Ismael Talke
Department of Mathematical Sciences
Montana State University, Bozeman

May 10, 2009

A writing project submitted in partial fulfillment
of the requirements for the degree

Master of Science in Statistics

APPROVAL

of a writing project submitted by

Ismael Talke

This writing project has been read by the writing project director and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the Statistics Faculty.

5/09/2009
Date

Mark Greenwood
Dr. Mark Greenwood
Writing Project Director

Contents

List of Tables

List of Figures

Abstract

The primary objective of this paper is to study the dynamics of Eritrean Malaria cases at the subzone level. Malaria in Eritrea varies both spatially and temporally. Two classes of models are considered in this study, the conventional Generalized Linear Models (GLM) and the Generalized Linear Mixed Models (GLMM) to model count data. A poisson GLM was fit, but due to lack of independence of the observations a high overdispersion parameter resulted. As a result, GLMM with two random intercepts was fit to accomodate the dependence within zones and subzones. Two different random effect structures are considered, one with random intercepts for subzones only and one that nests those random intercepts within a zone level random effect. The GLMM was fit using an offset based on the population size of subzones and considering between a 1 and 4 month rainfall lag to impact malaria rates. The GLMM model for a 3 month rainfall lag with nested random effects was found to be the top model as per the AIC selection criterion. 4 programs for estimating GLMM were considered to fit the model and the estimates and performance of the top model were compared and evaluated across those programs where possible.

Chapter 1

Introduction

1.1 Background

Eritrea is a country located in the Horn of Africa with a total area of approximately 124,000km^2. It shares borders with Sudan, Ethiopia and Djibouti. Most of this is defined as semi-arid tropical convergence zone, indicating low, sporadic rainfall, ranging between 250 millimeters (mm) annually in the lowland areas and 500 mm in the highland region [1]. Since 1996, the country has been divided into six administrative zones. It is further divided into subzones (groups of villages). Currently there are 58 subzones. The 6 zones and their corresponding subzones are presented in Table 1.1.

Eritrea has three epidemiologic strata [6, 9]. There are four lowland zones. The first is western lowland, with altitude 600-1000 meters above sea level and 2 zones (GBarka and Debub). This area is prone to high malaria cases. The second is also lowland located in the east of the country (coastal plain) along the Red Sea at 0-1000 meters above sea level. Northern Red Sea (NRS) and Southern Red Sea (SRS) belong to this strata. The third one includes the remaining two highland zones (Maekel and Anseba), with an average elevation of 1500-2000 meters. These three distinct climate systems affect the rainfall and hence malaria in

Zone	Subzones
Anseba	Aditekelezan, Asmat, Elabered, Geleb, Habero, Hagaz, Halhal, Hamelmalo, Keren, Kerkebet and Sela
Debub	Adikeih, Adiquala, Areza, Dekemhare, Dubaruwa, Emnihaili, Maiaini, Maimine, Mendefera, Segeneiti, Senafe and Tsorona
GBarka	Agordat, Barentu, Dighe, Forto, Gogne, Guluj, Haikota, Lalaygash, Logoanseba, Mensura, Mogolo, Mulki, Shambuko and Tesseney
Maekel	Berik, Galanefhi, NE, NW, SE, Serejeka and SW
NRS	Adobha, Afabet, Dahlak, Foro, Gelalo, Ghinda, Karura, Massawa, Nakfa and Shieb
SRS	Araeta, Assab, So.Denkel and Ce.Denkel

Table 1.1: Zones and Subzones

Eritrea. The months March-May are a season of short rains, which fall mainly in the Eritrean highlands, the July-October rains are seasons that usually bring heavy rains to the south-west of the country including the western escarpments, and the third season is between December - February that occurs mainly in the eastern lowlands and escarpments. In Eritrea, 67% of the population live in malaria endemic areas [6]. The predominant malaria parasite is *Plasmodium Flaciparum* and is mainly transmitted by *Anopheles Arabiensis* 84% and *P. Vivax* 16% [6].

1.2 Significance of the Study

The absence of informative models for malaria counts in Eritrea can make it hard for policy makers to intervene and allocate the right resources. This is because without the model it becomes hard if not impossible to seperate seasonal changes from other patterns of malaria dynamics. Therefore, studying the dynamics of malaria at the subzone level and modeling it will help the Malaria Control Program at the Ministry of Health in planning their work and to evaluate success of control efforts.

1.3 Objectives of the Study.

The objectives of this study are to:

1. Understand the dynamics of Eritrean malaria cases at the Sub-Zone level.
2. Study the relationship between rainfall and malaria cases and determine the optimal rainfall lag to predict increases in malaria cases.
3. Study the distribution of malaria cases between zones and subzones over time.

This is done by understanding the factors that contribute to the incidence of malaria cases and how the seasonal changes relate to the patterns of malaria cases. Generalized Linear Mixed Models (GLMM) are used to identify the optimal lag between rainfall and malaria counts and estimate trends in cases. After identifying an optimal rainfall lag, we can then estimate the impact of rainfall on malaria incidence.

Various methods have been used in modeling malaria dynamics by different authors [8] and [10]. [10] used weekly data and applied robust poisson regression by categorizing the districts into two climatic zones, hot and cold. [8] used time series analysis models such as the Seasonal Multiplicative Autoregressive Integrated Moving Average (SARIMA) model. In this study, GLMM are considered. Different versions of GLMM are available in R, glmer, lmer, glmmPQL and glmmML can call estimate versions of GLMMs. The rest of the paper is organized as follows. The next section describes the data set used in this study and visually explores the data. This is followed by the description of the models used in this study. A presentation and interpretation of the results are then discussed in some detail and finally conclusions and implications together with some suggestions for future work will be presented.

Chapter 2

Data and Descriptive Data Analysis

The data consists of monthly records on malaria outpatients and inpatients from all the health facilities, reported monthly to the National Malaria Control Program of the Ministry Health of Eritrea, by subzone, for the period January 2000-December 2007. In Eritrea, there are 6 zones and 58 subzones and the prevalence of malaria varies spatially and temporally. The 2000 population estimate for each subzone obtained from National Health Management Information System (NHMIS), Ministry of Health, will be used in this study for adjustment of malaria incidence at the subzone level, as was used in [5]. This population size might not necessarily be accurate and does not reflect the population growth over the period of study. The way adjustment is made for the population size will be discussed in the subsequent chapters.

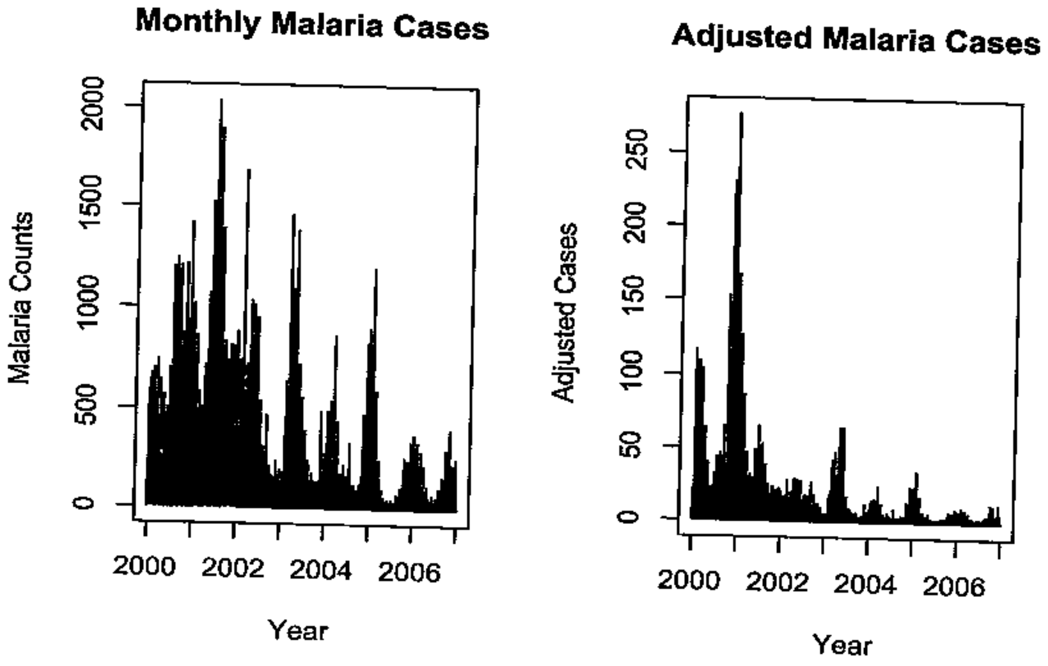


Figure 2.1: Time Series Plot of Malaria Cases

Fig. 2.1 shows the time series plot of the adjusted and unadjusted monthly malaria cases in each subzone over the period of 8 years. In total there are 5557 monthly observations used in this study. Explanations on how adjustment is made will follow in the subsequent chapters. Figures 2.2, 2.3, 2.4, 2.5, 2.6, and 2.7 show the overall trend of the plot at the subzone level for each zone. As seen in the plots, the malaria incidence declines as we progress over the study period. The decrease in the incidence of malaria in Eritrea can be attributed to a number of factors. The achievement might be due to the increase in health facilities, training given by the Malaria Control Program to the village health agents, coordinated endeavors made to control the infection, the distribution of sufficient and effective anti-malaria medicine, the provision of malaria nets to the population free of charge, as well as the active popular participation in maintaining environmental sanitation.

