UNICAMP

UNIVERSIDADE ESTADUAL DE
CAMPINAS

Instituto de Matemática, Estatística e
Computação Científica

MARÍLIA GABRIELA ROCHA

Modeling Sequential Data from Multiple Sources using Variable Length Markov Chains and Exogenous Covariates

Modelagem de Dados Sequenciais de Múltiplas Fontes com Cadeias de Markov de Alcance Variável e Covariáveis Exógenas

Campinas

2024

Marília Gabriela Rocha

Modeling Sequential Data from Multiple Sources using Variable Length Markov Chains and Exogenous Covariates

Modelagem de Dados Sequenciais de Múltiplas Fontes com Cadeias de Markov de Alcance Variável e Covariáveis Exógenas

Dissertação apresentada ao Instituto de Matemática, Estatística e Computação Científica da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestra em Estatística.

Dissertation presented to the Institute of Mathematics, Statistics and Scientific Computing of the University of Campinas in partial fulfillment of the requirements for the degree of Master in Statistics.

Supervisor: Nancy Lopes Garcia

Este trabalho corresponde à versão final da Dissertação defendida pela aluna Marília Gabriela Rocha e orientada pela Profa. Dra. Nancy Lopes Garcia.

Campinas

2024

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

R582m

Rocha, Marília Gabriela, 1998-
Modeling sequential data from multiple sources using variable length Markov chains and exogenous covariates / Marília Gabriela Rocha. – Campinas, SP : [s.n.], 2024.

Orientador: Nancy Lopes Garcia.
Dissertação (mestrado) – Universidade Estadual de Campinas, Instituto de Matemática, Estatística e Computação Científica.

1. Cadeias de Markov. 2. Regressão logística. 3. Dengue - Brasil. I. Garcia, Nancy Lopes, 1964-. II. Universidade Estadual de Campinas. Instituto de Matemática, Estatística e Computação Científica. III. Título.

Informações Complementares

Título em outro idioma: Modelagem de dados sequenciais de múltiplas fontes com cadeias de Markov de alcance variável e covariáveis exógenas
Palavras-chave em inglês:
Markov chains
Logistic regression
Dengue - Brazil
Área de concentração: Estatística
Titulação: Mestra em Estatística
Banca examinadora:
Nancy Lopes Garcia [Orientador]
Adriano Zanin Zambom
Victor Freguglia Souza
Data de defesa: 05-03-2024
Programa de Pós-Graduação: Estatística

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0002-4394-2423
- Currículo Lattes do autor: http://lattes.cnpq.br/7522127827888881

Dissertação de Mestrado defendida em 05 de março de 2024 e aprovada pela banca examinadora composta pelos Profs. Drs.

Prof(a). Dr(a). NANCY LOPES GARCIA

Prof(a). Dr(a). ADRIANO ZANIN ZAMBOM

Prof(a). Dr(a). VICTOR FREGUGLIA SOUZA

A Ata da Defesa, assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria de Pós-Graduação do Instituto de Matemática, Estatística e Computação Científica.

Acknowledgements

I want to express my gratitude to Prof. Dr. Nancy Lopes Garcia, my supervisor, for her guidance and support throughout the completion of this work. Her understanding of the need to balance study with work has been invaluable, allowing for a more flexible pace. Witnessing a woman researcher in the exact field has been truly inspiring, and I greatly admire her work.

I extend my thanks to my friends from my master's program, as well as those from various aspects of my life and my boyfriend, for their continuous encouragement and support. Your companionship has made every path lighter, and I am truly grateful. Special appreciation goes to Adenilson Rodrigues Júnior, Lucas Santos, Alexandre Reggiolli, Pedro Nasevicius, and Rafael Ricardi Branco for our collaborative studies. Your contributions significantly impacted the navigation through the challenges of this master's degree period, especially during the tumultuous times of the COVID-19 pandemic.

A heartfelt thanks goes out to my family, with a special mention to my brother Marco Aurélio, for their unwavering encouragement.

Additionally, I express my gratitude to all the dedicated workers at IMECC and Unicamp whose efforts have made my studies possible in recent years.

Lastly, I appreciate all the authors whose works I consulted and read during the course of my research. Your contributions have enriched my understanding and shaped the development of this work.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

Resumo

As Cadeias de Markov de Alcance Variável com Covariáveis Exógenas são modelos estocásticos inseridos no contexto das Cadeias de Markov de Alcance Variável. Estes modelos empregam Modelos Lineares Generalizados para calcular as probabilidades de transição, considerando tanto o histórico de estados quanto as covariáveis exógenas dependentes do tempo. O algoritmo beta-contexto é utilizado para selecionar um sufixo finito relevante, ou contexto, para prever o próximo símbolo. Este algoritmo estima modelos flexíveis em forma de árvore, agregando estados irrelevantes no histórico do processo e permitindo que o modelo incorpore covariáveis exógenas ao longo do tempo.

Nossa pesquisa amplia o algoritmo beta-contexto com alcance variável para incorporar tanto covariáveis exógenas dependentes quanto invariantes no tempo, utilizando dados de múltiplas fontes. Dentro dessa abordagem, temos uma cadeia de Markov distinta para cada fonte de dados, o que possibilita uma compreensão do comportamento do processo em diversas situações, como diferentes localizações geográficas. Apesar do uso de dados provenientes de diferentes fontes, pressupomos que todas as fontes são independentes e compartilham parâmetros idênticos - exploramos os contextos dentro de cada fonte de dados e os combinamos para calcular as probabilidades de transição, resultando em uma árvore unificada. Essa abordagem elimina a necessidade de considerações relacionadas à dependência espacial dentro do modelo. Além disso, também incorporamos modificações no procedimento de estimação para lidar com contextos que ocorrem com baixa frequência.

Nossa motivação foi investigar o impacto das taxas anteriores de dengue, condições climáticas e fatores socioeconômicos nas taxas subsequentes da doença em diversos municípios do Brasil, fornecendo percepções sobre a dinâmica de transmissão da doença.

Palavras-chave: Cadeias de Markov. Regressão logistica. Dengue - Brasil.

Abstract

Variable Length Markov Chains with Exogenous Covariates (VLMCX) are stochastic models in the framework of Variable Length Markov Chains (VLMC) that use Generalized Linear Models (GLM) to compute transition probabilities, taking into account the state history and time-dependent exogenous covariates. The beta-context algorithm is used to select a relevant finite suffix, or context, for predicting the next symbol. This algorithm estimates flexible tree-structured models by aggregating irrelevant states in the process history and enables the model to incorporate exogenous covariates over time.

Our research extends the beta-context model with variable length to incorporate both time-dependent and time-invariant exogenous covariates, using data from multiple sources. Within this approach, we have a distinct Markov chain for every data source, allowing for a comprehensive understanding of the process behavior across multiple situations, such as different geographic locations. Despite the use of data from different sources, we assume that all sources are independent and share identical parameters - we explore contexts within each data source and combine them to compute transition probabilities, deriving a unified tree. This approach eliminates the necessity for spatial-dependent structural considerations within the model. Furthermore, we incorporate modifications in the estimation procedure to address contexts that appear with low frequency.

Our motivation was to investigate the impact of previous dengue rates, weather conditions, and socioeconomic factors on subsequent dengue rates across various municipalities in Brazil, providing insights into dengue transmission dynamics.

Keywords: Markov Chains. Logistic regression. Dengue - Brazil.

List of Figures

Figure 1 – Trees for the string 10001 using Context Algorithm 21

Figure 2 – Tree representation of the context function in Example 1 28

Figure 3 – Context tree τ and associated parameters $\boldsymbol{\theta}$ for Model 1 (numbers in parenthesis represent the values of α^u for each context u). 44

Figure 4 – Context tree τ and associated parameters $\boldsymbol{\theta}$ for Model 2 (numbers in parenthesis represent the values of α^u for each context u). 47

Figure 5 – Context tree τ and associated parameters $\boldsymbol{\theta}$ for Model 3 (numbers in parenthesis represent the values of α^u for each context u). 49

Figure 6 – Context tree τ and associated parameters $\boldsymbol{\theta}$ for Model 4. Numbers in parenthesis represent the values of (α_1^u, α_2^u) from each context u and $\boldsymbol{\beta}_1^u$ and $\boldsymbol{\beta}_2^u$ are the coefficient vectors of the covariates. $\boldsymbol{\beta}^{1*}$ means the context is 1 preceded by other state that is not 1. 51

Figure 7 – Context tree τ and associated parameters $\boldsymbol{\theta}$ for Model 5 (numbers in parenthesis represent the values of α^u for each context u). 53

Figure 8 – Context tree τ and associated parameters $\boldsymbol{\theta}$ for Model 6. Numbers in parenthesis represent the values of (α_1^u, α_2^u) from each context u and $\boldsymbol{\beta}_1^u$ and $\boldsymbol{\beta}_2^u$ are the coefficient vectors of the covariates. $\boldsymbol{\beta}^{1*}$ means the context is 1 preceded by other state that is not 1. 55

Figure 9 – Context tree τ and associated parameters $\boldsymbol{\theta}$ for Model 7 (numbers in parenthesis represent the values of α^u for each context u). 57

Figure 10 – Correlation matrix between socioeconomic factors 71

Figure 11 – Final tree and estimated parameters for dengue incidence in Brazilian municipalities 73

Figure 12 – Confusion matrix for prediction results (using all available observations from previous months) - Brazil municipalities 74

Figure 13 – Confusion matrix for prediction results by month (using all available observations from previous months) - Brazil municipalities 74

Figure 14 – Confusion matrix for prediction results (using previously predicted dengue incidence instead of real values) - Brazil municipalities 75

Figure 15 – Confusion matrix for prediction results by time ahead (using previously predicted dengue incidence instead of real values) - Brazil municipalities 76

Figure 16 – Confusion matrix for prediction results - Rio de Janeiro districts 79

Figure 17 – Final tree and estimated parameters for dengue incidence in Rio de Janeiro districts (fitted for years 2012-2015) 80

Figure 18 – Final tree and estimated parameters for dengue incidence in Rio de Janeiro districts (fitted for years 2013-2016) 81

Figure 19 – Final tree and estimated parameters for dengue incidence in Rio de Janeiro districts (fitted for years 2014-2017) 81
Figure 20 – Final tree and estimated parameters for dengue incidence in Rio de Janeiro districts (fitted for years 2015-2018) 82
Figure 21 – Final tree and estimated parameters for dengue incidence in Rio de Janeiro districts (fitted for years 2016-2019) 82
Figure 22 – Context tree τ and associated parameters $\boldsymbol{\theta}$ for Model 8 (numbers in parenthesis represent the values of α^u for each context u). 92
Figure 23 – Context tree τ and associated parameters $\boldsymbol{\theta}$ for Model 9 (numbers in parenthesis represent the values of α^u for each context u). 95
Figure 24 – Context tree τ and associated parameters $\boldsymbol{\theta}$ for Model 10. Numbers in parenthesis represent the values of (α_1^u, α_2^u) from each context u and $\boldsymbol{\beta}_1^u$ and $\boldsymbol{\beta}_2^u$ are the coefficient vectors of the covariates. $\boldsymbol{\beta}^{1*}$ means the context is 1 preceded by other state that is not 0. 96
Figure 25 – Comparison of average monthly dengue cases per 100,000 inhabitants in municipalities during dengue epidemic and non-epidemic years (2008 to 2023) 100
Figure 26 – Comparison of average monthly temperature (ºC) in municipalities during dengue epidemic and non-epidemic years (2008 to 2023) 101
Figure 27 – Association between average yearly temperature (ºC) and yearly dengue cases per 100,000 inhabitants in municipalities during epidemic and non-epidemic years (2008 to 2023) 102
Figure 28 – Comparison of monthly rainfall (mm) averages in municipalities during dengue epidemic and non-epidemic years (2008 to 2023) 103
Figure 29 – Comparison of monthly rainfall (days) averages in municipalities during dengue epidemic and non-epidemic years (2008 to 2023) 104
Figure 30 – Association between the percentage of population in low-income situations and yearly dengue cases per 100,000 inhabitants for municipalities during epidemic years (2008 to 2023) 105
Figure 31 – Distribution of yearly dengue cases per 100,000 inhabitants based on percentage of population in low-income situations in municipalities during epidemic years (2008 to 2023) 106
Figure 32 – Association between GDP per capita in Brazilian Real (R$) and yearly dengue cases per 100,000 inhabitants for municipalities in epidemic years (2008 to 2023) 107
Figure 33 – Distribution of yearly dengue cases per 100,000 inhabitants based on GDP per capita (R$) in municipalities during epidemic years (2008 to 2023) 108

Figure 34 – Association between the percentage of population with sewage system and yearly dengue cases per 100,000 inhabitants for municipalities during epidemic years (2008 to 2023) 109
Figure 35 – Distribution of yearly dengue cases per 100,000 inhabitants based on percentage of population with sewage system in municipalities during epidemic years (2008 to 2023) 110
Figure 36 – Association between the percentage of population living in urban areas and yearly dengue cases per 100,000 inhabitants for municipalities during epidemic years (2008 to 2023) 111
Figure 37 – Distribution of yearly dengue cases per 100,000 inhabitants based on percentage of population living in urban areas in municipalities during epidemic years (2008 to 2023) 112
Figure 38 – Distribution of yearly dengue cases per 100,000 inhabitants based on demographic density in municipalities during epidemic years (2008 to 2023) 113

List of Tables

Table 1 – Simulation results for Model 1 with time-varying exogenous covariates generated from N(0,1) and t(2) distributions (average over 100 simulations). 45

Table 2 – Estimated τ trees for Model 1, with n = 1000 and N(0, 1) distribution (frequency of occurrences over 100 simulations). 46

Table 3 – Estimated τ trees for Model 1, with n = 1000 and t(2) distribution (frequency of occurrences over 100 simulations). 46

Table 4 – Estimated τ trees for Model 1, with n = 2000 and N(0, 1) distribution (frequency of occurrences over 100 simulations). 46

Table 5 – Estimated τ trees for Model 1, with n = 2000 and t(2) distribution (frequency of occurrences over 100 simulations). 46

Table 6 – Differences between real and estimated values for Model 1 (average over 100 simulations). 47

Table 7 – Simulation results for Model 2 with time-varying exogenous covariates generated from a N(0, 1) distribution (average over 100 simulations). . . . 48

Table 8 – Estimated τ trees for Model 2, with n = 1000 (frequency of occurrences over 100 simulations). 48

Table 9 – Estimated τ trees for Model 2, with n = 2000 (frequency of occurrences over 100 simulations). 48

Table 10 – Differences between real and estimated values for Model 2 (average over 100 simulations). 49

Table 11 – Simulation results for Model 3 with time-varying exogenous covariates generated from a N(0, 1) distribution (average over 100 simulations). . . . 50

Table 12 – Estimated τ trees for Model 3, with n = 1000 (frequency of occurrences over 100 simulations). 50

Table 13 – Estimated τ trees for Model 3, with n = 2000 (frequency of occurrences over 100 simulations). 50

Table 14 – Differences between real and estimated values for Model 3 (average over 100 simulations). 51

Table 15 – Simulation results for Model 4 with time-varying exogenous covariates generated from a N(0, 1) distribution (average over 100 simulations). . . . 52

Table 16 – Estimated τ trees for Model 4, with n = 4000 (frequency of occurrences over 100 simulations). 52

Table 17 – Estimated τ trees for Model 4, with n = 8000 (frequency of occurrences over 100 simulations). 52

Table 18 – Differences between real and estimated values for Model 4 (average over 100 simulations). 53

Table 19 – Simulation results for Model 5 with time-varying exogenous covariates generated from a N(0, 1) distribution (average over 100 simulations). . . 54
Table 20 – Estimated τ trees for Model 5, with n = 1000 (frequency of occurrences over 100 simulations). 54
Table 21 – Estimated τ trees for Model 5, with n = 2000 (frequency of occurrences over 100 simulations). 54
Table 22 – Differences between real and estimated values for Model 5 (average over 100 simulations). 54
Table 23 – Simulation results for Model 6 with time-varying exogenous covariates generated from a N(0, 1) distribution (average over 100 simulations). . . 55
Table 24 – Estimated τ trees for Model 6, with n = 4000 (frequency of occurrences over 100 simulations). 56
Table 25 – Estimated τ trees for Model 6, with n = 8000 (frequency of occurrences over 100 simulations). 56
Table 26 – Differences between real and estimated values for Model 6 (average over 100 simulations). 57
Table 27 – Simulation results for Model 7 with time-varying exogenous covariates generated from a N(0, 1) distribution (average over 100 simulations). . . 58
Table 28 – Estimated τ trees for Model 7, with $n_1 = 500$ and $n_2 = 500$ (frequency of occurrences over 100 simulations). 58
Table 29 – Estimated τ trees for Model 7, with $n_1 = 1000$ and $n_2 = 1000$ (frequency of occurrences over 100 simulations). 59
Table 30 – Estimated τ trees for Model 7, with $n_1 = 1000$ and $n_2 = 2000$ (frequency of occurrences over 100 simulations). 59
Table 31 – Estimated τ trees for Model 7, with $n_1 = 2000$ and $n_2 = 2000$ (frequency of occurrences over 100 simulations). 59
Table 32 – Estimated τ trees for Model 7, with $n_1 = 200$ and $n_2 = 3800$ (frequency of occurrences over 100 simulations). 59
Table 33 – Estimated τ trees for Model 7, with $n_1 = 4000$ and $n_2 = 4000$ (frequency of occurrences over 100 simulations). 60
Table 34 – Differences between real and estimated values for Model 7 (average over 100 simulations). 60
Table 35 – Distribution of dengue cases for all the 126 municipalities 72
Table 36 – Features used as input to the model - Rio de Janeiro, Brazil (January 2012 to October 2020) 77
Table 37 – Simulation results for Model 8 with time-varying exogenous covariates generated from a N(0, 1) distribution (average over 100 simulations). . . 93
Table 38 – Estimated τ trees for Model 8, with n = 1000 (frequency of occurrences over 100 simulations). 93

Table 39 – Estimated τ trees for Model 8, with n = 2000 (frequency of occurrences over 100 simulations). 94
Table 40 – Differences between real and estimated values for Model 8 (average over 100 simulations). 94
Table 41 – Simulation results for Model 9 with time-varying exogenous covariates generated from a N(0, 1) distribution (average over 100 simulations). . . 95
Table 42 – Estimated τ trees for Model 9, with n = 1000 (frequency of occurrences over 100 simulations). 95
Table 43 – Estimated τ trees for Model 9, with n = 2000 (frequency of occurrences over 100 simulations). 95
Table 44 – Differences between real and estimated values for Model 9 (average over 100 simulations). 96
Table 45 – Simulation results for Model 10 with time-varying exogenous covariates generated from a N(0, 1) distribution (average over 100 simulations). . . 97
Table 46 – Estimated τ trees for Model 10, with n = 4000 (frequency of occurrences over 100 simulations). 97
Table 47 – Estimated τ trees for Model 10, with n = 8000 (frequency of occurrences over 100 simulations). 97
Table 48 – Differences between real and estimated values for Model 10 (average over 100 simulations). 98
Table 49 – Number of municipalities per year categorized as epidemic and non-epidemic 99
Table 50 – Distribution of the percentage of population in low-income situations for all the 126 municipalities 101
Table 51 – Distribution of GDP per capita (R$) for all the 126 municipalities . . . 102
Table 52 – Distribution of the percentage of population with sewage system for all the 126 municipalities 103
Table 53 – Distribution of the percentage of population living in urban areas for all the 126 municipalities 104
Table 54 – Distribution of the demographic density for all the 126 municipalities . . 105

List of abbreviations and acronyms

VLMCX	Variable Length Markov Chains with Exogenous Covariates
VLMC	Variable Length Markov Chains
GLM	Generalized Linear Model
WHO	World Health Organization
modified-beta-VLMC	Modified Beta-Context Algorithm
beta-VLMC	Beta-Context Algorithm
SINAN	Notifiable Diseases Information System (Sistema de Informação de Agravos de Notificação)
INMET	National Institute of Meteorology (Instituto Nacional de Metereologia)
NTD	Neglected Tropical Disease
GDP	Gross Domestic Product
IBGE	Brazilian Institute of Geography and Statistics (Instituto Brasileiro de Geografia e Estatística)
SARIMA	Seasonal Autoregressive Integrated Moving Average

Contents

Introduction **17**

1 Context Algorithm **19**

2 Variable Length Markov Chain: a statistical perspective **23**

2.1 Essential definitions to comprehend VLMC and VLMCX 23

2.2 Variable Length Markov Chain 26

2.2.1 Tree representation of minimal state space 26

2.2.2 Context algorithm from a statistical perspective 28

2.3 Variable Length Markov Chains with Exogenous Covariates (VLMCX) . . 30

2.3.1 The Beta-Context Algorithm 31

2.3.2 Consistency of the Beta-Context Algorithm 34

2.4 Variable Length Markov Chains with Time-Varying and Time-Invariant Exogenous Covariates 36

2.4.1 The Modified Beta-Context Algorithm 36

2.4.2 Highlighting modifications in Beta-Context Algorithm 40

2.4.3 Consistency of the Modified Beta-Context Algorithm 41

3 Simulations **43**

3.1 Simulations with a single source and without time-invariant exogenous covariates 44

3.2 Simulations with multiple independent sources and time-invariant exogenous covariates 56

4 Application: Predicting Dengue outbreaks **61**

4.1 Dengue: an overview of the disease and its global impact 61

4.2 Dengue as an endemic disease in Brazil 62

4.3 Dengue as a Neglected Tropical Diseases (NTDs) 62

4.4 Reviewing forecasting models for dengue incidence 63

4.4.1 Factors correlated to the number of dengue cases 64

4.4.2 Common forecasting methods for dengue incidence 66

4.5 Applying the VLMC with time-varying and time-invariant exogenous covariates for dengue outbreak prediction 68

4.5.1 Scenario 1: National Analysis - Brazil (January 2008 to July 2023) . 69

4.5.2 Scenario 2: City-Level Analysis - Rio de Janeiro, Brazil (January 2011 to September 2020) 76

5 Conclusion and future work **83**

BIBLIOGRAPHY **84**

Appendix 91
APPENDIX A Extra simulations 92
APPENDIX B Descriptive analysis - dengue incidence in Brazilian municipalities 99

Introduction

Stochastic chains with variable-length memory constitute an economic class of high-order Markov chains. These chains were introduced by Rissanen (1983) as a data compression tool, capable of efficiently compressing long strings without prior source knowledge. To accomplish this, the Context Algorithm was proposed to estimate adaptable tree-structured models by lumping together irrelevant states within the historical context of the process. Rissanen (1983) demonstrated that when the tree's order is bounded, regardless of the sample size, this algorithm consistently estimates both the context length and the corresponding transition probabilities.

Subsequently, Bühlmann e Wyner (1999) expanded upon these chains from a statistical perspective, leading to substantial theoretical advancements in both classical Markov chain techniques and their variable-length variants. They demonstrated that the context algorithm remains consistent even when the order of the chain is permitted to increase with the sample size.

In addition to full Markov chains with a finite order being one of the most comprehensive models for a stationary process, the statistical interest in Variable Length Markov Chains (VLMC) arises to address two challenges of full Markov chains that limit their suitability from an estimation perspective:

- **Problem 1**: The class of all finite-order Markov chains lacks structural richness, resulting in abrupt dimensionality increases as the order grows. This makes it difficult to achieve an effective trade-off between bias (minimized with numerous parameters) and variance (minimized with fewer parameters) for a predictor.
- **Problem 2**: The curse of dimensionality is particularly problematic when dealing with high-order models, as the dimensionality expands exponentially with the order.

Both of these issues can be resolved by allowing the memory of a stationary Markov chain to have variable length, which implies that some transition probabilities of the Markov chain are lumped together if they are equal.

Nonetheless, these methods often do not incorporate available time-dependent covariates, which can substantially influence transition probabilities. Numerous studies have explored Markov chains with exogenous covariates (for instance, MacRae (1977), Muenz e Rubinstein (1985), Azzalini (1994), López, Fernández e Velasco (1995), Cook e Ng (1997), Vermunt, Langeheine e Bockenholt (1999), Aalen, Borgan e Fekjær (2001), Heagerty (2002), Fokianos e Kedem (2003), Islam e Chowdhury (2006), Paroli e Spezia

(2007), Browning e Carro (2010), Meligkotsidou e Dellaportas (2011), Gao et al. (2017), Rubin et al. (2017), Sirdari e Islam (2018), Bray (2019), Liu et al. (2021), among others), but formal inference on Markov chains with variable length and exogenous covariates was a relatively unexplored area until the study conducted by Zambom, Kim e Garcia (2022), introducing the concept of Variable Length Markov Chains with Exogenous Covariates (VLMCX) and the beta-context algorithm.

VLMCX represents an extension of VLMC where the transition probabilities are influenced by the state history and exogenous covariates and are modeled using a Generalized Linear Model (GLM). The primary objective of VLMCX is not just to estimate the context of the process, which comprises the relevant historical information for predicting the next state, but also to estimate the coefficients associated with significant exogenous variables. Zambom, Kim e Garcia (2022) demonstrate the method's consistency, meaning that as the sample size increases, the probability of the estimated context and coefficients matching the true data-generating mechanism tends to 1.

The beta-context algorithm offers a solution selecting a relevant finite suffix, or context, for predicting the next symbol. This algorithm estimates flexible tree-structured models by aggregating irrelevant states in the process history and enables the model to incorporate exogenous covariates over time.

The primary goal of this study was to extend the beta-context algorithm to accommodate not only time-dependent but also time-invariant exogenous covariates, along with data collected from different sources. Despite utilizing historical data from multiple sources, our approach assumes that all sources share identical parameter estimates and we do not currently impose a spatial-dependent structure on the model. In addition to the adaptations for integrating time-invariant exogenous covariates and data from various sources, we encountered challenges in handling rare events and limited data during certain simulations. To address these issues, we introduced modifications to accommodate scenarios with limited data that may not be suitable for asymptotic parametric methods.

Our methodology was motivated by analysing a real dataset focused on monthly dengue cases across multiple municipalities in Brazil. Time-dependent covariates include temperature and precipitation levels of these regions, while time-invariant covariates include fixed attributes of municipalities, such as poverty rates and urban population percentages. Our primary aim is to investigate the impact of previous dengue rates - segmented into four distinct categories to establish a finite alphabet for the Markov chain -, weather conditions, and socioeconomic factors on subsequent dengue rates across various municipalities, providing insights into dengue transmission dynamics. This is crucial given the significant global threat posed by dengue fever, recognized by the World Health Organization (WHO), with an estimated 390 million dengue virus infections annually and its endemic status in over 100 countries.

1 Context Algorithm

The Context Algorithm was introduced by Rissanen (1983) as a data compression tool, capable of efficiently compressing long strings without prior source knowledge, and subsequently improved in the works of Rissanen (1986) and Furlan (1990). It focuses on capturing *contexts*, which are relevant subsets of the past string, along with the corresponding counts of conditional occurrences. Since all methods related to stochastic chains with variable-length memory rely on the Context Algorithm, we explore it in detail in this chapter.

To identify these relevant contexts, the algorithm integrates a parameter that assesses past symbols' influence on determining symbol values. In general, better data compression results from larger contexts, which increases both their quantity and the model's complexity. However, adding each new context carries a cost, which must be balanced against the additional compression gain it provides. Consequently, with a reasonable selection of the parameter, the model's complexity does not exceed that of the original data source.

The following is a detailed explanation of the algorithm for the binary case, as presented in Rissanen (1983):

Consider a permutation $i \rightarrow t_i$ of natural numbers, where for any string $s = y(1)...y(t-1)$, we define another string $\sigma(s) = y(t-t_1)...y(t-t_{t-1})$. In this discussion, we will consider only the identity permutation, where $t_i = i$, implying that $\sigma(s)$ is essentially the same as the sequence s but written in reverse order.

The next step involves the creation of two binary trees. One tree represents the scenario where the current symbol $y(t)$, denoted as u, equals 0, and the other tree represents the case where it equals 1:

1. Start with the initial context tree for the first symbol $y(1)$ in the string. This tree, denoted as $T(0)$, consists of a single leaf, which serves as the root. This root node is associated with a count pair of $(c(0,\kappa), c(1,\kappa)) = (1,1)$, where κ represents a empty string.

2. As you progress through the string, recursively build the subsequent trees. If $T(t-1)$ is the most recently constructed tree, and you observe the next symbol $u = y(t)$, create the next tree $T(t)$ as follows:

   - Traverse the tree $T(t-1)$, starting at the root. For each symbol in the past sequence $\sigma(y(1)...y(t-1)) = y(t-1)...y(1) = z_1 z_2 ... z_{t-1}$, take the left branch for 0 and the right branch for 1.

- While visiting each node z, increase the count $c(u, z)$ by one, and continue until you reach a node w where $c(u, w) = 1$ before the update.

3. When you encounter node w, consider the following scenarios:

- If w is an internal node with two successors, $w0$ on the left and $w1$ on the right, increase the component counts $c(u, w0)$ and $c(u, w1)$ by one. This adjustment defines the resulting tree as $T(t)$.
- If w is a leaf node, expand the tree by creating two new leaves, $w0$ and $w1$. Assign identical counts to both new leaves: $c(u, w0) = c(u, w1) = 1$, and set $c(u', w0) = c(u', w1) = 0$, where u' represents the opposite symbol to u. The modified tree is named $T(t)$.

For instance, let's examine the binary string 10001. Employing the identity permutation, we can depict the trees as illustrated in Figure 1.

The algorithm constructs a tree that accumulates significant contexts and their associated symbol statistics as the string length grows. However, which node should serve as the context for $y(t)$?

