International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:4, No:1, 2010

Forecasting Malaria Cases in Bujumbura

Hermenegilde Nkurunziza , Albrecht Gebhardt and Juergen Pilz

***Abstract*—**The focus in this work is to assess which method allows a better forecasting of malaria cases in Bujumbura (Burundi) when taking into account association between climatic factors and the disease. For the period 1996-2007, real monthly data on both malaria epidemiology and climate in Bujumbura are described and analyzed. We propose a hierarchical approach to achieve our objective. We first fit a Generalized Additive Model to malaria cases to obtain an accurate predictor, which is then used to predict future observations. Various well-known forecasting methods are compared leading to different results. Based on in-sample mean average percentage error (MAPE), the multiplicative exponential smoothing state space model with multiplicative error and seasonality performed better.

***Keywords*—**Burundi, Forecasting, Malaria, Regression model, State space model.

I. Introduction

IN Burundi, malaria is still a major public health problem in terms of both morbidity and mortality with around 2.5 millions clinical cases and more than 15.000 deaths each year. In 2001, Burundi was the world's most affected country by malaria [1]. Malaria is the single main cause of mortality among pregnant women and children below the age of five, accounting for more than 50 % of all cases. It continues to ravage millions of rural Burundians, despite concerted efforts to reduce malaria mortality [2],[3]. This is often attributed to a number of factors such as the limited access to basic health care due to poverty, limited specialized health facilities, the cost-sharing system, and underfunding of the health sector by the government. Currently the government only allocates 2%-4% of its national budget towards supporting the health sector. The direct economic costs of malaria that result from treatment and from time away from work or school are enormous, but the overall economic impact of malaria is likely to be much more substantial than suggested by estimates of direct costs alone [4].

Healthcare managers in Burundi need simple, accurate and reliable methods for forecasting malaria so that more effective control measures can be undertaken. Decisions on drugs purchases and health plans require predictions of future observations. Stakeholders can gain useful information trough models which are capable of predicting malaria. Very few research works have been proposed in this regard. The authors in [1] proposed the ARIMA model to forecast malaria incidence in Karuzi (a province in central-eastern of Burundi), but they were not able to motivate their choice. In this paper we compare various methods and we base our choice on one of the most reliable criteria, namely the mean average percentage error (MAPE), to choose the most accurate forecasting method.

H. N is a PhD student in the Department of Statistics at the University of Klagenfurt, Austria(corresponding author ; e-mail: hnkurunz@ edu.uni-klu.ac.at).

A. G. is a Lecturer in the Department of Statistics, University of Klagenfurt, Austria(e-mail: albrecht.gebhardt@uni-klu.ac.at).

J. P. is the Head of the Department of Statistics, University of Klagenfurt, Austria(e-mail: juergen.pilz@uni-klu.ac.at).

II. Methods

1. Study Area

Bujumbura, the capital and largest city of Burundi is located in the western part of the country; bordering Tanganyika Lake. The city of Bujumbura has now an area of 11,000 hectares with an average altitude of 820 meters. Bujumbura has a tropical climate and has a dominant sunshine all the year, with an average temperature of 23°c, peaking at 28°c-30°c during the hottest periods (July - September). The population of Bujumbura was estimated at 500,000 inhabitants in 2005. Malaria is the main health threat in Bujumbura [5]. The majority of the Health workers in Burundi are concentrated in Bujumbura.

2. Data description

The goal in this study is to propose a more accurate method for forecasting malaria in Bujumbura, the capital of Burundi, when taking into account the influence of climatic factors. Data on monthly malaria cases in Burundi were obtained from EPISTAT (Epidemiology and Statistics) [6], a department of the Burundi Ministry of health, collecting and storing data on epidemiology all over the country. We collected malaria morbidity data from 1996 to 2007. This is the period where complete data were available from EPISTAT.

Data on monthly cumulative precipitation, for 1996-2007 were obtained from the geographic institute of Burundi (IGEBU) [7]. Data on monthly averages of maximum temperature, minimum temperature, maximum humidity and minimum humidity were also obtained from the same institute. The record of these variables from 1996 to 2007 has remained uniform, with the same measurement instruments, the same calibration and the same precision [1].