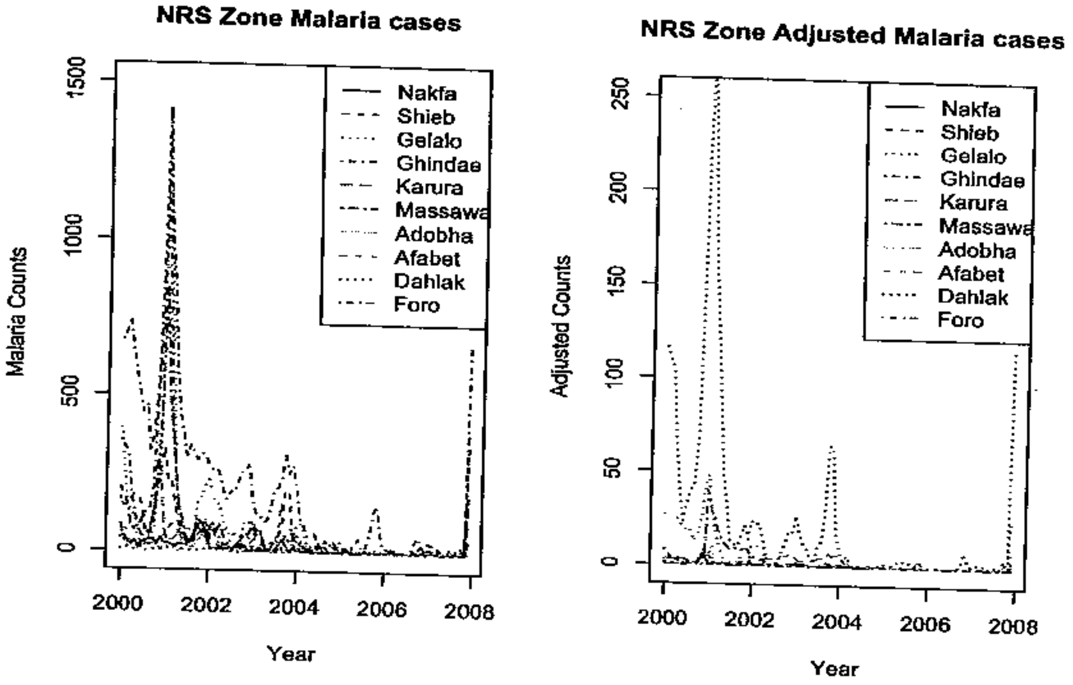


Figure 2.2: NRS Subzones Monthly Malaria Times series plot

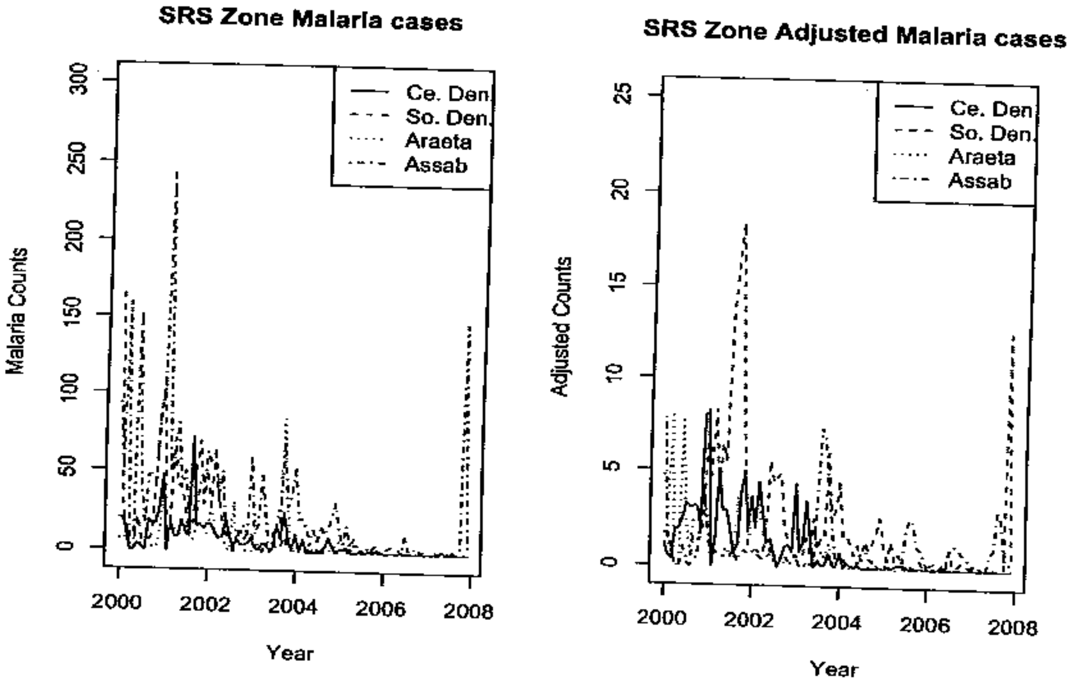


Figure 2.3: SRS Subzones Monthly Malaria Time Series Plot

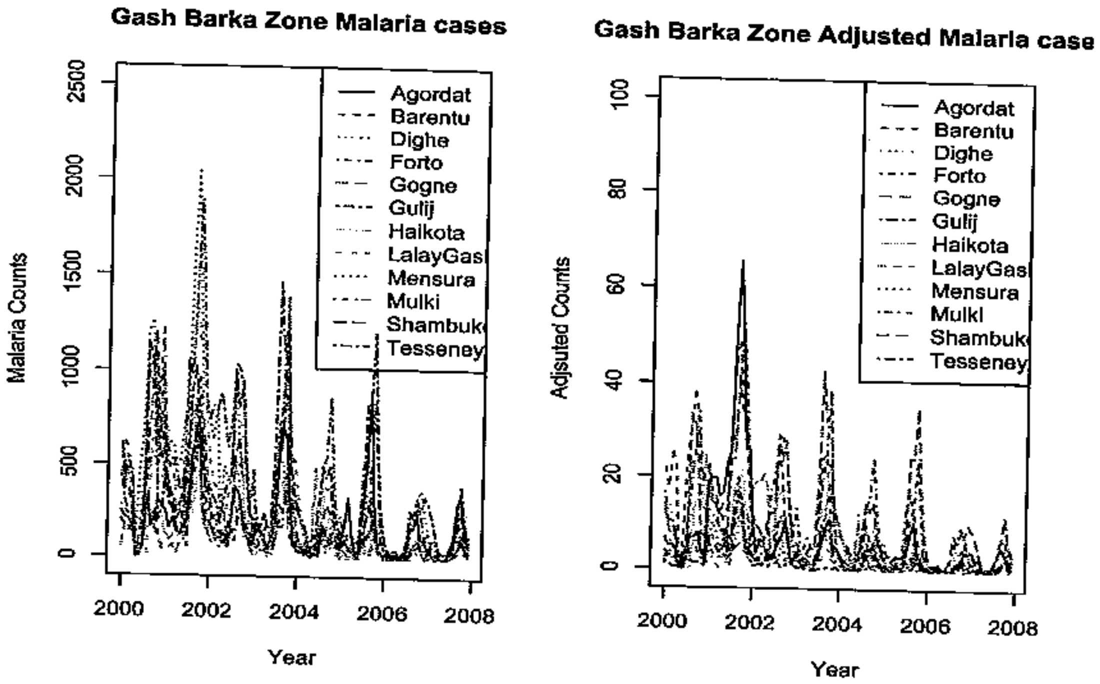


Figure 2.4: Gash Barka Subzones Monthly Malaria Time Series plot

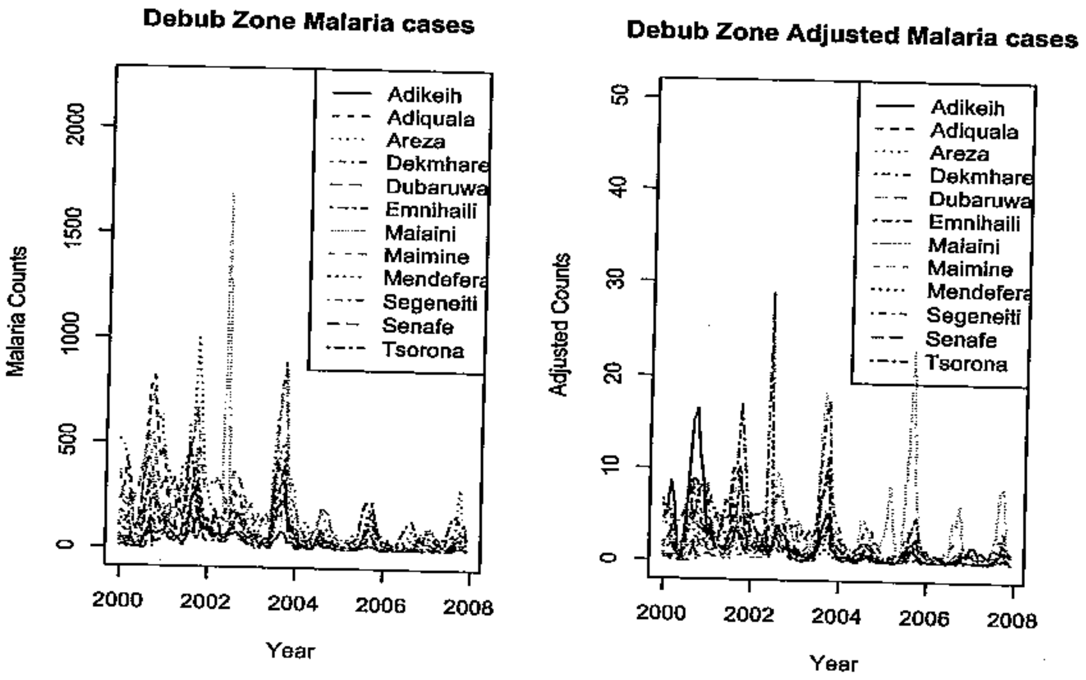


Figure 2.5: Debub Subzones Monthly Malaria Time Series plot

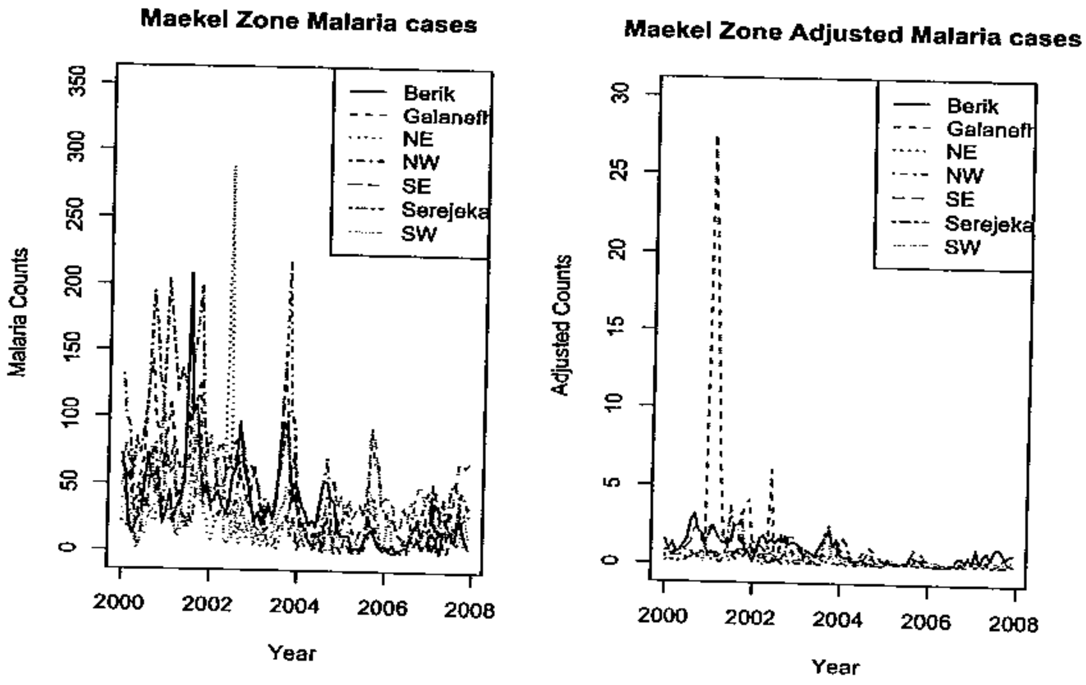


Figure 2.6: Maekel Subzones Monthly Malaria Time Series plot

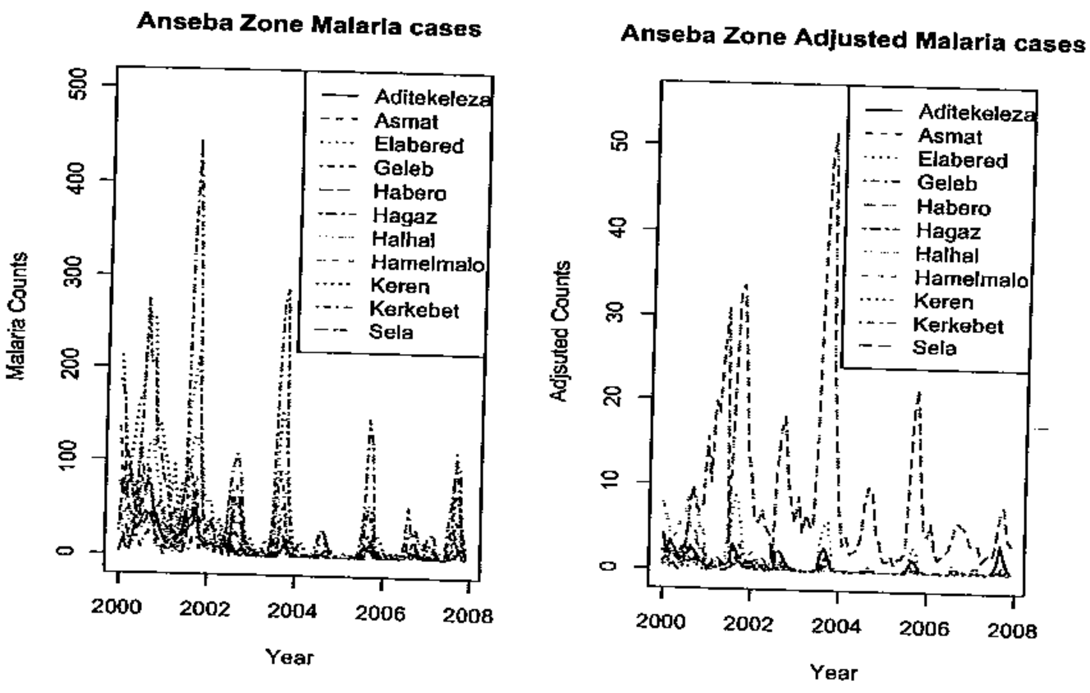


Figure 2.7: Anseba Subzones Monthly Malaria Time Series plot

Rainfall in Eritrea varies both spatially and temporally as explained previously. The rainfall data consists of the monthly average in mm for each subzone for the period January 2000- December 2007. The rainfall data is provided along with the malaria counts by the Ministry of Health Control Program.

In this Chapter, the distribution of malaria cases and rainfall will be described. Fig. 2.8 shows the time series of the monthly average rainfall in millimeter over the period of 8 years for all the subzones and is plotted by