Temporal Analysis and Prediction of Malaria Dynamics Using Meteorological Data in Southeast of Senegal: Suplementary informations

April 30, 2025

Contents

S1 Statistical description of meteorological data

The meteorological data used in this study were drawn from NASA platform from 2018 to 2022. These data were aggregated into weekly counts.

Variables	Mean	St.dev	Q1	Median	Q3	Min	Max
PRECTOTCORR : Weekly rainfall(mm)	4.6	8.3	0.0	0.1	6.5	0.0	56.1
days_with_precipitation : number of rainy days per week	3.3	3.2	0.0	2.5	7.0	0.0	7.0
WS10M_MAX : weekly mean of daily maximum wind speed at 10m (m/s)	5.0	2.0	4.1	5.1	5.9	2.5	8.1
WS10M_MIN :weekly mean of daily minimum wind speed at 10m(m/s)	1.6	0.6	1.1	1.5	2.1	0.6	3.5
WS10M: weekly mean of daily average wind speed at 10m (m/s)	3.2	0.8	2.5	3.3	3.7	1.6	5.2
WD10M : Weekly average of wind direction	168.7	63.9	101.1	187.6	223.9	53.6	268.7
T2M : weekly mean of daily average temperature (°C)	28.2	3.5	25.8	26.8	31.1	22.2	36.1
T2M_MIN :weekly mean of daily minimum (°C)	22.4	3.2	22.8	23.1	24.1	14.2	28.9
T2M_MAX : weekly mean of daily maximum temperature (°C)	34.7	4.9	30.2	33.5	39.9	28.0	44.1
TS : Weekly average of the Earth's surface temperature (°C)	29.0	4.6	25.8	26.8	32.8	22.1	38.9
PS : weekly mean of daily average atmospheric pressure (hPa)	99.4	0.15	99.3	99.5	99.5	98.9	99.7
QV2M : weekly mean of daily average specific humidity (%)	12.2	5.9	6.4	12.4	18.3	2.1	19.4
RH2M : weekly mean of daily average relative humidity (%)	54.7	28.9	23.1	61.2	85.0	10.2	91.3

Table 1: List of meteorological variables with their abbreviations and descriptive statistics. St.dev is the standard deviation, Min is the minimum, Q1 is the first quartile,Q3 is the third quartile dans Max is the maximum.

S2 Correlation analysis of meteorological variables

Correlations between variables play an important role in a descriptive analysis. The correlation matrix presented above was obteint by using Pearson correlation coefficient.

The correlation coefficients are represented by a color palette ranging from blue (positive correlations) to red (negative correlations). Values close to 1 are shown in blue hues, representing strong positive correlations, while values close to -1 are colored red, indicating strong negative correlations. Values close to 0 are represented in gray, signaling an absence of significant linear correlation.