The algorithm addresses this by assigning a cost to each context. Let Z denote the set of leaves defining a complete subtree. A context is accepted into the set Z only if its impact on reducing conditional entropy exceeds its cost. This determination involves computing the change in conditional entropy that results from merging two elements, z_0 and z_1 into a parent node z. The new set created through this merging is termed Z'.

The increase in conditional entropy is given by:

$$\Delta(t, z) = H(U \mid Z') - H(U \mid Z) = P(z)H(U; z) - P(z0)H(U; z0) - P(z1)H(U; z1) \quad (1.1)$$

where $P(z) = c(z)/c(\kappa)$, with $c(\kappa)$ representing the root count and

$$P(u \mid z) = \begin{cases} c(u, z)/c(z), & \text{if} \quad c(u, z) > 0 \\ 1/(c(z) + 1), & \text{if} \quad c(u, z) = 0, \end{cases} \quad (1.2)$$

$$H(U; z) = -p\log(p) - (1-p)\log(1-p), \quad p = P(0 \mid z). \quad (1.3)$$

The context selection rule is as follows: the context $z^*(t)$ for the symbol $y(t)$ is determined as the node in $T(t-1)$ with the longest length $|z|$ along the path defined by the past sequence $z_1, z_2, ...$, satisfying the condition that:

$$\Delta(t, z) > (1/t)\log(t), \quad (1.4)$$

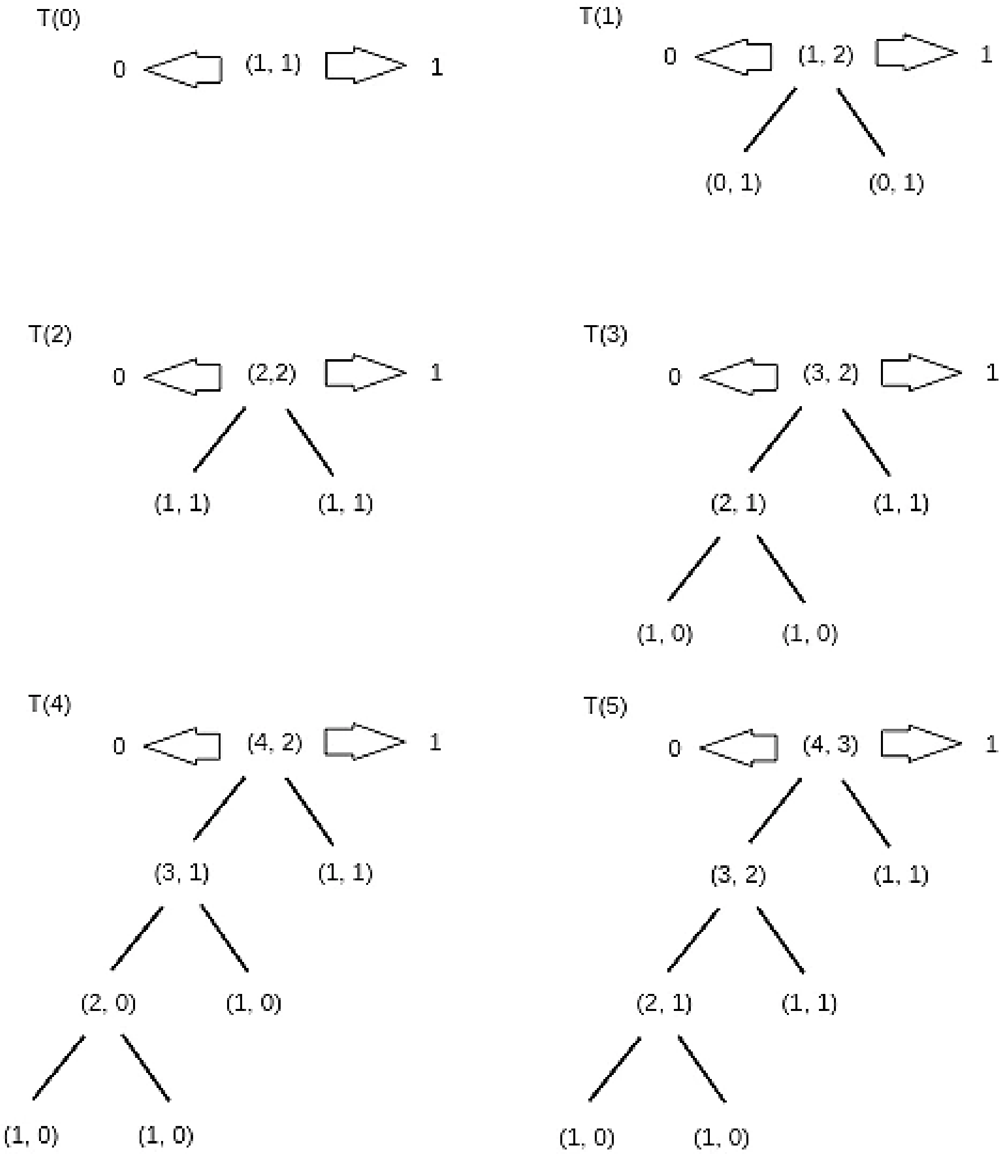


Figure 1 – Trees for the string 10001 using Context Algorithm

while

$$|z| \leqslant \beta \log(t), \tag{1.5}$$

and

$$\min\{c(z0), c(z1)\} \geqslant 2\alpha t/\sqrt{\log(t)}, \tag{1.6}$$

or, if no such node exists, take z to be the root node. Here, α and β are positive numbers chosen to determine a suitable range for searching nodes in a finite string. However, for infinite strings, any values assigned to them will do. This selection rule will minimize locally the combined cost

$$H(U/Z) + (|Z| \log(t))/t, \tag{1.7}$$

where $|Z|$ denotes the number of elements in Z.

For a more comprehensive understanding of this selection rule, the rationale behind assigning a cost of $(\log(t))/t$ to each context, and the proof demonstrating that the algorithm will find asymptotically any stationary ergodic finitely generated source from its samples, consult Rissanen (1983). When the tree's order is bounded, irrespective of the sample size, this algorithm consistently estimates both the context length and the corresponding transition probabilities.

2 Variable Length Markov Chain: a statistical perspective

Before exploring the description of VLMC by Bühlmann e Wyner (1999), it is essential to establish some fundamental definitions to comprehend VLMC, VLMCX, and the modified version proposed in this study. To achieve this, we will provide definitions in the broader scenario, considering exogenous time-dependent and time-invariant covariates, as well as multiple independent sources. The definitions for VLMC and VLMCX are specific cases derived from these broader scenarios.

2.1 Essential definitions to comprehend VLMC and VLMCX

Let $\mathscr{Y} = \{1, ..., p\}$ represent the finite set of possible state spaces and $\mathscr{S} = \{1, ..., q\}$ denote the finite set of potential sources. For each independent source [1], $s \in \mathscr{S}$, consider a stochastic process $(Y_t(s))_{t\in\mathbb{Z}}$ with values in the finite state space $\mathscr{Y}$. Denote by $y_i^j = y_j, y_{j-1}, ..., y_i$ $(i < j, i, j \in \mathbb{Z} \cup \{-\infty, \infty\}, y_j \in \mathscr{Y})$ the string that represents the states visited by the process from time i to time j. Notice that it is written in reverse order.

Suppose that the transition probabilities of the process depends on the previous states through a set of parameters, d time-varying exogenous covariates and m time-invariant exogenous covariates (fixed characteristics). For each source, $s \in \mathscr{S}$, denote the b-th time-varying covariate value at time t by $x_{tb}(s)$, $b = 1, ..., d$, and let $\mathbf{x}_t(s) = (x_{t1}(s), ..., x_{td}(s))$. Also, denote the v-th time-invariant covariate value by $z_v(s)$, $v = 1, ..., m$, and let $\mathbf{z}(s) = (z_1(s), ..., z_m(s))$.

Further, define

$$\mathbf{x}_i^j(s) = (1, \mathbf{x}_j(s), \mathbf{x}_{j-1}(s), ..., \mathbf{x}_i(s))$$

as the vector of time-varying covariates for source s from time i to time j. In order to estimate the intercept of the regression we need to include the number one. Notice that we are considering all the covariates to be deterministic.

The proposed model (given by 2.1) writes the transition probabilities into the p states by a multinomial linear regression with parameters that can depend on the previous history.

[1] In our motivating example, where we aim to predict dengue rates in Brazil, the sources refer to different geographic locations, specifically municipalities.

Definition 1 (Context)**.** *For each independent source, $s \in \mathscr{S}$, let $(Y_t(s))_{t\in\mathbb{Z}}$ be a stationary process with values $Y_t(s) \in \mathscr{Y}$, $(\mathbf{x}_t(s))_{t\in\mathbb{Z}}$ a d-dimensional vector of deterministic time-varying exogenous covariates and $(\mathbf{z}(s))$ a m-dimensional vector of deterministic time-invariant exogenous covariates, both in a compact set. Denote by $c : \mathscr{Y}^\infty \to \bigcup_{l=0}^{\infty} \mathscr{Y}^l$ a (projection) function which maps $c : y_{-\infty}^0 \to y_{-l+1}^0$, where l is defined by*

$$\begin{aligned} l = l(y_{-\infty}^0) &= min\{k; \prod_{s=1}^{q} P[Y_1(s) = y_1 \mid Y_{-\infty}^0(s) = y_{-\infty}^0, \mathbf{x}_{-\infty}^0(s), \mathbf{z}(s)] \\ &= \prod_{s=1}^{q} P[Y_1(s) = y_1 \mid Y_{-k+1}^0(s) = y_{-k+1}^0, \mathbf{x}_{-k+1}^0(s), \mathbf{z}(s)] \\ &\quad \textit{for all } \mathbf{x}_{-\infty}^0(s), \mathbf{z}(s) \textit{ and for all } y_1 \in \mathscr{Y}\}, \end{aligned}$$

where $l \equiv 0$ corresponds to independence.

Letting $u := y_{-l+1}^0$ and $\pi_j := P_{\boldsymbol{\theta}}(Y_1 = j \mid Y_{-l+1}^0 = u, \mathbf{x}_{-l+1}^0, \mathbf{z})$, where $\mathbf{x}_{-l+1}^0$ and $\mathbf{z}$ will represent the values of the covariates according to each source,

$$\boldsymbol{\pi} = \mathbf{g}(\mathbb{X}^\intercal \boldsymbol{\theta}^u), \tag{2.1}$$

where

$$\mathbb{X} = \begin{bmatrix} (\mathbf{x}_{-l+1}^0, \mathbf{z})^\intercal & 0 & \dots & 0 \\ 0 & (\mathbf{x}_{-l+1}^0, \mathbf{z})^\intercal & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & (\mathbf{x}_{-l+1}^0, \mathbf{z})^\intercal \end{bmatrix}, \boldsymbol{\theta}^u = \begin{bmatrix} \boldsymbol{\theta}_1^u \\ \vdots \\ \boldsymbol{\theta}_p^u \end{bmatrix},$$

$\boldsymbol{\pi} = (\pi_1, ..., \pi_p)$. Note that $\mathbb{X}^\intercal$ is a p by $(1 + dl + m)p$ matrix and $\boldsymbol{\theta}^u$ is a $(1 + dl + m)p$ dimensional vector. The link function $\mathbf{g}$ is a one-to-one mapping from a p-dimensional region $D \subset \mathbb{R}^p$ to the set $\{(\pi_1, ..., \pi_p)^\intercal, \pi_j > 0, \sum \pi_j = 1\}$. We denote by $\boldsymbol{\theta} := \boldsymbol{\theta}^u = (\boldsymbol{\theta}_1^u, ..., \boldsymbol{\theta}_p^u)^\intercal$, with $\boldsymbol{\theta}_j^u = (\alpha_j^u, \boldsymbol{\beta}_{j,0}^u, ..., \boldsymbol{\beta}_{j,(-l+1)}^u, \boldsymbol{\gamma}_j^u)^\intercal$, the vector of coefficients associated with the past states $u = y_{-l+1}^0$ for transitioning into state $j \in \mathscr{Y}$, where $\boldsymbol{\beta}_{j,t}^u = (\beta_{j,t1}^u, ..., \beta_{j,td}^u)^\intercal$ is the vector of coefficients corresponding to the d time-varying covariates at time $t = 0, ..., -l + 1$ and $\boldsymbol{\gamma}_j^u = (\gamma_{j,1}^u, ..., \gamma_{j,m}^u)^\intercal$ is the vector of coefficients corresponding to the m time-invariant covariates.

Then $c(\cdot)$ is called the beta-context function, and $c(y_{-\infty}^0)$ is called the beta-context for the transition at time 1 with associated parameter vector $\boldsymbol{\theta}^u$.

Remark 1. *If the link function $\mathbf{g}$ is chosen to be multinomial logistic (softmax function), then*

$$\begin{aligned} \pi_j &:= P_{\boldsymbol{\theta}}(Y_1 = j \mid Y_{-l+1}^0 = u, \mathbf{x}_{-l+1}^0, \mathbf{z}) \\ &= \frac{exp(\alpha_j^u + \sum_{t=-l+1}^{0} \sum_{b=1}^{d} x_{tb} \beta_{j,tb}^u + \sum_{v=1}^{m} z_v \gamma_{j,v}^u)}{\sum_{i=1}^{p} exp(\alpha_i^u + \sum_{t=-l+1}^{0} \sum_{b=1}^{d} x_{tb} \beta_{i,tb}^u + \sum_{v=1}^{m} z_r \gamma_{j,v}^u)} \quad \textit{for all} \quad j \in \mathscr{Y}. \end{aligned}$$

Definition 2 (Order of the time-varying covariate parameters)**.** *Consider a parameter vector $\boldsymbol{\beta}_j^u$, $j = 1, ..., p$ associated with a beta-context u. The length of $\boldsymbol{\beta}_j^u$, which represents the number of steps where covariate values have a significant contribution to the model, is defined as*

$$h := h_u = \left|\boldsymbol{\beta}_j^u\right| = 1 - \underset{k=0,...,-l+1}{min}\{k : \boldsymbol{\beta}_{j,k}^u \neq 0\}, j = 1, ..., p.$$

If $\boldsymbol{\beta}_{j,0}^u = ... = \boldsymbol{\beta}_{j,(-l+1)}^u = 0$, then $\left|\boldsymbol{\beta}_j^u\right| = 0$.

The proposed model accommodates situations where $h \leqslant l$, including cases where $h < l$, that is, the transition probability may rely on a more extended history of state transitions and time-invariant covariates while considering only the more recent history of time-varying covariates.

Definition 3 (Order of the state transitions and time-invariant covariates)**.** *For each independent source, $s \in \mathscr{S}$, let $(Y_t(s))_{t\in\mathbb{Z}}$, $(\mathbf{x}_t(s))_{t\in\mathbb{Z}}$, $\mathbf{z}(s)$, $c(\cdot)$ and $l(\cdot)$ be defined as in Definition 1. Let $0 \leqslant \eta < \infty$ be the smallest integer such that*

$$\left|c(y_{-\infty}^0)\right| = l(y_{-\infty}^0) \leqslant \eta, \quad \textit{for all} \quad y_{-\infty}^0 \in \mathscr{Y}^\infty.$$

Then $c(\cdot)$ is called a beta-context function of order η and we have a beta-context model of order η.

Definition 4 (Beta-context tree)**.** *Let $c(\cdot)$ be a beta-context function of a beta-context model of order η. The ($|\mathscr{Y}| - ary$) beta-context rooted tree τ is defined as*

$$\tau := \tau_c = \{u : u = c(y_{-\eta+1}^0), y_{-\eta+1}^0 \in \mathscr{Y}^\eta\}$$

with an associated parameter tree

$$\tau_\theta = \{(u, \boldsymbol{\theta}^u) : u \in \tau\}$$

where $\boldsymbol{\theta}^u$ is defined in Definition 1.

Besides the definitions above, to understand the statistical perspective of the context algorithm and the beta-context algorithm, it is essential to have a clear understanding of the terms siblings, children, and parents used in this text, which are defined as follows.

Definition 5 (Siblings, children and parents)**.** *Let $u_1 = uw \in \tau$ and $u_2 = uw' \in \tau$, for w, $w' \in \mathscr{Y}$ and $u \in \mathscr{Y}^\infty$, be two contexts differing only by the last nodes. Then u_1 and u_2 are called siblings in τ and this relationship is denoted by $u_1 \wr u_2$. In addition, u is called the parent of u_1 and u_2 and u_1 and u_2 are called the children of u.*

2.2 Variable Length Markov Chain

Consider a stationary Markov chain $(Y_t)_{t\in\mathbb{Z}}$ of finite order η with values in a finite categorical space $\mathscr{Y}$. Thus,

$$P[Y_1 = y_1 \mid Y_{-\infty}^{0} = y_{-\infty}^{0}] = P[Y_1 = y_1 \mid Y_{-k+1}^{0} = y_{-k+1}^{0}] \text{ for all } y_{-\infty}^{1} \in \mathscr{Y}^{\infty}. \tag{2.2}$$

A variable length memory allows us to lump together irrelevant states in the history y_{-k+1}^{0} in 2.2. In that regard, only some values from the history are relevant and can be considered a *context* for Y_1. These *contexts* are defined similarly to Definition 1, but without time-varying and time-invariant covariates and with only one source, meaning $\mathscr{S} = 1$. Therefore:

$$\begin{aligned} l = l(y_{-\infty}^{0}) = \min\{k; P[&Y_1 = y_1 \mid Y_{-\infty}^{0} = y_{-\infty}^{0}] \\ &= P[Y_1 = y_1 \mid Y_{-k+1}^{0} = y_{-k+1}^{0}] \text{ for all } y_1 \in \mathscr{Y}\}. \end{aligned} \tag{2.3}$$

Here, $c(\cdot)$ is referred to as a context function, and for any $t \in \mathbb{Z}$, $c(y_{-\infty}^{t-1})$ is considered a context for the variable y_t.

The order of the VLMC is defined similarly to Definition 3. Consequently, $c(\cdot)$ is called a context function of order η, and if $\eta < \infty$, we have a VLMC of order η.

As a solution to Problem 1, mentioned earlier in the Introduction, the class of context functions with order η is structurally diverse enough to include a wide range of Markov chains. Moreover, in addressing Problem 2, certain context functions $c(\eta)$ significantly reduce the number of states compared to a full Markov chain of the same order. When the context function $c(\eta)$ of order η corresponds to the complete projection $y_{-\infty}^{0} \to y_{-k+1}^{0}$ for all $y_{-\infty}^{0}$, the VLMC is essentially a full Markov chain of order η.

2.2.1 Tree representation of minimal state space

VLMCs are commonly characterized by their probability distribution P_c on $\mathscr{Y}^{\mathbb{Z}}$ and, due to stationarity, can be entirely specified by their transition probabilities,

$$\mathbb{P}_{P_c}[Y_1 = y_1 \mid Y_{-\infty}^{0} = y_{-\infty}^{0}] = p(y_1 \mid c(y_{-\infty}^{0})),\ y_{-\infty}^{0} \in \mathscr{Y}^{\infty}. \tag{2.4}$$

The states that determine these transition probabilities are represented by the context function's values $c(\cdot)$. To facilitate this representation, it is convenient to visualize these states using a tree structure, where the ($\mathscr{Y}$-ary) context tree τ can be defined similarly to Definition 4 and terminal node context tree τ^T is defined as

$$\tau^T = \tau_c^T = \{w; w \in \tau_c \text{ and } wu \notin \tau_c \text{ for all } u \in \mathscr{Y}\}. \tag{2.5}$$

Trees are built following the steps:

1. Start with a root node positioned at the top of the tree.

2. The branches will grow downward from this root node so that each internal node has at most $|\mathscr{Y}|$ children.

3. Each value of a context function $c(\cdot) : \mathscr{Y}^{-\infty} \to \mathscr{Y}^{\eta}$ can be represented as a branch (or terminal node) on the tree.

4. The context $w = c(y_{-\infty}^{0})$ is represented by a branch, where each branch level represents a different part of the context. The topmost subbranch corresponds to y_0, the next subbranch to y_{-1} and so on. The terminal subbranch corresponds to $y_{-l(y_{-\infty}^{0})+1}$. Context trees do not have to be complete, meaning that not every internal node must have precisely $|\mathscr{Y}|$ children nodes.

Note that only the terminal nodes within the tree τ are recognized as elements in the terminal node context tree τ^{T}. The states $w \in \tau_c$ are not required to be terminal nodes in τ_c. However, it is possible to reconstruct the context function $c(\cdot)$ from either τ_c or τ_c^{T}.

The context tree τ_c essentially serves as the minimal state space for the VLMC P_c. An internal node with $b < N = |\mathscr{Y}|$ children nodes implicitly lumps the non-present children nodes $N - b$ together into a single new terminal node. This terminal node, in turn, represents a single state within τ_c.

The following example is drawn from Bühlmann e Wyner (1999).

Example 1. $\mathscr{Y} = \{0, 1, 2, 3\}, \eta = 2$. *The function,*

$$c(y_{-\infty}^{0}) = \begin{cases} 0, & \textit{if} \quad y_0 = 0,\ y_{-\infty}^{-1} \quad \textit{arbitrary}, \\ 1, & \textit{if} \quad y_0 = 1,\ y_{-\infty}^{-1} \quad \textit{arbitrary}, \\ 2, & \textit{if} \quad y_0 = 2,\ y_{-\infty}^{-1} \quad \textit{arbitrary}, \\ 3, & \textit{if} \quad y_0 = 3,\ y_{-1} \in \{0,1,2\}, y_{-\infty}^{-2} \quad \textit{arbitrary}, \\ 3, 3, & \textit{if} \quad y_0 = 3,\ y_{-1} = 3, y_{-\infty}^{-2} \quad \textit{arbitrary} \end{cases}$$

can be represented by the tree τ_c on Figure 2. The round-edge rectangle, which is typically omitted, represents the missing nodes 0, 1 and 2 at a depth of 2. This can be seen as a way of completing the tree by lumping nodes together. In terms of transition probabilities, it means that $p(y \mid 3z), y \in \mathscr{Y}$, is the same for all $z \in \{0, 1, 2\}$. The terminal nodes context tree is $\tau_c^{T} = \{0, 1, 2, 33\}$, whereas the context tree is $\tau_c = \{0, 1, 2, 3, 33\}$.

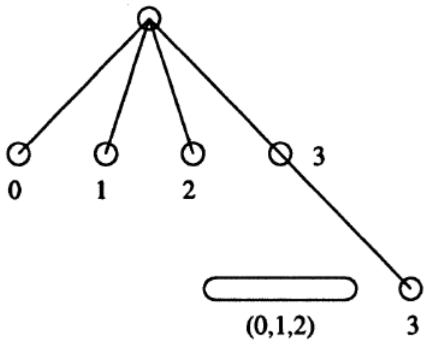


Figure 2 – Tree representation of the context function in Example 1

Source: Bühlmann e Wyner (1999)

2.2.2 Context algorithm from a statistical perspective

Let

$$N(w) = \sum_{t=1}^{n} \mathbb{1}_{[Y_t^{t+|w|-1}=w]}, \quad w \in \bigcup_{m=1}^{\infty} \mathscr{Y}^m, \tag{2.6}$$

denote the number of occurrences of the string w in the sequence Y_1^n and

$$\hat{P}(w) = \frac{N(w)}{n}, \quad \hat{P}(y \mid w) = \frac{N(yw)}{N(w)}, \quad y, w \in \bigcup_{m=1}^{\infty} \mathscr{Y}^m. \tag{2.7}$$

Based on the methodology presented in Bühlmann e Wyner (1999), the context algorithm operates according to the following steps:

1. **Step 1 - Finding Maximal Context Tree:**
   - Given $\mathscr{Y}$-valued data $Y_1, ..., Y_n$, fit a maximal context tree, that is, search for the context function $c_{\max}(\cdot)$ with terminal node context tree representation $\tau_{\max}^T$.
   - $\tau_{\max}^T$ is the bigest tree such that every element (terminal node) w in it must have been observed at least twice in the data ($N(w) \geqslant 2$). Additionally, for any other tree τ^T, where $w \in \tau^T$ implies $N(w) \geqslant 2$, τ^T is a sub-tree of $\tau_{\max}^T$ ($\tau^T \preceq \tau_{\max}^T$, which means that $w \in \tau^T \Rightarrow wu \in \tau_{\max}^T$ for some $u \in \cup_{m=0}^{\infty} \mathscr{Y}^m (\mathscr{Y}^0 = \varnothing)$). This maximal tree may not necessarily be full, meaning that the final nodes' contexts may have varying lengths.
   - Set $\tau_{(0)}$ as $\tau_{\max}$.
2. **Step 2 - Pruning the Tree:**

- Examine if each element (terminal node) of $\tau_{(0)}^T$ with a context function $c(\cdot)$ can be pruned (the order of examining is irrelevant).
- Let

$$wu = y_{-l+1}^0 = c(y_{-\infty}^0), \quad u = y_{-l+1}, \quad w = y_{-l+2}^0,$$

  where wu is an element (terminal node) of $\tau_{(0)}$. Prune wu to w (if $l = 1$, the pruned version is the empty branch $\varnothing$, that is, the root node) if

$$\Delta_{wu} = \sum_{y \in \mathscr{Y}} \hat{P}(y \mid wu) \log \left(\frac{\hat{P}(y \mid wu)}{\hat{P}(y \mid w)} \right) N(wu) < K,$$

  with $K = K_n \sim C \log(n), C > 2|\mathscr{Y}| + 4$ a cutoff to be chosen by the user and $\hat{P}(\cdot \mid \cdot)$ as defined in 2.7.
- The outcome of this step is a potentially smaller tree, $\tau_{(1)} \preceq \tau_{(0)}^T$. Construct the terminal node context tree $\tau_{(1)}^T$. Note that the context tree do not have to be complete, that is, every internal node does not need to have exactly $|\mathscr{Y}|$ offspring.

3. **Step 3 - Iterative Pruning:**

   - Repeat the pruning process from Step 2 with $\tau_{(i)}$, $\tau_{(i)}^T$ instead of $\tau_{(i-1)}^T$ $(i =, 2, ...)$.
   - Continue this iterative pruning until no more elements can be pruned.
   - Denote this maximal pruned context tree by $\hat{\tau} = \tau_{\hat{c}}$ and its corresponding context function by $\hat{c}(\cdot)$.

4. **Step 4 - Probability Estimation:**

   - If interested in probability distributions, estimate the transition probabilities $p(y_1 \mid c(y_{-\infty}^0))$ by $\hat{P}(y_1 \mid \hat{c}(y_{-\infty}^0)$, where $\hat{P}(\cdot \mid \cdot)$ is defined as in 2.7.

The decision to prune in Step 2 is similar to a likelihood ratio test:

- Denote the estimated likelihood function (conditioned on the first η states), based on the context function $c(\cdot)$ by

$$\hat{P}_c(Y_1^n) = \prod_{t=\eta+1}^{n} \hat{P}(Y_t \mid c(Y_{-\infty}^{t-1})), \tag{2.8}$$

  where η is the order of $c(\cdot)$ and $\hat{P}(Y_t \mid c(Y_{-\infty}^{t-1}))$ in defined in 2.7.
- Now, we have the context function $c(\cdot)$ of the unpruned context tree and $c'(\cdot)$ of the subtree after pruning a single terminal node $wu = y_{-l+1}^0$ to its parent node $w = y_{-l+2}^0$.

- In the likelihood ratio test, which is essentially the pruning criterion, many terms in the likelihood function cancel out due to its multiplicative structure. What remains is a term specifically associated with the node under consideration for pruning:

$$\Delta_{wu} = \log \frac{\hat{P}_c(Y_1^n)}{\hat{P}_{c'}(Y_1^n)}. \tag{2.9}$$

In essence, the pruning criterion, as described in the 2.9, is fundamentally a likelihood ratio test. The algorithm effectively conducts numerous likelihood ratio tests to determine which nodes should be pruned.

Still in Step 2, the choice of the cutoff value $K_n \sim C\log(n)$ is determined through an asymptotic perspective. It is important to note that small cutoff values would lead to the construction of large context trees, increasing the risk of overfitting the data. The estimation of this cutoff value is elaborated in detail in a related work by Bühlmann (2000).

It is worth highlighting that the algorithm imposes no a priori length restrictions on contexts. This contrasts with approaches like the one proposed by Weinberger, Rissanen e Feder (1995), which used context lengths limited by $\log(n)/\log(|\mathscr{Y}|)$. Such restrictions can be quite limiting in practical applications.

For more detailed information and proofs regarding the algorithm's consistency when dimensionality increases, refer to Bühlmann e Wyner (1999). Two important results are presented: the first demonstrates the consistency in discovering the minimal state spaces, while the second outlines properties of the estimated probability distributions.

2.3 Variable Length Markov Chains with Exogenous Covariates (VLMCX)

Let $(Y_t)_{t\in\mathbb{Z}}$ be a stochastic process taking values in the finite state space $\mathscr{Y} = \{1, ..., p\}$. Suppose that the transition probabilities of the process may depend on the previous states through a set of parameters and d time-variant exogenous covariates. Therefore, VLMCX extends VLMC by incorporating the influence of both state history and exogenous covariates on transition probabilities.

VLMCX was introduced by Zambom, Kim e Garcia (2022) through the *beta-context* algorithm. This algorithm constructs a context model and generates a context tree, where each branch represents a history of response states and is associated with parameters defining transition probabilities within a Generalized Linear Model (GLM). The term *beta* in *beta-context* refers to the parameters in the GLM.