subzone in zones in Figures 2.9, 2.10, and 2.11. The plot does not show a clear trend in the distribution of rainfall but there were some "wet" subzone months in the last year of study. The plots show that there is a difference in the amount of rainfall received by the zones.

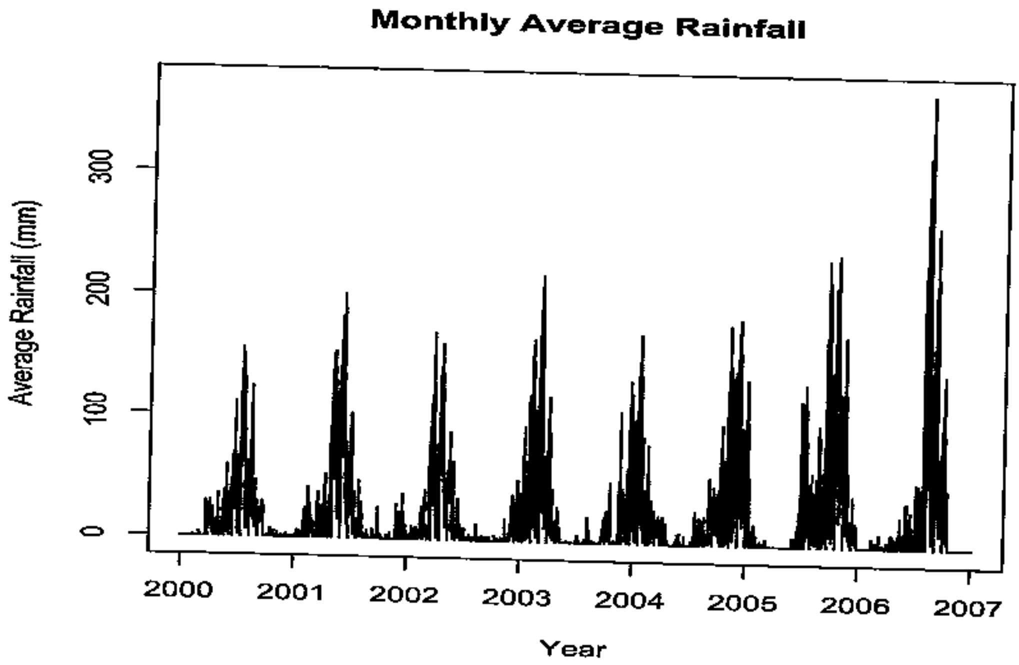


Figure 2.8: Time Series of Rainfall

Fig. 2.12 shows the monthly variation in the number of malaria cases and the monthly variation of monthly average rainfall. Fig. 2.12 shows clearly that the highest number of malaria incidence occur in the month of October and the highest season for malaria is between August-November. This is in line with the fact that the peak of malaria transmission occurs at the end of rainy seasons. The monthly average rainfall plot on the right side of Fig. 2.12 shows that the rainy season is June-September and the effect of rainfall starts to be observed after these months. As was explained in the introduction, these are seasons of heavy rain in the south-west of the country and western escarpments.

The left panel of Fig. 2.13 depicts the general decline of malaria cases over the study period of 8 years.

