



# Spatio-temporal modeling and risk ratio assessment of adult and egg mosquitoes abundance with consideration of environmental data in the island of Mayotte

Solym Manou-Abi, Lema Logamou Seknewna, Sophie Dabo-Niang, Julien Balicchi, Ambdoul-Bar Idaroussi, Maoulide Saindou

## ► To cite this version:

Solym Manou-Abi, Lema Logamou Seknewna, Sophie Dabo-Niang, Julien Balicchi, Ambdoul-Bar Idaroussi, et al.. Spatio-temporal modeling and risk ratio assessment of adult and egg mosquitoes abundance with consideration of environmental data in the island of Mayotte. 2023. hal-04224216

**HAL Id: hal-04224216**

**<https://hal.science/hal-04224216>**

Preprint submitted on 1 Oct 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Public Domain

# Spatio-temporal modeling and risk ratio assessment of adult and egg mosquitoes abundance with consideration of environmental data in the island of Mayotte

Solym M. Manou Abi <sup>1,2,3 \*</sup>, Lema L. Seknewna <sup>1</sup> , Sophie Dabo<sup>5</sup> , Julien Balicchi <sup>4 ‡</sup>, Ambdoul-Bar Idaroussi <sup>4 ‡</sup>, Maoulide Saindou <sup>4 ‡</sup>


**1** Centre Universitaire de Formation et de Recherche, 8 Rue de l'Université, 97660, Dembeni, Mayotte, France

**2** Institut Montpelliérain Alexander Grothendieck, UMR CNRS 5149, Place Eugène Bataillon, 34090, Montpellier, France

**3** Laboratoire de Mathématiques et Applications, UMR CNRS 7031, Futuroscope Chasseneuil, 86000, Poitiers, France

**4** Agence Régionale de Santé de Mayotte, 97660, Mamoudzou, Mayotte, France

**5** Laboratoire Paul Painlevé, UMR CNRS 8524, Cité scientifique, Villeneuve d'ascq, 59653, France

 These authors contributed equally to this work.

<sup>‡</sup>These authors also contributed equally to this work.

\*solum-mawaki.manou-abi@umontpellier.fr

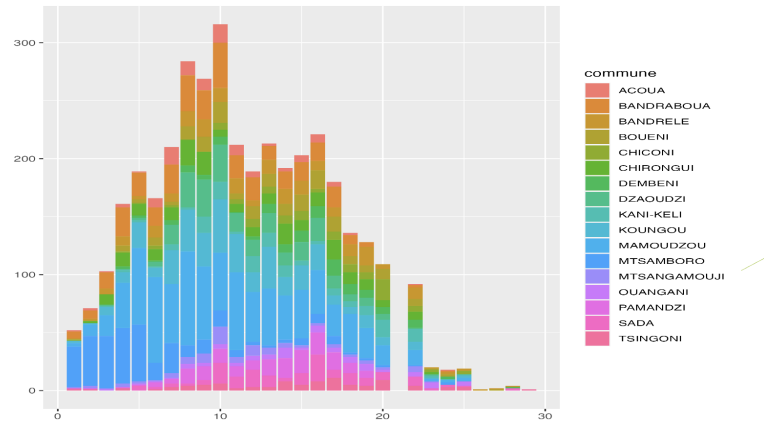
## Abstract

This work provides a spatio-temporal statistical modeling for egg and adult *Aedes* mosquitoes count data with consideration of environmental data in the context of mosquito epidemiology. For a given spatio-temporal incomplete mosquito count data, we derive predictions assuming that all of the spatio-temporal dependence can be accounted by potential factors influencing the development of *Aedes* mosquitoes such as, rainfall, temperature (including many delay) and waste data. In this paper, after a data analysis with the entomological and environmental data, we apply a LASSO regression to perform a variable selection strategy (we are in the presence of a large number of explanatory variables). We highlight the relevant factors that explain the abundance of our egg and adult mosquitoes count data. We define a spatio-temporal risk ratio which is a probability of exceeding a given threshold value of mosquito abundance. We propose two spatio-temporal modeling approaches for the *Aedes* mosquitoes' count data. The first is based on a spatio-temporal kernel smoother and the second on a generalized additive model. The paper conclude with a detailed discussion that follows not only, the spatio-temporal prediction and model performance measures, but also the obtained spatio-temporal risk ratio in order to highlight potential space-time areas of threshold exceedances of mosquitoes' abundance where health services could apply vector surveillance and control measures. We also discuss a forthcoming work concerning the theoretical developments of a spatio-temporal model for simulation purposes and an R Shiny application called May'Aedes, that aims to be a flexible and efficient tool to predict spatio-temporal risk ratio.

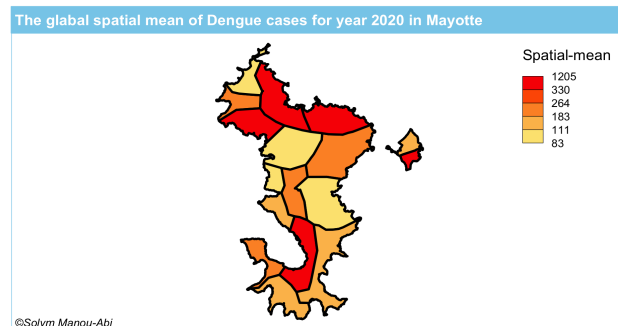
# Introduction

In the island of Mayotte, mosquitoes can transmit serious diseases: chikungunya, dengue fever and malaria [1]. Dengue is the world's most prevalent mosquito-borne viral disease that spreads mainly from *Aedes (aegypti and albopictus)* to people. It is endemic in the island of Mayotte and also in many tropical and subtropical countries. This represents a significant global health burden and, its dynamic is seasonal with peaks during the wet-hot months. Note that, during the year 2020, more than 3533 confirmed cases of dengue fever have been reported on the island [2]. We show in Figure 2 the global spatio-temporal variation of the detected cases. Vector monitoring, recommended by the World Health organisation, is a routine practice in many dengue endemic countries to provide quantifiable measures of fluctuations in time and space. The tropical climate in the island of Mayotte is particularly favorable to vector-borne diseases [3]. The presence of various mosquitoes vectors represents a major health risk for the population of Mayotte. This has encouraged reflection on the establishment of a concerted effort for the monitoring and prevention of related vector of *Aedes*. The increase in mosquito population following climate and environmental conditions caused a major threat to humans because of mosquitoes' ability to carry disease-causing pathogens. The entire population is affected by vector control since approximately 80% of mosquito breeding sites are created by humans around their homes. Among the planned efforts are vector surveillance and control approaches. A practical strategy to minimize dengue and other borne diseases transmission commonly relies on vector control which aims to maintain *Aedes* mosquito density below a theoretical threshold. One of the simple strategies is the monitoring of mosquitoes' densities abundance and, regular mosquito control operations. However, mosquito abundance modeling can help to identify areas with higher risk assessment of disease transmission since one can lower their risk of dengue by avoiding especially strong risk ratio of abundance in space (villages or communes) and time (weeks). In general vector abundance data are limited or restricted for most parts of the world due to the cost and effort required to collect such data. We consider in this work, adult and egg *Aedes* mosquito population collected in certain areas during a given period and in common spaces trough traps, by a monitoring team (to be described in the sequel) in the island of Mayotte. This is a great opportunity we took here to highlight power full of spatial data models as well as spatial dynamics aspects with such available entomological data. We recall that, the life cycle from an egg to an adult, for *Aedes* mosquitoes, typically takes up to two weeks, but depending on conditions (water, temperature, food) and type of mosquito, it can range from 4 days to as long as a month [4]. The adult mosquito emerges onto the water's surface and flies away, ready to begin its lifecycle. Different modeling studies of their distribution and dynamics have been developed recently [5–7] in the neighbouring islands of Reunion and Mauritius. However, on the island of Mayotte and up to our knowledge, none of this previous research has been carried out. Moreover none of this research adress directly the question of a spatio-temporal modeling with the models we propose in this paper. Our objective in this study, is to analyse the egg and adult mosquito count data in the island of Mayotte and to establish statistical methods to highlight a number of potential predictors for the abundance of our egg and adult mosquito count data, in order to produce spatio-temporal models for prediction. We define a spatio-temporal risk ratio of abundance that can be seen as a probability of exceeding a given threshold value of mosquitoes' abundance. After that we propose two spatio-temporal models that can take into account potential predictors (temperature, rainfall, waste), [8–10]. The first model is a modified and adapted spatio-temporal kernel smoother which reveals to be suitable for our eggs count data. The second model is an application of the Generalized Additive Model (GAM) in a spatio-temporal setting both for egg and adult count data. We discuss the performance measures of such data

process models as well as some relevant spatio-temporal predictions that can be subject of a control measure in the monitoring of the mosquito abundance and therefore the risk of contracting dengue fever. The paper end by the announcement of future projects. For instance, a forthcoming R Shiny application called May'Aedes, that aims to be a flexible and efficient tool to produce an online spatio-temporal risk ratio for control measures. Also a forthcoming theoretical developments of a spatio-temporal simulation model based on a mechanistic Partial Differential Equation (PDE) model.



**Fig 1.** Dengue reported cases by week during the year 2020 in the island of Mayotte.



**Fig 2.** Total spatial mean variation of Dengue reported cases during the year 2020 in the island of Mayotte.

## Materials

The material of this work deals with eggs and adults *Aedes* (*Albopictus* and *Aegypti*) mosquitoes count data between 2018 and 2021 captured per trap per day/week in areas chosen to meet specific criteria including accessibility, and geographic isolation in the island of Mayotte. The data analysis and processing is done in R programming language.

