Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Bayesian Spatio-Temporal Bell Model for Count Data: Malaria Incidence in the Brazilian Legal Amazon

Moisés Augusto Farias Silva
Federal University of Bahia
Paulo Henrique Ferreira
Federal University of Bahia
Rosemeire Leovigildo Fiaccone
Federal University of Bahia
Madhuchhanda Bhattacharjee
University of Manchester
Nixon Jerez-Lillo
nixon.jerez@usm.cl
Federico Santa María Technical University

Article

Keywords: Bayesian modeling, Bell distribution, malaria, spatio-temporal analysis

Posted Date: June 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9692257/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Bayesian Spatio-Temporal Bell Model for Count Data: Malaria Incidence in the Brazilian Legal Amazon

Moisés Augusto Farias da Silva[1], Paulo Henrique Ferreira[1], Rosemeire Leovigildo Fiaccone[1], Madhuchhanda Bhattacharjee[2], Nixon Jerez-Lillo[3*]

[1]Department of Statistics, Federal University of Bahia, Rua Barão de Jeremoabo, s/n, Ondina, Salvador, 40170-110, Bahia, Brazil.
[2]Department of Mathematics, University of Manchester, Oxford Road, Manchester, M13 9PL, Greater Manchester, United Kingdom.
[3*]Departamento de Matemática, Universidad Técnica Federico Santa María, Av. España 1680, Cerro Placeres, Valparaíso, 2390123, Valparaíso, Chile.

*Corresponding author(s). E-mail(s): nixon.jerez@usm.cl;
Contributing authors: moisesdff8@gmail.com; paulohenri@ufba.br; fiaccone@ufba.br; madhuchhanda.bhattacharjee@manchester.ac.uk;

Abstract

Effective control of epidemics and pandemics relies on the monitoring of reported cases, enabling rapid situational evaluations and timely interventions to confirm or adjust actions. This study proposes a predictive model for the incidence of malaria cases (*Plasmodium vivax* and *Plasmodium falciparum*) per thousand inhabitants across the microregions of the Legal Amazon region in Brazil. The analysis incorporates temporal, spatial, and spatio-temporal structures using data from 2003 to 2018. A key focus of this work is the application of the Bell distribution for Bayesian modeling count data, accounting for spatio-temporal dependencies. Additionally, the performance of the Bell model was compared with established models, including the Negative Binomial and Poisson distributions, using several evaluation metrics such as Mean Bias Error and Relative Absolute Error, among others. The results show that the proposed model effectively identifies microregions and months with increased malaria incidence. In particular, the models accurately forecast the timing and location of

incidence surges for both types of malaria, demonstrating their capability for spatio-temporal prediction. These findings indicate the usefulness of the Bell distribution for modeling spatio-temporal count data and its potential to support public health surveillance.

Keywords: Bayesian modeling, Bell distribution, malaria, spatio-temporal analysis.

1 Introduction

Malaria remains a prevalent disease in tropical and subtropical regions, continuing to pose a significant global public health challenge despite being treatable. In 2020, the World Health Organization (WHO) estimated 241 million cases and 627,000 deaths across 85 countries (Ministério da Saúde, Brasil 2022). In South America, most malaria cases are concentrated in four countries: Brazil, Colombia, Peru, and Venezuela, which represented about 80% of the regional burden between 2016 and 2019 (Barreto et al. 2019). Although major progress has been achieved through control programs and vector management strategies, malaria remains endemic in many rural and forested areas. Factors such as limited access to healthcare, population mobility, and environmental conditions that favor mosquito breeding continue to hinder eradication efforts (de Aguiar Barros et al. 2022; Torres et al. 2022).

In Brazil, the Legal Amazon region, which comprises the states of Acre, Amazonas, Amapá, Maranhão, Mato Grosso, Pará, Rondônia, Roraima, and Tocantins, bears the highest malaria burden (Lapouble et al. 2015). The Brazilian government's Malaria Control Action Intensification Plan - PIACM (in Portuguese: "Plano de Intensificação das Ações de Controle da Malária"), launched in 2000, has been credited with reducing the incidence of malaria in the region, although the impact has varied across states. Amazonas saw the greatest reduction (71%), followed by Acre (67%), while Rondônia and Amapá experienced the smallest declines (9% and 15%, respectively) (Ministério da Saúde, Brasil 2003). The main causative agents are *Plasmodium vivax* and *Plasmodium falciparum*. *Plasmodium vivax* infections account for around 83% of cases in the Amazon region (Garcia et al. 2024).

Ladislau (2006) emphasizes the importance of efficient monitoring and timely adjustment of health data systems to detect changes in disease transmission dynamics and guide interventions. Brazil experienced peak in malaria cases, with over 600,000 reported in 2005. However, the incidence has generally declined since then, with an exception in 2017-2018. In 2016, approximately 130,000 cases were reported, and since 2017, the total number of cases, including those of *Plasmodium falciparum*, has never reached the levels seen in 2016 (Ministério da Saúde, Brasil 2022). Although Brazil saw an approximate 4% decline in new malaria cases in 2021 compared to 2020, the number of people affected remains considerable, underscoring the ongoing need for robust monitoring and control efforts. While control programs have played a key role in curbing transmission, the disease continues to pose a serious threat. Consequently, gaining insight into the interplay between environmental, socioeconomic, and epidemiological factors is essential for the development of effective prevention and intervention

strategies. Therefore, understanding the relationships between environmental, socioeconomic, and disease factors is crucial to developing effective prevention and control strategies.

Monitoring malaria involves not only tracking reported cases, but also assessing their spatial and temporal evolution. This allows for identifying vulnerable microregions within the Legal Amazon, both in terms of current and future risk. Given the spatial nature of many infectious diseases, including malaria, analyzing the spatial distribution of cases and their relationship with socio-environmental factors is essential to detect outbreaks and understand seasonal patterns, supporting more effective control efforts. Recent advances in malaria mapping and fine-resolution modeling underscore the importance of integrating spatial, epidemiological, and environmental data to inform elimination strategies (Alegana et al. 2016). Similarly, improved characterization of malaria seasonality through longitudinal datasets offers key insights into temporal transmission dynamics, strengthening surveillance and intervention planning (Sanna et al. 2025). Moreover, identifying clusters of epidemic municipalities can help prioritize areas for malaria control in the Brazilian Amazon, and time series analysis can forecast future malaria cases, aiding in the development of targeted control strategies in these regions.

The primary objective of this study is to model and predict malaria incidence rates in the Legal Amazon region, focusing on monthly monitoring and forecasting of reported cases per thousand inhabitants. A key methodological innovation of this work is the use of the Bell distribution, proposed by Castellares et al. (2018), to account for the spatio-temporal variation in malaria incidence.

The Bell distribution has several noteworthy properties that make it a promising alternative for modeling count data in epidemiological contexts. Similar to the Negative Binomial distribution, it is capable of handling overdispersion, a common feature in disease surveillance data where the variance exceeds the mean. This behavior has been observed, for instance, by Sena et al. (2022), who reported substantial overdispersion in malaria incidence data from comparable contexts.

The performance of the Bell model is compared with well-established approaches, such as the Negative Binomial and Poisson distributions, to predict the number of cases per hundred thousand inhabitants at the microregional level. Unlike the Negative Binomial, the Bell distribution handles overdispersion without requiring an additional dispersion parameter, which enhances parsimony and computational efficiency. At the same time, it preserves a natural connection with the Poisson model, maintaining a uniparametric structure and interpretability, while allowing for a broader range of variance-to-mean relationships. This combination of flexibility, simplicity, and interpretability suggests that the Bell model may provide a robust framework for predicting malaria incidence even under heterogeneous transmission dynamics across space and time. As noted by Castellares et al. (2018), these properties make the Bell distribution a particularly appealing choice for epidemiological modeling.

The paper is structured as follows. Section 2 describes the data used in the analysis, including its sources, temporal and spatial coverage. Section 3 describes the use of the Bell distribution for modeling spatio-temporal malaria incidence, along with the benchmark Negative Binomial and Poisson models, and the criteria employed to

assess model performance. Section 4 presents a simulation study designed to examine the behavior of the proposed methodology. Section 5 presents the findings, including model fit evaluations and predictive performance comparisons, highlighting the ability of the Bell model to capture overdispersion and heterogeneity across microregions. Finally, Section 6 reviews the main findings, highlights their implications for malaria surveillance, and identifies avenues for future research.

2 The Data

In this study, two databases were used. The first was obtained through SIVEP - Malaria (in Portuguese: "Sistema de Informação da Vigilância Epidemiológica"), consisting of 5,972,715 individual records of confirmed cases of malaria (*Plasmodium falciparum*, *Plasmodium vivax* or others) in their respective locations (microregions belonging to the states of the Legal Amazon) and period (between 2003 and 2018). For the purposes of this analysis, case counts were summed across municipalities to obtain the total number of cases per microregion.

