FITTING PROBABILITY DISTRIBUTION FUNCTION TO MALARIA INCIDENCE DATA

Drinold Mbete[1], Kennedy Nyongesa[2], Joseph Rotich[3]

1; 2; Department of Mathematics Masinde Muliro University of Science and Technology, Box 190-50100, Kakamega, Kenya.
3, Vice-chancellor, Laikipia University, Box 1100-20300, Nyahururu, Kenya

Abstract : *Malaria remains a major infectious disease that affects millions of people. Once infected with Plasmodium parasites, a host can develop a broad range of clinical presentations, which result from complex interactions between factors derived from the host, the parasite and the environment. Malaria has historically been a very serious health problem and currently poses the most significant threat to the health of Masinde Muliro University of Science and Technology students, data shows that more than 70% percent of pediatric cases are due to malaria.*
Methodology: Hence, the study aimed to fit malaria incidences dataset for the period 1st January, 2013 to 31st December, 2015. Data on monthly malaria incidence was obtained from the Masinde Muliro University of Science and Technology health service. Gamma, Weibull and Lognormal Distribution were employed to fit the malaria incidence dataset using R-software.
Results: High malaria incidences were observed in the months of August, September and November. AIC values results showed that lognormal distribution had the lowest AIC value of 185.9875 followed by the Gamma distribution with a value of 187.8815 and then the Weibull distribution with a value of 188.7271. This confirmed the lognormal distribution to be the best fitting distribution for the malaria incidence dataset
Conclusion: The Poisson regression did not accurately fit the data on malaria incidences due to over dispersion in the data but lognormal was a better fit compared to gamma and Weibull distribution.

Keywords – *Malaria incidence, dataset, fit, lognormal, Weibull, Gamma*

INTRODUCTION

Malaria is parasitic disease caused by a unicellular protozoan of the genus Plasmodium which is transmitted by the anopheles mosquito. There are about 120 species of Plasmodium parasites found in the blood of mammals, reptiles and birds. Four of these species commonly infect humans, i.e. *Plasmodium falciparum*, *Plasmodium vivax*, *Plasmodium ovale* and *Plasmodium malariae*. *Plasmodium falciparum* found throughout the tropics and sub-tropics is largely responsible for most of malaria's global morbidity and mortality. The parasites are transmitted to humans through the bites of infectious female mosquito (vectors). Following an infective bite, symptoms appear in about 9-14 days. According to WHO [1], the latest world malaria report released in November 2017 shows that the number of malaria cases reported in the year 2016 was 216 million up from 211 million cases reported in 2015. The report also shows that malaria death estimates in 2016 stood at 445,000 compared to 446,000 deaths in 2015. The high burden of malaria cases in 2016 was in Africa at 90% with 91% cases of deaths reported. Malaria is among the top ten disease burden in Kenya, it was ranked 3rd cause of premature death in 2010 by Global Burden of Diseases (GBD). Of the total population in Kenya, approximately 70% are at risk. Fourteen million and 17 million are in endemic and epidemic areas respectively. Malaria has historically been a very serious health problem and currently poses the most significant threat to the health of student at Masinde Muliro University of Science and Technology (MMUST) which is in malaria prone area. The malaria situation in MMUST is typical of Sub-Saharan Africa making its transmission an all year round affair and seasonal variation. The MMUST Health facility records show that between 300-700 cases of malaria

are reported each month which constitutes 75% of all out- patient cases. Transmission of malaria in takes place throughout the year but peaks during the long rainy season from April to August (MMUST health facility records). Malaria transmission is driven by several factors including climatic, geographic, and socio-economic variables it is also affected by climate variability at both seasonal and inter-annual scales Yeka et al [2]. According to Gemperli [3], the main effect the environment has on the malaria vector is the influence factors such as temperature and rainfall have on the mosquito's survival and the duration of the parasites life cycle in the vector. Malaria transmission will thus depend on whether the mosquito vector and parasite had the ability to coexist long enough for transmission to occur. The main malaria vectors found in MMUST are Anopheles *gambiae sensu stricto*, *An. Arabiensis* and *An. Funestus*. Anopheles gambiae generally increases in density after the start of the long rains, while *An. funestus* density is seen to vary in direct proportion to the proximity of permanent breeding grounds rather than rainfall. The malaria cases in MMUST can either be complicated malaria or uncomplicated malaria.

The study Area
The research was conducted in Masinde Muliro University of Science and Technology (MMUST) located in Kakamega Town, Kakamega County with an altitude of 1561m above the sea level and a student population of approximately 15000.

Literature Review

Malaria remains one of the leading health problems of the developing world, and Kenya bears a particularly large burden from the disease. The prevalence of malaria infection, incidence of disease, and mortality from severe malaria all remain very high despite efforts from different multi-agency like WHO in fighting the burden of malaria disease. Several studies have been carried out on malaria incidence. Bayesian analysis of an epidemiological model of Plasmodium falciparum malaria infection in Ndiop, Senegal was analyzed by Nicole *et al* [4]. The model describes the application of Bayesian calibration of malaria transmission model using longitudinal data gathered from 176 subjects in Ndiop from 1st July, 1993 to 1st August, 1994. The model was able to adequately predict Plasmodium falciparum parasitaemia prevalence in the study population that is, during the dry season, the estimated fraction of non-immune subjects went down to 20% and increases upto 80%. The model was also able to predict time-weighted average incidence contributed by non-immune and immune individuals as 0.2 and 0.47 case/day respectively. Poisson and Negative binomial regression models was used to study the trend of malaria prevalence in Minna, Nigeria by Patience and Osagie [5]. Their results showed that the prevalence of malaria is still on increase by 6% on monthly basis while Martens *et al* [6] examined the relationship between malaria and environmental and socio-economic variables in Sudan using health production modified model. The used regression analysis method in analysis of their results, the regression results showed significant relationships between malaria, rainfall and water bodies while other variables such as Human Development Index, temperature, population density and percent of cultivated areas were not significant. Teklehaimanot *et al* [7] used robust Poisson regression model to model the daily average number of cases in 10 districts of Ethiopia that was associated with rainfall, minimum temperature and maximum temperature as explanatory variable in a polynomial distributed lag model. To improve reliability and generalizability within similar climatic conditions, the districts were grouped into two climatic zones, hot and cold. The results showed that malaria was associated with rainfall and minimum temperature in Ethiopia. In cold districts, rainfall was associated with a delayed increase in malaria cases while the association in the hot districts occurred at relatively shorter lags. The results also showed that in cold districts, minimum temperature was associated with malaria cases with a delayed effect while in hot districts, the effect of minimum temperature was non-significant at most lags, and much of its contribution was relatively immediate.