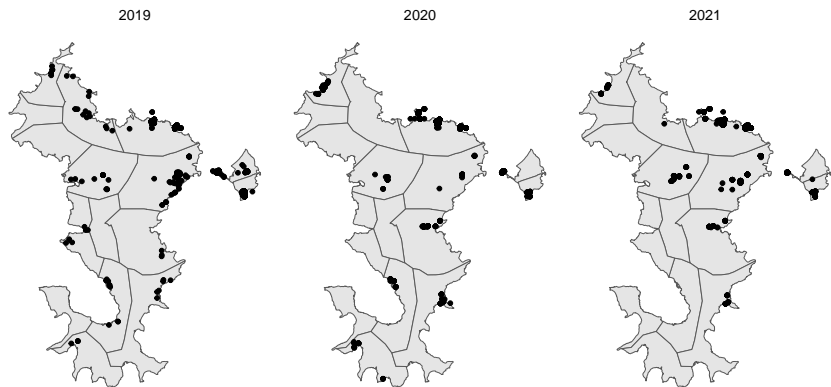


## The Vector Control Team and Department in the island of Mayotte

Vector surveillance is an integral component of an Integrated Vector Management (IVM) program and is the primary tool for quantifying virus transmission and human risk. This principal function of a mosquito-based surveillance program in the island of Mayotte is set up by the Vector Control (LAV) Department of the french Regional Health Agency (ARS) in the island of Mayotte carries out entomological surveillance and research in order to anticipate the risk of vectors being transmitted to the human population. LAV agents carry out surveillance of arboviroses and operations in areas surrounding for example, reported cases of mosquito-borne viral disease (chikungunya, dengue, malaria). To do this, they capture mosquitoes (through traps) and identify the species involved in the transmission of the disease; eliminate or treat all situations where mosquitoes proliferate (eggs and adults breeding grounds, pots, waste, stockpiles of tyres, stagnant water as well as other spatial relevant environmental conditions). This allows insecticide and impregnation treatments [11,12], to be put in place and inform the local population about the potential space-time risk of transmission of mosquitoes disease and how to protect themselves against mosquito bites. The statistical analysis and modelling approaches proposed in this work are part of the vector control actions against the *Aedes* mosquito. We hope to contribute to the orientation or recommendation of LAV department actions.

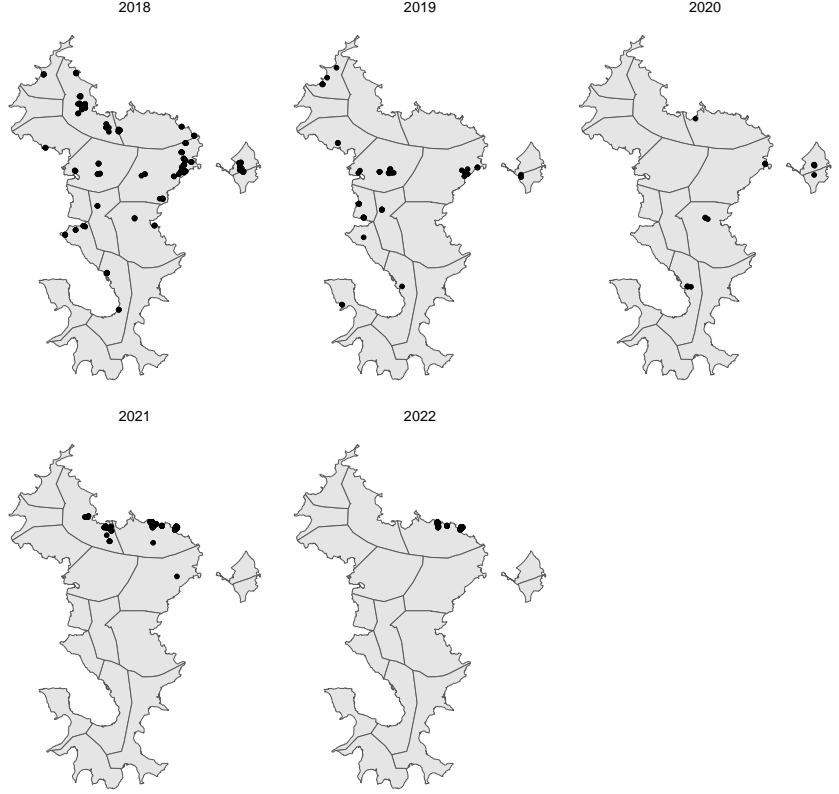
### Study area and entomological Data

The study area as mentioned above is located in the island of Mayotte, which is a french department of 374 km<sup>2</sup> composed of two islands (the largest and main island called Grande-Terre) with a total of 17 communes and located in the Indian Ocean, East of Africa, in the Mozambique Channel, separating Madagascar from Africa. Data monitoring sites of adult and egg mosquito year are shown in Figures 3 and 4. The laying of mosquito traps (eggs and adults) was carried out by a team from the Regional Health Agency (ARS in French) of Mayotte, which confirm the presence of *Aedes* mosquitoes. The monitoring of egg data were available only for years between 2019 and 2021, but for adult data, the available years are range from 2018 to the beginning of 2022.



**Fig 3. Spatial distribution of egg mosquito collection sites from 2019 to 2021.**

The weekly distribution of the entomological Data is shown in S1 Fig and S2 Fig. We can observe that some collection sites have more than one time of collection in a



**Fig 4. Spatial distribution of adult mosquito collection sites available from 2018 to 2022.**

week with missing observations (both temporally and spatially). Moreover, the Data is not evenly distributed in all the communes, i.e. there was no data collected over some period of time in some places. The yearly percentage of egg and adult data observations for all represented communes is shown in S1 Table. The years 2020 and 2021 are under-represented in the adult mosquito dataset because only 5 communes were concerned with data collection. We see a big void in 2020 (which can be explained by the lack of collection activities or stolen traps or Covid-19 constraints) from week 8 to week 38 then from week 44 to the end of the year. The year 2020 will be issued due to a lack of data. The more the years increase, the less we see the communes. The second largest block of data in terms of percentage is concentrated in Koungou and Bandraboua in 2021 (37%). The range of data from Koungou in 2021 is broader than that of other communes in previous years. In 2022, there was no collection almost in the first three weeks of the year except for weeks 13 and 15. Only two communes are present even if more than a third of observations are concentrated there.

## Environmental Data

The environmental data considered in this work concerns climate (for instance temperature, rainfall) and waste data (waste containers such as poor housing, sewage deposit, tire waste, waste drain). Such co-variate data are well known factors in the proliferation of mosquitoes population in a given area [3, 13–17].

## Climate Data

The climate is tropical with two distinct seasons. A rainy season globally from November to March (austral winter from Weeks 1 to 16 and 40 to 52 or 53). The temperature is particularly high and the humidity is also high for this austral summer and concentrates most of the annual precipitation. The dry season (austral summer) lasts from April to October (weeks 17 to 39). The temperature and humidity are lower than in the Austral summer, with less precipitation. There are not enough weather stations in Mayotte and the meteorological french service (<https://publitheque.meteo.fr>) provide the daily temperature (minimum and maximum) and rainfall records from 2018 to 2021 at few available weather stations and some of them are closed to surveillance areas where entomological data were collected. Such areas of entomological data collection were defined before the start of this study. For instance, rainfall data come from stations located in 6 communes, namely Bandré, Dembeni, Mamoudzou, Mtsamboro, Ouangani and Pamandzi. Note that before 2018, the Mtsamboro station did not yet exist, or we did not have data recorded for that commune. Figure 5 shows the map of stations and missing data by station. In 2022, we had data for rainfall from 5 stations (Mtsamboro, Mamoudzou, Ouangani, Pamandzi and Bandré only). No data for Dembeni. Lastly, we have data for rainfall from 4 stations (Mtsamboro, Mamoudzou, Ouangani, Pamandzi, Dembeni and Bandré). No data for Dembeni and Ouangani in 2022. In 2020, We had data for rainfall from 6 stations (Mtsamboro, Mamoudzou, Ouangani, Pamandzi, Dembeni and Bandré) except the commune of Dembeni which has around 67% of missing records for temperature. In 2018, we had complete data for rainfall from 5 stations (Mamoudzou, Ouangani, Pamandzi, Dembeni, and Bandré) except the commune of Bandré which has no data record for temperature. The following year, i.e 2019, we have data for rainfall from 6 stations (Mtsamboro, Mamoudzou, Ouangani, Pamandzi, Dembeni and Bandré) except the commune of Ouangani which has around 34% of missing records for temperature (minimum and maximum temperature). As the remaining 11 communes did not have climate data recorded, we shall be concerned with data imputation methods.

## Waste Data

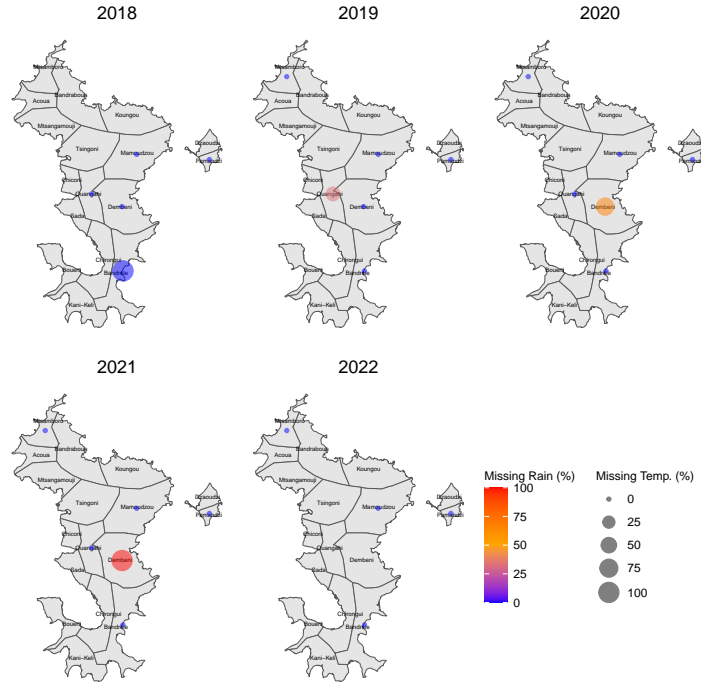
The following picture in Fig 6 shows the yearly spatial variation of the waste type.

