THE EFFECT OF DROUGHT ASSOCIATED INDICATORS ON MALARIA IN THE CHOMA DISTRICT OF ZAMBIA

by
Anton Kvit

A thesis submitted to Johns Hopkins University in conformity with the requirements for the degree of Master of Science (Sc.M.)

Baltimore, Maryland

April, 2017
Abstract

**Background:** Malaria is an infectious disease endemic to Zambia that is transmitted by mosquitos. Due to this reliance on mosquitos, malaria incidence is strongly influenced by environmental variables, like vegetation and elevation, and climatic variables like rainfall and temperature. While previous studies have looked at the associations of many of these variables with malaria, few papers have looked at the effect of drought indicators. Furthermore, most papers assumed point temporal lags between weather variables and the malaria incidence, however, the weather occurring over an interval of time within a lag might have a stronger influence on the outcome.

**Methods:** Weekly malaria counts from 14 rural health centers (RHCs) in Choma, in the Southern Province of Zambia, were recorded between 2012 and 2016. All of the climatic and environmental parameters considered in this study were remotely sensed, except relative humidity and daily rainfall, which were recorded via a weather station. For initial analysis, the weekly data from all 14 RHCs was aggregated. Optimal temporal point lags, as well as lag intervals, were determined using cross correlation maps (CCMs). Univariate and multivariate negative binomial regressions were carried out to determine the best predictors of malaria counts. In order to account for nonlinear relationships between malaria counts and predictor variables, models using restricted cubic splines were also considered. A sensitivity analysis was performed by applying the final models to data from two individual RHCs – Macha Hospital and Chitongo.

**Results:** While both the drought index (IRR: 1.017) and the soil moisture (IRR 1.001) were found to be significantly associated with malaria counts, they weren’t the strongest predictors. The point lag model that best predicted malaria counts included rainfall, night temperature, and
Normalized Difference Vegetation Index (NDVI), while the best lag interval model included average rainfall and minimum night temperature. The point lag and lag interval models resulted in similar prediction quality, and while both predicted 2015 malaria counts from 2012-2014 data well (RMSE: 7.577) and (RMSE: 7.13) respectively, both failed to predict the 2016 spike (RMSE: 46.10) and (RMSE: 47.94) respectively. When the nonlinear association of temperature and malaria counts were considered in the model, the prediction of the 2016 malaria spike improved, with (RMSE: 41.10) for the point lag model, and (RMSE: 39.86) for the lag interval model.

**Conclusion:** Drought index and soil moisture were found to be significantly associated with malaria counts, and could be used in future studies that focus on the effects of drought on malaria. However, climatic and environmental variables alone were unable to entirely explain the 2016 increase in malaria counts, even after accounting for nonlinearity.

Thesis Advisor: Frank Curriero, PhD

Thesis Reader: Clive Shiff, PhD
Acknowledgements

Thank you to Dr. Curriero for serving as both my thesis and academic advisor, and helping me so much during my two years at JHSPH. Also, thank you to Dr. Shiff for serving as my thesis reader, and Dr. Moss for his invaluable guidance. I would like to thank the ICEMR team in Macha, especially Dr. Thuma, Dr. Stevenson, and Jailos Lubinda, for all their help during my stay there. Thank you to Tim Shields for teaching me about GIS and his help, and Dr. Kanyangarara for her advice about the thesis. Lastly, I would like to thank the Global Health office at JHSPH for providing the funding for my trip to Macha.
Table of Contents

Abstract...........................................................................................................................................ii
Acknowledgements..........................................................................................................................iv
Table of Contents..............................................................................................................................v
List of Tables.......................................................................................................................................vi
List of Figures.....................................................................................................................................vii
List of Abbreviations..........................................................................................................................viii
Introduction.......................................................................................................................................1
Methods..............................................................................................................................................6
Results...............................................................................................................................................13
Discussion.........................................................................................................................................18
Conclusion..........................................................................................................................................27
References..........................................................................................................................................29
Tables and Figures..............................................................................................................................34
Curriculum Vitae.................................................................................................................................56
List of Tables

Table 1: Units, sources, as well as spatial and temporal resolutions of the variables used in this study

Table 2: Summary statistics all the time-varying variables included in the study, stratified by year and season

Table 3: Univariate regression coefficients adjusted for season using point lags and lag intervals

Table 4: Best fit multivariate regression coefficients using point lags and lag intervals

Table 5: RMSE from predicting 2015, 2016, and 2015-2016 malaria counts from point lag and lag interval models with and without splines
List of Figures

Figure 1: Map of the 14 rural health centers (RHC) included in the study, their Thiessen polygon defined catchment areas, and total malaria counts between 2012 and 2016.

Figure 2a-l: Cross correlation maps of all the used time-varying parameters.

Figure 3a-o: Scatter plots with loess curves and fitted restricted cubic spline curves of all the used time-varying parameters.

Figure 4: A plot predicting 2015-2016 malaria counts using final point lag and lag interval models fit on 2012-2014 data, without accounting for nonlinearity using splines.

Figure 5: A plot predicting 2015-2016 malaria counts using final point lag and lag interval models fit on 2012-2014 data, accounting for nonlinearity using restricted cubic splines.

Figure 6: A plot predicting 2015-2016 malaria counts using 2012-2014 data from Macha Hospital.

Figure 7: A plot predicting 2015-2016 malaria counts using 2012-2014 data from the Chitongo Rural Health Center.
List of Abbreviations

ITN: Insecticide-treated mosquito nets

IRS: Indoor residual spraying

RDT: Rapid diagnostic test

RS: Remote sensing

CCM: Cross correlation map

ICEMR: International Center of Excellence for Malaria Research

RHC: Rural health center

RFE: Rainfall estimate

LST: Land surface temperature

NDVI: Normalized difference vegetation index

VIC: Variable infiltration capacity

MODIS: MODerate-resolution Imaging Spectroradiometer

SMOS: Soil Moisture and Ocean Salinity

IRR: Incidence Risk Ratio

AIC: Akaike Information Criterion

RMSE: Root mean squared error
Introduction

Malaria is a life threatening infectious disease caused by plasmodium parasites and transmitted by mosquitoes, which, despite international efforts, still threatens nearly half of the world’s population.\(^1\) In 2015, there were approximately 212 million new cases of malaria around the world, the vast majority of which occurred in Africa, that resulted in an estimated 429,000 deaths.\(^2\) Children under five years of age are particularly susceptible to malaria, and approximately 303,000 of the malaria deaths in 2015 occurred in children of that age, 292,000 of which happened in Africa. While these numbers are still very large, there has been a 31% reduction in malaria mortality that occurred in Africa alone between 2010 and 2015. This reduction was largely accomplished due to interventions such as insecticide-treated mosquito nets (ITNs) and indoor residual spraying (IRS), better treatment opportunities, along with improved case detection using rapid diagnostic testing (RDT), which allows to quickly and efficiently detect malaria cases even in remote areas by local community health workers.\(^3\)

A major obstacle in reducing the burden of malaria is that it is most prevalent in the nations of Sub-Saharan Africa, many of which struggle economically, and therefore require a cost-efficient allocation of resources and interventions. This makes accurate surveillance of malaria outbreaks and control efforts very important, and numerous malaria risk maps have been developed to help identify hot-spots and areas that most require intervention\(^4-8\). However, this can be a difficult task, since malaria transmission and incidence has been found to be strongly heterogeneous in its distribution across space and time,\(^9,10\) since malaria is characterized by the complex interaction of the plasmodium parasites, mosquito vectors, human populations, and social, environmental, and climatic variables\(^5,11\). Each of these factors adds a layer of complexity
when characterizing and attempting to predict malaria risk, and needs to be considered to obtain the most accurate results.