The second database is similar to the one considered by Sena et al. (2022) and concerns the average monthly humidity and the average monthly temperature of the microregions. These data were obtained through the National Oceanic and Atmospheric Administration (NOOA), given the established association between climatic factors and malaria transmission (Prothero 1999). Daily and municipal measurements were first aggregated at the monthly level (mean) and then at the microregion level (also mean). Exclusion criteria included periodic control records after treatment (1,041,397 cases), inconsistent date information (10,436 lines), data without complete information (957,879 records), other types of malaria or mixed types of malaria (44,920 cases). Therefore, 3,918,083 individual records were analyzed, with 3,232,766 observations with *Plasmodium vivax* infection and 685,317 with *Plasmodium falciparum* infection.

3 Methodology

This section presents the methodological framework of the study. First, we introduce the Bell model. Then, we describe the spatio-temporal models considered in the analysis, along with the selection criteria and evaluation metrics used to compare different model specifications.

3.1 Bell Model

The Bell distribution is a one-parameter model suitable for discrete random variables (Castellares et al. 2018). In contrast to the Poisson distribution, the Bell distribution does not impose the constraint of equal mean and variance. Similarly to the Negative Binomial, the Bell distribution is suitable for situations with overdispersion (i.e., variance $>$ mean), but it does so more parsimoniously. Suppose Y is a random variable following the Bell distribution. The probability mass function (PMF) of Y is expressed

as:

$$p(y \mid \theta) = \frac{\theta^y e^{-e^\theta+1} B_y}{y!}, \quad y \in \mathbb{N},\, \theta > 0,$$

where B_y is the Bell number (Bell 1934a,b), defined as:

$$B_y = \frac{1}{e} \sum_{k=0}^{\infty} \frac{k^y}{k!}, \quad \text{with } B_0 = B_1 = 1.$$

Some notable properties of the Bell distribution are as follows (Castellares et al. 2018):

1. The Bell distribution is strongly unimodal;
2. It is identifiable for all $\theta > 0$ and is infinitely divisible;
3. If $X_1, X_2, \ldots, X_N$ are independent and identically distributed (IID) random variables where X_i follows a zero-truncated Poisson distribution with parameter $\theta > 0$, and if $Y = X_1 + X_2 + \ldots + X_N$, with $N \sim \text{Poisson}(e^\theta - 1)$, independent of the X_i's, then $Y \sim \text{Bell}(\theta)$.

From a frequentist perspective, the model parameters can be estimated using the maximum likelihood (ML) method, which is well known for its desirable asymptotic properties (consistency, efficiency, and asymptotic normality). This method consists in finding the parameter values that maximize the log-likelihood function based on a random sample $\mathbf{y} = (y_1, \ldots, y_n)$, which is defined by:

$$\ell(\theta \mid \mathbf{y}) = \log(\theta) \sum_{i=1}^{n} y_i - n e^\theta + n \log\left(\frac{B_y}{y!}\right) + n. \tag{1}$$

Maximizing (1) with respect to θ yields the ML estimator:

$$\overline{Y} = \hat{\theta} e^{\hat{\theta}} \implies \hat{\theta} = W_0(\overline{Y}),$$

where $\overline{Y}$ is the sample mean, and $W_0(\cdot)$ denotes the Lambert W function (Corless et al. 1996).

The Bell distribution, similarly to the Poisson distribution, belongs to the exponential family of distributions. This property allows it to be naturally incorporated within the framework of generalized linear models (GLMs) (McCullagh and Nelder 2019). To facilitate such inclusion, it is convenient to express the distribution in terms of its mean parameter. Specifically, by defining:

$$\mu = \theta e^\theta \implies \theta = W_0(\mu).$$

Under this parametrization, the PMF of the Bell distribution can be written as (Castellares et al. 2018):

$$p(y \mid \mu) = \exp\left\{1 - e^{W_0(\mu)}\right\} \frac{[W_0(\mu)]^y B_y}{y!}, \quad y \in \mathbb{N},\, \mu > 0.$$

The corresponding mean and variance of the distribution are given by

$$\mathbb{E}[Y] = \mu \quad \text{and} \quad \mathbb{V}\text{ar}[Y] = \mu\,[1 + W_0(\mu)]\,,$$

which show us the variance is always larger than the mean, with the degree of overdispersion determined by the value of $W_0(\mu)$. Within the GLM framework, we can link the covariates to the mean of the response variable through a logarithmic link function:

$$\log(\mu) = \mathbf{x}^\top \boldsymbol{\beta},$$

where $\mathbf{x}$ denotes the vector of covariates and $\boldsymbol{\beta}$ the corresponding vector of regression coefficients.

3.2 Proposed Models

In general, when working with data modeling, it is an interesting practice to consider information that can influence the phenomenon being evaluated. This is commonly done through covariates incorporated into a regression model. In the literature, it is possible to find several approaches to modeling phenomena that involve counts. In particular, in the case of counts of a given disease, several studies consider temporal and/or spatial dimensions (Achcar et al. 2011; Sena et al. 2022; Li and Dey 2022; Ayala et al. 2022).

In this study, we analyze malaria incidence rates as the response variable of interest. For each microregion i and period j, the incidence rate is defined as the ratio between the number of reported malaria cases and the corresponding population size, and is expressed per thousand inhabitants as $1{,}000 \times y_{i,j}/\text{population}_{i,j}$, with predicted values $1{,}000 \times \hat{y}_{i,j}/\text{population}_{i,j}$.

Besides the Bell model, two additional response distributions are considered (see A) for modeling the number of malaria cases across microregions and time periods:

$$Y_{i,j} \mid \mu_{i,j} \sim \text{Bell}(\mu_{i,j}), \quad Y_{i,j} \mid \mu_{i,j} \sim \text{Poisson}(\mu_{i,j}), \quad Y_{i,j} \mid \mu_{i,j} \sim \text{NB2}(\mu_{i,j}, \phi_{i,j}),$$

where $\mu_{i,j}$ denotes the expected number of malaria cases in microregion i and period j, and NB2 represents the Negative Binomial type 2 distribution (Gardner et al. 1995). For each distributional assumption, four model specifications are explored, which differ in how spatial and temporal dependence structures are incorporated. Model fitting is performed using the Integrated Nested Laplace Approximation (INLA) method, implemented in the `R-INLA` package (Lindgren and Rue 2015).

Specification 1

The first model spectification is largely based on the work of Ayala et al. (2022), with malaria cases modeled according to the following functional form:

$$\log\left(\mu_{i,j}\right) = \text{offset}\left(\log\left(\text{population}_{i,j}\right)\right) + \gamma_k + \iota_l + \nu_i + \delta_j, \tag{2}$$

where:

- γ_k, for $k \in \{1, 2, \ldots, 12\}$, represents a monthly effect modeled as a second-order random walk, as described in Lindgren and Rue (2008);
- ι_l, for $l \in \{1, 2, \ldots, 16\}$, represents an annual effect modeled as a first-order random walk;
- ν_i, for $i \in \{1, 2, \ldots, 107\}$, is a spatial effect modeled using a reparameterized Besag-York-Mollie (BYM2) specification, as detailed in Simpson et al. (2017);
- δ_j, for $j \in \{1, 2, \ldots, 192\}$, represents a temporal effect modeled as a first-order random walk for each time period in the historical series, where $j = 1$ corresponds to January 2003 and $j = 192$ corresponds to December 2018.

Predictions for this and the other models discussed in this paper were obtained by computing the mode of the posterior predictive distribution. For priors, vague log-Gamma$(0, 0.00005)$ priors were specified for the unstructured precision of the first- and second-order random walks, and a log-Gamma$(0, 0.005)$ prior was assigned for the unstructured precision, as well as for the log-transformed precision of the spatial effect BYM2. These prior specifications were chosen according to the standards set by the `R-INLA` package (Rue et al. 2009).

Specification 2

The second model builds upon the first specification by incorporating two meteorological covariates: humidity and temperature. These variables were previously analyzed by Sena et al. (2022) and are included here to explicitly account for environmental variation across both space and time. Formally, the model is given by:

$$\log(\mu_{i,j}) = \text{offset}\left(\log\left(\text{population}_{i,j}\right)\right) + \gamma_k + \iota_l + \nu_i + \delta_j + \beta_1 X_{1,i,j} + \beta_2 X_{2,i,j},$$

where $X_{1,i,j}$ and $X_{2,i,j}$ denote the average relative humidity and mean temperature, respectively, for municipality i and month j. The coefficients β_1 and β_2 quantify the marginal effects of these covariates, for which weakly informative priors were specified as $\mathcal{N}(0, 10^6)$.