## Merged Data

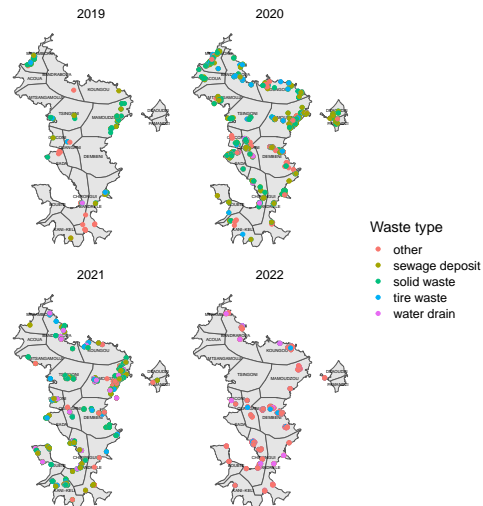
The weather and waste data, identified in the literature as potentially influencing *Aedes* mosquito abundance, were merged to the egg and adult count data for the modeling process. The set of co-variables (see also [7] for meteorological variables) that will be considered for our modeling purpose is given in Table 1. When dealing with adults mosquitoes count data, we also include the eggs number as a probable explanatory variable. In the following section, we shall describe the main purpose of this paper and the methods we have developed.

## Data Analysis and Imputation

In the presence of missing co-variate observations, there are suitable imputations methods well known in the literature, ranging from the most rudimentary techniques (median, mean) to those by statistical learning. This section part describes a classic imputation technique in order to have a complete co-variate data of meteorological data.



**Fig 5.** Map of meteorological stations in Mayotte and missing data



**Fig 6.** Yearly distribution of waste in Mayotte

### k-Nearest Neighbors Algorithm

We deal firstly, with the imputation of meteorological Data. We consider the k-Nearest Neighbors (k-NN) ([18]) method, using coordinates to determine the proximity between communes or village, in order to impute the above climate data is reused in the following lines. This is also in line with Tobler's First Law of Geography which says that things that are closer tend to be much more alike than things that are far away [19]. The k-NN method is a widely used machine learning algorithm that can be

**Table 1.** Set of potential explanatory variables as potentially influencing Aedes mosquito abundance

Variable	Description
$Z_{1,N} = MTN_N$	Daily minimum temperature (last N days)
$Z_{2,N} = MTX_N$	Daily maximum temperature (last N days)
$Z_{3,N} = RRcum_N$	Rain accumulation (last N days)
$Z_{4,N} = MRRN$	Maximum rainfall (last N days)
$Z_{5,N} = MNDwtRR_N$	Maximum number of consecutive days without rain (last N days)
$Z_{6,N} = MNDwRR_N$	Maximum number of consecutive days with rain (last N days) where the rainfall accumulation is greater than a given threshold value as it's percentile $S \in [10, 20, \dots, 90]$
$Z_7 = \text{waste\_number}$	Daily observed waste number
Week, Year	Temporal variables
$(X, Y)$	Spatial coordinates
N	Retroactive period that starts at the capture date: : $N \in \{7, 14, 21, 28, 35, 42, 49, 56, 65\}$

used also for classification and regression tasks. It was first proposed by [20] in 1951 for classification problems and later extended to regression problems by [18] in 1967. The k-NN algorithm works by finding the  $k$  nearest neighbors to a query point based on a chosen distance metric. This will allow the choice of the observed values of those neighbors to make a prediction value for the query point. The choice of distance metric and the choice of  $k$  can have a significant impact on the performance of the algorithm. The k-NN algorithm is a non-parametric method, which means that it does not make any assumptions about the underlying distribution of the data. It can work well in high-dimensional spaces and can be adapted for use with different distance metrics. However, it can be computationally expensive for large data sets. One popular distance metric for the k-NN algorithm is the Euclidean distance. Another commonly used distance metric is the Manhattan distance, which is calculated as the sum of the absolute differences between corresponding elements of two vectors. The  $k$ -NN algorithm requires the following steps:

**Data:**  $X$ , a set of  $n$  query coordinates points with unobserved values of rain or temperature and  $R$ , a set of reference coordinates points with observed values of rain or temperature.

**Result:** The set of query coordinates points  $X$  with complete observed values of rain or temperature

**for**  $i = 1, \dots, n$  **do**

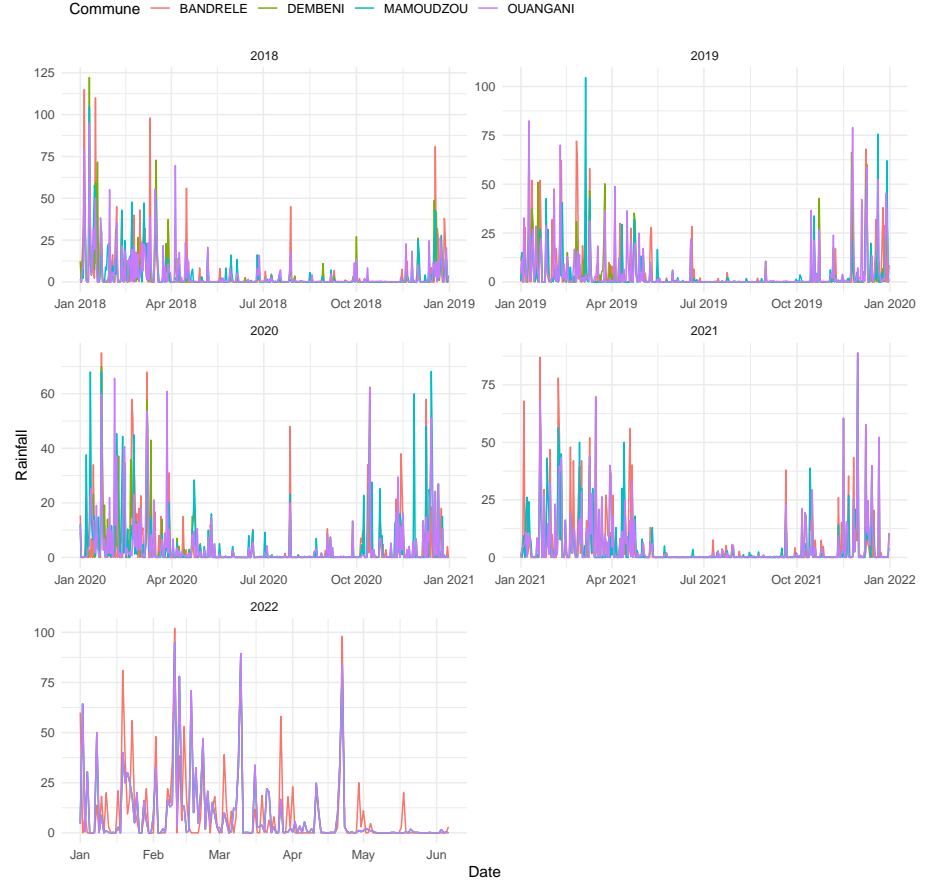
- Step 1 Compute the Euclidean distance  $d(X_i, R)$ .
- Step 2 Arrange the calculated  $n$  Euclidian distances in non-decreasing order.
- Step 3 Count the number of  $k$  small distances
- Step 4 Replace the missing value by the value computed from the  $k$  closest point.

**end**

**Algorithm 1:** K-NEAREST NEIGHBORS ALGORITHM

We apply the k-NN algorithm to impute the missing daily temperatures and rainfall for each commune or village from the neighboring observation. As an example of the

imputation result, Figure 7 show the partially observed rainfall data in Dembeni (around 50%) for the year 2020 and completed by the observed data in the neighboring commune (here Ouangani was almost chosen by the k-NN algorithm). The same was also done and check for year 2021. In 2022, climate data of Mamoudzou were used by the algorithm to fill the gaps in Dembeni (Dembeni is the southern commune near Mamoudzou) and Ouangani (a central commune also close to Mamoudzou), confirmed by the comparison with the others years where the data was available. This applies also to the imputation of the minimum temperature in Figure 8. In our case, we choose  $k$  to be 1 for the k-NN method. One can choose  $k \geq 1$  and the quantity of interest, say the rainfall in millimeters, will be the average, min, max, median, etc. of the figure observed in the  $k$ -neighboring communes.



**Fig 7.** Comparison of rainfall between closest communes, after data imputation

A Complete map of rainfall about the maximum number of consecutive days with rain is shown in S5 Fig for year 2019. We observe that weeks 42 to 53 and week 1 to 18 record more maximum number of consecutive days with rain than the others. This confirm the observed seasonal variation in the island of Mayotte. Temperatures are pleasant all year round and the mean A temperatures range from 28 degrees (June) to 31 degrees (February).

This completeness of climatic data by the k-NN algorithm could be improved by taking into account the relief, in particular the terrain attributes (slope, orientation), terrain variability (roughness, curvature, convexity of the profile and also altitudes but



**Fig 8.** Comparison of minimum temperature between closest communes, after data imputation

no data available) in order to improve the climatic data by geographical proximity.

### Outliers detection in entomological Data

We discussed the issue of outliers due to the high frequency of data collection in some few communes. Therefore, a commune where data was not collected frequently may be underrepresented spatially. The commonly used and popular method for outliers detection is the boxplot method. It is a method for graphically depicting groups of numerical data through quartiles (first (Q1), median, third (Q3), minimum and maximum) to easily detect outliers and how the data is skewed. The difference between the minimum and maximum tells us about the range of dataset. The difference between Q3 and Q1 is called the Inter-Quartile Range (IQR). Not every outlier is a wrong value and it is not acceptable to drop an observation just because it is an outlier. When we shouldn't drop an outlier, we can apply some transformation methods. For instance an application of the IQR method with a boxplot tell us more or less about the distribution of the data, [21]. Any data point less than the Lower Bound ( $Q1 - 1.5 \times IQR$ ) or more than the Upper Bound ( $Q3 + 1.5 \times IQR$ ) is considered as an outlier. Hence the number 1.5 (hereinafter scale) controls the sensitivity and the decision rule. This scale, depends on the distribution followed by the data. But this scale depends on the distribution followed by the data. For instance, we can calculate the IQR decision range in terms of the standard deviation of a Gaussian Distribution. To deal with the issue of outliers, we are going to perform some data transformations, thank to the Central Limit Theorem which grants us the liberty to assume the gaussian distribution without any guilt. Figure in S3 Fig and S4 Fig shows the outliers detected for egg and adult mosquito count data, respectively. Note that for the last two years (2021 and 2022), data collection was mostly performed in Koungou and we notice high figures as compared to

the counts in the previous years. We have chosen to exclude the high number of adult mosquitoes in the Koungou area in 2021 and 2022 because the other communes are not represented. In the latter case, we use data from years 2018, 2019 and 2020.