The plasmodium parasites that are known to cause malaria are *P.falciparum, P.malariae, P.ovale, P.vivax*, and *P.knowlesi*, with *P.falciparum* and *P.vivax* being the most common. A malaria infection begins when a malaria-infected female *Anopheles* mosquito injects sporozoite-stage plasmodium parasites into the blood stream of a human. The sporozoites proceed to infect liver cells, where they mature and replicate for a period of 8-30 days, and burst out back into the blood stream as merozoites, where they infect red blood cells, subsequently creating even more merozoites. This process of replicating and destroying blood cells continues, causing the symptoms of malaria, and also creating the sexual stage of plasmodium – the gametocytes. Another mosquito then bites the infected individual, ingesting the gametocytes, thus becoming infected itself. The gametocytes eventually develop into sporozoites inside the mosquito, perpetuating the malaria cycle.11,12

*Anopheles arabiensis, Anopheles gambiae, and Anopheles funestus*, are some of the major malaria vectors in Africa.13 In order to breed, mosquitoes require water collected in pools, puddles, or slow moving streams.14 Their breeding sites can therefore be associated with parameters such as proximity to water and vegetation indices, which are also indicative of water availability in the area.15 Furthermore, climatic variables such as rainfall, temperature, and humidity have been shown to be associated with malaria transmission through the effect of these factors on mosquito populations and foraging behavior.11,16-18 It can be more difficult for mosquitoes to obtain blood meals when the conditions are dry and hot, while increased humidity can ease foraging behavior. *Anopheles* mosquitoes can most successfully survive and develop at temperatures between 16 °C and 33 °C, with their foraging frequency and behavior affected by
temperature as well\textsuperscript{17,18}. The optimal temperature for malaria transmission is thus currently still a contested topic\textsuperscript{19}. Because temperatures tend to be lower at higher elevations, altitude has been shown to be inversely related with malaria incidence as well.\textsuperscript{20} Relative humidity below 10\% has been shown to result in quick mosquito deaths\textsuperscript{21}, while the optimal relative humidity for parasite development was shown to be between 55\% and 80\%\textsuperscript{22}, and positive associations between humidity and mosquito survival were found only within certain humidity ranges in a laboratory setting\textsuperscript{24}. Many studies have shown that increased rainfall is associated with malaria incidence\textsuperscript{16,25}, however, excessive rainfall has also been shown to wash away breeding sites, suggesting that there is likely a rainfall range optimal for malaria transmission.\textsuperscript{26} The effects of rainfall, humidity, and temperature likely interact\textsuperscript{22,23}, so for example, mosquito foraging during times of low rainfall might be increased by cooler temperatures, since those can result in higher humidity, while in cases of high rainfall, low temperatures might reduce foraging.

Drought is another environmental variable that has been shown to be associated with malaria, though the evidence regarding the exact nature of this relationship is mixed\textsuperscript{27}. Researchers in South America found that malaria mortality is increased by drought in the previous year\textsuperscript{28,29}, likely due to fast-flowing streams reducing to more stationary puddles, that are better suited for mosquito breeding. Other studies in Senegal and Niger found a significant decrease in malaria parasite prevalence in children following prolonged droughts, likely due to the disappearance of \textit{An. funestus}.\textsuperscript{30} A study in Zambia, Southern Africa found reduced \textit{An. arabiensis} and \textit{An. funestus} numbers during the 2004-2005 drought, with a significant increase in \textit{An. arabiensis} when the rains returned next year, but not in \textit{An. funestus}.\textsuperscript{13} While drought can be measured indirectly through rainfall or the remotely sensed Normalized Difference Vegetation Index (NDVI), other methods exist as well, such as measuring the amount of
moisture in the soil, either via satellite\textsuperscript{31} or through mathematical modeling\textsuperscript{32}. A study in Kenya showed that including soil moisture significantly improved the prediction of mosquito biting rates compared to rainfall and NDVI, and thus can enhance predictive modeling of malaria\textsuperscript{33}. Despite the findings that drought has strong effects on malaria, drought and soil moisture indices are included in studies of environmental variables and malaria much less frequently than rainfall, temperature, or NDVI\textsuperscript{34}, leaving them relatively understudied.

Since malaria burden is often highest in developing countries and rural areas that have little infrastructure, collecting environmental or climatic data over long periods of time or over large areas through traditional weather stations and other ground based methods can be difficult or impossible. Proxy measurements through remote sensing (RS) satellites are a way to continuously collect data even in the most inaccessible areas\textsuperscript{34,35}. RS has been successfully utilized in numerous previous studies,\textsuperscript{10,14,15,34,35} and are the source of climatic variables in this study.

Mosquito and parasite life cycles, as well as infected people seeking treatment, take time, adding to the complexity of analyses designed to characterize and predict malaria risk. Therefore, there are temporal lags between environmental circumstances that are most optimal for mosquito and parasite development, and outbreaks of malaria cases recorded in medical facilities. Previous studies have found 1 to 4 month lags between rainfall and peak malaria incidence resulted in the highest correlations, as well as 1-2 month lags for maximum temperature and 2-5 month lags for minimum temperature\textsuperscript{34,36}. Optimal NDVI correlations with malaria incidence were found at 4-6 week lags\textsuperscript{41,38} and optimal relative humidity correlation at a 7 week lag\textsuperscript{39}. No lag information regarding drought index was found, and the optimal lag for the correlation of soil moisture and mosquito biting rates has been reported to be 2 weeks\textsuperscript{33}.33
Much of the research that incorporated time lags in measuring the association of malaria and environmental factors focused on lags restricted to single points in times (for example a week). However, the biology of mosquito and parasite life cycles suggests that environmental factors persisting over periods of time could also affect mosquito populations, and subsequently malaria incidence rates. For example, the average precipitation over a period of five to ten weeks ago could have a different effect on mosquito populations than precipitation exactly ten weeks ago. In order to help determine the best lag intervals, Curriero et al. developed the cross correlation map (CCM), a method to visually identify the lag interval combination with the highest correlation (or other measure of association) to mosquito counts. The researchers found that such an approach helped predict daily mosquito population dynamics better than traditional single time point lags.

This work aims to look at the association of climatic factors related to drought and counts of malaria cases, while filling in some gaps in weather variable choice and temporal lag implementation left by previous studies outlined above. Of particular interest are soil moisture and drought indices that have been rarely looked at before, but could influence the mosquito foraging behaviors, thus complementing more traditional parameters. Furthermore, in addition to single temporal lags, this study will employ cross correlation maps, to determine the influence of climatic variables aggregated over interval time lags on predictive modeling efforts for malaria risk. Malaria cases will be taken from a location in southern Zambia.
Methods

Study Area

The study location consists of 14 rural health centers (RHCs), located in the Choma District of the Southern Province of Zambia. This country in Southern Africa has scaled up its malaria control program in the late 1990’s, through a partnership with the World Health Organization (WHO) and funding of various other donors \(^{43}\), which resulted in significant decline in malaria prevalence in many districts. However, malaria still kills more children under five than any other disease in Zambia, and results in about 8,000 deaths annually, and thus remains a major healthcare challenge.\(^{44}\) The site was determined by The International Center of Excellence for Malaria Research (ICEMR) in Southern Africa to be an area of successful malaria control\(^{45}\), and in 2015 the Southern Province was reported to have the lowest parasite prevalence (0.6\%) of any other province in the country\(^{46}\). The Southern Province was also shown to have the lowest annual rainfall in Zambia (650-800mm)\(^{47}\), with distinct dry seasons spanning approximately May through October, and wet seasons from November to April. Most of the malaria transmission in that region happens through the *Anopheles arabiensis*, and to a lesser extent, *Anopheles funestus*, and the primary malaria parasite is *P.falciparum*.

Malaria and Environmental Data

The malaria cases are confirmed through rapid diagnostic testing and reported by each RHC on a weekly basis via mobile telephone SMS\(^{48}\). In order to estimate the catchment areas for each RHC, first, a buffer was built around each center, with a radius equivalent to the distance to the nearest neighboring RHC. All the buffers were subsequently combined into a single geometry. Thiessen polygons, where boundaries define an area that is closest to each RHC, relative to all other RHC locations, were used to estimate a catchment area for each health center.
In addition to catchment areas, the geographic coordinates of each RHC were recorded as well. All the geographic data management and analysis was conducted using the WGS_1984_UTM_Zone_35S projection.

The rainfall estimate (RFE) data was obtained from the Famine Early Warning Systems Network (FEWS NET). The RFE 2.0 rainfall product is derived from the National Oceanic and Atmospheric Administration (NOAA) African Precipitation Estimation Algorithm, which incorporates infrared data collected by the Meteosat 7 satellite, and passive microwave data from the Special Sensor Microwave/Imager (SSM/I), and the Advanced Microwave Sounding Unit (AMSU) of the Defense Meteorological Satellite Program (DMSP) satellites. The RFE estimates are generated by linearly combining the data from the various satellites through maximum likelihood estimation, which is then bias corrected with the Global Telecommunication System (GTS) rain gauge station data. The resulting rasters were downloaded in dekadal, ten-day average format and 8km spatial resolution.