Beta-context algorithm uses the concepts defined in Definition 1, Definition 2, Definition 3, and Definition 4, with the exception of time-invariant covariate parameters and considering only one source, denoted as $\mathscr{S} = 1$. Note that VLMCX accommodates situations where $h \leqslant l$, including cases in which the transition probability may rely on a more extended history of state transitions while considering only the recent history of covariates. If all coefficients $\boldsymbol{\beta}_j^u$, $j = 1, ..., p$, are set to zero, the model essentially becomes a VLMC.

2.3.1 The Beta-Context Algorithm

In this section we will outline the process of identifying the underlying beta-context function $c(\cdot)$ and the transition probability parameters $\boldsymbol{\theta}^u$ based on the methodology presented in Zambom, Kim e Garcia (2022).

The beta-context algorithm is a backward elimination procedure based on previous approaches such as Bühlmann e Wyner (1999) and Rissanen (1983). It iteratively prunes the final nodes based not only on the significance of the context but also on the potential influence of exogenous covariates as determined by the coefficients $\boldsymbol{\beta}$.

The likelihood of the data Y_1^n, conditioning upon knowing y_1^η and $\boldsymbol{x}_1^\eta$, based on Definition 1 under a beta-context tree of order η is

$$\begin{aligned} &L(\tau_\theta; Y_1^n = y_1^n, \boldsymbol{x}_1^n) = P(Y_1^n = y_1^n \mid Y_1^\eta = y_1^\eta, \boldsymbol{x}_1^\eta, \tau_\theta) \\ &= P(Y_{\eta+1} = y_{\eta+1} \mid Y_1^\eta = y_1^\eta, \boldsymbol{x}_1^\eta, \tau_\theta) \; ... \; P(Y_n = y_n \mid Y_1^{n-1} = y_1^{n-1}, \boldsymbol{x}_1^{n-1}, \tau_\theta), \end{aligned} \tag{2.10}$$

where each probability follows from 2.1[2]. For each context $u = y_{-k+1}^0 = y_0, y_{-1}, ..., y_{-k+1}$ with k steps into the past there are $(1 + dk)$ parameters to estimate for each $j = 1, ..., p - 1$. This includes one parameter corresponding to α_j^u and dk parameters associated with the d exogenous covariates at each of the k steps. To ensure the likelihood ratio test in the algorithm performs effectively, a minimum of $f \geqslant 1$ observations per parameter must be available. Therefore, the algorithm needs a total of $f(1 + dk)(p - 1)$ observations to estimate the $(1 + dk)(p - 1)$ parameters for each context u.

Given data $(Y_1^n, \boldsymbol{x}_1^n)$, the algorithm follows these steps:

1. **Step 1 - Finding Maximal Context Tree:**

   - Similarly to Step 1 of context algorithm by Bühlmann e Wyner (1999), start by fitting the largest possible beta-context tree, denoted as $\tau_{\max}$. However, in this case, every element $u \in \tau_{\max}$ must be observed a minimum of $f(1 + dk)(p - 1)$

[2] A correction has been made to 2.10 in comparison to the original formulation presented in Zambom, Kim e Garcia (2022).

times. To define this tree, identify the beta-context function $c(\cdot)$ such that its corresponding context tree $\tau_{\max}$ is formed as follows:

$$\tau_{\max} = \{u = y_{-k+1}^{0} : N(u) \geqslant f(1 + dk)(p - 1)\},$$

where $N(u)$ is defined in 2.6.

- Set the initial beta-context tree as $\tau^{(0)} = \tau_{\max}$. Let r denote the order of this context tree. This initial tree $\tau^{(0)}$ may not necessarily be full, meaning that the final nodes' contexts may have varying lengths.
- Compute $\hat{\tau}_{\theta}^{(0)}$, which is the associated estimated parameter tree. The parameters in this tree are estimated by maximizing the likelihood function $L(\tau_{\theta}^{(0)} \mid y_1^n, \boldsymbol{x}_1^n)$.

2. **Step 2 - Pruning the Tree (Influence of the exogenous covariates):**

- For each context u in the initial tree $\tau^{(0)}$ with length r, apply a likelihood ratio test to examine the significance of the parameter vector associated with past covariates from the node $(-r + 1)$ to any of the outcomes (excluding the baseline). This test is expressed as:

$$H_0^u : \boldsymbol{\beta}_{j,(-r+1)}^{u} = 0, j = 1, ..., p - 1.$$

- Compute the deviance statistic λ_{-r+1}^{u} for testing H_0^u:

$$\lambda_{-r+1}^{u} = -2[\log L(\tilde{\tau}_{\theta}^{u} \mid y_1^n, \boldsymbol{x}_1^n) - \log L(\hat{\tau}_{\theta}^{(0)} \mid y_1^n, \boldsymbol{x}_1^n)] \tag{2.11}$$

where $L(\cdot)$ represents the likelihood function as defined in 2.10.

- Calculate estimators

$$\begin{aligned} \tilde{\tau}_{\theta}^{u} = \{&(w, \tilde{\boldsymbol{\theta}}^{w}) : w \in \tau^{(0)}, w \neq u\} \\ &\cup \{(u, (\tilde{\alpha}_j^u, \tilde{\boldsymbol{\beta}}_{j,0}^{u}, \tilde{\boldsymbol{\beta}}_{j,-1}^{u}, ..., \tilde{\boldsymbol{\beta}}_{j,(-r+2)}^{u}, 0)), j = 1, ..., p - 1\} \end{aligned} \tag{2.12}$$

by maximizing the likelihood under the null hypothesis H_0^u.

- Compute the p-value $\pi_{-r+1}^{u} = 1 - \Psi_{d(p-1)}(\lambda_{-r+1}^{u})$, where $\Psi_{d(p-1)}(\cdot)$ is the cumulative distribution function of a χ^2 random variable with $d(p - 1)$ degrees of freedom.
- If $\pi_{-r+1}^{u} > \delta_n$, where δ_n is a chosen significance level, update the estimated parameter tree $\hat{\tau}_{\theta}^{(0)}$ with $\tilde{\tau}_{\theta}^{u}$.

3. **Step 3 - Pruning the Tree (Influence of the context):** Concerning the tests performed in **Step 2**:

a) **Substep 3.1** - If no $H_0^{u_k}, k = 1, ..., p$, for $u_1, ..., u_p$ siblings in $\tau_{(0)}$, was rejected:

- Calculate the test statistic

$$\lambda^{u_{1\ldots p}} = -2[\log L(\tilde{\tau}_\theta^{u_{1\ldots p}} \mid y_1^n, \boldsymbol{x}_1^n) - \log L(\hat{\tau}_\theta^{(0)} \mid y_1^n, \boldsymbol{x}_1^n)], \tag{2.13}$$

  where $\tilde{\tau}_\theta^{u_{1\ldots p}} = \{(w, \tilde{\boldsymbol{\theta}}^w) : w \in \tau^{(0)}, w \neq u_1, ..., u_p\} \cup \{(u, \tilde{\boldsymbol{\theta}}^w) : u$ is parent of $u_1, ..., u_p\}$ is estimated under the null hypothesis that the sibling nodes are all replaced by their parent, reducing the parameters. The parameters in vectors $\boldsymbol{\theta}^{u_1}, ..., \boldsymbol{\theta}^{u_p}$, which have been together reduced to size $\mathbb{R}^{(1+d(r-1))p(p-1)}$ in Step 2, are merged into parameters $\boldsymbol{\theta}^u$ of size $\mathbb{R}^{(1+d(r-1))(p-1)}$.

- Compute the p-value $\pi^{u_{1\ldots p}} = 1 - \Psi_{(1+d(r-1))(p-1)^2}(\lambda^{u_{1\ldots p}})$. If $\pi^{u_{1\ldots p}} \geqslant \delta_n$, replace sibling nodes $u_1, ..., u_p$ with their parent node u. Update both $\tau^{(0)}$ and $\hat{\tau}_\theta^{(0)}$ with $\tilde{\tau}_\theta^{u_{1\ldots p}}$.

b) **Substep 3.2** - If at least one of $H_0^{u_k}, k = 1, ..., p$, for $u_1, ..., u_p$ siblings in $\tau^{(0)}$, was rejected:

- The nodes corresponding to the rejected tests are retained in the model.
- Test if the nodes corresponding to tests that were not rejected in Step 2 can be lumped together. Use the test statistic defined in 2.13 to do this.
- Lastly, sequentially test to prune the past most parameters in $\boldsymbol{\beta}^{u_k}$ (for $k = 1, ..., p$) which had its hypothesis not rejected in Step 2 up to the root. Consider $k = 1$, $u_1 = y_{-r+2}^0$ so that $\boldsymbol{\beta}_j^{u_1} = (\boldsymbol{\beta}_{j,0}^{u_1}, ..., \boldsymbol{\beta}_{j,(-r+2)}^{u_1}), j = 1, ..., p-1$.
  - Test the null hypothesis $H_{0m}^{u_1} : \boldsymbol{\beta}_{j,m}^{u_1} = 0$ for all $j = 1, ..., p-1$, where $m = -r+2$ is the past most index of $\boldsymbol{\beta}_j^{u_1}$. Calculate the corresponding test statistic $\lambda_{-r+2}^{u_1}$ using 2.11.
  - Compute the p-value $\pi_{-r+2}^{u_1} = 1 - \Psi_d(\lambda_{-r+2}^{u_1})$. If $\pi_{-r+2}^{u_1} \geqslant \delta_n$, prune the past most parameters from $\boldsymbol{\beta}_j^{u_1}$ and set $m = -r+3$. Continue this process for parameters $m = -r+2, ..., 0$ or until the parameters are reduced to zero, keeping all parameters that do not meet the criteria.
  - Repeat the above process for $\boldsymbol{\beta}^{u_k}$ with $k = 2, ..., p$.

At the end of Step 3, the tree structure and parameters are possibly pruned based on the significance of context and covariate influence, leading to a potentially smaller context tree $\tau^{(1)} \subseteq \tau^{(0)}$ and its updated parameter tree $\hat{\tau}_\theta^{(1)}$ at level r.

4. **Step 4 - Iterative Pruning:**

- Repeat Steps 2 and 3 for contexts of lengths $r-1, r-2, ..., 1$, using the updated trees $\tau^{(1)}$ and $\hat{\tau}_\theta^{(1)}$.
- If $u_1, ..., u_p$ are siblings and at least one of them has children, no context pruning is performed (both nodes are retained). As in Step 3, the pruning of covariate parameters is executed sequentially, working from the most distant past to the root.

Denote this pruned beta-context tree by $\hat{\tau}_n$ with associated parameter tree $\hat{\tau}_\theta$ and corresponding beta-context function $\hat{c}(\cdot)$.

The pruning approach in Step 3 differs from that of Bühlmann e Wyner (1999). In their method, each terminal node wu_j, $j = 1, ..., p$, for $u_1, u_2, ..., u_p$ siblings in the respective tree, is individually compared with its pruned version $w = y^0_{-l+2}$, where l represents the context length. This involves comparing the probability $\hat{P}(y \mid wu_j)$ with $\hat{P}(y \mid w)$ for each terminal node wu_j, $j = 1, ..., p$. In this case, the procedure never leaves only one terminal node in a branch. To illustrate this, consider a scenario with three terminal nodes: wu_1, wu_2, and wu_3. If it happens that $\hat{P}(y \mid wu_1) = \hat{P}(y \mid w)$ and $\hat{P}(y \mid wu_2) = \hat{P}(y \mid w)$, it logically follows that $\hat{P}(y \mid wu_3) = \hat{P}(y \mid w)$.

In contrast, in the approach presented in Zambom, Kim e Garcia (2022), we test whether all terminal nodes that had a parameter β cut in the previous step can be merged together. To achieve this, consider three terminal nodes, wu_1, wu_2 and wu_3, where only wu_1 and wu_2 had their beta values cut in the previous step. In this case, we compare the probability $\hat{P}(y \mid wu_1)$ with $\hat{P}(y \mid wu_2)$. Consequently, node merging occurs when there is no rejection of the hypothesis that $\hat{P}(y \mid wu_1) = \hat{P}(y \mid wu_2)$, even if $\hat{P}(y \mid wu_1) \neq \hat{P}(y \mid w)$ and $\hat{P}(y \mid wu_2) \neq \hat{P}(y \mid w)$. After merging nodes wu_1 and wu_2, we need to estimate parameters for the combined node, considering the scenarios where either u_1 or u_2 precede w.

2.3.2 Consistency of the Beta-Context Algorithm

Zambom, Kim e Garcia (2022) demonstrated that the beta-context algorithm ensures strong consistency in estimating the beta-context tree, regression parameters, and transition probabilities (Theorem 1) under conditions C1-C3 and A1-A4:

C1: $\delta_n \to 0$ such that $n\delta_n = o(1)$

C2: $\delta_n \to 0$ such that $(1/n)\log(1/\delta_n) = o(1)$

C3: The order of the initial maximal tree $\tau_{\max}$ is $r = O(\log(n))$

Condition C3 ensures that the size of the estimated initial maximal tree does not grow too rapidly with the sample size, preventing the number of test statistics that are under the null hypothesis to go to infinity faster than the significance level δ_n, which is required in condition C1. Condition C2 ensures that the test statistics under alternative hypotheses remain larger than the increasing boundary of the rejection region, determined by the decreasing significance level δ_n.

Besides conditions C1-C3, it is necessary a lower bound for the test statistic corresponding to the regression parameter under the alternative hypothesis, which will be given in Lemma 1.

Lemma 1. *Assume the following conditions:*

A1: The parameter vectors $\{\boldsymbol{\theta}^u, u \in \tau\}$ *are in the (open) admissible set* B*.*

A2: The link function $\mathbf{g}$ *is two times continuously differentiable.*

A3: The possible values $\mathbf{x}^0_{-h+1}$ *lie in a compact set* C *such that* $\mathbb{X}^\intercal \boldsymbol{\theta}^u$ *lies within the domain of* $\mathbf{g}$*, for all* $\mathbf{x}^0_{-h+1} \in C$*,* $\{\boldsymbol{\theta}^u, u \in \tau\} \in B$*.*

A4: For any $h \leqslant l$*, the smallest eigenvalue of* $\sum_{t=h}^{n} \mathbf{x}^t_{t-h+1}(\mathbf{x}^t_{t-h+1})^\intercal$ *diverges.*

Let λ^u_{-r+1} *be the test statistic defined as in 2.11 for the hypothesis* H^u_0 : $\boldsymbol{\beta}^u_{j,(-r+1)} = 0$*, for all* $j \in \{1, ..., p\}$ *vs* H^u_a : $\boldsymbol{\beta}^u_{j,(-r+1)} \neq 0$*, for at least one* $j \in \{1, ..., p\}$*. Then under the alternative* H^u_a

$$\lambda^u_{-r+1} \geqslant O_p(n).$$

Let $\hat{\tau}_n$ and $\hat{\tau}_{\theta_n}$ be the estimated beta-context tree and its associated parameter tree. Theorem 1 establishes that $\hat{\tau}_n$ converges almost surely to the the true data generating mechanism denoted by the tree τ and that $\hat{\tau}_{\theta_n}$ is strongly consistent for τ_θ.

Theorem 1. *Assume the beta-context tree* τ *has finite order. Then, under conditions C1-C3 and A1-A4, there exists an integer-valued variable* N *with* $P(N < \infty) = 1$ *such that*

a) $\hat{\tau}_n = \tau \ \forall n \geqslant N$ *with probability 1,*

b) $\mid \hat{\boldsymbol{\theta}}^u_n \mid = \mid \boldsymbol{\theta}^u \mid \ \ \forall u \in \tau, \ \forall n \geqslant N$ *with probability 1,*

c) $\hat{\boldsymbol{\theta}}^u_n \to \boldsymbol{\theta}^u \ \forall u \in \tau$*, as* $n \to \infty$ *with probability 1.*

The proof for Theorem 1 is available in Zambom, Kim e Garcia (2022). As for the proof of Lemma 1, it relies on arguments similar to those presented in Fahrmeir (1987) and Fahrmeir e Kaufmann (1987).

Through simulation results, Zambom, Kim e Garcia (2022) demonstrated the superior performance of the proposed beta-context algorithm compared to the VLMC in the presence of exogenous covariates. Even in the case that there are no exogenous covariates, it presents competitive results when compared to VLMC, with improving performance as the sample size grows. In contrast, the VLMC frequently underestimates the true tree by failing to recognize the impact of covariates on transition probabilities.

2.4 Variable Length Markov Chains with Time-Varying and Time-Invariant Exogenous Covariates

In this section, we introduce modifications to the beta-context algorithm to accommodate time-dependent and time-invariant exogenous covariates, considering data collected from multiple independent sources. Our approach assumes that all sources share identical parameter estimates, and we do not currently impose a spatial-dependent structure on the model. Additionally, we adapt the algorithm to handle scenarios with limited data that may not be suitable for asymptotic parametric methods.

Given that the concepts of contexts, order of the time-varying covariate parameter, order of the state transitions and time-invariant covariates, the beta-context tree and the terms siblings, children, and parents were already defined in Definition 1, Definition 2, Definition 3, Definition 4, and Definition 5, respectively, we will now proceed to define the algorithm itself, skipping the repetition of the model definitions.

2.4.1 The Modified Beta-Context Algorithm

The likelihood of the data Y_1^n, conditioning upon knowing y_1^η, $\boldsymbol{x}_1^\eta$ and $\boldsymbol{z}$, based on Definition 1 under a beta-context tree of order η is

$$\begin{aligned} &L(\tau_\theta; Y_1^n = y_1^n, \boldsymbol{x}_1^n, \boldsymbol{z}) = P(Y_1^n = y_1^n \mid Y_1^\eta = y_1^\eta, \boldsymbol{x}_1^n, \boldsymbol{z}, \tau_\theta) \\ &= P(Y_{\eta+1} = y_{\eta+1} \mid Y_1^\eta = y_1^\eta, \boldsymbol{x}_1^\eta, \boldsymbol{z}, \tau_\theta) \ \dots \ P(Y_n = y_n \mid Y_{n-\eta}^{n-1} = y_{n-\eta}^{n-1}, \boldsymbol{x}_{n-\eta}^{n-1}, \boldsymbol{z}, \tau_\theta), \end{aligned} \tag{2.14}$$

where each probability follows from 2.1. For each context $u = y_{-k+1}^0 = y_0, y_{-1}, ..., y_{-k+1}$ with k steps into the past there are $(1 + dk + m)$ parameters to estimate for each $j = 1, ..., p-1$. This includes one parameter corresponding to α_j^u, dk parameters associated with the d time-varying exogenous covariates at each of the k steps and m parameters associated with the m time-invariant exogenous covariates.

Given data $(Y_1^n(s), \boldsymbol{x}_1^n(s), \boldsymbol{z}(s))$ for each independent source $s = 1, ..., q$, the algorithm follows these steps:

1. **Step 1 - Finding Maximal Context Tree:**
   - Similar to Step 1 of beta-context algorithm proposed by Zambom, Kim e Garcia (2022), initiate the construction of the largest possible beta-context tree, denoted as $\tau_{\max}$. However, in this case, the tree includes all contexts $u \in \tau_{\max}$ that have been observed at least $f(p-1)$ times. Define the beta-context function $c(\cdot)$ to form the corresponding context tree $\tau_{\max}$:

$$\tau_{\max} = \{u = y_{-k+1}^0 : N(u) \geqslant f(p-1)\}.$$

Note that, given the presence of multiple sources, $N(u)$ now represents the sum of occurrences of the sequence u in each $Y_1^n(s)$, $s = 1, ..., q$:

$$N(u) = \sum_{s=1}^{q} \sum_{t=1}^{n-|u|+1} I(Y_t^{t+|u|-1}(s) = u), u \in \mathscr{Y}^{\infty}.$$

- Set the initial beta-context tree as $\tau^{(0)} = \tau_{\text{max}}$. Similar to Step 1 of beta-context algorithm, the variable r denotes the order of this context tree and the initial tree $\tau^{(0)}$ may not be complete.
- Compute $\hat{\tau}_\theta^{(0)}$, the associated estimated parameter tree. Unlike the beta-context algorithm, we allow for the possibility that some parameters may not be estimated. For each context u, proceed as follows:
  - If $N(uj) \geqslant f(1 + d|u| + m)$ $\forall j \in \mathscr{Y}$, calculate all the parameters (α_j^u, $d|u|$ parameters associated with the d time-varying exogenous covariates at each of the $|u|$ steps, and m parameters associated with the m time-invariant exogenous covariates). Estimate the parameters by maximizing the likelihood function $\prod_{s=1}^{q} L(\tau_\theta^{(0)} \mid y_1^n(s), \boldsymbol{x}_1^n(s), \boldsymbol{z}(s))$.
  - If $N(uj) \geqslant f(1 + m)$ $\forall j \in \mathscr{Y}$ and $N(uj) < f(1 + d|u| + m)$ for some $j \in \mathscr{Y}$, calculate only α_j^u and m parameters associated with the m time-invariant exogenous covariates. Estimate the parameters by maximizing the likelihood function $\prod_{s=1}^{q} L(\tau_\theta^{(0)} \mid y_1^n(s), \boldsymbol{z}(s))$.
  - If $0 < N(uj) < f(1 + m)$ for some $j \in \mathscr{Y}$, calculate only α_j^u. Estimate the parameters by maximizing the likelihood function $\prod_{s=1}^{q} L(\tau_\theta^{(0)} \mid y_1^n(s))$.
  - If $N(uj) = 0$ for some $j \in \mathscr{Y}$, calculate only α_j^u. Estimate the parameters by maximizing the likelihood function $\prod_{s=1}^{q} L(\tau_\theta^{(0)} \mid y_1^n(s))$ using a single-hidden-layer neural network with a softmax function. In a neural network, we initialize estimates with random values and iteratively adjust them to approximate the real ones. This allows us to estimate parameters even for events that never occur, resulting in estimates for very low probabilities. This differs from traditional Generalized Linear Models (GLMs), which cannot estimate parameters for events that never happen. For a more comprehensive understanding, please refer to Venables e Ripley (2002).
- It is worth noting that while the initial tree $\tau^{(0)}$ may not be complete, all terminal nodes must possess all possible children. This constraint is imposed because, in the presence of children, we refrain from estimating parameters for the parent node. Consequently, without this constraint, there might be

insufficient occurrences for a child node to be represented in the tree, leading to a lack of parameter estimates for that particular node.

2. **Step 2 - Pruning the Tree (Influence of time-dependent exogenous covariates):**

   - For each context u in the initial tree $\tau^{(0)}$ with length r, where parameters associated with the d time-varying exogenous covariates have been estimated, apply a likelihood ratio test to examine the significance of the parameter vector associated with past covariates from the node $(-r+1)$ to any of the outcomes (excluding the baseline). This test is performed similar to Step 2 of beta-context algorithm, with the deviance statistic λ^u_{-r+1} defined as:

   $$\begin{aligned}\lambda^u_{-r+1} = &-2\sum_{s=1}^{q}[\log L(\tilde{\tau}^u_\theta \mid y^n_1(s), \boldsymbol{x}^n_1(s), \boldsymbol{z}(s)) \\ &- \log L(\hat{\tau}^{(0)}_\theta \mid y^n_1(s), \boldsymbol{x}^n_1(s), \boldsymbol{z}(s))]\end{aligned} \tag{2.15}$$

     where $L(\cdot)$ represents the likelihood function as defined in 2.14.

3. **Step 3 - Pruning the Tree (Influence of the context and time-independent exogenous covariates):**

   Concerning the tests performed in **Step 2**:

   a) **Substep 3.1** - If at least two $H_0^{u_k}$ $(k = 1, ..., p)$ for $u_1, ..., u_p$ siblings in $\tau_{(0)}$ were not rejected in Step 2, or if they did not have parameters associated with the d time-varying exogenous covariates estimated in Step 1, and any children in this node were not lumped together yet:

      - The nodes corresponding to the rejected tests are retained in the model.
      - Calculate the test statistic for each pair of siblings (u_i, u_j) $(i, j \in \{1, ..., p\})$, where $H_0^{u_i}$ and $H_0^{u_j}$ were not rejected in Step 2, or u_i and u_j did not have parameters associated with the d time-varying exogenous covariates estimated in Step 1:
        - If $N(u_i a) \geqslant f$ and $N(u_j a) \geqslant f$, $\forall a \in \mathscr{Y}$ and f defined in Step 1,

        $$\begin{aligned}\lambda^{u_i,u_j} = -2\sum_{s=1}^{q}&[\log L(\tilde{\tau}^{\tilde{u}}_\theta \mid y^n_1(s), \boldsymbol{x}^n_1(s), \boldsymbol{z}(s)) - \\ &\log L(\hat{\tau}^{(0)}_\theta \mid y^n_1(s), \boldsymbol{x}^n_1(s), \boldsymbol{z}(s))],\end{aligned} \tag{2.16}$$

          where $\tilde{\tau}^{\tilde{u}}_\theta = \{(w, \tilde{\boldsymbol{\theta}}^w) : w \in \tau^{(0)}, w \neq u_i, u_j\} \cup \{(\tilde{u}, \tilde{\boldsymbol{\theta}}^w) : \tilde{u} = (u_i \cup u_j)u,\ u \text{ is parent of } u_1, ..., u_p\}$ is estimated under the null hypothesis that the sibling nodes u_i and u_j are lumped together, reducing the parameters. The parameters in vectors $\boldsymbol{\theta}^{u_i}$ and $\boldsymbol{\theta}^{u_j}$, which have been together reduced to size $\mathbb{R}^{2(1+d(r-1)+m)(p-1)}$ in Step 2, are merged into parameters $\boldsymbol{\theta}^{\tilde{u}}$ of size $\mathbb{R}^{(1+d(r-1))(p-1)}$.

  - If $(N(u_ia) < f$ and $N(u_ja) \geqslant f)$ or $(N(u_ja) < f$ and $N(u_ia) \geqslant f)$, for some $a \in \mathscr{Y}$, and one of the nodes u_i or u_j has estimates for the exogenous covariates parameters, then the corresponding exact test statistic for the likelihood ratio test is the Cochran-Mantel-Haenszel statistic (for additional details refer to Agresti (2012)). To compute this statistic, for the node that lacks covariate estimates, sum the event occurrences across all sources. For the other node, segregate event occurrences for each source. Subsequently, each contingency table will represent a source, and the row of the node without exogenous covariate estimates will be consistent across all contingency tables.
  - If $N(u_ia) < f$ and $N(u_ja) < f$, for some $a \in \mathscr{Y}$, the corresponding exact test statistic for the likelihood ratio test is the Fisher statistic (for additional details refer to Agresti (2012)). To compute the Fisher statistic, aggregate the event occurrences across all sources for both nodes, resulting in a single contingency table. Each row in this table corresponds to a node.
- For the smallest λ^{u_i,u_j}, compute the p-value $\pi^{(u_i,u_j)}$. If $N(u_ia) \geqslant f$ and $N(u_ja) \geqslant f$, $\forall a \in \mathscr{Y}$, $\pi^{(u_i,u_j)} = 1 - \Psi_{(1+d(r-1))(p-1)}(\lambda^{(u_i,u_j)})$. If $\pi^{(u_i,u_j)} \geqslant \delta_n$, lump siblings nodes (u_i, u_j) together in $\tilde{u}$. Update both $\tau^{(0)}$ and $\hat{\tau}_\theta^{(0)}$ with $\tilde{\tau}_\theta^{\tilde{u}}$.

b) **Substep 3.2** - If any two children were lumped together in Substep 3.1 and still have at least one of $H_0^{u_k}(k = 1, ..., p)$ not rejected in Step 2, where u_k were not lumped together in $\tilde{u}$ and $u_1, ..., u_p$ siblings:

- The nodes corresponding to the rejected tests are retained in the model.
- Calculate the test statistic $\lambda^{u_k,\tilde{u}}$ (following the guidelines provided in Substep 3.1) for each u_k $(k = 1, ..., p)$ which had not been lumped together in $\tilde{u}$ and had $H_0^{u_k}$ not rejected in Step 2.
- For the smallest $\lambda^{u_k,\tilde{u}}$ compute the p-value $\pi^{(u_k,\tilde{u})}$ (following the guidelines provided in Substep 3.1). If $\pi^{(u_k,\tilde{u})} \geqslant \delta_n$, lump node u_k together in $\tilde{u}$. Update both $\tau^{(0)}$ and $\hat{\tau}_\theta^{(0)}$.
- Repeat Substep 3.2 until no more u_k $(k = 1, ..., p)$ can be lumped together in $\tilde{u}$.

c) **Substep 3.3** Lastly, sequentially test to prune the past most parameters in $\boldsymbol{\beta}^{u_k}$ $(k = 1, ..., p)$ which had its hypothesis not rejected in Step 2 up to the root, similar to Substep 3.2 of beta-context algorithm. Calculate the corresponding test statistic $\lambda_{-r+2}^{u_1}$ using 2.15.

At the end of Step 3, the tree structure and parameters are possibly pruned based on the significance of context and covariate influence, leading to a potentially smaller context tree $\tau^{(1)} \subseteq \tau^{(0)}$ and its updated parameter tree $\hat{\tau}_{\theta}^{(1)}$ at level r.

4. **Step 4 - Iterative Pruning:**

   - Repeat Steps 2 and 3 for contexts of lengths $r-1, r-2, ..., 1$, using the updated trees $\tau^{(1)}$ and $\hat{\tau}_{\theta}^{(1)}$.
   - However, if $u_1, ..., u_p$ are siblings and at least one of them has children, no context pruning is performed (both nodes are retained). But, as in Step 3, the pruning of covariate parameters is executed sequentially, proceeding from the most distant past to the root.