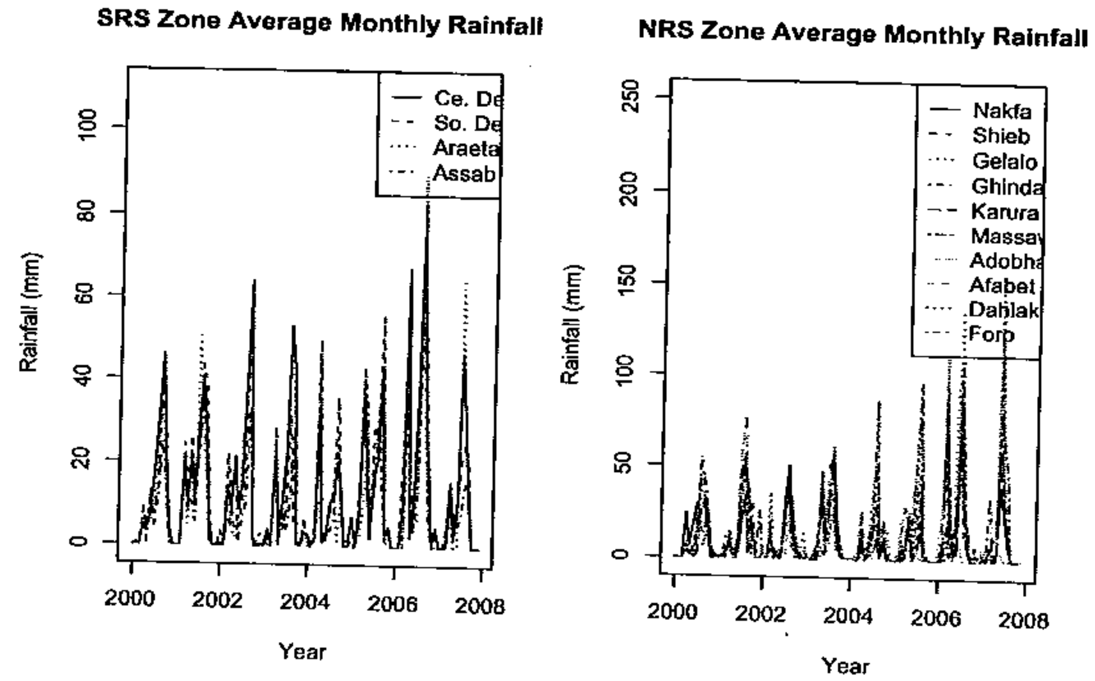


Figure 2.9: SRS and NRS Zones Time Series of Rainfall

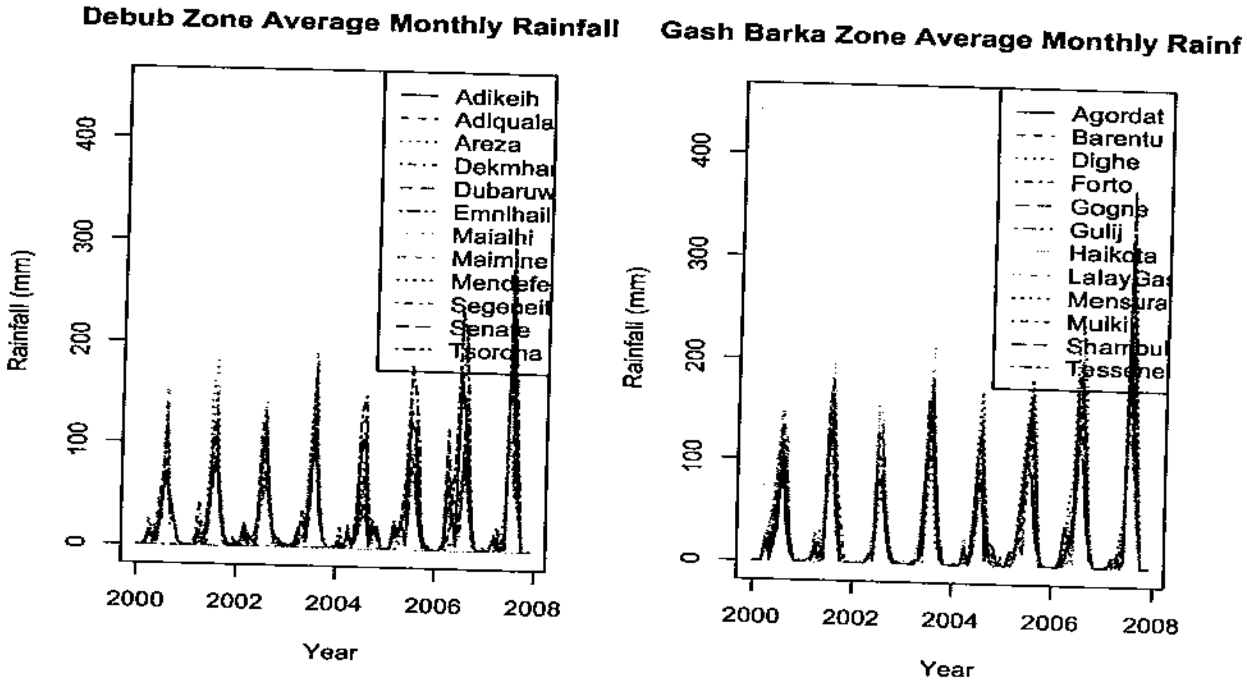


Figure 2.10: Debub and Gash Barka and Zones Time Series of Rainfall

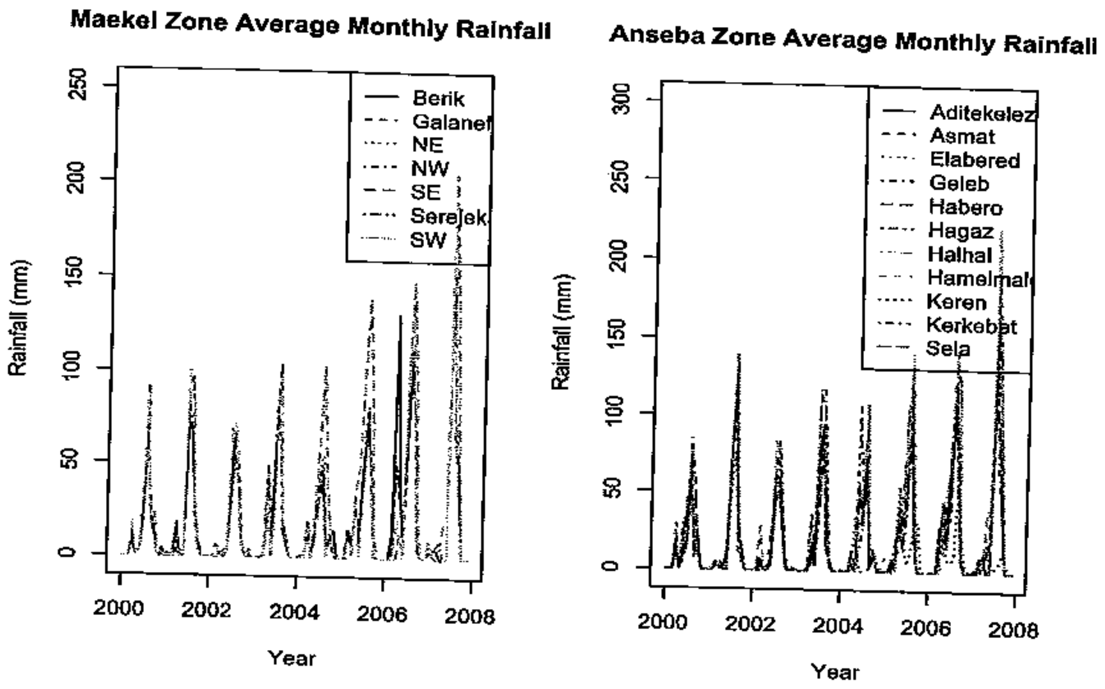
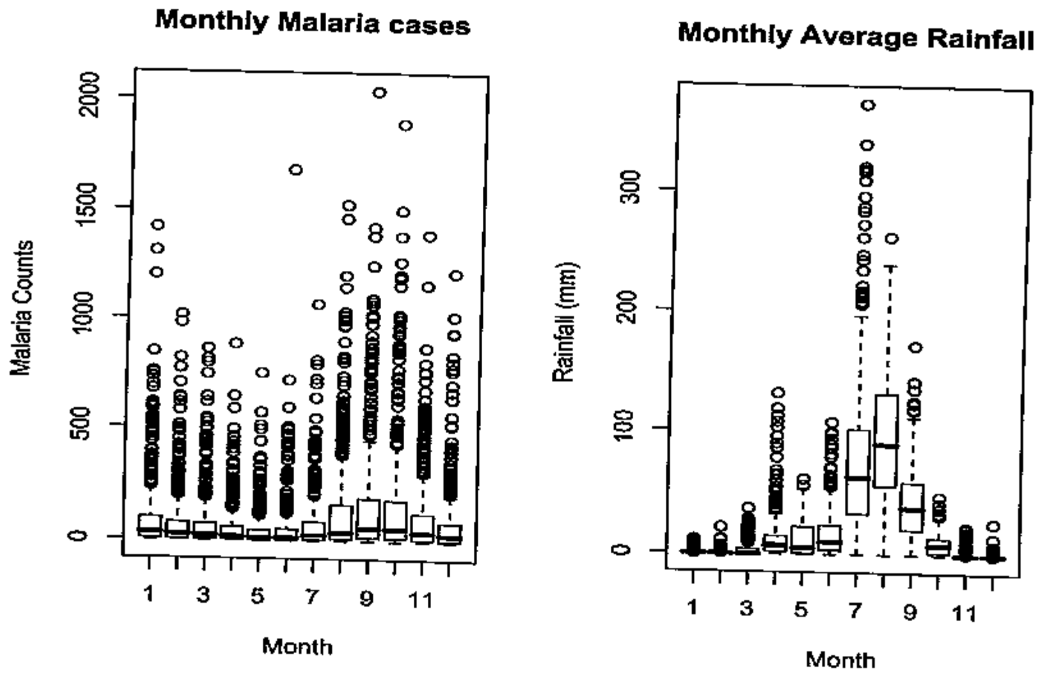


Figure 2.11: Maekel and Anseba Zones Time Series of Rainfall

Figure 2.12: Distribution of Monthly Malaria Cases and Average Rainfall



In addition, the right panel of the plot is the distribution of yearly average rainfall in millimeters and, as can be seen from the plot, rainfall varies with 2005 and 2006 being the years with the highest amount of rain. Fig. 2.13 shows a general decline in the malaria incidence and 2001 was the year with the highest number of incidence. The median rainfall was highest in 2006. There is no clear pattern in the distribution of rainfall. However, comparing the distribution of rainfall and malaria incidence in Fig. 2.13, there is an indication that there are some other driving forces that affect malaria incidence other than rainfall.