In order to determine whether rainfall patterns, rather than just weekly rainfall amount, has an effect on malaria counts, hourly rainfall measurements where obtained from a HOBO Micro Station (Onset Computer Corporation, Bourne, MA), located near the Macha Mission Hospital, which is one of the centrally located RHC’s included in the study. For every hour of the week, it was determined whether any rainfall occurred (defined as more than 0.1 mm of precipitation), and the proportion of hours with rain was calculated for every week. This measurement helped determine whether most of the weekly rain occurred in short spurts, or happened over a long period of time. To further explore the effect of rainfall pattern, a binary indicator was assigned to every week that was found to have at least one full day with no rainfall (defined as less than or equal to 0.1 mm of precipitation). Similar binary indicators were
assigned to weeks that had at least two, three, or four consecutive days without rainfall. This measurement was added to explore the hypothesis that excessive rain can wash away mosquito breeding sites\textsuperscript{36}, and therefore, gaps of at least several days between cases of rainfall are needed in order for puddles with optimal mosquito breeding conditions to form. Both the proportion of hours of rainfall in a week (further referred to as rainfall proportion per week), and the consecutive number of days with no rainfall were derived from the hourly precipitation data obtained by the HOBO Micro Station in Macha, and were therefore assumed to be the same throughout the entire study area.

Day and night land surface temperature (LST), which can be used as a proxy for air temperature, was obtained from the MOD11A2 version 6 product of the MODerate-resolution Imaging Spectroradiometer (MODIS) tool aboard NASA’s Terra satellite. The data consisted of 8 day averages, at 1km spatial resolution. The normalized difference vegetation index (NDVI) was obtained from the MOD13A2 MODIS product, and came as 16-day averages at 1km spatial resolution. NDVI is derived from atmosphere-corrected, bidirectional surface reflectance which is translated into vegetation canopy greenness. Visible and near-infrared wavelengths are measured, and the difference in their intensity helps quantify the photosynthetic capacity of vegetation in each given pixel, and thus estimate the amount of vegetation present.\textsuperscript{51} Both NDVI and LST data was reprojected into the WGS_1984 projection using the MODIS Reprojection Tool before subsequent analysis.

Elevation data was obtained from the Shuttle Radar Topography Mission (SRTM) digital elevation model at 90m spatial resolution. Additionally, the ArcHydro Tools module of the ArcGIS software was used with the elevation data to develop a stream network using the Strahler classification\textsuperscript{52}. This classification assigns a value of 1, 2, 3 etc., to streams based on a tributary
hierarchy, where category 1 is assigned to streams at their very beginning, category 2 is assigned when two category 1 streams converge, category 3 is assigned when two category 2 streams converge, and so on.

The drought index used in this analysis is derived from a Variable Infiltration Capacity (VIC) land surface hydrological model, which is forced by precipitation, temperature, and wind speed data, obtained from satellites\textsuperscript{32}. Other required variables in this model are surface and longwave radiation, humidity, and pressure, which are derived from the precipitation and temperature data using empirical regressions. The index is calculated by determining the percentile of the daily average relative soil moisture at each 0.25 degree pixel, and informed by a cumulative probability distribution function derived from historical simulations that range from 1950 to 2008. Finally, the drought index, which is estimated daily, is augmented by satellite-measured soil moisture and vegetation indices.

Soil moisture was taken from the MIR_SMUDP2 tool of the European Space Agency’s Soil Moisture and Ocean Salinity (SMOS) satellite. In order to derive it, an interferometric radiometer aboard the satellite detects wavelengths at the 1400-1427 MHz protected band, and subsequently measures earth surface emissivity, which is related to moisture content in the first few centimeters of soil. The resulting data has a spatial resolution of approximately 40km, and produces soil moisture maps with an accuracy of 0.04 m$^3$/m$^3$ and an approximately 3 day temporal resolution.\textsuperscript{31}

Finally, relative humidity was measured hourly on the ground using the same HOBO Micro Station near the Macha Mission Hospital as used for the hourly measure of rainfall described above.
All the remotely sensed variables (day and night LST, NDVI, RFE, soil moisture, drought index, elevation, stream density) were downloaded as rasters, overplayed over the catchment areas, and the average value of the raster within each catchment area and time interval was recorded. Since the malaria count data was provided on a weekly basis, all the environmental variables were aggregated or disaggregated into weekly time intervals. Disaggregation was conducted by first disaggregating all the climatic parameters into daily measures, assuming that the value for each day was equal to the most recent previous value actually measured, and then aggregating the resulting time series into weekly averages.

Due to cloud cover and satellite malfunctions, raw day LST was missing 8.35% of values, night LST was missing 11.16% of values, soil moisture was missing 5.99% of values, and drought index was missing 0.52% of values. Once these data were aggregated into weeks, about 1.72% of the soil moisture data was missing, and approximately 9.44% of the weather-station measured relative humidity and rainfall data was missing. The missing values that remained after aggregating the data into weeks were imputed through predictive mean matching, which has been previously used to impute missing data in climatology and public health studies. A binary season variable was added to the analysis, with May to September defined as the dry season, and October to April defined as the wet season. Clinic catchment areas, elevation, geographic location, and stream network were all calculated using ArcGIS 10.4.1 (ESRI, Redlands, California), while all other data management and analysis was conducted using R 3.3.2 Statistical Computing Environment.
**Statistical Analysis**

The averages and ranges of malaria counts, as well as all the climatic and environmental variables, were calculated. These summaries were stratified by year and season. The mean differences in environmental variables between the fourteen RHCs were analyzed using ANOVA, stratifying for both season and year. No significant difference between any of the variables was found during the wet season, and significant differences were found for soil moisture (p<0.01) and NDVI (p<0.05) during the dry season. ANOVA analyses conducted for each individual year showed no significant differences between clinics during any of the years. This is likely due to the relatively small study area (approximately 117km by 80km). A significant difference between RHCs was only found for the malaria counts. Additionally, while the Thiessen polygon derived catchment areas were considered to be a reasonable approximation to estimate the weather and environment around each health center, little information was available regarding the population within each catchment area. Since the population density likely varies from RHC to RHC, estimating the catchment area population based on area is not feasible. As a result, all the climatic variables for each week were averaged between the RHC’s, and total weekly malaria counts for the entire study area were considered for analysis (not offset by any population or catchment area denominator).

The remaining analysis focused on identifying and quantifying environmental determinants for the weekly malaria counts with particular focus on the lag timing of environmental effects. Due to persistent over-dispersion, negative binomial regression was used in place of Poisson regression. Both univariate and multivariate models assessments were considered and a manual stepwise backwards elimination approach was used to select the optimal variables for multivariate models. The Akaike Information Criterion (AIC), where the
model resulting in the lowest AIC is considered the best fit, was used as the model selection tool. Regression effect estimates were expressed in terms of incidence risk ratios (IRR), quantifying the expected change in malaria counts when predictor variables changed, and a p-value of <0.05 was considered statistically significant. Point and interval lag summaries of environmental determinants were considered in the regression analysis. Both optimal point lags and interval lags were selected using CCMs, which were built for each independent variable. Each CCM plotted a grid of correlations between the dependent variable of interest, lagged at all possible intervals from zero to thirty weeks, and malaria counts. The brightness and color of each cell in the grid was indicative of the strength and direction of the correlation, which helped to visualize and select the lags that resulted in the highest correlation.

Model fit was further assessed by breaking the data into training and test components, and measuring the agreement of observed and predicted malaria counts in the test component using the root mean square error (RMSE). Since the average weekly malaria count in 2016 (58.92) was found to be over four times the average weekly malaria count in 2012-2015 (13.66), only data from 2012 to 2015 was considered in initial model building. In order to assess the predictive quality of models, the 2012-2014 interval was designated to be the training set, and 2015 was designated as the test set. Models that included only the point lags, as well as models that included lag intervals were both considered. Once the optimal models were selected, they were used to predict malaria counts in both 2015 and 2016.

In order to explore the patterns in the association between predictor variables and malaria counts, scatterplots of each association were built, with an overlaying loess curve to ease interpretation (Figure 3a-o). In cases where it was found to be necessary, restricted cubic splines were selected for the predictor variables that resulted in a curve as close as possible to the loess
curve. Multivariate models incorporating the spline terms were considered in addition to the models described above.

