typically have better drainage and aeration compared to clay soils. This means that water does not stagnate as easily, creating conditions that are more suitable for snail habitation. Stagnant water can become anoxic (depleted of oxygen), which is detrimental to snail populations.

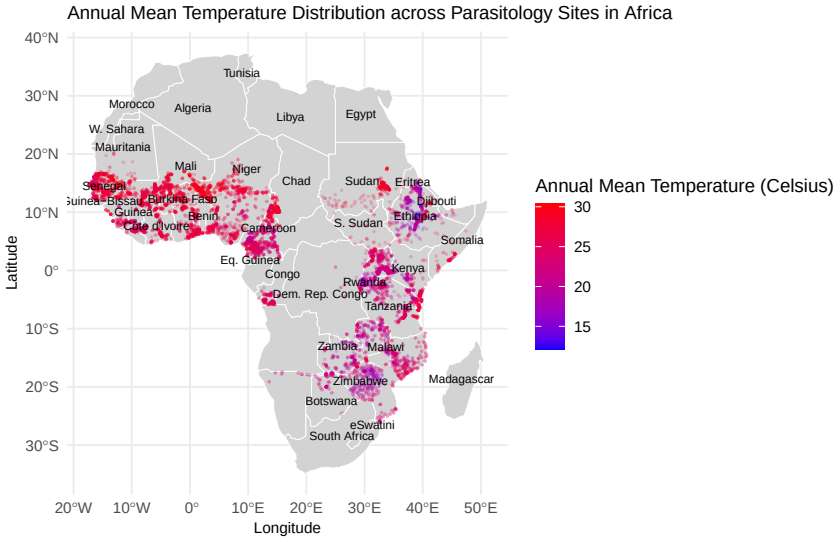

Figure 6: The distribution of different parisotology sites scattered across Africa. A blue color indicates lower annual mean temperature (celsius), while a red color indicates higher annual mean temperature (celsius).

Majority of observations are influenced by higher temperatures. This provides some insight into the causal relationship between temperature and spread of schistosomiasis. A possible explanation is that the snails that act as vectors for the spread of the disease prefer to live in warmer climates. Reviewing literature highlights that this is because the warmer temperature allows the snails to have higher rates of metabolic activity that supports their survival and reproduction.

Almost all observations seem to be recorded with lower annual mean precipitation. Even though I will be considering precipitation related features as part of my indices when I build models to predict schistosomiasis prevalence, it is important to note that an underlying factor impacted by precipitation most is the prevalence of seasonal water bodies. With more water bodies, there is more freshwater and stillwater available for snail populations to use as a habitat and grow in, which could lead to increased spread of the disease.

Deeper Analysis and Modelling: Reducing Dimensionality

Now that we have a more comprehensive understanding of the data, I will be trying to answer my research question, which I will restate here as "what environmental and terrain features predict the presence of Schistoma mansoni and Schistosoma haematobium most effectively?"

Before modelling, it is important to test for multicollinearity in order to detect relationships that might impact model performance. By identifying multicolinear relationships, I can reduce the dimensionality of the data that is being inputted in my model.