## Methodology

We adopted a spatio-temporal modeling approach to investigate the eggs and adults *Aedes vector* abundance in Mayotte island. Spatio-temporal models are useful tools applied in many research fields dealing with empirical data. They can connect spatially, an outcome variable (eggs and adults mosquitoes' values) to one or several variables (co-variables) and quantify the strength of association between them and the outcome variable with given performance measures. First of all, we propose a novel and simple spatio-temporal risk ratio in order to evaluate the predictions results of the proposed models in the first part of this section. Next we turn into a variable selection method through a suitable regression approach. Given that we are handling large sets of relevant co-variables, we deemed necessary to consider the LASSO (Least Absolute Shrinkage and Selection Operator [22]) to obtain a better selection of the potential factors that mostly influence the abundance of our mosquito eggs and adults count datasets. In the third part of this section, we investigate a non-parametric model that may be appropriate for missing observations. More precisely, we improve a spatio-temporal kernel smoother introduced in [23], by adapting it to take into account the selected relevant variables obtained by the LASSO. We apply the obtained modified model to the egg mosquito count data. In the fourth part of this section, we explore Generalized Additive Models (GAM) [24] in order to take into account the non linearity and obtain better predictions. We end by recalling spatio-temporal goodness measures to evaluate the models' performance.

### The proposed spatio-temporal risk ratio

To evaluate the predictions results of the models, we introduce a spatio-temporal relative risk ratio as follows. We aim to define a probability of the presence of mosquitoes in the egg and adult stages by the proposed training and prediction models. In the sequel, we will use  $\hat{Y}$  to denote either the Egg or Adult stage mosquito value estimated. Let  $\hat{Y}(s, t)$  be the space-time sequence of predicted mosquitoes values at a spatial location  $s$  and time  $t$ . To take care of out-of-range predictions, we defined a simple spatio-temporal risk ratio function that includes a threshold  $x^*$  such that all predictions greater than or equal to  $x^*$  will have 1 as ratio. For example, values of  $x^*$  can be the maximum observed value or other practical threshold value.

$$Ratio(s, t) = \frac{\hat{Y}(s, t) - \min_{s, t}(\hat{Y}(s, t))}{x^* - \min_{s, t}(\hat{Y}(s, t))} \mathbb{I}_{\{\hat{Y}(s, t) < x^*\}} + \mathbb{I}_{\{\hat{Y}(s, t) \geq x^*\}} \quad (1)$$

with  $t$  measured on a weekly scale  $t \in \{t_1, \dots, t_T\}$  and  $s$  on a spatial grid  $s \in \{s_1, \dots, s_m\}$  of size  $m$ .

### The variable selection method

The variable selection procedure we adopt here is the Least Absolute Shrinkage and Selection Operator (LASSO) which is mainly useful when dealing with large sets of explanatory variables such that some of them are suspected of being less relevant. Hence, we use this method to perform variable selection, since we are in the presence of a large number of variables because of the delay combinations taken into account in the target variables. Note that, the LASSO regression method is a regularization and popular



technique for achieving regularized parameter estimates and reducing their variability through a penalty term [25]. Consider a penalized regularization in which a penalty term is added to the Residual Sum of Square (RSS, defined in 4) that effectively shrinks the regression parameter estimates  $\beta_j$ ,  $j = 1, \dots, p$  towards zero (bickel2009simultaneous). Specifically, consider estimates of the vector parameter  $\beta$  given by :

$$\hat{\beta}_L = \arg \min \left( RSS + \lambda \sum_{j=1}^p |\beta_j| \right),$$

where:

- $\lambda \sum_{j=1}^p |\beta_j|$  represents the penalty function
- $\lambda$  represents the tuning parameter that determines the  $\hat{\beta}$  that shrunk towards 0.

If  $\lambda = 0$ , then there is no shrinkage at all, if  $\lambda < 0$  ie huge, the function can be minimized by shrinking all the parameters back towards zero. The shrinkage factor, which may be calculated using bootstrapping, is essentially an estimate of the extent of overfitting [26]. The cross-validation method is used to try out the different values of  $\lambda$ . The LASSO regression can be defined as a shrinkage method that can actively select from a broad and potentially multicollinear collection of covariables in the regression, yielding a more relevant and interpretable set of predictors. The unique feature of LASSO is that it penalizes the absolute value of a regression coefficient, hence controlling the impact of a coefficient on the total regression. The stronger the penalization, the smaller the coefficients get, with some approaching zero, removing unnecessary influential factors automatically [22]. LASSO regression works like a feature selector that picks out the most important coefficients, i.e. those that are most predictive (and have the lowest p-values). The key drawback of this model is that if there are two or more highly collinear variables then LASSO regression selects one of them randomly which is not good for the interpretation of data also if the number of predictors  $p$  is greater than the number of observations  $n$ , Lasso will pick at most  $n$  predictors as non-zero, even if all predictors are relevant. In the CARET package of R, variable importance can be calculated using different methods depending on the type of model being used.

## The proposed spatio-temporal kernel smoother predictor

We introduce, in this section, the proposed spatio-temporal model to predict the unobserved egg data. It is based on the so-called Inverse Distance Weighted (IDW) model (see [23]). As the name suggests, it predicts the attribute value of a variable at positions where no samples are available based on the spatial distance between that position and other positions where samples have been collected. An advantage of the IDW model is that it takes into account available observations to make the predictions of mixing observations. Assume that we have  $m$  observed values of a mosquito population value  $Y$  in a given stage at time  $t_j : j = 1, \dots, T$ , and locations points  $\{s_i : i = 1, \dots, m\}$ . The spatio-temporal IDW predictor of the mosquitoes population  $\hat{Y}$  on the unobserved location  $s_0$  and time  $t_0$ , is given as follows:

$$\hat{Y}(s_0; t_0) = \sum_{j=1}^T \sum_{i=1}^m K_{ij}(s_0; t_0) Y(s_i, t_j) \text{ with } K_{ij}(s_0; t_0) = \frac{\tilde{K}_{ij}(s_0; t_0)}{\sum_{k=1}^T \sum_{l=1}^m \tilde{K}_{lk}(s_0; t_0)},$$

where

$$\tilde{K}_{ij}(s_0, t_0) = \frac{1}{d((s_i; t_j), (s_0; t_0))^\theta} \quad (2)$$

and  $d$  is a well-chosen distance between the spatio-temporal location  $(s_{ij}; t_j)$  and the forecast spatio-temporal location  $(s_0, t_0)$ . The transition kernel function  $K$  depends on the bandwidth parameter  $\theta$  that specifies the redistribution of weights for the observed process according to the above process. More precisely  $\theta$  controls the amount of smoothing. This spatio-temporal data process is a simple weighted average of the data points, giving the closest locations more weight by requiring the weights to sum to one. We can also take a look at other kernel, for instance the Gaussian kernel :

$$\tilde{K}_{ij}(s_0, t_0) = \exp\left(-\frac{1}{\theta}d((s_i; t_j), (s_0; t_0))^2\right),$$

where the bandwidth parameter  $\theta$  is proportional to the variance parameter in a Gaussian distribution. Many other kernels exist in the literature but in this work, we focus on the two previous kernels. We aimed to improve the IDW model structure by including the selected relevant variables by the Lasso regression, through the calculation of the distance  $d$ . To this end, assume that, at locations points  $\{s_i : i = 1, \dots, m\}$  we consider the associated  $k$  features  $c_{ij}^{(1:k)} = (c_{ij}^{(1)}, \dots, c_{ij}^{(k)})$  at time  $t_j$  and spatial location  $s_i$  corresponding to measurement values of  $k$  relevant explanatory variables that may explain the abundance of the mosquitoes' population. More precisely, we may write for example

$$\tilde{K}_1((s_i; t_j), (s_0; t_0)) = \frac{1}{d((s_i, c_{ij}^{(1:k)}; t_j), (s_0; c_0^{(1:k)}, t_0))^\theta}. \quad (3)$$

This modified process can be seen as a weighted average of the data points, giving the closest locations with the closest co-variables values. More specifically, these are the observations, that are close to the unobserved point to be predicted not only spatially, but also according to their associate features values, that receive a greater weight in the prediction, while distant observations in space or in characteristics will have a relatively weak influence on the prediction. At this stage, we have practical implications considering the kernel in 2, because not only data may lead to very small values approaching zero which causes problems in the above equation, but one needs also a convenient choice of this distance to take into account the effects of features in time and space. The mix-max scaling or normalization helps to solve the latter situation. After that if values of 0 are present in the data, we derive the following generalization where  $N$  is the number of observations:

$$Y^*(s; t) = \frac{\tilde{Y}(s; t)(N - 1) + 1/C}{N}$$

and  $C$  the total number of co-variables. This compresses the data symmetrically around .5 from a range of 1 to  $(N - 1)/N$ , so extreme values are affected more than values lying close to 1/2. Additionally, we see that for  $N$  large enough the compression vanishes, that is, larger data sets are less affected by this transformation.

## Generalized Additive Model

This part illustrates how Generalized regression Models can be use for spatio-temporal data. A Generalized Additive Model (GAM), [24, 27, 28], is a type of Generalized Linear Model (GLM) where the linear predictor has a linear relationship with predictor variables and smooth functions  $s(\cdot)$  of predictor variables. Recall that, in a usual GLM, prediction, assume that the dependence of the target variable can be accounted for, by "trend" (i.e. co-variate) terms linearly through a specified monotonic link function  $g$  :

$$g(Y(s, t)) = \beta_0 + \beta_1 X_1(s, t) + \dots + \beta_p X_p(s, t)$$

where  $\beta_0$  is the intercept and  $\beta_k, k > 0$ , is a regression coefficient associated with the  $j$ -th co-variate  $X_j(s, t)$  at spatial location  $s$  and time  $t$ . This regression model can be fitted via ordinary least squares (OLS) in which case we find estimates of the parameters  $\beta_0, \beta_1, \dots, \beta_p$  that minimize the residual sum of squares (RSS):

$$RSS = \sum_{j=1}^T \sum_{i=1}^m \left( Y(s_i, t_j) - \hat{Y}(s_i, t_j) \right)^2. \quad (4)$$

A GAM is a generalized linear model with a linear predictor involving a sum of smooth functions  $s$  of co-variables (predictor variables):

$$g(Y(s, t)) = \beta_0 + f_1(X_1(s, t)) + \dots + f_p(X_p(s, t))$$

where the functions  $f_j$  are smooth functions with a specified form (polynomial basis function, cubic spline basis, etc.). GAM prediction, thus require some smoothing terms used for prediction. It is necessary to represent the smooth functions in some way and to choose how smooth they should be. Natural cubic splines are proven to be the smoothest interpolators, as shown in [24]. This approach defines the splines by  $s$  in terms of their values at some knots  $k$ , which sets up the dimensionality of the smoothing. Various smooth classes are available and smooth terms are specified in a GAM formula using  $s$ ,  $te$ ,  $ti$  and  $t2$  for different modelling tasks terms linked to in the R `MGCV` package [29]. For spatio-temporal count data, one can examine the GAM with the mean response  $g(Y(s, t))$  related to some family distribution such as Poisson or Negative Binomial.