Finally, to further explore the effect of stream density on the association of climactic factors and malaria counts, and as a way to conduct sensitivity analysis regarding the aggregation of area-wide RHC counts, the multivariate regressions using point and interval lags of environmental determinants were repeated using only data from two specific RHCs: Macha Hospital and Chitongo. These were selected since both had relatively high total malaria counts, 525 and 461 respectively, while also significantly differing in stream density (Chitongo stream density is about 59% larger than the one around Macha).

Results

Location of the 14 RHCs along with their catchment areas are presented in Figure 1. The analytical sample included data from these 14 RHC’s, recorded weekly starting from April 22nd, 2012 and ending on September 26th, 2016. A total of 4,993 RDT confirmed malaria cases were recorded during this time at all the health centers, occurring more frequently during the rainy seasons than dry seasons. The list of environmental and climatic variables considered in this analysis are summarized in Table 1 with respect to their unit of measurement, source and spatial and temporal resolution.

In order to assess the effect of elevation, geographic location, and total length of streams within an RHC catchment area, total malaria counts that occurred during the entire study at each clinic were regressed against these variables, using negative binomial regression. No significant associations were found, suggesting the analysis to use weekly weather and environmental parameters averaged over all the 14 RHCs, with the independent variable being the total weekly
sum of malaria cases across all health centers. As expected, the day and night temperatures, as well as RFE, NDVI, drought index, soil moisture, weekly rain proportion, and humidity were all found to be higher during the wet seasons compared to the dry seasons. The rainfall and drought index both decreased starting from the 2014 wet season, suggesting a drought onset at that time that lasted through 2016. These data, stratified by year, as well as wet and dry season, are summarized in Table 2.

For an assessment including only point lags, both univariate models adjusted for season and multivariate models were considered. Optimal point lags were selected using the CCMs as visual aids, by selecting the highest reasonable correlation on the CCM diagonal. Optimal lag points for each variable are marked by red squares in Figure 2a to Figure 2l, while optimal lag intervals are marked by green squares. In univariate analysis adjusted for season, all of the climactic variables considered at appropriate lags were found to be statistically significantly associated with malaria counts after negative binomial regression at 95% confidence level (Table 3). The variables with the highest effects on malaria counts were day and night LST, proportion of rain per week, three or four consecutive days with no rain, and season. An increase in day LST (land surface temperature) by 1°C resulted in a 11.8% decrease in malaria counts in the same week (IRR 0.88, 95% CI 0.864-0.901), while a 1°C increase in night LST resulted in a 27.8% increase in malaria counts 11 weeks later (IRR 1.28, 95% CI 1.233-1.324). A 1% increase in the weekly proportion of hours with rain was associated with a 13.1% increase in malaria counts (IRR 1.131, 95% CI 1.096 1.169) 13 weeks later. RFE, NDVI, drought index, soil moisture, and relative humidity resulted in less than 10% change in malaria count, but were still all positively and statistically significantly associated with it.
Optimal lag intervals were selected using CCMs as visual aids as well, with results summarized in Table 3. As with the point lags, the night LST had one of the largest effects, as the minimum night LST within a 5 to 17 week lag interval resulted in a 28.7% increase in counts for every 1°C increase in temperature (IRR 1.29, 95% CI 1.251-1.325). Since the day LST correlation with malaria counts was the highest when no point lag was applied, no lag interval was calculated for it either. The optimal lag interval for RFE was determined to be 8 to 17 weeks, for NDVI it was 4 to 7 weeks, for soil moisture it was 8 to 12 weeks, for the drought index it was 8 to 14 weeks, and for relative humidity it was 0 to 11 weeks. As with the point lags, all of these variables had significant positive association with malaria counts, and resulted in effects that were less than 10%.

Weeks including one day and two consecutive days with no rainfall were found to have no association with malaria counts 9 weeks later, possibly because there were very few weeks that didn’t have at least two consecutive days with no rain. Controlling for season, three and four consecutive days of no rain in a week were found to reduce malaria counts 8 weeks later by 53% (IRR 0.47, 95% CI 0.3-0.722) and 60% (IRR 0.40, 95% CI 0.269-0.59) respectively. Finally, a 88.9% increase in malaria counts was observed during the wet season as compared to the dry season (IRR 1.88, 95% CI 1.383-2.568).

The final multivariate point lag model, with an AIC of 1054.37 included RFE, NDVI, and night LST at their optimal lags, as described above. When lag interval parameters were considered as well, the model with the lowest AIC of 1050.13 included the average RFE lag interval and the minimum night LST lag interval (Table 4). When 2012-2014 data was used to fit these models, the point lag model resulted in AIC of 735.33 and lag interval model resulted in AIC of 733.74. Predicting 2015 malaria counts using the 2012-2014 fit models resulted in 7.577
RMSE for the point lag model, and 7.132 RMSE for the lag interval model. When both 2015 and 2016 malaria counts were predicted, the point lag model resulted in an RMSE of 46.10, and the lag interval model resulted in an RMSE of 47.93. The RMSE values from predicting 2015, 2016 and 2015-2016 malaria counts from the point lag and lag interval models fit on 2012-2014 and 2012-2015 data are summarized in Table 5. The 2012-2014 fit and 2015-2016 predictions of both the point and lag interval models are depicted in Figure 4.

Including the drought index in the point lag model increased the AIC slightly to 737.11, while slightly reducing the RMSE for predicting 2015 malaria counts to 6.872, but increasing the RMSE for predicting 2015-2016 counts to 47.03. These changes were not significant. Soil moisture was found to be highly correlated with NDVI (0.878) and relative humidity (0.750), and was not found to improve the model AIC or RMSE.

In order to account for the nonlinear associations between predictor variables and malaria counts, tail-restricted cubic splines were considered for each variable and plotted along with scatter plots and overlaying loess curves to visually determine the optimal number of knots (Figure 3a-o). When considering point lags, day LST, night LST, proportion of rain in a week, drought index, and relative humidity all appeared to have distinct non-linear patterns in the association with malaria counts. Consequently, restricted cubic splines with four knots were considered for day LST and night LST, eight knots were considered for proportion of rain in a week, and seven knots were considered for both drought index and relative humidity. Incorporating lag intervals instead of point lags resulted in a reduction of non-linear patterns in the association of predictor variables and malaria counts. However, restricted cubic splines with four knots were still considered for night LST, and splines with seven knots were considered for the drought index. Due to the instability of the model if more than one parameter with splines
was included, spline-adjusted variables were considered one at a time, with no-spline parameters added using a manual forward stepwise selection approach. RMSE from predicting the 2015-2016 malaria counts was used for model selection, resulting in a point lag model that included night LST with splines, RFE, and NDVI. If lag interval parameters were allowed, the model included average lag interval night LST with splines, RFE, and season. The point lag model resulted in an RMSE of 43.20, and the lag interval model resulted in an RMSE of 41.53 (Figure 5) for the 2015-2016 malaria count prediction. The RMSE from predicting 2015 malaria counts from 2012-2014 data was 15.96 for the point lag model, and 7.64 for the lag interval model (Table 5).

The final point lag model including RFE, NDVI, and night LST, and the final lag interval model including the average RFE lag interval and the minimum night LST lag interval, were applied to data from Macha and Chitongo RHCs. Models including splines were not applied to these data, since there were very few weekly malaria counts for most weeks, with an average of only 2.24 cases of malaria per week in Macha, and 1.97 in Chitongo. In Macha, the point lag model resulted in an RMSE of 1.438 for predicting 2015 malaria counts, and an RMSE of 7.241 for predicting malaria counts in 2015-2016. The lag interval model resulted in an RMSE of 1.385 for predicting 2015 counts, and 7.394 for predicting 2015-2016 ones (Figure 6). In Chitongo, the point lag model resulted in an RMSE of 1.769 when predicting 2015 counts, and 3.310 when predicting 2015-2016 counts. The lag interval model RMSE for 2015 prediction was 2.229 and for the 2015-2016 prediction, it was 3.468 (Figure 7).
Discussion

While the majority of time-varying parameters, summarized in Table 3, were found to be significantly associated with malaria counts, no significant associations were found between the time-constant, clinic-varying parameters such as geographic coordinates, elevation, catchment area, and the sum of stream lengths within the catchment area and malaria counts. This can be partially explained by the fact that only 14 RHCs were involved in the study and by the relatively small study area of approximately 8,960 square kilometers. The elevation range in the area was found to be only about 274 meters. This, in addition to the fact that the remotely sensed weather variables did not differ significantly between RHCs, lead the study to focus on temporal within clinics variation, rather than spatial between clinic variation. The season-adjusted univariate regressions using both point lag and lag interval parameters found that all of the time varying parameters were significantly associated with malaria counts, except the binary indicators for one a two consecutive days with no rain in a week. However, there were very few weeks that did not have at least two consecutive days without any rain, which significantly skewed the binary indicator. Once the consecutive number of days with no rain was increased to three and four, at which point only about half of the weeks met the criteria, a significant association with malaria counts became apparent.