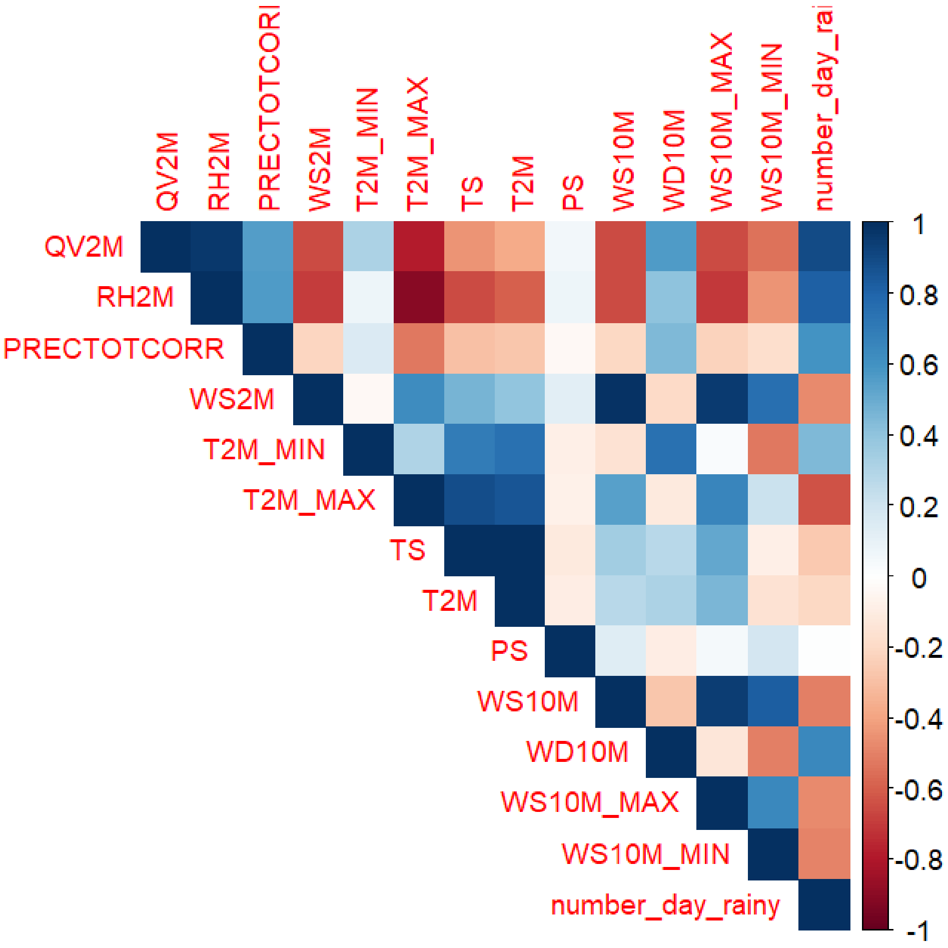

Figure 1: Correlation matrix of the meteorological variables.

S3 Principale Component Analysis

The table below presents the contributions of each variable as well as the correlation coefficients with the principal components derived from the principal component analysis.we have the coefficients that indicate the association between a variable and the principal component, the sign indicates the direction of this association The first principal component is associated among other things with a high maximum temperature (T2M_MAX) but also with high maximum wind speeds; Furthermore, it is associated with low relative and specific humidity and low precipitation frequencies The second component is first associated with high minimum temperatures (at 10m) and the earth's surface temperature and positively associated with the wind direction The third component is mainly associated with an increase in precipitation intensity.

Variables	Dim.1		Dim.2		Dim.3	
	Contribution	correlation	Contribution	correlation	Contribution	correlation
PRECTOTCORR	4.23 %	-0.55	0.43%	0.13	19.61%	0.59
WS10M _MAX	10.12%	0.85	0.07%	-0.05	12.52%	0.47
WS10M_MIN	4.48%	0.57	11.44 %	-0.67	8.89 %	0.40
T2M	5.43%	0.62	15.26%	0.77	0.06%	-0.03
T2M_MIN	0.01%	-0.02	23.81%	0.96	1.78%	0.18
PS	3.47%	-0.50	11.09%	-0.66	3.85%	0.26
QV2M	11.66%	-0.91	1.61%	0.25	2.79%	0.22
T2M_MAX	11.61%	0.91	3.14%	0.35	1.38%	-0.16
WS10M	9.22%	0.81	1.76%	-0.26	14.68%	0.51
WS2M	9.98%	0.85	0.45%	-0.13	14.31%	0.50
days_with_precipitation	8.71 %	-0.79	3.06%	0.34	9.07%	0.40
WD10M	1.64 %	-0.34	14.57 %	0.75	9.10%	0.40
TS	6.53%	0.68	13.29%	0.72	0.01%	-0.01
RH2M	12.91%	-0.96	0.01%	0.02	1.97%	0.19

Table 2: Table showing the contributions of each variable and the correlation coefficients with the principal components derived from the principal component analysis.

S4 Cluster Analysis and meteorological variable variation based on PCA Results

After performing principal component analysis (PCA), a clustering analysis was applied to the PCA results. The graphs below illustrate the variation of the variables across the different identified clusters