Nkurunziza *et al* [8] modeled the effects of climate on malaria in Burundi using generalized linear models and generalized additive mixed models. The results showed that there was a strong positive association between malaria incidence in a given month and minimum temperature of the previous month. In contrast, the results also showed that rainfall and maximum temperature in a given month have possible negative effect on malaria incidence of the same month. Kakchapati and Ardkaew [9] carried out a study to model the spatial and trends of malaria incidence in Nepal. They used Poisson and negative binomial regression models to _t malaria incidence rates as a function of year and location. Their study showed a steady decreasing trend in malaria incidence, but the numbers of malaria cases were still very high. Sriwattanapongse *et al* [10] used Spearman's correlation

between weekly climatic variables (temperatures, relative humidity and rainfall) and malaria to analyze the bivariate relationships between types of malaria para-sites and potential climatic factors. A discrete Poisson model was used to identify purely spatial clusters of malaria incidence in the high risk areas. A Poisson regression model combined with distributed lag non-linear model was also used to examine the effects of temperature, relative humidity and rainfall on the number of malaria cases. The residuals were checked to evaluate the adequacy of the model. Sensitivity analysis was performed to ensure that the associations between climate variables and malaria incidences did not change substantially when the degrees of freedom for climate variables were changed.

Drebel [11] carried out a study using logistic regression to estimate and assess malaria prevalence and the use of malaria risk reduction measures and their association with selected background characteristics in South Sudan. The results suggest that educational attainment need not be very advanced to affect practices of malaria prevention and treatment. Primary school attendance was a stronger predictor for use of malaria risk reduction measures than any other selected background characteristics. Chanda *et al* [12] conducted a study on malaria vector control in South Sudan. The study revealed that the peak of malaria transmission season lasting 7 to 8 months of the year south of the country and 5 to 6 months in the north. Wardrop *et al* [13] studied malaria incidence over time and its association with temperature and rainfall in four counties of Yunnan province, China. Seasonal trend decomposition was used to examine secular trends and seasonal patterns in malaria incidence, a Poisson regression with Distributed lag non-linear models were used to estimate the weather drivers of malaria seasonality. The study revealed that there was a declining trend in malaria incidence in all four counties. Kim et al [14], estimated the effects of climate factors on P. vivax malaria transmission using generalized linear Poisson models and distributed lag non-linear models. Their findings suggested that malaria transmission in temperate areas was highly dependent on climate factors. Nath and Mwchahary [15], analyzed the temporal correlation between malaria incidence and climatic variables using malaria incidence rates in Kokrajhar district of Assam over the period 2001 to 2010. Linear regressions were used to obtain linear relationships between climatic factors and malaria incidence. Temperature was found to be negatively correlated with non-forest malaria incidence while relative humidity was positively correlated with forest malaria incidence.

3.0 Material and Methods
3.1 Data
The malaria incidence records over the period 2013- 2015 were considered for analysis, records of malaria incidence were collected from Masinde Muliro University of Science and Technology (MMUST) health facility. The malaria transmission in MMUST occurs all year round with peak seasons following the long rains in March to June and Short rains in October to December. The annual inoculation rates are estimated to be 300 infective bites per person per annum [2]. The data considered were from 50,723 cases reported to health facility between 1st January, 2013 and 31th December, 2015.

Table 3.1: Malaria Incidence between 2013 and 2015

	MONTH	YEAR		
		2013	2014	2015
1	JANUARY	655	2417	2467
2	FEBRUARY	919	2518	1958
3	MARCH	505	1460	2566
4	APRIL	368	1798	2294
5	MAY	390	2151	1102
6	JUNE	445	2730	1130
7	JULY	306	3195	1447
8	AUGUST	330	656	205
9	SEPTEMBER	2216	1671	496
10	OCTOBER	2057	1603	1080
11	NOVEMBER	1802	2842	201
12	DECEMBER	662	1919	162

Monthly malaria incidence rate (MIR) for MMUST was calculated by the formula-

$$MIR = \frac{Total\ malaria\ positive\ cases}{Total\ number\ of\ cases\ tested\ for\ malaria}$$

3.2: METHODS

The response variable is the malaria incidence, and the explanatory variable is time in months for three years. The generalized linear models were applied. The basic count data regression models can be represented and understood using generalized linear models (GLM) framework. Poisson regression is commonly used for modeling the number of cases of disease in a specific population with a certain time. Poisson regression is a special case of (GLM) where the response variable follows Poisson distribution. Poisson models for disease counts are often over-dispersed, in which case the negative binomial regression is more appropriate. The negative binomial model is an extension of Poisson model for incidence rates that allows for the over-dispersion that commonly occurs for disease count [16].

