FUNDAÇÃO GETULIO VARGAS
SCHOOL OF APPLIED MATHEMATICS

EZEQUIEL DE BRAGA SANTOS

A JOINT MODELLING APPROACH FOR LEISHMANIASIS IN THE STATE OF SÃO PAULO

Rio de Janeiro
2024

EZEQUIEL DE BRAGA SANTOS

A JOINT MODELLING APPROACH FOR LEISHMANIASIS IN THE STATE OF SÃO PAULO

Bachelor dissertation presented to the School of Applied Mathematics (FGV/EMAp) to obtain the Bachelor's degree in Applied Mathematics.

Area of study: Biostatistics

Advisor: Prof. Luiz Max Carvalho

Rio de Janeiro

2024

Ficha catalográfica elaborada pela BMHS/FGV

Santos, Ezequiel de Braga
A joint modelling approach for leishmaniasis in the state of São Paulo/ Ezequiel de Braga Santos. – 2024.
27f.

Bachelor Dissertation (Undergraduation) – School of Applied Mathematics.

Advisor: Prof. Luiz Max Carvalho.
Includes bibliography.

1. . 2. . 2. . I. Carvalho, Luiz Max. II. School of Applied Mathematics. III. A joint modelling approach for leishmaniasis in the state of São Paulo

EZEQUIEL DE BRAGA SANTOS

A JOINT MODELLING APPROACH FOR LEISHMANIASIS IN THE STATE OF SÃO PAULO

Bachelor dissertation presented to the School of Applied Mathematics (FGV/EMAp) to obtain the Bachelor's degree in Applied Mathematics.

Area of study: Biostatistics

Approved on December 11th, 2024
By the organizing committee

I dedicate this dissertation to my mother, Elisabete, who always encouraged me to study; and to my cousin, Isaac (in memoriam), for having been part of my childhood and student life.

Acknowledgements

I thank my parents, Elisabete and Eraldo, especially my mother, for all the encouragement and teachings throughout my student life. And to my sister, Aline, and my brother-in-law, Marcelo, for all the support during this journey.

I thank my aunt, Maria, for all the advice and encouragement. And to my godparents Arnaldo and Berenice, who always supported me.

I thank all my friends who contributed to my education, especially: my elementary and high school friends, Hicaro and Sávio; my classmates Pedro and Iara for all the fun and support; and my former roommate, Jairon, for all the teachings and companionship throughout these years.

I thank the FGV's Center for the Development of Sciences and Mathematics (CDMC) for making this dream possible.

I thank my high school Math teacher, Josué, and my PIC teacher, Messias, for sparking my interest in Mathematics. I would also like to thank all the professors at FGV EMAp for all their contributions throughout the course.

Finally, I am extremely grateful to my advisor and mentor, Luiz Max Carvalho, for believing in my potential, for always being available to help me with my studies, and for always helping me become a better person academically and personally.

Abstract

This study aims to investigate the relationship between environmental covariates and leishmaniasis, especially how they contribute to the emergence of the vector in the city. For this purpose, a joint modeling is employed using a Bayesian approach. The data used are from the state of São Paulo between the years 2003 and 2017, and were constructed by combining several sources.

Keywords: joint modelling, jmbayes2, leishmaniasis, joint models.

Resumo

Este estudo se dedica a investigar a relação entre covariaveis ambientais com a leishmaniose, especialmente como elas contribuem para o surgimento do vetor transmissor na cidade. Para isso, uma modelagem conjunta é empregada usando uma abordagem Bayesiana. Os dados utilizados são do estado de São Paulo entre os anos de 2003 e 2017, e foi construido combinando diversas fontes.

Palavras-chave: modelagem conjunta, jmbayes2, leishmaniose, modelos conjuntos.

List of Figures

List of Tables

Contents

1 Introduction

Leishmaniasis is a disease caused by a parasitic protozoan of the *Leishmania* species, transmitted through the bite of infected sandflies. It can manifest in three main forms: visceral (the most severe), cutaneous (the most common, usually causing skin ulcers), and mucocutaneous (affecting the mouth, nose, and throat) (WORLD HEALTH ORGANIZATION, 2024). Our study will focus on visceral leishmaniasis (VL), which is fatal in over 95% of cases if untreated, according to the World Health Organization (WHO) (WORLD HEALTH ORGANIZATION, 2024). In the state of São Paulo, data reveal a low incidence over the past decade, but there is a concentration of cases in specific regions that deserve attention.

Due to the complexity and diversity of factors influencing its spread, it is important to explore advanced techniques that can capture the effects of these different aspects and provide deeper insights into the disease's dynamics. In this context, "joint models" stand out as a powerful statistical tool for the joint modeling of longitudinal data and time-to-event data.

In Alsefri et al. (2020), the use of this methodology is highlighted in 75 articles, with applications in 14 of them, utilizing different submodels for both longitudinal measurements and time-to-event data. For the former, models range from generalized linear models (GLM) to more complex ones, assuming different association structures with the second submodel.

Although this approach has been applied to various diseases, its application to leishmaniasis appears to be underexplored. Thus, this study will not only broaden the scope of joint modeling applications but also provide a better understanding of the effectiveness of this methodology in leishmaniasis analysis.

2 Methodology

2.1 Joint Models

In the original formulation (HENDERSON; DIGGLE; DOBSON, 2000), we observe m subjects over a time interval $[0, \tau)$ with quantitative measurements y_{ij} at time t_{ij} for $j = 1, \ldots, n_i$ and ith subject. Moreover, we consider a function $\delta_i(u)$ indicating whether the subject is at risk of experiencing an event at time u. The formulation proposed for the joint modelling of the measurements and events consider two sub-models driven by an unobserved zero-mean Gaussian process $W_i(t) = \{W_{1i}(t), W_{2i}(t)\}$ for subject i as follows.

The model for measurements y_{ij} at time t_{ij} is given by

$$Y_{ij} = \mu_i(t_{ij}) + W_{1i}(t_{ij}) + Z_{ij},$$

with $\mu_{ij}(t) = x_{1i}^{\top}(t)\beta_1$ and $Z_{ij} \sim \text{Normal}(0, \sigma_Z^2)$ independent measurement errors.

For the event intensity process at time t, the model is given by

$$h_i(t) = \delta_i(t) h_0(t) \exp[x_{2i}^{\top}(t)\beta_2 + W_{2i}(t)],$$

for some baseline function $\alpha_0(t)$.

