International Journal of
Environmental Research and Public Health

MDPI

Article

Generalized Linear Models to Forecast Malaria Incidence in Three Endemic Regions of Senegal

Ousmane Diao [1,*,†,‡], P.-A. Absil [1,‡] and Mouhamadou Diallo [2]

[1] ICTEAM Institute, UCLouvain, B-1348 Louvain-la-Neuve, Belgium; pa.absil@uclouvain.be
[2] Molecular Biology Unit/Bacteriology-Virology Lab, CNHU A. Le Dantec/Université Cheikh Anta Diop, Dakar Fann P.O. Box 5005, Senegal; mouhamdiallo@gmail.com
* Correspondence: ousmane.diao@uclouvain.be or diaoousmane1@gmail.com
† The first author is supported by a fellowship awarded by UCLouvain's Conseil de l'action internationale.
‡ These authors contributed equally to this work.

Abstract: Affecting millions of individuals yearly, malaria is one of the most dangerous and deadly tropical diseases. It is a major global public health problem, with an alarming spread of parasite transmitted by mosquito (Anophele). Various studies have emerged that construct a mathematical and statistical model for malaria incidence forecasting. In this study, we formulate a generalized linear model based on Poisson and negative binomial regression models for forecasting malaria incidence, taking into account climatic variables (such as the monthly rainfall, average temperature, relative humidity), other predictor variables (the insecticide-treated bed-nets (ITNs) distribution and Artemisinin-based combination therapy (ACT)) and the history of malaria incidence in Dakar, Fatick and Kedougou, three different endemic regions of Senegal. A forecasting algorithm is developed by taking the meteorological explanatory variable X_j at time $t - \ell_j$, where t is the observation time and ℓ_j is the lag in X_j that maximizes its correlation with the malaria incidence. We saturated the rainfall in order to reduce over-forecasting. The results of this study show that the Poisson regression model is more adequate than the negative binomial regression model to forecast accurately the malaria incidence taking into account some explanatory variables. The application of the saturation where the over-forecasting was observed noticeably increases the quality of the forecasts.

Keywords: epidemiological data; meteorological data; generalized linear models; parameters estimation; forecasting

Citation: Diao, O.; Absil, P.-A.; Diallo, M. Generalized Linear Models to Forecast Malaria Incidence in Three Endemic Regions of Senegal. *Int. J. Environ. Res. Public Health* **2023**, *20*, 6303. https://doi.org/10.3390/ijerph20136303

Academic Editor: Antonio G. Oliveira

Received: 15 April 2023
Revised: 29 June 2023
Accepted: 1 July 2023
Published: 5 July 2023

1. Introduction

Malaria is a disease caused by a parasitic infection transmitted by a mosquito (female Anophele). It can also be passed to human by blood transfusion, sharing needles or congenitally [1]. According to the 2020 World Health Organization (WHO) report, malaria caused 627,000 deaths, 95% of which were registered in African Region. The number of malaria cases decreased from 2000 (238 million reported cases) to 2019 (229 million reported cases). In spite of this diminution, malaria is still endemic in many countries in the world, particularly in Africa. In Senegal, it constitutes a major public health problem, according to the "Programme national de lutte contre le paludisme (PNLP)".

Several mathematical or statistical models were developed for predicting malaria case incidence. Generalized Linear Models (GLM) [2–4] were used in the literature. Examples of GLM include the Poisson regression developed first by Nelder and Wedderburn [5], the negative binomial (NB) regression [3], the quasi-Poisson regression [5] and the zero-inflated regression [6]. In general, the Poisson regression is very popular for data fitting but its mean-equal-variance property can limit its application on over-dispersed data [5–7]. Also in [8], a multivariate generalized Poisson regression model was defined and studied.

In [6], a model that adapts to malaria incidence using the zero-negative binomial was developed based on climate variables and mosquito density in Limpopo province, South

Africa. The results in [6] show how rain and average temperature affect the incidence of malaria. In [9], authors introduced a model that takes into account the incidence of malaria morbidity and mortality in Akure, Nigeria. In that work, the negative binomial regression model, with log as link function, was used to express the malaria morbidity and mortality incidence as functions of climatic variables. Then, the autoregressive integrated moving average (ARIMA (p, d, q)) model was used to fit the residuals. The findings in [9] revealed that an increase in minimum temperature and relative humidity at a 1-month lag significantly increases the chance of malaria transmission and thereby leads to an increase in the number of inpatient and outpatient individuals, as well as the total number of malaria cases. In another study [10], a Bayesian spatiotemporal analysis has been made to describe year-to-year variation of malaria incidence data from Zimbabwe, and in relation to variation in climate risk factors to enhance our ability of developing an operational malaria early warning system (MEWS) and determine areas prone to climate-driven epidemics. As methods in [10], the authors used the annual proportion of monthly malaria cases and Markham's seasonality index to display between-year variation in the data. Then, the data were fitted with the Bayesian negative binomial models such as the non-spatial model, the spatial model and the spatiotemporal model. In addition, a Markov Chain Monte Carlo (MCMC) simulation was applied to estimate the model parameters. As a result in [10], it was found that a high annual malaria incidence coincides with high rainfall and relatively warm conditions while low incidence years coincide only with low rainfall. In conclusion, all models indicated that the mean annual temperature, rainfall, vapour pressure and normalized difference vegetation index (NDVI) were strong positive predictors of increased annual incidence rate. In [11], the authors applied and compared a Bayesian and classical methods of parameter estimation on the effect of climatic factors in the context of modelling malaria incidence in Limpopo Province, South Africa. In that work, the authors estimated the parameters from a negative binomial model by a Bayesian estimation and maximum likelihood estimation. As result, in [11] the Bayesian method appears more robust than the classical method in analyzing malaria incidence. In [12], the authors include the link between CD4 cell count and influencing covariates of biometric and demographic factors from negative binomial mixed models. A GLM is applied in [13] where authors provide spatially explicit burden estimates of malaria in Senegal using the Senegal Malaria Indicator Survey (SMIS) data and Bayesian geostatistical Zero-Inflated Binomial models based on variable selection methods for spatial data.

A comparative study of existing models was carried out across six countries of Sub-Saharan Africa—Burkina Faso, Nigeria, DRC, Mali, Cameroon, and Niger—over a period of 28 years on malaria incidence in [14]. It is reported in [14] that the SARIMA model was found to work best with time series data that exhibited periodic or seasonal characteristics and was able to predict the seasonal trend of malaria. That model type is only suitable for a stationary or seasonal process. The negative binomial model correctly identified the association between climate variables (taken as explanatory variables) and the rate of malaria transmission. That last model type can make good short-term forecasts, but is not ideal for prediction in subsequent years.

In this paper, a GLM is used in the context of forecasting falciparum malaria incidence count per month based on climate variables and history of falciparum malaria incidence count per month in three endemic regions of Senegal: Dakar (hypoendemic zone), Fatick (endemic zone), and Kedougou (hyperendemic zone). The choice of these three regions is motivated by data availability (notably the presence of villages where longitudinal studies have been conducted), but also by geographical differences: influence of the ocean in the Dakar peninsula, tropical climate in Kedougou, and savanna landscape in Fatik. These fundamental geographical differences allow us to test the applicability of GLMs under drastically different evolutions of the climate variables. A machine learning approach is developed, based on a separation of the data into a train-set to estimate the parameters by maximum likelihood and a test-set to assess the forecast accuracy. Addition and ablation studies are developed to show the influence of each explanatory variable in the forecasts.

A Vuong test reveals that the Poisson distribution is preferred to the negative binomial distribution to model the malaria incidence given the explanatory variables. The forecast accuracy of GLMs with various distributions (Poisson, negative binomial, and Gaussian) and link functions (identity, log, and sqrt) is compared in terms of several model performance metrics. Whereas the best distribution-link combination varies according to the endemic region of interest and the performance metric, the experiments lead to the conclusion that the Poisson distribution with the identity link is overall the most suitable combination. In addition, a saturation method is introduced on the rainfall variable to remedy some overestimations observed during the forecasts. This method has reduced by 4% in the sense of MARE, the over-estimation occurring at the end of 2015 in Dakar.

The paper is organized as follows. The available data are presented in Section 2.1. The models are described in Section 2.2. The estimation and forecasting method containing the train-test machine learning method, the principles of forecasting, the algorithmic protocols, and the saturation concept are presented in Section 2.3. Experimental results and discussion are reported in Section 3 and conclusions are drawn at the end.

2. Materials and Methods

2.1. Data and Notation

There are two principal malaria transmission zones in Senegal:

- "Faciès tropical": corresponding the regions of Ziguinchor, Kolda, Tambacounda, and Kedougou. In that zone, the raining season is the longest and most intensive in the country and covers 5 to 6 months. Malaria cases are observed between 4 to 6 months and the transmission is high (20 to 100 infected bites/human/year).
- "Faciès sahélien": corresponding the regions such as Kaolack, Fatick, Diourbel, Dakar, Thies, Louga, Saint-Louis, and Matam with a less intensive rainy season and covers 2 to 3 months. The transmission is very low in general (0 to 20 infected bites/human/year).

We are interested in three regions of Senegal: Dakar (hypoendemic zone), Fatick (endemic zone), and Kedougou (hyperendemic zone) located in the map presented in Figure 1.

The historical data, such as the monthly falciparum malaria incidence count, the distributed insecticide-treated bed-nets (ITNs), and the distributed Artemisinin-based combination therapy (ACT), between 2008 and 2016, come from the "Programme national de lutte contre le paludisme (PNLP)" of Senegal (https://www.dropbox.com/s/0p4uc2dihfhr9cb/Dakar.csv?dl=0 (accessed on 3 July 2023)). In this study, we consider as malaria cases, the cumulative number of confirmed tests by Rapid diagnostic tests (RDTs) during the month, in all individual groups [15,16]. Malaria cases are confirmed by the methods validated by the "Programme national de lutte contre le paludisme (PNLP)" of Senegal in accordance to the WHO guidelines. The main method is the rapid diagnostic test (RDT) even if it has been recently discovered that this test could miss up to 20% of malaria cases [17]. Due to unavailability of the RDT in some deep localities and the lack of materials to keep it, some districts use the "goutte épaisse", which is a very old method and less sensitive than the RDT. There also is a more sensitive but very expensive test: polymerase chain reaction (PCR), used only in some high level medical research institute in Senegal. Then, we grouped the data from all the different big sanitary districts in each region such as Dakar, Fatick, and Kedougou. The hourly meteorological data, such as the temperature, the relative humidity, and the rainfall, between 2008 and 2016, come from meteoblue (https://www.meteoblue.com/historyplus (accessed on 3 July 2023)). To obtain adequate meteorological data for our study (monthly time unit) we add up all the measured values of the month for rainfall, and we calculate the mean of all the measured values of the month for temperature and humidity.

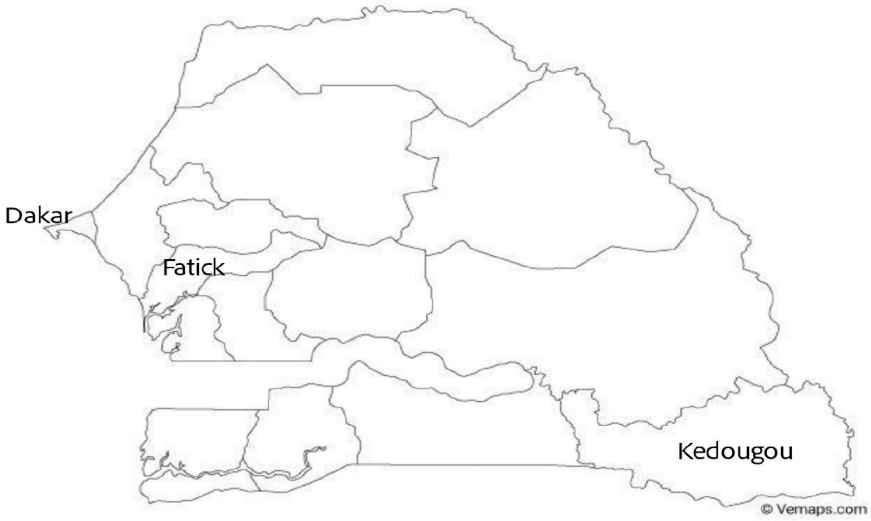

Figure 1. Location of Dakar, Fatick, and Kedougou in Senegal. The choice of these three regions is motivated by data availability (notably the presence of villages where longitudinal studies have been conducted), but also by geographical differences: influence of the ocean in the Dakar peninsula, tropical climate in Kedougou, and savanna landscape in Fatick. These fundamental geographical differences allow us to test the applicability of GLMs under drastically different evolutions of the climate variables.

2.1.1. Response Variable

In this study, the response variable (or explained variable) is the falciparum malaria incidence count per month noted by $Y_o(t)$.

2.1.2. Independent Variables

The available explanatory variables of this study are the history of falciparum malaria incidence count per month ($Y_o(t-\delta)$), the rainfall ($R(t-\delta)$, mm per month), the average temperature ($T(t-\delta)$, °C per month), the relative humidity ($H(t-\delta)$, % per month), the number of insecticide treated bed-nets distributed per month ($B(t-\delta)$), the number of anti-malarial drugs distributed per month ($A(t-\delta)$) where we consider $\delta = h, h+1, \ldots, 6$ (h represents a forecast horizon), and an artificial vector that we call intercept vector I equal to 1 all t.