Specification 3

The third model extends the first specification by incorporating a spatio-temporal interaction term that captures the joint variability between space and time. This interaction allows for spatially structured deviations that evolve over time, thereby improving the model's ability to represent local temporal dynamics in malaria incidence. The model is expressed as:

$$\log(\mu_{i,j}) = \text{offset}\left(\log\left(\text{population}_{i,j}\right)\right) + \gamma_k + \iota_l + \nu_i + \delta_j + (\nu_i \times \delta_j), \tag{3}$$

where $(\nu_i \times \delta_j)$ denotes the spatio-temporal interaction between the spatial effect ν_i and the temporal effect δ_j. Following Blangiardo et al. (2013), this term is modeled as $\xi_{i,j} \sim \mathcal{N}(0, 1/\tau_{\nu\times\delta})$, where $\tau_{\nu\times\delta}$ is the precision parameter for the interaction term.

Specification 4

The fourth model extends last specification by including the meteorological covariates considered in second specification, namely average relative humidity and mean temperature. Thus, it combines the spatiotemporal interaction structure with the climatic covariates, allowing for both environmental and space–time dependencies. The model is defined as:

$$\log(\mu_{i,j}) = \text{offset}\left(\log\left(\text{population}_{i,j}\right)\right) + \gamma_k + \iota_l + \nu_i + \delta_j + (\nu_i \times \delta_j) + \beta_1 X_{1,i,j} + \beta_2 X_{2,i,j}. \tag{4}$$

The coefficients β_1 and β_2 are assigned weakly informative priors, $\mathcal{N}(0, 10^6)$, while all other latent components and prior specifications remain as previously described.

3.3 Model Comparison Criteria

Model selection was performed in two stages. First, for each distribution evaluated, a model was selected based on the Deviance Information Criterion (DIC) and the Widely Applicable Information Criterion (WAIC), as defined by Watanabe and Opper (2010).

The DIC is given by:

$$\text{DIC} = 2\bar{D} - D(\hat{\boldsymbol{\varphi}}), \qquad D(\boldsymbol{\varphi}) = -2\sum_{i,j} \log p(y_{i,j} \mid \mu_{i,j}(\boldsymbol{\varphi}))\,,$$

where $\bar{D}$ is the posterior mean of $D(\boldsymbol{\varphi})$ and $D(\hat{\boldsymbol{\varphi}})$ is the deviance evaluated at the posterior mean $\hat{\boldsymbol{\varphi}}$.

The WAIC is defined as:

$$\text{WAIC} = -2\,(\text{lppd} - p_{\text{WAIC}})\,,$$

where

$$\text{lppd} = \sum_{i,j} \log\left(\frac{1}{S}\sum_{s=1}^{S} p\Big(y_{i,j} \mid \mu_{i,j}(\boldsymbol{\varphi}^{(s)})\Big)\right),$$

and

$$p_{\text{WAIC}} = \sum_{i,j} \text{Var}\Big[\log p\Big(y_{i,j} \mid \mu_{i,j}(\boldsymbol{\varphi}^{(s)})\Big)\Big]\,,$$

where $\{\boldsymbol{\varphi}^{(s)}\}_{s=1}^{S}$ denote the posterior samples of the parameter vector $\boldsymbol{\varphi}$.

These criteria assess predictive performance while accounting for model complexity through different formulations of the effective number of parameters. Preference was given to WAIC, as it has been shown to offer advantages over DIC, particularly under reparameterization (Spiegelhalter et al. 2014).

In the second stage, the best models from each distribution were compared using strictly predictive metrics, namely, the Mean Bias Error (MBE), Relative Absolute Error (RAE), Normalized Root Mean Squared Error (NRMSE), Root Mean Squared Logarithmic Error (RMSLE), and Residual Standard Error (RSE). This comparison

was based on a set of tests, including the incidences of all microregions in the Legal Amazon during the months following 2015 (excluding 2015 itself). The model that performed best on these criteria was selected. The evaluation metrics used for this purpose are defined below, as discussed in the book by Cichosz (2014).

3.3.1 Evaluation Metrics

Let y_m denote the observed incidence, and $\hat{y}_m$ the corresponding model prediction, of *vivax* or *falciparum* malaria cases per thousand inhabitants for observation $m = 1, \ldots, M$, where $M = 107 \times 36$ corresponds to all combinations of microregions and time points in the test dataset (i.e., from 2016 onwards).

Table 1 presents commonly used evaluation metrics in predictive modeling, together with their formulas and brief descriptions. These measures capture different aspects of accuracy, scale, and relative performance, providing a broader perspective for evaluating model performance.

Table 1 Summary of the error metrics used to evaluate predictive performance.

Metric	Formula	Description
MBE	$\frac{1}{M}\sum_{m=1}^{M}(y_m - \hat{y}_m)$	Measures the average prediction bias; positive values indicate overestimation and negative values indicate underestimation.
RAE	$\frac{\sum_{m=1}^{M}\lvert y_m - \hat{y}_m\rvert}{\sum_{m=1}^{M}\lvert y_m - \bar{y}\rvert}$	Compares absolute prediction errors relative to deviations from the mean; values near zero indicate better performance.
NRMSE	$\frac{\sqrt{\frac{1}{M}\sum_{m=1}^{M}(y_m - \hat{y}_m)^2}}{\max(y_m)}$	Normalized root MSE, making the error scale-independent; sensitive to large deviations.
RMSLE	$\sqrt{\frac{1}{M}\sum_{m=1}^{M}[\log(y_m + 1) - \log(\hat{y}_m + 1)]^2}$	Error in logarithmic scale; penalizes underestimations more heavily and is suitable for count data.
RSE	$\frac{\sum_{m=1}^{M}(y_m - \hat{y}_m)^2}{\sum_{m=1}^{M}(y_m - \bar{y})^2}$	Scale-independent squared error; lower values indicate greater predictive accuracy.

A global view of the metrics shows that each measure captures a different aspect of predictive performance. The bias-based indicators, such as MBE, reflect the direction and magnitude of systematic errors, while scale-adjusted measures like NRMSE and RSE provide a sense of overall accuracy relative to the variability of the data. The ratio-based metrics, including RAE, emphasize how the model compares with simple reference predictions, and RMSLE helps detect whether the model tends to underpredict or overpredict on a multiplicative scale.

4 Simulation Study

This section presents a Monte Carlo simulation study designed to examine the behavior of the proposed methodology under controlled scenarios. This is an important aspect that has been omitted in several similar articles, but it is necessary for conducting a systematic assessment of the empirical performance of the estimators and analyzes how the evaluation metrics evolve as the sample size increases.

Our study has two main objectives. First, it aims to evaluate how well the spatio-temporal model recovers the fixed effects when the data follow a Bell distribution, one of the response specifications considered in our work as an alternative to the Poisson and Negative Binomial models. Second, it seeks to analyze the behavior of the evaluation metrics as the length of the temporal series increases, providing insight into how the model's predictive performance improves with additional temporal information.

For this purpose, we generated $R = 1{,}000$ independent datasets for each considered duration. Specifically, we created 1,000 replicates spanning 2 years, 1,000 replicates spanning 4 years, and similarly for 6, 8, and 10 years. Each replicate contains monthly observations for each spatial unit. The spatial structure was based on the administrative map of the state of Amazonas, providing a realistic network of neighboring regions for the spatial random effects. To ensure comparability across areas, the population size was kept constant, so that the relative scale of the incidence rates remained consistent.

The counts were generated according to the following specification:

$$\begin{aligned} Y_{i,k} &\sim \text{Bell}(\theta_{i,k} = W_0(\mu_{i,k})), \\ \log(\mu_{i,k}) &= \nu_i + \delta_k + 1.5\, X_{1,i,k} - 2.5\, X_{2,i,k}, \end{aligned}$$

where ν_i represents the spatial effect for area $i \in \{1, 2, \ldots, 107\}$, modeled using the BYM2 reparameterization; δ_k denotes a first-order random walk capturing the temporal structure across the k-th month, $k \in \{1, 2, \ldots, 12\}$; and X_1 and X_2 are covariates drawn from a uniform distribution on $(0, 1)$.

For each of the R replicates, we fitted the proposed model using INLA with its default prior distributions: the fixed effects were assigned Gaussian priors, $(\beta_1, \beta_2) \overset{\text{iid}}{\sim} \mathcal{N}(0, 10^3)$, and the precision of the monthly effect was given a Gamma prior, $\tau \sim \text{Gamma}(1, 5 \times 10^{-5})$. The mean and standard deviation of the posterior modes across replicates are reported in Table 2.