## Performance : spatio-temporal goodness of fit results

To evaluate the performance of our prediction models, to draw the best conclusion and interpretation for the data, we'll make use of the following performance metrics and better selection issue. One of the most common scalar validation statistics for continuous-valued spatio-temporal processes is the mean squared prediction error (MSPE) given by

$$MSPE = \frac{1}{Tm} \sum_{j=1}^T \sum_{i=1}^m |Y(s_i; t_j) - \hat{Y}(s_i; t_j)|^2.$$

It is useful when one wishes to protect against the influence of outliers. The Root Mean Squared Error (RMSE) is defined as the root of MSPE. The Adjusted Coefficient of Determination (Adjusted R-squared)  $R^2$  [30], is another measure that provides information about the goodness of fit of a model in the context of regression. The use of a cross-validation principle will be necessary for a better selection (for instance the bandwidth parameter  $\theta$  for the spatio-temporal kernel predictor). In the case of Leave-One-Out Cross-Validation (LOOCV), only a single observation is used for validation and the remaining observations are used to make up the training set. In an  $u$ -fold cross-validation, this method requires the model to be fitted  $u$  times because this is repeated for all  $m$  observations. The LOOCV score is then evaluated as follows:

$$CV_{(u)} = \frac{1}{u} \sum_{i=1}^u MSPE_i.$$

## Results and Discussion

Predicting vector abundance is an essential part of modelling vector-borne disease spread such as the dengue disease caused by *Aedes* vector. In the island of Mayotte, the

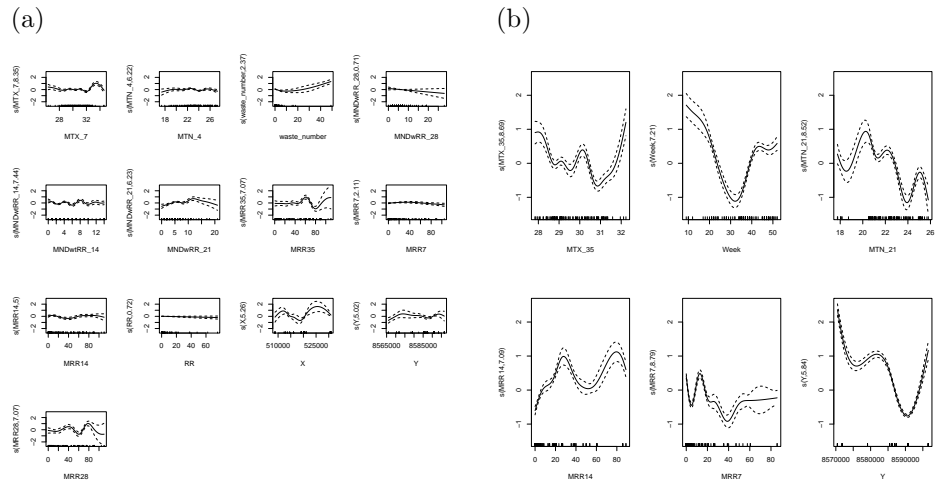
above dengue outbreak in 2020 had make an urgent need for operational actions of the vector control team. Following the data analysis, we present in this section the results of the modeling methodology for the eggs and adults mosquitoes' count data together with a detailed discussion.

## Variables and Models Selection

From a large set of potential factors described above, we obtained relevant variables by the LASSO regression through and their importance is given in S3 Table and S4 Table for egg and adult mosquito data count respectively. Our results reveals the importance of environmental conditions as in [7]. For example, the main factors identified as influencing the abundance of adult mosquitoes are as mainly among others : cumulative rainfall over the last 5 days, daily maximum rainfall over the last 35 days ; daily minimum temperature, maximum number of days without rainfall, maximum daily temperature over the last 35 days. In the case of eggs data, we retain among others, the cumulative rainfall over the last 28 days, daily maximum rainfall over the last 6 days, daily minimum (resp. maximum) temperature over the last 35 (resp. 28) days and, mainly the maximum number of days without rainfall. This observed delay time period could probably be explained by the biological cycle of the *Aedes* mosquito development.

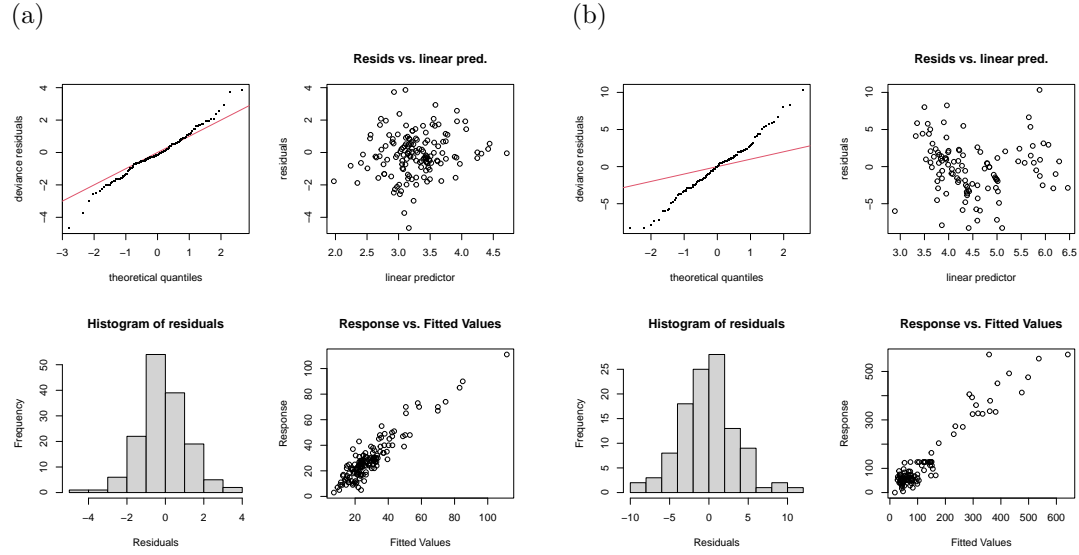
In the sequel and for simplicity, we consider the selected important variables obtained from the LASSO for the data models training, more precisely the GAM models (for eggs and adults mosquitoes) and also the above proposed spatio-temporal kernel smoother. We apply these two models described in the Methodology for prediction and comparing they performance measure. Our results demonstrate that such spatio-temporal modeling approach can provide consistent and efficient tools for vector population dynamics surveillance both in space and time. The explicit formulae of the selected GAM models (Model1, Model 2 and Model 3) selected as well as the performance measures achieved by the models are given in S2 Table. Model 1 is for egg mosquito data and trained with 2019-2020 eggs data (since they have the same range) whereas Model 2 is trained with 2021 data. Model 3 is the best selected GAM models for adult mosquito data.

The best Models 1 and 2 have in common the maximum rainfall over the last 7 and 14 days and the latitude as relevant variables. Notice that the first model which has the least adjusted  $R^2$  has the least RMSE. In Figure 9, we display the important variables and analyse their performance in each model. From this diagnostic plot, we see that

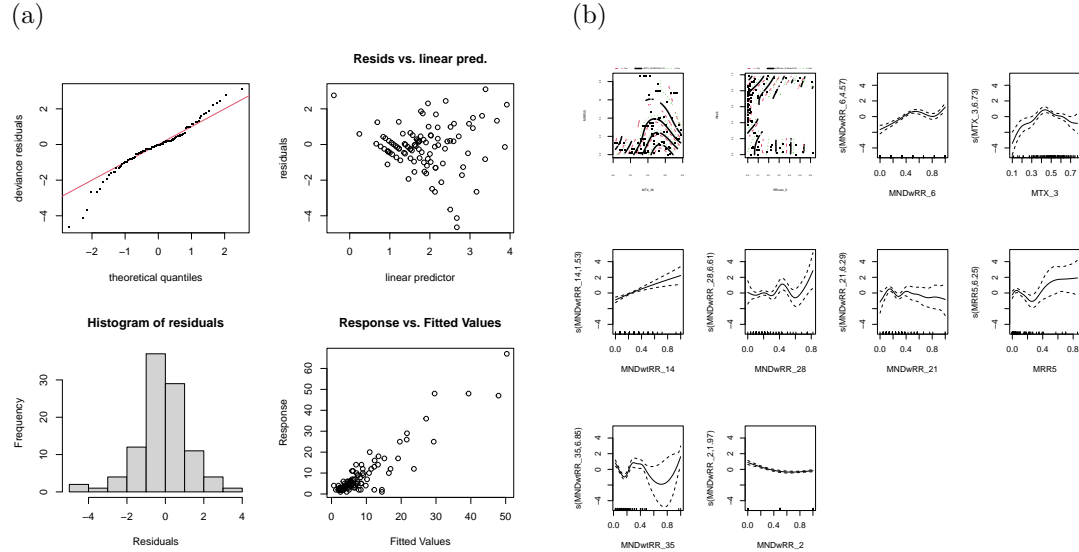


**Fig 9.** Non-linearity check for important variables of Model 1 (a) and Model 2 (b)

rainfall could be linear in the model according to its empirical distribution function. All the other variables have a non-linear contribution to the predictions of the egg counts. The plot of fitted egg counts data (Figure 10) and observed counts seems to be linear. The normality test for residuals failed for Model 1 and 2. The residual of histogram does not show symmetry, which also confirms the non-normality of the residuals. All predictors variables for Model 2 have a non-linear relationship with respect to the response variable (egg counts values). The goodness of fit and the non-linearity check with Model 3 for adult mosquito data is given in Figure 11. The maximum number of days with rain over the last two days seems to be linear.