Denote this pruned beta-context tree by $\hat{\tau}_n$ with associated parameter tree $\hat{\tau}_{\theta}$ and corresponding beta-context function $\hat{c}(\cdot)$.

2.4.2 Highlighting modifications in Beta-Context Algorithm

1. **Time-invariant exogenous covariates:** Time-fixed parameters (γ_j^u) for time-invariant exogenous covariates were introduced for each terminal node. Here, u denotes a terminal node, and $j = 1, ..., p$ represents the transition to state j. Note that during the tree-pruning process, the time-fixed parameter (γ_j^u) is pruned simultaneously with the intercept (α_j^u) when nodes are lumped together.

2. **Observations from various independent sources:** This aspect involves integrating data from diverse sources, allowing for a comprehensive understanding of the process behavior across multiple situations, such as different geographic locations. In this context, time-invariant exogenous covariates may vary for each source, providing a more nuanced representation of fixed conditions. Despite these variations, model parameters are assumed to be the same across sources, enhancing the predictive capabilities of the model. The current model does not incorporate a spatial-dependent structure, assuming sources are independent.

3. **Rare events:** To ensure the model's robustness in handling rare events, particularly in scenarios with finite observations, we have implemented several strategies:

   - **Minimum observation requirement:** Each category within every terminal node is subjected to a minimum observation threshold, ensuring a sufficient number of occurrences for reliable parameter estimation.
   - **Partial parameter estimation:** Terminal nodes have the flexibility to possess incomplete sets of parameter estimates. If the available observations are not enough for estimating all parameters, only a subset is computed. The intercept remains a consistently estimated parameter.

- **Intercept Estimation with Neural Network:** In cases where a specific event never occurs for a terminal node, the intercept can be estimated using a single-hidden-layer neural network (Venables e Ripley (2002)).
- **Exact Tests for Rare Events:** The model incorporates exact tests, such as the Fisher and Cochran-Mantel-Haenszel tests, which are particularly useful in situations involving rare events.

These measures collectively increase the model's adaptability to scenarios characterized by infrequent events and a limited number of observations.

An alternative approach considered was to set a fixed probability for the rare event while estimating all parameters for the other possible categories. However, challenges arose in determining how to calculate estimates with the bound on the fixed probability. Additionally, complications emerged when the rare event served as the baseline for other branches.

4. **Terminal node completeness requirement for initial tree:** Although the initial tree $\tau^{(0)}$ may lack completeness, a crucial constraint is imposed on all terminal nodes — they must have all possible children. This constraint is imposed because, in the presence of children, we refrain from estimating parameters for the parent node. Consequently, without this constraint, there might be insufficient occurrences for a child node to be represented in the tree, leading to a lack of parameter estimates for that particular node.

2.4.3 Consistency of the Modified Beta-Context Algorithm

It can be shown, following a methodology similar to Zambom, Kim e Garcia (2022), that the modified beta-context algorithm ensures strong consistency in estimating the beta-context tree, regression parameters, and transition probabilities (Theorem 1). The conditions C1-C3, as previously described, remain unchanged but with $n = \sum_{s=1}^{q} n_s$. Lemma 1 requires some adjustments, outlined in Lemma 2:

Lemma 2. *Assume the following conditions:*

A1: The parameter vectors $\{\boldsymbol{\theta}^u, u \in \tau\}$ are in the (open) admissible set B.

A2: The link function $\mathbf{g}$ is two times continuously differentiable.

A3: The possible values $(\mathbf{x}^0_{-h+1}, \mathbf{z})$ lie in a compact set C such that $\mathbb{X}^\intercal \boldsymbol{\theta}^u$ lies within the domain of $\mathbf{g}$, for all $(\mathbf{x}^0_{-h+1}, \mathbf{z}) \in C$, $\{\boldsymbol{\theta}^u, u \in \tau\} \in B$.

A4: For any $h \leqslant l$ and $s = 1, ..., q$, the smallest eigenvalue of

$$\sum_{t=h}^{n_s} \mathbf{x}^t_{t-h+1}(s)(\mathbf{x}^t_{t-h+1}(s))^\intercal$$

diverges.

Let λ^{u}_{-r+1} *be the test statistic defined as in Equation 2.15 for the hypothesis* $H^{u}_{0} : \boldsymbol{\beta}^{u}_{j,(-r+1)} = 0$, *for all* $j \in \{1, ..., p\}$ *vs* $H^{u}_{a} : \boldsymbol{\beta}^{u}_{j,(-r+1)} \neq 0$, *for at least one* $j \in \{1, ..., p\}$. *Then under the alternative* H^{u}_{a}

$$\lambda^{u}_{-r+1} \geqslant O_p(n).$$

The proofs for Theorem 1 in this case follow a similar demonstration as provided in Zambom, Kim e Garcia (2022), with $n = \sum_{s=1}^{q} n_s$.

3 Simulations

In this chapter, we evaluate the performance of the proposed modified beta-context algorithm (referred to as *modified-beta-VLMC* in the tables) across various scenarios, including some adapted from Zambom, Kim e Garcia (2022). The generated data follows models derived from context trees of orders 2, 3, and 4, with varying lengths of the univariate time-varying covariate parameter vector $\boldsymbol{\beta}^u$. Additionally, we explore models with and without a univariate time-invariant covariate, considering scenarios with one or more sources. For comparative analysis, in cases where the simulation involves a single source and lacks time-invariant exogenous covariates, we present results obtained using the beta-context algorithm introduced by Zambom, Kim e Garcia (2022) (referred to as *beta-VLMC* in the tables).

In all scenarios, the values of the tuning parameters δ_n and f were selected to minimize the BIC criterion. Consequently, the same model may have different tuning parameters for the original and the modified version of the beta-context algorithm. Since tuning parameters directly affect tree length, discrepancies in tree lengths between algorithms can arise from either the algorithm itself or the tuning parameter values. However, we considered this approach better than fixing the same tuning parameters for both algorithms, as they have different constructions.

Various metrics were employed to evaluate the performance of the methods in estimating the context function. These metrics include the average values of several measures over 100 simulations: BIC, AIC, log-likelihood, the number of parameters $\hat{\alpha}^u$ and $\hat{\boldsymbol{\gamma}}^u$ (number of final nodes u in $\hat{\tau}$), the number of parameters $\hat{\boldsymbol{\beta}}^u$ (total number of coefficients estimated different from 0 in all vectors $\hat{\boldsymbol{\beta}}^u$, $\forall u$), the order of the $\hat{\tau}$ tree, the order of the time-varying exogenous covariate (maximum length of $\hat{\boldsymbol{\beta}}^u$, $\forall u$), the number of missing nodes in $\hat{\tau}$, the number of extra nodes in $\hat{\tau}$, whether the τ tree is identified exactly (no missing and no extra nodes in $\hat{\tau}$), and whether the $\tau_{\boldsymbol{\theta}}$ tree is identified exactly (no longer nor shorter estimated parameter vector $\boldsymbol{\beta}^u$, $\forall u$).

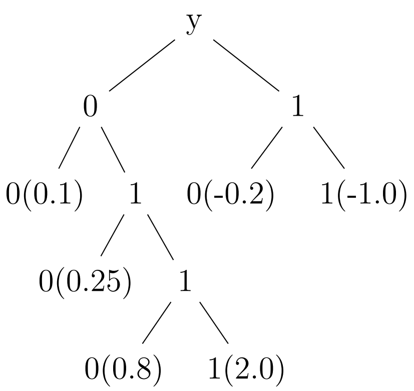
Furthermore, to assess the methods' performance in estimating the coefficient vector, the mean difference between real and estimated values was calculated over 100 simulations for all parameters when the context was correctly identified.

Additional simulations can be found in Appendix A, where all exogenous covariates parameters were set to zero.

3.1 Simulations with a single source and without time-invariant exogenous covariates

The simulations presented in this section were generated for a single source, and no time-invariant covariates were taken into account.

Figure 3 provides details on Model 1, where time-varying exogenous variables were generated from both a standard Normal distribution (N(0, 1)) and, to evaluate performance under heavy-tailed distributions, a t-Student distribution with 2 degrees of freedom (t(2)). Samples of $n = 1000$ and $n = 2000$ state transitions were considered.



$$\boldsymbol{\beta}^{00} = (2, 0)'$$
$$\boldsymbol{\beta}^{010} = (-1, 1, 0)'$$
$$\boldsymbol{\beta}^{0111} = (1.5, 2, 0, 0)'$$
$$\boldsymbol{\beta}^{0110} = (4, 3, 2, 1)'$$
$$\boldsymbol{\beta}^{10} = (0, 0)'$$
$$\boldsymbol{\beta}^{11} = (0, 0)'$$

Figure 3 – Context tree τ and associated parameters $\boldsymbol{\theta}$ for Model 1 (numbers in parenthesis represent the values of α^u for each context u).

Performance metrics for both the proposed modified beta-context algorithm and the beta-context algorithm are presented in detail in Table 1. To enhance the clarity of the estimated trees, the frequency of occurrences for each context tree across 100 simulations are presented in Table 2, Table 3, Table 4, and Table 5. Each table delineates potential context trees in the first column, followed by the respective number of contexts within each tree. The last two columns indicate the frequency of occurrence for each context tree, distinguished between modified and original beta-VLMC models. Bolded lines indicate trees that were precisely estimated.

In cases where $n = 1000$, the modifications, which require a larger number of observations to estimate exogenous covariate parameters, lead to smaller estimated trees in the proposed method compared to the real ones. Furthermore, for trees generated with exogenous time-varying covariates following a normal distribution with mean 0 and standard deviation 1, the modified version achieves a higher percentage of identical τ trees compared to the original version. However, it is important to note that both methods exhibit low accuracy.

For $n = 2000$, the modified version achieved 72% and 83% of identical τ trees for covariates following N(0,1) and t(2) distributions, respectively, while the original version attained 57% and 79%. However, the original beta-context algorithm achieved a higher percentage of identical $\tau_{\boldsymbol{\theta}}$, potentially because the modified version requires a larger number of observations to estimate covariate parameters. It is noteworthy that, in general, the original version of the beta-context algorithm generates larger trees. Both methods exhibit improved accuracy in identifying the exact same nodes as τ for $n = 2000$, supporting the consistency theory of the estimators.

Note that even when the correct identification of $\tau_{\boldsymbol{\theta}}$ is relatively low, $\hat{\tau}$ can be correctly estimated for both methods. Yet, when the covariates are generated with a heavy-tailed t(2) distribution, the performance of both methods is not significantly affected - in fact, the accuracy of recovering $\tau_{\boldsymbol{\theta}}$ is improved.

Table 1 – Simulation results for Model 1 with time-varying exogenous covariates generated from N(0,1) and t(2) distributions (average over 100 simulations).

	Model 1 (n = 1000, N(0, 1) distr.)		Model 1 (n = 1000, t(2) distr.)	
	modified-beta-VLMC	beta-VLMC	modified-beta-VLMC	beta-VLMC
BIC	1154.46	1153.34	1083.69	1079.17
AIC	1099.74	1090.62	1031.69	1016.86
LogLik	-538.72	-532.53	-505.25	-495.73
# par. $\hat{\alpha}^u$	5.05	5.54	4.64	5.19
# par. $\hat{\boldsymbol{\beta}}^u$	6.10	7.24	5.96	7.50
order $\hat{\tau}$	3.35	3.74	3.25	3.74
order-Cov	2.94	3.10	2.94	3.37
# Missing $\hat{\tau}$	2.20	1.95	2.77	2.02
# Extra $\hat{\tau}$	0.30	0.43	0.04	0.30
Identical τ	0.16	0.06	0.15	0.18
Identical $\tau_{\boldsymbol{\theta}}$	0.00	0.01	0.00	0.07
	Model 1 (n = 2000, N(0, 1) distr.)		Model 1 (n = 2000, t(2) distr.)	
	modified-beta-VLMC	beta-VLMC	modified-beta-VLMC	beta-VLMC
BIC	2252.61	2260.75	2079.33	2073.82
AIC	2176.88	2159.99	1998.40	1985.44
LogLik	-1074.92	-1062.0	-984.75	-976.94
# par. $\hat{\alpha}^u$	5.95	7.15	6.04	6.39
# par. $\hat{\boldsymbol{\beta}}^u$	7.57	10.84	8.41	9.39
order $\hat{\tau}$	3.85	4.63	3.97	4.17
order-Cov	3.57	4.07	3.85	3.89
# Missing $\hat{\tau}$	0.46	0.10	0.20	0.10
# Extra $\hat{\tau}$	0.36	2.17	0.28	0.78
Identical τ	0.72	0.57	0.83	0.79
Identical $\tau_{\boldsymbol{\theta}}$	0.17	0.30	0.49	0.63

To evaluate the performance of the proposed method in estimating the vector of coefficients, Table 6 presents the mean differences between real and estimated parameters for both the modified and original beta-context algorithms when the context is correctly

Table 2 – Estimated τ trees for Model 1, with n = 1000 and N(0, 1) distribution (frequency of occurrences over 100 simulations).

Contexts	# Contexts	modified-beta-VLMC	beta-VLMC
0 0 - 0 1 0 - 0 1 1 - 1	4	28	13
0 0 - 0 1 0 - 0 1 1 - 0 1 1 0 - 1	5	0	15
0 0 - 0 1 0 - 0 1 1 - 1 0 - 1 1	5	40	20
0 0 - 0 1 0 - 0 1 1 0 - 0 1 1 1 - 1	5	7	6
0 0 - 0 1 0 - 0 1 1 - 0 1 1 0 - 0 1 1 0 1 - 1	6	0	1
0 0 - 0 1 0 - 0 1 1 - 0 1 1 0 - 1 0 - 1 1	6	0	27
0 0 - 0 1 0 - 0 1 1 - 1 0 0 - 1 0 1 - 1 1	6	1	1
0 0 - 0 1 0 - 0 1 1 0 - 0 1 1 1 - 1 0 - 1 1	**6**	**16**	**6**
0 0 - 0 1 0 0 - 0 1 0 1 - 0 1 1 - 1 0 - 1 1	6	1	0
⩾ 7 contexts	⩾ 7	7	11

Table 3 – Estimated τ trees for Model 1, with n = 1000 and t(2) distribution (frequency of occurrences over 100 simulations).

Contexts	# Contexts	modified-beta-VLMC	beta-VLMC
0 0 - 0 1 0 - 0 1 1 - 1	4	53	19
0 0 - 0 1 0 - 0 1 1 - 0 1 1 0 - 1	5	0	3
0 0 - 0 1 0 - 0 1 1 - 1 0 - 1 1	5	20	13
0 0 - 0 1 0 - 0 1 1 0 - 0 1 1 1 - 1	5	9	34
0 0 - 0 1 0 - 0 1 1 - 0 1 1 0 - 1 0 - 1 1	6	0	2
0 0 - 0 1 0 - 0 1 1 0 - 0 1 1 0 1 - 0 1 1 1 - 1	6	0	1
0 0 - 0 1 0 - 0 1 1 0 - 0 1 1 1 - 1 0 - 1 1	**6**	**15**	**18**
0 0 - 0 1 0 - 0 1 1 0 0 - 0 1 1 0 1 - 0 1 1 1 - 1	6	0	1
0 0 - 0 1 0 0 - 0 1 0 1 - 0 1 1 0 - 0 1 1 1 - 1	6	0	1
0 0 0 - 0 0 1 - 0 1 0 - 0 1 1 - 1 0 - 1 1	6	1	1
0 0 0 - 0 0 1 - 0 1 0 - 0 1 1 0 - 0 1 1 1 - 1	6	1	1
⩾ 7 contexts	7	0	5

Table 4 – Estimated τ trees for Model 1, with n = 2000 and N(0, 1) distribution (frequency of occurrences over 100 simulations).

Contexts	# Contexts	modified-beta-VLMC	beta-VLMC
0 0 - 0 1 0 - 0 1 1 - 1 0 - 1 1	5	22	3
0 0 - 0 1 0 - 0 1 1 0 - 0 1 1 1 - 1 0 - 1 1	**6**	**72**	**57**
⩾ 7 contexts	⩾ 7	6	40

Table 5 – Estimated τ trees for Model 1, with n = 2000 and t(2) distribution (frequency of occurrences over 100 simulations).

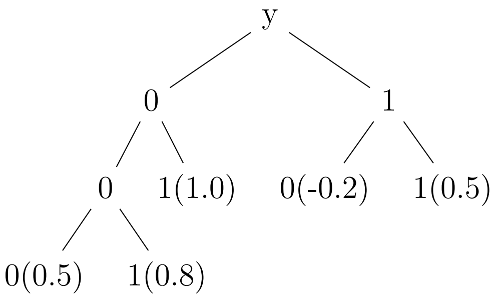
Contexts	# Contexts	modified-beta-VLMC	beta-VLMC
0 0 - 0 1 0 - 0 1 1 - 1 0 - 1 1	5	6	1
0 0 - 0 1 0 - 0 1 1 0 - 0 1 1 1 - 1	5	4	3
0 0 - 0 1 0 - 0 1 1 0 - 0 1 1 1 - 1 0 - 1 1	**6**	**83**	**79**
⩾ 7 contexts	⩾ 7	7	17

identified. For parameters corresponding to short contexts, such as $u = (0, 0)$, estimation accuracy is high for both methods. However, larger contexts like (0,1,1,0) and (0,1,1,1) were infrequent in the data generated by Model 1, especially for $n = 1000$, leading to insufficient observations for accurate estimation of these six parameters. Generally, when the context is present in the final tree, the modified version provides better estimates.

Table 6 – Differences between real and estimated values for Model 1 (average over 100 simulations).

	Model 1 (n = 1000, N(0, 1) distr.)		Model 1 (n = 1000, t(2) distr.)	
	modified-beta-VLMC	beta-VLMC	modified-beta-VLMC	beta-VLMC
α^{00}	0.14	0.14	0.14	0.14
α^{010}	0.14	0.14	0.18	1.18
α^{0110}	0.68	1.34	2.11	1.83
α^{0111}	0.81	0.81	1.07	4.16
α^{10}	0.08	0.08	0.09	0.09
α^{11}	0.13	0.13	0.14	0.15
$\boldsymbol{\beta}^{00}$	(0.26, 0)	(0.26, 0)	(0.26, 0)	(0.25, 0)
$\boldsymbol{\beta}^{010}$	(0.16, 0.22, 0)	(0.16, 0.22, 0)	(0.18, 0.18, 0)	(0.18, 0.18, 0)
$\boldsymbol{\beta}^{0110}$	(1.50, 2.00, 0, 0)	(1.10, 1.53, 0, 0)	(1.50, 2.00, 0, 0)	(4.65, 4.88, 1.17, 0)
$\boldsymbol{\beta}^{0111}$	(2.10, 1.88, 1.38, 1.42)	(2.89, 2.40, 1.51, 1.54)	(9.14, 7.82, 5.11, 2.49)	(7.39, 5.82, 3.87, 2.27)
$\boldsymbol{\beta}^{10}$	(0, 0)	(0, 0)	(0, 0)	(0, 0)
$\boldsymbol{\beta}^{11}$	(0, 0)	(0, 0)	(0,0)	(0, 0)
	Model 1 (n = 2000, N(0, 1) distr.)		Model 1 (n = 2000, t(2) distr.)	
	modified-beta-VLMC	beta-VLMC	modified-beta-VLMC	beta-VLMC
α^{00}	0.10	0.10	0.09	0.09
α^{010}	0.10	0.10	0.12	0.12
α^{0110}	0.28	0.27	0.38	0.37
α^{0111}	0.66	1.14	0.69	1.03
α^{10}	0.06	0.06	0.06	0.06
α^{11}	0.08	0.08	0.09	0.09
$\boldsymbol{\beta}^{00}$	(0.16, 0)	(0.16, 0)	(0.16, 0)	(0.15, 0)
$\boldsymbol{\beta}^{010}$	(0.13, 0.12, 0)	(0.13, 0.12, 0)	(0.13, 0.15, 0)	(0.13, 0.15, 0)
$\boldsymbol{\beta}^{0110}$	(0.10, 1.28, 0, 0)	(0.63, 1.36, 0.32, 0.31)	(0.82, 1.09, 0, 0)	(0.82, 1.32, 0, 0)
$\boldsymbol{\beta}^{0111}$	(0.75, 0.58, 0.47, 0.63)	(0.76, 0.56, 0.45, 0.71)	(0.98, 0.72, 0.49, 0.41)	(0.96, 0.70, 0.48, 0.44)
$\boldsymbol{\beta}^{10}$	(0, 0)	(0, 0)	(0, 0)	(0, 0)
$\boldsymbol{\beta}^{11}$	(0, 0)	(0, 0)	(0, 0)	(0, 0)

Model 2 is described in Figure 4 and involves generating time-varying exogenous covariates from a N(0, 1) distribution. We create samples with $n = 1000$ and $n = 2000$ state transitions.



$$\boldsymbol{\beta}^{000} = (3, 1, 2)'$$
$$\boldsymbol{\beta}^{001} = (1, 0, 0)'$$
$$\boldsymbol{\beta}^{01} = (-1, -2)'$$
$$\boldsymbol{\beta}^{10} = (-1.2, 0)'$$
$$\boldsymbol{\beta}^{11} = (0, 0)'$$

Figure 4 – Context tree τ and associated parameters $\boldsymbol{\theta}$ for Model 2 (numbers in parenthesis represent the values of α^u for each context u).

The performance metrics for the estimated context function in Model 2 are detailed in Table 7. To aid in the visualization of the estimated trees, Table 8 and Table 9 provide the count of estimated occurrences for each context. In both instances with $n = 1000$ and $n = 2000$, the modified beta-VLMC algorithm demonstrated a superior percentage of identical τ trees and exact covariate vectors $\tau_{\boldsymbol{\theta}}$ when compared to the original algorithm introduced by Zambom, Kim e Garcia (2022). It is noteworthy that, overall, the original algorithm tends to produce larger trees. Both methods fits better Model 2 than Model 1, either by under or overfitting, probably because it has a larger sample size when compared to the number of parameters to be estimated.

Table 7 – Simulation results for Model 2 with time-varying exogenous covariates generated from a N(0, 1) distribution (average over 100 simulations).

	Model 2 (n = 1000, N(0, 1) distr.)		Model 2 (n = 2000, N(0, 1) distr.)	
	modified-beta-VLMC	beta-VLMC	modified-beta-VLMC	beta-VLMC
BIC	1160.27	1183.37	2261.52	2264.89
AIC	1108.0	1118.47	2185.40	2183.90
logLik	-543.35	-545.31	-1079.03	-1077.49
# par. $\hat{\alpha}^u$	4.77	5.89	5.50	5.76
# par. $\hat{\boldsymbol{\beta}}^u$	5.88	7.37	8.09	8.70
order $\hat{\tau}$	2.79	3.64	3.37	3.49
order-Cov	2.76	3.53	3.32	3.23
# Missing $\hat{\tau}$	0.50	0.44	0	0
# Extra $\hat{\tau}$	0.04	2.09	1.00	1.40
Identical τ	0.76	0.58	0.87	0.76
Identical $\tau_{\boldsymbol{\theta}}$	0.35	0.29	0.86	0.76

Table 8 – Estimated τ trees for Model 2, with n = 1000 (frequency of occurrences over 100 simulations).

Contexts	# Contexts	modified-beta-VLMC	beta-VLMC
0 0 - 0 1 - 1	3	1	1
0 0 - 0 1 - 1 0 - 1 1	4	21	16
0 0 0 - 0 0 1 - 0 1 - 1	4	1	1
0 0 0 - 0 0 1 - 0 1 - 1 0 - 1 1	**5**	**76**	**58**
0 0 - 0 1 0 - 0 1 1 0 - 0 1 1 1 - 1 0 - 1 1	6	1	0
>= 7 contexts	7	0	24

Table 9 – Estimated τ trees for Model 2, with n = 2000 (frequency of occurrences over 100 simulations).

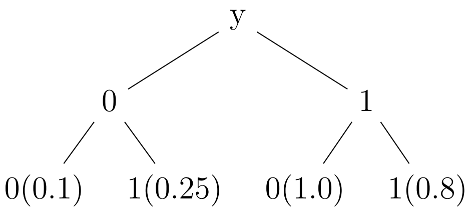
Contexts	# Contexts	modified-beta-VLMC	beta-VLMC
0 0 0 - 0 0 1 - 0 1 - 1 0 - 1 1	**5**	**87**	**76**
0 0 0 0 - 0 0 0 1 - 0 0 1 - 0 1 - 1 0 - 1 1	6	0	2
>= 7 contexts	7	13	22

Table 10 shows the mean differences between real and estimated parameters for both the modified and original beta-context algorithms when the context is correctly identified. Overall, both methods yield comparable results.

Table 10 – Differences between real and estimated values for Model 2 (average over 100 simulations).

	Model 2 (n = 1000, N(0, 1) distr.)		Model 2 (n = 2000, N(0,1) distr.)	
	modified-beta-VLMC	beta-VLMC	modified-beta-VLMC	beta-VLMC
α^{000}	0.57	0.56	0.28	0.28
α^{001}	0.19	0.19	0.15	0.15
α^{01}	0.17	0.17	0.11	0.10
α^{10}	0.12	0.12	0.08	0.08
α^{11}	0.09	0.09	0.06	0.06
$\boldsymbol{\beta}^{000}$	(1.02, 0.66, 0.84)	(1.04, 0.67, 0.87)	(0.54, 0.27, 0.37)	(0.53, 0.26, 0.36)
$\boldsymbol{\beta}^{001}$	(0.66, 0, 0)	(0.65, 0, 0)	(0.17, 0, 0)	(0.16, 0, 0)
$\boldsymbol{\beta}^{01}$	(0.17, 0.21)	(0.18, 0.22)	(0.10, 0.15)	(0.10, 0.15)
$\boldsymbol{\beta}^{10}$	(0.15, 0)	(0.16, 0)	(0.10, 0)	(0.10, 0)
$\boldsymbol{\beta}^{11}$	(0,0)	(0,0)	(0,0)	(0,0)

In Model 3, as outlined in Figure 5, we assess the effectiveness of the proposed algorithm when data is generated from a full tree of fixed order. Time-varying exogenous variables were generated from a N(0,1) distribution. We generate samples with $n = 1000$ and $n = 2000$ state transitions.



$$\boldsymbol{\beta}^{00} = (1.2, 0.3)'$$
$$\boldsymbol{\beta}^{01} = (-1, 1)'$$
$$\boldsymbol{\beta}^{10} = (1.5, -2)'$$
$$\boldsymbol{\beta}^{11} = (-0.2, -0.9)'$$

Figure 5 – Context tree τ and associated parameters $\boldsymbol{\theta}$ for Model 3 (numbers in parenthesis represent the values of α^u for each context u).

The performance metrics for the estimated context function in Model 3 are outlined in Table 11. To aid in the visualization of the estimated trees, Table 12 and Table 13 provide the count of estimated occurrences for each context. For $n = 1000$, the original beta-context algorithm demonstrated superior results, achieving 100% identical τ trees, while the modified beta-context algorithm produced larger trees than the real ones. For $n = 2000$, improved outcomes were observed with the modified beta-context algorithm. Interestingly, in this case, the original beta-context algorithm generated larger trees than the real ones - 23 trees with more than seven contexts when the real tree has only 4 contexts.

The mean discrepancies between real and estimated parameters, when the context were correctly identified, are showed in Table 14 for both the proposed modified beta-context algorithm and the beta-context algorithm. Both methods exhibit comparable results in this scenario.

Table 11 – Simulation results for Model 3 with time-varying exogenous covariates generated from a N(0, 1) distribution (average over 100 simulations).

	Model 3 (n = 1000, N(0, 1) distr.)		Model 3 (n = 2000, N(0, 1) distr.)	
	modified-beta-VLMC	beta-VLMC	modified-beta-VLMC	beta-VLMC
BIC	1123.18	1120.03	2184.90	2195.48
AIC	1065.71	1066.29	2116.06	2115.72
logLik	-521.15	-522.19	-1045.74	-1043.62
# par. $\hat{\alpha}^u$	4.25	4.00	4.36	4.99
# par. $\hat{\boldsymbol{\beta}}^u$	7.46	6.95	7.93	9.25
order $\hat{\tau}$	2.25	2.00	2.36	2.93
order-Cov	2.23	2.00	2.25	2.56
# Missing $\hat{\tau}$	0	0	0	0
# Extra $\hat{\tau}$	0.50	0	0.72	1.78
Identical τ	0.90	1.00	0.86	0.71
Identical $\tau_{\boldsymbol{\theta}}$	0	0	0.13	0.11

Table 12 – Estimated τ trees for Model 3, with n = 1000 (frequency of occurrences over 100 simulations).

Contexts	# Contexts	modified-beta-VLMC	beta-VLMC
00 - 01 - 10 - 11	**4**	**90**	**100**
00 - 01 - 10 - 110 - 111	5	1	0
00 - 01 - 100 - 1010 - 1011 - 11	6	1	0
00 - 0100 - 0101 - 011 - 10 - 11	6	1	0
0000 - 0001 - 001 - 01 - 10 - 11	6	1	0
>= 7 contexts	7	6	0

Table 13 – Estimated τ trees for Model 3, with n = 2000 (frequency of occurrences over 100 simulations).