We are interested in the meteorological explanatory variables such as the rainfall (R), the average temperature (T) and the relative humidity (H) because they are known to influence the mosquitoes (Anopheles) ecology by affecting its distribution, seasonality, and transmission intensity [18,19]. For example, the temperature influences the sporogonic development duration of the parasite and many parameters related to the mosquitoes such as: the biting rate, the egg deposition rate, and the death rate of immature and adult mosquitoes [19]. The rainfall influences the availability and the quality of the larval breeding grounds [20] and the maturation of immature mosquitoes [19]. As for the bed-net (B) and the drugs (A) distributed, we took them because we suppose that they constitute the main factors fighting against malaria [16], reducing the morbidity and the mortality of the disease. Needless to say, the number of bed-nets actually used would be a more suitable explanatory variable, but these data are not available. Note that, in contrast with the physical explanatory variables (R, T, and H), the human explanatory variables (B and A) may depend on the incidence in the past, and they may also be influenced by models used by the health authorities. For this reason, in most of our experiments, we only consider the physical explanatory variables.

All simulations and data analysis are carried out with Jupyter Notebook (Anaconda 3) and Spyder (Python 3.7). Figures 2–7 (reproduce with Malaria_inci_and_variable_2022_07_23.py) present the plots of some variables in order to illustrate their annual distribution. Figures 2–4

show the variations of malaria cases in relation to the rainfall and Figures 5–7 show the variations of malaria cases in relation to the bed-net.

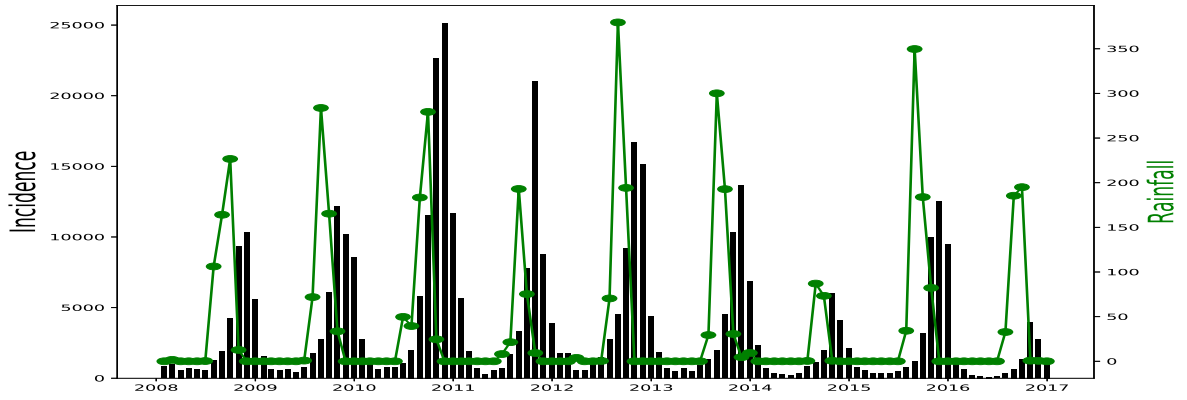

Figure 2. Malaria (falciparum malaria incidence count per month, black) and rainfall (mm per month, green) in Dakar, 2008–2016.

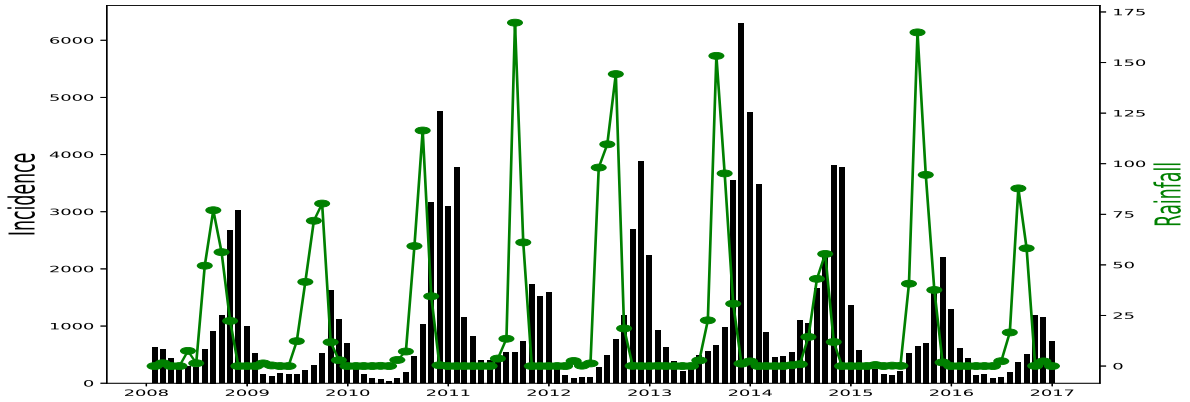

Figure 3. Malaria (falciparum malaria incidence count per month, black) and rainfall (mm per month, green) in Fatick, 2008–2016.

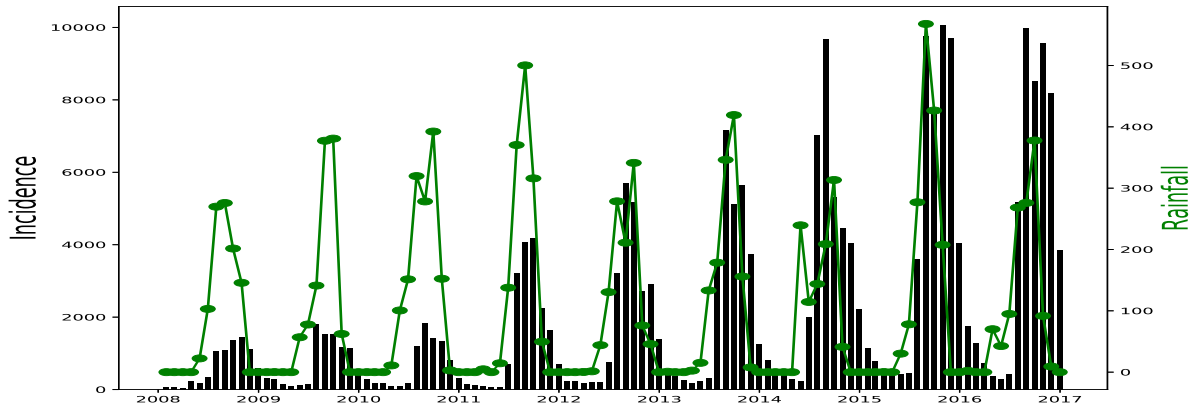

Figure 4. Malaria (falciparum malaria incidence count per month, black) and rainfall (mm per month, green) in Kedougou, 2008–2016.

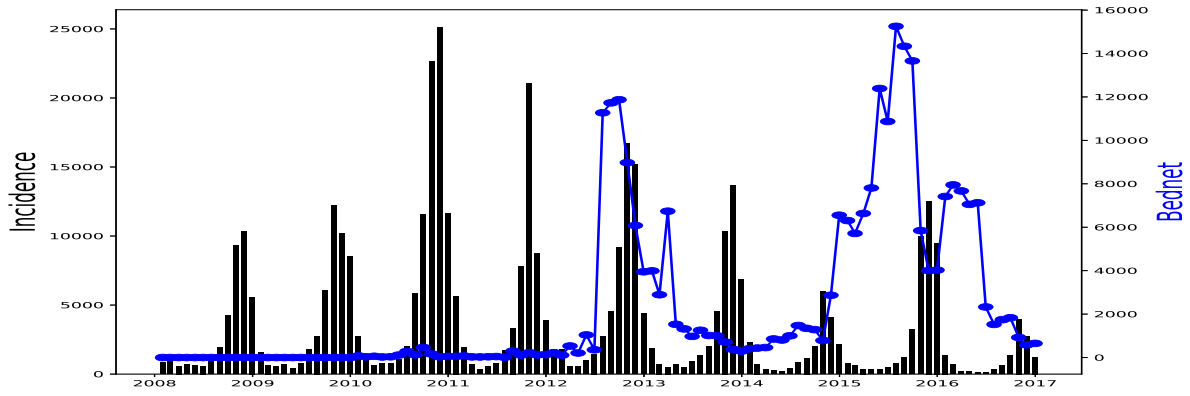

Figure 5. Malaria (falciparum malaria incidence count per month, black) and Bed-net distributed (the number of insecticide treated bed-nets distributed per month, blue) in Dakar, 2008–2016.

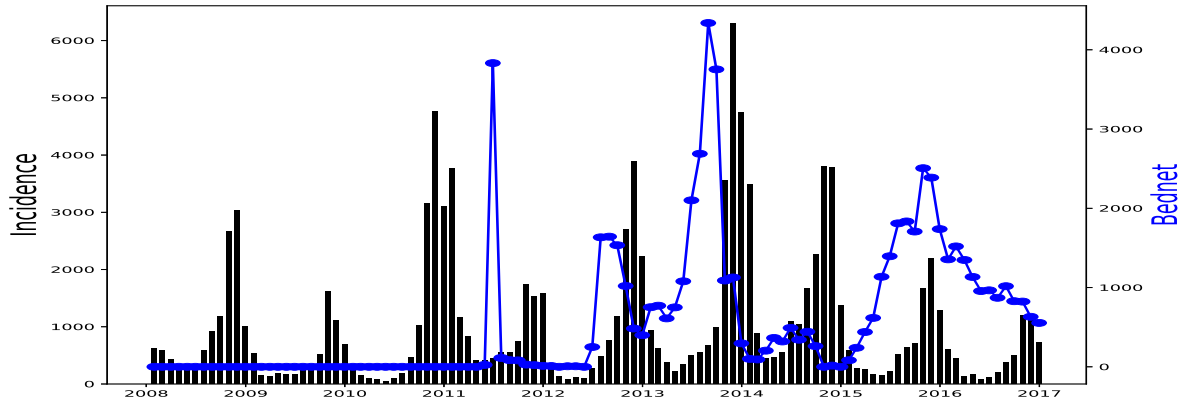

Figure 6. Malaria (falciparum malaria incidence count per month, black) and Bed-net distributed (the number of insecticide treated bed-nets distributed per month, blue) in Fatick, 2008–2016.

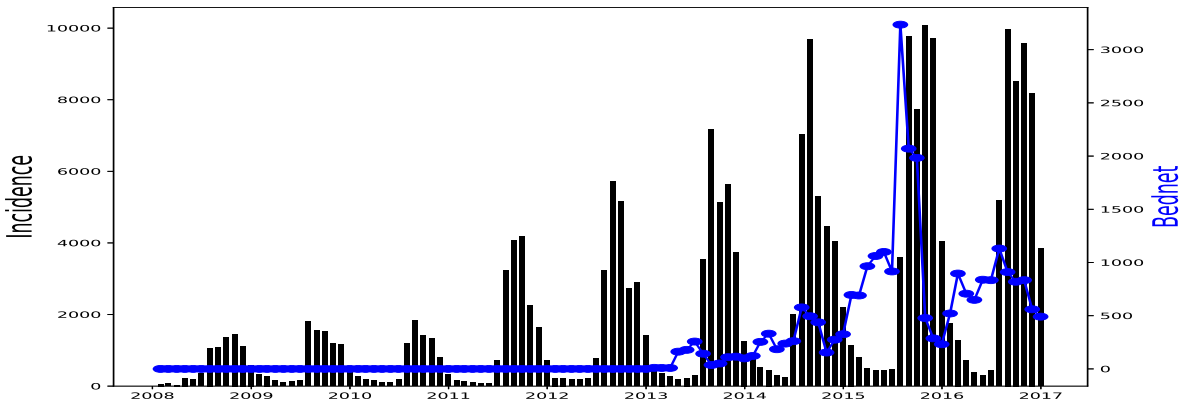

Figure 7. Malaria (falciparum malaria incidence count per month, black) and Bed-net distributed (the number of insecticide treated bed-nets distributed per month, blue) in Kedougou, 2008–2016.

Our experiments use the epidemiological data from PNLP and the meteorological data from meteoblue from 31 January 2008 to 31 December 2016, in Dakar, Fatick and Kedougou.

2.2. Models

We have n data points $(X_{t1}, X_{t2}, ..., X_{tk}, Y_t) \in \mathbb{R}^{k+1}$ for $t = 1, ..., n$ where k is the number of explanatory variables (including the intercept vector) and n is the number of months. We want to build a generalized linear model (GLM) of the response vector Y using the k explanatory variables $X_1, ..., X_k$, according to the diagram Equation (1), where we denote by R-v: Random variable, L-f: Link function and D-c: Deterministic component. According to [3], the link function permits the mean (μ) of the t th observation and its linear predictor (η) to be related. We let $X_{t1} = 1$, $t = 1, \ldots, n$ making β_1 the intercept. In the D-c block, the regression coefficients $\beta_1, \ldots, \beta_k$ are to be estimated on the train-set.

$$\text{R-v} : Y_t \sim f(Y_t; \mu_t) \xleftarrow{\mu_t} \text{L-f} : g(\mu_t) = \eta(X_t) \xleftarrow{\eta} \text{D-c} : \eta(X_t) = \sum_{j=1}^{k} \beta_j X_{tj}. \tag{1}$$

According to the studies in [3,12,21], candidate distributions for viable modeling include the Poisson and negative binomial (NB) distributions. Now, we are going to present these two regression models in the following Sections 2.2.1 and 2.2.2.