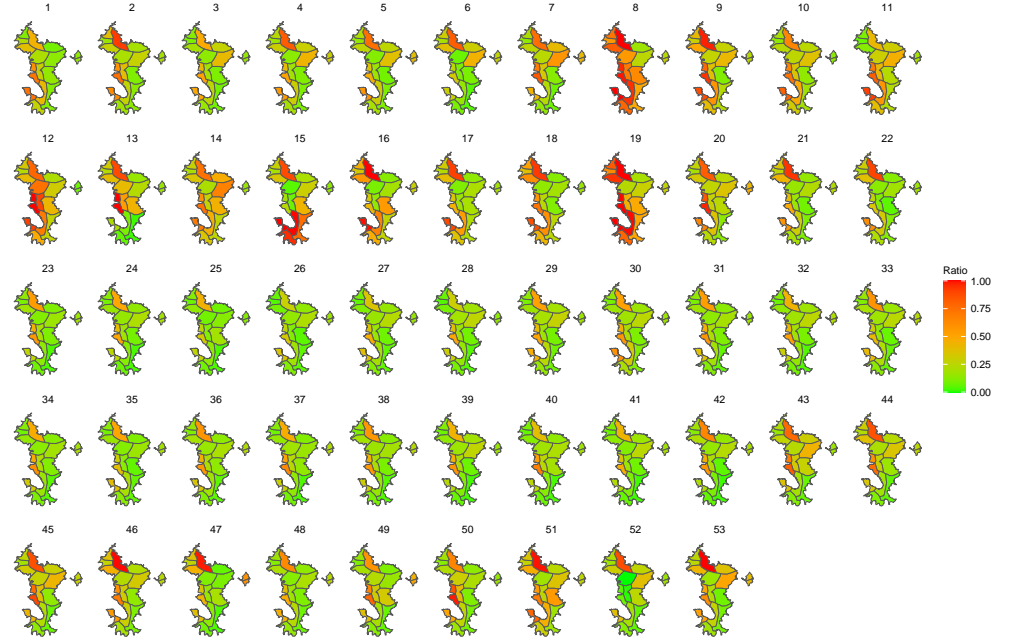
**Fig 10.** Goodness of fit plots for Model 1 (a) and Model 2 (b) with egg mosquito data.



**Fig 11.** Goodness of fit plots and non-linearity check for important variables of Model 3

The predicted spatio-temporal risk ratio of egg abundance is given in Figures 12 and 13 for Model 1 (2019-2020 data) and in Figure 14 for Model 2 (2021 data). The

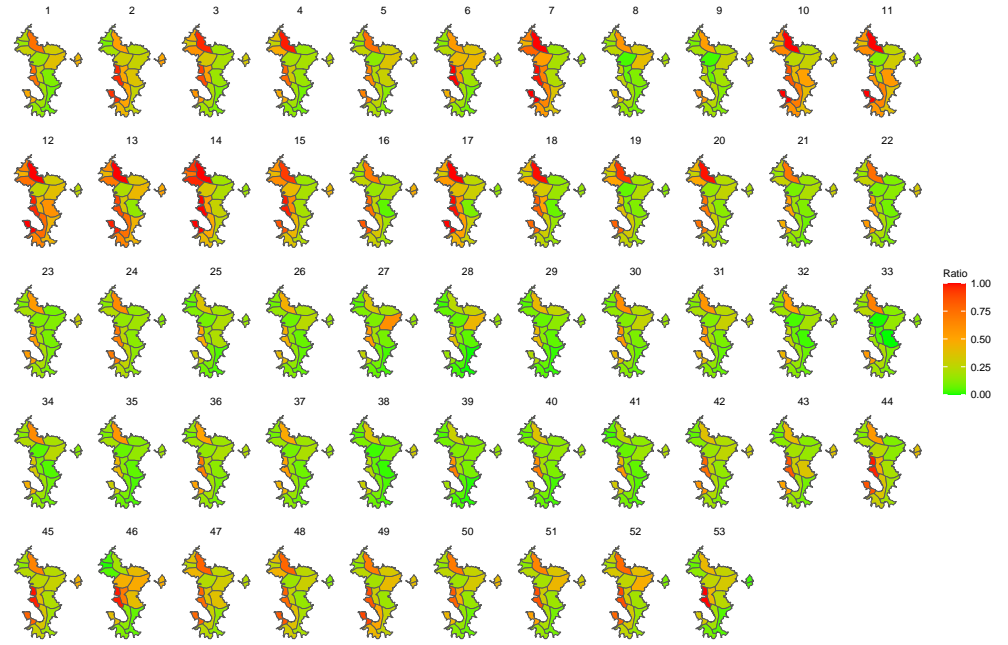
prediction were correlated with the seasonal variations. For instance in 2021, it can be observed that, the models predicts spatio-temporal threshold exceedances of egg abundance with high probability, mainly at the beginning of the year, particularly in weeks 1 to 22, and also towards the end of the year in weeks 41 to 52 or 53 in general. Such period correspond to the austral winter in the island of Mayotte. The same trends were confirmed in 2019 and 2020, but with spatio-temporal risk ratio less than the one in 2021. We also remark a different variability in the climatic data during this period. Thus, the year 2021 recorded more spatio-temporal threshold exceedances. More precisely during the year 2021 the model predict higher risk ratio (whenever the predicted GAM egg counts data is greater than the maximum predicted count eggs values) for most of the communes in the north, south, center and west. On the other hand, weeks 23 to 41/41 (austral summer) shows lower risk ratio trends. Similar results are observed for for larvae and pupae mosquitoes count data in the neighbour island, La Réunion [7].



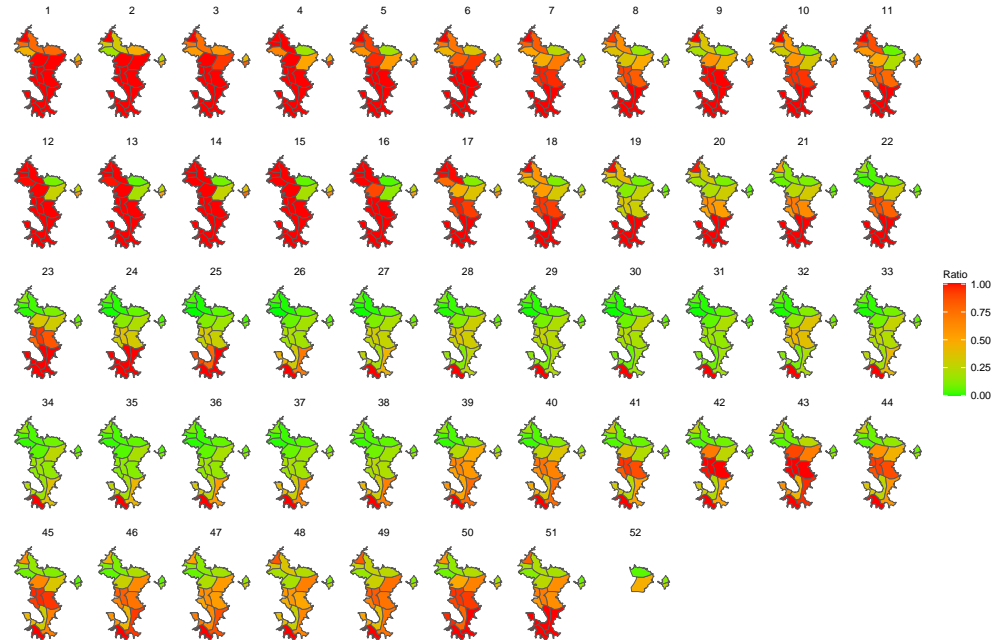
**Fig 12.** Spatio-temporal relatif eggs risk ratio based on GAM predictions for year 2019

Additionally, we apply the above proposed spatio-temporal kernel smoother (IDW) model with the same selected explanatory variables to egg data count. We use a 5-fold cross validation to perform the simulation. The spatio-temporal risk ratio prediction Figures 15, 16, 17 match globally with the prediction with the GAM models (Model 1 and 2) but with a risk ratio less important. The prediction reveals important variability in the last few weeks across the different models in the communes of Ouangani, Mamoudzou and Brandrele. Overall, there seems almost no period with zero probability of egg mosquitoes' abundance, which is consistent with the LAV monitoring team terrain observation.

For adult mosquito count data, we have included the egg count number as a potential explanatory variable. Unfortunately, this did not appear to be relevant. The predicted spatio-temporal risk ratio of adults mosquitoes abundance (Model 3) is given in Figures 18, 19, 20. In the year 2019, the model predict spatio-temporal threshold exceedances and hence a strong risk ratio of adult mosquito abundance in weeks 4 to 13 (middle of the austral winter season) and in weeks 40 to 50-52 (austral winter season).