Before fitting one or more distributions to a data set, it is generally necessary to choose good candidates among a predefined set of distributions. This choice may be guided by the knowledge of stochastic processes governing the modelled variable, or, in the absence of knowledge regarding the underlying process, by the observation of its empirical distribution. First of all, it is common to start with plots of the empirical distribution function and the histogram (or density plot), which can be obtained with the plotdist function of the fitdistrplus package. This function provides two plots: the left-hand plot is by default the histogram on a density scale (or density plot of both, according to values of arguments histo and demp) and the right-hand plot the empirical cumulative distribution function (CDF). In addition to empirical plots, descriptive statistics may help to choose candidates to describe a distribution among a set of parametric distributions. Especially the Skewness and kurtosis, linked to the third and fourth moments, are useful for this purpose. A non-zero Skewness reveals a lack of symmetry of the empirical distribution, while the kurtosis value quantifies the weight of tails in comparison to the normal distribution for which the kurtosis equals 3. A Skewness-kurtosis plot such as the one proposed by Cullen and Frey (1999) is provided by the descdist function for the empirical distribution. On this plot, values for common distributions are displayed in order to help the choice of distributions to fit to data. For some distributions (normal, uniform, logistic, exponential), there is only one possible value for the Skewness and the kurtosis. Thus, the distribution is represented by a single point on the plot. For other distributions, areas of possible values are represented, consisting in lines (as for gamma and lognormal distributions), or larger areas (as for beta distribution). Skewness and kurtosis are known not to be robust. In order to take into account the uncertainty of the estimated values of kurtosis and Skewness from data, a nonparametric bootstrap procedure (Efron and Tibshirani, 1994) can be performed by using the argument boot.

3.3 RESULTS

Table 1: Summary statistics for Malaria incidence

Year	minimum	Mean	Standard Deviation	Standard error	Coefficient of variation	1st quartile	3rd quartile	maximum
2013	306.0	887.92	712.44	205.66	0.8024	384.5	1139.75	2216.0
2014	656	2080	705.13	203.55	0.3390	1654	2571	3195.0
2015	162	1259	897.06	258.96	0.7125	423.25	2042	2566.0

It is observed from the Table 1 that the average number of malaria incidence in the year 2013 was 887.92. This rose to 2080 in the year 2014 and reduce to 1259 in the year 2015. This was a drastic decrease after the year 2014. The table also shows that the minimum number of malaria incidence in a month is 162 for the three years and 3195 as the maximum number of malaria incidence in a month for the three years.
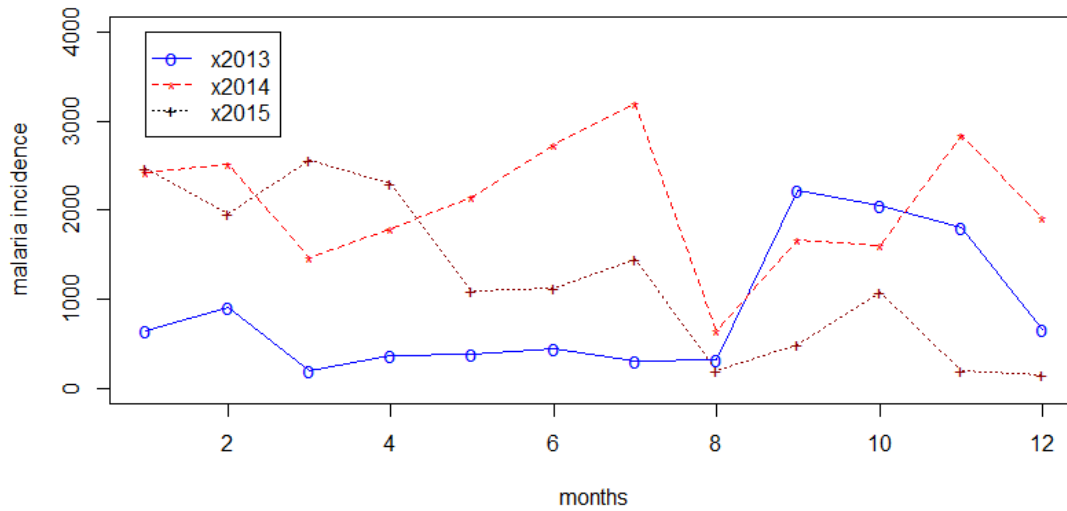
Figure 1: Monthly Malaria incidence

Figure 1 Shows that High endemic occurred in the month of July and lowest endemic in the month of August and December. The results also shows that there was no clear trend of malaria incidence.

In order to choose good candidates among a predefined set of distributions to fit to the malaria incidence data set, an observation of empirical distribution of the dataset was done. Empirical distribution function and histogram was plotted. In addition to empirical plot, descriptive statistics were also done to help in choosing candidates to describe a distribution among a set of parametric distributions. Especially, skewness and kurtosis were useful for this purpose. A skewness-kurtosis plot like the one which was proposed by Cullen and Frey in 1999 was then done as shown in figure 2a,b,3a,b and 4a,b.The function provides two plots; a histogram on a density scale and the plot for the empirical distribution function (CDF). Since skewness and kurtosis are known to be robust, therefore, to take into account, the uncertainty of the estimated values of kurtosis and skewness from data a non-parametric bootstrap procedure is performed by use of argument boot stored in R software. Also the values of skewness and kurtosis get computed on bootstrap samples and get reported on skewness-kurtosis plot.