2.2.1. Poisson Regression Model

The Poisson distribution is probably the most used discrete distribution because of its simplicity, according to [11]. Its conditional probability mass function is defined as in [2,11] by

$$\begin{aligned} f(Y_t; \mu_t) &= \frac{\mu_t^{Y_t} \exp(-\mu_t)}{Y_t!} \\ &= \exp[Y_t \log(\mu_t) - \mu_t - \log(Y_t!)]. \end{aligned}$$

According to Equation (1), we have

$$\begin{aligned} \mu_t &= g^{-1}(X_t^T \beta) \\ &= g^{-1}(\beta_1 X_{t1} + \ldots + \beta_k X_{tk}). \end{aligned} \tag{2}$$

2.2.2. NB Regression Model

The Poisson–Gamma mixture distribution is the negative binomial distribution, according to [6]. Its probability mass function is given as in [5–7] by

$$f(Y_t; \mu_t, \alpha) = \frac{\Gamma(Y_t + \frac{1}{\alpha})}{\Gamma(\frac{1}{\alpha})\Gamma(Y_t + 1)} \left(\frac{1}{1 + \alpha\mu_t}\right)^{\frac{1}{\alpha}} \left(\frac{\alpha\mu_t}{1 + \alpha\mu_t}\right)^{Y_t}, \tag{3}$$

where Γ is the gamma function. Its mean is μ_t and its variance is $\mu_t + \alpha\mu_t^2$, where α is termed the distribution parameter. Note that, if $\alpha \to 0$, the negative binomial converges to the Poisson distribution.

The Section Appendix A describes how the regression coefficients are computed.

2.3. Estimation and Forecasting Methods

2.3.1. Train and Test Sets

We have $t_s < t_i < t_c < t_e$, where t_s is the initial time of the data, t_i is the initial time of the observed malaria incidence, t_c is the end time of the train set, t_e is the end time of the test set.

2.3.2. Parameter Estimation and Principles of Forecasting

We train the model by taking the observed malaria incidence ($Y_o(t)$) in $[t_i, t_c]$ and the explanatory variables ($X_{tj}, j = 1, \ldots, k$) in $[t_i - \delta, t_c - \delta]$. We did this in order to have the regression coefficients βs and the dispersion parameter α (only with NB regression model).

In [2], the authors suggest to calculate α using a technique that they call auxiliary Ordinary Least Squares (OLS) regression without a constant. In the negative binomial

case, a first estimation of the βs is obtained by the procedure of the Poisson case. Then, α is computed by OLS method. Finally, the βs are re-estimated by maximizing the log likelihood (Equation (A3)) wrt β.

Then, we make the forecasts, according to Algorithm 1, with the coefficients found in the train period. We assess the model accuracy in the test period $[t_c + 1, t_e]$ by comparing the theoretical (mean μ_t) and the observed ($Y_o(t)$) incidences.

Algorithm 1 describes the train-test procedure. For the link function g, the choices identity, log and sqrt are available in the Python library `statsmodels.genmod.families.links`. The forecasts are obtained with the formula

$$\mu_t = g^{-1}(\beta_1 X_1(t) + \sum_{j=2}^{k} \beta_j X_j(t - \delta)). \tag{4}$$

Algorithm 1: Forecasting Algorithm

Input: t_s, t_i, t_c, $t_e \leftarrow$ times of Section 2.3.1;
$\ell \leftarrow$ vector (Section 3.1);
$h \leftarrow$ forecast horizon ($h \geq 1$);
$Y_o \leftarrow$ observed malaria incidence (dependant variable);
$X = \{X_1, X_2, \ldots, X_k\} \leftarrow$ set of explanatory variables;
$f \leftarrow$ distribution (Poisson or NB);
$g \leftarrow$ link (identity, log, or sqrt);

Output: $\hat{Y} \leftarrow$ the forecasted vector of malaria incidence;

1 $Y_{train} = Y_o(t_i : t_c)$;
2 $X_{train} = \{X_1(t_i : t_c), X_2(t_i - \ell_2 : t_c - \ell_2), \ldots, X_k(t_i - \ell_k : t_c - \ell_k)\}$;
3 Fit the GLM with distribution f and link g in the train period using the Python library `statsmodels.genmod.families`;
4 Get the regression coefficients β_j, $j = 1, \ldots, k$;
5 **for** $t \in [t_c + 1, t_e]$ **do**
6 $\quad \hat{Y}(t) = g^{-1}\left(\beta_1 X_1(t) + \sum_{j=2}^{k} \beta_j X_j(t - \ell_j)\right)$.

2.3.3. Saturation Method

We would like to test a saturation in the explanatory variables, in particular rainfall. The motivation is that additional rainfall should have less impact on the malaria incidence in a wet period than in a dry period. The saturation can be simply a hyperbolic tangent function. We posit that, instead of being linear, the contribution of rainfall to $\eta(X)$ is affected by a saturation, which we model by

$$R_{sat} = \gamma \tanh(R/\gamma). \tag{5}$$

We estimate the parameter γ by making a research in many initial values of γ in order to find the more adapted value which gives the low RMSE_train after fitting the GLM. This procedure is entirely described in Algorithm 2.

Algorithm 2: Forecasting Algorithm with Saturation

Input: t_s, t_i, t_c, $t_e \leftarrow$ times of Section 2.3.1;
$\ell \leftarrow$ vector (Section 3.1);
$h \leftarrow$ forecast horizon ($h \geq 1$);
$Y_o \leftarrow$ observed malaria incidence (dependant variable);
$X = \{X_1, X_2, \ldots, X_{k-1}, R\} \leftarrow$ set of explanatory variables;
$f \leftarrow$ distribution (Poisson or NB);
$g \leftarrow$ link (identity, log, or sqrt) ;

Output: $\hat{Y} \leftarrow$ the forecasted vector of malaria incidence;

$Y_{\text{train}} = Y_o(t_i : t_c)$;

$X_{\text{train}} = \{X_1(t_i : t_c), X_2(t_i - \ell_2 : t_c - \ell_2), \ldots, X_{k-1}(t_i - \ell_{k-1} : t_c - \ell_{k-1}), R(t_i - \ell_R : t_c - \ell_R)\}$;

for $\gamma \in [\gamma_{\min}, \gamma_{\max}]$ **do**
  $R_{\text{sat}} = \gamma \tanh(R(t_i - \ell_R : t_c - \ell_R)/\gamma)$ // calculate the saturated rainfall;
  $X_{\text{train-sat}} = \{X_1(t_i : t_c), X_2(t_i - \ell_2 : t_c - \ell_2), \ldots, X_{k-1}(t_i - \ell_{k-1} : t_c - \ell_{k-1}), R_{\text{sat}}\}$;
  Fit the Poisson distribution and link g with Y_{train} and $X_{\text{train-sat}}$ in order to obtain the predicted mean (μ);
  $\text{RMSE}(\gamma) = \sqrt{\frac{\sum_{t=t_i}^{t_c} (Y_o(t) - \mu_t)^2}{t_c - t_i + 1}}$ // Calculte the RMSE for every value of γ

$\gamma_{\text{opt}} = \arg\min \text{RMSE}$ // Get the γ_{opt};

$R_{\text{sat-opt}} = \gamma_{\text{opt}} \tanh(R(t_i - \ell_R : t_c - \ell_R)/\gamma_{\text{opt}})$ // Re-calculate the saturated rainfall with γ_{opt};

$X_{\text{train-opt}} = \{X_1(t_i : t_c), X_2(t_i - \ell_2 : t_c - \ell_2), \ldots, X_{k-1}(t_i - \ell_{k-1} : t_c - \ell_{k-1}), R_{\text{sat-opt}}\}$;

Fit the GLM with distribution f and link g with Y_{train} and $X_{\text{train-opt}}$ using the Python library `statsmodels.genmod.families`;

Get the regression coefficients β_j, $j = 1, \ldots, k$;

for $t \in [t_c + 1, t_e]$ **do**
  $\hat{Y}(t) = g^{-1}\left(\beta_1 X_1(t) + \sum_{j=2}^{k-1} \beta_j X_j(t - \ell_j) + \gamma_{\text{opt}} \tanh(R(t - \ell_R)/\gamma_{\text{opt}})\right)$.

3. Results and Discussion

In this section, we first present and discuss the results of the correlation between the explained variable and each explanatory variable. We define metrics. Then, we present the forecast results from the Algorithm 1. Finally, we present the results from other methods such as addition study, ablation study, and saturation.

3.1. Determination of Lags

For Dakar data in Figure 2, we observe, every year, an increase in malaria cases in the rainy season (from May or June to October or November), reaching a peak around the month of October or November, and a decrease to become stationary along the dry season. This situation is also observed in Fatick (Figure 3) and Kedougou (Figure 4), and proves the seasonality of the malaria cases from these three regions.

The plots of the malaria cases and the explanatory variables (e.g., Figures 2–4) reveal that there is a delay between the maximum of malaria cases and the maximum of the explanatory variable) in the three data sets. This delay is called `lag` and represented by ℓ in the formulae. We set

$$\ell_j := \underset{\delta \in \{h, h+1, \ldots, 6\}}{\arg\max} \left| r(Y_o(t_i : t_e), X_j(t_i - \delta : t_e - \delta)) \right|, \tag{6}$$

where r denotes the sample Pearson correlation coefficient. The statistical results are presented in Table 1 (reproduce with Determination_of_lag_2022_07_23.py) and the evo-

lution of correlations as function of `lags` is illustrated in Figures 8–10 (reproduce with Correlation_Plots_2022_07_23.py).

Table 1. Sample Correlations between the falciparum malaria incidence count per month and the explanatory variables in Dakar, Fatick, and Kedougou, from 2008 to 2016. These correlations are the maximum values in absolute value obtained at index ℓ. The statistical significance of these correlations is tested by calculating the *p*-value associated with the Pearson correlation coefficient by using the `Scipy pearsonr()` function, which returns the Pearson correlation coefficient along with the two-tailed *p*-value. Correlation, lag and *p*-values are reported.

	Dakar	**Fatick**	**Kedougou**
$Y_{o\{t\}}$ and $R_{t-\ell}$	0.79 $\ell = 2$ 4.48×10^{-23}	0.62 $\ell = 3$ 2.14×10^{-12}	0.61 $\ell = 1$ 9.93×10^{-12}
$Y_{o\{t\}}$ and $T_{t-\ell}$	0.65 $\ell = 1$ 8.73×10^{-14}	0.43 $\ell = 4$ 5.04×10^{-06}	0.56 $\ell = 5$ 4.86×10^{-10}
$Y_{o\{t\}}$ and $H_{t-\ell}$	0.56 $\ell = 5$ 7.43×10^{-10}	0.69 $\ell = 3$ 5.88×10^{-16}	0.58 $\ell = 1$ 1.50×10^{-10}
$Y_{o\{t\}}$ and $B_{t-\ell}$	0.15 $\ell = 6$ 0.129	0.38 $\ell = 3$ 5.99×10^{-05}	0.60 $\ell = 3$ 3.16×10^{-11}
$Y_{o\{t\}}$ and $A_{t-\ell}$	0.85 $\ell = 0$ 8.61×10^{-30}	0.88 $\ell = 0$ 5.84×10^{-35}	0.92 $\ell = 0$ 6.48×10^{-43}
$Y_{o\{t\}}$ and $Y_{o\{t-\ell\}}$	0.72 $\ell = 1$ 9.55×10^{-18}	0.75 $\ell = 1$ 1.31×10^{-19}	0.80 $\ell = 1$ 4.28×10^{-24}

Since these p-values are less than 0.05, we could conclude that there is a statistically significant correlation between the malaria incidence and the explanatory variable at the delay considered. The only exception is the bed-nets distributed in Dakar where the *p*-value is 0.129 > 0.05.

The rainfall is highly and positively correlated, meaning that if the rainfall increases, the malaria incidence will increase. These results are also found in [11], revealing that rainfall was a strong positive predictor of increasing the annual incidence rate. The bed-net distribution is weakly correlated in Dakar. Then, it is positively correlated in Fatick (even if the value is low) and in Kedougou with a `lag` of three months. Indeed, the situation in Fatick and Kedougou can be explained by the fact that the authorities anticipate the bed-net distribution three months before the beginning of the rainy season. On the other hand, in Figure 7, we observe that the bed-net distribution is not regular in Kedougou because its behavior is very different in the last two years of the dataset. For all these reasons, we do not use the bed-net distribution (B) as an explanatory variable. The optimal `lag` between the malaria cases and the drugs (A) is equal to 0 in all the regions. This result means that the drugs is not a real predictor because it is usually taken after appearing some malaria symptoms from a patient. That reason leads us to do not consider the drug as an explanatory variable. But it can help to cure some sick people and to reduce future malaria cases. In Figures 8–10, the correlation between malaria cases at time t and the malaria in the past ($t - \delta$) decreases for all values of δ meaning that the malaria in the past, as an explanatory variable, becomes less and less important when δ becomes larger. The

explanatory variables for $Y(t)$ are thus finally $Y_o(t-\ell_{Y_o}), R(t-\ell_R), T(t-\ell_T), H(t-\ell_H)$, and 1 for the intercept.