The final multivariate models, both point lag and lag interval based, included RFE and night LST, with the point model also including NDVI. These parameters being the most significant predictors of malaria incidence is consistent with previous findings. While drought index and soil moisture were not found to improve models based on these established predictors, both were found to be significantly associated with malaria counts, and did not make the predictions significantly worse either. The drought index successfully captured the drought
that started in Zambia in the fall of 2015, and therefore, if data from a longer time span was available, it could help predict drought-associated malaria spikes that have lags of a year or longer, since previous research has found that drought might influence malaria incidence even a year later. Since soil moisture was highly associated with NDVI, one could be used in place of the other, if one of the measurements isn’t available.

The IRRs using lag intervals and point lags were generally quite similar, as can be seen in Table 3. The biggest difference occurs for the proportion of rain in a week, where the point lag results in a 13% increase in malaria counts with every 1% rain proportion increase, while the lag interval results in a 23% increase in malaria counts. Generally, lag interval parameters where more highly correlated with malaria counts than point lag parameters by around 5% percent, which could explain the small difference between the point lag and lag interval AIC and RMSE values. The insignificant improvement observed when using lag intervals versus point lags could be partially explained by the fact that the correlations grew and decayed very gradually as one increased the lag by one week at a time. Generally, within the 30 week potential lag window considered, there was one peak of maximum correlation between 4 and 14 weeks, at which point the correlation started gradually decreasing. This can be contrasted with the previous studies that used cross correlation maps that dealt with a significantly more variable correlation pattern within the lag interval window, which included several negative and positive peaks. Such a pattern resulted in a less smooth and homogenous CCM, and therefore, areas of unusually high correlation could appear that wouldn’t have been noticed without the tool. As can be seen in Figure 2a-l, in cases where the predictor variable was continuous, the resulting CCM curve was smooth and consistent. In cases of the binary variables, a very distinct pattern formed. In both
cases, there were few distinct “hot-spots” of higher correlation that would significantly improve the prediction performance beyond what was obtained by point lags.

This smoothness in the correlation pattern of the current study might be explained by the seasonality of most of the parameters. Since Zambia has a distinct cool and dry season and a distinct hot and wet season, all of the predictor variables follow this pattern as well, to a certain extent. Therefore, the correlation peaks and troughs are strongly season-dependent, which is consistent with observing a single correlation peak within a 30 week study window. The lag interval approach could be more useful when studying day to day changes in malaria counts within a single season, where the correlation would likely have a less predictable pattern.

The rainfall estimate had both the largest point lag at 14 weeks, and the largest lag interval at 8 to 17 weeks. While this number is on the larger end of the typical lags, previous studies have found associations between malaria incidence and rainfall or temperature at approximately 16 week lags, and lags of up to three months generally have been considered in similar studies. Mosquitos take one to several weeks to develop from eggs, depending on temperature, rainfall, and other weather factors, and several weeks pass before a bitten patient notices the infection and seeks treatment, further expanding the lag. Lags of this length might result from two or more generations of mosquitos needing to be born before peak malaria incidence is recorded. Other parameters, such as NDVI and relative humidity, also had optimal lags consistent with previous studies.

Both the point lag and lag interval models predicted 2015 malaria counts fairly accurately, but failed to predict the spike in 2016 (Figure 4). The proportion of rain per week also failed to predict the 2016 increase, as did the binary variables indicating one through four
consecutive days with no rainfall. The proportion of rain per week was found to be positively associated with malaria counts, controlling for both season and rainfall amount, suggesting that the extent of the rains throughout the week, rather than simply the amount of rain, might influence mosquito development, and thus, malaria counts. The results from three and four consecutive days of no rain show a similar trend, since as the number of days with no rain in a wet-season week increased, the malaria count nine weeks later fell. All of these parameters, however, while varying between wet and dry seasons, did not change significantly in 2016, as was seen with the malaria counts, suggesting that other variables might be driving that increase.

One of the possible explanations for the inability of the weather parameters to predict the malaria spike in 2016 is that ordinary negative binomial regression does not account for the nonlinear patterns in the association between predictor variables and malaria counts. These patterns, summarized for both point lag and lag interval predictors in Figure 3a-o, are particularly prominent for the point lag variables, since lag interval variables are generally derived from averaging the parameters within the entire interval, and therefore, have lower variance. The most accurate 2015-2016 malaria count prediction was based on a model including RFE, NDVI, and restricted cubic spline terms of night LST. While the increase in malaria counts as RFE (Figure 3:a) and NDVI (Figure 3:e) increased appears fairly constant, a more complicated pattern is apparent for night temperature. No association appears to exist between malaria count and night LST when temperature is below 15°C, a positive association exists between 15°C and 19°C, and a negative association occurs when temperature is above 19°C, which is consistent with previous literature\textsuperscript{18}. Accounting for this pattern in the model lead to an improved prediction for the 2016 spike, even though the prediction of 2015 malaria counts was subsequently slightly worse than the prediction based on a the models with no splines (Figure 5). Though other parameters, such
as proportion of rain per week (Figure 3b), day LST (Figure 3d), drought index (Figure 3f), and relative humidity (Figure 3h), also had distinct nonlinear patterns, accounting for them did not lead to an improvement in the prediction of 2015-2016 malaria counts, likely because these variables were initially shown to be less predictive of malaria counts, while night LST was a component of every final model.

Selecting optimal restricted cubic splines for each predictor allows to account for the nonlinearity of that predictor, however, once two variables with splines are included in the model, the spline curve might stop being accurate, since it does not account for the association of malaria counts and the parameter it is describing in the presence of other variables with splines. Therefore, in this study, only one variable with splines was considered a time, and when attempted, models with several splines lead to very inaccurate predictions. Computationally intensive methods exist that allow to fit splines, accounting for the presence of other spline variables in the model, however, these were not performed in this study. Furthermore, as can be seen from Table 5, the quality of the models that included splines was also heavily dependent on the training and test sets. While the point lag spline model was significantly better than the non-spline models at predicting 2015-2016 and 2016 outcomes, it preformed significantly worse at predicting 2015 malaria counts. On the other hand, the lag interval spline model performed similarly well to the non-spline models when predicting 2015 counts, but overestimated malaria counts in 2016 by a very wide margin when it was fit on the 2012-2015 data. The instability of the spline models is also the reason why they were not implemented in predicting malaria counts in Macha and Chitongo.

While the inclusion of splines for the night temperature variable helped improve the prediction of malaria counts in 2016, the count was still significantly underestimated. As can be
seen in Table 1, besides the slight increases in night LST and NDVI during the wet season of 2016, there were slight increases in soil moisture and relative humidity, both of which have nonlinear associations with malaria counts. A complex interaction of these terms could have driven malaria counts up in 2016, however, was not considered in this study. Changes in non-environmental or climatic variables that were not included here could also contribute to the sudden increase. For example, insecticide-treated-nets, many of which have been distributed in the area over three years ago, might have become less effective. Furthermore, the ICEMR team in Choma has caught a larger number of *Anopheles squamosus* mosquitos in 2016, which are considered as potential secondary malaria vectors\(^5\). However, this increase might have also been due to a change in the trapping and sampling strategy. Since malaria is dependent on an intricate combination of climatic, environmental, mosquito, and human factors, it is likely that a combination of them accounted for the spike in 2016.