Table 2: Skewness and Kurtosis for Malaria incidence

Year	Skewness	Kurtosis
2013	1.1772	-0.3092
2014	-0.3434	0.0827
2015	0.1996	-1.4230

Figure 2a below shows a positively skewed distribution and the summery from table 2 give estimated skewness as 1.1772 and estimated kurtosis as -0.3092. The nonzero skewness reveal lack of symmetry for the empirical distribution. The kurtosis is below three hence show that the distribution is platykurtic. Figure 3a below shows a negatively skewed distribution and the summery from table 2 give estimated skewness as -0.3434 and estimated kurtosis as 0.0827. The nonzero skewness reveal lack of symmetry for the empirical distribution. The kurtosis is below three hence show that the distribution is platykurtic. Figure 4a below shows a positively skewed distribution and the summery from table 2 give estimated skewness as 0.1996 and estimated kurtosis as -0.4230. The nonzero skewness reveal lack of symmetry for the empirical distribution. The kurtosis is below three hence show that the distribution is platykurtic. The nonzero skewness revealed lack of symmetry of the empirical distribution. The kurtosis value quantifies the weight of the tails in comparison to the normal distribution for which the kurtosis equals 3.From the Cullen and Gray graph (figure 2b, 3b and 4b), the observed value is outside the normal distribution. Positively skewed functions (Gamma, Lognormal and Weibull) were then considered to be the possible candidate distributions to fit.
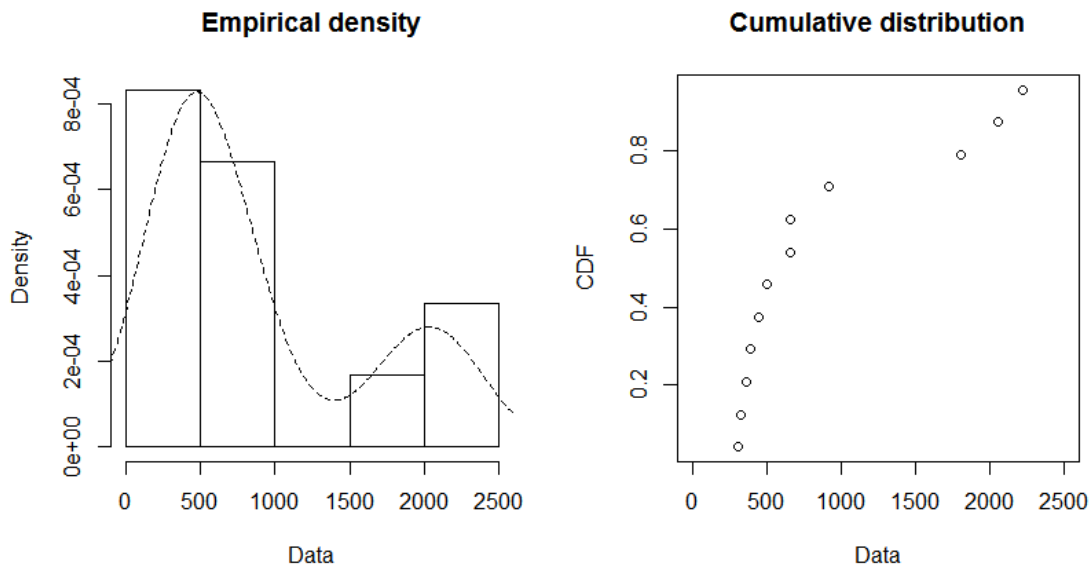
Empirical density

Cumulative distribution

Figure 2a: Histogram and CDF for malaria incidence in 2013 dataset

Cullen and Frey graph
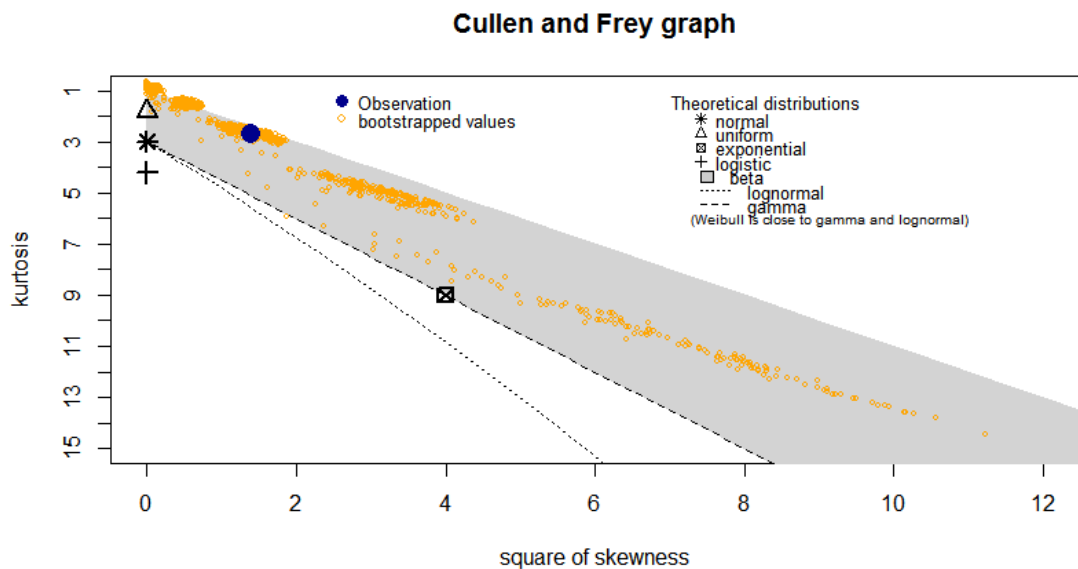
Figure 2b: Skewness-Kurtosis plot for 2013 malaria incidence dataset