On average, the posterior modes remain close to the true coefficient values, 1.5 and -2.5, for all time horizons considered. Moreover, as the number of years increases, the variability of the posterior modes decreases, reflecting the improved stability and precision afforded by longer temporal series. These results demonstrate that the proposed spatio-temporal model is capable of accurately recovering the fixed effects when the response follows a Bell distribution.

Table 3 reports the evaluation metrics across different time horizons. MBE and RAE decrease as the temporal sample size increases, indicating reduced bias, while NRMSE and RMSLE remain relatively stable, reflecting their limited sensitivity to the

Table 2 Mean and standard deviation of the posterior mode estimates of β_1 and β_2 across simulation replicates.

n_{year}	mean($\hat{\beta}_1$)	sd($\hat{\beta}_1$)	mean($\hat{\beta}_2$)	sd($\hat{\beta}_2$)
2	1.502	0.114	-2.505	0.126
4	1.500	0.081	-2.497	0.087
6	1.503	0.063	-2.500	0.064
8	1.503	0.056	-2.498	0.059
10	1.498	0.049	-2.503	0.050

length of the time series. In addition, RSE generally decreases with longer temporal horizons, indicating improved variability control.

Table 3 Mean performance metrics (with standard deviations in parentheses) for predicted values.

n_{year}	MBE	RAE	NRMSE	RMSLE	RSE
2	0.085 (0.005)	0.060 (0.027)	0.751 (0.174)	0.480 (0.016)	0.560 (0.294)
4	0.046 (0.003)	0.051 (0.022)	0.730 (0.171)	0.489 (0.014)	0.518 (0.273)
6	0.032 (0.002)	0.050 (0.021)	0.739 (0.163)	0.492 (0.013)	0.529 (0.256)
8	0.024 (0.002)	0.050 (0.020)	0.746 (0.151)	0.495 (0.013)	0.543 (0.246)
10	0.020 (0.001)	0.046 (0.020)	0.728 (0.153)	0.495 (0.013)	0.505 (0.250)

Overall, the results indicate that the model gains precision with longer time series, supporting its suitability for both inference and prediction in Bell-distributed spatio-temporal incidence modeling.

5 Results of the Spatio-Temporal Malaria Analysis

This section presents the main findings of the study based on a spatio-temporal analysis of malaria incidence in the Legal Amazon region. As previously mentioned in Section 2, the results rely on two complementary data sources: epidemiological records from the SIVEP–Malaria surveillance system, covering malaria cases observed between 2003 and 2018, with infections predominantly caused by *Plasmodium vivax*, and climate information obtained from the NOAA database.

Figure 1 shows the average incidence of *Plasmodium vivax* malaria cases per thousand inhabitants in the Legal Amazon region over the study period.

The data reveal a clear seasonal pattern in malaria incidence. After an initial increasing trend, the incidence significantly declined until 2015. However, by late 2017, the region again recorded more than one case per thousand inhabitants. This temporal pattern supports the inclusion of time-dependent structures in the models presented in this study.

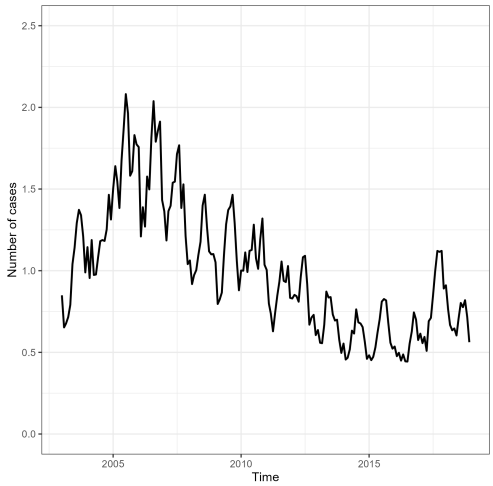


Fig. 1 Incidence of *Plasmodium vivax* malaria per thousand inhabitants in the Legal Amazon region.
Source: SIVEP - Malaria

Figure 2 illustrates the spatial distribution of malaria cases in November 2017. The clustering of high-incidence microregions suggests a spatial dependence, reinforcing the importance of incorporating spatial structures into the modeling approach.

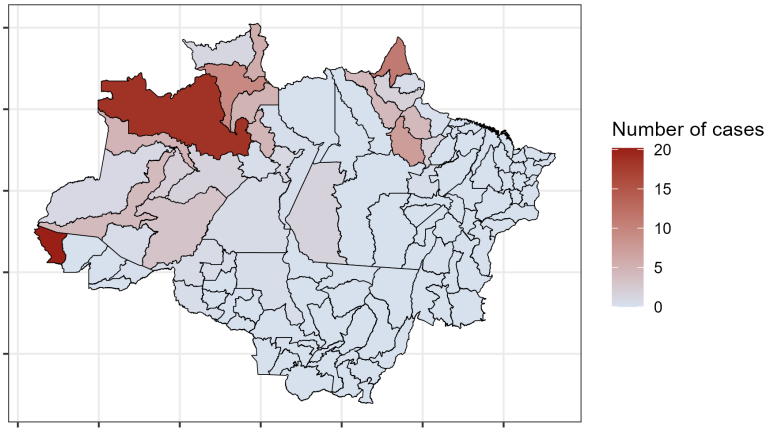


Fig. 2 Incidence of *Plasmodium vivax* malaria per thousand inhabitants in the Legal Amazon region in November 2017.
Source: SIVEP - Malaria

Taking Figures 1 and 2 as reference, two key characteristics are desirable for the predictions of the proposed model, particularly for the test data:

1. Positive correlation: Predictions should be higher in locations (microregions) and periods (historical series points) where real incidences are higher.

2. Magnitude consistency: The predicted values should approximate the order of magnitude of the actual incidence rates.

Achieving the second criterion is particularly challenging for malaria incidence data spanning a historical series, especially when training data is restricted to the year 2015, as in this study. Until 2015, malaria cases exhibited a decreasing trend, making it unlikely that a model trained up to this point would anticipate a reversal in this trend unless explicitly informed by exogenous factors.

Thus, even if the model cannot precisely predict the magnitude of malaria incidence (which may fluctuate due to external influences such as the implementation of a new disease control plan), it is crucial that it correctly identifies when and where incidence rates will rise. As will be demonstrated, even if criterion 2 is not fully met, criterion 1 — combined with a careful error analysis — allows for a reasonable estimation of the order of magnitude of incidence rates. However, without satisfying the first criterion, resource allocation to mitigate potential epidemics may be compromised.

Before assessing whether the selected model meets these desired characteristics, it is useful — both for a methodological discussion of the Bell distribution in spatiotemporal modeling and for model evaluation — to present intermediate results. Specifically, we report the DIC and WAIC values for each proposed model and the predictive metrics for the selected model under each considered distribution.

5.1 Model Selection

As previously discussed, four functional forms were fitted for each of the distributions considered for the different types of malaria. The following tables present the model selection results in terms of DIC and WAIC, which were used to determine the best model for each case: $Y_{ij} \sim$ Bell, $Y_{ij} \sim$ Poisson, and $Y_{ij} \sim$ NB2, respectively.

Table 4 DIC and WAIC of fitted models assuming the Bell distribution.

P. vivax		
Proposal	DIC	WAIC
Model 1	233,902.10	1,018,134.63
Model 2	2,342,679.50	1,028,955.96
Model 3	**111,270.15**	**109,604.89**
Model 4	111,385.85	109,743.70

P. falciparum		
Proposal	DIC	WAIC
Model 1	136,848.26	331,275.93
Model 2	137,073.47	324,996.72
Model 3	**72,939.07**	**72,108.23**
Model 4	73,069.97	72,275.74

In Table 4, the DIC and WAIC values (low is better) are displayed for both *Plasmodium vivax* and *Plasmodium falciparum* cases. The selected model for the Bell distribution is highlighted with bold. In this case, Model 3 was chosen for both malaria types.

A key takeaway from this selection is that Model 3 includes an interaction term between the spatial effect ν_i and the temporal effect δ_j, as defined in Equation (3). Compared to Equation (2), which lacks this interaction, the inclusion of a spatiotemporal effect significantly improved both DIC and WAIC metrics. Interestingly, none

of the three previous studies reviewed incorporated such an interaction: Sena et al. (2022) modeled only the temporal dimension, while Achcar et al. (2011) and Ayala et al. (2022) handled spatial and temporal dimensions separately. Additionally, Model 3 excludes temperature and humidity covariates, indicating that these environmental variables may not improve predictive accuracy. While Prothero (1999) suggested that including environmental covariates would improve model fit, Sena et al. (2022) found that these variables were not strong predictors of malaria incidence.