While most of the independent variables considered did not differ significantly between the RHCs, stream density was an exception, leading to the choice of Macha and Chitongo for individual clinic analysis, as these two health centers had the biggest difference in stream density within their catchment areas. Point lags and lag intervals chosen for both of the centers did not vary significantly from each other or from the point lags and lag intervals selected for the combined analysis. Both Macha and Chitongo, like all the other clinics, had a fairly consistent pattern of seasonal malaria count increases, and an unusually high spike during the wet season of 2016, even though it was less severe when compared to the one seen in the combined analysis. For Macha (Figure 6), the spike was more pronounced than for Chitongo (Figure 7), and therefore, the predictive quality of the 2015-2016 malaria counts in Chitongo was better than the one observed for Macha or the combined model. The point lag model and the lag interval model
preformed similarly as when applied to the combined data, despite the overall low malaria counts at both RHCs and the difference in stream densities, indicating that the selected final models are relatively robust.

There are several important limitations to consider in the context of this study. One of the main constrains was the lack of information on the population within each catchment area. A previous study in the same area, relied on population estimates based on a 2000 census, which at this point was deemed to be highly inaccurate. The rural location of these fourteen RHCs, and the limited resources of the Zambian government do not allow to obtain more recent population data. This would be a significant issue for a multilevel analysis, where each clinic would be assessed independently; however, since this study primarily relied on combining the data from all the health centers, the issue had a limited effect. In order to understand how malaria counts at individual clinics are affected by weather and environmental parameters, remote sensing could be utilized in future studies to estimate the populations within catchment areas, for example by counting the number of households around each health center.

Another limitation that would have made a multilevel analysis more difficult was the similarity of remotely sensed weather factors, such as RFE, temperature, and NDVI, between the rural health centers. This is likely due to the fact that the entire study area wasn’t very big, and generally experienced the same weather. However, since accurate weekly RHC malaria counts were only available for the 14 centers included in the study, there were few options to expand to a larger area. Obtaining data from a larger study area, with more diverse elevation, vegetation, and weather patterns would make the results of this study more broadly applicable.
Due to the rural and low resource setting of this study, satellite remote sensing was the best option to obtain weather information. Most of the obtained variables have been ground truthed and extensively used in previous studies\textsuperscript{10,14,15}, offering accurate results. However, it has been previously noted that high day LST is often overestimated by the MODIS tool, with the degree of error depending on seasonality, ecosystem, solar radiation and cloud cover\textsuperscript{60}. This high overestimation rate has been observed in this study, with maximum day LST values reaching unreasonable outliers of 49°C. Fortunately, this overestimation was consistent during the entire study period, and thus would be consistently correlated with malaria counts. The day LST variable also wasn’t used for any final prediction models, and thus has a minimal impact on the implications of the study results.

A more prevalent limitation of the remotely sensed variables were their spatial and temporal resolutions. While spatial resolutions ultimately mattered less once the data across all RHCs was combined, low temporal resolution likely does have an impact on model accuracy. Since the malaria counts were recorded weekly, parameters such as temperature, NDVI, and rainfall had to be disaggregated, since they were recorded in intervals longer than a week. However, the disaggregation of temperature, which was measured in eight day increments, into week-long increments, likely resulted in minimal damage to data integrity, due to only facing a one day difference. The disaggregation of NDVI, which was measured in 16 day intervals would appear more problematic, however, NDVI, unlike temperature or rainfall, takes a significantly longer time to change, and is likely to be fairly constant within the entire 16 day interval. The 10 day rainfall aggregates provided the biggest concern, especially since such long intervals would not allow to explore the more detailed patterns of rainfall that occur within a week. However, the hourly rainfall data collected near Macha Hospital allowed to circumvent this issue. Applying
rainfall and relative humidity data measured in Macha to the entire area was considered reasonable, since as the remotely sensed weather parameters indicated, the weather in the entire study area was relatively similar. The biggest potential difference in relative humidity might occur in RHCs that are located in the flood plain, such as Chitongo. Remotely sensed relative humidity should be used in future studies in order to properly quantify this difference.

Another limitation of this study is the relatively short temporal range of collected data. Since data collection for all the clinics began in April of 2012, and an outstanding malaria count spike occurred in early 2016, only data from 2012 to 2015 could be used to build and test the models. Additionally, the occurrence of drought in the fall of 2015 could have had an effect on the malaria counts, however, too few years of data were available to analyze this hypothesis. With a longer dataset, it would become clearer as to how unusual or significant the 2016 spike is, and if more than one drought could be included in the model, parameters such as the drought index and soil moisture would become more useful. Finally, the low number of malaria cases, especially when it comes to the analysis of individual clinics, makes the models presented in this study more unstable and dependent on outliers.

Most of the outlined limitations can be addressed in future studies. ICEMR in Southern Africa continues collecting weekly malaria counts at all of the fourteen rural health centers, and therefore, future analyses would have larger datasets for model selection. In order to estimate population within health center catchment areas, remote sensing or surveys could be utilized. To further explore the effects of variable nonlinearity, splines that account for the presence of other variables in the model could be determined using computationally heavy approaches. Finally, since the current study area proved to be homogenous in terms of weather and environmental variables, a larger, more diverse area could be selected for future analysis.
Conclusion

The presented study showed that while drought index and soil moisture aren’t the best predictors of malaria case counts, both are significantly associated with it. Soil moisture, being strongly correlated with one of the best predictors – NDVI, could serve as a proxy when NDVI values are not available. The drought index successfully captured the drought that began in the fall of 2015, and therefore could be useful in a longer-term study of the effect of drought on malaria incidence, where at least several episodes of drought are included.

The point lag and lag interval approaches resulted in similar outcomes and prediction quality in this study, likely due to the heavily season-dependent nature of the considered predictor variables. Utilizing lag intervals might be more effective in studies that take place within a single season, where correlations would have a much more unpredictable pattern.

Precipitation and night temperature, and to a lesser extent, NDVI, were the three main predictors of malaria counts. While the prediction of 2015 count was accurate, both of the models initially failed to predict the 2016 malaria spike. Rain proportion per week, and consecutive days with no rainfall per week, both of which aimed to take a more detailed look at the rain pattern within a week, failed to predict the spike as well. However, when the nonlinear nature of the malaria count and night temperature association was accounted for with restricted cubic splines, the model did a significantly better job at predicting the 2016 increase, though still resulting in a substantial error. Beyond the variables included in this study, the spike might have been the result of ineffective interventions, changes in mosquito populations, or a combination of several climatic, environmental, and human factors.
Overall, the findings presented in this study could inform malaria prediction efforts in Choma district, and provided a basis for future models based on larger and more climatically diverse areas.
References