Empirical density

Cumulative distribution
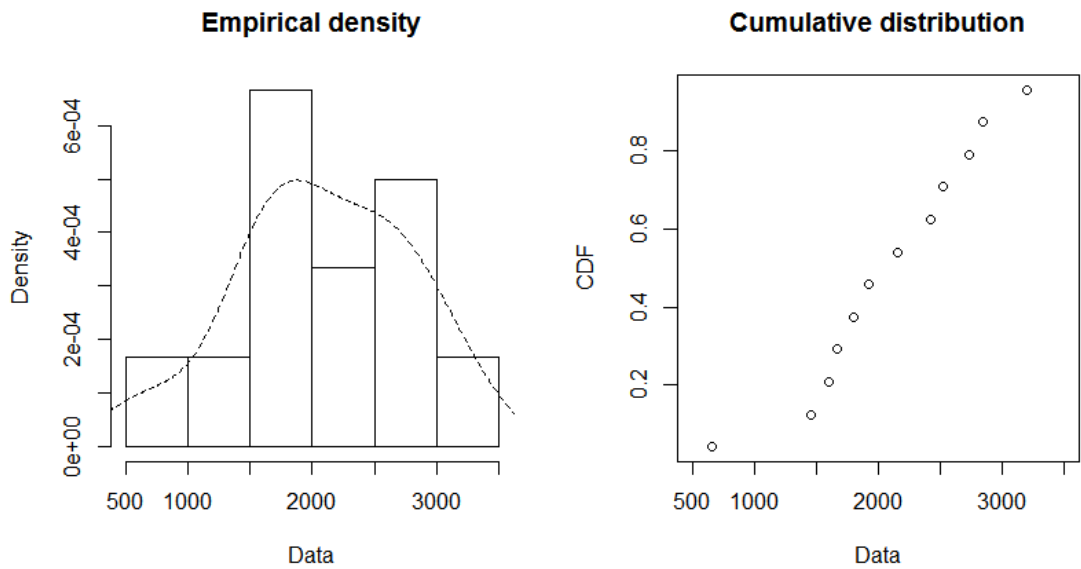
Figure 3a: Histogram and CDF for malaria incidence in 2014 dataset

Cullen and Frey graph
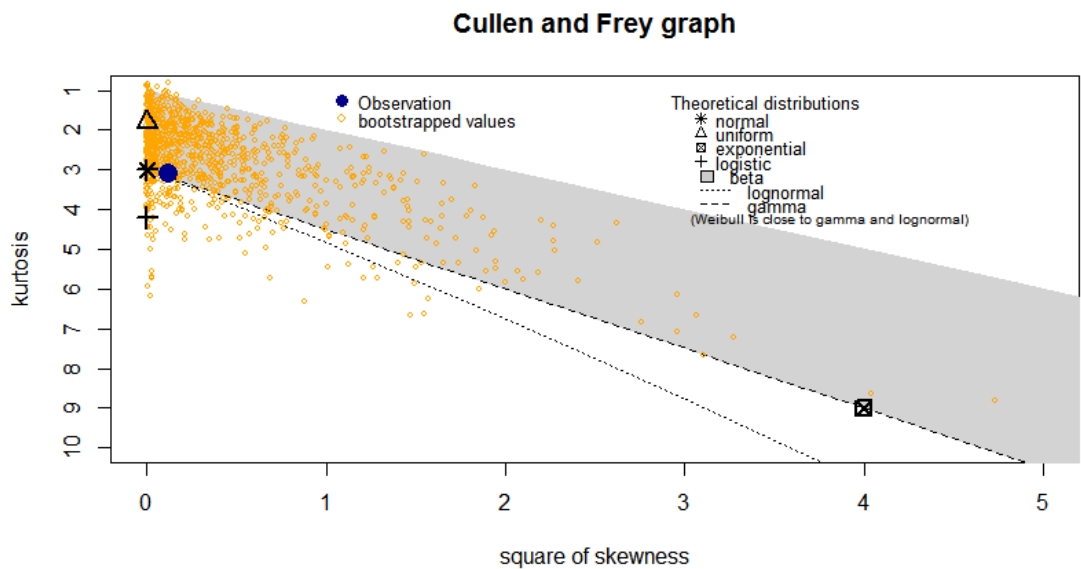
Figure 3b: Skewness-Kurtosis plot for 2014 malaria incidence dataset