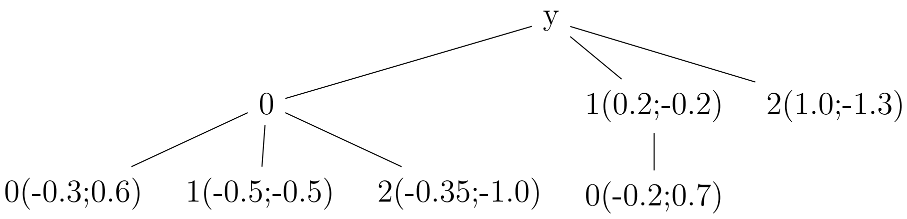
Contexts	# Contexts	modified-beta-VLMC	beta-VLMC
00 - 01 - 10 - 11	**4**	**86**	**71**
00 - 01 - 100 - 101 - 11	5	2	1
00 - 010 - 011 - 10 - 11	5	1	1
00 - 01 - 10 - 1100 - 1101 - 111	6	1	1
00 - 01 - 100 - 1010 - 1011 - 11	6	1	3
00 - 01 - 1000 - 1001 - 101 - 11	6	1	0
>= 7 contexts	7	8	23

In Model 4, as outlined in Figure 6, we evaluate the effectiveness of the proposed algorithm in situations where the state space is not binary. Time-varying exogenous variables were generated from a N(0,1) distribution. Given its larger number of parameters to be estimated, we generate samples with $n = 4000$ and $n = 8000$ state transitions.

The performance metrics for the estimated context function in Model 4 are detailed in Table 15. To aid in the visualization of the estimated trees, Table 16 and Table 17 provide the count of estimated occurrences for each context. For $n = 4000$, both algorithms yielded similar results, with the original beta-context algorithm having

Table 14 – Differences between real and estimated values for Model 3 (average over 100 simulations).

	Model 3 (n = 1000, N(0, 1) distr.)		Model 3 (n = 2000, N(0,1) distr.)	
	modified-beta-VLMC	beta-VLMC	modified-beta-VLMC	beta-VLMC
α^{00}	0.13	0.13	0.10	0.10
α^{01}	0.13	0.14	0.09	0.09
α^{10}	0.18	0.18	0.13	0.13
α^{11}	0.11	0.11	0.07	0.07
$\boldsymbol{\beta}^{00}$	(0.15, 0.3)	(0.16, 0.3)	(0.13, 0.27)	(0.14, 0.27)
$\boldsymbol{\beta}^{01}$	(0.15, 0.15)	(0.15, 0.17)	(0.10, 0.11)	(0.10, 0.11)
$\boldsymbol{\beta}^{10}$	(0.23, 0.26)	(0.23, 0.25)	(0.15, 0.15)	(0.15, 0.15)
$\boldsymbol{\beta}^{11}$	(0.10, 0.11)	(0.10, 0.12)	(0.07, 0.08)	(0.06, 0.08)



$$
\begin{array}{ll}
\boldsymbol{\beta}_1^{00} = (0.4, 0)' & \boldsymbol{\beta}_2^{00} = (0.4, 0)' \\
\boldsymbol{\beta}_1^{01} = (0.9, 0)' & \boldsymbol{\beta}_2^{01} = (0.3, 0)' \\
\boldsymbol{\beta}_1^{02} = (1.5, -2)' & \boldsymbol{\beta}_2^{02} = (0.5, -0.85)' \\
\boldsymbol{\beta}_1^{10} = (0, 0)' & \boldsymbol{\beta}_2^{10} = (0, 0)' \\
\boldsymbol{\beta}_1^{1*} = (-0.4)' & \boldsymbol{\beta}_2^{1*} = (0.6)' \\
\boldsymbol{\beta}_1^{2} = (0)' & \boldsymbol{\beta}_2^{2} = (0)'
\end{array}
$$

Figure 6 – Context tree τ and associated parameters $\boldsymbol{\theta}$ for Model 4. Numbers in parenthesis represent the values of (α_1^u, α_2^u) from each context u and $\boldsymbol{\beta}_1^u$ and $\boldsymbol{\beta}_2^u$ are the coefficient vectors of the covariates. $\boldsymbol{\beta}^{1*}$ means the context is 1 preceded by other state that is not 1.

a higher percentage of identical τ trees (98% compared to 95% for the modified beta-context algorithm) but a lower percentage of exact covariate vectors (24% compared to 38% for the modified beta-context algorithm). For $n = 8000$, the modified beta-context algorithm demonstrated superior results, achieving 100% and 76% identical τ trees and exact covariate vectors, respectively. Both methods showed better results in achieving exact covariate vectors with $n = 8000$. In general, the original beta-context algorithm resulted in larger trees.

The mean discrepancies between real and estimated parameters, when the context were correctly identified, are showed in Table 18. Both methods exhibit comparable results in this scenario.

Model 5, described in Figure 7, shares similarities with Model 3 but includes

Table 15 – Simulation results for Model 4 with time-varying exogenous covariates generated from a N(0, 1) distribution (average over 100 simulations).

	Model 4 (n = 4000, N(0, 1) distr.)		Model 4 (n = 8000, N(0, 1) distr.)	
	modified-beta-VLMC	beta-VLMC	modified-beta-VLMC	beta-VLMC
BIC	7661.21	7658.67	15160.71	15189.83
AIC	7526.39	7527.75	15010.34	15008.16
logLik	-3741.78	-3743.08	-7483.65	-7478.08
# par. $\hat{\alpha}^u$	12.32	12.14	12	14.04
# par. $\hat{\boldsymbol{\beta}}^u$	9.10	8.66	9.52	11.96
order $\hat{\tau}$	2.14	2.06	2	2.67
order-Cov	2.05	2.05	2	2.4
# Missing $\hat{\tau}$	0	0	0	0
# Extra $\hat{\tau}$	0.18	0.07	0	1.08
Identical τ	0.95	0.98	1	0.79
Identical $\tau_{\boldsymbol{\theta}}$	0.38	0.24	0.76	0.70

Table 16 – Estimated τ trees for Model 4, with n = 4000 (frequency of occurrences over 100 simulations).

Contexts	# Contexts	modified-beta-VLMC	beta-VLMC
0 0 - 0 1 - 0 2 - 1 - 1 0 - 2	**6**	**95**	**98**
0 0 - 0 0 2 - 0 0 2 0 - 0 1 - 0 2 - 1 - 1 0 - 2	8	1	0
0 0 - 0 1 - 0 1 1 - 0 1 1 1 - 0 2 - 1 - 1 0 - 2	8	1	0
0 0 - 0 0 1 - 0 0 1 1 - 0 0 1 1 2 - 0 1 - 0 2 - 1 - 1 0 - 2	9	0	1
⩾ 10 contexts	10	3	1

Table 17 – Estimated τ trees for Model 4, with n = 8000 (frequency of occurrences over 100 simulations).

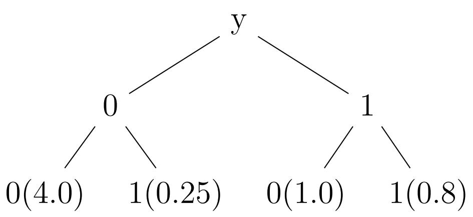
Contexts	# Contexts	modified-beta-VLMC	beta-VLMC
0 0 - 0 1 - 0 2 - 1 - 1 0 - 2	**6**	**100**	**79**
0 0 - 0 0 0 - 0 0 0 1 - 0 0 0 1 2 - 0 1 - 0 2 - 1 - 1 0 - 2	9	0	1
0 0 - 0 0 1 - 0 0 1 1 - 0 0 1 1 2 - 0 1 - 0 2 - 1 - 1 0 - 2	9	0	1
0 0 - 0 0 2 - 0 0 2 1 - 0 0 2 1 0 - 0 1 - 0 2 - 1 - 1 0 - 2	9	0	1
0 0 - 0 1 - 0 1 1 - 0 1 1 2 - 0 1 1 2 0 - 0 2 - 1 - 1 0 - 2	9	0	2
0 0 - 0 1 - 0 1 1 - 0 1 1 2 - 0 1 1 2 1 - 0 2 - 1 - 1 0 - 2	9	0	1
0 0 - 0 1 - 0 1 2 - 0 1 2 1 - 0 1 2 1 0 - 0 2 - 1 - 1 0 - 2	9	0	1
0 0 - 0 1 - 0 2 - 0 2 1 - 0 2 1 1 - 0 2 1 1 1 - 1 - 1 0 - 2	9	0	1
0 0 - 0 1 - 0 2 - 1 - 1 0 - 2 - 2 0 - 2 0 0 - 2 0 0 2	9	0	1
0 0 - 0 1 - 0 2 - 1 - 1 0 - 2 - 2 0 - 2 0 1 - 2 0 1 0	9	0	1
⩾ 10 contexts	10	0	11

$\alpha^{00} = 4$ and $\boldsymbol{\beta}^{00} = (0, 0)'$ to simulate scenarios with rare events. We generate samples with $n = 1000$ and $n = 2000$ state transitions for a single source.

The performance metrics for the estimated context function in Model 5 are detailed in Table 19. To aid in the visualization of the estimated trees, Table 20 and Table 21 provide the count of estimated occurrences for each context. For both $n = 1000$ and $n = 2000$, the modified beta-context algorithm demonstrated superior results, achieving 99% and 94% identical τ trees and exact covariate vectors, respectively, along with lower values of BIC and AIC. In general, the original beta-context algorithm results in larger

Table 18 – Differences between real and estimated values for Model 4 (average over 100 simulations).

	Model 4 (n = 4000, N(0, 1) distr.)		Model 4 (n = 8000, N(0,1) distr.)	
	modified-beta-VLMC	beta-VLMC	modified-beta-VLMC	beta-VLMC
$(\alpha_1^{00}, \alpha_2^{00})$	(0.09, 0.08)	(0.09, 0.08)	(0.09, 0.07)	(0.09, 0.07)
$(\alpha_1^{01}, \alpha_2^{01})$	(0.09, 0.09)	(0.09, 0.09)	(0.06, 0.05)	(0.07, 0.05)
$(\alpha_1^{02}, \alpha_2^{02})$	(0.15, 0.20)	(0.15, 0.20)	(0.10, 0.13)	(0.10, 0.13)
$(\alpha_1^{10}, \alpha_2^{10})$	(0.13, 0.11)	(0.13, 0.11)	(0.09, 0.07)	(0.09, 0.07)
$(\alpha_1^{1*}, \alpha_2^{1*})$	(0.05, 0.06)	(0.06, 0.06)	(0.04, 0.04)	(0.04, 0.04)
$(\alpha_1^{2}, \alpha_2^{2})$	(0.06, 0.09)	(0.06, 0.09)	(0.04, 0.06)	(0.04, 0.07)
$\boldsymbol{\beta}_1^{00}$	(0.29, 0)	(0.34, 0)	(0.16, 0)	(0.12, 0)
$\boldsymbol{\beta}_2^{00}$	(0.28, 0)	(0.34, 0)	(0.15, 0)	(0.11, 0)
$\boldsymbol{\beta}_1^{01}$	(0.10, 0)	(0.10, 0)	(0.08, 0)	(0.08, 0)
$\boldsymbol{\beta}_2^{01}$	(0.09, 0)	(0.09, 0)	(0.07, 0)	(0.07, 0)
$\boldsymbol{\beta}_1^{02}$	(0.19, 0.22)	(0.19, 0.22)	(0.13, 0.19)	(0.13, 0.18)
$\boldsymbol{\beta}_2^{02}$	(0.19, 0.20)	(0.19, 0.19)	(0.14, 0.15)	(0.13, 0.15)
$\boldsymbol{\beta}_1^{10}$	(0, 0)	(0, 0)	(0, 0)	(0, 0)
$\boldsymbol{\beta}_2^{10}$	(0, 0)	(0, 0)	(0, 0)	(0, 0)
$\boldsymbol{\beta}_1^{1*}$	(0.06)	(0.06)	(0.04)	(0.04)
$\boldsymbol{\beta}_2^{1*}$	(0.07)	(0.07)	(0.05)	(0.05)
$\boldsymbol{\beta}_1^{2}$	(0)	(0)	(0)	(0)
$\boldsymbol{\beta}_2^{2}$	(0)	(0)	(0)	(0)



$$\boldsymbol{\beta}^{00} = (0, 0)'$$
$$\boldsymbol{\beta}^{01} = (-1, 1)'$$
$$\boldsymbol{\beta}^{10} = (1.5, -2)'$$
$$\boldsymbol{\beta}^{11} = (-0.2, -0.9)'$$

Figure 7 – Context tree τ and associated parameters $\boldsymbol{\theta}$ for Model 5 (numbers in parenthesis represent the values of α^u for each context u).

trees.

The differences between the real and estimated parameters, when the context were correctly identified, are shown in Table 22 for both the proposed modified beta-context algorithm and the original beta-context algorithm. Both methods generally provide similar results for most parameters, except for α^{00}. This discrepancy arises because, in cases with insufficient observations, the original beta-context algorithm tends to estimate $\alpha^{00} = 0$, implying equal probabilities for events '0' and '1' after the context '00'. In reality, event '0' is a rare occurrence following this context. The modified version can estimate α^{00}, but it may yield higher values. This leads to a larger difference observed in Table 22. However, opting for higher values is preferable as it maintains the behavior of the rare event, contrasting with the original algorithm, which assigns the same probability to both

Table 19 – Simulation results for Model 5 with time-varying exogenous covariates generated from a N(0, 1) distribution (average over 100 simulations).

	Model 5 (n = 1000, N(0, 1) distr.)		Model 5 (n = 2000, N(0, 1) distr.)	
	modified-beta-VLMC	beta-VLMC	modified-beta-VLMC	beta-VLMC
BIC	991.98	1029.64	1941.36	1965.18
AIC	942.51	968.39	1881.32	1901.55
logLik	-461.18	-471.72	-929.94	-939.41
# par. $\hat{\alpha}^u$	4.03	5.15	4.23	5.54
# par. $\hat{\boldsymbol{\beta}}^u$	6.05	7.33	6.49	6.82
order $\hat{\tau}$	2.03	3.08	2.23	2.55
order-Cov	2.03	2.84	2.23	2.42
# Missing $\hat{\tau}$	0	0	0	0
# Extra $\hat{\tau}$	0.06	2.09	0.46	0.99
Identical τ	0.99	0.79	0.94	0.88
Identical $\tau_{\boldsymbol{\theta}}$	0.99	0.79	0.94	0.88

Table 20 – Estimated τ trees for Model 5, with n = 1000 (frequency of occurrences over 100 simulations).

Contexts	# Contexts	modified-beta-VLMC	beta-VLMC
0 0 - 0 1 - 1 0 - 1 1	**4**	**99**	**79**
0 0 - 0 1 0 0 - 0 1 0 1 0 - 0 1 0 1 1 - 0 1 1 - 1 0 - 1 1	7	1	0
>= 8 contexts	8	0	21

Table 21 – Estimated τ trees for Model 5, with n = 2000 (frequency of occurrences over 100 simulations).

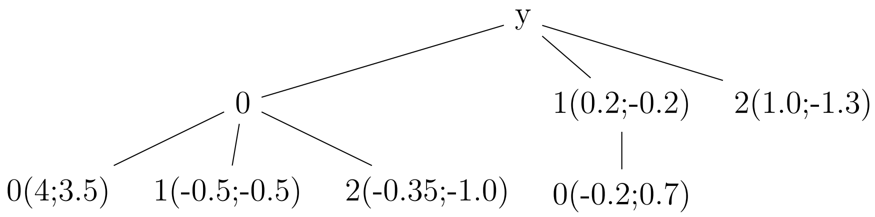
Contexts	# Contexts	modified-beta-VLMC	beta-VLMC
0 0 - 0 1 - 1 0 - 1 1	**4**	**94**	**88**
0 0 - 0 1 - 1 0 0 - 1 0 1 0 - 1 0 1 1 - 1 1	6	1	0
>= 8 contexts	8	5	12

events in such situations.

Table 22 – Differences between real and estimated values for Model 5 (average over 100 simulations).

	Model 5 (n = 1000, N(0, 1) distr.)		Model 5 (n = 2000, N(0,1) distr.)	
	modified-beta-VLMC	beta-VLMC	modified-beta-VLMC	beta-VLMC
α^{00}	6.79	0.96	5.34	0.58
α^{01}	0.12	0.13	0.10	0.10
α^{10}	0.17	0.17	0.13	0.13
α^{11}	0.10	0.10	0.07	0.07
$\boldsymbol{\beta}^{00}$	(0, 0)	(0, 0)	(0, 0)	(0, 0)
$\boldsymbol{\beta}^{01}$	(0.14, 0.14)	(0.14, 0.14)	(0.09, 0.10)	(0.10, 0.10)
$\boldsymbol{\beta}^{10}$	(0.23, 0.26)	(0.23, 0.27)	(0.17, 0.17)	(0.17, 0.17)
$\boldsymbol{\beta}^{11}$	(0.09, 0.11)	(0.09, 0.11)	(0.06, 0.07)	(0.06, 0.07)

Model 6, described in Figure 8, shares similarities with Model 4 but includes $\alpha^{00} = (4, 3.5)$, $\boldsymbol{\beta}_1^{00} = (0, 0)'$ and $\boldsymbol{\beta}_2^{00} = (0, 0)'$ to simulate scenarios with rare events. We generate samples with $n = 4000$ and $n = 8000$ state transitions for a single source.



$$\begin{array}{ll}
\boldsymbol{\beta}_1^{00} = (0,0)' & \boldsymbol{\beta}_2^{00} = (0,0)' \\
\boldsymbol{\beta}_1^{01} = (0.9,0)' & \boldsymbol{\beta}_2^{01} = (0.3,0)' \\
\boldsymbol{\beta}_1^{02} = (1.5,-2)' & \boldsymbol{\beta}_2^{02} = (0.5,-0.85)' \\
\boldsymbol{\beta}_1^{10} = (0,0)' & \boldsymbol{\beta}_2^{10} = (0,0)' \\
\boldsymbol{\beta}_1^{1*} = (-0.4)' & \boldsymbol{\beta}_2^{1*} = (0.6)' \\
\boldsymbol{\beta}_1^{2} = (0)' & \boldsymbol{\beta}_2^{2} = (0)'
\end{array}$$

Figure 8 – Context tree τ and associated parameters $\boldsymbol{\theta}$ for Model 6. Numbers in parenthesis represent the values of (α_1^u, α_2^u) from each context u and $\boldsymbol{\beta}_1^u$ and $\boldsymbol{\beta}_2^u$ are the coefficient vectors of the covariates. $\boldsymbol{\beta}^{1*}$ means the context is 1 preceded by other state that is not 1.

The performance metrics for the estimated context function in Model 6 are outlined in Table 23. To aid in the visualization of the estimated trees, Table 24 and Table 25 provide the count of estimated occurrences for each context. For both $n = 4000$ and $n = 8000$, the modified beta-context algorithm exhibited superior results, achieving 100% identical τ trees and exact covariates vectors and lower values of BIC and AIC. In general, the original beta-context algorithm resulted in larger trees, with an average of 0.56 and 1.13 extra nodes, respectively, for $n = 4000$ and $n = 8000$.

Table 23 – Simulation results for Model 6 with time-varying exogenous covariates generated from a N(0, 1) distribution (average over 100 simulations).

	Model 6 (n = 4000, N(0, 1) distr.)		Model 6 (n = 8000, N(0, 1) distr.)	
	modified-beta-VLMC	beta-VLMC	modified-beta-VLMC	beta-VLMC
BIC	7452.25	7465.61	14757.67	14792.52
AIC	7326.37	7326.51	14617.92	14619.93
logLik	-3643.18	-3641.15	-7288.96	-7285.27
# par. $\hat{\alpha}^u$	12.00	13.04	12.00	14.18
# par. $\hat{\boldsymbol{\beta}}^u$	8.00	9.06	8.00	10.52
order $\hat{\tau}$	2.00	2.40	2.00	2.79
order-Cov	2.00	2.29	2.00	2.55
# Missing $\hat{\tau}$	0	0	0	0
# Extra $\hat{\tau}$	0	0.56	0	1.13
Identical τ	1.00	0.89	1.00	0.74
Identical $\tau_{\boldsymbol{\theta}}$	1.00	0.89	1.00	0.74

Table 24 – Estimated τ trees for Model 6, with n = 4000 (frequency of occurrences over 100 simulations).

contexts	# contexts	modified-beta-VLMC	beta-VLMC
0 0 - 0 1 - 0 2 - 1 - 1 0 - 2	**6**	**100**	**89**
0 0 - 0 1 - 0 2 - 1 - 1 0 - 1 0 1 - 1 0 1 1 - 1 0 1 1 2 - 2	9	0	2
0 0 - 0 1 - 0 2 - 1 - 1 0 - 2 - 2 0 - 2 0 2 - 2 0 2 1	9	0	1
>= 10 contexts	10	0	8

Table 25 – Estimated τ trees for Model 6, with n = 8000 (frequency of occurrences over 100 simulations).

Contexts	# Contexts	modified-beta-VLMC	beta-VLMC
0 0 - 0 1 - 0 2 - 1 - 1 0 - 2	**6**	**100**	**74**
0 0 - 0 1 - 0 2 - 0 2 2 - 1 - 1 0 - 2	7	0	1
0 0 - 0 0 1 - 0 0 1 0 - 0 1 - 0 2 - 1 - 1 0 - 2	8	0	1
0 0 - 0 0 1 - 0 0 1 1 - 0 1 - 0 2 - 1 - 1 0 - 2	8	0	1
0 0 - 0 1 - 0 2 - 1 - 1 0 - 1 0 2 - 1 0 2 0 - 2	8	0	1
>= 9 contexts	9	0	22

The differences between the real and estimated parameters, when the context were correctly identified, are presented in Table 26 for both the proposed modified beta-context algorithm and the original beta-context algorithm. Both methods yield similar results for all parameters. Since we have a larger number of observations, we do not encounter the issue explained earlier for Model 5, where insufficient observations could impede the estimation of rare events.

3.2 Simulations with multiple independent sources and time-invariant exogenous covariates

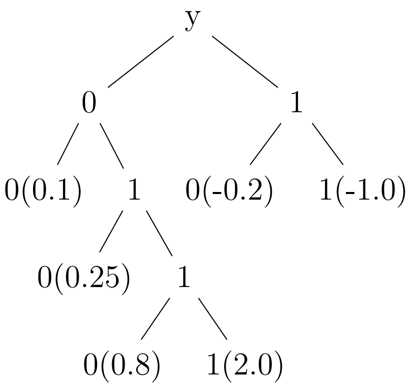
Model 7, outlined in Figure 9, involve multiple independent sources and time-invariant exogenous covariates. Consequently, only the modified version of the beta-context algorithm was employed, as it includes adaptations to accommodate these characteristics. It was considered univariate time-varying and time-invariant exogenous covariates, both generated by a N(0, 1) distribution.

Model 7 shares similarities with Model 1 (Figure 3), incorporating a univariate time-invariant exogenous covariate and multiple sources. Different simulations were conducted for Model 7, varying sample sizes and considering sources with both equal and varying numbers of observations. Additionally, scenarios were simulated with an identical overall sample size, but with varying sizes for each source. This approach enables an evaluation of the algorithm's performance as sample sizes increase across all sources or only for specific ones. Furthermore, it allows an assessment of cases where there is a high total number of observations but fewer observations per source, reflecting situations with numerous sources, each contributing a limited number of observations.

Table 26 – Differences between real and estimated values for Model 6 (average over 100 simulations).

	Model 6 (n = 4000, N(0, 1) distr.)		Model 6 (n = 8000, N(0,1) distr.)	
	modified-beta-VLMC	beta-VLMC	modified-beta-VLMC	beta-VLMC
$(\alpha_1^{00}, \alpha_2^{00})$	(0.48, 0.49)	(0.48, 0.49)	(0.31, 0.32)	(0.36, 0.36)
$(\alpha_1^{01}, \alpha_2^{01})$	(0.09, 0.09)	(0.09, 0.09)	(0.07, 0.06)	(0.07, 0.06)
$(\alpha_1^{02}, \alpha_2^{02})$	(0.16, 0.19)	(0.16, 0.19)	(0.10, 0.14)	(0.10, 0.14)
$(\alpha_1^{10}, \alpha_2^{10})$	(0.10, 0.09)	(0.10, 0.09)	(0.08, 0.07)	(0.08, 0.07)
$(\alpha_1^{1*}, \alpha_2^{1*})$	(0.05, 0.06)	(0.05, 0.06)	(0.04, 0.04)	(0.04, 0.04)
$(\alpha_1^{2}, \alpha_2^{2})$	(0.05, 0.12)	(0.05, 0.12)	(0.04, 0.07)	(0.04, 0.07)
$\boldsymbol{\beta}_1^{00}$	(0, 0)	(0, 0)	(0, 0)	(0, 0)
$\boldsymbol{\beta}_2^{00}$	(0, 0)	(0, 0)	(0, 0)	(0, 0)
$\boldsymbol{\beta}_1^{01}$	(0.11, 0)	(0.11, 0)	(0.07, 0)	(0.07, 0)
$\boldsymbol{\beta}_2^{01}$	(0.09, 0)	(0.09, 0)	(0.07, 0)	(0.07, 0)
$\boldsymbol{\beta}_1^{02}$	(0.18, 0.21)	(0.18, 0.21)	(0.15, 0.16)	(0.14, 0.17)
$\boldsymbol{\beta}_2^{02}$	(0.20, 0.20)	(0.20, 0.20)	(0.14, 0.15)	(0.14, 0.15)
$\boldsymbol{\beta}_1^{10}$	(0, 0)	(0, 0)	(0, 0)	(0, 0)
$\boldsymbol{\beta}_2^{10}$	(0, 0)	(0, 0)	(0, 0)	(0, 0)
$\boldsymbol{\beta}_1^{1*}$	(0.06)	(0.06)	(0.04)	(0.04)
$\boldsymbol{\beta}_2^{1*}$	(0.05)	(0.06)	(0.05)	(0.05)
$\boldsymbol{\beta}_1^{2}$	(0)	(0)	(0)	(0)
$\boldsymbol{\beta}_2^{2}$	(0)	(0)	(0)	(0)

We chose a model similar to Model 1 because it exhibited worse estimation results, and we aimed to assess the accuracy of the model with multiple independent sources and time-invariant exogenous covariates in challenging scenarios.



$$\begin{aligned} \boldsymbol{\beta}^{00} &= (2,0)' \\ \boldsymbol{\beta}^{010} &= (-1,1,0)' \\ \boldsymbol{\beta}^{0111} &= (1.5,2,0,0)' \\ \boldsymbol{\beta}^{0110} &= (4,3,2,1)' \\ \boldsymbol{\beta}^{10} &= (0,0)' \\ \boldsymbol{\beta}^{11} &= (0,0)' \end{aligned}$$

$$\begin{aligned} \gamma^{00} &= 3 \\ \gamma^{010} &= 1.5 \\ \gamma^{0111} &= 1 \\ \gamma^{0110} &= 0.5 \\ \gamma^{10} &= -2 \\ \gamma^{11} &= -1 \end{aligned}$$

Figure 9 – Context tree τ and associated parameters $\boldsymbol{\theta}$ for Model 7 (numbers in parenthesis represent the values of α^u for each context u).

The performance metrics for the estimated context function in Model 7 are outlined in Table 27. To aid in the visualization of the estimated trees, Table 28, Table 29, Table 30, Table 31, Table 32 and Table 33 provide the count of estimated occurrences for each context. As the overall sample size increases, there is an improvement in the percentage of identical τ trees and exact covariate vectors $\tau_{\boldsymbol{\theta}}$. When comparing balanced and imbalanced sample sizes with the same overall size, a higher percentage of exact covariate vectors $\tau_{\boldsymbol{\theta}}$ is observed in the balanced sample. This might occur because in sources with small samples, we cannot obtain a large number of large contexts, thus lacking sufficient data in only one source to estimate all the parameters of the larger contexts. Additionally, the variances observed could also arise from differences in the samples themselves. It is worth noting that for $n_1 = 4000$ and $n_2 = 4000$, not only do we achieve better results, but we also obtain larger trees. This is expected because now we have larger contexts with enough data to enter the initial tree (maximal tree), but not enough data to comprehensively understand the behavior within these contexts.

Table 27 – Simulation results for Model 7 with time-varying exogenous covariates generated from a N(0, 1) distribution (average over 100 simulations).

	Model 7 (N(0, 1) distr.)					
	$n_1 = 500$	$n_1 = 1000$	$n_1 = 1000$	$n_1 = 2000$	$n_1 = 200$	$n_1 = 4000$
	$n_2 = 500$	$n_2 = 1000$	$n_2 = 2000$	$n_2 = 2000$	$n_2 = 3800$	$n_2 = 4000$
BIC	937.74	1783.47	2657.37	3431.95	3359.24	6752.52
AIC	875.65	1696.38	2554.49	3304.05	3254.51	6581.69
logLik	-425.20	-832.60	-1260.10	-1631.70	-1610.60	-1639.79
# par. $\hat{\alpha}^u$	5.04	5.43	5.77	7.34	5.48	7.42
# par. $\hat{\boldsymbol{\beta}}^u$	3.60	5.18	5.81	7.06	5.85	9.92
order $\hat{\tau}$	3.18	3.47	3.93	4.25	3.66	4.77
order-Cov	2.01	2.46	2.51	3.14	2.70	3.98
# Missing $\hat{\tau}$	3.10	2.33	1.79	1.27	1.69	0.51
# Extra $\hat{\tau}$	1.50	1.22	1.34	3.96	0.72	3.36
Identical τ	0.02	0.11	0.36	0.41	0.42	0.57
Identical $\tau_{\boldsymbol{\theta}}$	0.00	0.00	0.00	0.15	0.05	0.46

Table 28 – Estimated τ trees for Model 7, with $n_1 = 500$ and $n_2 = 500$ (frequency of occurrences over 100 simulations).