Figure 2.13: Distribution of Yearly Malaria Cases and Average Rainfall

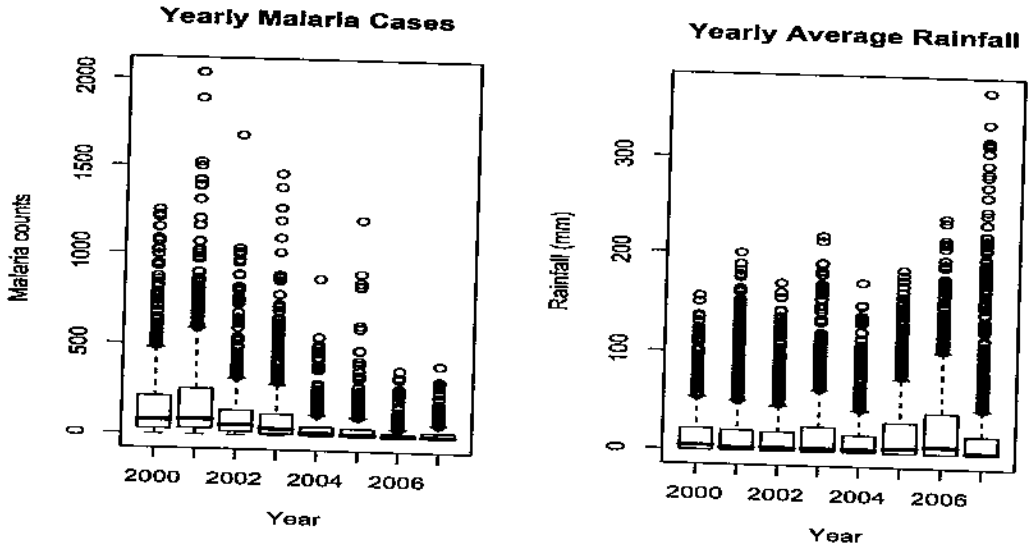


Figure 2.14: Distribution of Malaria Cases

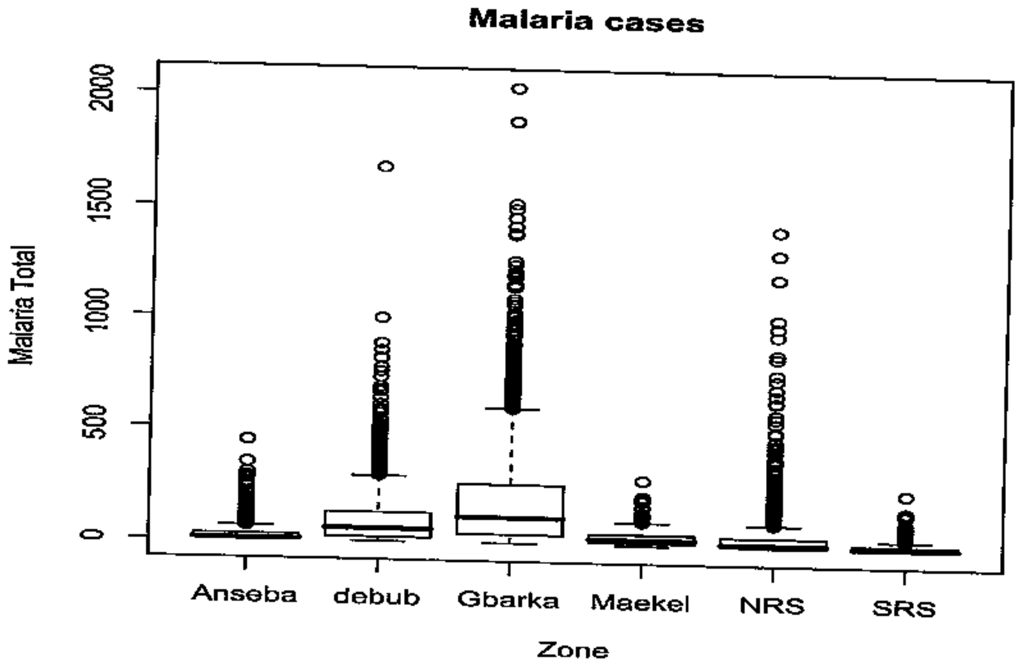


Figure 2.15: Distribution of Adjusted Cases

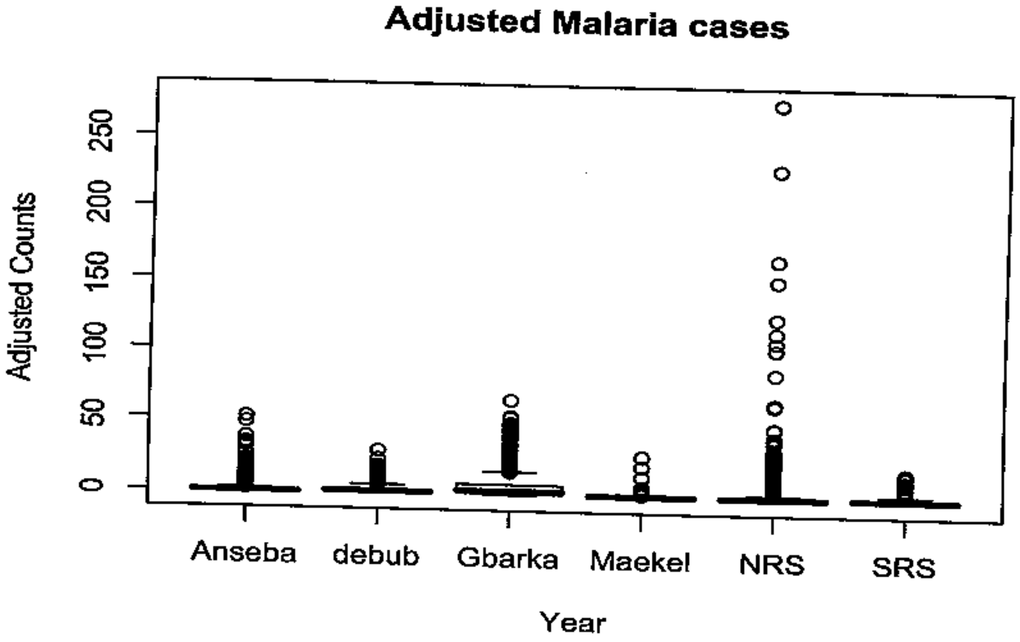


Fig. 2.14 is the distribution of the variation of the raw monthly counts across the zones. Gash Barka, Debub and Anseba are the zones with highest monthly counts. Gash Barka has the highest prevalence rate, followed by Debub and then Anseba. These are areas where the most active agricultural work in the country is located.

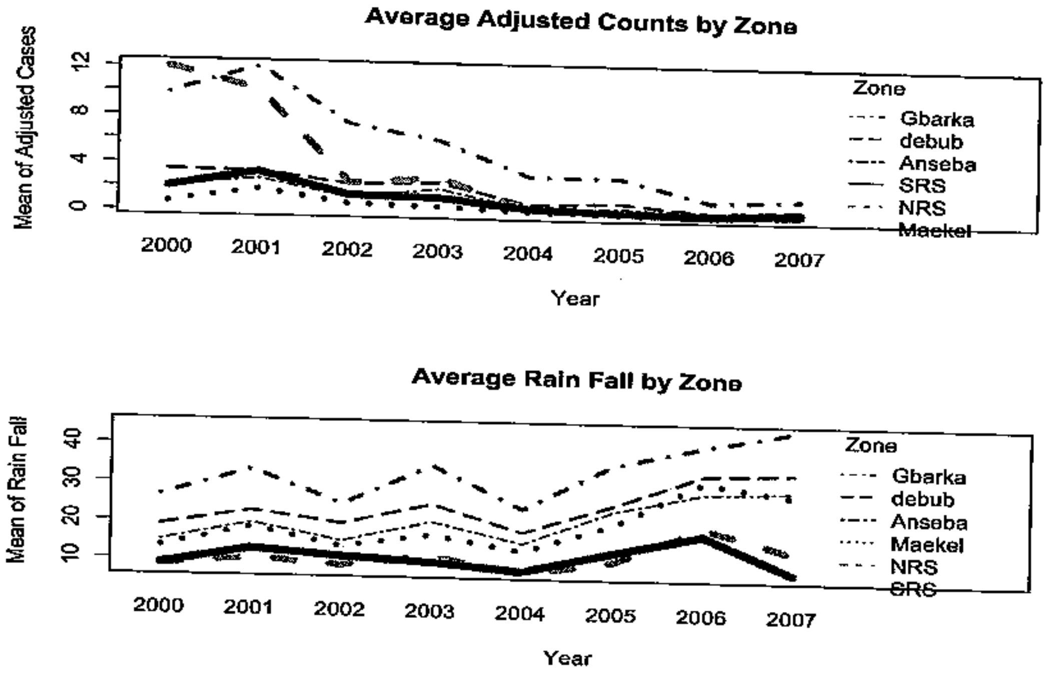
However, in order to compare the distribution of malaria at the zone or subzone level, it is better to compare the adjusted counts which have been adjusted based on population size, with more details on the adjustment below. Fig 2.15 and Table 2.1 show the order of prevalence rate has now changed after the adjustment is made. As a result, Gash Barka, NRS (Northern Red Sea), and Debub are the high prevalence rate zones. Note also that NRS has many outliers. The variation from zone to zone will be incorporated in the model in chapter 4.

Fig. 2.16 contains the average adjusted counts by zone and average rainfall in millimeters by zone. The mean adjusted yearly rate from 2000-2003 was highest in Northern Red Sea and Gash Barka. After 2004, the mean adjusted rates seems to be similar in the remaining 5 zones. The average rainfall by zone is plotted in Fig. 2.16 and shows that Gash Barka, Debub, Anseba and Maekel are the zones receiving the highest amount of rainfall. Southern and Northern Red Sea had the lowest rainfall. It is also worth noting that now in Fig. 2.16 the rainfall shows an increasing trend where as malaria cases are decreasing over the period of study. However, our model does show that there is a positive relationship between rainfall and malaria rates.

	SRS	Anseba	Maekel	NRS	Debub	GBarka
Mean before adjusting for Popsize	13.2	23.1	35	55.7	102.4	211
Mean after adjusting for Popsize	1.2	1.3	0.6	3.7	1.9	5.6
Median before adjusting for Popsize	3.0	7.0	26.0	11.0	57.0	114.5
Median before adjusting for Popsize	0.4	0.2	0.3	0.2	1.0	2.6

Table 2.1: Measure of center for adjusted and unadjusted cases

Figure 2.16: Average adjusted counts and Rainfall by zone



Figures 2.17 and 2.18 show the distribution of monthly rainfall and monthly counts of malaria at the subzone level over the study period of 8 years. The lines in each panel represent a particular year. The zones with the highest number of malaria incidences and heavy rainfall are once again shown at the subzone level. For example, Teseney, Haikota, Agordot Guluj, Shambuko etc. all recieve heavy rainfall in the summer and all belong to the Gash Barka zone. The same is true of the subzones in the Debub Zone.

Fig. 2.18 also shows the same trend that we observed at the zone level. The subzones with the highest malaria counts belong to Gash Barka and those with the lowest number of cases belong to the coastal zones, in particular SRS (Southern Red Sea). Gulij is the subzone with the highest malaria incidence and subzones from SRS, in particular Dahlak have the lowest malaria counts. However, since we are interested in modeling at the subzone level, the adjusted data should be used. The subzone with the highest prevalence rate of malaria is Dahlak, with a population under 4000. It seems that, due to its low population size or other related factors, this subzone get less attention from the health officials (Malaria Control Program).

The trend of the other subzones can not be seen clearly in Fig. 2.19 as the scale in the plot is dictated by that one subzone. Fig. 2.20 is a similar plot to Fig. 2.19 after removing Dahlak to enable us to see the seasonal/trend in other subzones. Hence, now the trends for the remaining subzones can be seen more clearly. The highest rates occur in the subzones that belong to Gash Barka.