Empirical density

Cumulative distribution
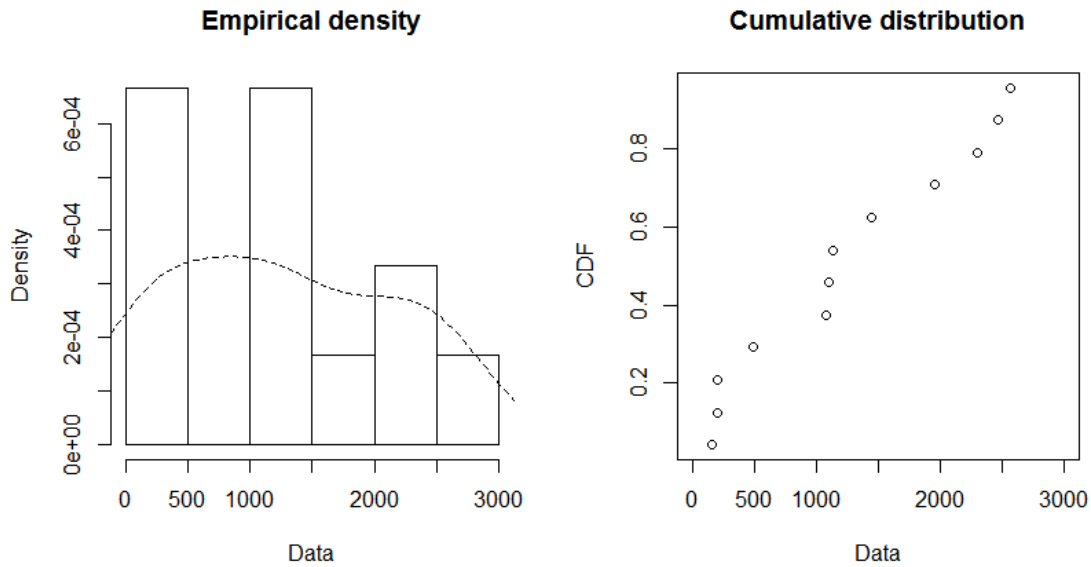
Figure 4a: Histogram and CDF for malaria incidence in 2015 dataset
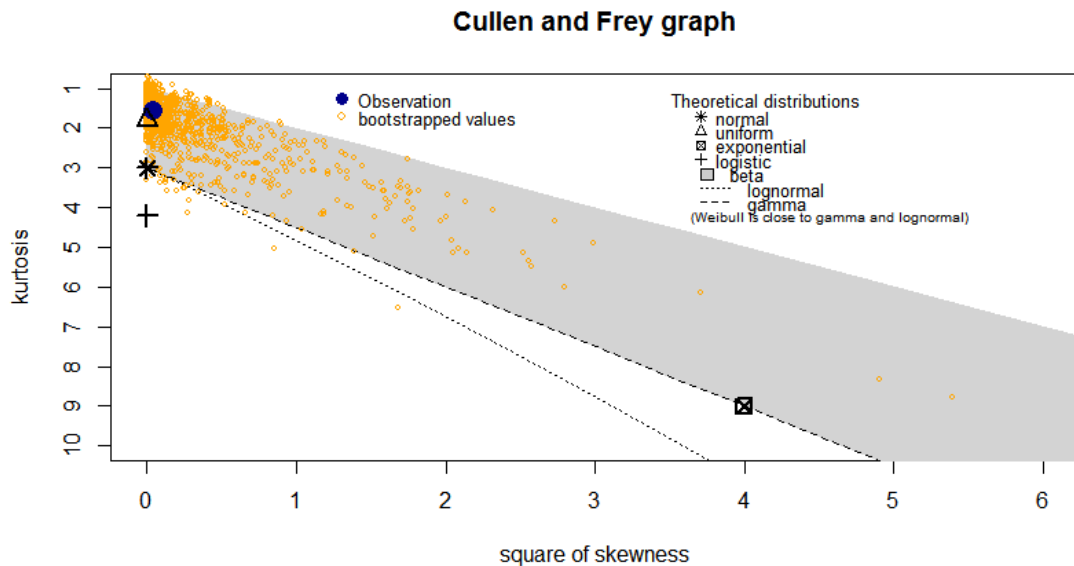
Cullen and Frey graph

Figure 4b: Skewness-Kurtosis plot for 2015 malaria incidence dataset

The Identification of probability distribution process was guided by observation of empirical plots and descriptive statistics which helped in identifying the candidate to describe a distribution among a set of parametric distributions. The candidate distributions identified were Gamma, Lognormal and Weibull. The distributions were then fitted to the malaria dataset.

The density functions of the fitted distributions along with the histogram was also plotted using the R software (Figure 5), as shown below. By looking at the results by balling, lognormal distribution fitted better. Goodness of fit of quantile-quantile plot for the comparison of the three fitted distributions and the empirical distribution was also done using R software (Figure 6). Also, by using AIC values results got from the table 2, lognormal distribution had the lowest AIC value of 185.9875 followed by the Gamma distribution with a value of 187.8815 and then the Weibull distribution with a value of 188.7271. This confirmed the lognormal distribution to be the best fitting distribution for the malaria incidence dataset. The Q-Q plot also in figure 6 showed that lognormal

distribution had most of its points lying along the straight line as compared to the Weibull distribution and the gamma distribution. The shape parameter is greater than one thus the lognormal distribution assumes a mounded but skewed shape. The mean of the distribution is greater than the standard deviation which is responsible for the very shape just mentioned.

Table 3: Distribution fitted to malaria incidence dataset

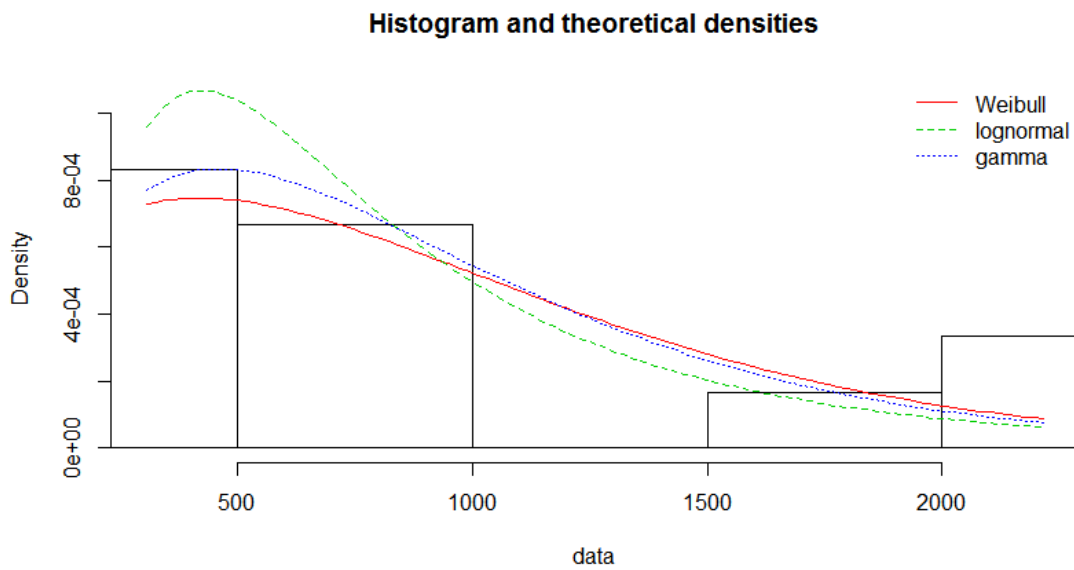| Distribution	Parameter	Estimate	Std Error	AIC	BIC	Loglikelihood
Gamma	Shape	2.0775	0.5194	187.8815	188.8513	-91.94073
	Parameter	0.0023	0.0005			
Weibull	Shape	1.4194	0.3075	188.7271	189.6969	-92.36357
	Scale	986.4877	213.3607			
Lognormal	Meanlog	6.5393	0.2003	185.9872	186.957	-90.9936
	Sdlog	0.6934	0.1416			