2.2 Data

The VL data were reported from the state of São Paulo and collected from different sources: the vector data were collected from Casanova et al. (2015) and the human and dog cases from Department of Information and Informatics of the Unified Health System (DATASUS) (MINISTÉRIO DA SAÚDE, n.d.) and São Paulo State Secretariat (CENTRO DE VIGILÂNCIA EPIDEMIOLÓGICA "PROF. ALEXANDRE VRANJAC" - CVE, n.d.). We got yearly cases in humans, dogs and vector and the yearly number of deaths in humans for the period from 2003 to 2017.

Climate and air pollution covariates were obtained from European Centre for Medium-Range Weather Forecasts (ECMWF). In particular, the first ones are from the ERA 5 model (ECMWF, 2024b) and the second ones are from the CAMs model (ECMWF, 2024a). Finally, fire outbreaks data were collected from INPE (INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS - INPE, n.d.).

For land use data obtained from MapBiomas ((SEEG), 2024) we created an identifier corresponding to the land use classification used in square kilometers. Moreover, we added geographic and demographic data from Brazilian Institute of Geography and

Statistics (IBGE) (INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE, n.d.). Lastly, we extracted the area of each biome from **geobr** package in R (PEREIRA; GONCALVES, 2024).

We aggregate each of these data in long format containing the information cited for each city in each year and we compute relevant metrics such as incidence, mortality rate and indicators of the presence of disease.

Finally, we create a variable of time-to-event for status_dogs, status_human and status_vector with year 2000 as base time and we separate data in time-to-event (with status variables, time and immutable covariates) and longitudinal outcomes (with remaining covariates). The variables are described in the table below.

Table 1 – **Description of variables**

Variable	**Description**
code_7d	IBGE code for city
name_city	City name
name_state	State name
year	Year
total_area_m2	City area in squares meters
prop_cer	Proportion of cerrado in the city
prop_ma	Proportion of atlantic forest in the city
prop_sc	Proportion of coastal system in the city
mb_i_j_k_l_m	Area of land use classification in levels i, j, k, l e m
co_ppb	Annual mean carbon monoxide concentration in ppb
no2_ppb	Annual mean nitrogen dioxide concentration in ppb
o3_ppb	Annual mean ozone concentration in ppb
pm25_ugm3	Annual mean concentration of particulate matter up to 2.5 micrometers in $\mu g/m^3$
so2_ugm3	Annual mean concentration of sulfur dioxide in $\mu g/m^3$
rainfall	Annual mean precipitation in mm
temperature_c	Annual mean temperature in ºC
relative_humidity_percent	Annual mean relative humidity in %
wind_direction_degree	Annual mean wind direction in degrees
wind_speed_ms	Annual mean wind speed in ms
fire_outbreaks	Annual number of fire outbreaks
status_dogs	Variable indicating the presence of an infected dog
status_human	Variable indicating the presence of an infected human
status_vector	Variable indicating the presence of a vector
n_cases	Number of cases in humans
n_deaths	Number of deaths in humans
pop	Population
incidence	Annual disease incidence (per 100K inhabitants)
death_rate	Annual death rate (per 100K inhabitants)

2.2.1 Data imputation

Problems were identified with two covariates in specific years: population and relative humidity. For the first one, data is missing for 2007 and we imputed the average of the years 2006 and 2008. For the second one, incorrect data was found for 2017. The solution used was to apply a linear regression on time from 2003 to 2016 for each city and predict the data for 2017.

2.3 Selection of Covariates

This process was carried out using by combining Variance Inflation Factor (VIF) with negative binomial model; correlation plot; and Least Absolute Shrinkage and Selection Operator (LASSO) with Poisson model. Based in these results and percentage of zeros/NA seen in the figure (1), we select the treatable variables: prop_cer for survival model; and temperature_c, wind_speed_ms and mb_1_2_4_4_5 for longitudinal models.

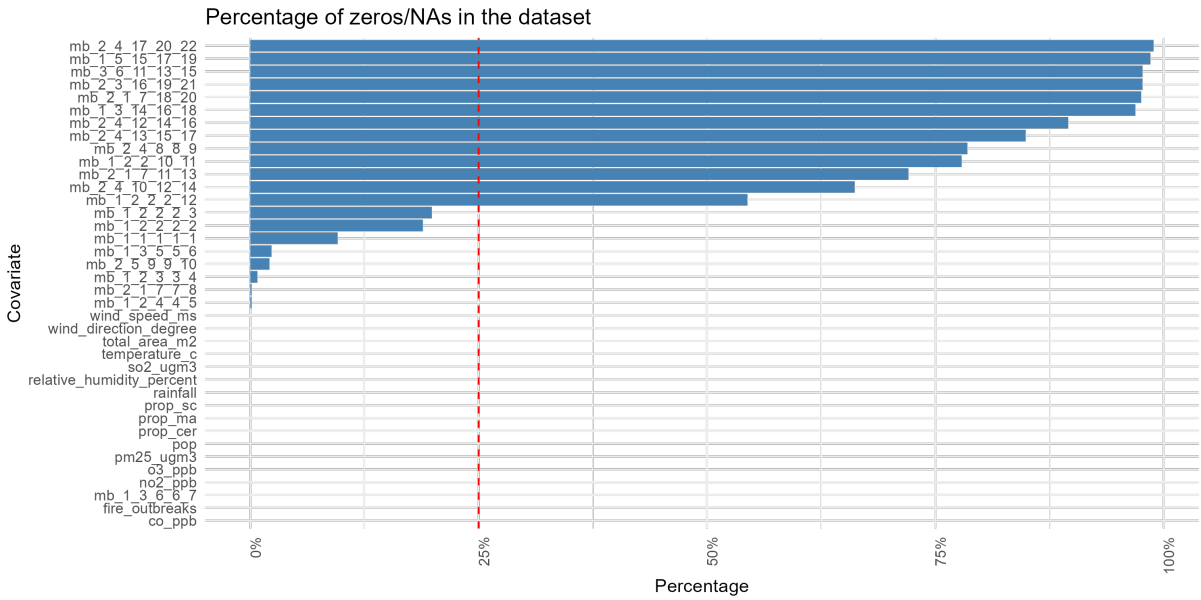


Figure 1 – Percentage of missing data (zeros, in the case of MapBiomas variables).