First, I will determine which features I want to consider for my

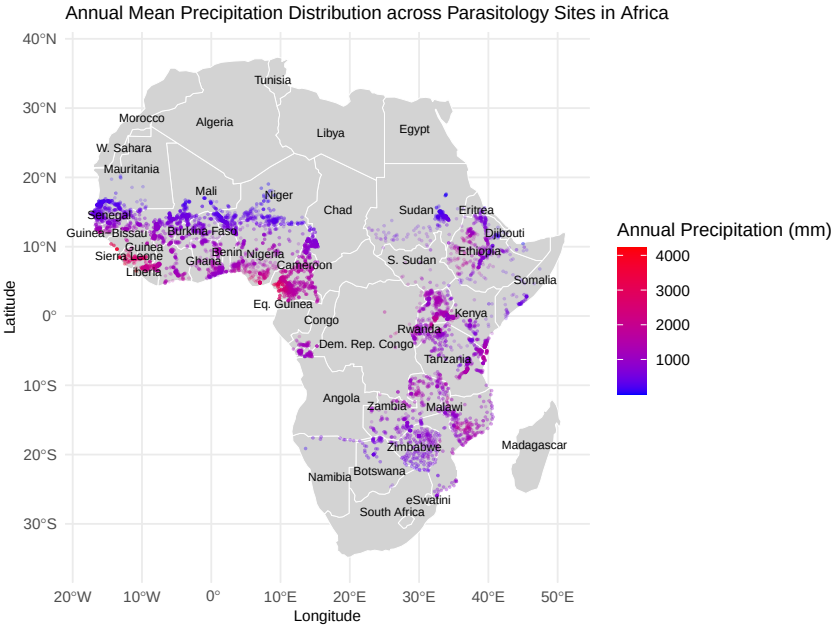

Figure 7: The distribution of different parisotology sites scattered across Africa. A blue color indicates lower annual mean precipitation (mm), while a red color indicates higher annual mean precipitation (mm).

modelling from the temperature related variables part of the environmental features.

Based on Fig. 8 representing the correlation matrix, I will be using bio01 ("Annual Mean Temperature in Degree Celsius"), bio03 ("Isothermality in Percentage"), bio04 ("Temperature Seasonality in Degree Celsius"), and bio07 ("Temperature Annual Range in Degree Celsius"). From my understanding of the topic area as well as the correlation matrix, these are the most important features that include the most information without skewing the colinearity. An example of this is the use of bio03 (Isothermality), which is derived from bio02 ("Mean Diurnal Range in Degree Celsius") and bio07 ("Temperature Annual Range in Degree Celsius"). bio07 ("Temperature Annual Range in Degree Celsius") is another example of this as it is derived from bio05 ("Maximum Temperature of Warmest Month in Degree Celsius") and bio06 ("Minimum Temperature of Coldest Month in Degree Celsius").

Next, I will determine which features I want to consider for my modelling from the precipitation related variables part of the environmental features.

The features that I consider most important based on the correlation matrix in Fig. 9 which is computed on the precipitation related variables include bio12 ("Annual Precipitation in Millimeters"), bio15 ("Precipitation Seasonality as a Coefficient of Variation"), bio16 ("Precipitation of Wettest Quarter in Millimeters"), bio17("Precipitation of Driest Quarter in Millimeters"), bio18("Precipitation of Warmest

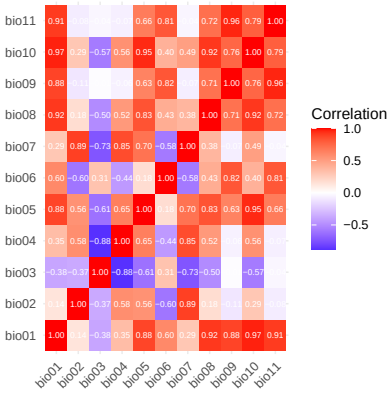

Figure 8: Correlation matrix of all temperature related variables

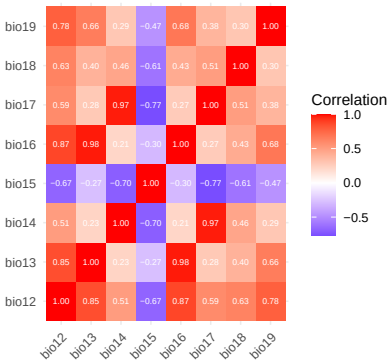

Figure 9: Correlation Matrix of All Precipitation Related Variables

Quarter in Millimeters"), and bio19("Precipitation of Coldest Quarter in Millimeters").

Lastly, I will determine which features I want to consider for my modelling from the terrain features.

I will be using all the terrain variables other than "sand." This is because "clay" and "sand", which both describe the percentage composition of the soil, are very heavily negatively correlated so using one of them will provide enough information for the models.