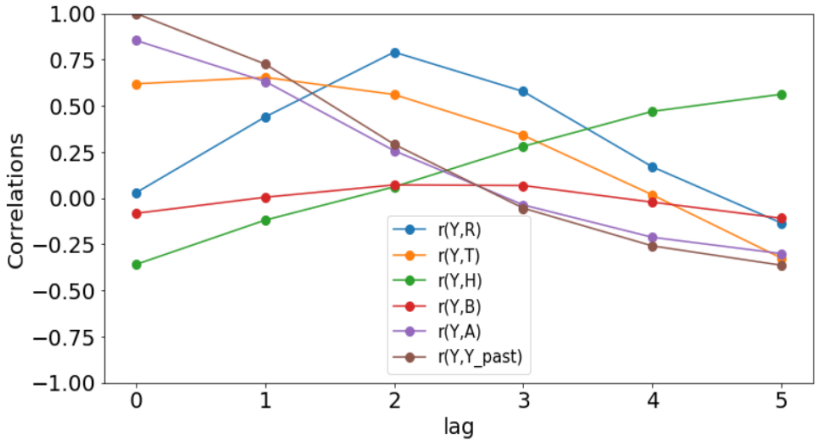

Figure 8. Correlation plots for Dakar.

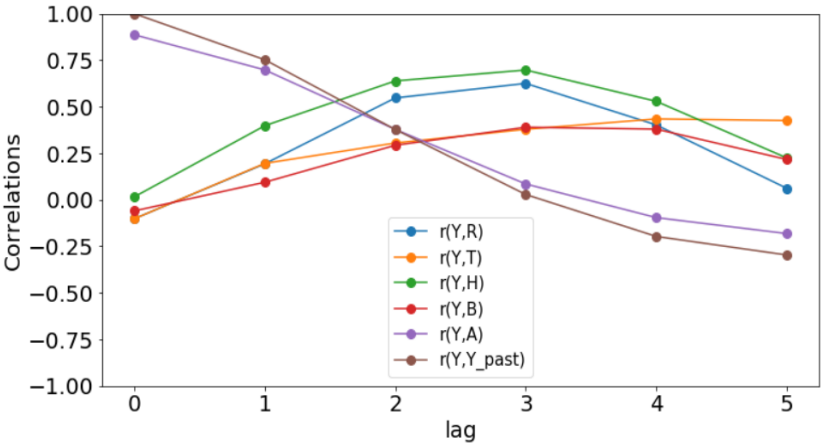

Figure 9. Correlation plots for Fatick.

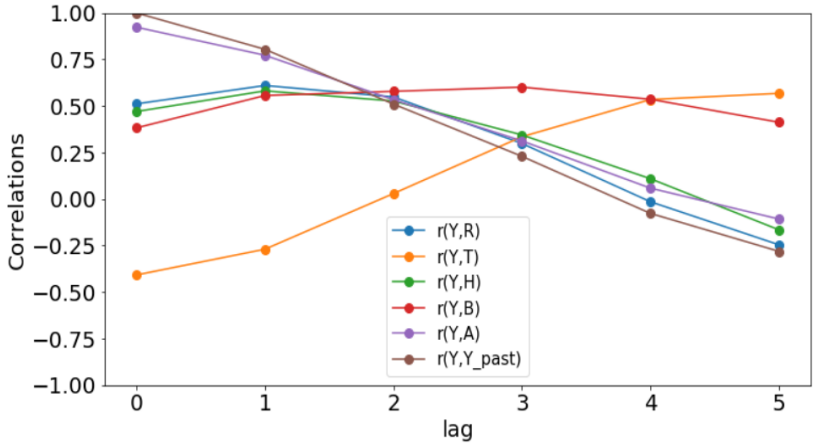

Figure 10. Correlation plots for Kedougou.

3.2. Model Performance Metrics

The output of the GLM is a probability distribution at each time instant. Accuracy measures such as the root mean square error (RMSE) and the mean absolute scaled error (MASE) defined in [22], and the mean absolute relative error (MARE) defined in [23], quantify the discrepancy between the mean μ_t of the distribution and the observed incidence $Y_o(t)$. We also introduce other statistical measures such as the scatter index (SI) and the reliability analysis (RA) defined in [24]. We conduct experiments for 108 months with a train and a test period.

- Root mean square error (RMSE):

$$\text{RMSE} = \sqrt{\frac{\sum_{t=t_i}^{t_c} (Y_o(t) - Y(t))^2}{t_c - t_i + 1}}. \tag{7}$$

- Mean absolute error (MAE):

$$\text{MAE} = \frac{1}{t_c - t_i + 1} \sum_{t=t_i}^{t_c} |Y_o(t) - Y(t)|. \tag{8}$$

- Mean absolute scaled error (MASE):

$$\text{MASE} = \frac{\text{MAE}}{\frac{1}{t_c - t_i} \sum_{t=t_i}^{t_c - 1} |Y_o(t+1) - Y_o(t)|}. \tag{9}$$

  It consists of the ratio between the MAE and the mean monthly variation of the observed values. A MASE value around 1 or below indicates an excellent accuracy.
- Mean absolute relative error (MARE):

$$\text{MARE} = \frac{1}{t_c - t_i + 1} \sum_{t=t_i}^{t_c} \frac{|Y_o(t) - Y(t)|}{|Y_o(t)|}. \tag{10}$$

- R-squared [2]:

$$R^2_{\text{COR}} = (\hat{COR}[Y_o, \mu])^2. \tag{11}$$

  It is the proportion of variation in the outcome that is explained by the predictor variables. The higher the R-squared, the better the model, in contrast to all the above metrics.

The Scatter index (SI) (also called the normalized root mean squared error (NRMSE)) and the reliability analysis (RA) are defined in [24].

- The SI presents the percentage of RMSE difference with respect to mean observation or it gives the percentage of expected error for the parameter. Lower values of the SI are an indication of better model performance.

$$\text{SI} = \frac{\sqrt{(1/n) \sum_{t=t_a}^{t_b} ((Y_t - \bar{Y}) - (Y_o(t) - \bar{Y_o}))^2}}{\sqrt{(1/n) \sum_{t=t_a}^{t_b} Y_o(t)}}. \tag{12}$$

- The reliability analysis (RA) is a statistical method for measuring the overall consistency of a model by determining if this suggested model achieves a permissible level of performance.

$$\text{RA} = \left(\frac{100\%}{t_b - t_a + 1}\right) \sum_{t=t_a}^{t_b} k(t), \tag{13}$$

where the ks are determined through two steps. First, the relative average error (RAE) is defined as a vector whose tth component is

$$\mathrm{RAE(t)} = |\frac{Y_o(t) - Y(t)}{Y_o(t)}|. \tag{14}$$

Next, if $\mathrm{RAE(t)} \leq \Delta$, then $k(t) = 1$, otherwise $k(t) = 0$, where Δ is a threshold value that is 0.2 (20%) based on Chinese standards.

3.3. Model Selection and Result Comparison

3.3.1. Model Selection by Using the Vuong Test

In order to assess the adequacy of the distribution, we apply the Vuong statistical test as in [25,26]. The Vuong test is defined as follows:

$$V = \frac{\sqrt{n}\frac{1}{n}\sum_t^n m_t}{\sqrt{\frac{1}{n}\sum_i^n (m_t - \bar{m})^2}}, \tag{15}$$

where $m_t = \log(f_1(Y_t|X_t)/f_2(Y_t|X_t))$ in which $f_1(Y_t|X_t)$ is the first probability mass function and $f_2(Y_t|X_t)$ is the second probability mass function. If $V > 1.96$, then the first model is preferred. If $V < -1.96$, then the second one is preferred. If $-1.96 < V < 1.96$, none of the models are preferred. In our case, we let f_1 be the Poisson distribution and f_2 the NB distribution. This statistical test permits to choose the most adequate between the two regression models in order to fit the data.

Figures 11–13 (reproduce with Vuong_test_2022_07_23.py) present the dependence of the Vuong test value (V) with respect to α (dispersion parameter). The figures show that the α computed by OLS is usually in the window where the Poisson model is preferable to the NB model. An exception is observed in Kedougou where none is preferred with log and sqrt. We can conclude that the use of the GLM with Poisson distribution is justified.

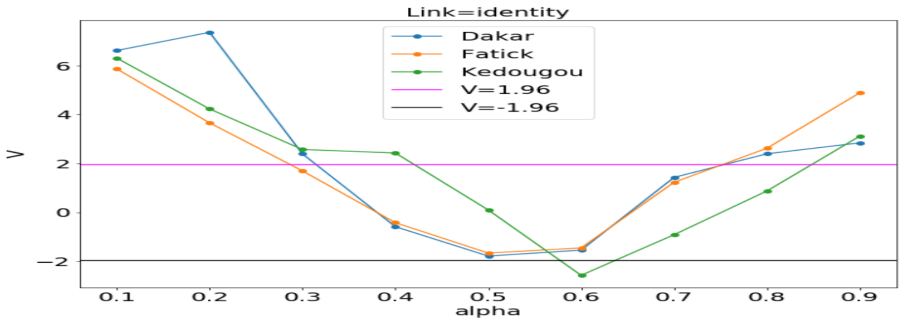

Figure 11. V vs. α between Poisson and NB distributions. Estimating α in each region by the ordinary least squares (OLS) method gives 0.109, 0.222 and 0.282, respectively, in Dakar, Fatick, and Kedougou.

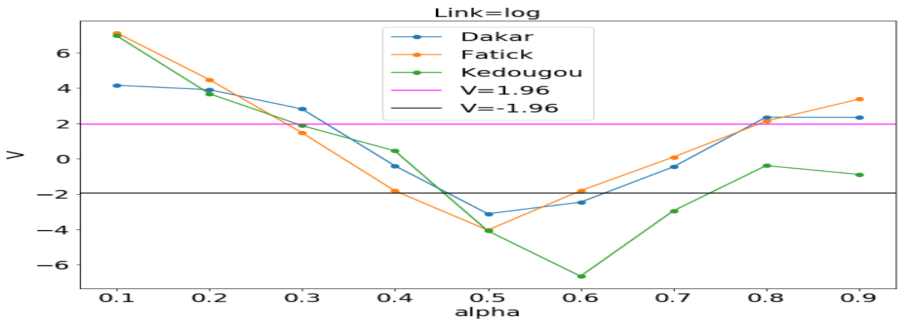

Figure 12. V vs. α between Poisson and NB distributions. Estimating α in each region by the ordinary least squares (OLS) method gives 0.142, 0.202 and 0.412, respectively, in Dakar, Fatick, and Kedougou.

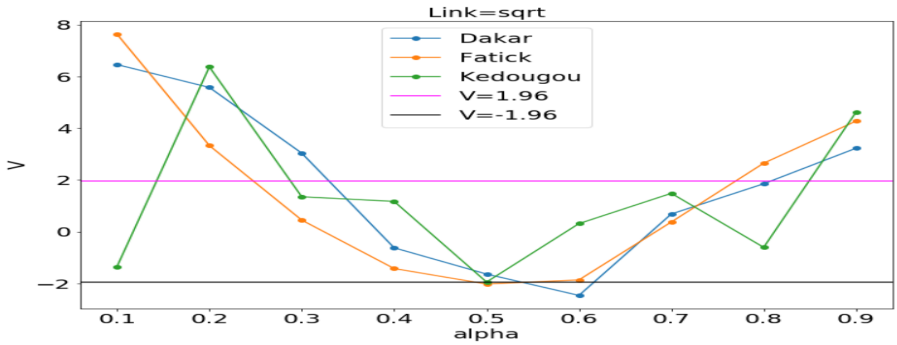

Figure 13. V vs. α between Poisson and NB distributions. Estimating α in each region by the ordinary least squares (OLS) method gives 0.119, 0.214 and 0.318, respectively, in Dakar, Fatick, and Kedougou.

3.3.2. Results Comparison by Using Metrics

In this part, we made a comparative analysis in Table 2 (reproduce with Comparative_results_with_all_models_2022_07_23.py) between the three models that are Gaussian (identity, log), Poisson (identity, log, sqrt), and negative binomial (identity, log, sqrt). Experiments are carried out with Algorithm 1 without saturation. We include the minimum values of the predicted mean obtained after fitting the model. A negative sign indicates that the model is not adequate for our count data. All these metric values permit us to validate and to compare the performance of the models.