2.4 Model

For each observation $y_{ij}(t)$ of biomarker $j = 1, \ldots, 3$ and city $i = 1, \ldots, 645$ on time t, we consider a longitudinal model given by

$$y_{ij}(t) = \beta_{0j} + b_{0ij} + b_{1ij}t + \varepsilon_i(t),$$

where $\varepsilon_i(t) \sim N(0, \sigma^2)$ and $b_i = \begin{bmatrix} b_{i1} \\ b_{i2} \\ b_{i3} \end{bmatrix} \sim N(0, D)$ for $b_{ij} = \begin{bmatrix} b_{0ij} \\ b_{1ij} \end{bmatrix}$, $j = 1, 2, 3$ and $i = 1, \ldots, 645$.

For event in the city $i = 1, \ldots, 645$, we consider a relative risk model for status_vector given by

$$h_i(t) = h_0(t) \exp \left[\gamma_1 \text{prop\_cer} + \sum_{j=1}^{3} \alpha_i \left(\beta_{0j} + b_{0ij} + b_{1ij} t \right) \right]$$

2.5 Likelihood

Given the true event time for city i, T_i^*; the observed time for city i, T_i; the event indicator, δ_i; and the longitudinal covariate y_i; the joint distribution is of the following form

$$\begin{aligned} p(y_i, T_i, \delta_i) &= \prod_j p(y_{ij} \mid b_{ij}) p(T_i, \delta_i \mid b_i), \\ p(y_i \mid b_i) &= \prod_j p(y_{ij} \mid b_{ij}). \end{aligned}$$

where b_i a vector of random effects that explains the interdependencies and $p(\cdot)$ density function.

2.6 Estimation

Under the Bayesian paradigm both θ and b_i are regarded as parameters and the inference is based on full posterior distribution

$$p(\theta, b \mid T, \delta, y) = \frac{\prod_i p(T_i, \delta_i \mid b_i; \theta) p(y_i \mid b_i; \theta) p(b_i; \theta) p(\theta)}{\prod_i p(T_i, \delta_i, y_i)}.$$

Given prior distributions on parameters and baseline function for survival model, Markov chains Monte Carlo (MCMC) is used to sample from posterior distribution.

3 Results

3.1 Exploratory data analysis

Firstly, when observing the incidence in the figure (2), we notice a greater concentration of cases along a regional strip.

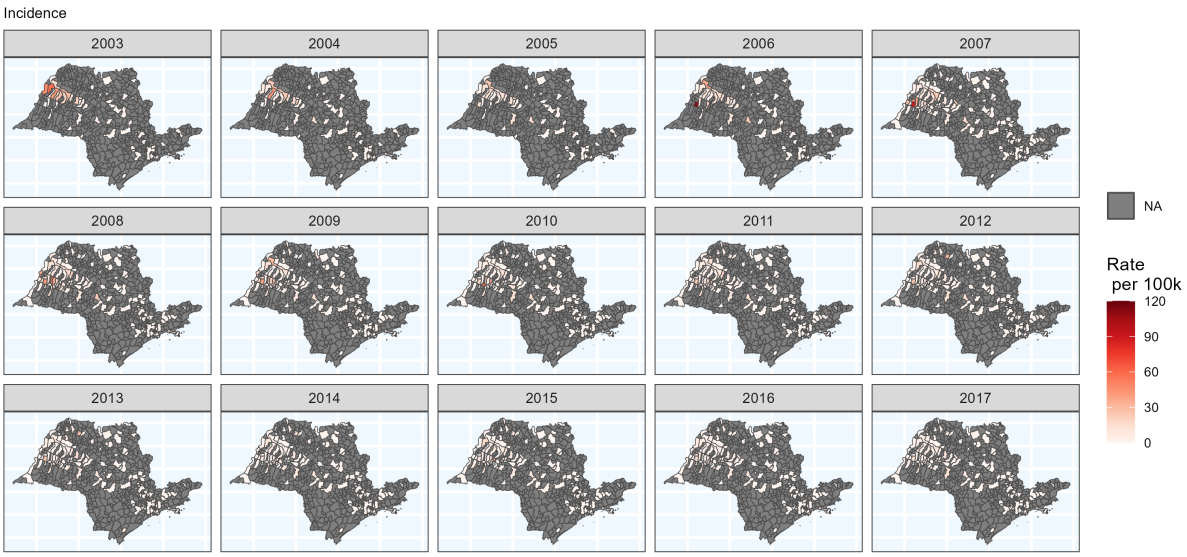


Figure 2 – Incidence per $100k$ inhabitants.

This is due to the fact that there is a greater concentration of the transmitting vector (*Lutzomyia*) in this region, as can be seen in the figure below:

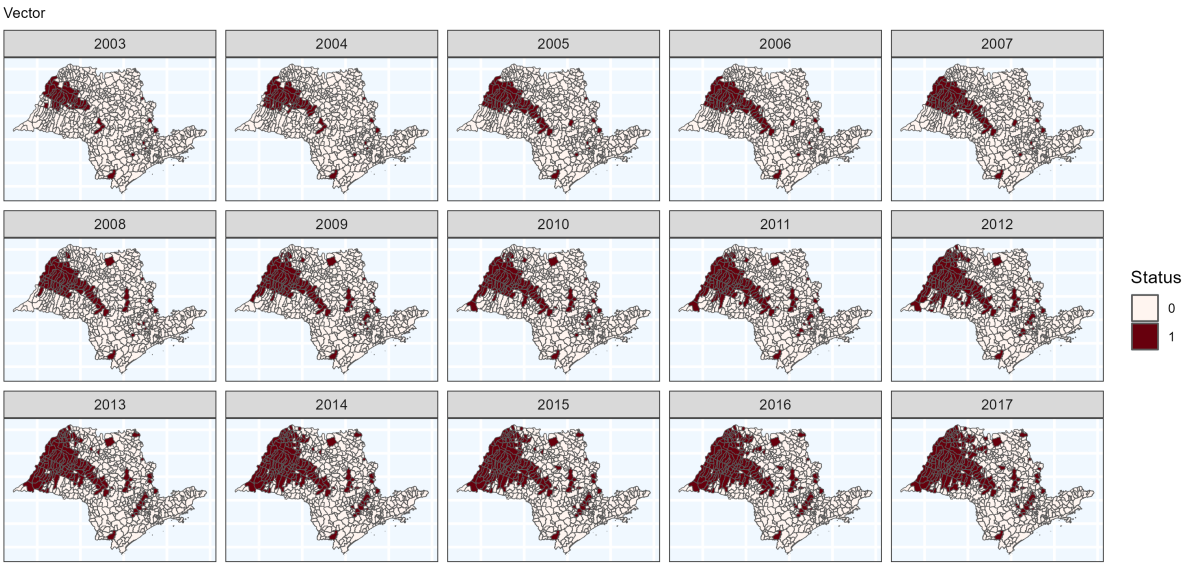


Figure 3 – Vector detection in the city: 1 for presence and 0 for absence.