Figure 2.17: Rain Fall for each Subzone Grouped Yearly

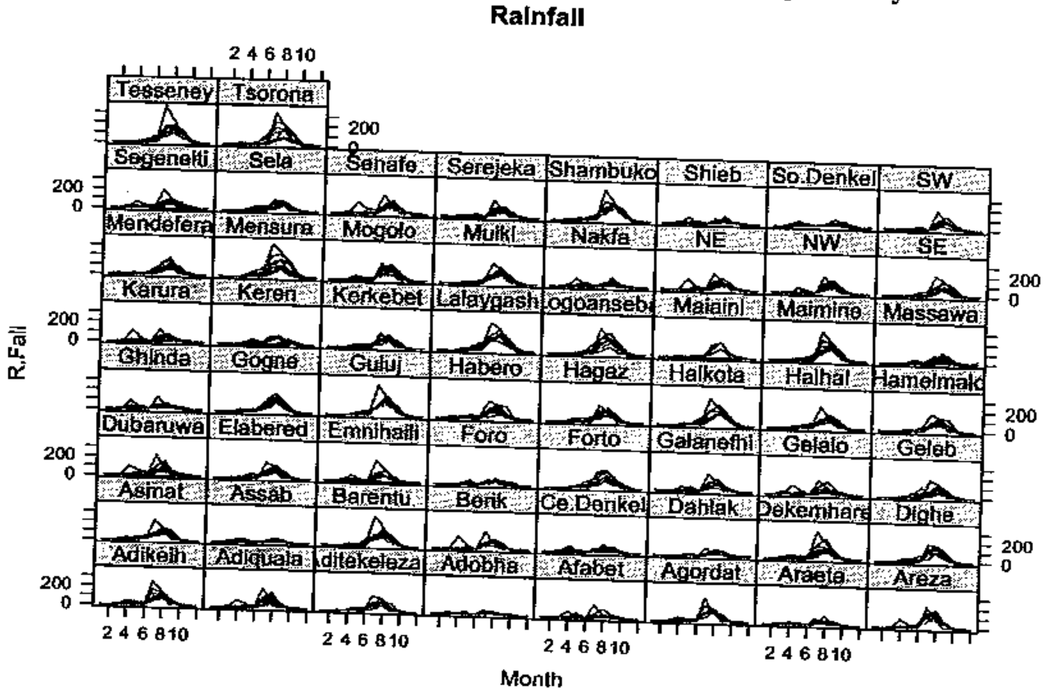


Figure 2.18: Plots of Malaria Counts for each Subzone Grouped Yearly

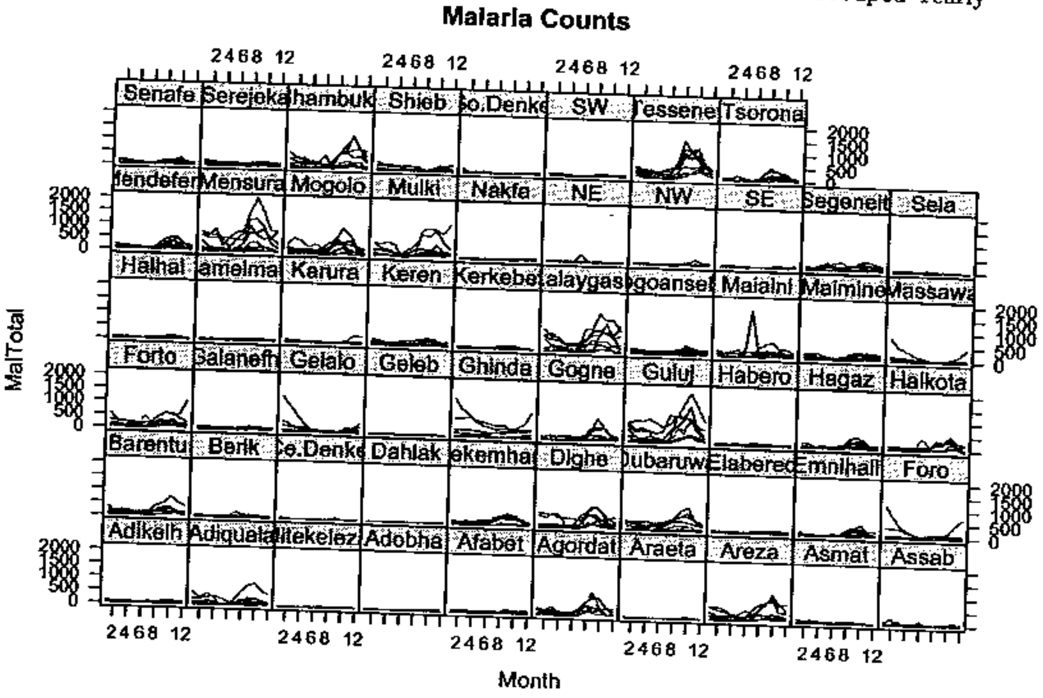


Figure 2.19: Plot of Adjusted Counts for each Subzone

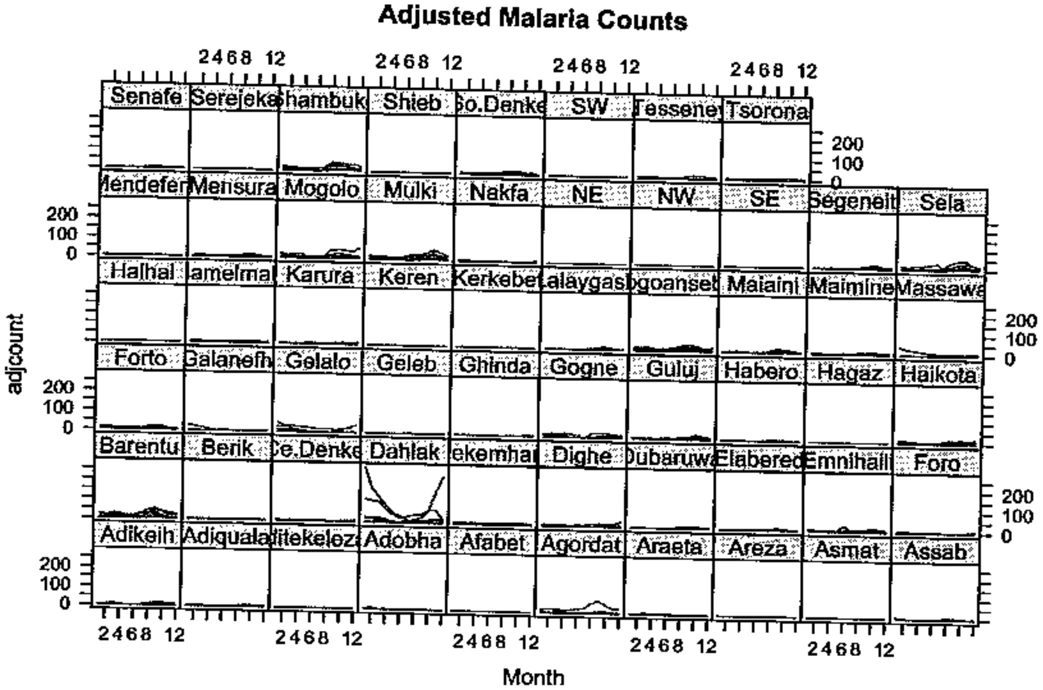
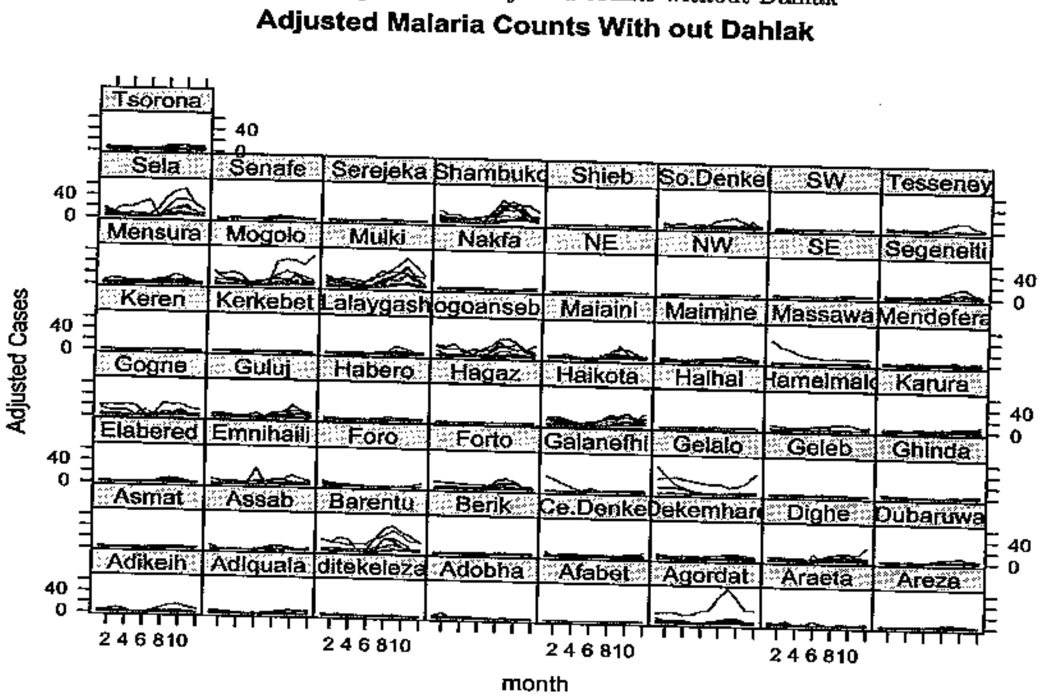


Figure 2.20: Adjusted counts without Dahlak



Chapter 3

Methods:

3.1 Generalized Linear Models (GLM)

Typical linear regression assumes that the responses, Y_i, follow a normal distribution, $Y_i \sim N(\mu_i, \sigma)$ with $\mu_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_q X_{iq}$. Generalized Linear Models (GLM) are an extension of linear regression that allows for non-normal distributions for the response variable. There are many choices for the link between the mean of the response variable and the explanatory variables. Depending on the nature of the data, one can use distributions such as Binomial, Negative Binomial and Poisson. In this study, since the responses are counts of malaria cases, GLM with a Poisson distribution is used. A Poisson GLM has 3 parts.

1. The response variable(Counts) $Y_i \sim Poisson(\mu_i)$, $E(Y_i) = \mu_i = Var(Y_i)$
2. The systematic part is given by the predictor function $\eta(X_{i1}, ..., X_{iq}) = \beta_0 + \beta_1 X_{i1} + ... + \beta_q X_{iq}$, and
3. The logarithmic link, $\log(\mu_i) = \eta(X_{i1}, ..., X_{iq})$ or $\mu_i = e^{\eta(X_{i1},....,X_{iq})}$.

This link function ensures the fitted values are non-negative. Good references on GLM include [7] and [12].

3.2 Poisson Rate Modeling

The response variable Y_{ijk}, is the number of malaria cases for each subzone i, each year j, and month k. Where i=1,...58, j=1,...8 and k=1,....12. Note that in this study there are 8 years (2000-2007), 6 zones, and 58 subzones. The number of monthly malaria incidences are obtained from each subzone and the subzones have different population sizes. Hence, an adjustment using the population size as an offset is needed to model the malaria rate in Counts/Person. The typical offset fixes the coefficient at 1 (which assumes proportionality between count and population size) and has this form: $\log(\mu/t) = \beta_0 + \beta_1 X$ where t is the population size. When using an offset, the proportionality can be checked by estimating a coefficient for the offset. To estimate the GLM with a typical offset, we use $\log(\mu) = \log(t) + \beta_0 + \beta_1 X$. Using $\log(\mu) = \alpha \log(t) + \beta_0 + \beta_1 X$, the coefficient for the offset dictates the adjustment that is made to the counts using $\log(\mu/t^{\alpha}) = \beta_0 + \beta_1 X$. For example, if the proportionality assumption is met, that is the offset coefficient α, estimate is 1, then dividing the counts by the population size is the ideal adjustment that needs to be made. Unfortunately, in this study, the offset coefficient, $\hat{\alpha} \approx 0.2$ in the top model. Thus, $Counts \propto t^{1/5}$ is used as an adjustment. This suggests that the malaria counts increase as a function of fifth root of population size which is a more complicated relationship than proportionality. Proportionality suggests that with more people, you see more malaria cases but our results suggest that counts increase as a nonlinear function of population size.

A key assumption in GLM is the statistical independence of the observations. However, in this study the malaria counts are related or dependent from one month to the next. There is variability from zone to zone, a spatial type of dependence, and variability from subzone to subzone. There is also variability from subzone to subzone within a zone. These variabilities need to be taken into account when modeling. A GLM with Poisson family was fit using a systematic component of $\log(\mu) = \alpha \log(t) + \beta_0 + \beta_1 R.Fall_k + \beta_2 Yr(01) + ... + \beta_8 Yr(07) + \beta_9 sin(2 * \pi * (Month)/12) + \beta_{10} cos(2 * \pi * (Month)/12)$ for the data with a search for the optimal lag of rainfall, k. In Chapter 2, monthly seasonality was observed and there was also a trend. To deal with that a harmonic sine and cosine function was used for within year variation around a yearly trend. When this model was fit, not surprisingly, the overdisperison parameter was large, $\hat{c} = 169.92$. The overdisperision parameter (variance inflation factor) is estimated using $\hat{c} = \frac{\chi^2}{df}$ and under the assumption of identically and independently distributed (iid), $\hat{c}$=1. χ^2 is the usual Pearson goodness-of-fit test statistic and df is the degrees of freedom for the test. In real biological data, $\hat{c}$ is often in the range between 1 and 3 [2]. If $\hat{c}$ is large, that means there are structural issues with the model. Due to the structural issues of the model in this study, a model that can account for some of the previously discussed random effects is discussed below.