3. Forecasting malaria in Bujumbura

The choice of Bujumbura for our study is dictated by the

following reasons. The capital of Burundi, Bujumbura, presents more reliable data as it comprises more than 80% of the health workers [8], and most of those who suffer from malaria seek for a medical service. Health facilities (hospitals, clinics and test material) are of better quality than in other provinces. The decomposition of malaria cases time series (Fig.1) in level, trend and seasonal components suggests a seasonal behavior and an increase of malaria cases, with a peak in 2001 followed by a decrease.

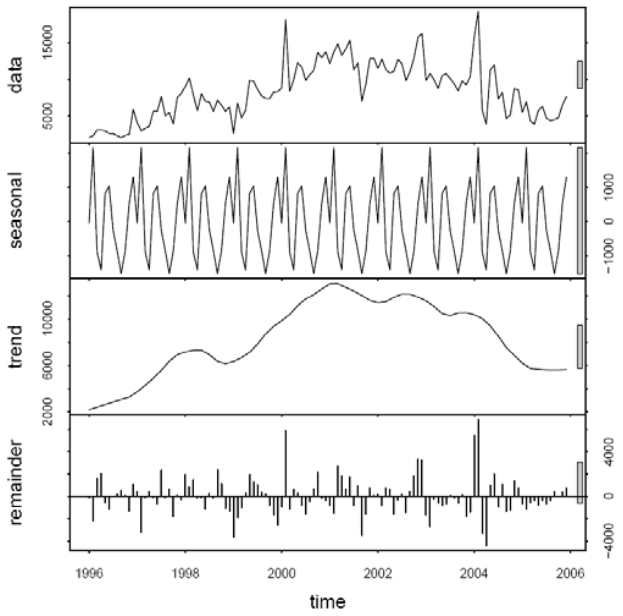

Fig. 1 Time series decomposition of Malaria cases in Bujumbura

The seasonality in the malaria time series might be explained by the seasonality in climatic variables (not shown here). Using only malaria series' own history for forecasting might lose some valuable information contained in the influencing factors [1]. Incorporating covariates might lead to a more accurate forecast compared to considering only the series' past values. The work in [9] shows the limitations of forecasting malaria incidence from historical morbidity patterns alone and indicates the need for improved epidemic forecasting by incorporating external predictors such as meteorological factors. Hence, we propose a hierarchical approach as follows: First we fit a generalized additive (regression) model (GAM) to malaria data to find the "best" predictor. Hereafter, forecasts are made on the basis of this predictor using different methods with the aim to choose the best one.

For our GAM, we have time series data (Y_t, X_t), $t = 1, \cdots, T$, where Y_t is the response variable (malaria cases) and X_t is a vector of covariates (rainfall, maximum and minimum temperature, maximum and minimum humidity). The evolution of Y_t is assumed to be driven by its own past as well as by the covariates X_t [10]. The conditional expectation of Y_t is modelled in the form $h(\eta_t)$, where h is an appropriate response function and η_t is a regression term containing the actual covariates as well as the previous observations of the response. We wish to forecast future observations $Y_{T+1}, Y_{T+2}, \cdots, Y_{T+p}$ if the process has been observed up to time T.

Taking into account the life cycle of the parasite and the incubation period [11], we assume that the number of malaria cases in a given month is associated with that of the previous month as well as climatic conditions of the same and previous months. Most of those who become sick in a given month were bitten by mosquitoes in the previous month. Let $D_t = (X_{t-1}, X_t, Y_{t-1}), \quad X_t = \{X_{it}\}, \quad i = 1, \cdots, 5$ where X_{it} are the five covariates for month t. Y_{t-1} is a variable representing the number of malaria cases of the previous month.