When observe the biomes, we noticed that most of the cities invaded by the vector are characterized by the absence of cerrado, as can be seen in the figure below:

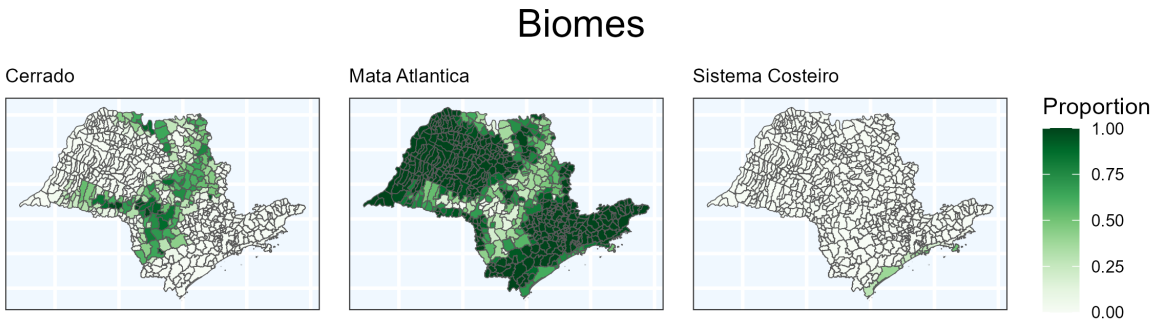


Figure 4 – Proportion of biomes in cities.

Another important factor analyzed was the concentration of molecules that indicate air pollution, such as carbon monoxide, for example. A greater presence of these variables can be seen in the metropolitan region of São Paulo, as shown in the figure below:

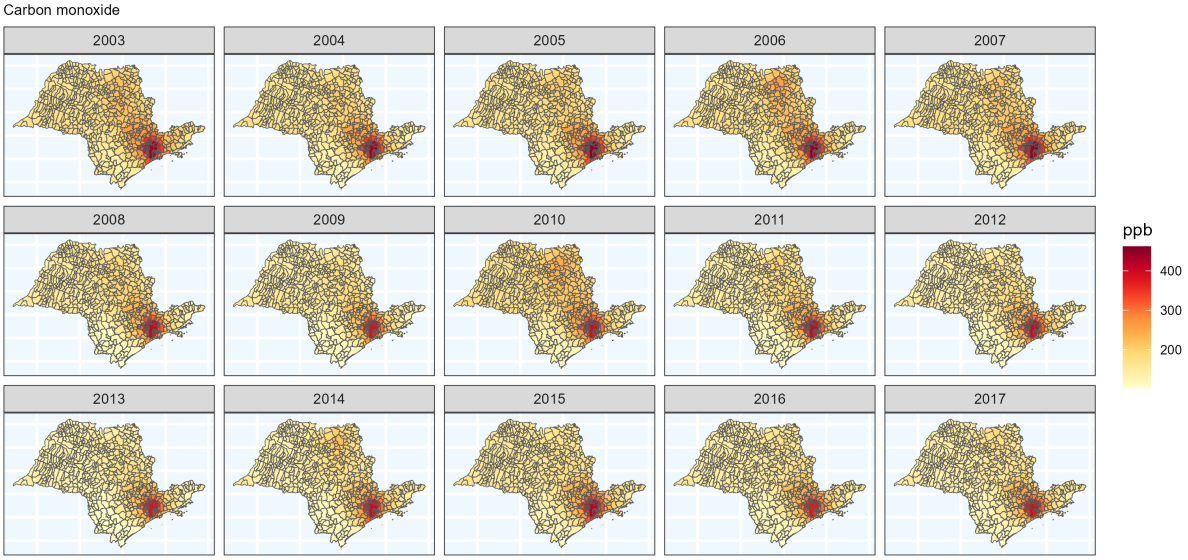


Figure 5 – Concentration of carbon monoxide.

Another example is concentration of nitrogen dioxide:

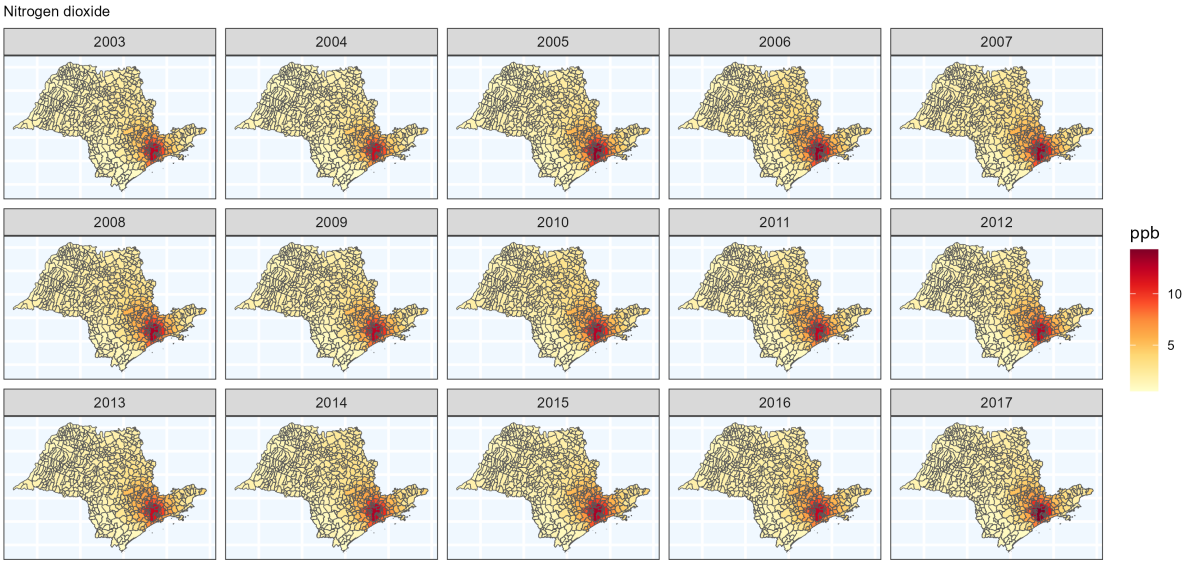


Figure 6 – Concentration of nitrogen dioxide.