The difference in the magnitude of the metrics is small, and it is worth mentioning that temperature and humidity are critical environmental factors for malaria transmission. Therefore, we cannot simply focus on predictive indicators for model selection within the scope of applied statistics. Consequently, the choice involves a trade-off between a model that fits the real data well and one that appropriately contextualizes the topic at hand.

Table 5 shows the DIC and WAIC for the adjustments that assumed a Poisson distribution for the number of malaria cases. A striking aspect of this table is the negative DIC for models without a spatiotemporal effect. While rare in the literature, Pan (2016) demonstrated via simulation that small variance values can yield negative DICs, even for well-specified models. Due to this, for *Plasmodium vivax* cases, model selection was based exclusively on WAIC. Once again, Model 3 (Equation (3)) was chosen for *Plasmodium vivax* malaria type. Whereas Model 4 (Equation (4)) was chosen for *Plasmodium falciparum*.

Table 5 DIC and WAIC of fitted models assuming the Poisson distribution.

P. vivax		
Proposal	DIC	WAIC
Model 1	-160,471.11	8,686,732.13
Model 2	-161,565.69	8,679,052.18
Model 3	**97,281.61**	**96,032.41**
Model 4	97,396.74	96,196.44

P. falciparum		
Proposal	DIC	WAIC
Model 1	97,856.71	1,841,198.06
Model 2	99,080.71	1,825,179.50
Model 3	65,829.69	66,950.01
Model 4	**64,967.01**	**65,353.24**

Unlike the Bell and Poisson cases, the best model for the NB2 distribution was Model 4 for both malaria types (Table 6).

Thus, from this point onward, the best model specified when $Y_{ij} \sim \text{Bell}(\mu_{ij})$ will be referred to as the Bell model (of course, we will specify whether we are referring to models for *Plasmodium vivax* or *Plasmodium falciparum* cases); similarly, the Poisson model and Negative Binomial model are defined. As previously stated, the selection of the final model was based on strictly predictive criteria, as outlined in Section 3.3. Table 7 presents the predictive metrics for the best models considering cases of both malaria types.

The best model for each malaria type is highlighted in bold. Notably, the Bell model achieved the best fit across all criteria for *Plasmodium vivax* cases, making it the preferred model for this scenario.

Table 6 DIC and WAIC of fitted models assuming the Negative Binomial type 2 (NB2) distribution.

P. vivax		
Proposal	DIC	WAIC
Model 1	132,334.23	132,393.67
Model 2	132,295.20	132,354.46
Model 3	130,230.69	130,628.16
Model 4	**97,266.65**	**96,034.04**

P. falciparum		
Proposal	DIC	WAIC
Model 1	84,406.37	84,474.61
Model 2	84,403.39	84,470.86
Model 3	83,266.79	83,529.92
Model 4	**83,259.50**	**83,522.76**

Table 7 Predictive metrics of the best models, with y_{ij} representing the number of cases.

Model	MBE	NRMSE	RAE	RMSLE	RSE
			P. vivax cases		
Bell	**0.619**	**0.101**	**0.604**	**0.509**	**0.983**
Poisson	0.639	0.103	0.621	0.534	1.020
Negative Binomial	0.651	0.104	0.632	0.557	1.030
			P. falciparum cases		
Bell	0.066	0.072	0.567	0.179	0.992
Poisson	0.068	0.073	0.577	0.183	1.010
Negative Binomial	**0.062**	**0.070**	**0.552**	**0.171**	**0.947**

One might argue that the widespread use of the Poisson distribution in practical applications and the literature justifies choosing the Poisson model over the Bell model, especially given the minimal differences in the metrics shown in Table 7 for the *Plasmodium vivax* cases. However, since both the Bell and Poisson distributions are uniparametric, there is no substantial loss of parsimony in opting for the Bell distribution.

This result is noteworthy, as the Bell distribution has not been previously employed in space-time modeling, particularly for malaria cases. Here, for *Plasmodium vivax* type cases, the Bell model outperformed the others in terms of predictive accuracy, as indicated by several error metrics. It is possible that studies such as those by Achcar et al. (2011) or Ayala et al. (2022) could have benefited from evaluating the Bell distribution, leading to even better performance.

For *Plasmodium falciparum* malaria cases, the best predictive performance was achieved with the Negative Binomial model, followed by the Bell model. This suggests that accounting for overdispersion improves predictive accuracy for *Plasmodium falciparum* cases. Moreover, the additional parameter in the Negative Binomial model provided a slight advantage over the Bell model.

However, it is important to highlight that the Bell model remains a strong alternative, particularly due to its simplicity and parsimony. Unlike the Negative Binomial model, which requires an additional dispersion parameter, the Bell distribution

achieves competitive results with a single parameter and no fit for covariates, making it more interpretable and easier to estimate.

To ensure the robustness and reliability of the model performance evaluation, a bootstrap approach was applied to the predictive metrics (Efron and Tibshirani 1994). Given the inherent variability in malaria incidence data and the complexity of spatio-temporal modeling, bootstrap resampling provides a comprehensive way to assess the stability of the estimates. By repeatedly sampling the data with replacement, this method allows for the construction of empirical distributions of the metrics, reducing the impact of potential outliers and sampling biases.

The results obtained can be visualized in Figures 3 and 4, which further support our previous assertions and strengthen the validity of our findings.

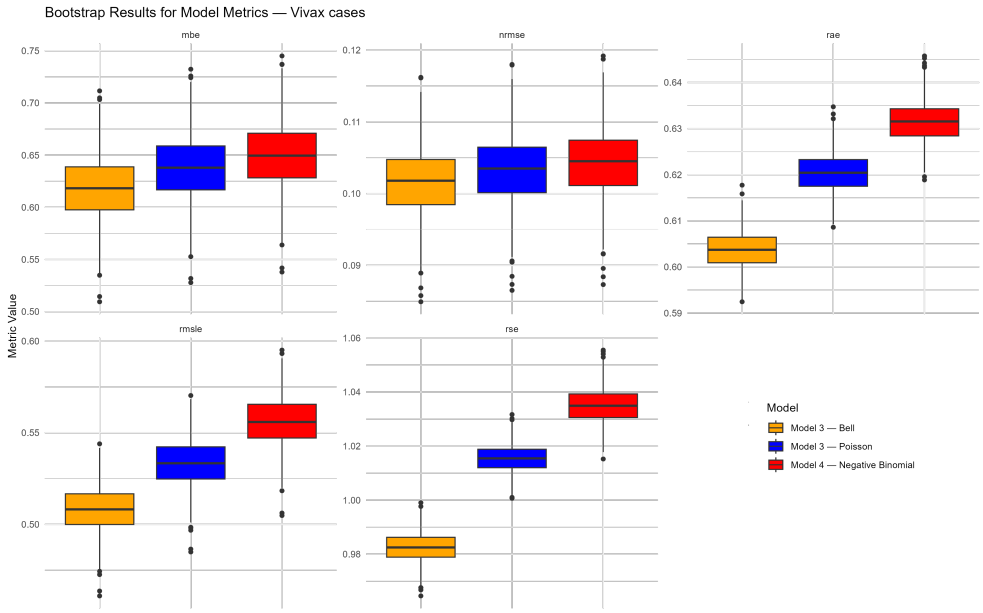


Fig. 3 Bootstrap analysis of *Plasmodium vivax* cases predictions.

Given its performance across both malaria types, the Bell model stands out as a viable and efficient alternative to the traditional models. While the Negative Binomial model showed a predictive advantage for *Plasmodium falciparum* cases, this came at the cost of increased complexity due to an additional parameter. The Bell model, on the other hand, consistently delivered competitive results with a simpler structure, making it a preferable choice when balancing accuracy, interpretability, and computational efficiency. Its success in modeling *Plasmodium vivax* cases and its strong performance for *Plasmodium falciparum* cases suggest that the Bell distribution deserves further exploration in future studies, particularly in space-time modeling applications where parsimony is a key consideration.