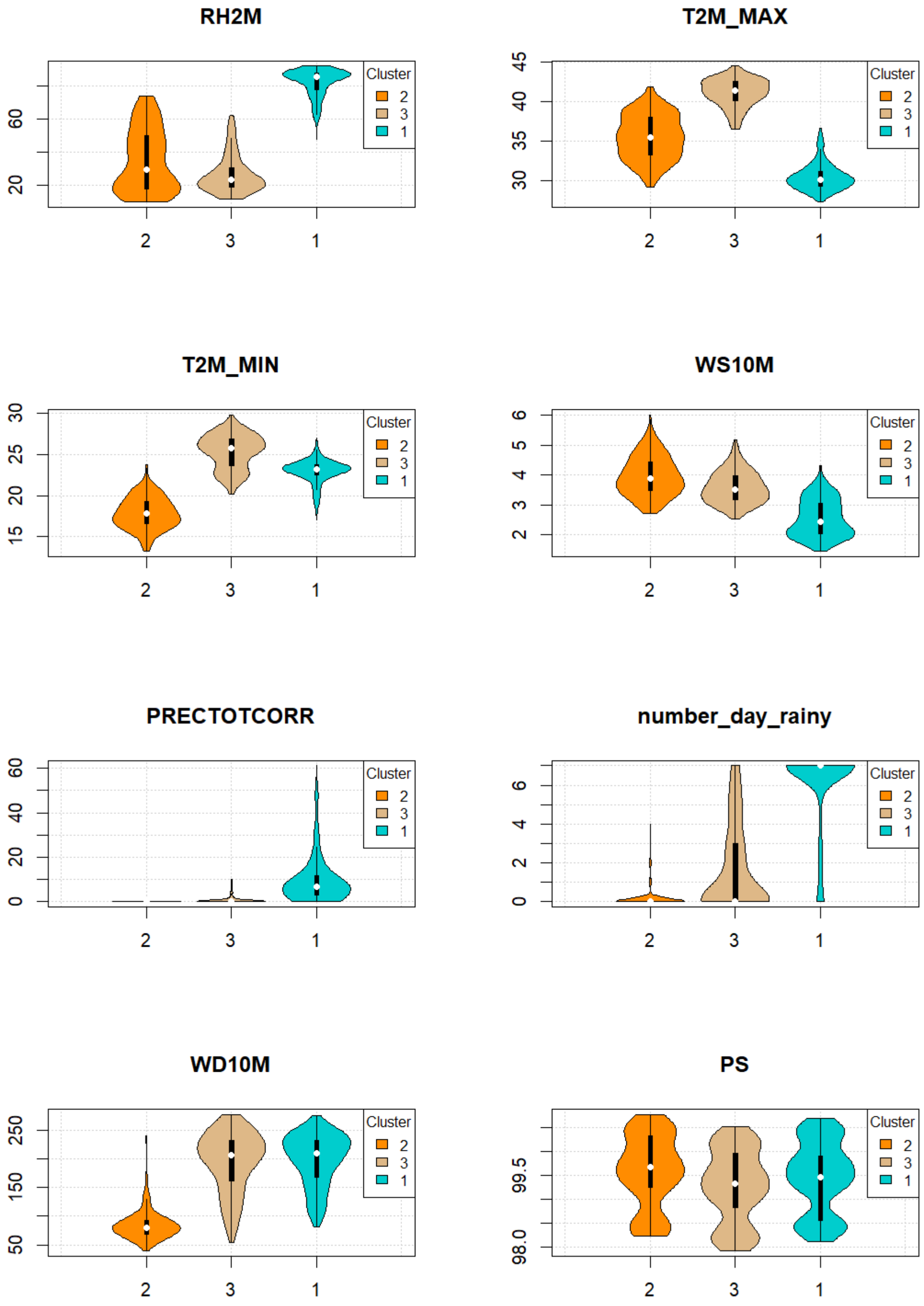

Figure 2: Variation on environnemental variables according cluster

S5 Model Comparison

To choose the best model among the three GAMs, you can compare the following criteria:the AIC, adjusted R^2 and deviance explained. The optimal model will have the lowest AIC, the highest adjusted R^2 (indicating better fit), and the highest deviance explained (reflecting the proportion of variability in the data captured by the model).

Models	AIC	Adjusted R^2	Deviance explained	RMSE train	RMSE test	df
Model 1	7746.715	92.2%	83.8%	117.605	111.872	31
Model 2	7714.471	91, 9%	84.3%	119.985	78.967	29
Model3	7745.266	92,1%	83.9%	117.653	110.509	34

Table 3: Comparison table of models for each district, using the GCV.Cp,adjusted R^2 and explained deviance criterion to evaluate the model's performance and complexity

Comparing these different measures, we can conclude that model 3 is more suitable.
This model, built using principal components with an interaction between meteorological clusters.