Figure 5: Density plots of Gamma, Weibull and lognormal of malaria dataset

The density plot and the CDF plot shown in figure 6 may be considered as the basic classical goodness-of-fit plots. The Q-Q plot emphasizes the lack-of-fit at the distribution tails while the P-P plot emphasizes the lack-of-fit at the distribution center. In figure 6, none of the three fitted distributions correctly describes the center of the distribution, but the lognormal distributions could be preferred for their better description of the right tail of the empirical distribution, especially if this tail is important in the use of the fitted distribution, as it is in the context of malaria incidence dataset.
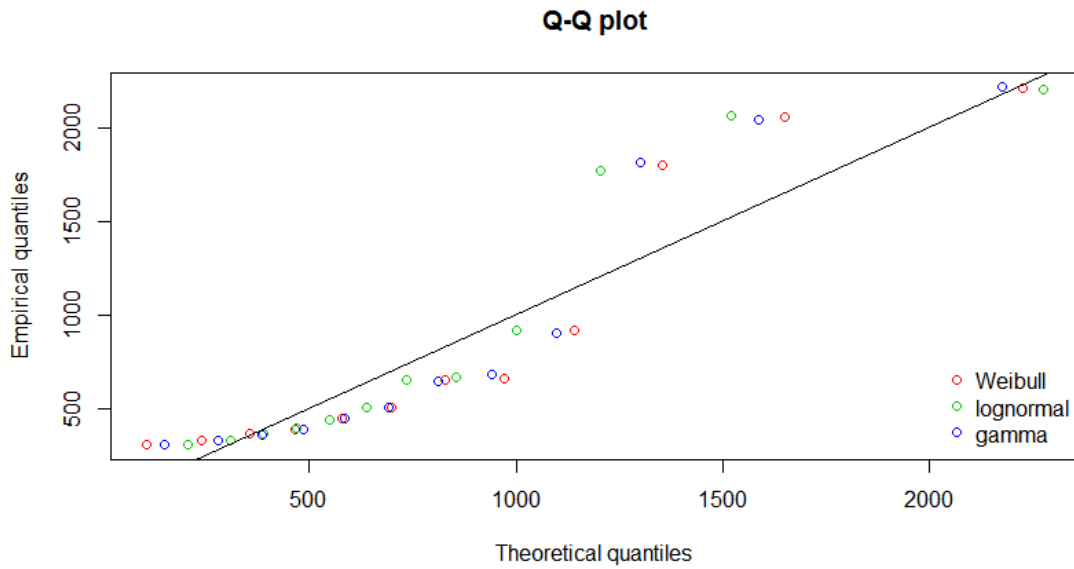
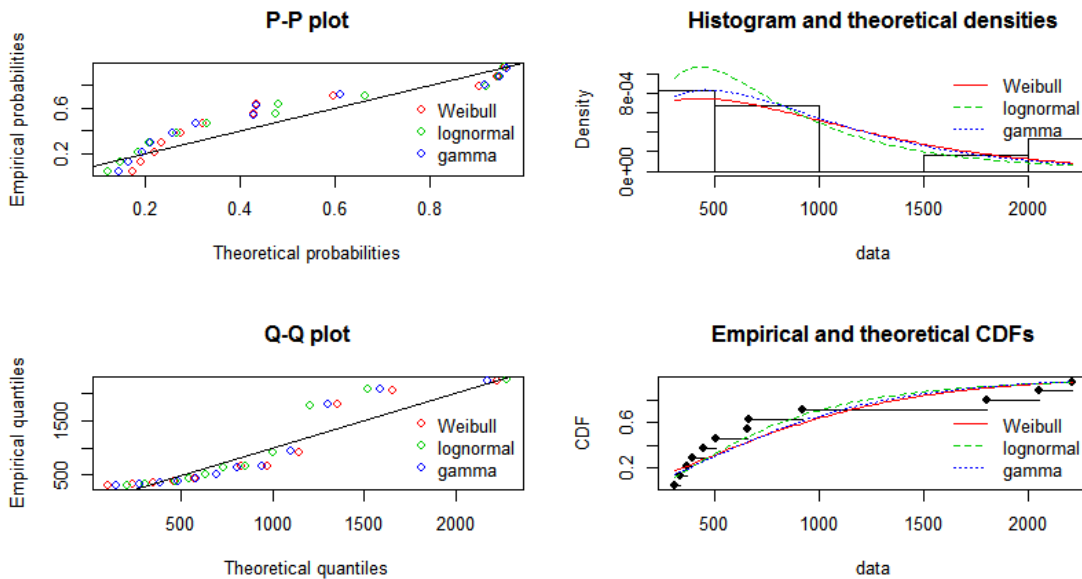**Q-Q plot**

Figure 6 a: Q-Q plot of Malaria Dataset

Figure 6 b: Four Goodness-of-fit plots for various distributions fitted to continuous data (Weibull, gamma and lognormal distribution)

Discussion

The Identification of probability distribution process was guided by observation of empirical plots and descriptive statistics which helped in identifying the candidate to describe a distribution among a set of parametric distributions. The Gamma, Weibull and Lognormal distributions were selected since the data was positively skewed. The parameters of the selected distributions were estimated by the maximum likelihood estimation method using the function 'fitdist' available in the 'fitdistrplus' package of the R software.