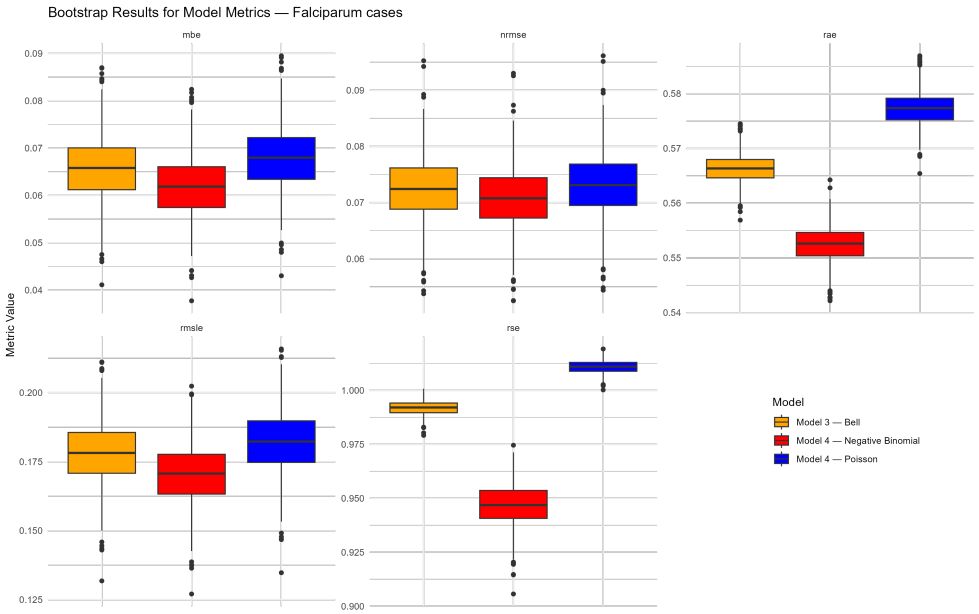


Fig. 4 Bootstrap analysis of *Plasmodium falciparum* cases predictions.

Finally, the following subsections will discuss the evaluations, focusing primarily on the errors of the best-selected models. As expected, this discussion will be based on the criteria defined in items 1 and 2 at the beginning of this section.

5.2 Evaluation of the Best Models

To contextualize the discussion on the predictions of the selected models, we refer to Figure 2, presented at the beginning of this section, which displays the incidence of malaria *Plasmodium vivax* per thousand inhabitants in November 2017 across the 107 microregions of the Legal Amazon. With this image in mind, let us examine Figure 5.

A direct comparison between the two images, focusing only on the intensity of incidence (ignoring its magnitude), suggests that the Bell model for *Plasmodium vivax* cases successfully identified microregions with a higher predicted incidence. Moreover, a closer inspection reveals that the model also captured the timing of this increase in incidence. However, upon reviewing the scales of the graphs, it becomes evident that item 2, as outlined in the beginning of Section 5, was not fully satisfied by this model alone.

Specifically, the scale in Figure 2 ranges from 0 to 20 cases per thousand inhabitants, while Figure 5 has a range from 0 to 0.6.

The decision to present the actual and predicted incidence for November 2017 was deliberate. As Figure 1 illustrates, this period marked the first time in several years that the average number of cases per thousand inhabitants in the region exceeded

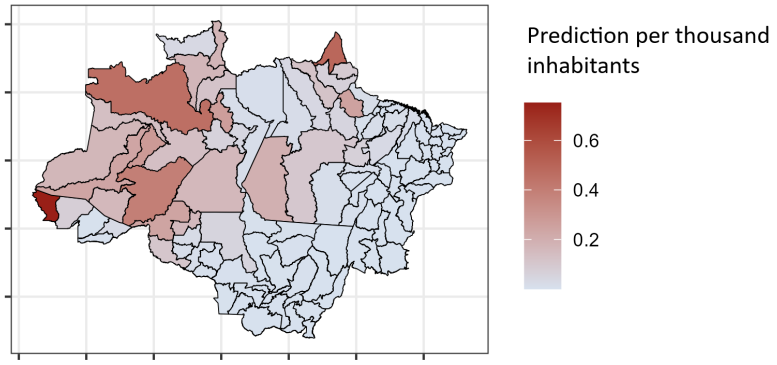


Fig. 5 Predicted incidence of *Plasmodium vivax* malaria per thousand inhabitants in the Legal Amazon in November 2017.

1. Furthermore, as shown in Figure 6 — which covers February, June, and November, selected due to seasonal factors and typical case volume — nominal errors are consistently larger in regions with higher disease incidence. This indicates that while the model can identify the location and timing of increased incidence (satisfying item 1), it struggles with the magnitude of these predictions, which increases with actual incidence (a failure to meet item 2).

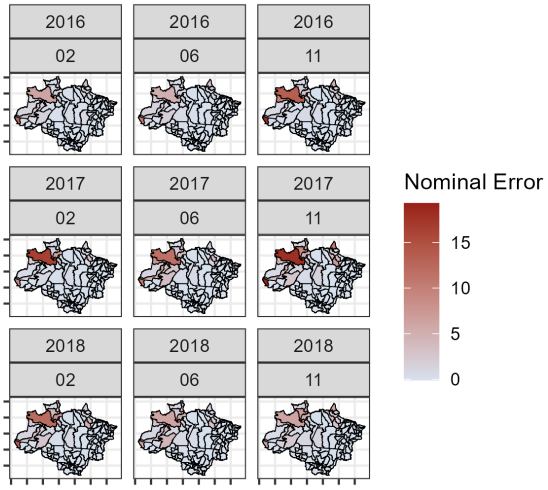


Fig. 6 Nominal prediction errors for *Plasmodium vivax* malaria over time. The labels 02, 06, and 11 denote the months of February, June, and November, respectively.

These observations suggest that reversing the downward trend in malaria cases has proven challenging for the model in terms of prediction. As stated earlier, an ideal model would satisfy both item 1 and item 2. This applies to both the model for *Plasmodium vivax* cases and the model for *Plasmodium falciparum* cases. Although

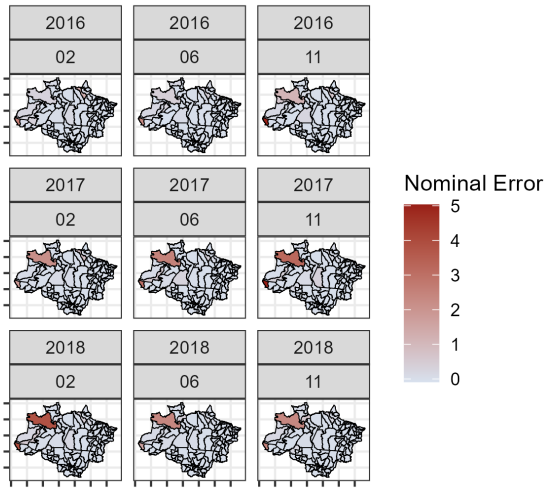


Fig. 7 Nominal prediction errors for *Plasmodium falciparum* malaria over time. The labels 02, 06, and 11 denote the months of February, June, and November, respectively.

the *Plasmodium falciparum* model performed better in terms of the scale-independent metrics presented in Table 7, it still struggled with predicting the magnitudes of incidences (as shown in Figure 7), especially considering that actual *Plasmodium falciparum* incidences are much less frequent than those for *Plasmodium vivax*.

Despite the *Plasmodium falciparum* model being better adjusted than the *Plasmodium vivax* model, Figures 8 and 9 illustrate that the RMSLE — a scale-independent measure that penalizes underestimation of the real value — is still higher in regions with higher incidence.

As noted earlier, even though item 2 was not fully satisfied, item 1 and the analysis of the errors provide valuable insights into the model's performance. Specifically, as shown in Figures 5-9, both models, especially the one for *Plasmodium vivax* cases, performed well in identifying regions (and times) with low or zero incidence. Using this information, a potential approach to correct for magnitude errors will be discussed in the following subsection.

5.3 Correction of Magnitudes

As previously mentioned, the adjusted models were effective at identifying both the timing and locations of increased malaria incidence, as well as areas with low or almost zero incidences. Using this information, we performed a linear regression considering the actual and predicted incidence values for the period between 2016 and 2017:

$$\text{actual}_m = \beta_1 \times \text{predicted}_m + \epsilon_m,$$

where β_1 represents the coefficient of the predicted values vector (note that this model does not include an intercept), and m indexes all combinations of microregions and time periods where the predicted values from the Bell model (for *Plasmodium vivax*

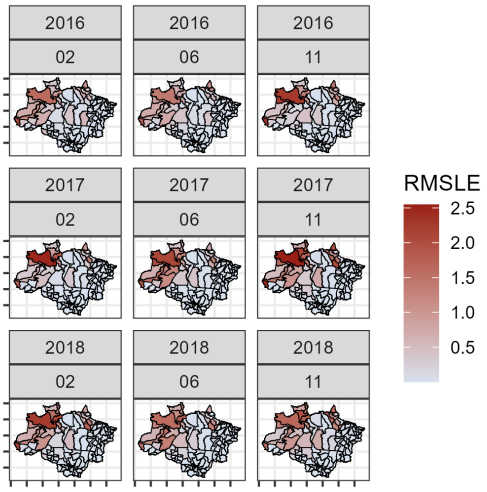


Fig. 8 RMSLE prediction for *Plasmodium vivax* malaria over time. The labels 02, 06, and 11 denote the months of February, June, and November, respectively.