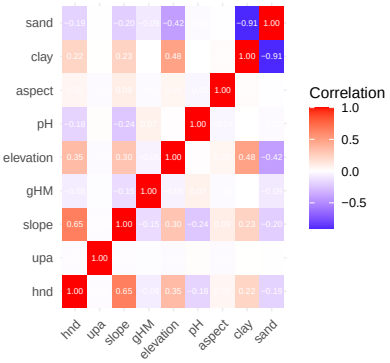

Figure 10: Correlation Matrix of All Terrain Related Variables

Deeper Analysis and Modelling: Building Models

I am going to be building different models to determine the feature importance for both parasite species to determine the best set of features to predict schistosomiasis prevalence. This will involve investigating both species individually.

Models for S. heamatobium

On a primary level, it is important to set benchmarks to see how well the model performs. In order to do this, I decided to run a linear regression on the features that I determined to be important in the earlier section.

According to Fig 12., the temperature related variables have a RMSE benchmark of 28.263%. The benchmark for precipitation related variables in an RMSE of 28.69%. Similarly, the benchmark for terrain related features is an RMSE of 28.512%. When clumping all features together, the overall RMSE benchmark obtained was 28.172%. Each of these benchmarks can be translated as the predictions having an error that is an equivalent percent away from the true value as in the RMSE column. For example, the benchmark for the overall model with features of all types is 28.172%, which means that we are 28.172% off the actual value on average.

To evaluate the importance of each factor that contributes towards predicting parasite prevalence in the overall model, I use the Random Forest ensemble method to evaluate which features are important. The two metrics that I use to decide which features hold most are: "Percentage Increase in Mean Squared Error (%IncMSE)", and "Increase in Node Purity (IncNodePurity)." Percent Increase in Mean Squared Error is calculated by determining the mean squared error (MSE) for different values by using a predictor and then calculating the percentage increase - higher values indicate more important features for making accurate predictions. The Increase in Node Purity is used to describe the homogeneity within a node of a decision tree with higher values referring to more significant features.

Benchmark	Error
Temperature related	28.263%
Precipitation related	28.690%
Terrain related	28.512%
Total	28.172%

Figure 11: Benchmarks of different feature catgories obtained by running a linear regression. The error refers to root mean squared error (RMSE) when predicting the percentage positivity of parasite presence for S. haematobium.

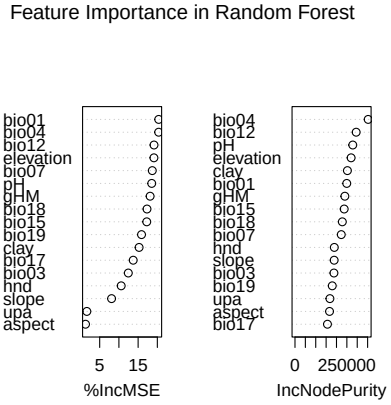

Figure 12: Feature importance graph to help determine the indices that contribute the most for predicting S. haematobium presence.

The RMSE for the random forest taking in all the variables is 23.671%, which is a ~4.5% improvement on the benchmark described in Fig. 12. Based on the plot in Fig. 13, I can tell that some of the most important features of interest for predicting the percentage prevalence of S. haematobium include bio01 ("Annual Mean Temperature in Degree Celsius"), bio04 ("Temperature Seasonality in Degree Celsius"), bio07 ("Temperature Annual Range in Degree Celsius"), bio12 ("Annual Precipitation in Millimeters"), elevation, pH, clay, and gHM. Hence, this is a combination of 3 temperature, 1 precipitation, and 4 terrain related variables.

When considering just the temperature related variables as features for my model, the importance of bio01 ("Annual Mean Temperature in Degree Celsius"), bio04 ("Temperature Seasonality in Degree Celsius"), and bio07 ("Temperature Annual Range in Degree Celsius") was emphasized.

A random forest model applied with just the precipitation related variables as features in showed that bio15 ("Precipitation seasonality") and bio18 ("Precipitation of warmest quarter") are features that I can consider adding to the set of indices of interest I established earlier.