We assume that the distribution of Y_t given D_t belongs to an exponential family, i.e.

$$p(Y_t / D_t, \theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi} \right) c(y, \phi) \qquad (1)$$

Here $b(.), \ c(.), \ \theta$ and ϕ determine the specific response distribution [12-14]. The mean $\mu_t = E(Y_t / D_t, \gamma)$ is linked to an additive predictor η_t by $\mu_t = h(\eta_t)$. Here h is a known response function, γ are unknown regression parameters, and η_t is given by:

$$\eta_t = f_1(Y_{t-1}) + \sum_{i=2}^{6} f_i(X_{it}) + \sum_{i=7}^{11} f_i(X_{i(t-1)}) + \alpha_t \qquad (2)$$

Here, as mentioned above, η_t is the predictor of malaria cases in month t, Y_{t-1} representing malaria cases of the previous month, X_{it} and $X_{i(t-1)}$ are the five covariates (rainfall, maximum temperature, minimum temperature, maximum humidity and minimum humidity) for months t and $(t-1)$, respectively. Further, α_t represents the effect of unobserved variables and f_i are unknown smooth functions of the covariates. We divide the data set into two parts, the first 10 years' data are used as test data (1996-2005) and those of the two remaining years as validation data (2006-2007). A common choice for count data is a Poisson distribution**.** We also assume that malaria cases are Poisson distributed i.e.

$$p(Y_t / D_t, \lambda) = \lambda^y \frac{\exp(-\lambda)}{y!} \qquad (3)$$

Here λ is the expected number of cases. The equation (3) is equivalent to (1) when setting

$\theta(\mu) = \log\lambda, b(\theta) = \lambda, \phi = 1$ (see [15] for more details). We then obtain

$$Y_t = \eta_t + \varepsilon_t \quad (4)$$

where η_t is the predictor of Y_t and ε_t represents the residuals. We checked the goodness of fit as follows:

(a) Fig.2 displays the Q-Q plot , the plot of the residuals and the plot of the response against fitted values.

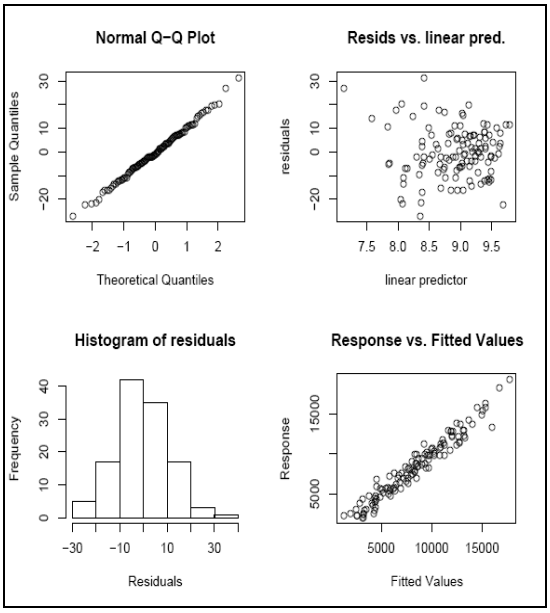

Fig. 2 Model checking plots

The upper left normal QQ-plot is very close to a straight line, suggesting that the distribution assumption in our modelling is realistic. The upper right plot suggests that the variance is almost constant as the mean increases. The histogram of residuals in the lower left plot is almost consistent with normality. The lower right plot of the response against fitted values suggests a positive linear relationship.

(b) In Fig.3, we plot the autocorrelation and the partial autocorrelation functions (ACF and PACF) of the residuals.

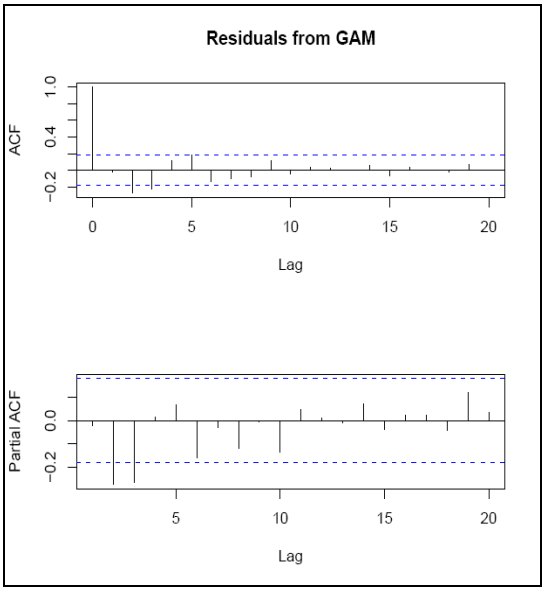

Fig. 3 ACF and PACF of residuals from the GAM

The plot of ACF and PACF of the residuals shows that the residuals are not correlated.

These two plots indicate that η_t in (2) is an accurate predictor.