Contexts	# Contexts	modified-beta-VLMC
0 0 - 0 1 - 1	3	17
0 0 - 0 1 - 1 0 - 1 1	4	4
0 0 - 0 1 0 - 0 1 1 - 1	4	21
0 0 - 0 1 0 - 0 1 1 - 1 0 - 1 1	5	34
0 0 - 0 1 0 - 0 1 1 0 - 0 1 1 1 - 1	5	2
0 0 0 - 0 0 1 - 0 1 - 1 0 - 1 1	5	1
0 0 0 0 - 0 0 0 1 - 0 0 1 - 0 1 - 1	5	2
0 0 - 0 1 0 - 0 1 1 0 - 0 1 1 1 - 1 0 - 1 1	**6**	**2**
0 0 - 0 1 0 0 0 - 0 1 0 0 1 - 0 1 0 1 - 0 1 1 - 1	6	1
⩾ 7 contexts	7	16

The disparities between the real and estimated parameters, when the context was correctly identified, are presented in Table 34. An interesting finding is that for samples

Table 29 – Estimated τ trees for Model 7, with $n_1 = 1000$ and $n_2 = 1000$ (frequency of occurrences over 100 simulations).

Contexts	# Contexts	modified-beta-VLMC
0 0 - 0 1 - 1	3	1
0 0 - 0 1 - 1 0 - 1 1	4	1
0 0 - 0 1 0 - 0 1 1 - 1	4	26
0 0 - 0 1 0 - 0 1 1 - 1 0 - 1 1	5	43
0 0 - 0 1 0 - 0 1 1 0 - 0 1 1 1 - 1	5	4
0 0 - 0 1 0 - 0 1 1 0 - 0 1 1 1 - 1 0 - 1 1	**6**	**11**
⩾ 7 contexts	7	14

Table 30 – Estimated τ trees for Model 7, with $n_1 = 1000$ and $n_2 = 2000$ (frequency of occurrences over 100 simulations).

Contexts	# Contexts	modified-beta-VLMC
0 0 - 0 1 - 1 0 - 1 1	4	1
0 0 - 0 1 0 - 0 1 1 - 1	4	23
0 0 - 0 1 0 - 0 1 1 - 1 0 - 1 1	5	18
0 0 - 0 1 0 - 0 1 1 0 - 0 1 1 1 - 1	5	8
0 0 - 0 1 0 - 0 1 1 0 - 0 1 1 1 - 1 0 - 1 1	**6**	**36**
⩾ 7 contexts	7	14

Table 31 – Estimated τ trees for Model 7, with $n_1 = 2000$ and $n_2 = 2000$ (frequency of occurrences over 100 simulations).

Contexts	# Contexts	modified-beta-VLMC
0 0 - 0 1 - 1	3	1
0 0 - 0 1 0 - 0 1 1 - 1	4	13
0 0 - 0 1 0 - 0 1 1 - 1 0 - 1 1	5	14
0 0 - 0 1 0 - 0 1 1 0 - 0 1 1 1 - 1	5	6
0 0 - 0 1 0 - 0 1 1 0 - 0 1 1 1 - 1 0 - 1 1	**6**	**41**
⩾ 7 contexts	7	25

Table 32 – Estimated τ trees for Model 7, with $n_1 = 200$ and $n_2 = 3800$ (frequency of occurrences over 100 simulations).

Contexts	# Contexts	modified-beta-VLMC
0 0 - 0 1 - 1 0 - 1 1	4	3
0 0 - 0 1 0 - 0 1 1 - 1	4	24
0 0 - 0 1 0 - 0 1 1 - 1 0 - 1 1	5	13
0 0 - 0 1 0 - 0 1 1 0 - 0 1 1 1 - 1	5	9
0 0 - 0 1 - 1 0 0 0 - 1 0 0 1 - 1 0 1 - 1 1	6	1
0 0 - 0 1 0 - 0 1 1 0 - 0 1 1 1 - 1 0 - 1 1	**6**	**42**
0 0 0 - 0 0 1 0 - 0 0 1 1 - 0 1 - 1 0 - 1 1	6	1
⩾ 7 contexts	7	7

with an overall sample size of 4000, imbalanced samples yield poorer results for parameter estimation, particularly for parameters related to time-invariant exogenous covariates.

Table 33 – Estimated τ trees for Model 7, with $n_1 = 4000$ and $n_2 = 4000$ (frequency of occurrences over 100 simulations).

Contexts	# Contexts	modified-beta-VLMC
0 0 - 0 1 - 1 0 - 1 1	4	1
0 0 - 0 1 0 - 0 1 1 - 1	4	3
0 0 - 0 1 0 - 0 1 1 0 - 0 1 1 1 - 1	5	14
0 0 0 - 0 0 1 - 0 1 0 - 0 1 1 - 1	5	1
0 0 - 0 1 0 - 0 1 1 0 - 0 1 1 1 - 1 0 - 1 1	**6**	**57**
$\geqslant$ 7 contexts	7	24

Better results are achieved for balanced data with a larger sample size.

Table 34 – Differences between real and estimated values for Model 7 (average over 100 simulations).

	Model 7 (N(0, 1) distr.)					
	$n_1 = 500$	$n_1 = 1000$	$n_1 = 1000$	$n_1 = 2000$	$n_1 = 200$	$n_1 = 4000$
	$n_2 = 500$	$n_2 = 1000$	$n_2 = 2000$	$n_2 = 2000$	$n_2 = 3800$	$n_2 = 4000$
α^{00}	0.67	0.76	0.53	0.35	0.76	0.36
α^{010}	0.76	0.94	0.84	0.34	0.79	0.55
α^{0110}	1.04	0.98	0.47	1.19	0.80	0.67
α^{0111}	0.22	0.87	1.93	0.72	1.82	1.00
α^{10}	0.66	0.43	3.96	0.22	0.28	0.31
α^{11}	0.63	0.32	0.42	0.15	0.60	0.17
γ^{00}	1.37	0.97	0.72	0.52	1.23	0.88
γ^{010}	1.56	1.18	1.18	0.61	1.34	1.16
γ^{0110}	8.97	2.47	1.09	1.47	2.77	1.48
γ^{0111}	1.00	1.67	1.73	0.99	4.89	2.48
γ^{10}	1.27	0.62	3.94	0.37	0.91	0.55
γ^{11}	1.53	0.40	0.66	0.29	0.76	0.40
$\boldsymbol{\beta}^{00}$	(0.38, 0)	(0.24, 0)	(0.16, 0)	(0.11, 0)	(0.16, 0)	(0.09, 0)
$\boldsymbol{\beta}^{010}$	(0.38, 0.46, 0)	(0.24, 0.30, 0)	(0.26, 0.28, 0)	(0.12, 0.14, 0)	(0.33, 0.36, 0)	(0.09, 0.09, 0)
$\boldsymbol{\beta}^{0110}$	(1.54, 1.67, 1.20, 1.00)	(0.80, 0.50, 0.75, 0.96)	(0.63, 0.50, 0.41, 1.00)	(0.63, 0.53,0.38, 0.71)	(0.51, 0.44, 0.34, 0.89)	(0.46, 0.37, 0.26, 0.38)
$\boldsymbol{\beta}^{0111}$	(1.50, 2.00, 0, 0)	(0.86, 1.28, 0, 0)	(1.22, 1.98, 0, 0)	(0.53, 0.71, 0, 0)	(0.46, 0.56, 0, 0)	(0.31, 0.37, 0, 0)
$\boldsymbol{\beta}^{10}$	(0, 0)	(0, 0)	(0, 0)	(0, 0)	(0, 0)	(0, 0)
$\boldsymbol{\beta}^{11}$	(0, 0)	(0, 0)	(0, 0)	(0, 0)	(0, 0)	(0, 0)

4 Application: Predicting Dengue outbreaks

4.1 Dengue: an overview of the disease and its global impact

Dengue is an urban arboviral disease [1] transmitted by infected female mosquitoes of the genus *Aedes* [2], primarily *Aedes aegypti*. Once infected, a female mosquito becomes a permanent disease vector, with 30 to 40% chance of transmitting the virus to its offspring (Rita, Freitas e Nogueira (2016)). These mosquitoes, known for their strictly urban behavior, lay eggs just above clean water surfaces in containers like cans, bottles, and tires, which hatch within minutes upon contact with rising water levels. On that account, the mosquito density is higher in the summer, when the periods of elevated rainfall increase the number of breeding sites. Additionally, higher temperatures during the summer accelerate the mosquito's development through the egg-larva-adult stages.

According to the World Health Organization (WHO) [3], besides transmission between humans involving mosquito vectors, there is also evidence of the possibility of maternal transmission from a pregnant mother to her baby. However, vertical transmission rates appear to be low, with the risk seemingly linked to the timing of the dengue infection during pregnancy - in cases where a mother has dengue infection while pregnant, babies may suffer from preterm birth, low birth weight, and fetal distress.

Since dengue is primarily transmitted amongst humans through mosquito vectors, preventing the spread of *Aedes aegypti* is essential. This is most effective during the insect aquatic phase, focusing on larvae and pupa removal or covering potential breeding sites (Rita, Freitas e Nogueira (2016)). As such, public awareness and continuous home monitoring are crucial stances to keep the disease rates under control, and both are heavily dependent on forethought public policies and institutional actions designed to create an effective sanitary culture.

Regarding global impact, dengue's incidence has grown dramatically in recent decades, with an estimated 390 million dengue virus infections annually, out of which 96 million manifest clinically (Brady et al. (2012)). According to the WHO, prior to the COVID-19 pandemic in 2020, dengue fever was ranked amongst the top ten global health threats. [4]. The disease is now endemic in over 100 countries in Regions of Africa, the Americas, the Eastern Mediterranean, South-East Asia, and the Western Pacific. The most

[1] Arboviral diseases are caused by viruses transmitted by mosquitoes. The most common arboviral diseases in urban environments are: Dengue, Chikungunya and Zika.

[2] These mosquitoes can be infected with four different serotypes of the Dengue virus: DENV-1, DENV-2, DENV-3, and DENV-4.

[3] https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue

[4] https://www.who.int/news-room/spotlight/ten-threats-to-global-health-in-2019

severely affected regions are the Americas, South-East Asia, and the Western Pacific, with Asia contributing to approximately 70% of the global disease burden. Notably, dengue is extending its reach to new areas, including Europe, leading to explosive outbreaks. In 2010, local transmission was reported for the first time in France and Croatia, and imported cases were detected in three other European countries [5].

4.2 Dengue as an endemic disease in Brazil

The incidence of dengue in Brazil follows endemic and epidemic cycles, with explosive outbreaks occurring approximately every 4 to 5 years. Since the introduction of the virus in the country (1981), more than seven million cases have been reported.

In the last two years (2022 and 2023), Brazil faced its highest recorded dengue-related deaths: at least 1016 [6] and 1079 [7] fatalities, respectively, as reported by the Notifiable Diseases Information System (Sistema de Informação de Agravos de Notificação - SINAN) [8]. Climate change, as noted by the Butantan Institute [9], has allowed the disease vector to adapt and spread to regions where it was not previously common.

According to National Institute of Meteorology (Instituto Nacional de Metereologia - INMET), temperatures in Brazil have consistently exceeded historical averages since the 1990s. This warming climate provides favorable conditions for the adaptation and proliferation of the dengue-transmitting mosquito. In the southern part of the country, increased rainfall and higher average temperatures led the region to be ranked second in 2022's dengue incidence rates, according to the Epidemiological Bulletin of the Ministry of Health [10] — until mid-2015, the presence of dengue in the region was variable and not very significant, as shown by Vecchia, Beltrame e D'Agostini (2018).

4.3 Dengue as a Neglected Tropical Diseases (NTDs)

Dengue is classified by the World Health Organization (WHO) as a Neglected Tropical Disease (NTD) – a diverse group of infirmities that gathers 20 conditions [11], often

[5] https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue

[6] https://www.gov.br/saude/pt-br/centrais-de-conteudo/publicacoes/boletins/epidemiologicos/edicoes/2023/boletim-epidemiologico-volume-54-no-01/

[7] https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/a/aedes-aegypti/monitoramento-das-arboviroses

[8] http://portalsinan.saude.gov.br/dengue

[9] https://butantan.gov.br/noticias/aumento-historico-de-temperatura-leva-a-disseminacao-da-dengue-em-todo-o-brasil

[10] https://www.gov.br/saude/pt-br/centrais-de-conteudo/publicacoes/boletins/epidemiologicos

[11] NTDs include: Buruli ulcer, Chagas disease, dengue and chikungunya, dracunculiasis (Guinea-worm disease), echinococcosis, foodborne trematodiases, human African trypanosomiasis (sleeping sickness), leishmaniasis, leprosy (Hansen's disease), lymphatic filariasis, mycetoma, chromoblastomycosis and other deep mycoses, onchocerciasis (river blindness), podoconiosis, rabies, scabies and other ectopara-

linked to environmental factors – and is responsible for approximately 20000 avoidable annual deaths. NTDs are usually vector-borne, involving animal reservoirs and exhibiting complex life cycles, which leads to challenges in their spread control. Predominantly, this category of disease affects impoverished communities in tropical and subtropical regions, with a disproportionate impact on women and children. Those NTDs have already affected over a billion people around the world, inflicting significant health, social, and economic challenges to the people, and often resulting in consequences such as disability, stigmatization, social exclusion, discrimination, plus the imposition of substantial financial burdens on patients and their families.

Horstick, Tozan e Wilder-Smith (2015) summarizes NTD definitions with key features:

- Poverty-related;
- Endemic to the tropics and subtropics;
- Lacking public health attention;
- Poor research funding and shortcomings in research and development (R&D);
- Usually associated with high morbidity but low mortality;
- Often having no specific treatment available.

They discuss dengue's classification as an NTD, since it does not meet some of the requirements listed above - although its prevalence is larger in countries with less economic power, it does not affect only the poor, and recently, dengue has been attracting public health attention and research funding, particularly for vaccine development. On the other hand, dengue epidemics are on the rise both in occurrence and severity and there is currently no specific treatment or widely accessible highly effective vaccine. Moreover, effective methods for surveillance and vector control remain elusive. Therefore, it is important that dengue be considered an NTD. For further details, Horstick, Tozan e Wilder-Smith (2015) provide an in-depth discussion on the classification of dengue as a NTD.

4.4 Reviewing forecasting models for dengue incidence

In addition to efforts in dengue vaccine development and combating *Aedes aegypti*, another strategy for dengue-endemic countries is the development of forecasting models for outbreak prediction - this way, public health systems can be prepared in terms

sitoses, schistosomiasis, soil-transmitted helminthiases, snakebite envenoming, taeniasis/cysticercosis, trachoma, and yaws and other endemic treponematoses.

of resource and protocols to the high influx of patients. WHO supports this strategy by providing financial assistance for innovative surveillance systems, enhancing prevention, control, and forecasting efficiency (Organization et al. (2012)). Stanaway et al. (2016) conducted a literature review, summarizing researchers' efforts to gather and analyze data, improving understanding of the relational factors influencing disease spread. The review also emphasizes the evolution of various predictive modeling methods, including statistical and mathematical analyses as well as machine learning techniques.

In this section, we will rely on the literature review conducted by Stanaway et al. (2016) to offer a brief overview of common relational factors used by researchers in dengue forecast modeling and the predictive modeling methods they employ.

4.4.1 Factors correlated to the number of dengue cases

There are direct and indirect factors that can be correlated to the number of dengue cases. Direct factors are the ones that directly affect the mosquitoes' live cycle. Indirect factors are those that do not have a direct impact on the number of mosquito larvae, but may be associated with the occurrence of a dengue epidemic.

Below, the direct and indirect factors mentioned are based on the enumeration by Stanaway et al. (2016), derived from their literature review on dengue forecast modeling.

- **Direct factors:**

  1. **Climate:** Rainfall significantly impacts the incubation period of mosquitoes, as they require still or standing water to complete their life cycle (Buczak et al. (2012)). This way, unusual weather conditions, such as drought and higher temperatures associated with phenomena like *El Niño*, adversely affect mosquito breeding habitats and populations. Interestingly, some studies suggest that excessive rainfall can lead to a decline in dengue epidemics. This may be attributed to the disruptive effects of fast-moving water flows on larvae survival and growth, as well as a reduction in the larval population (Hsu, Wen e Yu (2013), Thammapalo et al. (2005), Arcari, Tapper e Pfueller (2007)).
  2. **Mosquito density:** Several researchers have used mosquito populations information through a mosquito larvae density index - the percentage of houses in an area infested with larvae, the percentage of water-holding containers infested with larvae, the number of positive containers per 100 houses inspected and the number of pupae per 100 houses inspected.
  3. **Dengue virus serotypes:** Limkittikul, Brett e L'Azou (2014) found that the distribution of dengue serotypes varies seasonally and geographically. This variation impacts dengue incidence given that anyone infected with one serotype

will gain lifelong immunity to that virus serotype, but only partial or temporary protection against other serotypes. Fluctuating serotype incidence can lead to periods of no immunity and continual infections, correlating with the density of the dengue virus in the region (Veeraseatakul, Saosathan e Chutipongvivate (2014)).

4. **Bite rate:** Different severity levels of the epidemic depend significantly on the biting rate of the mosquitoes, which is one of the factors significantly associated with disease outbreaks and is a factor that could be used in the prediction of outbreaks. The bite rate of mosquitoes will vary from season to season and depends on weather and other climatic conditions, and mosquito density (Chompoosri et al. (2012)).

- **Indirect factors:**

  1. **Geography:** Topography plays a crucial role in providing an appropriate environment for mosquito growth, reproduction, and virus transmission. According to a study by the Center for Disease Control and Prevention (Control, (CDC et al. (2007)), border areas are identified as having a higher risk for dengue outbreaks compared to other regions. Additionally, rural areas in developing countries have experienced the rapid spread of dengue infections, possibly due to insufficient public health resources.

  2. **Spatial and spatial-temporal information:** There are many works in the literature that study the need to incorporate spatial analysis in to the modeling of outbreaks of dengue specially in terms of prediction. For example, Yu et al. (2014) proposed a spatial-temporal model incorporating population density, environmental conditions, and infrastructure factors. However, Costa et al. (2015) argued that relying solely on spatial information may not be sufficient, as *Ae. aegypti* is strongly influenced by local climate triggers. Thiruchelvam et al. (2018) found that dengue incidences were localized, and feedback models tailored to specific regions did not benefit from data from neighboring areas.

  3. **Population movement** Population movements and migration significantly contribute to the spread of dengue epidemics, whether for tourism, work, or other reasons. Additionally, international trade plays an indirect role in dengue outbreaks. Hawley et al. (1987) highlighted the spread of dengue from Asia to North America and Europe through the international trade of items like used tires and bamboo home decorative items, which serve as habitats for *Ae. albopictus*. Additionally, Jr, Stoddard e Scott (2014) found that specific dates such as religious holidays or festivities where there is great mobility of family members, tended to be associated with dengue outbreaks.

4. **Environment:** The majority of dengue outbreaks are observed in communities with unhygienic housing conditions and malnourished populations, contributing to lower immunity and an increased risk of disease infection (Organization et al. (2014)). Additionally, urban growth creates numerous breeding areas for dengue vectors - mosquitoes adapt well to this environment as they have a preference for feeding on humans over animals. Moreover, the ability of mosquitoes to fly within a range of 100-500 meters to find food and breeding sites increases the risk of mosquito-borne infections in urban areas, as houses fall within this flight range.
5. **Immunology:** Some researchers found that patients with dengue virus infection are often have poor nutrition, as nutritional deficiency negatively impacts the body's immune response to the dengue virus (Halstead, Nimmannitya e Cohen (1970), Waidab, Suphapeetiporn et al. (2008), Thisyakorn e Nimmannitya (1993)). Additionally, individuals with low levels of antibodies against the dengue virus, such as children when first exposed to it, are more likely to recurrent dengue virus infections. Oki e Yamamoto (2012) discovered that a decline in population immunity correlates with an increase in the severity of dengue outbreaks.

4.4.2 Common forecasting methods for dengue incidence

In their survey, Stanaway et al. (2016) identified common methods employed for detecting dengue outbreaks, forecasting future dengue cases and determining critical regions. They reported a total of 966 models created for the analysis of dengue epidemics, with 545 using regression methods, 220 using temporal series, 76 using neural networks, 50 using decision trees, 23 using suppot cector machine, 20 using k-means, 17 using association rules, 9 using lattice models and 6 using K-Nearest Neighbor. Some of these models are mentioned below:

- **Decision trees:** Decision trees have been employed both as a classifier for dengue cases (Tanner et al. (2008)) and to determine the choice between inpatient and outpatient treatment regimens for dengue infection (Lee et al. (2009)).
- **Regression analysis:** Regression models were employed for various purposes: predicting the duration of government intervention to control dengue epidemics and prevent further public health damage (Hii et al. (2012), Xu et al. (2014)), identifying and correlating factors contributing to dengue disease (Siriyasatien et al. (2016)), determining the relationship between dengue outbreaks and variables such as weather and dengue cases (Sang et al. (2014)), and predicting the age of dengue virus-infected mosquitoes (Hugo et al. (2014)).

- **Artificial Neural Network:** ANN, specifically Multilayer Feed-forward Neural Network, was utilized to identify individuals at risk of dengue outbreaks through Bio-electrical Impedance Analysis (Ibrahim et al. (2010)). Additionally, Ibrahim et al. (2005) created a system for predicting the peak day of dengue-induced fever in a patient, a crucial point associated with a higher risk of shock.

- **Support Vector Machine:** Support Vector Machine was used in conjunction with the Radial Basis Function for predicting the human mortality rate of dengue infection (Kesorn et al. (2015)).

- **K-Nearest Neighbor:** Spatial information on the risk areas of dengue infection was analyzed using the Nearest Neighbor Index, focusing on dengue hemorrhagic data from 1998 to 2004. The study revealed dengue movement patterns from rural to urban communities in Trinidad (Sharma et al. (2014)).

- **K-means:** K-means clustering is employed to classify distinct groups of genes, analyzing the relationship between the genetics of the virus in *Ae. aegypti* and the virus in patients (Chauhan et al. (2012)). Additionally, K-means identifies hotspots and localized regions of high dengue incidences in Malaysia (Mathur et al. (2015)). The K-medoids algorithm, related to K-means, diagnoses dengue outbreaks in India using mosquito species data, predicting the number and age groups of potential dengue patients (Manivannan e Isakki (2017)).

- **Time series analysis:** Time series methods, such as those mentioned by Hii et al. (2012), Johansson et al. (2016), Gharbi et al. (2011), Bhatnagar et al. (2012), Ho e Ting (2015), Lal et al. (2013), Lin et al. (2012), Siregar, Makmur e Saprin (2018), are widely employed for dengue prediction. These methods utilize data collected periodically over time, such as the number of patients each month over several years.

- **Association rules:** Buczak et al. (2014) utilized Fuzzy Association Rule Mining, a method to extract rules that relate variables like economic, social and weather conditions to develop predictive models for dengue epidemics.

- **Lattice models:** A lattice model enable researchers to explore the transmission dynamics of vector-borne diseases, considering both human mobility and vector movement in space and time. Botari, Alves e Leonel (2011) used vector movement to model registered dengue cases in Rio de Janeiro, Brazil, from 2006 to 2008, effectively explaining the unusually high number of cases in 2008. Barmak et al. (2011), Barmak, Dorso e Otero (2016) investigated the mobility patterns of human populations for dengue epidemic prediction.

An important limitation of the existing works is that they provide little or no explanations for the predictions, especially in the case of more complex models, as noted by Stanaway et al. (2016). Another challenge is how to compare different methods since they use distinct metrics, and different datasets, often unavailable or hard to obtain.

Addressing these limitations, Aleixo et al. (2022) utilized a so-called explainable boosted decision tree model - CatBoost (Prokhorenkova et al. (2018)) for dengue outbreak detection in Rio de Janeiro, Brazil. This approach allows practitioners to comprehend how the model uses available information for predictions. Importantly, not only they made all data and code publicly available, but also evaluated the model using multiple error metrics for regression and classification, and a detailed analysis per district, month of the year, and prediction span.

Aligned with the goal of providing an explainable and comparable model for dengue outbreak detection, the following section details both the results of applying the proposed modified VLMCX to predict dengue outbreaks and the comparison of these results with those presented by Aleixo et al. (2022).

4.5 Applying the VLMC with time-varying and time-invariant exogenous covariates for dengue outbreak prediction

We applied the proposed modified VLMCX model in two distinct scenarios:

1. **National Analysis - Brazil (January 2008 to July 2023):**
   - **Dataset:** Monthly dengue cases across multiple municipalities in Brazil.
   - **Covariates:** Time-dependent climate factors and time-invariant socioeconomic and demographic attributes.
   - **Objective:** Investigate the influence of previous dengue rates, weather conditions, and socioeconomic factors on subsequent dengue rates across diverse municipalities, providing insights into dengue transmission dynamics.
   - **Limitations:** The current model does not incorporate a spatial-dependent structure. However, considering that mosquitoes can fly over a range of 100-500 meters (Siriyasatien et al. (2018)), municipalities are assumed to be independent.
2. **City-Level Analysis - Rio de Janeiro, Brazil (January 2012 to September 2020):**
   - **Dataset:** Historical data series of monthly dengue cases for each district in the city of Rio de Janeiro, Brazil. The dataset utilized in this analysis is the same as the one used by them [12].

[12] https://gitlab.com/interscity/health/dengue-prediction

- **Covariates:** Time-dependent variables like temperature, precipitation, air humidity, the number of Chikungunya and Zika cases and the *Aedes aegypti* infestation index. Time-invariant covariates include fixed attributes of districts, such as the sum of dengue cases in neighboring districts, demographic density, and the number of health facilities.
- **Objective:** Compare results with those presented in Aleixo et al. (2022).
- **Limitations:** The current model does not incorporate a spatial-dependent structure. To address this limitation, the sum of dengue cases in neighboring districts in the past months is incorporated as an exogenous covariate in the model.

4.5.1 Scenario 1: National Analysis - Brazil (January 2008 to July 2023)

To explore the analysis of dengue transmission across multiple municipalities in Brazil, we collected the following variables for each municipality, based on the early literature review of main factors correlated with the number of dengue cases:

- **Dengue Incidence:**
  - **Dengue Cases:** Monthly dengue notifications recorded in the Notifiable Diseases Information System (Sistema de Informação de Agravos de Notificação - SINAN) [13] from January 2008 to July 2023.
- **Climate factors:**
  - **Temperature:** Monthly average temperature (ºC).
  - **Rainfall:** Monthly rainfall (mm).
  - **Days of Rainfall:** Monthly total days of rainfall.

    Climate factors were collected from National Institute of Meteorology (Instituto Nacional de Meteorologia - INMET) [14] from automatic weather stations from January 2008 to July 2023.
- **Socioeconomic factors:**
  - **Low-Income Population:** Percentage of the population in a low-income situation (< 1/2 minimum wage).
  - **Gross Domestic Product (GDP) per Capita:** GDP per capita.

[13] https://datasus.saude.gov.br/informacoes-de-saude-tabnet/

[14] https://portal.inmet.gov.br/dadoshistoricos

  - **Population with Sewage System:** Percentage of the population with access to a sewage system. Socioeconomic factors were collected from national demographic census from 2010 [15]

- **Demographic factors:**
  - **Population in Urban Area:** Percentage of the population living in urban areas according to the national demographic census from 2010 [15].
  - **Municipality Area:** Area of the municipality (km^2) according to IBGE [16].
  - **Total Population:** Total population of the municipality according to the national demographic census from 2010 [15] and 2022 [17].

First, we considered only municipalities with complete information on dengue incidence from 2008 to 2023, totaling 1179 municipalities. Out of these, only 280 had identified automatic weather stations and were considered in the sample.

For these 280 automatic weather stations, we applied a moving average approach for each year and station to address missing values in temperature and rainfall data. This was done exclusively for automatic weather stations with at least seven months of complete information in the corresponding year. Additionally, to obtain complete historical information for each municipality, years with more than four months without information (incomplete years) were managed by considering weather information either before the minimum incomplete year or after the maximum incomplete year, depending on which situation would yield more data for that municipality. At the end of this approach, we had 237 municipalities. Finally, we included only municipalities with at least 36 months of weather information, totaling 126 municipalities from all regions of Brazil.

Based on the descriptive analysis (Appendix B) and literature review, we decided to include the following exogenous covariates in the model fitting:

- **Time-varying exogenous covariates:** Average monthly temperature and monthly days of precipitation.
- **Time-invariant exogenous covariates:** Percentage of population in low-income situation (2010) divided by 10 and percentage of population living in urban areas (2010) divided by 10 and [18].

[15] https://sidra.ibge.gov.br/pesquisa/censo-demografico/demografico-2010/inicial

[16] https://www.ibge.gov.br/geociencias/organizacao-do-territorio/estrutura-territorial/15761-areas-dos-municipios.html

[17] https://sidra.ibge.gov.br/pesquisa/censo-demografico/demografico-2022/inicial

[18] We divided these covariates by 10 to standardize their scale and ensure that the exogenous variables are more comparable.

The choice of using days of precipitation instead of total precipitation stems from the fact that mosquitoes require stagnant water to complete their life cycle (Buczak et al. (2012)). Therefore, it is not only about the volume of rain but also the constancy of rainfall over time, allowing for the presence of standing water after warm periods. Additionally, the decision to use the percentage of the population in a low-income situation instead of the percentage of the population with a sewage system or GDP (R$) per capita is based on the belief that it provides a better representation of municipal poverty and social inequalities. GDP (R$) per capita can be influenced by individuals with incomes well above the average, while having or lacking a sewage system is not always indicative of high or low-income situations in contemporary settings. Moreover, the percentage of the population in a low-income situation shows stronger correlation with the other two social factors (Figure 10).