When considering the results obtained from a random forest model with just terrain features as features, I can see that pH, elevation, clay, and gHM may be important to consider for the set of indices that predict the prevalence of S. haematobium best.

After compiling all my insights, I have determined that the set of indices that predict S. haematobium presence best have an RMSE of 23.469%, which is a ~4.7% improvement on my benchmark. The set of indices include bio01 ("Annual Mean Temperature in Degree Celsius"), bio04 ("Temperature Seasonality in Degree Celsius"), bio07 ("Temperature Annual Range in Degree Celsius"), bio12 ("Annual Precipitation in Millimeters"), bio15 ("Precipitation seasonality"), bio18 ("Precipitation of warmest quarter"), elevation, pH, clay, and gHM. Hence, this is a combination of 3 temperature, 3 precipitation, and 4 terrain related variables. A visual representation of these indices is depicted in Fig. 14.

Models for S. mansoni

I will be using the same methodology as used for S. haematobium to develop indices for S. mansoni.

For a first step, I will establish benchmarks using the features I decided upon after dimensionality reduction. The feature categorization and associated benchmarks are depicted in Fig. 19. From the table, The benchmark for the temperature related features is an RMSE of 23.618%. Additionally, the benchmark for the precipitation related

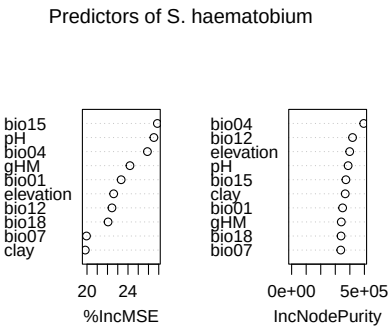

Figure 13: Feature importance graph to help determine indices for predicting S. haematobium presence. The graph was obtained after deciding features based on results from the random forest models run with each category individually considered.

Indices
bio01 (Annual Mean Temperature in Degree Celsius)
bio04 (Temperature Seasonality in Degree Celsius)
bio07 (Temperature Annual Range in Degree Celsius)
bio12 (Annual Precipitation in Millimeters)
bio15 (Precipitation seasonality)
bio18 (Precipitation of warmest quarter)
elevation
pH
clay
gHM

Figure 14: Indices for S. haematobium

Benchmark	Error
Temperature related	23.618%
Precipitation related	23.472%
Terrain related	23.503%
Total	23.039%

Figure 15: Benchmarks of different feature catgories obtained by running a linear regression. The error refers to root mean squared error (RMSE) when predicting the percentage positivity of parasite presence for S. mansoni.

features is 23.472%. Similarly, the benchmark for the terrain related variables is an RMSE of 23.503%. When all of the features of the model are used together, the benchmark RMSE is 23.039%.

To evaluate the importance of each factor, I will use the same approach as with S. haematobium, where I will determine feature importance in a random forest model using %IncMSE and IncNodePurity.

The RMSE for the random forest taking in all the variables is 17.931%, which is a ~5.1% improvement on the benchmark. Based on the plots in Fig. 20, we can tell that some of the most important features of interest for predicting the percentage prevalence of S. mansoni include bio01 ("Annual Mean Temperature in Degree Celsius"), bio07 ("Temperature Annual Range in Degree Celsius"), bio12 ("Annual Precipitation in Millimeters"), bio18 ("Precipitation of warmest quarter"), bio19 ("Precipitation of coldest quarter"), elevation, pH, and clay. Hence, this is a combination of 2 temperature, 2 precipitation, and 3 terrain related variables.

Applying the random forest to just the temperature related variables as features emphasizes the importance of bio01 ("Annual Mean Temperature in Degree Celsius"), bio03 ("Isothermality in Percentage"), bio04 ("Temperature Seasonality in Degree Celsius"), and bio07 ("Temperature Annual Range in Degree Celsius").

The random forest model applied with the precipitation related variables as the only features per shows that bio12 ("Annual precipitation"), bio15 ("Precipitation seasonality"), bio18 ("Precipitation of warmest quarter"), and bio19 ("Precipitation of coldest quarter") are features that I can consider adding to the set of indices of interest I established earlier.