Table 2. Results of accuracy measures with Algorithm 1 and the explanatory variables $Y_o(t-h)$, $R(t-\ell_R)$, $T(t-\ell_T)$, $H(t-\ell_H)$, and 1 for the intercept. We have considered $t_s = 0$, $t_i = 5$, $t_c = 84$ and $t_e = 108$ and $h = 1$. Refer to Table 1 for ℓ_R, ℓ_T and ℓ_H values. Train/test values are reported and $\min_{t\in\{t_i,\ldots,t_c\}} \mu(t)$. We denote by G: Gaussian, P: Poisson, id: identity, Dk: Dakar, Ft: Fatick, and Kd: Kedougou.

	Model	**Link**	**RMSE**	**MASE**	**MARE**	R^2_{COR}	**min**	**RA**
Dk	G	id	2197.29/2384.26	0.52/1.01	0.68/1.54	0.84/0.79	−517.03	28.75/20
		log	2466.29/2352.81	0.6/1.38	0.85/2.66	0.79/0.78	511.81	22.5/4
	P	id	2245.27/2689.74	0.52/1.02	0.54/1.28	0.83/0.78	282.6	32.5/16
		log	2523.01/2297.45	0.57/1.23	0.65/2.05	0.79/0.77	341.22	27.5/4
		sqrt	2354.97/2303.14	0.54/1.08	0.62/1.77	0.81/0.8	279.49	30/12
	NB	id	2558.87/3555.94	0.58/1.28	0.5/1.24	0.82/0.76	460.63	27.5/16
		log	3736.91/2424.02	0.74/1.1	0.61/1.89	0.73/0.79	471.42	28.75/8
		sqrt	3409.99/3234.53	0.73/1.23	0.55/1.56	0.79/0.79	482.42	28.75/16
Ft	G	id	768.52/578.88	0.83/1.33	0.66/0.59	0.67/0.75	87.78	27.5/24
		log	721.66/484.07	0.81/1.31	0.73/0.7	0.71/0.83	83.29	25/16
	P	id	772.32/543.56	0.82/1.26	0.65/0.6	0.67/0.72	99.12	31.25/24
		log	741.92/491.42	0.83/1.37	0.85/0.88	0.69/0.78	215.44	25/8
		sqrt	763.78/526.94	0.84/1.32	0.79/0.74	0.68/0.73	215.13	26.25/20
	NB	id	839.53/527.32	0.87/1.22	0.61/0.59	0.64/0.67	69.22	30/20
		log	861.98/466.03	0.91/1.21	0.84/0.79	0.63/0.7	268.19	21.25/20
		sqrt	942.27/504.91	0.98/1.13	0.75/0.59	0.62/0.64	180.39	22.25/28
Kd	G	id	1230.61/2467.27	1.01/0.92	1.19/0.63	0.62/0.62	29.85	10/16
		log	1409.28/2720.94	1.27/1.01	2.18/0.73	0.51/0.56	872.92	13.75/20
	P	id	1241.89/2431.91	0.99/0.87	0.9/0.47	0.61/0.63	126.76	16.25/24
		log	1488.67/3009.49	1.15/1.1	1.47/0.52	0.46/0.59	510.81	21.25/16
		sqrt	1352.9/2523.28	1.03/0.89	1.07/0.42	0.56/0.62	324.21	18.75/12
	NB	id	1287.33/2346.07	1.07/0.83	0.9/0.44	0.61/0.65	−52.81	16.25/28
		log	2160.53/7024.86	1.42/2.47	1.08/0.73	0.4/0.58	275.72	16.25/8
		sqrt	1567.23/3037.79	1.19/1.14	0.91/0.44	0.54/0.63	132.16	20/16

Results in Table 2 show that the most of the MASEs and MAREs in the train and test windows are around 1 in the three regions. Then, we also have the R^2_{COR}s indicator whose values are high (>50 %) indicating a good contribution of the explanatory variables in the forecasts. Additionally, predictions provided by Poisson (identity) are globally more reliable in in the three regions when compared to other models based on RA values.

We can conclude that the Poisson (identity) model can be nicely used to fit our data for parameter estimation and to make the forecasts with these found parameters in the three regions. All the additional studies will base on this model.

3.4. Forecasts Results by Various Sets of Explanatory Variables

Using Algorithm 1 with Poisson for f, identity for g, $t_s = 0$, $t_i = 5$, $t_c = 84$, $t_e = 108$, and $h = 1$, we validate the model with datasets from Dakar, Fatick, and Kedougou.

3.4.1. Forecasts Results Using History of Malaria Incidence Only

For transmissible diseases, the incidence at a given time $t - \delta$ is very important to predict the expected incidence at a given time t. That reason leads us to first consider a set of explanatory variables composed of two variables that are the history of malaria incidence ($Y_o(t - \delta)$) and the intercept vector ($I(t)$). We defined this set as follows:

$$\text{start\_set} \leftarrow \{Y_o(t - \delta), I(t)\}. \tag{16}$$

Thus, our linear predictor becomes like a simple Markov model with the Gaussian distribution that is called first order autoregressive AR(1) [3] and defined by

$$\mu_{t|t-\delta} = g^{-1}(\beta_1 I(t) + \beta_2 Y_o(t - \delta)). \tag{17}$$

The estimates returned by the model in Figures 14–16 are very accurate because of the low standard errors and the tight 95% confidence interval and statistically significant due to the p-values that are less than 0.05. That is why we decide not to show the confidence intervals of the predictions as they do not clearly appear. In Figures 14–16, (reproduce with Addition_study_2022_07_23.py) we present the forecast results (noted by A) and the curves of $\beta_j X_j$ (noted by B). These results are obtained from Equation (17) with Poisson (identity) model when we take $h = 1$. The figures noted by B permit to show which variable is highly ($\beta_j X_j(t) \gg 0$) or weakly ($\beta_j X_j(t) \sim 0$) used during the forecasting process. In all three regions, we observe that the sole use of the history of malaria incidence at time $t - 1$ gives good results. That can be explained by the fact that it is highly used based on figures (B) compared to the intercept whose values are close to 0. This situation is biologically true because with the transmissible diseases like malaria the history of cases is very important to estimate the future cases.

In addition, in Table 3, (reproduce with Various_forecasting_horizon_2022_07_23.py) we remark that the model gives less accurate predictions when the values of h increases corresponding a weak use of the history of malaria incidence ($Y_o(t - h), h > 1$). We also have collected, in Table 3, the result of SI in the training and testing periods with various value of h. These values of SI reveal that the model is the most accurate when $h = 1$ because they are the lowest in all the three regions.

The conclusion of these results is that malaria in the past was a very good explanatory variable when $h = 1$.

```
                 Results: Generalized linear model
================================================================
Model:              GLM              AIC:            150148.7739
Link Function:      identity         BIC:            149056.8686
Dependent Variable: y                Log-Likelihood: -75072.
Date:               2023-06-05 16:36 LL-Null:        -2.0802e+05
No. Observations:   79               Deviance:       1.4939e+05
Df Model:           1                Pearson chi2:   1.62e+05
Df Residuals:       77               Scale:          1.0000
Method:             IRLS
----------------------------------------------------------------
            Coef.    Std.Err.      z      P>|z|     [0.025    0.975]
----------------------------------------------------------------
x1          0.8854     0.0020  443.8495  0.0000     0.8815     0.8893
const     518.2042     5.7730   89.7637  0.0000   506.8894   529.5191
================================================================
```

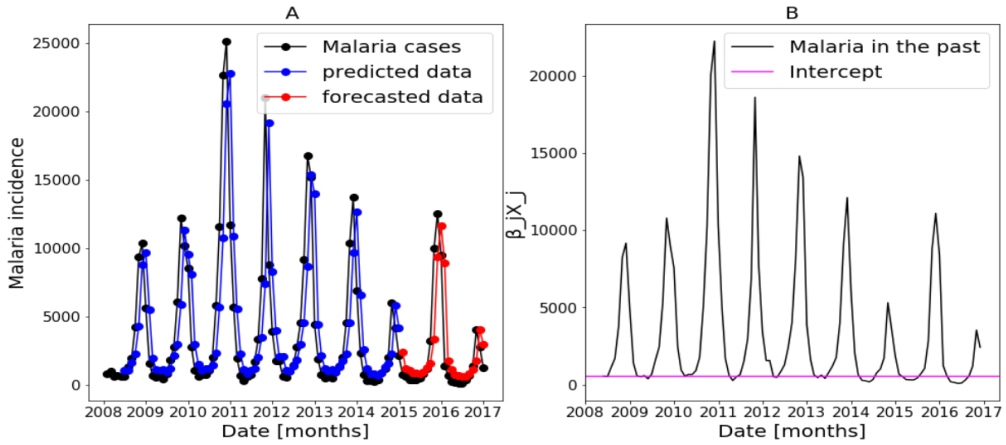

Figure 14. Statistical (in **top**) and forecasting (in **bottom**) results in Dakar. Malaria incidence means the falciparum malaria incidence count per month. The train/test accuracy measures are RMSE: 3886.43 /2367.55, MASE: 0.95/1.06, MARE: 0.85/1.59, and R^2_{COR}: 0.51/0.54. We present the forecast results (noted by **A**) and the curves of $\beta_j X_j$ (noted by **B**).

```
                 Results: Generalized linear model
================================================================
Model:              GLM              AIC:            34841.7973
Link Function:      identity         BIC:            33843.7791
Dependent Variable: y                Log-Likelihood: -17419.
Date:               2023-06-05 17:01 LL-Null:        -48787.
No. Observations:   79               Deviance:       34180.
Df Model:           1                Pearson chi2:   3.74e+04
Df Residuals:       77               Scale:          1.0000
Method:             IRLS
----------------------------------------------------------------
            Coef.    Std.Err.      z      P>|z|     [0.025    0.975]
----------------------------------------------------------------
x1          0.8912     0.0040  224.6363  0.0000     0.8834     0.8989
const     143.5645     3.2865   43.6830  0.0000   137.1231   150.0059
================================================================
```

Figure 15. *Cont.*

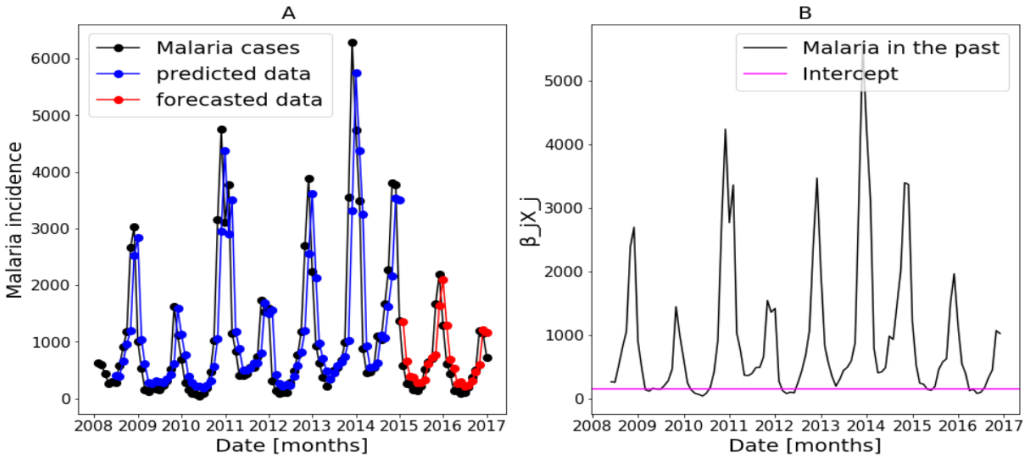

Figure 15. Statistical (in **top**) and forecasting (in **bottom**) results in Fatick. Malaria incidence means the falciparum malaria incidence count per month. The train/test accuracy measures are RMSE: 916.35/408.29, MASE: 0.96/1.11, MARE: 0.76/0.75, and R^2_{COR}: 0.55/0.5. We present the forecast results (noted by **A**) and the curves of $\beta_j X_j$ (noted by **B**).

Results: Generalized linear model

Model:	GLM	AIC:	56734.4846
Link Function:	identity	BIC:	55725.4986
Dependent Variable:	y	Log-Likelihood:	-28365.
Date:	2023-06-05 17:03	LL-Null:	-79005.
No. Observations:	79	Deviance:	56062.
Df Model:	1	Pearson chi2:	7.78e+04
Df Residuals:	77	Scale:	1.0000
Method:	IRLS		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
x1	0.8338	0.0032	259.6079	0.0000	0.8275	0.8401
const	293.5650	3.6741	79.9012	0.0000	286.3639	300.7661

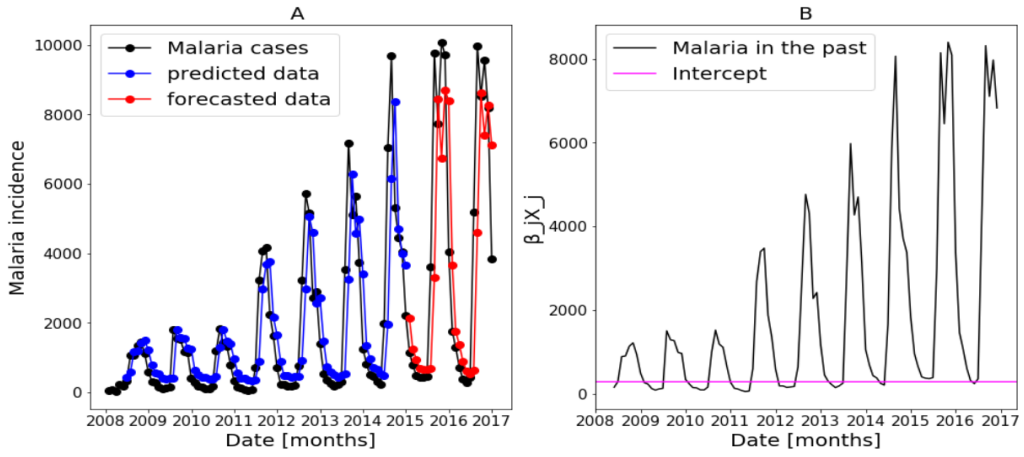

Figure 16. Statistical (in **top**) and forecasting (in **bottom**) results in Kedougou. Malaria incidence means the falciparum malaria incidence count per month. The train/test accuracy measures are RMSE: 1250.53/2523.74, MASE: 1.02/0.93, MARE: 1.06/0.61, and R^2_{COR}: 0.61/0.59. We present the forecast results (noted by **A**) and the curves of $\beta_j X_j$ (noted by **B**).