3.3 Generalized Linear Mixed Model Fitting

Generalized Linear Mixed Models (GLMM) are methods used when the data are hierarchically structured [12]. GLMM accommodates the lack of independence of the observations and differences from zone to zone and subzone to subzone using the random effects of the model. That is, unlike the GLM, GLMM relaxes the assumption of independence through the introduction of random effects. This is because there is variation between the subzones and the subzones within zones. Thus, in this study, two different random intercept structures are considered. This is in order to accomodate the month to month dependency of the observations in the subzones and existing variation between zones and variation between the subzones even in the same zones. Subzones are to be nested within zones and a random intercept will be estimated for each zone and for each subzone within the zones.

3.4 Programs Used

Generalized Linear Mixed Models were fit using 4 different programs. glmer, lmer, glmmPQL and glmmML. [12].

1. glmer is a function in R from the package lme4 [3]
2. lmer is also a function in R from the package lme4 [3].
3. glmmPQL is also a function used for mixed models from the R package MASS. Since this function maximizes a penalized quasi-likelihood rather than the full likelihood, it does not provide AIC [11].
4. glmmML is a function in R from the the package glmmML [4]. This estimates the model parameters using maximum likelihood and provides AIC.

Note that lmer and glmer are functions that are nearly interchangeable. The reason for that is if lmer with the family different from the default (normal) distribution is used then the call for lmer is replaced by a call to glmer [3]. In this study, since the poisson family is used in both lmer and glmer both fits give the same result and hence they are interchangeable.

3.5 Model Selection

Akaike (1973) used Kullback-Leibler (K-L) Information as the basis for model selection. K-L information is the distance between the approximating model and the reality or measures the information lost when a model is used to approximate the reality. Akaike found the relationship between the K-L information and the maximum log-likelihood. He showed that the expected K-L information can be estimated and it is related to log-likelihood function at its maximum [2]. However, the maximum log-likelihood is biased and he found that the bias correction factor is approximately equal to k, where k is the number of parameters in the model. That is, K-L information and E(K-L information)$=L-k$ and multiplying this by -2 yields what is defined as Akaike's An Information Criteria (AIC), $AIC=-2L+2k$

where L is the maximum log-likelihood of the model [1]. He multiplied by -2 because the ratio of 2 maximized likelihood values is asymptotically distributed as a chi-square under certain assumptions and conditions. The model with the smallest AIC is selected because the selected model minimizes the information lost in approximating the reality by a fitted model. Thus, a model with the lowest AIC is selected as the top model. Ranking the AIC and taking the difference, ∇_i defined as $\nabla_i = AIC_i - AIC_{min}$ is used. These values are the distance between the best selected model and the i^{th} model. The best model is the one with $\nabla_i = 0$.

The Bayesian Information Criteria (BIC) is based on Bayesian approach and attempts to find the model with the highest posterior probability of being correct model. It is defined as $BIC = -2L + Klog(n)$, where n is the number of observations. AIC and BIC differ in their penalty term. As described in [2], if $n > 8$ the penality in BIC is a little larger and as a result selects smaller dimensioned models than AIC. In Table 4.1, we have reported the AIC, BIC and ranking of the difference of the estimated values.

Chapter 4

Results and Discussion

4.1 Lagging Rainfall

Biological considerations indicate that a lag exists between rainfall and associated increases in malaria cases. With the availability of temperature as in [10], lags of 4-12 weeks are considered due to biological considerations. The number of lags could be shorter for high temperature areas [10]. The assumption used in most studies is that changes in temperature and rainfall at a particular time do have an important influence on the reproduction rate of mosquitoes. Temperature data is not available in this study and thus lags between 1 and 4 months were considered for rainfall. Note also that in this study that the data set is constrained to use the same number of responses regardless of lag of rainfall considered to allow direct comparisons of models. This means that 232 observations are dropped from the analysis, the 4 months from January-April, 2000.

4.2 Results

In this study a rainfall lag of 1 to 4 months was used to fit the GLM model described in section 3.1. Since there was variability from zone to zone and subzone within zone, the overdispersion parameter for the optimal model from the GLM fit was large i.e $\hat{c} = 169.92$. As a result a GLMM that can accommodate this type of dependence was used.

In the GLMM, once again a rainfall lag of 1 to 4 months was used in order to determine the optimal lag for the top model. As was explained in Chapter 2, there was a nonlinear trend across years and as a result natural splines based on year with 2 to 6 degrees of freedom were considered. In addition, no trend and a unique value for each year or 7 degree of freedom spline were also considered. Additionally, two forms of random effects were considered. First, a random intercept for the subzones was used. Second, a random intercept for the zones and a random intercept for the subzones nested within the zones was used. Note also that harmonic functions (sine and cosine) are used in the model to accomodate monthly seasonality.

The GLMM model results using year as a linear quantitative (natural spline with 1 degree of freedom) or the unique (7 degrees of freedom spline is the same as using year as a factor) for the rainfall lag of 1 to 4 months and two forms of random intercept model structures are presented in Table 4.1.

Model	Year DF	Rainfall lag	AIC	BIC	∇AIC	∇BIC
Nested	7	3	220380	220420	0	0
Not Nested	"	3	220420	220505	40	33
Nested	"	2	223459	223552	3079	3080
Not Nested	"	2	223500	223585	3120	3113
Nested	"	4	227853	227945	7473	7473
Not Nested	"	4	229894	229979	7514	7507
Nested	"	1	229831	229923	9451	9451
Not Nested	"	1	229873	229958	9493	9486
Nested	1	3	245631	245684	25251	25212
Not Nested	"	3	245672	245718	25292	25246
Nested	"	2	249708	249760	29328	29288
Not Nested	"	2	249749	249795	29369	29323
Nested	"	4	255245	255298	34865	34826
Not Nested	"	4	255288	255334	34908	34862
Nested	"	1	258536	258589	38156	38117
Not Nested	"	1	258580	258626	38200	38154

Table 4.1: GLMM Fit

Table 4.1 shows the AIC and BIC for the 1 and 7 degrees of freedom for the trend. However, all other possible combinations of these variables as well as different degrees of freedom 2-6 for the year trend were also fit but not presented here. Regardless of which rainfall lag you consider, the nested random effect structure is the preferred one, as shown in Table 4.1. That is, nesting subzone within zone resulted in the lowest AIC. It is also worth mentioning that the population size as an offset with coefficient 1 was also fit, however the AIC was much worse. The GLMM model nesting the subzone within zone provided the top AIC model. This is an indication that the introduction of the random effects in the GLMM may have have tackled some of the problem of dependence that was originally present in the conventional GLM. Table 4.1 indicates that regardless of which random effect structure used, the optimal rainfall lag is found to be three months. The finding of this study is that, with the absence of temperature, the optimal rainfall lag to consider is three months. This is a little bit different result from previous studies done in this area. The optimal lag of rainfall considered in previous studies was two months [8]. However, the finding of this study is in-line with the descriptive analysis of the data in chapter 2. That is, the effect of rainfall on the malaria incidence starts to be felt most strongly three months later.

	Methods			
	glmer		glmmPQL	
Parameter	Estimates	St. Error	Estimates	St. errors
Intercept (Y(00))	2.01200	0.68270	-2.38330	1.54470
Log(Popsize)	0.19420	0.05440	0.61170	0.14206
R.Fall	0.00380	0.00004	0.00381	0.00028
Y(01)	0.21730	0.00450	0.21640	0.00333
Y(02)	-0.26020	0.00499	-0.26043	0.00369
Y(03)	-0.40010	0.00515	-0.40040	0.03805
Y(04)	-1.18400	0.00660	-1.18400	0.04880
Y(05)	-1.24800	0.00671	-1.24800	0.49600
Y(06)	-2.1100	0.00922	-2.11100	0.06813
Y(07)	-1.82000	0.00820	-1.82010	0.06055
$\sin(2 * \pi * (Month)/12)$	0.28690	0.00223	0.28690	0.01647
$\cos(2 * \pi * (Month)/12)$	-0.33490	0.00304	-0.33420	0.02250

Table 4.2: Comparison of Methods For the Top Model

In addition, Table 4.1 shows that three months lag and nested model with 7 degrees of freedom (unique mean for each year) is found to be the best. The models found using a natural spline with 7 degrees of freedom for the year and treating year as a categorical are exactly the same.

Table 4.2 contains the estimated parameters for the top model and optimal rainfall lag using the glmer and glmmPQL fit. The lmer fit is not presented here as both the glmer and lmer yield the same result as explained in Chapter 3. As can be seen in Table 4.2, both glmer and glmmPQL gave the same results except for the glmer/lmer result on the estimated intercept and offset coefficient. In this study, we used AIC selection criterion; glmmPQL does not provide AIC since it does not use a likelihood, instead relying on quasi-likelihood methods. glmmML was also considered in this study. However, glmmML was not available for the top model as it did not allow nesting subzone within zone.

The estimated coefficient for the offset is 0.1942 ≈ 0.2 and it is this value that determined the adjustment that was made when comparing the malaria cases between the different subzones. Therefore, the adjustment used in Chapter 2 is simply the counts divided by the fifth root of the population size.

The performance of the top model also needs to be checked and different components assessed. For that, the observed and estimated cases were plotted in Fig. 4.1. Though difficult to observe in this plot, the model captures much of the variation over time and between subzones.