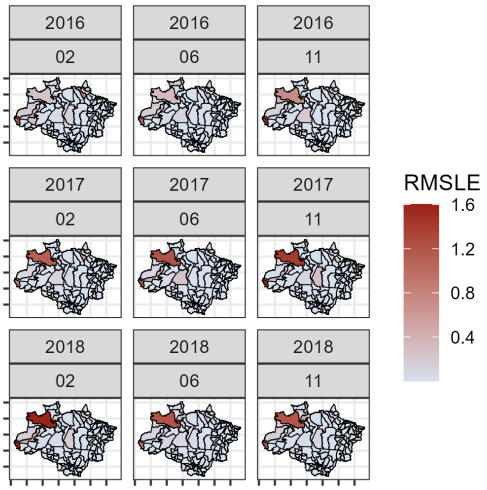


Fig. 9 RMSLE prediction for *Plasmodium falciparum* malaria over time. The labels 02, 06, and 11 denote the months of February, June, and November, respectively.

malaria cases — similar construction applies for *Plasmodium falciparum* cases) were greater than 0.007.

This threshold of 0.007 was intentionally chosen because it corresponds to the maximum magnitude where the model did not need correction (using 2016 and 2017 data). Predictions below this threshold often closely matched the real value, and therefore, we set this cutoff. The linear regression produced a coefficient estimate of $\hat{\beta}_1 = 10.493$. Using this coefficient, we defined a new prediction vector for 2018 (a year not used in

the regression) as follows:

$$\text{new_predicted}_m = \begin{cases} \text{predicted}_m, & \text{if predicted}_m < 0.007, \\ 10.493 \times \text{predicted}_m, & \text{otherwise.} \end{cases}$$

The nominal errors of this new prediction vector can be compared with those of the original prediction vector, as shown in Figures 10 and 11. The nominal errors appear to have been significantly reduced by the correction, though they remain high in some areas.

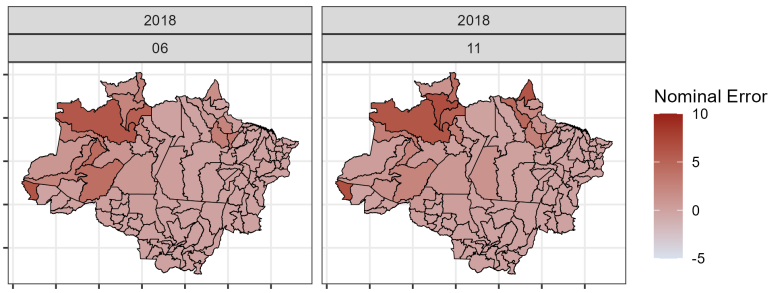


Fig. 10 Nominal prediction errors (original) for *Plasmodium vivax* malaria. The labels 06 and 11 denote the months of June and November, respectively.

However, this approach to correcting incidence rates has its limitations. For instance, there is no guarantee that the relationship between real and predicted values will remain consistent in the future.

More advanced methods than linear regression could potentially address this issue, particularly considering the temporal dimension of the data, which likely introduces significant autocorrelation in the predicted values — violating the assumptions of the linear model. Nevertheless, any such method would still require continuous reevaluation of the mechanism used to generate the *new predicted* vector.

Despite these challenges, the use of more suitable covariates, which better explain malaria incidence, could improve the correction of case magnitudes. This approach would likely be more appropriate and elegant than the one proposed here. However, the search for such covariates is hindered by the difficulty in obtaining reliable and up-to-date historical data for microregions (or municipalities) in the Legal Amazon region. This challenge, along with other related issues, will be discussed in greater detail in the following section.

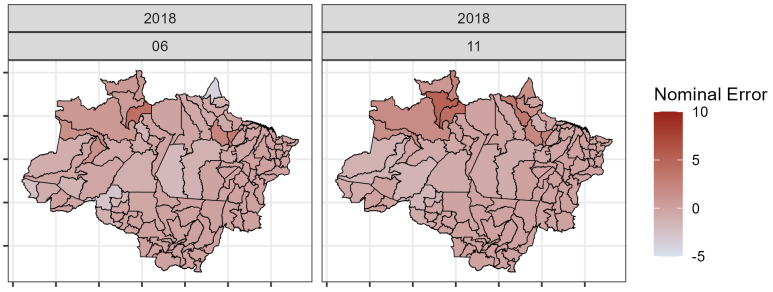


Fig. 11 Nominal prediction errors (with corrected vector) for *Plasmodium vivax* malaria. The labels 06 and 11 denote the months of June and November, respectively.

6 Conclusion

Malaria remains a significant global health threat, with hundreds of thousands of new cases and deaths reported annually. Monitoring malaria is crucial to track its temporal and spatial evolution, which helps identify patterns and predict future outbreaks. Over the years, several approaches have been developed to explain and predict malaria cases, risks, and incidence, especially in the context of the Brazilian Legal Amazon region. In this work, we focused on three notable methodologies, including two recent studies by Achcar et al. (2011), Sena et al. (2022) and Ayala et al. (2022).

Building upon these and other space-time models used for count data, our goal was to develop a space-time model that could predict the monthly malaria case rate for each microregion in the Legal Amazon, with an emphasis on assessing the incidence per 1,000 inhabitants. A key innovation of this study was the application of the Bell distribution to model case counts. This distribution, previously untested in space-time contexts, proved to be a valuable alternative to traditional models such as Poisson and Negative Binomial, particularly for *Plasmodium vivax* malaria data. Moreover, while previous studies assessed the evolution of the disease over time and space, none explored the interaction between the spatial and temporal components in modeling. Our findings indicate that incorporating this interaction significantly improves the model's fit, highlighting its importance. The `R` code for the procedures implemented is available on GitHub, as referenced in da Silva (2023).

The results were promising. Both models — one for *Plasmodium vivax* and one for *Plasmodium falciparum* malaria — successfully predicted when and where the incidence of malaria would increase for new data. Furthermore, the models effectively identified regions with low or no incidence. However, they were less successful in accurately predicting the magnitude of the incidence, tending to underestimate it. This issue was illustrated in the comparison between Figures 2 and 5. To address this, we proposed a method to correct the magnitudes, although it was acknowledged that this

approach might not be the most robust. One potential improvement involves identifying additional covariates to enhance the functional form of the models (Equations (3) and (4)), but obtaining relevant and high-quality data remains a major challenge.

A significant limitation faced in this work is the difficulty in accessing up-to-date malaria data at the monthly or microregional level. At the time of writing, only records up to 2018 were consolidated, with more recent data still in the preliminary stages. This limited data availability presents a major hurdle to practical implementation of the model, as predictions would have to rely on incomplete or outdated information.

Furthermore, obtaining exogenous variables (covariates) with sufficient granularity poses an even greater challenge. Ideally, multiple factors influencing malaria incidence should be integrated into the models, but the scarcity of reliable and detailed data significantly restricts this possibility. Notably, even previous studies on this topic have been limited to a small number of exogenous variables, many of which are fixed over time.

In light of these challenges, future work should focus on a systematic search for reliable predictors of malaria incidence at the microregional level, ideally monthly (or even weekly/daily). Furthermore, evaluating the performance of the model with preliminary data will be essential for its real-world implementation, as any model must provide accurate predictions based on the available data at the time. Another direction for future research is to explore bivariate modeling approaches, where records of different *Plasmodium* species (*vivax* and *falciparum*) could be modeled jointly, as suggested by Colborn et al. (2018). This could be compared with traditional univariate modeling approaches. Additionally, incorporating control charts, as done by Sena et al. (2022), would be valuable for monitoring case incidence, although it would be crucial to account for the space-time dependencies in the data, as discussed by Qiu (2019). Recent work by Zhou et al. (2025) further highlights how machine learning-based control charts can effectively monitor complex spatio-temporal data streams, significantly outperforming traditional methods.

In conclusion, the space-time model based on the Bell distribution proved effective in predicting malaria incidence and demonstrated the importance of incorporating spatial-temporal interactions in count data models. This approach holds promise for improving predictions of malaria cases and may contribute to a more accurate forecasting and better management of malaria outbreaks in the future.

Funding. Paulo H. Ferreira acknowledges support from CNPq (Grant 307221/2022-9).

Conflict of interest. No potential conflict of interest was reported by the authors.

Data availability. The analyses conducted in this study are based on a real dataset composed of information obtained from two independent databases. The first source consists of malaria notifications from the SIVEP - Malaria system (http://200.214.130.44/sivep_malaria/). Access to the SIVEP - Malaria data requires a valid login and password. The second database contains information on the average monthly humidity and average monthly temperature of the microregions. These data were obtained from the National Oceanic and Atmospheric Administration (https://www.noaa.gov).

Code availability. All procedures in this article were implemented in R, and the code is available in the GitHub repository: `https://github.com/mkyou/malaria`.