Table 3. Accuracy measures in the three regions: Algorithm 1 with Poisson for f and identity for g where the set of explanatory variables is Equartion (16) and $h = 1, 2, 3$. Train/test values are reported.

Regions	h	RMSE	MASE	MARE	R^2_{COR}	SI
Dakar	1	3886.43/2367.55	0.95/1.06	0.85/1.59	0.51/0.54	33.93/49.35
Dakar	2	5265.41/3565.7	1.47/2.15	2.4/5.97	0.08/0.06	41.5/69.91
Dakar	3	5399.89/4075.83	1.55/2.79	3.15/9.38	0.01/0.01	60.31/56.57
Fatick	1	916.35/408.29	0.96/1.11	0.76/0.75	0.55/0.5	22.22/18.83
Fatick	2	1250.88/741.17	1.47/2.26	1.99/2.21	0.14/0.03	24.92/24.75
Fatick	3	1339.06/807.71	1.67/2.71	2.85/3.32	0.0/0.01	25.23/25.19
Kedougou	1	1250.53/2523.74	1.02/0.93	1.06/0.61	0.61/0.59	30.69/36.47
Kedougou	2	1790.88/3770.72	1.58/1.59	2.67/1.23	0.19/0.18	41.54/41.69
Kedougou	3	1972.37/4427.78	1.87/1.84	3.65/1.43	0.02/0.01	42.41/47.11

3.4.2. Forecasts Results by Using all Explanatory Variables

In this part, we use the whole explanatory variables and the results from Algorithm 1 and Equation (4) are presented in Figures 17–19 (reproduce with Saturated_and_non_saturated_rainfall_2022_07_23.py). The confidence intervals of the predictions are not shown based on the same observations made in the previous Section 3.4.1.

Figure 17B of Dakar data shows with some peaks in the rainfall distribution. While, in Figure 18B about Fatick data, the rainfall variable is weakly used and this variable is negatively used in Figure 19 of Kedougou data. The peaks observed in Dakar and the situation in Kedougou leads us to develop the method of saturation in Section 2.3.3. Then, in Figure 17B of Dakar data, it is also shown that the humidity is weakly used to mean that this variable has a little contribution in the forecasts. Contrary to Figure 18B about Fatick data, the humidity is highly used, so it has a big participation in the forecasts. As for Kedougou, in Figure 19B, this variable (humidity) is moderately used. As for temperature, its contribution changes depending on the region. For example, in Dakar and Fatick, this variable is negatively used as shown in Figures 17 and 18. However, in Kedougou, it is positive and highly used in Figure 19.

We can summarize here that each explanatory variable does is not used the same in the three regions. Thus, these analyses and observations lead to the development of the methods in Sections 3.5 and 3.6.

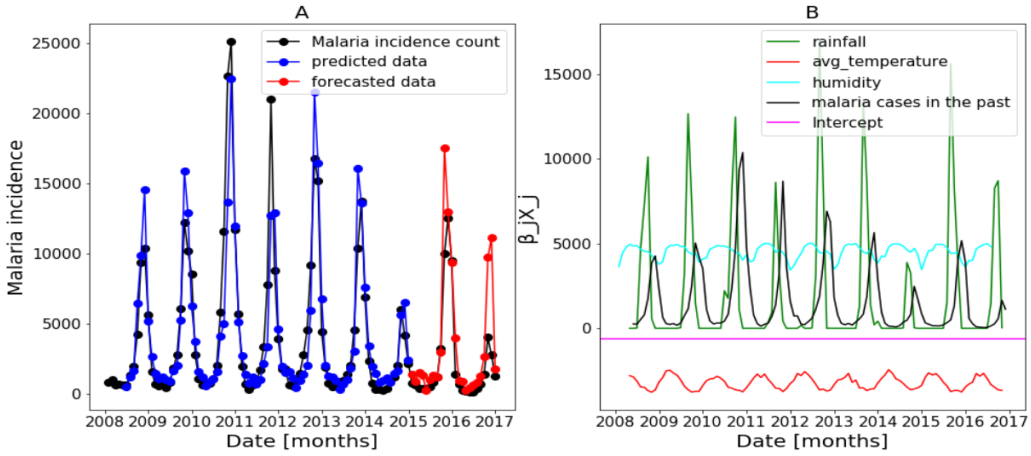

Figure 17. Forecasting results in Dakar, no saturation applied. Malaria incidence means the falciparum malaria incidence count per month. We present the forecast results (noted by **A**) and the curves of $\beta_j X_j$ (noted by **B**).

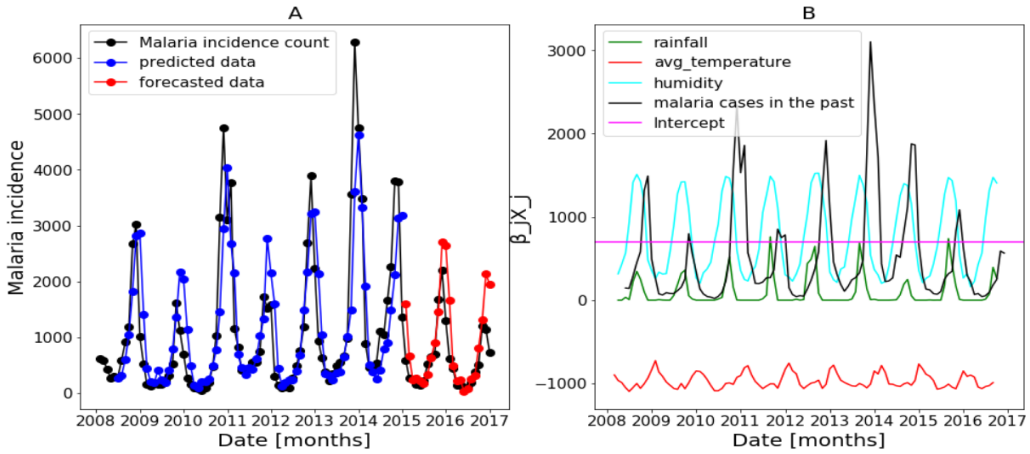

Figure 18. Forecasting results in Fatick: no saturation applied. Malaria incidence means the falciparum malaria incidence count per month. We present the forecast results (noted by **A**) and the curves of $\beta_j X_j$ (noted by **B**).

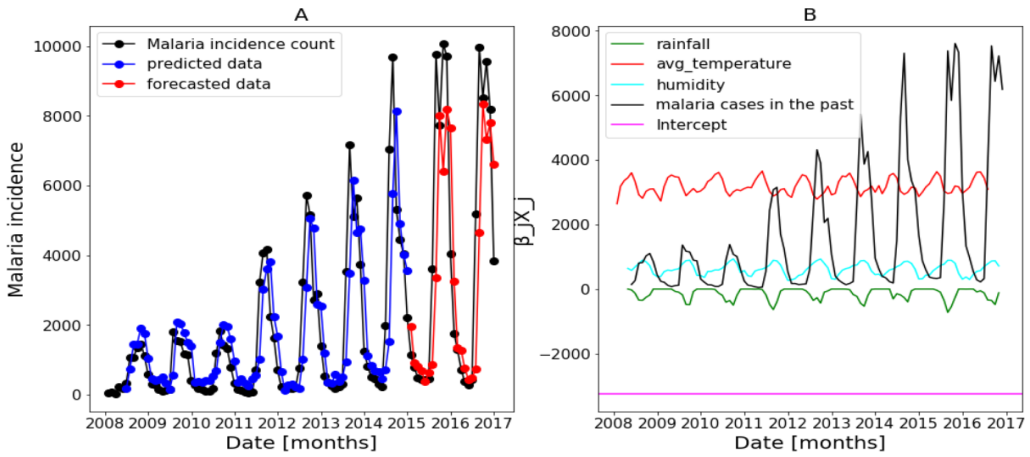

Figure 19. Forecasting results in Kedougou: no saturation applied. Malaria incidence means the falciparum malaria incidence count per month. We present the forecast results (noted by **A**) and the curves of $\beta_j X_j$ (noted by **B**).

3.5. Addition Study

In this section, we experiment with different combinations of sets from Equation (16) and we define them as follows

$$\{\text{start\_set} \cup X_j, j = 3, \ldots, k\}.$$

The purpose is to investigate the influence of each additional variable compared to the observation in Section 3.4.1.

The Table 4 (reproduce with Addition_study_2022_07_23.py) reveals that, in Kedougou, forecast results are improved by adding to start_set explanatory variables such as temperature or humidity. In contrast, the addition of the rainfall (R) gives some less good results in the sense of the RMSE, MASE and MARE, in Dakar and Fatick, even if the higher values of R^2_{COR} are observed there.

3.6. Ablation Study

We now present ablation studies where we start with all the available explanatory variables and investigate the effect of removing any one of them. The sets of explanatory variables after ablation are thus

$$\{X_1, X_2, \ldots, X_k\} \setminus X_j, j = 2, \ldots, k.$$

Table 4. Results of the addition study in the three regions: Algorithm 1 with Poisson for f and identity for g, $t_s = 0$, $t_i = 5$, $t_c = 84$, $t_e = 108$ and $h = 1$. We denote by w: situation in the test period with the added variable and wo: situation in the test period without the indicated variable. Ratios of w/wo are reported. Note that, for the metrics such as RMSE, MASE and MARE, a ratio lower than 1 implies an improvement of the model when the variable is added as an explanatory variable. A ratio of R^2_{COR} higher than 1 indicates that adding the variable improves the forecasts.

Regions	Variable	RMSE w/wo	MASE w/wo	MARE w/wo	R^2_{COR} w/wo
Dakar	Rainfall	1.18	1.02	0.99	1.43
	Temperature	0.99	1.05	1.15	1.04
	Humidity	0.94	1.04	1.06	1.11
Fatick	Rainfall	1.24	1.22	1.04	1.34
	Temperature	1	1.06	1.08	1.1
	Humidity	1.18	1.05	0.77	1.34
Kedougou	Rainfall	0.99	0.97	0.92	1.02
	Temperature	0.96	0.95	0.88	1.05
	Humidity	0.98	0.93	0.7	1.03

By doing that, we can know which variable is more responsible of the over-forecasting (or under-forecasting) observed.

The results in Table 5 (reproduce with Ablation_study_2022_07_23.py) show a low accuracy in terms of RMSE, MASE, and MARE in the whole three regions data sets when the malaria incidence in the past was deleted. This situation was expected when we refer to the forecast results in Section 3.4.1. It is also shown there that the deletion of one of the others variables such as rainfall, temperature and humidity generally reduces a bit the performance of the models in the sense of RMSE, MASE, and MARE.

Table 5. Results of the ablation study in the three regions: Algorithm 1 with Poisson for f and identity for g, $t_s = 0$, $t_i = 5$, $t_c = 84$, $t_e = 108$ and $h = 1$. We denote by wo: situation in the test period without the indicated variable and w: situation in the test period with all the explanatory variables. Ratios of wo/w are reported. Note that, for the metrics such as RMSE, MASE and MARE, a ratio lower than 1 implies that the result is good without the indicated variable. A ratio of R^2_{COR} higher than 1 indicates adding the variable improves the forecasts. We denote by Mcp: Malaria cases in the past.

	Variable	RMSE wo/w	MASE wo/w	MARE wo/w	R^2_{COR} wo/w
Dakar	Mcp	1.5	1.74	2.04	0.82
	Rainfall	0.83	1.06	1.27	0.75
	Temperature	1.05	1.07	1.11	0.98
	Humidity	1.04	1.05	1.25	0.98
Fatick	Mcp	1.48	1.63	1.58	1.11
	Rainfall	0.94	0.95	1.06	0.93
	Temperature	0.96	0.98	0.96	1.01
	Humidity	0.92	1.05	1.3	0.98
Kedougou	Mcp	1.5	1.64	1.25	1.21
	Rainfall	1	0.99	0.92	1.01
	Temperature	1.02	1.02	0.92	0.97
	Humidity	1.01	1.03	1.14	0.98

3.7. Forecasts Results Using Saturation

In this section, we made a main modification to the rainfall variable. We call this method by saturation and the procedure is entirely detailed in Section 2.3.3 and the al-

gorithm therein (Algorithm 2). Statistical results of this method are presented in Table 6. They permit us to distinguish the performance given by this novelty compared to Table 2. The confidence interval of the predictions in Figures 20 and 21 (reproduce with Saturated_and_non_saturated_rainfall_2022_07_23.py) are not shown based on the same observations made in the previous Section 3.4.1.