The decomposition of η_t into mean level, trend and seasonality produces the same behavior as that shown in Fig.1. Since η_t predicts Y_t fairly well, we apply various forecasting methods to η_t, to assess which method is more accurate to forecast malaria cases in Bujumbura. In Table 1 we present values of various in-sample error measures [2],[16] as well as the AIC and BIC for various forecasting methods. In Table 1,

ME is the mean error , $$ME = \frac{1}{m}\sum_{t=1}^{m} e_{n+t} \quad (5.1)$$

RMSE is the root mean squared error,

$$RMSE = \sqrt{\frac{1}{m}\sum_{t=1}^{m} e_{n+t}^2} \quad (5.2)$$

MAE is the mean absolute error

$$MAE = \frac{1}{m}\sum_{t=1}^{m} |e_{n+t}| \quad (5.3)$$

MPE is the mean percent error

$$MPE = \frac{100}{m}\sum_{t=1}^{m} \frac{e_{n+t}}{Y_{n+t}} \quad (5.4)$$

MAPE is the Mean Average Percentage Error,

$$MAPE = \frac{100}{T}\sum_{t=1}^{T}\left|\frac{e_t}{Y_t}\right| \tag{5.5}$$

MASE is the mean averaged scaled error ,

$$MASE = mean(|q_t|)$$

$$\text{with } q_t = \frac{e_t}{\frac{1}{n-1}\sum_{i=2}^{n}|Y_t - Y_{t-i}|} \tag{5.6}$$

Here $e_t = Y_t - \hat{Y}_{t-1}$ is the one-step ahead forecasting error, T is the forecasting horizon, Y_t is the observed value and $\hat{Y}_{t-1}$ is the one-step ahead forecasting from Y_{t-1} (see [16], [3] for more details). AIC is the Akaike Information Criterion $AIC = 2k - 2\ln(L)$ where k is the number of parameters in the model and L is the maximum value of the likelihood function for the estimated model. BIC is the Bayesian Information Criterion $BIC = -2\ln L + k\ln(n)$, where n is the number of observations. ARIMA is the Autoregressive integrated moving average model, ETS stands for Error, Trend, Seasonality or ExponenTial Smoothing in some literature. In its argument, the first letter denotes the error type (A=Additive, M=Multiplicative), the second letter denotes the trend type (N=None, A=Additive, M=multiplicative), and the third letter denotes the season type (N=None, A=Additive, M=Multiplicative). SE is the Simple Exponential, HWA is the HoltWinters model with additive seasonality . HWM is the HoltWinters model with multiplicative seasonality [2-4],[16],[17]. Our model selection is based on the accuracy measured by MAPE. Many authors have suggested the model selection method based on AIC [16],[17]. This method penalizes a model with too many parameters. However, with the current computer's speed, the number of parameters should not be a subject of much concern nowadays. In the forecasting process, a method that minimizes the errors is more valuable. Our choice of MAPE is due to the following reasons:

(a) its simplicity (the test is easy to understand);
(b) MAPE is scale-independent [17];
(c)it quantifies clearly the deviation, in terms of percentage, from the true value (when all the observations are not zero) [16-18];
(d) it is unit free [16];
(e) Hyndman [19] recommends that if the data are positive and much greater than zero, then MAPE is to be preferred.

In our study, amongst all the attempted forecasting methods, the multiplicative exponential smoothing state space with multiplicative error, referred to as (M,N,M) in the literature [17] , has produced small MAPE (see Table 1). This method can be summarized as follows.

II.4. State Space model

The general form of state space model assumes a state vector $x_t = (l_t, b_t, s_t, s_{t-1}, \cdots, s_{t-(m-1)})$ containing unobserved components that describe the level (l_t), trend (b_t) and seasonality (s_t) and state equations of the form

$$Y_t = h(x_{t-1}) + k(x_{t-1})\varepsilon_t \tag{6}$$

$$x_t = f(x_{t-1}) + g(x_{t-1})\varepsilon_t \tag{7}$$

Here Y_t denotes the observation at time t, $\{\varepsilon_t\}$ is a Gaussian white noise process with mean zero and variance σ^2, m is the number of seasons per year. In our study, $m = 12$ as we are dealing with monthly data. Defining $e_t = k(x_{t-1})\varepsilon_t$ and $\mu_t = h(x_{t-1})$ leads to $Y_t = \mu_t + e_t$. The model with multiplicative error is written as