In order to choose good candidates among a predefined set of distributions to fit to malaria incidence dataset, an observation of empirical distribution of the data set was done. Empirical distribution function and histogram. The function provided two plots according to values of arguments 'histo' and 'demp' stored in the package 'fitdistrplus';one plot is a histogram drawn on a density scale and the other is the empirical cumulative

distribution function (CDF). A skewness-kurtosis plot was done, the function provides two plots; a histogram on a density scale and the plot for the empirical distribution function (CDF). On such a plot, values for common distributions are displayed in order to help in the choice of distributions to fit to data. Values of skewness and kurtosis get computed on bootstrap samples and reported on skewness-kurtosis plot. The non-zero skewness revealed lack of symmetry of the empirical distribution. The kurtosis value quantifies the weight of the tails in comparison to the normal distribution for which the kurtosis equals. From the Cullen and Gray graph, the observed value is outside the normal distribution. Positively skewed functions (Gamma, Lognormal and Weibull) were considered.

Conclusion

The results obtained suggest that malaria incidence is increasing with time therefore there is need for study that incorporates more explanatory variables. Determination of malaria trend is critical for the design of policies aimed at reducing malaria incidence and improvement of healthcare system. The Poisson regression did not accurately fit the data on malaria incidences due to over dispersion in the data but lognormal was a better fit compared to gamma and Weibull distribution.

Acknowledgements

REFERENCES

1. WHO (2017).Malaria Fact Sheet 2017 Report.
2. Yeka, A., Gasasira, A., Mpimbaza, A., Achan, J., Nankabirwa, J., Nsobya, S., Staedke, S.G., Donnelly, M.J., Wabwire-Mangen, F., Talisuna, A. (2012) Malaria in Uganda: Challenges to Control on the Long Road to Elimination: I. Epidemiology and Current Control Efforts. *Acta Tropica*, 121, 184-195.
3. Gemperli A (2003*). Development of spatial methods for modeling point-referenced spatial data in malaria epidemiology*. PhD thesis, University of Basel, Switzerland.
4. Nicole Cancre (2000). Bayesian analysis of an epidemiological model of Plasmodium falciparum malaria infection in Ndiop, Senegal. *American Journal of Epidemiology*, Oxford University Press, 152(8), pp 760-770.
5. Patience, E.O. and Osagie, A.M. (2014) Modeling the Prevalence of Malaria in Niger State: An Application of Poisson Regression and Negative Binomial Regression Models. *International Journal of Physical Sciences*, 4, 061-068.
6. Martens, P., Kovats, R.S., Nijhof, S., De Vries, P., Livermore, M.T.J., Bradley, D.J., Cox, J. and McMichael, A.J. (1999) Climate Change and Future Populations at Risk of Malaria. *Global Environmental Change*, 9, S89-S107.
7. Teklehaimanot, H.D., Lipsitch, M., Teklehaimanot, A. and Schwartz, J. (2004) Weather-Based Prediction of Plasmodium Falciparum Malaria in Epidemic-Prone Regions of Ethiopia I. Patterns of Lagged Weather Effects Reect Biological Mechanisms. *Malaria Journal*, 3, 41.
8. Nkurunziza, H., Gebhardt, A. and Pilz, J. (2010) Bayesian Modelling of the Effect of Climate on Malaria in Burundi. *Malaria Journal*, 9, 114.
9. Kakchapati, S. and Ardkaew, J. (2011) Modeling of Malaria Incidence in Nepal. Journal of Research in Health Sciences, 11, 7-13.Malaria Vector Control Challenges in South Sudan: Proposed Recommendations. *Malaria Journal*,
10. Sriwattanapongse, W., Meead, S. and Khanabsakdi, S. (2011) Forecasting Malaria Incidence Based on Monthly Case Reports and Climatic Factors in Ubon Ratchathani Province, Thailand, *Science Journal* Ubon Ratchathani University, 2, 17-24.
11. Drebel, T., Kueil, B.G. and Meyrowitsch, D.W. (2013) Prevalence of Malaria and Use of Malaria Risk Reduction Measures among Resettled Pregnant Women in South Sudan. *International Health*, 5, 211-216.
12. Chanda, E., Doggale, C., Pasquale, H., Azairwe, R., Baba, S. and Mnzava, A. (2013) Addressing Malaria Vector Control Challenges in South Sudan: Proposed Recommendations. *Malaria Journal*, 12, 59

13. Wardrop, N.A., Barnett, A.G., Atkinson, J.-A and Clements, A.C.A. (2013) Plasmodium Vivax Malaria Incidence over Time and Its Association with Temperature and Rainfall in Four Counties of Yunnan Province, China. *Malaria Journal*, 12, 452.

14. Kim, Y.-M, Park, J.W and Cheong, H.-K. (2012) Estimated Effect of Climatic Variables on the Transmission of Plasmodium Vivax Malaria in the Republic of Korea. *Environmental Health Perspectives*, 120, 1314-1319.

15. Nath, D.C. and Mwchahary, D.D. (2013) Association between Climatic Variables and Malaria Incidence: A Study in Kokrajhar District of Assam, India: Climatic Variables and Malaria Incidence in Kokrajhar District. *Global Journal of Health Science*, 5, 90.

16. Vennables, W. N. and Ripley, B. D. 2002. *Modern Applied Statistics with S*. 4th Ed. New York, Springer