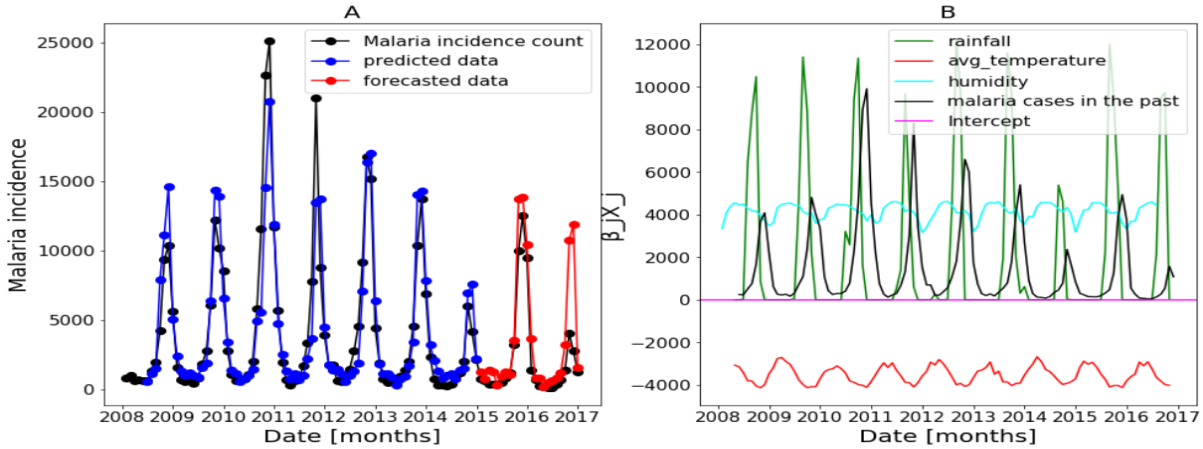

Figure 20. Forecasting results of the saturation in Dakar. Malaria incidence means the falciparum malaria incidence count per month. We present the forecast results (noted by **A**) and the curves of $\beta_j X_j$ (noted by **B**).

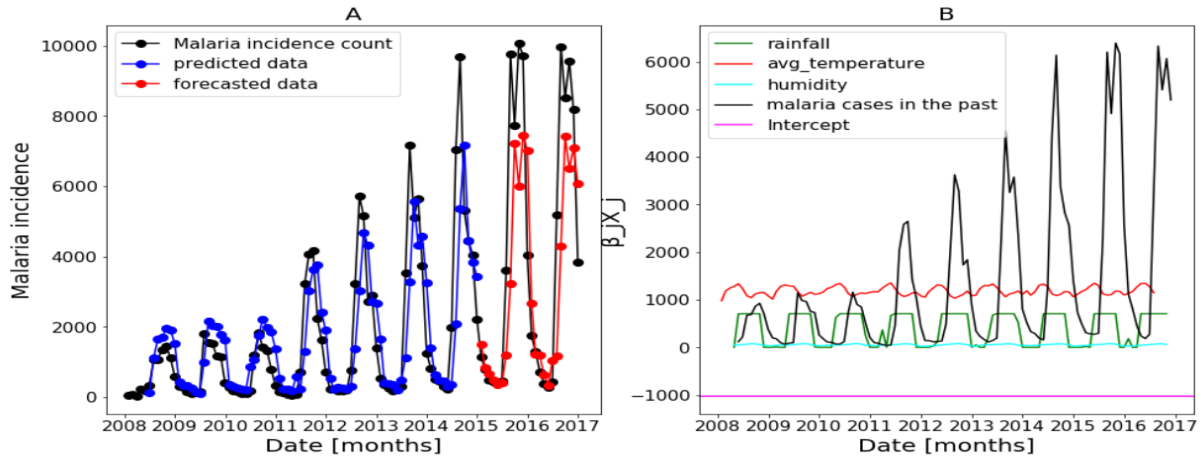

Figure 21. Forecasting results of the saturation in Kedougou. Malaria incidence means the falciparum malaria incidence count per month. We present the forecast results (noted by **A**) and the curves of $\beta_j X_j$ (noted by **B**).

As a result, an improvement of the forecasts is observed based on Table 6. For Dakar, we observed an improvement of the results after applying the saturation according to all the metrics. It is interesting to see that, in Dakar, the introduction of the saturated rainfall has reduced the over-estimation occurring at the end of 2015; compare Figures 17 and 20. Then, in Fatick, all the values are 1, meaning that the saturation method does not improve the results. That can be explained by the fact that the rainfall is less used in this region, according to Figure 18B. As for Kedougou, we have a ratio of MARE smaller than its value before applying the method (0.87) and a ratio of R^2_{COR} higher than the value before applying the method (1.04). There is no improvement in terms of ration of the RMSE and MASE. That may be explained by the fact that there was not a peak as that is illustrated in Figure 19 or the forecasts have been already good. Another favorable effect of the saturation is that,

in Kedougou, the non-saturated rainfall is negatively used (Figure 19B) while the saturated rainfall is positively used (Figure 21B).

Table 6. Comparison results after and before the saturation: Algorithm 2 with Poisson for f and identity for g, $t_s = 0$, $t_i = 5$, $t_c = 84$, $t_e = 108$ and $h = 1$. We denote by w: the metric with saturation and wo: the metric without saturation, in the test period. Ratios of w/wo are reported. Note that, for the metrics such as RMSE, MASE, and MARE, a ratio lower than 1 implies an improvement of the model to make good forecasts with saturation. A ratio of R^2_{COR} higher than 1 indicates that applying the saturation improves the forecasts.

	RMSE w/wo	MASE w/wo	MARE w/wo	R^2_{COR} w/wo
Dakar	0.95	0.97	0.96	1.01
Fatick	1	1	1	1
Kedougou	1.01	1.02	0.87	1.04

4. Conclusions

For three endemic regions of Senegal, we have investigated the accuracy, in the sense of the metrics such as RMSE, MASE and MARE, of the falciparum malaria incidence count per month forecasts obtained with GLM's by using meteorological data and history of falciparum malaria incidence count per month as explanatory variables. Using the Vuong test, we have compared the adequacy of Poisson-based and NB-based GLM's. And the Poisson with identity as a GLM link function is in practice the more adequate regression model to make forecasts of malaria incidence based on meteorological factors. We have observed that the choice of the GLM's link function and the use of adequate `lags` in the explanatory variables may have a considerable impact on the forecast accuracy. We also have observed that the application of saturation in the rainfall increases the quality of the forecasts in Dakar and Kedougou.

Ablation study shows that removing the history of malaria cases from the explanatory variables has a strong adverse effect on the forecast accuracy.

This study is led with a monthly malaria incidence count due to the unavailability of the daily malaria incidence reports that could have helped us to understand better the influence of the climatic data. This study is a step towards providing the authorities with decision-making tools for the optimal dispatch of resources.

This proposed GLM gives some overestimations in the forecasts (end of 2016 in Dakar in Figure 17 and Fatick in Figure 18). These peaks can be caused by an inadequacy of this model due to its linear character even if a non-linearity is then applied for rainfall. This may also be due to a variable ratio of infected people being sufficiently ill to go to the hospital and be confirmed, thus causing a bias in the collected data. Another and important explanation to the over-forecasting (or under-forecasting) observed can be the unavailability of some explanatory variables: the distance to water bodies, the normalized differenced vegetation index (NDVI), the night and day LST (land surface temperature), the ownership and use of insecticide treated nets (ITNs) and the intermittent preventive treatment distributed for pregnant women (IPTp) that are considered to fit malaria incidence [13]. Providing the enhanced vegetation index (EVI), and actual evapotranspiration (ETa) could help to improve the malaria incidence fitting model [27]. In our available data, there was the ITNs distributed because it constitutes the main factor fighting against malaria according to [16] but its distribution was not very regular (see Figures 5–7). Needless to say, the ownership and use of ITNs would be a more suitable explanatory variable.

Author Contributions: Conceptualization, O.D., P.-A.A. and M.D.; methodology, O.D. and P.-A.A.; software, O.D.; validation, O.D., P.-A.A. and M.D.; formal analysis, O.D.; writing—original draft preparation, O.D.; writing—review and editing, O.D.; visualization, O.D.; supervision, P.-A.A. and

M.D.; funding acquisition, O.D. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by ARES-CCD and UCLouvain's "Conseil de l'action internationale" through PhD scholarships.

Data Availability Statement: Source codes can be found at https://www.dropbox.com/s/7vdxopgeshlxhrc/Python%20codes%20of%20malaria%20model.zip?dl=0 (accessed on 3 July 2023).

Acknowledgments: This work benefited from discussions with colleagues: Loïc Van Hoorebeeck and Hazan Daglayan Sevim. We are grateful to the PNLP for providing us with historical data, such as the monthly number of malaria cases, bed-nets, and therapy distributed from 2008 to 2016 and to meteoblue for providing us with meteorological data, such as temperature, humidity and rainfall from 2008 to 2016. This work has received funding from a fellowship awarded by UCLouvain's Conseil de l'action internationale.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Appendix A. Method for Parameter Estimation

Appendix A.1. Poisson Log-Likelihood

1. The likelihood function is defined as follows

$$\begin{aligned}\ell(\beta) &= \prod_{t=1}^{n} f(Y_t|X_t;\beta)\\ &= \prod_{t=1}^{n} \exp[Y_t \log(\mu_t) - \mu_t - \log(Y_t!)],\end{aligned}$$

   where the right-hand side depends on β through Equation (2).
2. The Log-likelihood is defined as follows

$$\log \ell(\beta) = \log \prod_{t=1}^{n} \exp[Y_t \log(\mu_t) - \mu_t - \log(Y_t!)].$$

$$\log \ell(\beta, Y, X) = \sum_{t=1}^{n} \{Y_t \log(\mu_t) - \mu_t - \log(Y_t!)\} \tag{A1}$$

   where μ_t depends on X and β and is defined as follows Equation (2). Observe that the term $\log(Y_t!)$ does not depend on the parameters to be estimated.
3. The first derivative of the log-likelihood is named the gradient. If $g = \log$, we have $\log(\mu_t) = \eta_t = X_t^T\beta$. So this function is defined as follows

$$\begin{aligned}\log \ell(\beta) &= \sum_{t=1}^{n} \{Y_t \log(\mu_t) - \mu_t - \log(Y_t!)\}\\ &= \sum_{t=1}^{n} \{Y_t\eta_t - e^{\eta_t} - \log(Y_t!)\}.\end{aligned}$$

   Then, the gradient at state j is defined by

$$\begin{aligned}S(\beta_j) &= \frac{\partial \log \ell(\beta_j)}{\partial \beta_j}\\ &= \sum_{t=1}^{n} \frac{\partial \{Y_t\eta_{tj} - e^{\eta_{tj}} - \log(Y_t!)\}}{\partial \eta_{tj}} \frac{\partial \eta_{tj}}{\partial \beta_j}\\ &= \sum_{t=1}^{n} \{Y_t - \mu_t\} X_{tj}\end{aligned}$$

Then, if g=identity, we have $\mu_t = \eta_t = X_t^T\beta$. So

$$\log \ell(\beta) = \sum_{t=1}^{n}\{Y_t \log(\eta_t) - \eta_t - \log(Y_t!)\}.$$

Then, the gradient at state j is defined by

$$\begin{aligned} S(\beta_j) &= \sum_{t=1}^{n} \frac{\partial\{Y_t \log(\eta_{tj}) - \eta_{tj} - \log(Y_t!)\}}{\partial \eta_{tj}} \frac{\partial \eta_{tj}}{\partial \beta_j} \\ &= \sum_{t=1}^{n} \{\frac{Y_t - \mu_t}{\mu_t}\} X_{tj}. \end{aligned}$$

Finally, if $g = \sqrt{}$, we have $\sqrt{\mu_t} = \eta_t = X_t^T\beta$. So

$$\log \ell(\beta) = \sum_{t=1}^{n}\Big\{Y_t \log(\eta_t^2) - \eta_t^2 - \log(Y_t!)\Big\}.$$

Then, the gradient at state j is defined as follows

$$\begin{aligned} S(\beta_j) &= \sum_{t=1}^{n} \frac{\partial\Big\{2Y_t \log(\eta_{tj}) - \eta_{tj}^2 - \log(Y_t!)\Big\}}{\partial \eta_{tj}} \frac{\partial \eta_{tj}}{\partial \beta_j} \\ &= 2\sum_{t=1}^{n} \{\frac{Y_t - \mu_t}{\sqrt{\mu_t}}\} X_{tj}. \end{aligned}$$