$$Y_t = \mu_t(1+\varepsilon_t) \tag{8}$$

Thus $k(x_{t-1}) = \mu_t$ for this model and $\varepsilon_t = e_t/\mu_t = (Y_t - \mu_t)/\mu_t$. Hence ε_t is a relative error for the multiplicative model [2-4]. Equations for the (M,N,M) model are given by :

$$l_t = l_{t-1}(1+\alpha\varepsilon_t) \tag{9}$$

$$s_t = s_{t-m}(1+\gamma\varepsilon_t) \tag{10}$$

$$Y_t = l_{t-1}s_{t-m}(1+\varepsilon_t) \tag{11}$$

Here α and γ are parameters controlling the smoothness, l_t represents the mean level, and s_t represents the seasonal component. Further, it is assumed that there is only one source of error, i.e. all the observation and state variables are driven by one single error sequence ε_t [2-4],[17],[18].

III. RESULTS

The aim in this study is to propose a more accurate method for forecasting malaria in Bujumbura, the capital of Burundi. Table 1 presents the values of various in-sample error measures as well as the AIC and BIC for various forecasting methods.

TABLE I IN-SAMPLE ERROR MEASURES WITH AIC AND BIC

Model	ME	RMSE	MAE	MPE	MAPE	MASE	AIC	BIC
ARIMA(1,0,1)	133	2242	1657.9	-7.3	24.65	0.98	2201	2212
ARIMA(1,1,1)	86.7	2172.7	1569.1	-5	22.57	0.93	2173	2182
ARIMA(1,0,0)	50	2297.3	1660.5	-9.7	25.24	0.99	2204	2213
ARIMA(0,1,0)	32.2	2423.2	1660.9	-4.1	23.21	0.99	2195	2198
ARIMA(0,1,1)	77.2	2251.3	1647.9	-5.6	23.83	0.98	2180	2185
ARIMA(1,1,0)	38.6	2385.9	1658.8	-4.7	37.33	1.31	2193.7	2199
ARIMA(0,0,1)	12.2	2728.2	2205.8	-19.1	37.33	1.31	2245.8	2254
ETS(M,M,M)	-169	1953.8	1431.7	-7.56	20.81	0.85	2406.7	2454
ETS(M,N,A)	86.3	1999.5	1500.3	-1.78	22.5	0.89	2540.2	2579
ETS(M,N,M)	84.2	1928.7	1459.7	-2.76	**20.38**	0.8	2405	2444
ETS(A,A,A)	-61	1884.2	1415.2	-5.08	21.70	0.84	2418.4	2465.9
ETS(A,N,A)	88.9	1898.1	1419.1	-1.86	21.23	0.84	2414.1	2453.2
SE	77.1	2251.2	1650.2	-5.74	23.91	0.98	2431.1	2436.6
HWA	-501	1989.3	1529.3	-11.69	24.06	0.91	2470.2	2514.8
HWM	-318	1928.9	1414.1	-9.89	21.18	0.84	2403.5	2448.1

Based on the Mean Average Percentage Error (MAPE), the multiplicative exponential smoothing state space with multiplicative error produced the smallest value. Fig.4 represents the malaria time series (dotted-line), predicted values and two years ahead forecast (solid-line) with 50% and 95% credible interval.

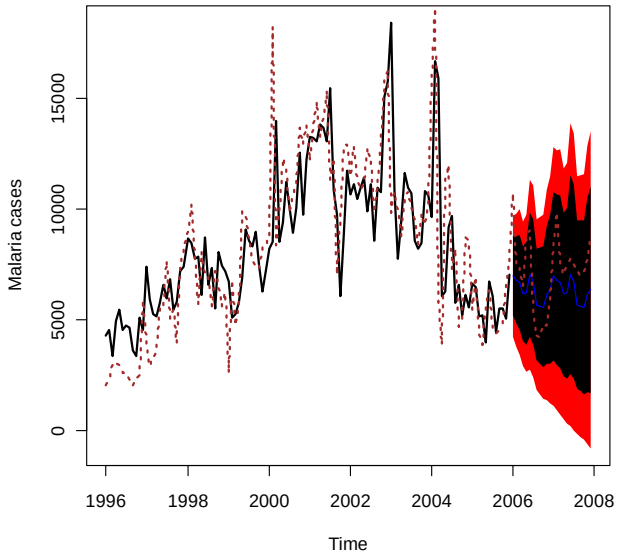

Fig. 4 True and forecasted malaria cases in Bujumbura

The large forecast interval suggests a very high stochasticity in the data. Our method was applied to all provinces of Burundi, leading to the same conclusions. The multiplicative exponential smoothing state space with multiplicative error produced the smallest value in each province.