When considering the random forest model with terrain features as the only features, I can see that pH, elevation, and clay may be important to consider for the set of indices that predict the prevalence of S. mansoni best.

After compiling all my insights, I have determined that the set of indices that predict S. mansoni presence best have an RMSE of 17.931%, which is the same analysis as I did in the first model for predicting S.mansoni prevalence. The set of indices include bio01 ("Annual Mean Temperature in Degree Celsius"), bio03 ("Isothermality"), bio04 ("Temperature Seasonality in Degree Celsius"), bio07 ("Temperature Annual Range in Degree Celsius"), bio12 ("Annual Precipitation in Millimeters"), bio15 ("Precipitation seasonality"), bio17 ("Precipitation of driest quarter"), bio18 ("Precipitation of warmest quarter"), bio19 ("Precipitation of coldest quarter"), hnd, upa, slope, gHM, elevation, pH, aspect, clay. A visual representation of these indices is depicted in Fig. 18.

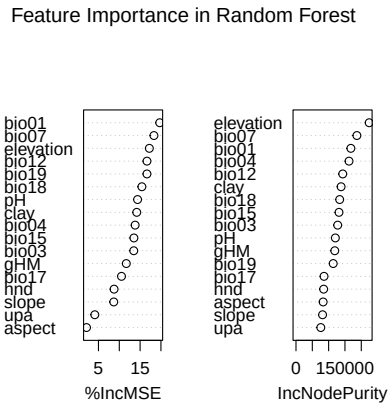

Figure 16: Feature importance graph to help determine the indices that contribute the most for predicting S. mansoni presence.

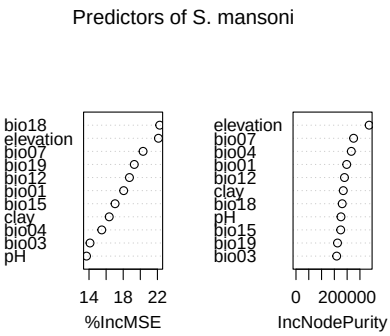

Figure 17: Feature importance graph to help determine indices for predicting S. mansoni presence. The graph was obtained after deciding features based on results from the random forest models run with each category individually considered.

bio01 (Annual Mean Temperature in Degree Celsius)
bio03 (Isothermality)
bio04 (Temperature Seasonality in Degree Celsius)
bio07 (Temperature Annual Range in Degree Celsius)
bio15 (Precipitation seasonality)
bio17 (Precipitation of driest quarter)
bio18 (Precipitation of warmest quarter)
bio19 (Precipitation of coldest quarter)
elevation
hnd
upa
slope
aspect
pH
clay

Figure 18: Indices for S. mansoni

Significance

There are different indices that can be used to predict S. mansoni and S. haematobium. These features range from precipitation related, temperature related, and terrain related variables. From the explanations provided in the exploratory data analysis section, the main relationship that seems to be dictating schistosomiasis prevalence is the impact of environmental features on schistosomiasis vector snail populations. Conditions that promote the growth of the snail populations correspond with a greater prevalence rate. With this correspondence seen across various features, such as precipitation, temperature, soil composition, and soil pH, I believe that the main underlying mechanism of cause and effect is through the various impacts of different environmental conditions on the vector snail species.

The set of indices that predict S. haematobium presence best have an RMSE of 23.469%, which is a ~4.7% improvement on my benchmark. The set of indices include bio01 ("Annual Mean Temperature in Degree Celsius"), bio04 ("Temperature Seasonality in Degree Celsius"), bio07 ("Temperature Annual Range in Degree Celsius"), bio12 ("Annual Precipitation in Millimeters"), bio15 ("Precipitation seasonality"), bio18 ("Precipitation of warmest quarter"), elevation, pH, clay, and gHM. Hence, this is a combination of 3 temperature, 3 precipitation, and 4 terrain related variables.