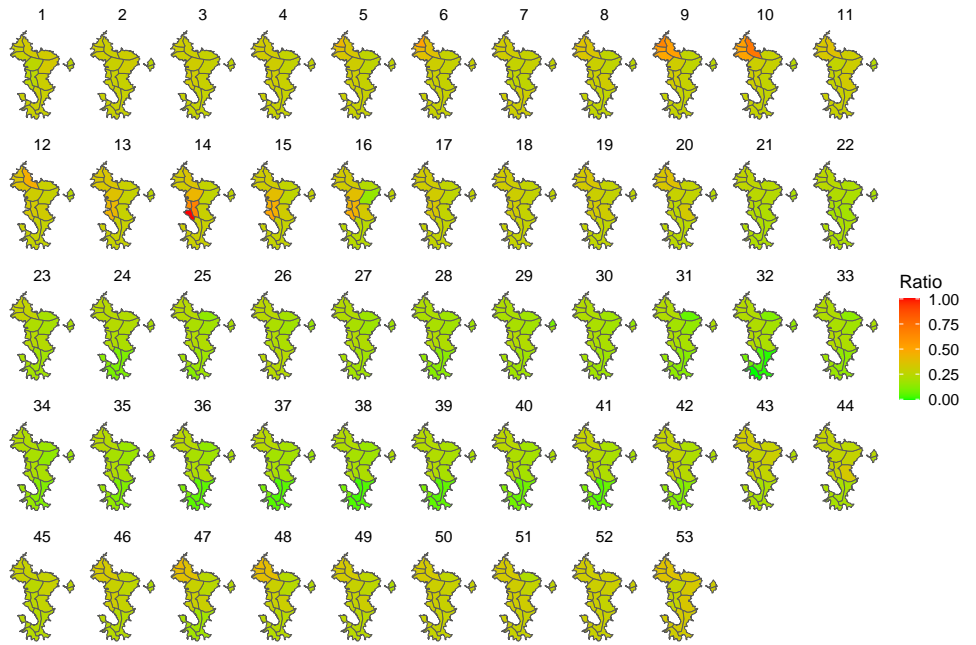


**Fig 13.** Spatio-temporal relatif eggs risk ratio based on GAM predictions for year 2020

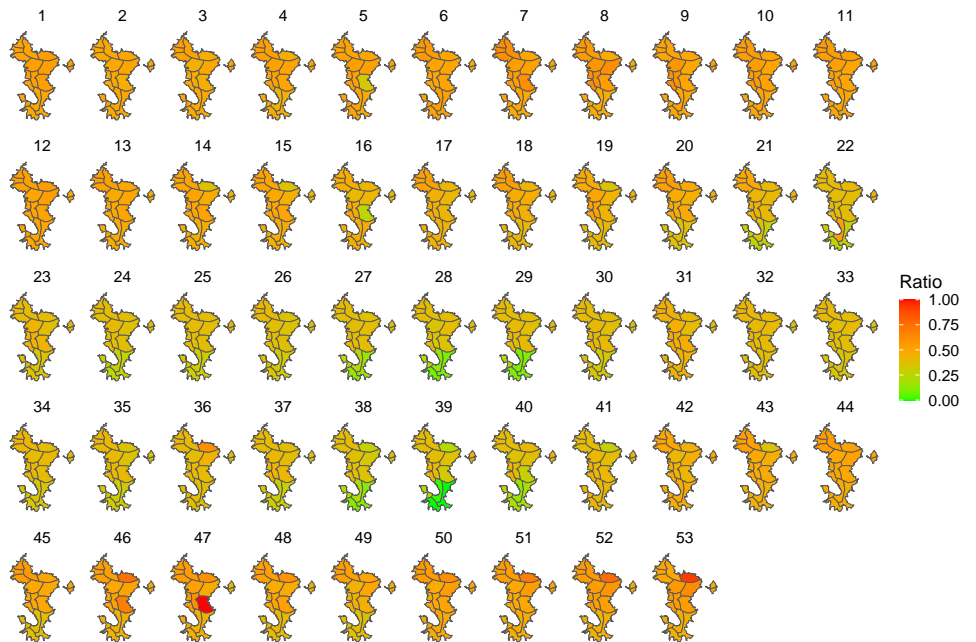


**Fig 14.** Spatio-temporal relatif eggs risk ratio based on GAM predictions for year 2021

Such period is characterised by strong record rainwater and low temperature. The risk ratio was not strong in weeks 17 to 23. We observe that there was a delay at the start of the year comparing to the eggs emergence in the same period, which can be explained by the development cycle between egg and adult stages and also environmental conditions. In 2020, adult mosquitoes risk ratio prediction show that they emerge globally, in week 28 but, early at week 24 (austral summer) in some communes in the



**Fig 15.** IDW predictions for year 2019

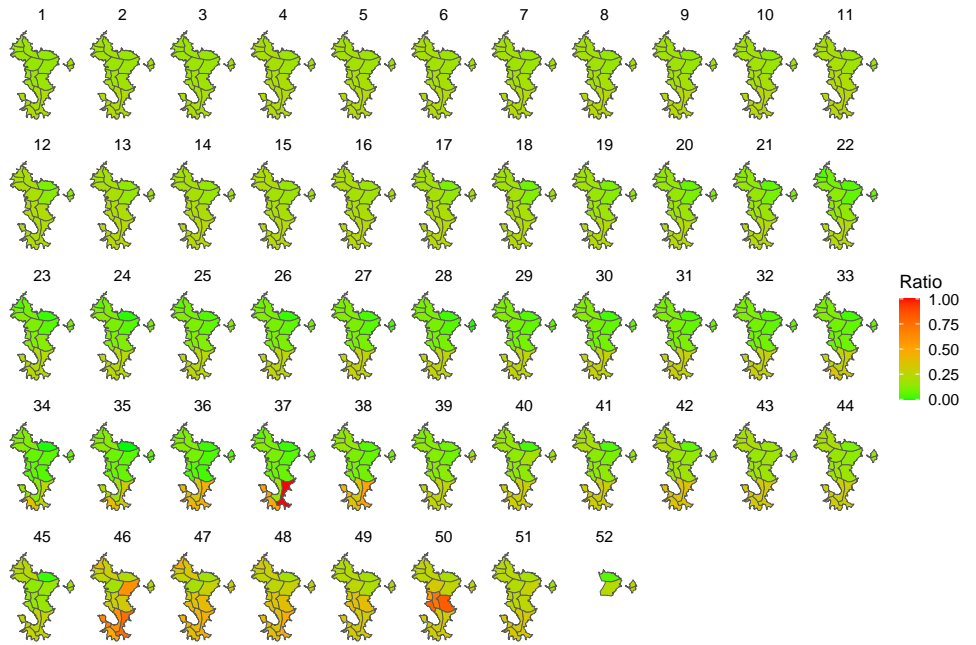


**Fig 16.** IDW predictions for year 2020

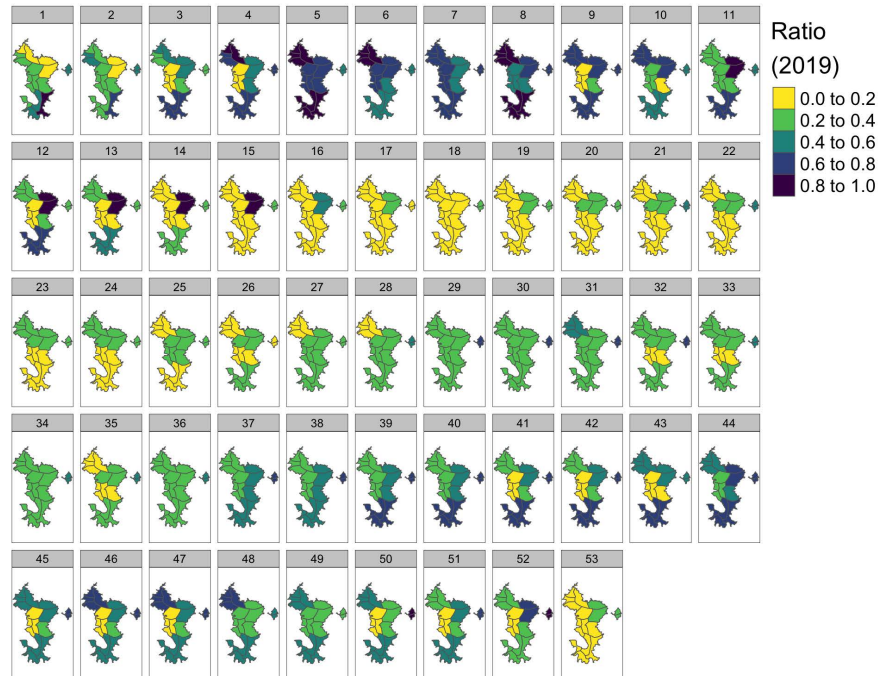
north and the center of the island. Like in 2020, the year 2021 record fairly high risk ratio of adult mosquito abundance in weeks 29 to 52. The above mentioned Figures still show medium relative risk ratio of mosquitoes during the austral summer over all years.

It should be noted that the models predict non-zero relative risk ratio as well as adult and egg mosquito abundance at all times in the island of Mayotte. These different predictive observations results will be used for ground truths during a forthcoming data