Tables and Figures

Table 1: The units, sources, as well as spatial and temporal resolutions of the variables utilized in this study. In cases where spatial or temporal resolutions are not relevant to the variable, the corresponding cell is marked with “NA”.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Units</th>
<th>Source</th>
<th>Spatial Resolution</th>
<th>Temporal Resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Day/Night LST</td>
<td>Celsius</td>
<td>MODIS (MOD11A2)</td>
<td>1 km</td>
<td>8-day</td>
</tr>
<tr>
<td>NDVI</td>
<td>%</td>
<td>MODIS (MOD113A2)</td>
<td>1 km</td>
<td>16-day</td>
</tr>
<tr>
<td>RFE</td>
<td>mm</td>
<td>Famine Early Warning Systems Network (Meteosat 7)</td>
<td>8 km</td>
<td>10-day</td>
</tr>
<tr>
<td>Soil Moisture</td>
<td>m³/m³</td>
<td>Soil Moisture and Ocean Salinity (SMOS)</td>
<td>40 km</td>
<td>3-day</td>
</tr>
<tr>
<td>Drought Index</td>
<td>%</td>
<td>Derived from a Variable Infiltration Capacity (VIC) land surface hydrological model</td>
<td>0.25 degrees (~28km)</td>
<td>Daily</td>
</tr>
<tr>
<td>Elevation</td>
<td>m</td>
<td>Shuttle Radar Topography Mission (SRTM)</td>
<td>90 m</td>
<td>NA</td>
</tr>
<tr>
<td>Stream Density</td>
<td>m</td>
<td>SRTM, ArcHydro Tools, Strahler classification (1-6)</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Malaria Case Count (Positive RDT)</td>
<td>count</td>
<td>14 RHC in Macha, Zambia (ICEMR)</td>
<td>NA</td>
<td>Weekly</td>
</tr>
<tr>
<td>Relative Humidity</td>
<td>%</td>
<td>Macha Hospital (ICEMR)</td>
<td>NA</td>
<td>Hourly</td>
</tr>
<tr>
<td>Rainfall Derived Parameters</td>
<td>%, Binary</td>
<td>Macha Hospital (ICEMR)</td>
<td>NA</td>
<td>Hourly, Weekly</td>
</tr>
<tr>
<td>Season</td>
<td>Binary</td>
<td>Dry (May - Sep), Wet (Oct - Apr)</td>
<td>NA</td>
<td>Weekly</td>
</tr>
</tbody>
</table>
Table 2: Summary statistics including the means and ranges of all the time-varying variables included in the study, stratified by year (2012-2016) and season.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dry</td>
<td>Wet</td>
<td>Dry</td>
<td>Wet</td>
<td>Dry</td>
</tr>
<tr>
<td>Malaria Count</td>
<td>11.18</td>
<td>19.2</td>
<td>5.82</td>
<td>13.63</td>
<td>12.77</td>
</tr>
<tr>
<td></td>
<td>(0-68)</td>
<td>(2-109)</td>
<td>(1-21)</td>
<td>(0-40)</td>
<td>(2-57)</td>
</tr>
<tr>
<td>RFE (mm)</td>
<td>0.22</td>
<td>43.03</td>
<td>0.28</td>
<td>43.57</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>(0-3.22)</td>
<td>(0-141.42)</td>
<td>(0-2.43)</td>
<td>(0-149.3)</td>
<td>(0-2.41)</td>
</tr>
<tr>
<td>Day LST (°C)</td>
<td>32.73</td>
<td>36.16</td>
<td>32.8</td>
<td>33.7</td>
<td>33.04</td>
</tr>
<tr>
<td></td>
<td>(26.67-43.3)</td>
<td>(28.61-45.5)</td>
<td>(26.79-44.9)</td>
<td>(25.57-49.4)</td>
<td>(28.86-41.42)</td>
</tr>
<tr>
<td>Night LST (°C)</td>
<td>12.3</td>
<td>18.48</td>
<td>12.33</td>
<td>18.4</td>
<td>12.7</td>
</tr>
<tr>
<td>NDVI</td>
<td>0.32</td>
<td>0.43</td>
<td>0.33</td>
<td>0.51</td>
<td>0.31</td>
</tr>
<tr>
<td></td>
<td>(0.27-0.45)</td>
<td>(0.3-0.57)</td>
<td>(0.27-0.44)</td>
<td>(0.27-0.68)</td>
<td>(0.25-0.44)</td>
</tr>
<tr>
<td>Drought Index (%)</td>
<td>36.84</td>
<td>55.45</td>
<td>48.98</td>
<td>56.97</td>
<td>28.42</td>
</tr>
<tr>
<td></td>
<td>(32.0-41.05)</td>
<td>(27.78-90.3)</td>
<td>(44.85-52.7)</td>
<td>(19.5-92.0)</td>
<td>(25.19-37.4)</td>
</tr>
<tr>
<td>Soil Moisture</td>
<td>0.05</td>
<td>0.16</td>
<td>0.05</td>
<td>0.18</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>(0.03-0.07)</td>
<td>(0.04-0.31)</td>
<td>(0.03-0.07)</td>
<td>(0.03-0.41)</td>
<td>(0.03-0.07)</td>
</tr>
<tr>
<td>Humidity (%)</td>
<td>53.24</td>
<td>68.88</td>
<td>47.63</td>
<td>69.8</td>
<td>51.4</td>
</tr>
<tr>
<td></td>
<td>(30.74-61.5)</td>
<td>(43.66-87.6)</td>
<td>(29.7-60.95)</td>
<td>(32.3-90.87)</td>
<td>(31.54-66.0)</td>
</tr>
<tr>
<td>Proportion of Rain in a Week (%)</td>
<td>0</td>
<td>4.44</td>
<td>0</td>
<td>5.33</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(0-0)</td>
<td>(0-13.1)</td>
<td>(0-0)</td>
<td>(0-25.6)</td>
<td>(0-0)</td>
</tr>
<tr>
<td>1 Day of no Rain</td>
<td>0.85</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2 Days of no Rain</td>
<td>0.85</td>
<td>0.91</td>
<td>1</td>
<td>0.9</td>
<td>0.95</td>
</tr>
<tr>
<td>3 Days of no Rain</td>
<td>0.85</td>
<td>0.82</td>
<td>1</td>
<td>0.57</td>
<td>0.95</td>
</tr>
<tr>
<td>4 Days of No Rain</td>
<td>0.85</td>
<td>0.73</td>
<td>1</td>
<td>0.37</td>
<td>0.95</td>
</tr>
</tbody>
</table>
Figure 1: The fourteen rural health centers (RHC) included in the study, their Thiessen polygon defined catchment areas, and total malaria counts between 2012 and 2016.
Figure 2a: Mean RFE CCM. The optimal point lag was 14 weeks (red square), while the optimal lag interval was 8 to 17 weeks (green square).
Figure 2b: Day LST CCM. The highest reasonable correlation required no temporal lags, thus both the point lag and lag interval was recorded as zero (red square).
Figure 2c: Minimum Night LST CCM. The optimal point lag was 11 weeks (red square), while the optimal lag interval was 5 to 17 weeks (green square).
Figure 2d: Mean NDVI CCM. The optimal point lag was 6 weeks (red square), while the optimal lag interval was 4 to 7 weeks (green square).
Figure 2e: Maximum drought index CCM. The optimal point lag was 11 weeks (red square), while the optimal lag interval was 8 to 14 weeks (green square).
Figure 2f: Mean soil moisture CCM. The optimal point lag was 9 weeks (red square), while the optimal lag interval was 8 to 12 weeks (green square).
Figure 2g: Mean relative humidity CCM. The optimal point lag was 4 weeks (red square), while the optimal lag interval was 0 to 11 weeks (green square).
Figure 2h: Mean proportion of rain per week CCM. The optimal point lag was 13 weeks (red square), while the optimal lag interval was 9 to 13 weeks (green square).
Figure 2i: One day of no rain per week CCM. The optimal point lag was 9 weeks (red square), while the optimal lag interval was 1 to 16 weeks (green square).
Figure 2j: Two consecutive days of no rain per week CCM. The optimal point lag was 9 weeks (red square), while the optimal lag interval was 1 to 16 weeks (green square).
Figure 2k: Three consecutive days of no rain per week CCM. The optimal point lag was 8 weeks (red square), while the optimal lag interval was 0 to 13 weeks (green square).
Figure 2i: Three consecutive days of no rain per week CCM. The optimal point lag was 8 weeks (red square), while the optimal lag interval was 0 to 13 weeks (green square).
Figure 3: Scatter plots of point lag and lag interval versions of predictor variables and malaria counts. Loess curves (blue) are added on top of the plots to help better visualize underlying patterns, and restricted cubic spline curves (black) are fitted that they are most similar to the loess curves.
Table 3: Univariate regression coefficients adjusted for season using both point lags and lag intervals with reported IRR and 95% confidence interval values. Since days of no rain per week are binary variables, they were not considered for the lag interval approach.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Point Lag</th>
<th>IRR</th>
<th>% CI</th>
<th>Lag Interval</th>
<th>IRR</th>
<th>% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>RFE (per 10 mm)</td>
<td>14 weeks</td>
<td>1.021</td>
<td>1.017-1.026</td>
<td>8 to 17 weeks (Avg)</td>
<td>1.03 *</td>
<td>1.026-1.034</td>
</tr>
<tr>
<td>Rain Proportion Per Week</td>
<td>13 weeks</td>
<td>1.131</td>
<td>1.096-1.169</td>
<td>9 to 13 weeks (Avg)</td>
<td>1.227 *</td>
<td>1.188-1.269</td>
</tr>
<tr>
<td>Day LST (per 1 C)</td>
<td>0 weeks</td>
<td>0.881</td>
<td>0.863-0.9</td>
<td>0 weeks</td>
<td>0.881 *</td>
<td>0.863-0.9</td>
</tr>
<tr>
<td>Night LST (per 1 C)</td>
<td>11 weeks</td>
<td>1.275</td>
<td>1.23-1.322</td>
<td>5 to 17 weeks (Min)</td>
<td>1.287 *</td>
<td>1.251-1.325</td>
</tr>
<tr>
<td>NDVI (per 1%)</td>
<td>6 weeks</td>
<td>1.056</td>
<td>1.049-1.063</td>
<td>4 to 7 weeks (Avg)</td>
<td>1.058 *</td>
<td>1.051-1.066</td>
</tr>
<tr>
<td>Drought Index (per 1%)</td>
<td>11 weeks</td>
<td>1.017</td>
<td>1.01-1.024</td>
<td>8 to 14 weeks (Max)</td>
<td>1.026 *</td>
<td>1.019-1.033</td>
</tr>
<tr>
<td>Soil Moisture (per 0.01 m3/m3)</td>
<td>9 weeks</td>
<td>1.001</td>
<td>1.001-1.001</td>
<td>8 to 12 weeks (Avg)</td>
<td>1.086 *</td>
<td>1.074-1.1</td>
</tr>
<tr>
<td>Relative Humidity (per 1%)</td>
<td>4 weeks</td>
<td>1.043</td>
<td>1.037-1.049</td>
<td>0 to 11 weeks (Avg)</td>
<td>1.053 *</td>
<td>1.047-1.06</td>
</tr>
<tr>
<td>1 Day of No Rain Per Week</td>
<td>9 weeks</td>
<td>0.807</td>
<td>0.201-2.208</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 Days of No Rain Per Week</td>
<td>9 weeks</td>
<td>0.768</td>
<td>0.349-1.49</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 Days of No Rain Per Week</td>
<td>8 weeks</td>
<td>0.47 *</td>
<td>0.3-0.722</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 Days of No Rain Per Week</td>
<td>8 weeks</td>
<td>0.400</td>
<td>0.269-0.59</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dry Season</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wet Season</td>
<td></td>
<td>1.888</td>
<td>1.383-2.568</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 4: Best fit multivariate point lag and lag interval regression model coefficients, with reported IRR and 95% confidence interval values.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Point Lag</th>
<th>IRR</th>
<th>% CI</th>
<th>Variable</th>
<th>Lag Interval</th>
<th>IRR</th>
<th>% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>RFE (per 10 mm)</td>
<td>14 weeks</td>
<td>1.006</td>
<td>* 1.001-1.01</td>
<td>RFE (per 10 mm)</td>
<td>8 to 17 weeks (Avg)</td>
<td>1.011</td>
<td>* 1.005-1.018</td>
</tr>
<tr>
<td>Night LST (per 1 C)</td>
<td>11 weeks</td>
<td>1.110</td>
<td>* 1.061-1.163</td>
<td>Night LST (per 1 C)</td>
<td>5 to 17 weeks (Min)</td>
<td>1.194</td>
<td>* 1.13-1.263</td>
</tr>
<tr>
<td>NDVI (per 1%)</td>
<td>6 weeks</td>
<td>1.034</td>
<td>* 1.019-1.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Root mean square error (RMSE) from predicting 2015, 2016, and 2015-2016 malaria counts from point lag and lag interval models with and without splines.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Point Lags (No Splines)</td>
<td>7.577083642</td>
<td>46.10092466</td>
<td>69.3797033</td>
<td>70.69676118</td>
</tr>
<tr>
<td>Lag Intervals (No Splines)</td>
<td>7.131843169</td>
<td>47.93889397</td>
<td>72.24675933</td>
<td>72.77899614</td>
</tr>
<tr>
<td>Point Lags (Splines)</td>
<td>15.95953613</td>
<td>41.09839174</td>
<td>55.99324661</td>
<td>56.08551925</td>
</tr>
<tr>
<td>Lag Intervals (Splines)</td>
<td>7.639992503</td>
<td>39.86304093</td>
<td>52.67176342</td>
<td>160.1833351</td>
</tr>
</tbody>
</table>
Figure 4: Predicting 2015-2016 malaria counts using final point lag and lag interval models fit on 2012-2014 data, without accounting for nonlinearity using splines.