The set of indices that predict S. mansoni presence best have an RMSE of 17.931%, which is the same analysis as I did above. The set of indices include bio01 ("Annual Mean Temperature in Degree Celsius"), bio03 ("Isothermality"), bio04 ("Temperature Seasonality in Degree Celsius"), bio07 ("Temperature Annual Range in Degree Celsius"), bio12 ("Annual Precipitation in Millimeters"), bio15 ("Precipitation seasonality"), bio17 ("Precipitation of driest quarter"), bio18 ("Precipitation of warmest quarter"), bio19 ("Precipitation of coldest quarter"), hnd, upa, slope, gHM, elevation, pH, aspect, clay.

I believe that the primary conclusion obtained from my research were the indices that can be used to predict the presence of the two different types of schistosomiasis. While it is challenging to extract the exact causal relationship between the indices and the mechanism of spread of the disease, I will discuss the general significance of the indices in terms of acting as predictors for schistosomiasis spread.

A notable observation to consider is that indices of both parasites are a combination of all types of features: temperature, precipitation, and terrain. This reveals that each feature holds significant importance in fostering the conditions that promote the spread of schistosomiasis.

When looking at temperature related features, the indices include

variables such as "Annual Mean Temperature in Degree Celsius," "Temperature Seasonality in Degree Celsius," and "Temperature Annual Range in Degree Celsius." The temperature related predictors all contribute towards a holistic understanding of the temperature and its behavior in that location. Temperature plays an important role in the mechanism of disease spread as the vectors of the disease, the snails, are heavily dependant on temperature for major life processes. Depending on the temperature, the reproductive cycles and growth rates of the snails are varied. According to Kubiriza et al. [5], snail growth depends on temperature as it was reduced at 22 degrees celsius. Furthermore, Barbosa et al. [6] reported that the maximal reproductive rate was observed at 19.9 degrees celsius. Based on the clear role that temperature plays in the life cycles and spread of the vectors of the schistosomiasis disease, there is an obvious connection with the prevalence of the disease. Since schistosomiasis prevalence depends on these snails, temperature conditions that do not favor their growth and spread will inhibit the spread of the disease and schistosomiasis will be less prevalent when compared with environments where snail populations are flourishing.

[5] Kubiriza G.K.et al. Effect of temperature on growth, survival and reproduction of Bulinus nyassanus (Smith, 1877) (Mollusca: Gastropoda) from Lake Malawi. Afr. Zool. 2010;45:315–320. doi: 10.3377/004.045.0210.

[6] Barbosa, N. et al. The effect of seasonal temperature and experimental illumination on reproductive rate in the snail Biomphalaria glabrata. Braz. J. Med. Biol. Res. 1986;20:685–696.

The second prominent indicator of schistosomiasis prevalence were the precipitation related features. These included variables such as "Annual Precipitation in Millimeters" and "Precipitation seasonality," which provide a detailed view of the change in behavior and total precipitation in the location. According to Matsumoto-Takahashi et al.[7], "increase in yearly precipitation decreased the prevalence of schistosomiasis, conversely, the increase in precipitation in the dry season increased the prevalence of schistosomiasis." This highlights the role that the amount of precipitation and its seasonality, especially "Precipitation of driest quarter," play in determining the disease prevalence.

[7] Matsumoto Takahashi. (2023). Impact of precipitation on the prevalence of schistosomiasis mekongi in Lao PDR: Structural equation modelling using Earth observation satellite data. One Health, 16, 100563. https://doi.org/10.1016/j.onehlt.2023.100563

The last predictor of disease prevalence were the terrain related variables, such as "elevation," "pH," and "clay." According to Sumboh et al.[8], soil factors, such as pH, carbon and sandy-loamy texture were associated with high larvae counts ($P<0.001$) while nitrogen and clay content were associated with low counts($P<0.001$)." This finding corresponds with the models used in this paper that have determined that soil factors such as "elevation," "pH," and "clay" are important in predicting the spread of schistosomiasis. It is interesting to see that at the core of all features lies the question of how the snails are being impacted. When the snail population is flourishing, schistosomiasis prevalence is increasing; however, if the snail lifecycle is interrupted, schistosomiasis is effectively curbed. This gives insight into how an optimal strategy to address schistosomiasis should take advantage of the seasons when the snails reproduce and try and interrupt the lifecycle

[8] Sumboh, J. G. et al. (2023). Investigating Environmental Determinants of Soil-Transmitted Helminths Transmission Using GPS Tracking and Metagenomics Technologies. https://doi.org/10.1101/2023.07.17.23292808

so that there are less snails to carry the disease.