When we observe land use covariates, the variable mb_1_2_4_4_5 is present in the regions of greatest incidence, as shown in the figure below:

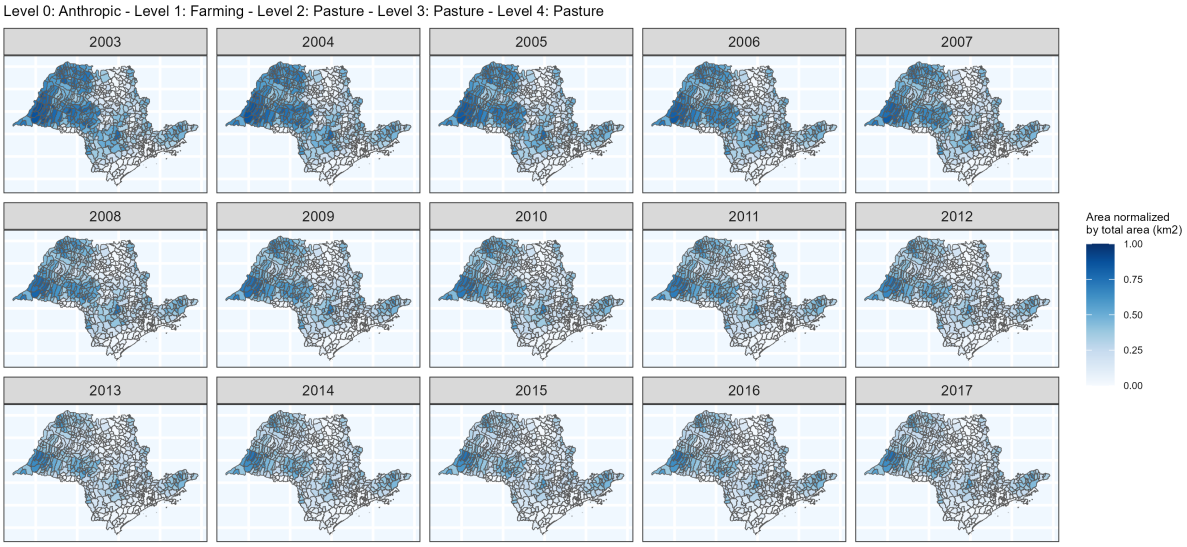


Figure 7 – Land use classification mb_1_2_4_4_5.

Furthermore, the correlation between the covariates was analyzed, highlighting a strong correlation between those indicating environmental pollution as well as between the urban area and these same covariates. Other expected results were the correlation of relative humidity and wind direction with forest formation and also with temperature. All correlations are summarized in the following figure.

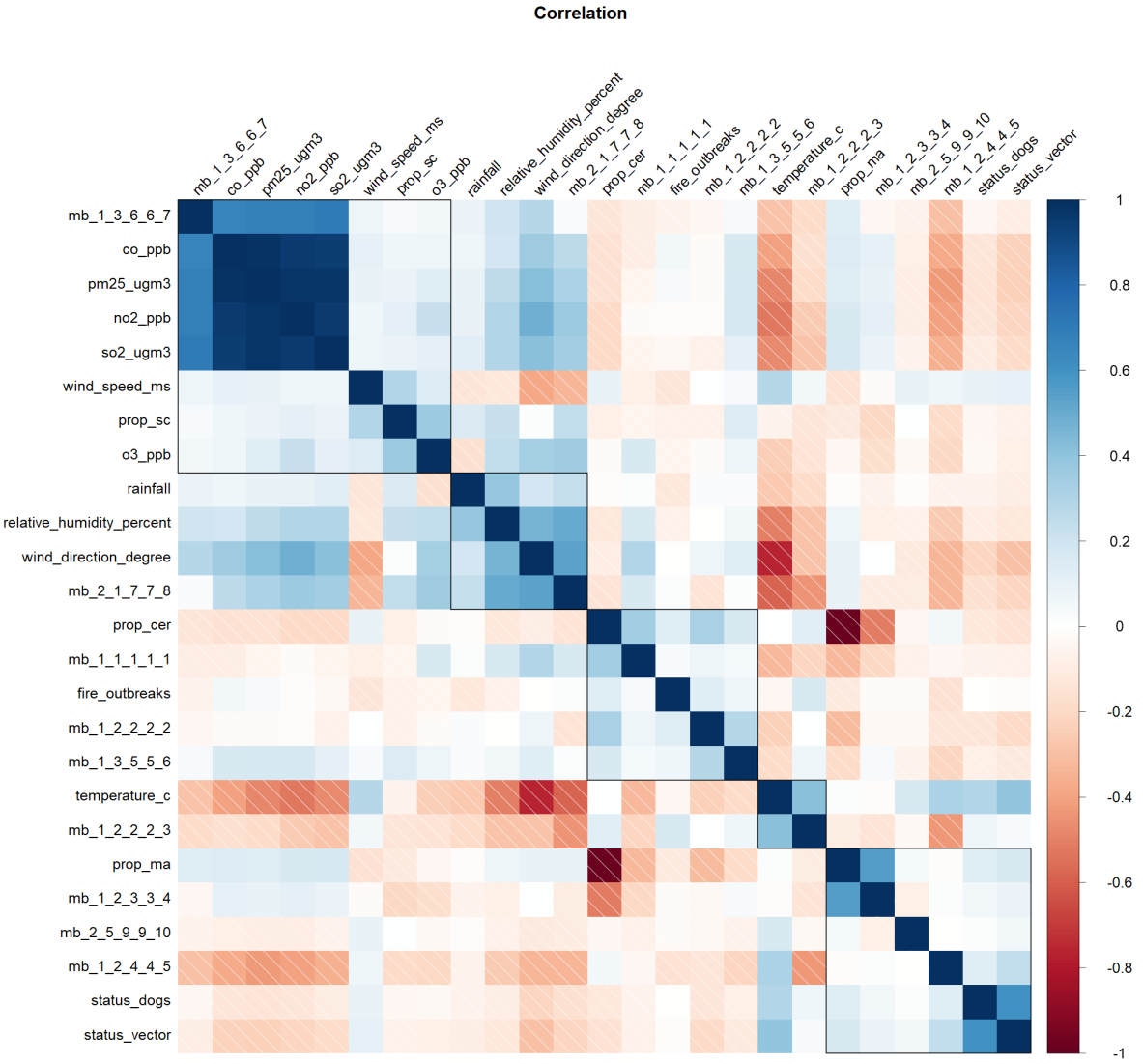


Figure 8 – Correlation of covariates.