Figure 5: Predicting 2015-2016 malaria counts using final point lag and lag interval models fit on 2012-2014 data, accounting for nonlinearity using restricted cubic splines.
Figure 6: Predicting 2015-2016 malaria counts using 2012-2014 data from Macha Hospital. The same point lag and lag interval models with no splines that were chosen using the combined clinic dataset are utilized here.
Figure 7: Predicting 2015-2016 malaria counts using 2012-2014 data from the Chitongo Rural Health Center. The same point lag and lag interval models with no splines that were chosen using the combined clinic dataset are utilized here.
Anton Kvit

325 East University Pkwy • Baltimore, MD, 21218 • (919) 457-3938 • antonkvit@gmail.com

Education
Master of Science: General Epidemiology and Methodology, GPA: 3.92/4.0 Anticipated May 2017

Johns Hopkins Bloomberg School of Public Health
Course Highlights: Statistical Methods in Public Health I-IV; Spatial Analysis I-IV; Advanced Data Science I & II, Longitudinal Data Analysis, Multilevel Model Analysis, Introduction to Data Management, Analysis of EHR Data

Bachelor of Science: Neurobiology, Psychology, GPA: 3.79/4.0 December 2013

University of Wisconsin – Madison
Course Highlights: Introduction to Programming (Java), Applied Statistics, Regression Analysis
Honors: Phi Beta Kappa

Technical Skills
• Data management and analysis (R, Stata, SQL, Excel, Access)
• Spatial analysis (ArcGIS, R, and SaTScan)
• Data collection and management (REDCap, DHIS2, Open Data Kit)
• GitHub (github.com/akvit1)
• COMSOL finite element analysis software

Data Analysis/Quantitative Experience
Center for a Livable Future
Graduate Research Assistant September 2016 – Present
• Build, manage, and analyze the Baltimore City Healthy Food Availability Index Survey dataset in collaboration with a diverse team of researchers, GIS experts, and city policy advisors, with the goals of informing policy and contributing to scientific publications
• Conduct exploratory data analysis, suggest and implement specific analyses for publications, create figures and tables, update the rest of the team through weekly reports
• Utilize multilevel and multivariate regression, spatial analysis, and latent class analysis approaches

Epidemiology Department
Graduate Research Assistant November 2015 – Present
• Collect, manage and analyze environmental and clinical data from satellites, weather stations, and health centers to better understand associations between environmental conditions and malaria incidence in Zambia
• Write R code for visualizing parameter correlations at varying time lag intervals
• Received a Global Health Established Field Program grant to conduct field work in Macha, Zambia

Covance
Quality Control Reviewer April 2015 – August 2015
• Ensured compliance with standard operation procedures and assessed the quality of data produced in the laboratories of the bioanalytical chemistry department

Laryngeal Physiology Laboratory
Undergraduate Research Assistant May 2011 – August 2013
• Independently chose, learned, and utilized commercial finite element analysis software in order to build mathematical models of vocal folds specific to the interests of the laboratory
• Designed and conducted novel experiments, resulting in several scientific publications
• Assisted in writing, researching literature, and prepared illustrations for grant proposals
Other Research Experience

Laboratory of Structural Biochemistry
Research Assistant
Institute of Bioorganic Chemistry (Moscow, Russia)
September 2013 – December 2014
- Assisted in developing a method to quantify RNA in cell nucleoli using confocal microscopy (LSM510DuaMeta)
- Managed the resulting data, selected and implemented proper statistical tools for analysis
- Learned and implemented cell culture maintenance, gel electrophoresis, qRT-PCR, and other protocols

Research Animal Resources Center
Student Hourly
University of Wisconsin – Madison
July 2010 – May 2011
- Assisted with animal necropsies as a part of a team in a fast-paced environment
- Worked in biosafety level 2 environment, assisted with histopathology slide preparation, learned and implemented parasitology techniques

Professional Development

“Say Yes” English School
English Teacher
Moscow, Russia
January 2014 – February 2015

Greater University Tutoring Service
Chemistry Tutor
Madison, WI
September 2012 – May 2013

Covance Alternative Spring Break Program
Intern
Madison, WI
March 2013
- Gained insight into the drug development and CRO industries, utilized laboratory techniques, and general company organization
- In a team with other interns developed and presented suggestions to improve company efficiency

Volunteering

City Ambulance Station
Emergency Medical Technician
Balashikha, Russia
December 2013 – July 2014

Town of Madison Fire Department
Emergency Medical Technician
Madison, WI
August 2012 – August 2013

University of Wisconsin Hospital – Psychiatric Unit
Group Coordinator
Madison, WI
May 2012 – August 2013

Publications
Presentations


Languages

- Russian (fluent), German (basic)