Discussion

A considerable limitation of this dataset acquired from the Global Neglected Tropical Diseases database is that it only contains studies for two types of schistosomiasis. This means that the results of this study can only be applied to studying the prevalence of *S. haematobium* and *S. mansoni.* The "Quality Control" section of the report mentions the data limitations. All missing entries that were transformed have all been documented in the section.

Additionally, the data is also representative of Africa alone, which is another limitation on the populations represented by the dataset. This means that the trends captured by the data and uncovered by my analysis may not be reflected in other regions. However, given how Africa is one of the regions where the disease is most prevalent, I believe that my results should be generalizable to a large extent even if the data is only representative of Africa.

Another limitation of the methodology presented in this project is recognizing the role played by seasonal water bodies. While precipitation is a good measure of determining how many seasonal water bodies are created during the rain season and how many bodies of still water exist in the dry season, one of the largest contributors to schistosomiasis prevalence is the existence of water bodies. The dataset did not have any information relating to the presence of water bodies and still water in the locations where the studies were conducted. This may lead to anomalous relationships, such as places with low rain displaying high schistosomiasis prevalence for which the underlying reason maybe the presence of many water bodies, which would be missed by the analysis and not accounted for in the analysis presented. Similarly, there is no documentation of important factors such as hygience and measures of sanitation. A study by Grimes et al. [9] "found that people with safe water and adequate sanitation have significantly lower odds of a Schistosoma infection." Therefore, some sort of quantification of the hygiene and sanitation of these locations would contribute to predicting the spread of the disease.

[9] Grimes, J. E. et al. (2015). The roles of water, sanitation and hygiene in reducing schistosomiasis: A Review. Parasites & Vectors, 8(1). https://doi.org/10.1186/s13071-015-0766-9

It will be difficult to overcome these challenges of sparse data and poor accessibility. The only solution that I can think of involves building on this current database by now tracking added variables such as number of seasonal water bodies, indicators of snail population growth, and signs of hygiene levels. By collecting this new data over time in different observational sites, current models such as the ones presented in this paper can be improved upon. Moreover, the scope of the study can be increased to other parts of the world where schis-

tosomiasis is present to enable a comparative study and determine if trends from one region can be applied to another. Lastly, the study can also collect data for other types of schistosomiasis so more comprehesensive guidelines can be developed.

Based on the proposed indices, future recommendations and steps can be implemented by African health authorities and organizational bodies to prevent the spread of schistosomiasis. Firstly, the developed seasonal indices and models presented in this paper can be implemented for targeted interventions during specific times of the year, potentially reducing the transmission of schistosomiasis significantly. An example of this could be treating freshwater bodies that have been collecting water for some time to get rid of the snail populations in a controlled manner. Secondly, there can be greater investment into providing enhanced public health education: the main goal would be to increase awareness about the schistosomiasis transmission and prevention methods, particularly in rural and impoverished areas. Lastly, there must be an improvement in water management and sanitation facilities. It is of utmost importance to develop infrastructure projects that enhance water quality and sanitation facilities, which disrupt the lifecycle of the parasite that thrives in pH and soil composition conditions as identified in the paper.

With the limitations, potential solutions, and future steps discussed, it is important to consider the ethical implications of the findings of this report. As discussed, most strategies to tackle schistosomiasis are centered around disrupting the vector snail populations. From an ethical standpoint, any intervention leveraging the findings of this report will need to be very carefully executed after consulting with environmental conservationists and experts. This will help in determining how much of the snail population is safe to eliminate to reduce spread of the disease while maintaining the balance of the food chains in the environment. Considering the environmental implications of reducing the snail population in the ecosystem is very ethically significant.