3.2 Fitting the model

Model fitting was done using the jm() function from **JMbayes2** R package (RIZOPOULOS; PAPAGEORGIOU; MIRANDA AFONSO, 2024). 4 chains with 150000 iterations and 50000 burn-in each were used to fit the model. The summary can be seen below.

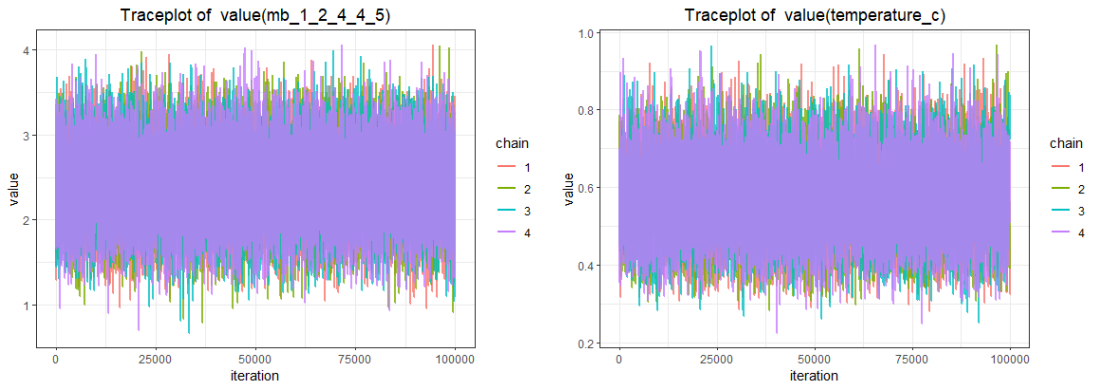
Table 2 – **Summary of survival outcome**

	Mean	StDev	2.5%	97.5%	P	Rhat
prop_cer	-2.3653	0.847	-4.1349	-0.8357	0.0005	1.0001
value(temperature_c)	0.5777	0.0870	0.4130	0.7548	0.0000	1.0002
value(wind_speed_ms)	0.9141	0.3746	0.1290	1.6558	0.0127	1.0004
value(mb_1_2_4_4_5)	2.4239	0.3952	1.6562	3.2027	0.0001	1.0002

Table 3 – **Summary of longitudinal outcomes**

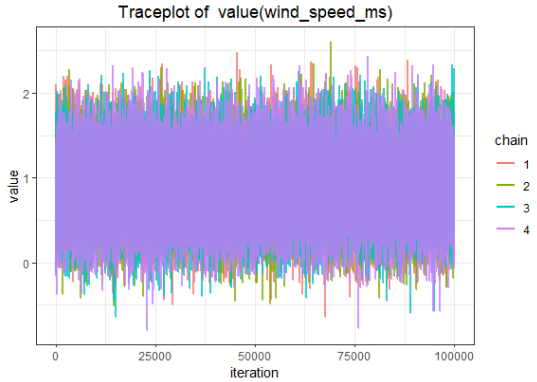
temperature_c	Mean	StDev	2.5%	97.5%	P	Rhat
(Intercept)	23.3356	0.1355	23.0694	23.5975	0.0000	1.0078
sigma	0.3918	0.0030	0.3859	0.3977	0.0000	1.0001
wind_speed_ms						
(Intercept)	2.7064	0.0370	2.6328	2.7785	0.0000	1.0087
sigma	0.0966	0.0007	0.0952	0.0980	0.0000	1.0001
mb_1_2_4_4_5						
(Intercept)	0.1853	0.0205	0.1453	0.2257	0.0000	1.0011
sigma	0.0219	0.0002	0.0216	0.0222	0.0000	1.0000

All rhat values are very close to 1, indicating convergence. When we observe trace plots in the figures (9), (10) and (11), we notice that the chains show good mixing and stability, suggesting the sampling is sampling effectively.



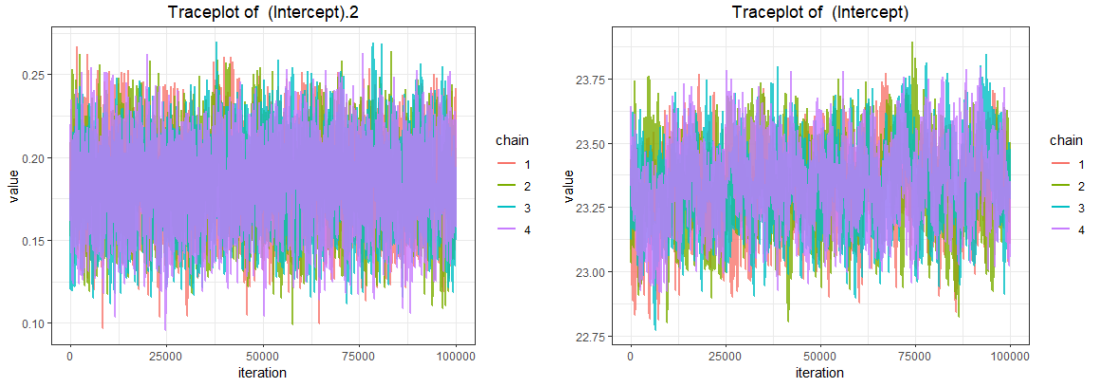
(a) Trace plot of association coefficient of mb_1_2_4_4_5.

(b) Trace plot of association coefficient of temperature_c.



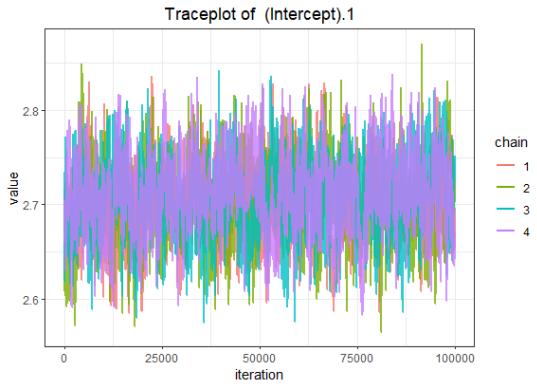
(c) Trace plot of association coefficient of wind_speed_ms.