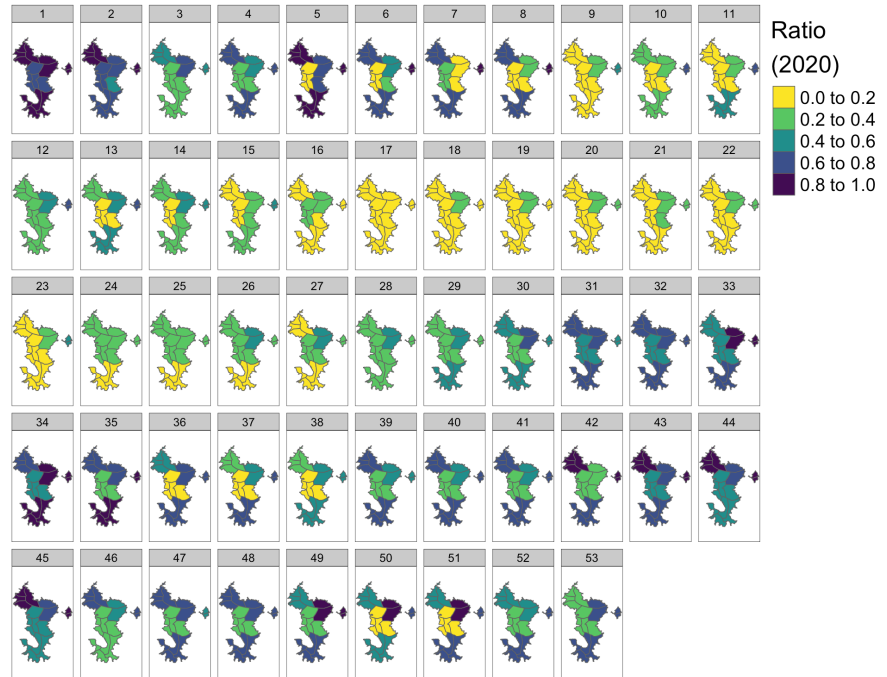


**Fig 17.** IDW predictions for year 2021

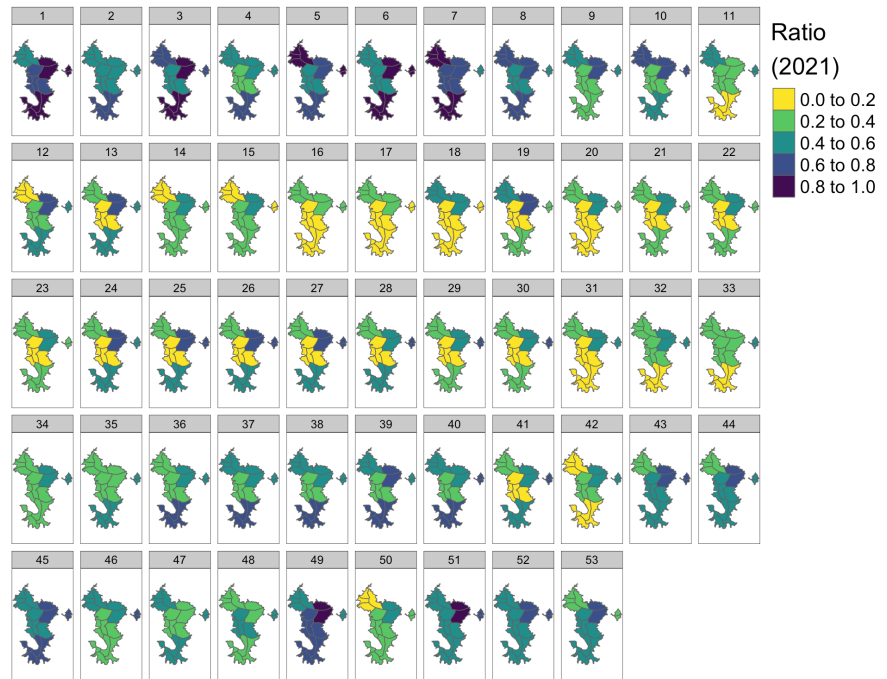


**Fig 18.** Spatio-temporal relatif adult risk ratio based on GAM predictions for year 2019

collection campaign; since the incomplete nature of our data and special observation patterns are clearly a limitation aspect of this work. This will enable us to improve our proposed spatio-temporal modeling approach, in order to built a robust simulation model available online in order to optimise monitoring activities of the LAV team as concerned the areas to be carried out according to abundance risk ratio periods.



**Fig 19.** Spatio-temporal relatif adult risk ratio based on GAM predictions for year 2020



**Fig 20.** Spatio-temporal relatif adult risk ratio based on GAM predictions for year 2021

## Conclusion and perspectives

In this study, several statistical analysis and spatio-temporal modeling tools have been proposed to analyse and model eggs and adults mosquitoes count data in the island of

Mayotte. Using the Lasso regression we highlight a number of explanatory variables (potential predictors) for the abundance of our egg and adult mosquito count data. We propose two spatio-temporal models that can take into account such potential predictors. The first model is a modified and adapted spatio-temporal kernel smoother which reveals to be suitable for our eggs count data. The second model is an application of the Generalized Additive Model in a spatio-temporal setting. We discuss the performance measures of such data process models and introduce a novel spatio-temporal relative risk of abundance that can be seen as a probability of abundance of mosquitoes. We discuss some relevant spatio-temporal predictive observations that can be subject of vector control teams to focus on suitable control measures in the monitoring of the mosquito abundance and therefore the risk of contracting dengue fever. We also discuss about a future R Shiny application called May'Aedes, that aims to be a flexible and efficient tool to produce predictive risk rate ratio for control measures according to a given threshold observation. In terms of forthcoming study, we are currently interested in theoretical development of a spatio-temporal simulation model based on a mechanistic PDE model.

## Supporting information

**S1 Fig. Weekly frequency of egg data collection by site.** This picture shows how many times data were collected in each of the 17 communes of Mayotte.

**S2 Fig. Weekly frequency of egg data collection by site.** This picture shows how many times data were collected in each of the 17 communes of Mayotte.

**S3 Fig. Outlier detection for mosquito egg data.** This picture discuss the issue of outliers in eggs data due to the high frequency of data collection in a few communes.

**S4 Fig. Outlier detection for adult mosquito count data.** This picture discuss the issue of outliers in adult mosquito count data due to the high frequency of data collection in a few communes.

**S5 Fig. Maximum number of consecutive days with rain after imputation.** This picture show the observed days with rain in the island of Mayotte.

**S1 Table. Yearly percentage of egg and adult data observations for all represented communes.**

**S2 Table. Selected relevant variable and their description for eggs count data.**

**S3 Table. Selected relevant variable and their description for adults count data.**

**S4 Table. Explicit GAM models formulas and their performance measures.**

## Data Availability Statement

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

## Acknowledgments

We thank the French Regional Agency (ARS) of Mayotte and the French National Weather Service Meteo-France for providing the needed data (entomological and meteorological data). This work was financially supported by ARS of Mayotte. We certify the the submission is an original work and is not under review at any other publication. The authors also thank Yousri Slaoui and Jean-Noël Bacro (during their visit in the island of Mayotte) for relevant discussions during the preparation of this paper.

## References

1. Indien CO. Dengue et chikungunya à La Réunion et à Mayotte. Bulletin de veille sanitaire. 2010;8.
2. Organization WH. Dengue fever-Mayotte, France. World Health Organ 2020. (accessed April 23, 2021);<https://www.who.int/emergencies/disease-outbreak-news/item/2020-DON265>.
3. Laurent J, Stéphane A, Leïla B. Impacts des changements climatiques sur les arboviroses dans une île tropicale en développement (Mayotte). VertigO-la revue électronique en sciences de l'environnement. 2011;10(3).
4. Anoopkumar A, Puthur S, Varghese P, Rebello S, Aneesh E. Life cycle, bio-ecology and DNA barcoding of mosquitoes *Aedes aegypti* (Linnaeus) and *Aedes albopictus* (Skuse). J Commun Dis. 2017;49(3):32–41.
5. Iyaloo DP, Degenne P, Elahee KB, Seen DL, Bheecarry A, Tran A. ALBOMAURICE: A predictive model for mapping *Aedes albopictus* mosquito populations in Mauritius. SoftwareX. 2021;13:100638.
6. Lebon C, Alout H, Zafihita S, Dehecq JS, Weill M, Tortosa P, et al. Spatio-temporal dynamics of a dieltrin resistance gene in *Aedes albopictus* and *Culex quinquefasciatus* populations from Reunion Island. Journal of Insect Science. 2022;22(3):4.
7. Tran A, Mangeas M, Demarchi M, Roux E, Degenne P, Haramboure M, et al. Complementarity of empirical and process-based approaches to modelling mosquito population dynamics with *Aedes albopictus* as an example—Application to the development of an operational mapping tool of vector populations. PloS one. 2020;15(1):e0227407.
8. Simoy MI, Simoy MV, Canziani GA. The effect of temperature on the population dynamics of *Aedes aegypti*. Ecological modelling. 2015;314:100–110.
9. Barrera R, Amador M, MacKay AJ. Population dynamics of *Aedes aegypti* and dengue as influenced by weather and human behavior in San Juan, Puerto Rico. PLoS neglected tropical diseases. 2011;5(12):e1378.
10. Marina CF, Bond JG, Hernández-Arriaga K, Valle J, Ulloa A, Fernández-Salas I, et al. Population dynamics of *Aedes aegypti* and *Aedes albopictus* in two rural villages in southern Mexico: Baseline data for an evaluation of the sterile insect technique. Insects. 2021;12(1):58.

11. Mussard R. Etat des lieux de la lutte anti-vectorielle à Mayotte et propositions de réduction des risques et impacts de la lutte chimique au profit d'une action intégrée contre les vecteurs;.
12. Pauline R. Mise en place des Moustiquaires Imprégnées d'Insecticide Longue Durée (MIILD) à Mayotte: Premiers éléments relatifs à la pertinence et à la faisabilité locales;.
13. Brady OJ, Johansson MA, Guerra CA, Bhatt S, Golding N, Pigott DM, et al. Modelling adult *Aedes aegypti* and *Aedes albopictus* survival at different temperatures in laboratory and field settings. *Parasites & vectors*. 2013;6(1):1–12.
14. Delatte H, Gimonneau G, Triboire A, Fontenille D. Influence of temperature on immature development, survival, longevity, fecundity, and gonotrophic cycles of *Aedes albopictus*, vector of chikungunya and dengue in the Indian Ocean. *Journal of medical entomology*. 2009;46(1):33–41.
15. Valdez LD, Sibona GJ, Condat C. Impact of rainfall on *Aedes aegypti* populations. *Ecological Modelling*. 2018;385:96–105.
16. Rocha NH, Codeço CT, Alves F, Magalhães MdAFM, Oliveira RLd, et al. Temporal Distribution of *Aedes aegypti* in Different Districts of Rio De Janeiro, Brazil, Measured by Two Types of Traps. 2009;.
17. Allman MJ, Slack AJ, Abello NP, Lin YH, O'neill SL, Robinson AJ, et al. Trash to treasure: how insect protein and waste containers can improve the environmental footprint of mosquito egg releases. *Pathogens*. 2022;11(3):373.
18. Cover T, Hart P. Nearest neighbor pattern classification. *IEEE transactions on information theory*. 1967;13(1):21–27.
19. Tobler W. On the first law of geography: A reply. *Annals of the association of American geographers*. 2004;94(2):304–310.
20. Fix E. Discriminatory analysis: nonparametric discrimination, consistency properties. vol. 1. USAF School of Aviation Medicine; 1985.
21. Spitzer M, Wildenhain J, Rappsilber J, Tyers M. BoxPlotR: a web tool for generation of box plots. *Nature methods*. 2014;11(2):121–122.
22. Bickel PJ, Ritov Y, Tsybakov AB. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of statistics*. 2009;37(4):1705–1732.
23. Wikle CK, Zammit-Mangion A, Cressie N. *Spatio-temporal statistics with R*. CRC Press; 2019.
24. Hastie TJ. Generalized additive models. In: *Statistical models in S*. Routledge; 2017. p. 249–307.
25. Tutz G, Gertheiss J. Regularized regression for categorical data. *Statistical Modelling*. 2016;16(3):161–200.
26. Harrell FE, et al. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. vol. 608. Springer; 2001.
27. Wood SN. *Generalized additive models: an introduction with R*. CRC press; 2017.
28. Sasieni P. Generalized additive models. TJ Hastie and RJ Tibshirani; 1990.

29. Wood S, Wood MS. Package ‘mgcv’. R package version. 2015;1(29):729.
30. Zhang D. A coefficient of determination for generalized linear models. The American Statistician. 2017;71(4):310–316.