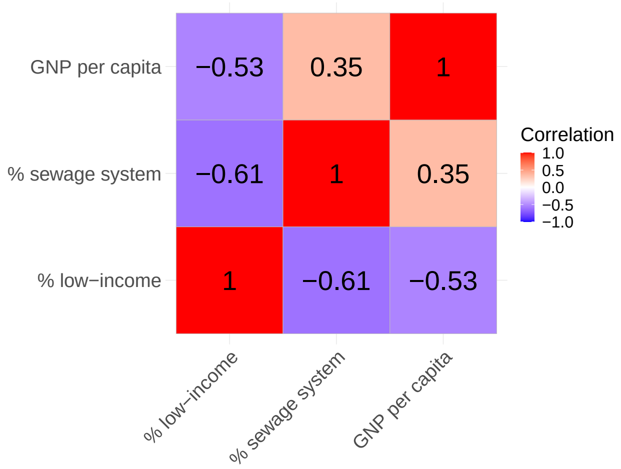


Figure 10 – Correlation matrix between socioeconomic factors

The monthly dengue cases were categorized based on their distribution (Table 35) and on the Ministry of Health's classification of dengue incidence (cases per 100,000 inhabitants):

- **Category 1** - Up to 5 cases per 100,000 inhabitants,
- **Category 2** - 5 to 25 cases per 100,000 inhabitants,
- **Category 3** - 25 to 75 cases per 100,000 inhabitants, and
- **Category 4** - Over 75 cases per 100,000 inhabitants.

To prepare the data for modeling, we excluded the most recent 12 months of data from each municipality, reserving them for prediction. For this fitting, we chose to employ multinomial logistic regression for parameter estimation in each context. Although

Table 35 – Distribution of dengue cases for all the 126 municipalities

Situation	Minimun	1^{st} Quartile	Median	Mean	3^{st} Quartile	Maximun
Epidemic	0	3.86	17.57	98.31	71.84	4478.30
Non-epidemic	0	0	0.89	2.93	3.53	56.76

we also experimented with ordinal logistic regression and found similar forecasting results, we opted to present the results for multinomial logistic regression fitting, as we believe it provides clearer estimations.

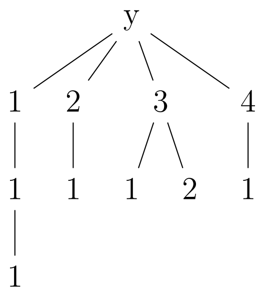
For tuning parameters (δ and f), we selected values that minimized the Bayesian Information Criterion (BIC), resulting in $\delta = 0.000001$ and $f = 4$. Additionally, we constrained the initial tree to have a maximum depth of 6, with Category 1 assumed as the baseline.

The estimated beta-context is depicted in Figure 11, where $\boldsymbol{\alpha}^u = (\alpha_2^u, \alpha_3^u, \alpha_4^u)$. Contexts 1* (1* means the context is 1 preceded by other state that is not 1), 11, and 41 lacked sufficient observations to estimate exogenous covariate parameters. Across all contexts, only time-varying exogenous covariates of the last month demonstrated a significant effect, with estimates for previous months being equal to zero.

The estimated context tree provides valuable insights into the trend of dengue incidence. According to the tree, when there are less than 5 cases per 100,000 inhabitants in the previous month, the next month's dengue incidence appears to be independent of the rest of the history, except when there are also fewer than 5 cases per 100,000 inhabitants in the history. In other words, when examining the preceding month, if there are fewer than 5 cases per 100,000 inhabitants, it is necessary to consider an additional previous month. If there are still fewer than 5 cases per 100,000 inhabitants, it becomes necessary to look back three steps in the past. Conversely, if there are more than 5 cases per 100,000 inhabitants, there is no need to look further back than one month ago. This observation suggests that having less than 5 cases per 100,000 inhabitants generally indicates a decreasing trend, and the specific events preceding that period may not significantly impact the subsequent month's incidence, except when there is a continued low incidence.

Similarly, when there are 25 to 75 cases per 100,000 inhabitants in the previous months, dengue incidence is independent of the remaining history unless there were less than 25 cases per 100,000 inhabitants in the month before last. This pattern could be explained by the likelihood that, when there are already 25 to 75 cases per 100,000 inhabitants in the previous months, the probability is higher that cases are decreasing. However, if there were fewer than 25 cases per 100,000 inhabitants, there might be a chance of the cases increasing.

Regarding time-varying exogenous covariates, it can be observed that, in general, increases in temperature and rainfall correspond to a higher probability of experiencing



$$\begin{aligned}
\hat{\boldsymbol{\alpha}}^{1*} &= (-1.27, -2.8, -4.45) & \hat{\boldsymbol{\alpha}}^{3*} &= (-6.65, -11.72, -13.43)\\
\hat{\boldsymbol{\alpha}} &= (-1.66, -3.66, -5.85) & \hat{\boldsymbol{\alpha}}^{31} &= (1.64, -8.39, 3.57)\\
\hat{\boldsymbol{\alpha}}^{111} &= (-6.51, -7.27, -11.88) & \hat{\boldsymbol{\alpha}}^{32} &= (2.21, -1.18, 1.87)\\
\hat{\boldsymbol{\alpha}}^{2*} &= (-6.68, -6.76, -8.43) & \hat{\boldsymbol{\alpha}}^{4*} &= (-3.19, -5.84, -9.68)\\
\hat{\boldsymbol{\alpha}}^{21} &= (-8.1, -8.65, -9.9) & \hat{\boldsymbol{\alpha}}^{41} &= (0, 0.09, 2.6)
\end{aligned}$$

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^{111}_{2,1} &= (0.2, 0.06)' & \hat{\boldsymbol{\beta}}^{111}_{3,1} &= (0.28, 0.07)' & \hat{\boldsymbol{\beta}}^{111}_{4,1} &= (0.37, 0.06)'\\
\hat{\boldsymbol{\beta}}^{2*}_{2,1} &= (0.21, 0.04)' & \hat{\boldsymbol{\beta}}^{2*}_{3,1} &= (0.35, 0.09)' & \hat{\boldsymbol{\beta}}^{2*}_{4,1} &= (0.4, 0.1)'\\
\hat{\boldsymbol{\beta}}^{21}_{2,1} &= (0.2, 0.07)' & \hat{\boldsymbol{\beta}}^{21}_{3,1} &= (0.2, 0.1)' & \hat{\boldsymbol{\beta}}^{21}_{4,1} &= (0.28, 0.1)'\\
\hat{\boldsymbol{\beta}}^{3*}_{2,1} &= (0.11, 0.03)' & \hat{\boldsymbol{\beta}}^{3*}_{3,1} &= (0.36, 0.08)' & \hat{\boldsymbol{\beta}}^{3*}_{4,1} &= (0.5, 0.13)'\\
\hat{\boldsymbol{\beta}}^{4*}_{2,1} &= (-0.10, -0.04)' & \hat{\boldsymbol{\beta}}^{4*}_{3,1} &= (0.07, -0.04)' & \hat{\boldsymbol{\beta}}^{4*}_{4,1} &= (0.31, 0.04)'
\end{aligned}$$

$$\begin{aligned}
\hat{\boldsymbol{\gamma}}^{111}_{2} &= (-0.01, -0.01)' & \hat{\boldsymbol{\gamma}}^{111}_{3} &= (-0.03, -0.04)' & \hat{\boldsymbol{\gamma}}^{111}_{4} &= (-0.02, -0.03)'\\
\hat{\boldsymbol{\gamma}}^{2*}_{2} &= (-0.01, 0.03)' & \hat{\boldsymbol{\gamma}}^{2*}_{3} &= (-0.05, -0.02)' & \hat{\boldsymbol{\gamma}}^{2*}_{4} &= (-0.06, -0.03)'\\
\hat{\boldsymbol{\gamma}}^{21}_{2} &= (0, 0.03)' & \hat{\boldsymbol{\gamma}}^{21}_{3} &= (-0.01, 0.02)' & \hat{\boldsymbol{\gamma}}^{21}_{4} &= (-0.01, 0.01)'\\
\hat{\boldsymbol{\gamma}}^{3*}_{2} &= (0.01, 0.06)' & \hat{\boldsymbol{\gamma}}^{3*}_{3} &= (-0.02, 0.05)' & \hat{\boldsymbol{\gamma}}^{3*}_{4} &= (-0.04, 0.03)'\\
\hat{\boldsymbol{\gamma}}^{31}_{2} &= (-0.02, -0.01)' & \hat{\boldsymbol{\gamma}}^{31}_{3} &= (0.05, 0.09)' & \hat{\boldsymbol{\gamma}}^{31}_{4} &= (-0.06, -0.01)'\\
\hat{\boldsymbol{\gamma}}^{32}_{2} &= (-0.03, 0)' & \hat{\boldsymbol{\gamma}}^{32}_{3} &= (-0.02, 0.05)' & \hat{\boldsymbol{\gamma}}^{32}_{4} &= (-0.04, 0.02)'\\
\hat{\boldsymbol{\gamma}}^{4*}_{2} &= (0.04, 0.08)' & \hat{\boldsymbol{\gamma}}^{4*}_{3} &= (0.03, 0.08)' & \hat{\boldsymbol{\gamma}}^{4*}_{4} &= (0, 0.08)'
\end{aligned}$$

Figure 11 – Final tree and estimated parameters for dengue incidence in Brazilian municipalities

higher dengue cases. However, concerning time-invariant exogenous covariates, the estimates are notably low, suggesting that there may be no significant impact on dengue incidence. This observation aligns with the discussion by Horstick, Tozan e Wilder-Smith (2015), who argue that although dengue predominantly affects resource-limited countries, it does not exclusively target the poor.

All available observations from previous months were used to forecast each of the 12 months, as the model relies on covariates from previous months. Figure 12 and Figure 13 display the confusion matrices of forecast results, both in total and broken down by the month of the year.

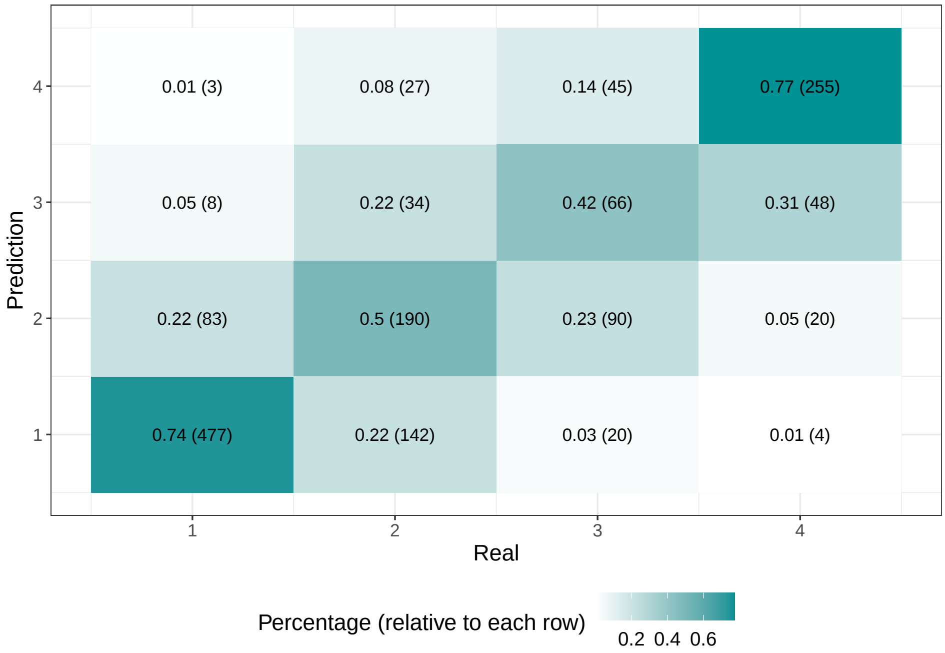


Figure 12 – Confusion matrix for prediction results (using all available observations from previous months) - Brazil municipalities

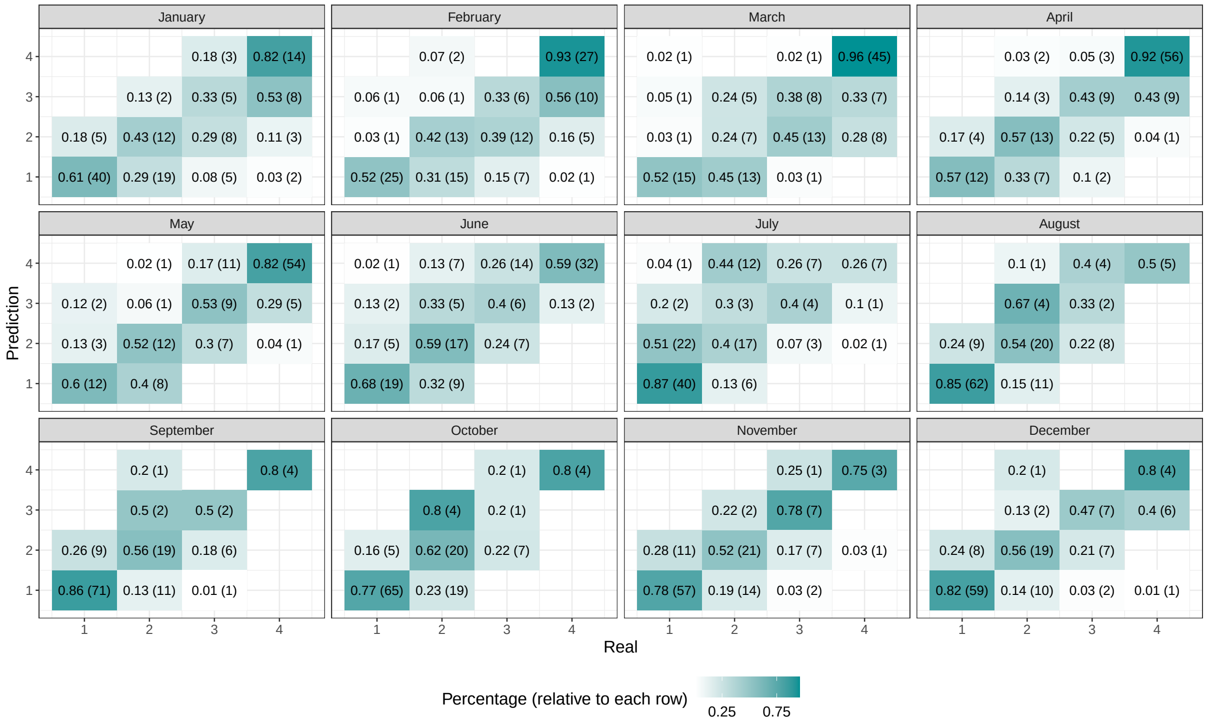


Figure 13 – Confusion matrix for prediction results by month (using all available observations from previous months) - Brazil municipalities

The model demonstrated satisfactory performance in predicting dengue incidence above 75 monthly cases/100,000 inhabitants, accurately identifying high incidence in 77% of the cases. It also showed the ability to recognize situations where the actual incidence may be lower than predicted, with 69% of cases falling within or below the predicted range of 25 to 75 cases/100,000 inhabitants. However, caution is advised when dengue incidence may be higher than initially predicted (31% for Category 3, 28% for Category 2 and 26% for Category 1).

Recognizing that dengue incidence data may not always be available for the previous month due to the time it takes for consolidation and availability, we also explored predicting using previously predicted dengue incidence instead of real values. In this approach, for the second month of prediction, we utilized the predicted values from the first month and continued this process consecutively. Therefore, to predict 12 months ahead, we used dengue incidence predictions for the previous 11 months instead of the actual dengue incidence in those months. Notice that we compared predictions using k, $k = 1, \ldots, 11$, months of predicted values and true values for further past. Figure 14 and Figure 15 present the confusion matrices of forecast results, both in total and broken down by the time ahead forecasted. As municipalities may have a different number of months, the time ahead may not always correspond to the same month of the year in each municipality. It is important to note that for time-varying exogenous covariates, we utilized only real values.

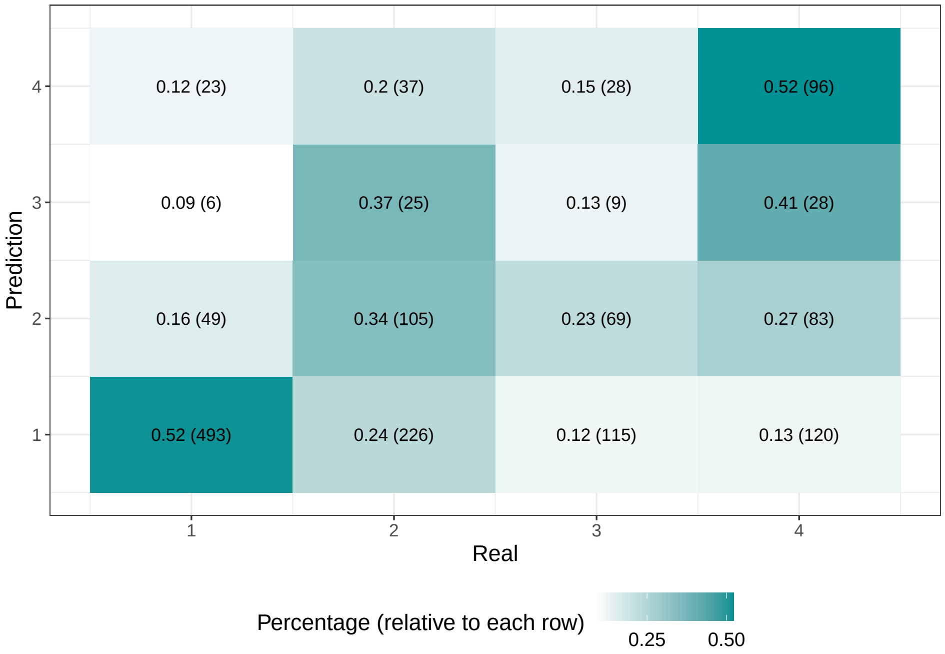


Figure 14 – Confusion matrix for prediction results (using previously predicted dengue incidence instead of real values) - Brazil municipalities

The model exhibited satisfactory performance in predicting dengue incidence up

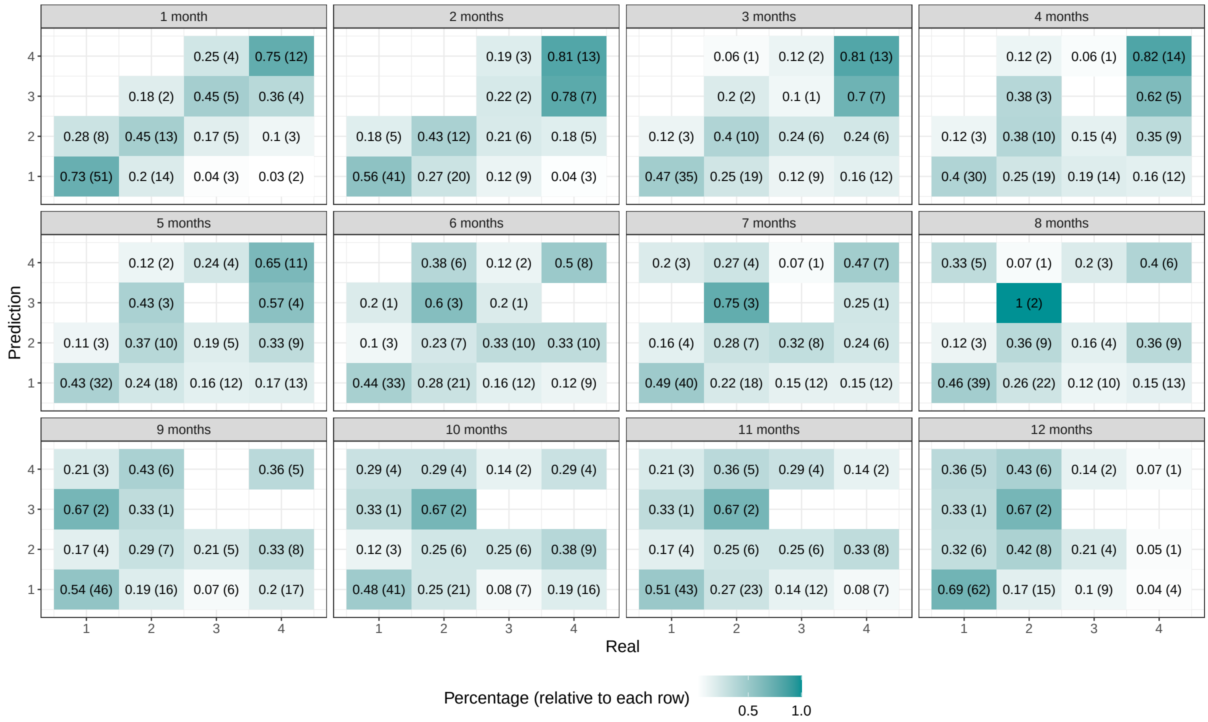


Figure 15 – Confusion matrix for prediction results by time ahead (using previously predicted dengue incidence instead of real values) - Brazil municipalities

to 2 months ahead but started to deviate more for predictions beyond 2 months, especially for intermediary categories. However, the model demonstrated satisfactory performance in predicting dengue incidence above 75 monthly cases per 100,000 inhabitants, even for more than 4 months ahead.

4.5.2 Scenario 2: City-Level Analysis - Rio de Janeiro, Brazil (January 2011 to September 2020)

For this study, we utilized the dataset provided by Aleixo et al. (2022), which is publicly accessible on GitLab [19]. Table 36 shows the full list of features.

As the purpose of this scenario is to compare results, variable selection was not required, as the same set of variables was used. Therefore, a descriptive analysis of the variables was not necessary.

Additionally, since there was no need to specify how far back to look, we included the number of dengue, Zika and Chikungunya cases, *Aedes aegypti* infestation index, total precipitation (mm), mean temperature (ºC) and mean air humidity (%) for all previous months. This approach allowed the model to automatically determine the relevant time lag for each variable.

[19] https://gitlab.com/interscity/health/dengue-prediction

Table 36 – Features used as input to the model - Rio de Janeiro, Brazil (January 2012 to October 2020)

Feature	Description
cases-n	Past number of dengue cases, where n = 1, 2, 3 is the number of months in the past (per district)
dengue_prevalence	Sum of dengue cases in the past, normalized to the 0 to 1 range (per district)
neighbor_cases	Sum of dengue cases in neighboring districts (per district)
precipitation (mm)	Total precipitation in the last month (per district)
temperature (ºC)	Mean temperature in the last month (per district)
air_humidity (%)	Mean air humidity in the last month (per district)
liraa	*Aedes aegypti* infestation index (for the city)
chikungunya	Number of chikungunya cases last month (per district)
zika	Number of zika cases last month (per district)
demographic density	Demographic density (per district)
num_health_unit	Number of health facilities (per district)

Source: Aleixo et al. (2022)

So, in summary, for each district, the modeling setup is as follows:

- **Response variable:** Number of dengue cases, categorized following Aleixo et al. (2022):
  - **Severe outbreak:** when the number of cases in a district in a given month is above 99% of all measurements in the training set.
  - **Mild outbreak:** when the number of cases in a district in a given month is above 95% of all measurements in the training set.
  - **No oubreak:** when the number of cases in a district in a given month is below or equal 95% of all measurements in the training set.
- **Time-varying exogenous covariates:** Total precipitation (mm), mean temperature (ºC), mean air humidity (%), sum of dengue cases in neighboring districts, number of Zika cases, number of Chikungunya cases, and *Aedes aegypti* infestation index.
- **Time-invariant exogenous covariates:** Demographic density, number of health facilities and sum of dengue cases in the past (normalized to the 0 to 1 range).

In their study, Aleixo et al. (2022) employed a boosted tree regression method (CatBoost) and compared its results with a Seasonal Autoregressive Integrated Moving Average (SARIMA) model, using individual time series for each district. For CatBoost, they conducted grid search to define tuning parameters, using 2015 as validation data and 2012 to 2014 as training data. In the case of SARIMA, tuning parameters were defined fitting the model to data from 2012 to 2015.

Regarding predictions, individual predictions were made for each district and month in the period 2016 to 2020. For CatBoost, they established a distinct model for each year, employing a 5-fold cross-validation approach. This involved using a single year as the test set and four years as the training set. For instance, when predicting the number of cases in 2017, the training data consisted of the years 2016, 2018, 2019, and 2020. In the case of the SARIMA model, they also created separate models for each year (2016 to 2020), using the previous four years as training data, as SARIMA requires a contiguous time series to extract trend and seasonality features.

To ensure comparability with Aleixo et al. (2022), we also created a separate model for each year (2016 to 2020). However, due to the requirement of a contiguous time series for Markov Chains, we used the previous four years as training data. For example, to predict the number of cases in 2016, we utilized data from 2015, 2014, 2013, and 2012 as training data. Additionally, Chikungunya and Zika cases, as well as the *Aedes aegypti* infestation index, were not used for the models predicting cases in 2016, 2017, and 2018, as these covariates lacked values before 2015. The tuning parameter δ was selected to minimize BIC in two training sets: 2012-2015 for models used to predict years 2016, 2017, and 2019, and 2015-2018 for models used to predict years 2019 and 2020, resulting in $\delta = 0.00001$ for both. The tuning parameter f was fixed at $f = 2$ to allow for the estimation of more parameters in long contexts.

Figure 16 display prediction results for our model and the ones fitted in Aleixo et al. (2022). The predictions were made one month ahead, and all available observations from previous months were used to forecast each of the 12 months, as the model relies on covariates from previous months. Our model shows better prediction results for extreme categories, but it performs less well for intermediary categories. In terms of the worst-case scenario where the model predicts no outbreak while there is an outbreak, our model and SARIMA have similar results for this situation in the mild category. Specifically, our model predicted no outbreak for 52% of the real values equal to mild, while this percentage is 57% for the SARIMA model.

The current categorization of dengue cases may be influencing predictions, as it does not take into account the population size of each district and changes for each year being predicted. Additionally, since the categorization is based on the training set used for Catboost fitting, future years are being utilized to categorize cases, potentially introducing biases into the predictions.

Figure 17, Figure 18, Figure 19, Figure 20, and Figure 21 present the estimated beta-contexts for each year. It is noticeable that the covariates with the most influence on future dengue cases are mean temperature (ºC), mean air humidity (%), total precipitation (mm), *Aedes aegypti* infestation index, and the sum of dengue cases in the past (normalized to the 0 to 1 range). Additionally, in general, only two previous months are needed to predict the next one. Larger contexts occur when there is a history of months with low incidence (category 1). These findings align with those presented in Aleixo et al. (2022),

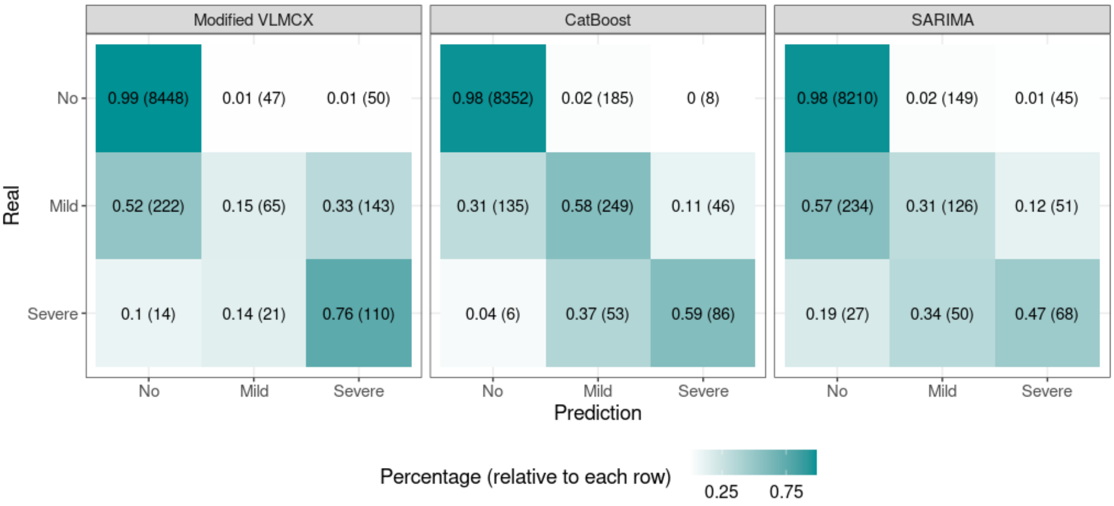
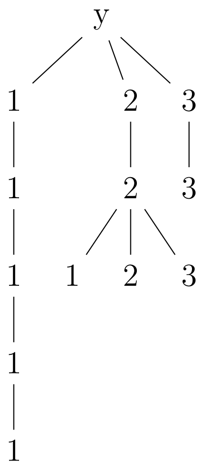


Figure 16 – Confusion matrix for prediction results - Rio de Janeiro districts

where the most relevant features for the CatBoost model were the number of cases in the previous month (cases $m - 1$), followed by the dengue prevalence of the district, precipitation, neighbor cases, and temperature.