Figure 9 – Trace plots of association coefficients.



(a) Trace plot of intercept of mb_1_2_4_4_5.

(b) Trace plot of intercept of temperature_c.



(c) Trace plot of intercept of wind_speed_ms.

Figure 10 – Trace plots of intercepts of longitudinal outcomes.

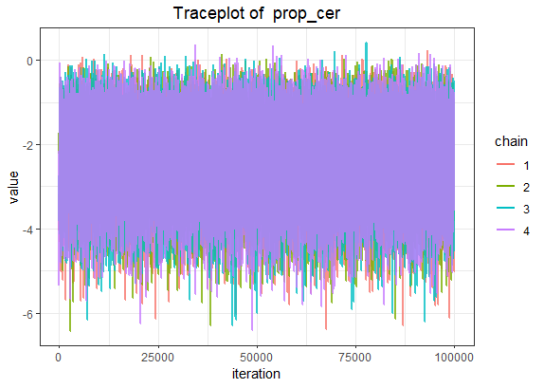
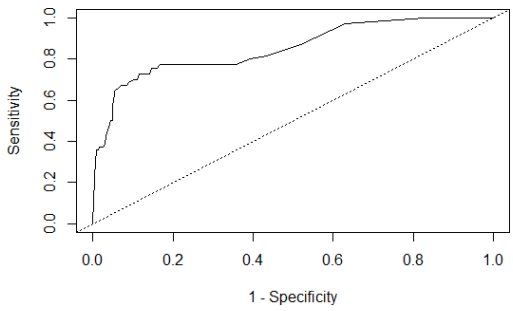


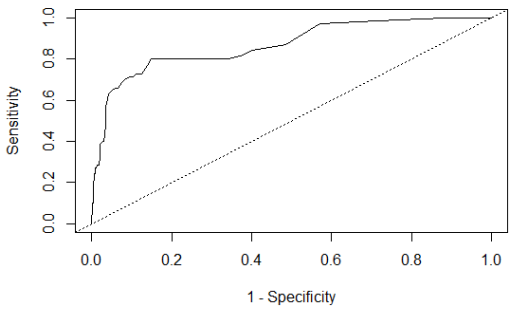
Figure 11 – Trace plot of survival covariate.

3.3 Predictive accuracy

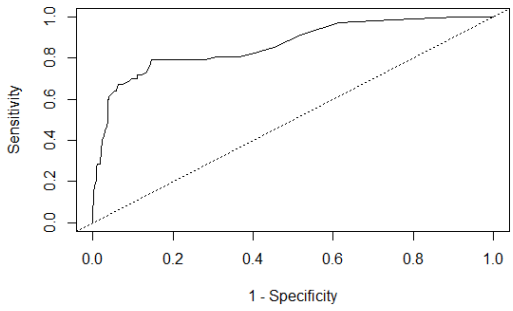
For evaluate the capability of the model, the ROC methodology can be used. Firstly, information up to the year 2008 and a five-year window were considered. Then the window was moved year by year until reaching 2016. The curves are summarized in the next figure.



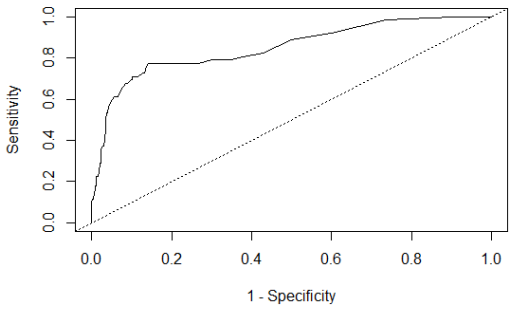
(a) Window: $(2008, 2013]$
AUC: 0.8505



(b) Window: $(2009, 2014]$
AUC: 0.8687



(c) Window: $(2010, 2015]$
AUC: 0.8631



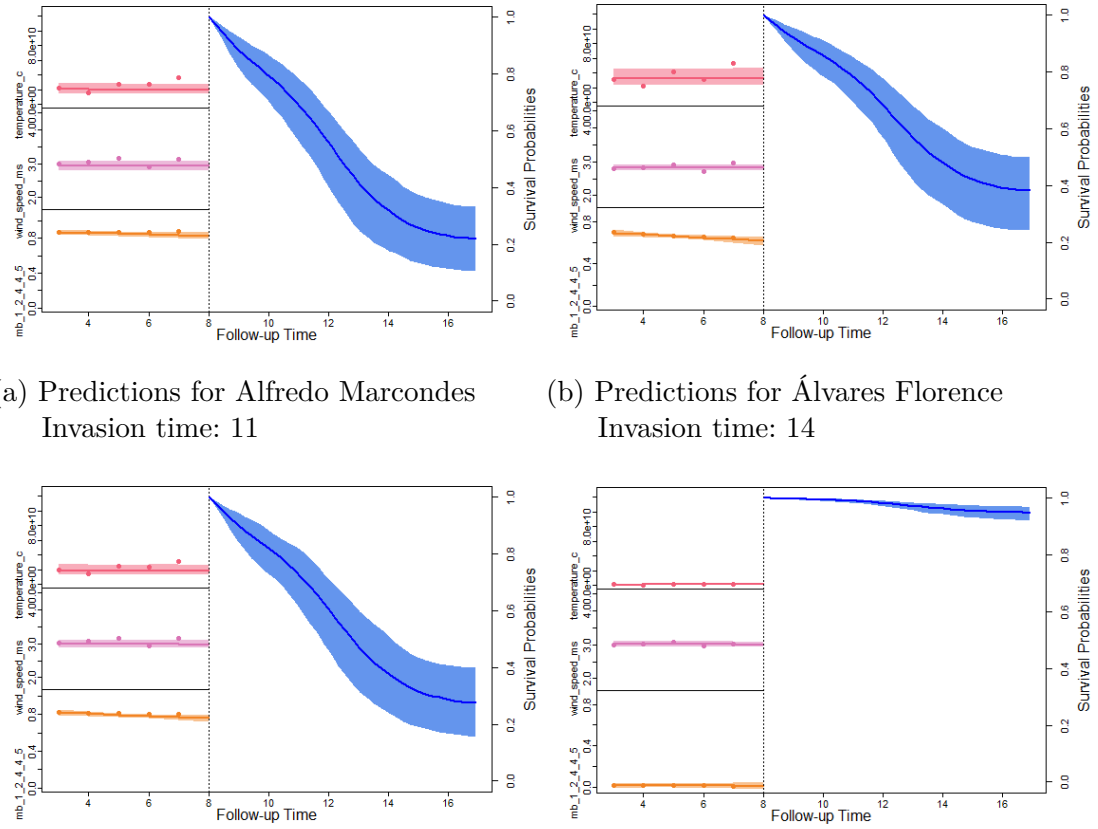
(d) Window: $(2011, 2016]$
AUC: 0.8492

Figure 12 – ROC curves.