4. The second derivative of the log-likelihood function named the Hessian is defined, if $g = \log$, by

$$\begin{aligned} H(\beta) &= \frac{\partial S(\beta)}{\partial \beta} \\ &= \frac{X^T\{Y - \mu\}}{\partial \beta} \\ &= -X^T \begin{bmatrix} \frac{\partial \mu_1}{\partial \eta_1} & 0 & \dots & 0 \\ 0 & \frac{\partial \mu_2}{\partial \eta_2} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \frac{\partial \mu_n}{\partial \eta_n} \end{bmatrix} \frac{\partial \mathbf{J}}{\partial \mathbf{fi}} \\ H(\beta) &= -X^T W X \end{aligned}$$

where $W = \frac{\partial \mu}{\partial \eta}$ is the diagonal matrix with $[\frac{\partial \mu_t}{\partial \eta_t}]_{tt} = [\mu_t]_{tt} = [\exp(\eta_t)]_{tt}$ and called the iterative weights.
Then, if g=identity so

$$\begin{aligned} H(\beta_j) &= \frac{\partial S(\beta_j)}{\partial \beta_j} \\ &= \frac{\partial \left[\sum_{t=1}^{n} \{\frac{Y_t}{\eta_{tj}} - 1\} X_{tj}\right]}{\partial \beta_j} \\ &= \frac{\partial \left[\sum_{t=1}^{n} \{\frac{Y_t}{\eta_{tj}} - 1\} X_{tj}\right]}{\partial \eta_{tj}} \frac{\partial \eta_{tj}}{\partial \beta_j} \\ &= -\sum_{t=1}^{n} \{\frac{Y_t}{\eta_{tj}^2} X_{tj}\} X_{tj} \\ &= -\frac{1}{\beta_j^2} \sum_{t=1}^{n} Y_t. \end{aligned}$$

Finally, if $g = \sqrt{}$ so

$$\begin{aligned} H(\beta_j) &= \frac{\partial S(\beta_j)}{\partial \beta_j} \\ &= 2\frac{\partial \left[\sum_{t=1}^{n} \{\frac{Y_t}{\eta_{tj}} - \eta_{tj}\} X_{tj}\right]}{\partial \beta_j} \\ &= 2\frac{\partial \left[\sum_{t=1}^{n} \{\frac{Y_t}{\eta_{tj}} - \eta_{tj}\} X_{tj}\right]}{\partial \eta_{tj}} \frac{\partial \eta_{tj}}{\partial \beta_j} \\ &= -2\sum_{t=1}^{n} \{\frac{Y_t}{\mu_t} + 1\} X_{tj}^2. \end{aligned}$$

Appendix A.2. NB Log-Likelihood

As in [11,26,28], we obtain the likelihood function

$$\begin{aligned} \ell(Y_t, \mu_t, \alpha) &= \prod_{t=1}^{n} f(Y_t; \mu_t, \alpha) \\ &= \prod_{t=1}^{n} [\frac{\Gamma(Y_t + \frac{1}{\alpha})}{\Gamma(\frac{1}{\alpha})\Gamma(Y_t + 1)} (\frac{1}{1 + \alpha\mu_t})^{\frac{1}{\alpha}} (\frac{\alpha\mu_t}{1 + \alpha\mu_t})^{Y_t}]. \end{aligned}$$

Then,

$$\log \ell(Y_t, \mu_t, \alpha) = \sum_{t=1}^{n} [Y_t \log(\frac{\alpha\mu_t}{1 + \alpha\mu_t}) - \frac{1}{\alpha} \log(1 + \alpha\mu_t) + \log\Gamma(Y_t + \frac{1}{\alpha}) - \log\Gamma(Y_t + 1) - \log\Gamma(\frac{1}{\alpha})]. \quad \text{(A2)}$$

Abusing notation, we omit the term $\log(\Gamma(y_i + 1))$ since it does not depend on the parameters to be estimated. This yields

$$\log \ell = \sum_{t=1}^{n} [Y_t \log(\frac{\alpha\mu_t}{1 + \alpha\mu_t}) - \frac{1}{\alpha} \log(1 + \alpha\mu_t) + \log(\Gamma(Y_t + \frac{1}{\alpha}) - \log(\Gamma(\frac{1}{\alpha}))].$$

According to [2] p. 81 and [28], and if Y_t is an integer, we have

$$L(Y_t) := \sum_{j=0}^{Y_t} \log(j + \frac{1}{\alpha}) = \log(\Gamma(Y_t + \frac{1}{\alpha}) - \log(\Gamma(\frac{1}{\alpha})).$$

We mention it in order to avoid using the gamma function that gives infinity when Y_t is high. Finally, the formula Equation (A2) yields

$$\log \ell(\alpha, \beta; Y, X) = \sum_{t=1}^{n} [Y_t \log(\frac{\alpha\mu_t}{1+\alpha\mu_t}) - \frac{1}{\alpha}\log(1+\alpha\mu_t) + L(Y_t)]. \quad \text{(A3)}$$

where μ_t depends on X and β according to Equation (2).

Appendix A.3. Optimization Algorithm

The βs are obtained with the iterative re-weighted least squares algorithm (IRLS) (Newton-Raphson method) in Algorithm A1 (details are in [3]) p. 202 by solving the following problem

$$\hat{\beta}^{i+1} = \hat{\beta}^i + (H^{-1})^i S(\hat{\beta}^i).$$

Algorithm A1: Newton-Raphson [29]

1. Choose initial parameter estimate $\beta^i = \beta^0$;
2. Calculate score $S(\beta) \mid_{\beta=\beta^i}$;
3. Calculate derivative of the function for which you want to calculate the roots;
4. Walk along first derivative until line (plane) of the derivative crosses zero;
5. Update the betas β^{i+1};
6. Iterate from step 2 to 5 until convergence.

References

1. Putri, R.G.; Jaharuddin.; Bakhtiar, T. SIRS-SI Model of Malaria Disease with Application of Vaccines, Anti-Malarial Drugs, and Spraying. *IOSR J. Math.* **2014**, *10*, 66–72. [CrossRef]
2. Cameron, A.C.; Trivedi, P.K. *Regression Analysis of Count Data-Second Edition*, 2nd ed.; Econometric Society Monographs, Cambridge University Press: Cambridge, UK, 2013. [CrossRef]
3. Lindsey, J.K., Generalized Linear Modelling. In *Applying Generalized Linear Models*; Springer: New York, NY, USA, 1997; pp. 1–26. [CrossRef]
4. McCullagh, P.; Nelder, J.A. *Generalized Linear Models*; Routledge: London, UK, 1983.
5. Lee, S.C. Delta Boosting Implementation of Negative Binomial Regression in Actuarial Pricing. *Risks* **2020**, *8*, 19. [CrossRef]
6. Abiodun, G.J.; Makinde, O.S.; Adeola, A.M.; Njabo, K.Y.; Witbooi, P.J.; Djidjou-Demasse, R.; Botai, J.O. A Dynamical and Zero-Inflated Negative Binomial Regression Modelling of Malaria Incidence in Limpopo Province, South Africa. *Int. J. Environ. Res. Public Health* **2019**, *16*, 2000. [CrossRef]
7. Nakashima, E. Some Methods for Estimation in a Negative Binomial Model. *Ann. Inst. Stat. Math.* **1997**, *49*, 101–115. [CrossRef]
8. Famoye, F. A Multivariate Generalized Poisson Regression Model. *Commun. Stat.-Theory Methods* **2015**, *44*, 497–511. [CrossRef]
9. Makindea, O.S.; Abiodun, G.J.; Ojo, O.T. Modelling of malaria incidence in Akure, Nigeria: Negative binomial approach. *GeoJournal* **2020**, *86*, 1327–1336. [CrossRef]
10. Mabaso, M.L.; Vounatsou, P.; Midzi, S.; Silva, J.D.; Smith, T. Spatio-temporal analysis of the role of climate in inter-annual variation of malaria incidence in Zimbabwe. *Int. J. Health Geogr.* **2006**, *5*, 20. [CrossRef] [PubMed]
11. Asnath, S.M.; Daniel, M.; Alexander, B. Modelling Malaria Incidence in the Limpopo Province, South Africa: Comparison of Classical and Bayesian Methods of Estimation. *Int. J. Environ. Res. Public Health* **2020**, *17*, 5016. [CrossRef]
12. Yirga, A.A.; Melesse, S.F.; Mwambi, H.G.; Ayele, D.G. Negative binomial mixed models for analyzing longitudinal CD4 count data. *Sci. Rep.* **2020**, *10*, 16742. [CrossRef] [PubMed]
13. Giardina, F.; Gosoniu, L.; Konate, L.; Diouf, M.B.; Perry, R.; Gaye, O.; Faye, O.; Vounatsou, P. Estimating the Burden of Malaria in Senegal: Bayesian Zero-Inflated Binomial Geostatistical Modeling of the MIS 2008 Data. *PLOS ONE* **2012**, *7*, e32625. [CrossRef]
14. Nkiruka, O.; Prasad, R.; Clement, O. Prediction of malaria incidence using climate variability and machine learning. *Inform. Med. Unlocked* **2021**, *22*, 100508. [CrossRef]
15. de lutte Contre le Paludisme, P.N. Bulletin Epidemiologique Annuel 2016 du Paludisme au Senegal. 2016. Available online: https://www.dropbox.com/scl/fi/n2w8hoi2ureubud7usc6e/Bulletin-Epidemiologique-Annuel-2016-du-Paludisme-au-Senegal-VF.pdf?rlkey=wryw7t3z4xt4ov3f5edwgaj33&dl=0 (accessed on 3 July 2023)
16. Faye, S.; Cico, A.; Gueye, A.B.; Baruwa, E.; Johns, B.; Ndiop, M.; Alilio, M. Scaling up malaria intervention "packages" in Senegal: using cost effectiveness data for improving allocative efficiency and programmatic decision-making. *Malar. J.* **2018**, *17*, 159 . [CrossRef] [PubMed]
17. Adepoju, P. Les Tests de Diagnostic Rapide Pourraient Omettre Jusqu'à 20% des cas de Paludisme. 2021. Available online: https://www.nature.com/articles/d44148-021-00087-0 (accessed on 14 June 2023).

18. Love, D.E.; Aseidu, L.J.; Adjei, L.E. *A Weather-Based Prediction Model of Malaria Prevalence in Amenfi West District, Ghana*; Hindawi Publishing Corporation Malaria Research and Treatment, London, UK, 2017.
19. Okuneye, K.; Gumel, A.B. Analysis of a temperature- and rainfall-dependent model for malaria transmission dynamics. *Math. Biosci.* **2017**, *287*, 72–92. [CrossRef] [PubMed]
20. Ndiaye, O.; Hesran, J.Y.L.; Etard, J.F.; Diallo, A.; Simondon, F.; Ward, M.N.; Robert, V. Variations climatiques et mortalité attribuée au paludisme dans la zone de Niakhar, Sénégal, de 1984 à 1996. *Cah. Santé* **2001**, 11, 25–33.
21. Maslen, B. How to Deal with Count Data? Technical Report; Stats Central: Mark Wainwright Analytical Centre, UNSW Sydney. 2019. Available online: https://www.analytical.unsw.edu.au/sites/default/files/document_related_files/2019April_Seminar_How%20to%20deal%20with%20count%20data_Maslen_1.pdf (accessed on 3 July 2023).
22. Absil, P.A.; Diao, O.; Diallo, M. Assessment of COVID-19 Hospitalization Forecasts from a Simplified SIR Model. *Lett. Biomath.* **2021**, *8*, 215–228.
23. Jin, C.; Liu, J.A. Applications of Support Vector Machine and Unsupervised Learning for Predicting Maintainability Using Object-Oriented Metrics. In Proceedings of the 2010 Second International Conference on Multimedia and Information Technology, Kaifeng, China, 24–25 April 2010; Volume 1, pp. 24–27. [CrossRef]
24. Saberi-Movahed, F.; Najafzadeh, M.; Mehrpooya, A. Receiving More Accurate Predictions for Longitudinal Dispersion Coefficients in Water Pipelines: Training Group Method of Data Handling Using Extreme Learning Machine Conceptions. *Water Resour. Manag.* **2020**, *34*, 529–561. [CrossRef]
25. Gowsar1, S.N.; Radha, M.; Devi, M.N. A Comparison of Generalized Linear Models for Insect Count Data. *Int. J. Stat. Anal.* **2019**, *9*, 1–9.
26. Hashim, L.H.; Dreeb, N.K.; Hashim, K.H.; Shiker, M.A.K. *An Application Comparison of Two Negative Binomial Models on Rainfall Count Data*; IOP Publishing: Bristol, UK, 2021; Volume 1818, p. 012100. [CrossRef]
27. Midekisa, A.; Senay, G.; Henebry, G.M.; Semuniguse, P.; Wimberly, M.C. Remote sensing-based time series models for malaria early warning in the highlands of Ethiopia. *Malar. J.* **2012**, *165*, 11. [CrossRef] [PubMed]
28. Cruyff, M.J.; van der Heijden, P.G. Point and interval estimation of the population size using a zero-truncated negative binomial regression model. *Biomed. J.* **2008**. [CrossRef]
29. Clemen, L. Poisson IRWLS. 2019. Available online: https://statomics.github.io/SGA2019/assets/poissonIRWLS-implemented.html (accessed on 3 July 2023).