Model 2: Although it has the best test RMSE, the large gap between training and test RMSE (-41 points) suggests a risk of overfitting to the test set or high variability (possibly due to the structure of the test set or excessive model flexibility).

Model 1: The gap is small, but the errors remain relatively high, which may indicate that the model is slightly underfitting.

Model 3: It has the smallest gap between training and test RMSE (-7.1), which indicates: Better generalization than Model 1, Lower variance than Model 2, Potentially a good balance between bias and variance.

S6 Model Evaluation and Diagnostics

To assess the quality of the best GAM model fit, we performed a gam.check to verify the residuals conformity with the model assumptions and ensure that the model has properly captured the temporal structures in the data.

- The QQ plot shows the normality of the residuals. The points should follow a straight line, indicating that the residuals are normally distributed.
- The Residual vs Fitted values plot shows the distribution of the residuals relative to the predicted values. A good fit is indicated by a random distribution of the residuals around zero.
- The Residual histogram visualizes whether the residuals follow a symmetric distribution, which is another sign of a good model fit.
- The Fitted values vs Response plot visualizes how the predicted values (fitted values) compare to the actual observed values (response). A good model fit will be indicated by points close to the identity line ($y = x$).

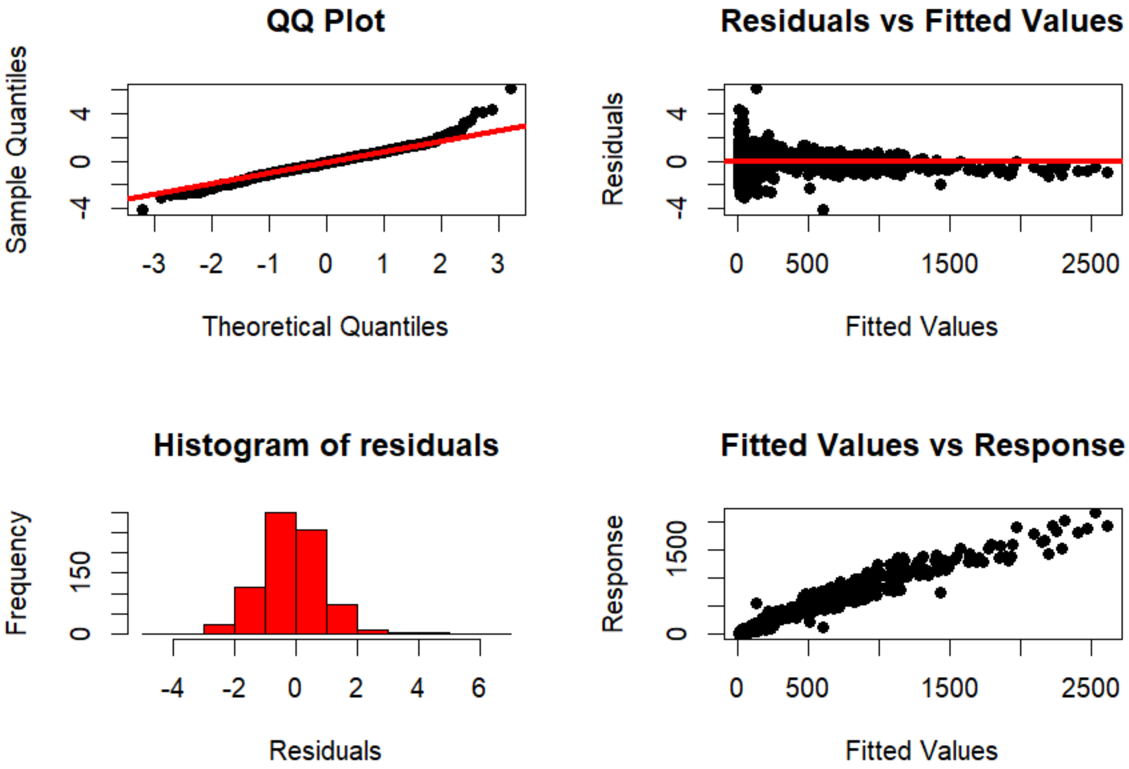

Figure 3: Diagnostic plots from the gam.check function).

These results, combined with the absence of residual correlations and the normal distribution of the residuals, suggest that the model is well-fitted and accurately captures the relationships in the data