The consistent high AUC values suggest the model is well-calibrated and performs effectively in predicting outcomes across varying time windows.

3.4 Computing survival probabilities

With this model it is possible to calculate the probability of survival for each city, that is, the probability of the municipality not being invaded up to time t. For example, the figure below shows these probabilities for four cities.



(a) Predictions for Alfredo Marcondes
Invasion time: 11

(b) Predictions for Álvares Florence
Invasion time: 14

(c) Predictions for Álvares Machado
Invasion time: 11

(d) Predictions for São Paulo
Invasion time: not invaded

Figure 13 – Predictions.

The figure above shows an expected behavior: the probability of non-invasion decreases as the time approaches the observed time and in the following times.

4 Conclusions

In this study, joint modelling was applied in unusual way, allowing an approach not only traditional survival data. This shows that this approach can be effectively adapted to various datasets, not just clinical trials, as is normally done.

The model developed was reasonably effective in capturing the effect of covariates on the probability of invasion of the transmitting vector, but it has limitations in making predictions for some cities. This difficulty may be related to the presence of the vector in isolated areas, where it is difficult to notice any invasion pattern.

Futhermore, the data used has many intractable covariates, making model convergence difficult and limiting the space of possible models. Another problem was the data format, because it did not present the structure time-to-event. Finally, there are few cities invaded by the vector throughout the study, especially in the early years.

Possible improvements include making a selecting of models with new covariates, although this has high computational cost. In addition, spatial components that capture the relationship between neighboring cities can be included. Finally, a calibration of priors can also be useful.

References

(SEEG), Greenhouse Gas Emissions Estimation System. **Projeto MapBiomas – Coleção 5.0 da Série Anual de Mapas de Cobertura e Uso de Solo do Brasil**. [S.l.: s.n.], 2024. Available from: <https://brasil.mapbiomas.org/en/>.

ALSEFRI, Maha et al. Bayesian joint modelling of longitudinal and time to event data: a methodological review. **BMC Medical Research Methodology**, v. 20, p. 465–480, Apr. 2020. ISSN 1471-2288. DOI: 10.1186/s12874-020-00976-2. Available from: <https://doi.org/10.1186/s12874-020-00976-2>.

CASANOVA, Claudio et al. Distribution of Lutzomyia longipalpis Chemotype Populations in São Paulo State, Brazil. **PLOS Neglected Tropical Diseases**, Public Library of Science, v. 9, n. 3, p. 1–14, Mar. 2015. DOI: 10.1371/journal.pntd.0003620. Available from: <https://doi.org/10.1371/journal.pntd.0003620>.

CENTRO DE VIGILÂNCIA EPIDEMIOLÓGICA "PROF. ALEXANDRE VRANJAC" - CVE. **Leishmaniose Visceral**. [S.l.: s.n.]. CVE website. Available from: <https://saude.sp.gov.br/cve-centro-de-vigilancia-epidemiologica-prof.-alexandre-vranjac/areas-de-vigilancia/doencas-de-transmissao-por-vetores-e-zoonoses/agravos/leishmaniose-visceral/>.

ECMWF. **Copernicus Atmosphere Monitoring Service (CAMS)**. [S.l.: s.n.], 2024. Available from: <https://atmosphere.copernicus.eu/>.

_____. **ECMWF Reanalysis v5 (ERA5)**. [S.l.: s.n.], 2024. Available from: <https://www.ecmwf.int/en/forecasts/dataset/ecmwf-reanalysis-v5>.

HENDERSON, Robin; DIGGLE, Peter; DOBSON, Angela. Joint modelling of longitudinal measurements and event time data. **Biostatistics**, v. 1, n. 4, p. 465–480, Dec. 2000. ISSN 1465-4644. DOI: 10.1093/biostatistics/1.4.465. eprint: https://academic.oup.com/biostatistics/article-pdf/1/4/465/655029/010465.pdf. Available from: <https://doi.org/10.1093/biostatistics/1.4.465>.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE. **Downloads**. [S.l.: s.n.]. IBGE website. Available from: <https://www.ibge.gov.br/geociencias/downloads-geociencias.html>.

INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS - INPE. **BDQUEIMADAS**. [S.l.: s.n.]. INPE website. Available from: <https://terrabrasilis.dpi.inpe.br/queimadas/bdqueimadas/>.

MINISTÉRIO DA SAÚDE. **LEISHMANIOSE VISCERAL - CASOS CONFIRMADOS NOTIFICADOS NO SISTEMA DE INFORMAÇÃO DE AGRAVOS DE NOTIFICAÇÃO - SÃO PAULO**. [S.l.: s.n.]. DATASUS website. Available from: <http://tabnet.datasus.gov.br/cgi/deftohtm.exe?sinanwin/cnv/leishvSP.def>.

PEREIRA, Rafael H. M.; GONCALVES, Caio Nogueira. **geobr: Download Official Spatial Data Sets of Brazil**. [S.l.: s.n.], 2024. R package version 1.8.2. Available from: <https://CRAN.R-project.org/package=geobr>.

RIZOPOULOS, Dimitris; PAPAGEORGIOU, Grigorios; MIRANDA AFONSO, Pedro. **JMbayes2: Extended Joint Models for Longitudinal and Time-to-Event Data**. [S.l.: s.n.], 2024. R package version 0.5-1, https://github.com/drizopoulos/JMbayes2. Available from: <https://drizopoulos.github.io/JMbayes2/>.

WORLD HEALTH ORGANIZATION. **Leishmaniasis**. [S.l.: s.n.], 2024. World Health Organization website. Available from: <https://www.who.int/news-room/fact-sheets/detail/leishmaniasis>. Visited on: 11 June 2024.