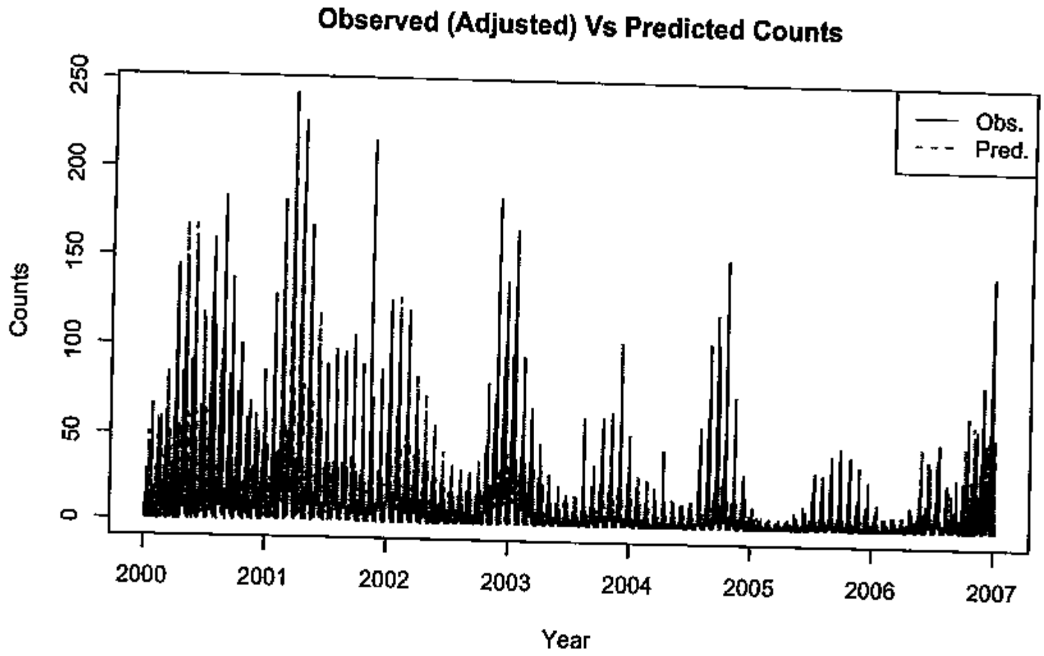


Figure 4.1: Observed vs Fitted from the best model

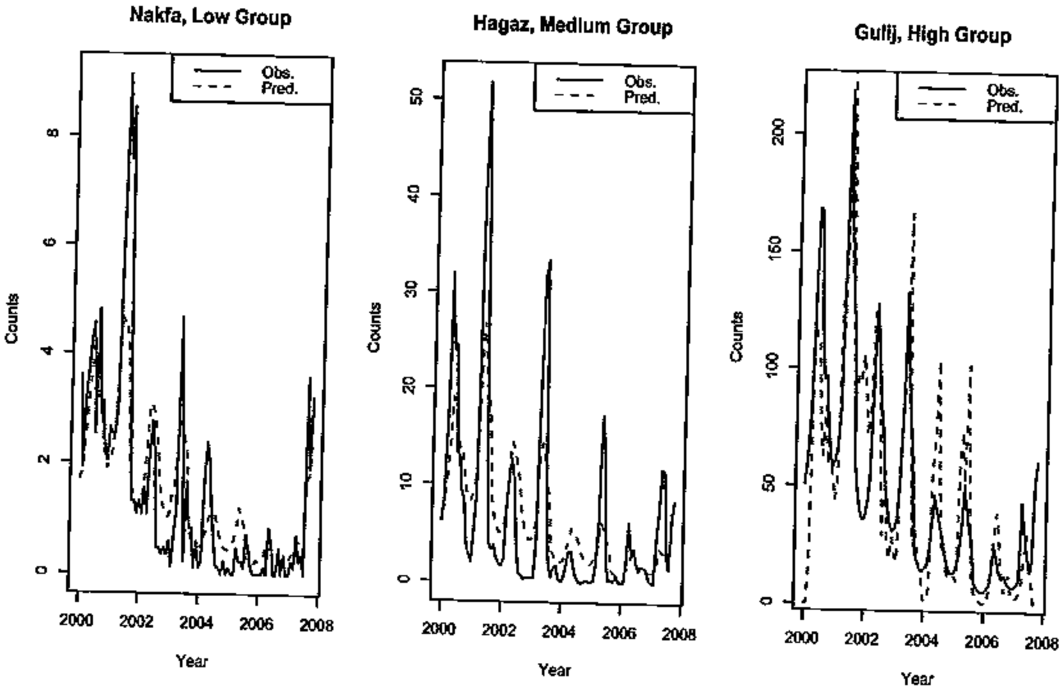


Figure 4.2: Goodness of Fit for the Low, Medium and High Groups

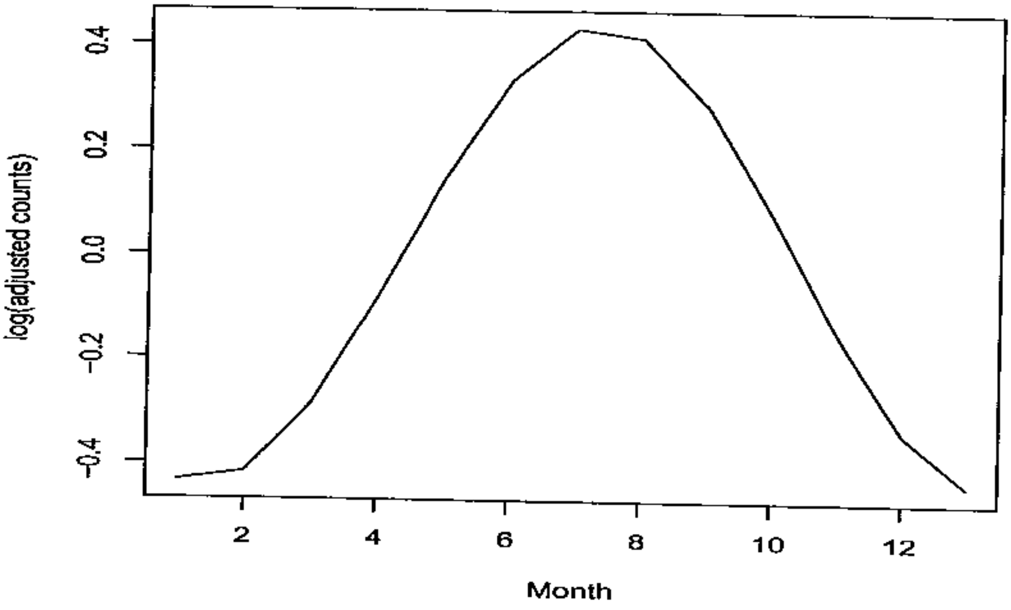


Figure 4.3: Seasonal effect

But since the aim is to study the malaria cases at the subzone level, the performance of the model has to be assessed at the subzone level. First, subzones are divided into 3 different groups, low, medium and high-incidence groups based on the mean incidence in each subzone. Then, one subzone is picked from each group and the observed versus fitted are plotted in Fig. 4.2 for subzones Nakfa(low), Hagaz(medium) and Gulij(High). The plot shows the top model has done a decent job of capturing seasonality and trend in these three subzones.

The seasonal effect of the harmonic functions, after accounting for the effect of rainfall, is shown in Fig. 4.3 and the peak month is found to be July. The peak of malaria rates in most subzones was October as described in Chapter 2, so the harmonic effect is picking up on a different aspect of seasonality in malaria rates that is not accounted for by rainfall (lagged 3 months).

4.3 Conclusion

Count data are modeled using a GLM using a poisson family with a log-link. However, this model requires the assumption of statistical independence of observations. In this study, both the ordinary GLM and GLMM using a poisson distribution were fit and due to the dependence from month to month and variability from zone to zone and subzone to subzone, the GLM model fit resulted in a large overdispersion parameter. The GLMM model accommodates the existing dependence and a model with random intercept for each zone and subzone in zone for the 3 month rainfall lag is found to be the best model. However, in most previous studies the optimal lag was found to be 2. This could be due to the effect of temperature on the life cycle of the reproduction of mosquitoes. Accounting for maximum temperature could shorten the number of rainfall lags to consider. Adjustment on the malaria count in each subzone was performed by dividing the counts by the fifth root of the population size. The estimated population size for the year 2000 for each subzone was obtained from National Health Management Information System (NHMIS), Ministry of Health in the work done by the Environment Health Project [5]. The 2000 population size used in this study might not be necessarily accurate and does not reflect the population growth over the period of study. However, that was the available data and we had to make use of that. 4 programs were also used to fit the model and this study found out that, especially for nested random effects, lmer/glmer needed to be used in R. This is because even though glmmPQL could also equally fit the GLMM, the estimation used is quasi-likelihood based and model selection based on AIC is not possible, but for glmmML, since it is based on maximum likelihood, AIC based model selection is possible.

As a conclusion, I would say, today's rainfall most usefully explains the malaria incidence to come after 3 months. This suggests a control strategy could target certain subzones based on high rainfall. After controlling for rainfall, malaria rates declined over the study period. Even though 2007 was the wettest year, malaria incidence was at its lowest level and this clearly shows the success of efforts made by the Malaria Control Program. This achievement might be due to the the increase in health facilities, training given by the Malaria Control program to the village health agents, coordinated endeavors made to control the infection, the distribution of sufficient and effective anti-malaria medicine, the provision of malaria nets to the population free of charge, as well as the active popular participation in maintaining environmental sanitation. The next stage in this research is to incorporate some other additional covariates such as temperature, bed nets used, amount of spray used and others into this model and study their effect on the malaria rates. Further analysis could consider differences between coastal and non-coastal zones.

The data was actually available as inpatient and outpatient cases of age under 5 and above 5. Modeling the malaria cases for under 5 and above 5 could also be considered. Moreover, a further look at the distribution of inpatient and outpatient malaria cases can be also done.

Acknowledgment

I would like to thank the the Head of Eritrean Malaria Control Program, Ministry of Health, Dr. Tewolde Ghebremeskel and Selam Mihreteab for providing the data used in this data. I would also like to thank Dr. Mark Greenwood for allowing me to work on this topic and for his helpful input and advice on this project.

Bibliography

[1] Akaike H (1973). Information theory as an extension of maximum likelihood principle. pp. 267-281 in BN Petrov, Csaki F.editors. Second International Symposium on the information theory. Budapest, Hungary.

[2] Anderson, D.R (2008). Model based inference in the life sciences. A Primer on Evidence. Springer.

[3] Bates D and Sarkar D (2006). lme4. R package.

[4] Brostorm G. (2008). glmmML: Generalized linear models with clustering. R package version 0.81-2.

[5] Graves, P.M, (2004), Eritrea: Malaria Surveillance, Epidemic preparedness, and Control Program Strengthening, Environmental Health Project.

[6] Josephat, S., Tewolde, G., Solomon, M., Helen, F., Mehari, Z., Charles, M., John, G., Weidong, G., Robert, N., and John, C.B, (2003). Distribution of Anopheline Mosquitoes in Eritrea. *Am. J. Trop. Med. Hyg.*, 69(3), 2003, pp. 295–302.

[7] McCullagh P., and Nelder J. (1989). *Generalized Linear Models*, Second edition, Champman and Hall, London, UK.

[8] Oliver J., Penelope, V., Dissanayake, M., Gawrie N.L, and Priyanie, H.A., (2008). Models for short term malaria prediction in Sri Lanka. *Malaria Journal* 2008, 7:76.

[9] Peter, M.N, Tewolde, G., Goitom, M., Jacob, M., Usman, A., Andom, O., Andrew, K., Andemariam, G. Disanayike, G., Yohannes, G., and Yohannes, O. (2006). A steep decline of malaria morbidity and mortality trends in Eritrea between 2000 and 2004: the effect of combination of control methods. *Malaria Journal* 2006. 5:33. 8.

[10] Teklehaimanot H.D, Lipsitch,M., Tekelehaimanot, A. and Schwartz, J., (2004). Weather-based prediction of Plasmodium falciparum malaria in epidemic-prone regions of Ethiopia I. patterns of lagged weather effect reflects biological mechanisms. *Malaria Journal* 2004. 3:41. 6.

[11] Venables, W. N. & Ripley, B. D. (2002) *Modern Applied Statistics with S*. Fourth Edition. Springer, New York.

[12] Zuur, A.F, Ieno, E.N., Walker, N.J, Saveliev, A.A., and Smith, G.M. (2009). *Mixed effect models and extensions in ecology with R*, Springer.