Clinical trial number. Not applicable

Appendix A Count Regression Models

Poisson model

The Poisson distribution is one of the most widely used models for count data and serves as a natural starting point for modeling the number of events occurring in any context. If $Y \sim \text{Poisson}(\mu)$, the PMF is given by:

$$p(y;\mu) = \frac{e^{-\mu}\mu^{y}}{y!}, \quad y \in \mathbb{N},\ \mu > 0,$$

where μ denotes both the expected value and the variance of the distribution, that is,

$$\mathbb{E}[Y] = \mu \quad \text{and} \quad \mathbb{V}\text{ar}[Y] = \mu.$$

This equality between the mean and variance is a defining characteristic of the Poisson model. Although it provides a parsimonious and interpretable framework, this property can be restrictive in real-world applications, where overdispersion (variance greater than the mean) is often observed.

Negative Binomial model

The Negative Binomial distribution provides a flexible alternative to the Poisson model for count data, particularly when the assumption of equality between the mean and variance does not hold. There are several parameterizations of this distribution, and in this study we employ the Negative Binomial type 2 (NB2) distribution. According to Gardner et al. (1995), if $Y \sim \text{NB2}(\mu, \phi)$, then the PMF is given by:

$$p(y;\mu,\phi) = \binom{y+\phi-1}{y}\left(\frac{\mu}{\mu+\phi}\right)^{y}\left(\frac{\phi}{\phi+\mu}\right)^{\phi}, \quad y \in \mathbb{N}, \mu > 0 \text{ and } \phi > 0,$$

where the mean and variance are:

$$\mathbb{E}[Y] = \mu \quad \text{and} \quad \mathbb{V}\text{ar}[Y] = \mu + \mu^2/\phi.$$

Here, the additional variance term $\mu^2/\phi > 0$ reflects the extra dispersion in the Negative Binomial compared to the Poisson distribution.

References

Achcar JA, Martinez EZ, Souza ADPd, Tachibana VM, Flores EF. Use of Poisson spatiotemporal regression models for the Brazilian Amazon Forest: malaria count data. Revista da Sociedade Brasileira de Medicina Tropical. 2011;44:749–754.

de Aguiar Barros J, Granja F, Pequeno P, et al. Gold miners augment malaria transmission in indigenous territories of Roraima state, Brazil. Malaria Journal. 2022;21:358.

Alegana VA, Atkinson PM, Lourenço C, Ruktanonchai NW, Bosco C, zu Erbach-Schoenberg E, et al. Advances in mapping malaria for elimination: fine resolution modelling of *Plasmodium falciparum* incidence. Scientific Reports. 2016;6:29628.

Ayala MJ, Bastos LS, Villela DA. On multifactorial drivers for malaria rebound in Brazil: a spatio-temporal analysis. Malaria Journal. 2022;21(1):52.

Barreto M, Bertoldo J, Sironi A, Sampaio V. Providing Malaria Analytics as a Service. In: CLOSER; 2019. p. 515–522.

Bell ET. Exponential numbers. The American Mathematical Monthly. 1934;41(7):411–419.

Bell ET. Exponential polynomials. Annals of Mathematics. 1934;p. 258–277.

Blangiardo M, Cameletti M, Baio G, Rue H. Spatial and spatio-temporal models with R-INLA. Spatial and spatio-temporal epidemiology. 2013;4:33–49.

Castellares F, Ferrari SL, Lemonte AJ. On the Bell distribution and its associated regression model for count data. Applied Mathematical Modelling. 2018;56:172–185.

Cichosz P. Data mining algorithms: explained using R. John Wiley & Sons; 2014.

Colborn KL, Mueller I, Speed TP. Joint modeling of mixed Plasmodium species infections using a bivariate Poisson lognormal model. The American Journal of Tropical Medicine and Hygiene. 2018;98(1):71.

Corless RM, Gonnet GH, Hare DE, Jeffrey DJ, Knuth DE. On the LambertW function. Advances in Computational mathematics. 1996;5(1):329–359.

Efron B, Tibshirani RJ. An introduction to the bootstrap. Chapman and Hall/CRC; 1994.

Garcia KKS, Soremekun S, Abrahão AA, Marchesini PB, Drakeley C, Ramalho WM. Is Brazil reaching malaria elimination? A time series analysis of malaria cases from 2011 to 2023. PLOS Global Public Health. 2024;4(1):1–14. https://doi.org/10.1371/journal.pgph.0002845.

Gardner W, Mulvey EP, Shaw EC. Regression analyses of counts and rates: Poisson, overdispersed Poisson, and negative binomial models. Psychological bulletin. 1995;118(3):392.

Ladislau JLdB. Avaliação do Plano de Intensificação das Ações de Controle da Malária na Amazônia Legal no Contexto da Descentralização. Epidemiologia e Serviços de

Saúde. 2006;15(2):9–20. https://bvsms.saude.gov.br/bvs/periodicos/rev_epi_vol15_n2.pdf.

Lapouble OMM, Santelli ACF, Muniz-Junqueira MI, et al. Situação epidemiológica da malária na região amazônica brasileira, 2003 a 2012. Revista Panamericana de Salud Pública. 2015;38:300–306.

Li X, Dey DK. Estimation of COVID-19 mortality in the United States using Spatio-temporal Conway Maxwell Poisson model. Spatial Statistics. 2022;49:100542.

Lindgren F, Rue H. On the second-order random walk model for irregular locations. Scandinavian journal of statistics. 2008;35(4):691–700.

Lindgren F, Rue H. Bayesian Spatial Modelling with R-INLA. Journal of Statistical Software. 2015;63(19). https://doi.org/10.18637/jss.v063.i19.

McCullagh P, Nelder JA. Generalized linear models. Routledge; 2019.

Ministério da Saúde, Brasil SdVeS.: Programa nacional de prevenção e controle da malária. Ministério da Saúde Brasília; 2003. https://bvsms.saude.gov.br/bvs/publicacoes/programa_nac_prev_malaria.pdf.

Ministério da Saúde, Brasil SdVeS. Boletim Epidemiológico Vol.53 Nº17. Boletim Epidemiológico. 2022;53(17). https://www.gov.br/saude/pt-br/centrais-de-conteudo/publicacoes/boletins/epidemiologicos/edicoes/2022/boletim-epidemiologico-vol-53-no17.pdf/view.

Pan Q. Is negative DIC possible? A simulation study on exploring the impact of negative DIC on ranking competing models.; 2016.

Prothero RM. Malaria, forests and people in Southeast Asia. Singapore Journal of Tropical Geography. 1999;20(1):76–85.

Qiu P. Some recent studies in statistical process control. Statistical quality technologies: theory and practice. 2019;p. 3–19.

Rue H, Martino S, Chopin N. INLA: A Package to do Bayesian Inference for Latent Gaussian Models; 2009, https://www.r-inla.org/, r package version 23.05.30-1.

Sanna F, Keddie SH, Boyhan T, Dzianach PA, McPhail M, Seitz J, et al. A malaria seasonality dataset for sub-Saharan Africa. Scientific Data. 2025;12:1703. Open access.

Sena JGSd, Ferreira PH, Fiaccone RL. Statistical Process Control as a Tool to Control and Prevent Malaria Epidemics in the Legal Amazon Region. Brazilian Journal of Biometrics. 2022;40(1). https://biometria.ufla.br/index.php/BBJ/article/view/538, https://doi.org/10.28951/bjb.v40i1.538.

da Silva M.: Legal Amazon region malaria analysis. GitHub; 2023. https://github.com/mkyou/malaria.

Simpson DP, Rue H, Riebler A, Martins TG, Sørbye SH. Penalising model component complexity: A principled, practical approach to constructing priors. Statistical Science. 2017;32(1):1–28. https://doi.org/10.1214/16-STS576.

Spiegelhalter DJ, Best NG, Carlin BP, Linde A. The deviance information criterion: 12 years on. Journal of the Royal Statistical Society Series B: Statistical Methodology. 2014;76(3):485–493.

Torres K, Ferreira MU, Castro MC, Escalante AA, Conn JE, Villasis E, et al. Malaria Resilience in South America: Epidemiology, Vector Biology, and Immunology Insights from the Amazonian International Center of Excellence in Malaria Research Network in Peru and Brazil. The American Journal of Tropical Medicine and Hygiene. 2022;107(4):168–181.

Watanabe S, Opper M. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. Journal of machine learning research. 2010;11(12).

Zhou Y, Jin R, Qiu P. Machine Learning Control Charts for Monitoring Spatio-Temporal Data Streams. Quality and Reliability Engineering International. 2025;41(6):2373–2384. https://doi.org/10.1002/qre.3809.