IV. CONCLUSION

The goal of this work was to assess which forecasting method is more accurate to predict future observations of malaria cases in Bujumbura from data collected over 12 years. We adopted a hierarchical approach to achieve our objective. We first fitted a Generalized Additive Model to find an accurate predictor of malaria cases. We then applied various well-known forecast methods to this predictor. The model selection was based on the mean average percentage error (MAPE). Amongst all the models, the multiplicative exponential smoothing state space with multiplicative error produced the smallest error and hence is the most appropriate to forecast malaria in Bujumbura.

REFERENCES

[1] A. Gomez-Elipe, Otero A, M. Van Herp and A.Aguirre-Jaime, "Forecasting malaria incidence based on monthly case reports and environmental factors in Karuzi, Burundi, 1997–2003," Malaria Journal 2007, 6:129, 1-10.
[2] R. J. Hynman, A. B. Koehler, J. K. Ord, and R. D.Snyder, Forecasting with exponential smoothing: the state space approach. Springer, 2008.
[3] R. J. Hyndman, M. Akram, and B. C. Archibal, " The admissible parameter space for exponential smoothing models," Annals of the Institute of Statistical Mathematics 2008, 60: 407–426.
[4] D. C. Medina, E. S. Findley, and S. Doumbia " State–Space Forecast of Schistosoma haematobium Time-Series in Niono, Mali," PLOS neglected tropical diseases 2008, 8: 1-12.
[5] http://fr.wikipedia.org/wiki/Bujumbura.
[6] Ministry of Health in Burundi , EPISTAT.
[7] Ministry of Planning and Environment in Burundi, IGEBU.
[8] WHO: Stratégie de coopération de l'OMS avec les pays. République du Burundi 2005-2009.
[9] T. A. Abeku, S. J. De Vlas, G. Borsboom, A.Teklehaimanot, A. Kebede, D. Olana, G. J. Van Oortmarssen, and J. D. Habbema, "Forecasting malaria incidence from historical morbidity patterns in epidemic-prone areas of Ethiopia: a simple seasonal adjustment method performs best," Tropical Meicine of International Health 2002, 7(10):851-7.
[10] H. Pruscha, "Residual and forecast methods in time series models with covariates," Collaborative Research Center 386, University of Munich, 1996.
[11] M. J. Bouma , C. Dye , and H. J. Van der Kaay , "Falciparum malaria and climate change in the North West Frontier Province of Pakistan," American Journal of Tropical Medicine and Hygiene 1996,55:131-137.
[12] A. J. Dobson. An Introduction to Generalized Linear Models. Second Edition, Chapman & Hall, 2002.
[13] P.J.Diggle, P. Heagerty , K.Y. Liang S., and Zeger, Analysis of Longitudinal Data. Oxford Science Publications, 1994.
[14] T. J. Hastie and R. J. Tibshirani, Generalized Additive Models. Chapman & Hall, 1997.
[15] L. Fahrmeir and G. Tutz, Multivariate Statistical Modelling based on generalized linear models. Springer, 2001.
[16] S. Wang, Exponential Smoothing for Forecasting and Bayesian Validation of Computer Models, Thesis, Georgia Institute of Technology , 2006.
[17] R. J. Hyndman, A. B. Koehler, R.D. Snyder, and S. Grose, "A state space framework for automatic forecasting using exponential smoothing methods," International Journal of Forecasting 2002, 18: 439– 454.
[18] http://www.ipredict.it/ErrorStatistics.aspx
[19] R. J. Hyndman and A. B. Koehler, "Another look at measures of forecast accuracy," International Journal of Forecasting 2006, 22: 679–688.