Conclusion

The utilization of seasonal indices and predictive models as outlined in this study presents a compelling opportunity for African nations to effectively combat the scourge of schistosomiasis. Through integration into existing health systems, these innovative approaches offer a cost-effective means to bolster disease prevention efforts and substantially reduce the incidence of this debilitating illness. Not only do these strategies align with the resource limitations often faced by developing nations, but they also lay the foundation for enduring health

advancements throughout affected regions.

Moreover, ongoing research and data collection will play a pivotal role in refining the predictive models discussed herein. By remaining adaptable to environmental and social shifts that influence disease dynamics, these methodologies can evolve to maintain optimal effectiveness. Furthermore, collaboration with international health organizations stands to amplify the impact of these strategies, fostering a united front against schistosomiasis and promoting sustainable health improvements across borders.

References

1. Ahmed, S. Schistosomiasis (Bilharzia), Medscape, 23 Mar. 2023, emedicine.medscape.com/article/228392-overview?form=fpf

2. Barbosa, N. et al. The effect of seasonal temperature and experimental illumination on reproductive rate in the snail Biomphalaria glabrata. Braz. J. Med. Biol. Res. 1986;20:685–696.

3. CDC. (2021, January 13). CDC - Schistosomiasis. Www.cdc.gov. https://www.cdc.gov/parasites/schistosomiasis/index.html#:~:text=Schistosomiasis%2C%20also%20known%20as%20bilharzia

4. Grimes, J. E. et al. (2015). The roles of water, sanitation and hygiene in reducing schistosomiasis: A Review. Parasites & Vectors, 8(1). https://doi.org/10.1186/s13071-015-0766-9

5. Kubiriza G.K. et al. Effect of temperature on growth, survival and reproduction of Bulinus nyassanus (Smith, 1877) (Mollusca: Gastropoda) from Lake Malawi. Afr. Zool. 2010;45:315–320. doi: 10.3377/004.045.0210.

6. Matsumoto, Takahashi. (2023). Impact of precipitation on the prevalence of schistosomiasis mekongi in Lao PDR: Structural equation modelling using Earth observation satellite data. One Health, 16, 100563. https://doi.org/10.1016/j.onehlt.2023.100563

7. Schur, N. et al. (2011). Geostatistical model-based estimates of Schistosomiasis prevalence among individuals aged 20 years in West Africa. PLoS Neglected Tropical Diseases, 5(6), e1194. https://doi.org/10.1371/journal.pntd.0001194

8. Schur, N. et al. Spatially explicit Schistosoma infection risk in eastern Africa using Bayesian geostatistical modelling. Acta Tropica, 128(2), 365–377. https://doi.org/10.1016/j.actatropica.2011.10.006

9. Sumboh, J. G. et al. (2023). Investigating Environmental Determinants of Soil-Transmitted Helminths Transmission Using GPS Tracking and Metagenomics Technologies. https://doi.org/10.1101/2023.07.17.23292808

Appendix

Affiliations

For the SURP-Stats cohort of Summer 2024, I will be researching with the De Leo lab. The De Leo lab is led by Principal Investigator Dr. Giulio De Leo in the Stanford Doerr School of Sustainability. One of the main focuses of the lab is the control and elimination of infectious diseases with an important environmental component in their transmission cycle, with the most notable being schistosomiasis in West Africa. In the lab, I will be working with Project Manager, Andy Chamberlin.

My project for DATASCI 120 is going to be related to the work that I will be doing with the De Leo Lab in the summer so that I can gain some subject matter knowledge and be able to contribute more greatly to the lab with prior experience. The datasets and some analysis techniques were provided by the De Leo lab as these were specific insights they were interested in deriving. Due to these reasons, this project is affiliated with the De Leo Lab.

For the course of this project, I will be receiving feedback from Andy, who can be reached at achamb@stanford.edu.

Data Information

My dataset is linked here.

The data dictionary that defines this dataset is linked here.