One advantage of our model over CatBoost (presented by Aleixo et al. (2022)) is that, while CatBoost allows the evaluation of the importance of variables in model predictions, our model provides the exact values of parameter estimates. This enables us to precisely understand how each covariate impacts dengue transmission in various scenarios of past dengue incidence. Furthermore, our model allows flexibility in examining the variable length of the past, providing precise insights into how past events influence future dengue transmission.



$$\hat{\boldsymbol{\alpha}}^{1*} = (-3.08, -5.22)$$
$$\hat{\boldsymbol{\alpha}}^{11*} = (-3.94, -37.32)$$
$$\hat{\boldsymbol{\alpha}}^{111*} = (-3.78, -36.21)$$
$$\hat{\boldsymbol{\alpha}}^{1111*} = (-3.72, -5.8)$$
$$\hat{\boldsymbol{\alpha}}^{11111} = (-3.37, -5.64)$$

$$\hat{\boldsymbol{\alpha}}^{2*} = (-18.11, -59.74)$$
$$\hat{\boldsymbol{\alpha}}^{221} = (-106.82, -164.7)$$
$$\hat{\boldsymbol{\alpha}}^{222} = (-1.9, -2.84)$$
$$\hat{\boldsymbol{\alpha}}^{223} = (-1.73, -11.42)$$
$$\hat{\boldsymbol{\alpha}}^{3*} = (2.56, 4.65)$$
$$\hat{\boldsymbol{\alpha}}^{33*} = (-9.52, -52.91)$$

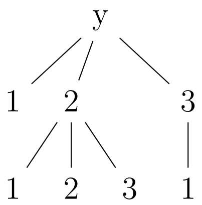
$$\hat{\boldsymbol{\beta}}_{2,1}^{2*} = (0.11, 0.67, 0.01, 0)'$$
$$\hat{\boldsymbol{\beta}}_{2,1}^{221} = (3.14, 9.62, 5.27, 0)'$$
$$\hat{\boldsymbol{\beta}}_{2,2}^{221} = (2.74, -7.98, -0.71, -0.03)'$$
$$\hat{\boldsymbol{\beta}}_{2,3}^{221} = (2.03, -6.7, -2.22, 0.06)'$$
$$\hat{\boldsymbol{\beta}}_{2,1}^{33*} = (-0.21, -0.48, 0.01, 0)'$$
$$\hat{\boldsymbol{\beta}}_{2,2}^{33*} = (0.03, 0.38, 0.17, 0)'$$

$$\hat{\boldsymbol{\beta}}_{3,1}^{2*} = (0.11, 1.78, 0.18, 0)'$$
$$\hat{\boldsymbol{\beta}}_{3,1}^{221} = (2.91, 10.2, 4.61, 0)'$$
$$\hat{\boldsymbol{\beta}}_{3,2}^{221} = (2.74, -6.81, -0.44, -0.02)'$$
$$\hat{\boldsymbol{\beta}}221_{3,3} = (2.02, -6.57, -1.65, 0.03)'$$
$$\hat{\boldsymbol{\beta}}_{3,1}^{33*} = (-0.11, 0.37, 0.2, 0)'$$
$$\hat{\boldsymbol{\beta}}_{3,2}^{33*} = (-0.03, 1.08, 0.04, 0)'$$

$$\hat{\boldsymbol{\gamma}}_{2}^{11111} = (0.01, 0, 3.37)'$$
$$\hat{\boldsymbol{\gamma}}_{2}^{2*} = (0.01, -0.01, 4.4)'$$
$$\hat{\boldsymbol{\gamma}}_{2}^{221} = (-0.01, 0.03, 145.09)'$$
$$\hat{\boldsymbol{\gamma}}_{2}^{222} = (0.05, 0.01, 2.74)'$$
$$\hat{\boldsymbol{\gamma}}_{2}^{33*} = (0, 0, 0.05)'$$

$$\hat{\boldsymbol{\gamma}}_{3}^{11111} = (-0.04, 0, 5.42)'$$
$$\hat{\boldsymbol{\gamma}}_{3}^{2*} = (0.02, 0.01, 8.22)'$$
$$\hat{\boldsymbol{\gamma}}_{3}^{221} = (0.02, 0.06, 138.42)'$$
$$\hat{\boldsymbol{\gamma}}_{3}^{222} = (0.07, 0.02, 1.46)'$$
$$\hat{\boldsymbol{\gamma}}_{3}^{33*} = (0.03, 0, 7.12)'$$

Figure 17 – Final tree and estimated parameters for dengue incidence in Rio de Janeiro districts (fitted for years 2012-2015)

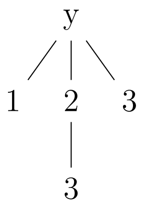


$$\hat{\boldsymbol{\alpha}}^{1*} = (-36.31, -49.46) \qquad \hat{\boldsymbol{\alpha}}^{23} = (-2.51, -12.08)$$
$$\hat{\boldsymbol{\alpha}}^{21} = (3.48, 8.81) \qquad \hat{\boldsymbol{\alpha}}^{3*} = (-5.83, -38.49)$$
$$\hat{\boldsymbol{\alpha}}^{22} = (-12.59, -91.4) \qquad \hat{\boldsymbol{\alpha}}^{31} = (6.92, 8.94)$$

$$\hat{\boldsymbol{\beta}}^{1*}_{2,1} = (0.03, 0.76, 0.17, 0.01)' \qquad \hat{\boldsymbol{\beta}}^{1*}_{3,1} = (0, 0.9, 0.26, 0.01)'$$
$$\hat{\boldsymbol{\beta}}^{21}_{2,1} = (-0.05, 1.44, 0.68, 0.01)' \qquad \hat{\boldsymbol{\beta}}^{21}_{3,1} = (-0.13, 2.44, 0.89, 0.01)'$$
$$\hat{\boldsymbol{\beta}}^{21}_{2,2} = (-0.01, -1.64, -0.63, -0.02)' \qquad \hat{\boldsymbol{\beta}}^{21}_{3,2} = (-0.04, -2.68, -0.89, -0.02)'$$
$$\hat{\boldsymbol{\beta}}^{22}_{2,1} = (0.14, 0.35, 0.02, 0)' \qquad \hat{\boldsymbol{\beta}}^{22}_{3,1} = (0.05, 2.39, 0.37, 0)'$$
$$\hat{\boldsymbol{\beta}}3*_{2,1} = (0.01, 0.09, 0.04, 0)' \qquad \hat{\boldsymbol{\beta}}^{3*}_{3,1} = (0.14, 0.89, 0.18, 0)'$$

$$\hat{\boldsymbol{\gamma}}^{1*}_{2} = (0.02, 0, 4.46)' \qquad \hat{\boldsymbol{\gamma}}^{1*}_{3} = (0.02, -0.02, 7.57)'$$
$$\hat{\boldsymbol{\gamma}}^{21}_{2} = (0.01, 0, 2.59)' \qquad \hat{\boldsymbol{\gamma}}^{21}_{3} = (0.03, 0, 3.76)'$$
$$\hat{\boldsymbol{\gamma}}^{22}_{2} = (0, 0, 4.93)' \qquad \hat{\boldsymbol{\gamma}}^{22}_{3} = (0.02, -0.03, 5.76)'$$
$$\hat{\boldsymbol{\gamma}}^{3*}_{2} = (0.04, -0.01, 2.54)' \qquad \hat{\boldsymbol{\gamma}}^{3*}_{3} = (0.05, 0, 6.72)'$$

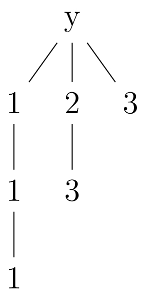
Figure 18 – Final tree and estimated parameters for dengue incidence in Rio de Janeiro districts (fitted for years 2013-2016)



$$\hat{\boldsymbol{\alpha}}^{1} = (-3.92, -7.23)$$
$$\hat{\boldsymbol{\alpha}}^{2*} = (-19.78, -32.16)$$
$$\hat{\boldsymbol{\alpha}}^{23} = (-2.04, -11.29)$$
$$\hat{\boldsymbol{\alpha}}^{3} = (1.87, 1.98)$$

$$\hat{\boldsymbol{\beta}}^{2*}_{2,1} = (0.07, 0.23, 0.18, 0)' \qquad \hat{\boldsymbol{\beta}}^{2*}_{3,1} = (0.18, 0.36, 0.24, 0)'$$
$$\hat{\boldsymbol{\gamma}}^{2*}_{2} = (0.01, 0, 2.27)' \qquad \hat{\boldsymbol{\gamma}}^{2*}_{3} = (0.03, -0.03, 6.08)'$$

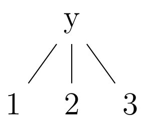
Figure 19 – Final tree and estimated parameters for dengue incidence in Rio de Janeiro districts (fitted for years 2014-2017)



$$\hat{\boldsymbol{\alpha}}^{1*} = (-2.92, -5.23)$$
$$\hat{\boldsymbol{\alpha}}^{11*} = (-3.4, -17.44)$$
$$\hat{\boldsymbol{\alpha}}^{111} = (-4, -6.92)$$
$$\hat{\boldsymbol{\alpha}}^{2*} = (-19.94, -53.96)$$
$$\hat{\boldsymbol{\alpha}}^{23} = (-0.81, -13.02)$$
$$\hat{\boldsymbol{\alpha}}^{3} = (0.95, 0.57)$$

$$\hat{\boldsymbol{\beta}}^{2*}_{2,1} = (0.03, 0.26, 0.12, 0, 0, -0.01, 4.56)' \qquad \hat{\boldsymbol{\beta}}^{2*}_{3,1} = (0.12, 0.67, 0.3, 0, 0, 0.01, 9.58)'$$
$$\hat{\boldsymbol{\gamma}}^{2*}_{2} = (0.01, 0, 2.56)' \qquad \hat{\boldsymbol{\gamma}}^{2*}_{3} = (0.01, -0.03, 6.81)'$$
$$\hat{\boldsymbol{\gamma}}^{3}_{2} = (0.02, 0.05, -1.64)' \qquad \hat{\boldsymbol{\gamma}}^{3}_{3} = (0.02, 0.04, 1.03)'$$

Figure 20 – Final tree and estimated parameters for dengue incidence in Rio de Janeiro districts (fitted for years 2015-2018)



$$\hat{\boldsymbol{\alpha}}^{1} = (-3.92, -7.44)$$
$$\hat{\boldsymbol{\alpha}}^{2} = (-39.13, -90.22)$$
$$\hat{\boldsymbol{\alpha}}^{3} = (1.73, 1.76)$$

$$\hat{\boldsymbol{\beta}}^{2}_{2,1} = (0.08, 0.6, 0.34, 0, 0, 0, -4.04)' \qquad \hat{\boldsymbol{\beta}}^{2}_{3,1} = (0.17, 1.64, 0.55, 0.01, 0, 0.01, -1.59)'$$
$$\hat{\boldsymbol{\gamma}}^{2}_{2} = (0.02, -0.01, 3.64)' \qquad \hat{\boldsymbol{\gamma}}^{2}_{3} = (0.05, -0.05, 10.35)'$$

Figure 21 – Final tree and estimated parameters for dengue incidence in Rio de Janeiro districts (fitted for years 2016-2019)

5 Conclusion and future work

This study has successfully expanded the capabilities of the beta-context algorithm, integrating both time-dependent and time-invariant exogenous covariates from multiple independent sources and addressing challenges related to rare events and limited data, providing a more versatile modeling framework. Our approach assumed identical parameter estimates for all sources.

The simulations conducted without time-invariant exogenous covariates and with only one source demonstrated that the modified algorithm outperforms the original version, particularly in situations with limited data. This was evident in the recovery of more accurate tree structures and covariate vectors. For simulations with time-invariant exogenous covariates and multiple independent sources, improvements were observed in the percentage of identical trees and exact covariate vectors with increasing sample size.

Our motivation comes from analyzing a real dataset focused on monthly dengue cases across multiple municipalities in Brazil. The inclusion of time-dependent covariates - temperature and precipitation levels, along with time-invariant covariates - poverty rates and urban population percentages, allowed us to investigate the complex dynamics of dengue transmission. The model demonstrated satisfactory performance in predicting dengue incidence, especially for high-incidence cases, showing its potential for early detection and proactive management of outbreaks. This is particularly significant given the alarming global threat of dengue fever, with an estimated 390 million infections annually, as recognized by the World Health Organization.

Looking forward, our future work includes incorporating additional model possibilities, including non-parametric methods and spatial correlation to account for non-independent multiple sources. Besides that, we aim to develop mechanisms to: handle missing values in exogenous covariates, address the challenge of impossible contexts, test the significance of time-invariant exogenous covariates and enable estimation of exogenous covariates parameters when not all categories/states for a specific context are available. In relation to dengue outbreaks prediction, we aim to conduct a more detailed study on how covariates influence dengue cases, exploring non-linear effects and interactions between factors. The overarching objective is to continually enhance the model's flexibility, accuracy, and applicability across diverse settings. Additionally, plans involve the publication of the algorithm implementation as a package on The Comprehensive R Archive Network (CRAN), further contributing to the broader scientific community.

Bibliography

AALEN, O. O.; BORGAN, Ø.; FEKJÆR, H. *Covariate adjustment of event histories estimated from Markov chains: the additive approach.* [S.l.]: Wiley Online Library, 2001. 993–1001 p. Citado na página 17.

AGRESTI, A. *Categorical data analysis.* [S.l.]: John Wiley & Sons, 2012. v. 792. Citado na página 39.

ALEIXO, R.; KON, F.; ROCHA, R.; CAMARGO, M. S.; CAMARGO, R. Y. D. Predicting dengue outbreaks with explainable machine learning. In: IEEE. *2022 22nd IEEE International Symposium on Cluster, Cloud and Internet Computing (CCGrid).* [S.l.], 2022. p. 940–947. Citado 6 vezes nas páginas 68, 69, 76, 77, 78, and 79.

ARCARI, P.; TAPPER, N.; PFUELLER, S. Regional variability in relationships between climate and dengue/dhf in indonesia. *Singapore Journal of Tropical Geography*, Wiley Online Library, v. 28, n. 3, p. 251–272, 2007. Citado na página 64.

AZZALINI, A. Logistic regression for autocorrelated data with application to repeated measures. *Biometrika*, Oxford University Press, v. 81, n. 4, p. 767–775, 1994. Citado na página 17.

BARMAK, D. H.; DORSO, C. O.; OTERO, M. Modelling dengue epidemic spreading with human mobility. *Physica A: Statistical Mechanics and its Applications*, Elsevier, v. 447, p. 129–140, 2016. Citado na página 67.

BARMAK, D. H.; DORSO, C. O.; OTERO, M.; SOLARI, H. G. Dengue epidemics and human mobility. *Physical Review E*, APS, v. 84, n. 1, p. 011901, 2011. Citado na página 67.

BHATNAGAR, S.; LAL, V.; GUPTA, S. D.; GUPTA, O. P. Forecasting incidence of dengue in rajasthan, using time series analyses. *Indian journal of public health*, Medknow, v. 56, n. 4, p. 281–285, 2012. Citado na página 67.

BOTARI, T.; ALVES, S.; LEONEL, E. D. Explaining the high number of infected people by dengue in rio de janeiro in 2008 using a susceptible-infective-recovered model. *Physical Review E*, APS, v. 83, n. 3, p. 037101, 2011. Citado na página 67.

BRADY, O. J.; GETHING, P. W.; BHATT, S.; MESSINA, J. P.; BROWNSTEIN, J. S.; HOEN, A. G.; MOYES, C. L.; FARLOW, A. W.; SCOTT, T. W.; HAY, S. I. Refining the global spatial limits of dengue virus transmission by evidence-based consensus. Public Library of Science San Francisco, USA, 2012. Citado na página 61.

BRAY, R. L. Markov decision processes with exogenous variables. *Management Science*, INFORMS, v. 65, n. 10, p. 4598–4606, 2019. Citado na página 18.

BROWNING, M.; CARRO, J. M. Heterogeneity in dynamic discrete choice models. *The Econometrics Journal*, Oxford University Press Oxford, UK, v. 13, n. 1, p. 1–39, 2010. Citado na página 18.

BUCZAK, A. L.; BAUGHER, B.; BABIN, S. M.; RAMAC-THOMAS, L. C.; GUVEN, E.; ELBERT, Y.; KOSHUTE, P. T.; VELASCO, J. M. S.; JR, V. G. R.; TAYAG, E. A. et al. Prediction of high incidence of dengue in the philippines. *PLoS neglected tropical diseases*, Public Library of Science San Francisco, USA, v. 8, n. 4, p. e2771, 2014. Citado na página 67.

BUCZAK, A. L.; KOSHUTE, P. T.; BABIN, S. M.; FEIGHNER, B. H.; LEWIS, S. H. A data-driven epidemiological prediction method for dengue outbreaks using local and remote sensing data. *BMC medical informatics and decision making*, BioMed Central, v. 12, n. 1, p. 1–20, 2012. Citado 2 vezes nas páginas 64 and 71.

BÜHLMANN, P. Model selection for variable length markov chains and tuning the context algorithm. *Annals of the Institute of Statistical Mathematics*, Springer, v. 52, p. 287–315, 2000. Citado na página 30.

BÜHLMANN, P.; WYNER, A. J. Variable length markov chains. *The Annals of Statistics*, Institute of Mathematical Statistics, v. 27, n. 2, p. 480–513, 1999. Citado 7 vezes nas páginas 17, 23, 27, 28, 30, 31, and 34.

CHAUHAN, C.; BEHURA, S. K.; DEBRUYN, B.; LOVIN, D. D.; HARKER, B. W.; GOMEZ-MACHORRO, C.; MORI, A.; ROMERO-SEVERSON, J.; SEVERSON, D. W. Comparative expression profiles of midgut genes in dengue virus refractory and susceptible aedes aegypti across critical period for virus infection. Public Library of Science San Francisco, USA, 2012. Citado na página 67.

CHOMPOOSRI, J.; THAVARA, U.; TAWATSIN, A.; ANANTAPREECHA, S.; SIRIYASATIEN, P. Seasonal monitoring of dengue infection in aedes aegypti and serological feature of patients with suspected dengue in 4 central provinces of thailand. *The Thai Journal of Veterinary Medicine*, Faculty of Veterinary Science, Chulalongkorn University, v. 42, n. 2, p. 185–193, 2012. Citado na página 65.

CONTROL, C. for D.; (CDC, P. et al. Dengue hemorrhagic fever–us-mexico border, 2005. *MMWR. Morbidity and mortality weekly report*, v. 56, n. 31, p. 785–789, 2007. Citado na página 65.

COOK, R. J.; NG, E. T. A logistic-bivariate normal model for overdispersed two-state markov processes. *Biometrics*, JSTOR, p. 358–364, 1997. Citado na página 17.

COSTA, A. C. C.; CODEÇO, C. T.; HONÓRIO, N. A.; PEREIRA, G. R.; PINHEIRO, C. F. N.; NOBRE, A. A. Surveillance of dengue vectors using spatio-temporal bayesian modeling. *BMC medical informatics and decision making*, Springer, v. 15, p. 1–12, 2015. Citado na página 65.

FAHRMEIR, L. Asymptotic testing theory for generalized linear models. *Statistics: A Journal of Theoretical and Applied Statistics*, Taylor & Francis, v. 18, n. 1, p. 65–76, 1987. Citado na página 35.

FAHRMEIR, L.; KAUFMANN, H. Regression models for non-stationary categorical time series. *Journal of time series Analysis*, Wiley Online Library, v. 8, n. 2, p. 147–160, 1987. Citado na página 35.

FOKIANOS, K.; KEDEM, B. Regression theory for categorical time series. *Statistical science*, Institute of Mathematical Statistics, v. 18, n. 3, p. 357–376, 2003. Citado na página 17.

FURLAN, G. *Contribution à l'étude et au développement d'algorithmes de traitement du signal en compression de données et d'images.* Tese (Doutorado) — Nice, 1990. Citado na página 19.

GAO, X.; CAO, Y. R.; OGDEN, N.; AUBIN, L.; ZHU, H. P. Mixture markov regression model with application to mosquito surveillance data analysis. *Biometrical Journal*, Wiley Online Library, v. 59, n. 3, p. 462–477, 2017. Citado na página 18.

GHARBI, M.; QUENEL, P.; GUSTAVE, J.; CASSADOU, S.; RUCHE, G. L.; GIRDARY, L.; MARRAMA, L. Time series analysis of dengue incidence in guadeloupe, french west indies: forecasting models using climate variables as predictors. *BMC infectious diseases*, BioMed Central, v. 11, n. 1, p. 1–13, 2011. Citado na página 67.

HALSTEAD, S. B.; NIMMANNITYA, S.; COHEN, S. Observations related to pathogenesis of dengue hemorrhagic fever. iv. relation of disease severity to antibody response and virus recovered. *The Yale journal of biology and medicine*, Yale Journal of Biology and Medicine, v. 42, n. 5, p. 311, 1970. Citado na página 66.

HAWLEY, W. A.; REITER, P.; COPELAND, R. S.; PUMPUNI, C. B.; JR, G. B. C. Aedes albopictus in north america: probable introduction in used tires from northern asia. *Science*, American Association for the Advancement of Science, v. 236, n. 4805, p. 1114–1116, 1987. Citado na página 65.

HEAGERTY, P. J. Marginalized transition models and likelihood inference for longitudinal categorical data. *Biometrics*, Wiley Online Library, v. 58, n. 2, p. 342–351, 2002. Citado na página 17.

HII, Y. L.; ZHU, H.; NG, N.; NG, L. C.; ROCKLÖV, J. Forecast of dengue incidence using temperature and rainfall. *PLoS neglected tropical diseases*, Public Library of Science San Francisco, USA, v. 6, n. 11, p. e1908, 2012. Citado 2 vezes nas páginas 66 and 67.

HO, C. C.; TING, C.-Y. Time series analysis and forecasting of dengue using open data. In: SPRINGER. *Advances in Visual Informatics: 4th International Visual Informatics Conference, IVIC 2015, Bangi, Malaysia, November 17-19, 2015, Proceedings 4.* [S.l.], 2015. p. 51–63. Citado na página 67.

HORSTICK, O.; TOZAN, Y.; WILDER-SMITH, A. Reviewing dengue: still a neglected tropical disease? *PLoS neglected tropical diseases*, Public Library of Science San Francisco, CA USA, v. 9, n. 4, p. e0003632, 2015. Citado 3 vezes nas páginas 63, 73, and 100.

HSU, W.-Y.; WEN, T.-H.; YU, H.-L. Analysis of impact of geographical environment and socio-economic factors on the spatial distribution of kaohsiung dengue fever epidemic. In: *EGU General Assembly Conference Abstracts.* [S.l.: s.n.], 2013. p. EGU2013–9056. Citado na página 64.

HUGO, L. E.; JEFFERY, J. A.; TREWIN, B. J.; WOCKNER, L. F.; YEN, N. T.; LE, N. H.; NGHIA, L. T.; HINE, E.; RYAN, P. A.; KAY, B. H. Adult survivorship of the dengue mosquito aedes aegypti varies seasonally in central vietnam. *PLoS neglected* *tropical diseases*, Public Library of Science San Francisco, USA, v. 8, n. 2, p. e2669, 2014. Citado na página 66.

IBRAHIM, F.; FAISAL, T.; SALIM, M. M.; TAIB, M. N. Non-invasive diagnosis of risk in dengue patients using bioelectrical impedance analysis and artificial neural network. *Medical & biological engineering & computing*, Springer, v. 48, p. 1141–1148, 2010. Citado na página 67.

IBRAHIM, F.; TAIB, M. N.; ABAS, W. A. B. W.; GUAN, C. C.; SULAIMAN, S. A novel dengue fever (df) and dengue haemorrhagic fever (dhf) analysis using artificial neural network (ann). *Computer methods and programs in biomedicine*, Elsevier, v. 79, n. 3, p. 273–281, 2005. Citado na página 67.

ISLAM, M. A.; CHOWDHURY, R. I. A higher order markov model for analyzing covariate dependence. *Applied Mathematical Modelling*, Elsevier, v. 30, n. 6, p. 477–488, 2006. Citado na página 17.

JOHANSSON, M. A.; REICH, N. G.; HOTA, A.; BROWNSTEIN, J. S.; SANTILLANA, M. Evaluating the performance of infectious disease forecasts: A comparison of climate-driven and seasonal dengue forecasts for mexico. *Scientific reports*, Nature Publishing Group UK London, v. 6, n. 1, p. 33707, 2016. Citado na página 67.

JR, R. C. R.; STODDARD, S. T.; SCOTT, T. W. Socially structured human movement shapes dengue transmission despite the diffusive effect of mosquito dispersal. *Epidemics*, Elsevier, v. 6, p. 30–36, 2014. Citado na página 65.

KESORN, K.; ONGRUK, P.; CHOMPOOSRI, J.; PHUMEE, A.; THAVARA, U.; TAWATSIN, A.; SIRIYASATIEN, P. Morbidity rate prediction of dengue hemorrhagic fever (dhf) using the support vector machine and the aedes aegypti infection rate in similar climates and geographical areas. *PloS one*, Public Library of Science San Francisco, CA USA, v. 10, n. 5, p. e0125049, 2015. Citado na página 67.

LAL, A.; IKEDA, T.; FRENCH, N.; BAKER, M. G.; HALES, S. Climate variability, weather and enteric disease incidence in new zealand: time series analysis. *PLoS One*, Public Library of Science San Francisco, USA, v. 8, n. 12, p. e83484, 2013. Citado na página 67.

LEE, V. J.; LYE, D.; SUN, Y.; LEO, Y. Decision tree algorithm in deciding hospitalization for adult patients with dengue haemorrhagic fever in singapore. *Tropical Medicine & International Health*, Wiley Online Library, v. 14, n. 9, p. 1154–1159, 2009. Citado na página 66.

LIMKITTIKUL, K.; BRETT, J.; L'AZOU, M. Epidemiological trends of dengue disease in thailand (2000–2011): a systematic literature review. *PLoS neglected tropical diseases*, Public Library of Science San Francisco, USA, v. 8, n. 11, p. e3241, 2014. Citado na página 64.

LIN, H.; YANG, L.; LIU, Q.; WANG, T.; HOSSAIN, S. R.; HO, S. C.; TIAN, L. Time series analysis of japanese encephalitis and weather in linyi city, china. *International journal of public health*, Springer, v. 57, p. 289–296, 2012. Citado na página 67.

LIU, H.; SONG, X.; TANG, Y.; ZHANG, B. Bayesian quantile nonhomogeneous hidden markov models. *Statistical Methods in Medical Research*, SAGE Publications Sage UK: London, England, v. 30, n. 1, p. 112–128, 2021. Citado na página 18.

LÓPEZ, M. B.; FERNÁNDEZ, M. M.; VELASCO, M. G. A goodness of fit test in markov models with dependence of covariates. *Extracta mathematicae*, Departamento de Matemáticas, v. 10, n. 2, p. 140–145, 1995. Citado na página 17.

MACRAE, E. C. Estimation of time-varying markov processes with aggregate data. *Econometrica: journal of the Econometric Society*, JSTOR, p. 183–198, 1977. Citado na página 17.

MANIVANNAN, P.; ISAKKI, D. P. Dengue fever prediction using k-medoid clustering algorithm. *International Journal of Innovative Research in Computer and Communication Engineering*, v. 5, n. 1, 2017. Citado na página 67.

MATHUR, N.; ASIRVADAM, V. S.; DASS, S. C.; GILL, B. S. Visualization of dengue incidences for vulnerability using k-means. In: IEEE. *2015 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*. [S.l.], 2015. p. 569–573. Citado na página 67.

MELIGKOTSIDOU, L.; DELLAPORTAS, P. Forecasting with non-homogeneous hidden markov models. *Statistics and Computing*, Springer, v. 21, p. 439–449, 2011. Citado na página 18.

MUENZ, L. R.; RUBINSTEIN, L. V. Markov models for covariate dependence of binary sequences. *Biometrics*, JSTOR, p. 91–101, 1985. Citado na página 17.

OKI, M.; YAMAMOTO, T. Climate change, population immunity, and hyperendemicity in the transmission threshold of dengue. *PLoS One*, Public Library of Science San Francisco, USA, v. 7, n. 10, p. e48258, 2012. Citado na página 66.

ORGANIZATION, W. H. et al. Global strategy for dengue prevention and control 2012-2020. World Health Organization, 2012. Citado na página 64.

______. *A global brief on vector-borne diseases*. [S.l.], 2014. Citado na página 66.

PAROLI, R.; SPEZIA, L. Bayesian variable selection in markov mixture models. *Communications in Statistics-Simulation and Computation*, Taylor & Francis, v. 37, n. 1, p. 25–47, 2007. Citado na página 18.

PROKHORENKOVA, L.; GUSEV, G.; VOROBEV, A.; DOROGUSH, A. V.; GULIN, A. Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, v. 31, 2018. Citado na página 68.

RISSANEN, J. A universal data compression system. *IEEE Transactions on information theory*, IEEE, v. 29, n. 5, p. 656–664, 1983. Citado 4 vezes nas páginas 17, 19, 22, and 31.

______. Complexity of strings in the class of markov sources. *IEEE Transactions on Information Theory*, IEEE, v. 32, n. 4, p. 526–532, 1986. Citado na página 19.

RITA, A. B.; FREITAS, R.; NOGUEIRA, R. M. R. *Dengue*. Fundação Oswaldo Cruz (Fiocruz), 2016. Disponível em: https://agencia.fiocruz.br/dengue-0. Citado na página 61.

RUBIN, M. L.; CHAN, W.; YAMAL, J.-M.; ROBERTSON, C. S. A joint logistic regression and covariate-adjusted continuous-time markov chain model. *Statistics in medicine*, Wiley Online Library, v. 36, n. 28, p. 4570–4582, 2017. Citado na página 18.

SANG, S.; YIN, W.; BI, P.; ZHANG, H.; WANG, C.; LIU, X.; CHEN, B.; YANG, W.; LIU, Q. Predicting local dengue transmission in guangzhou, china, through the influence of imported cases, mosquito density and climate variability. *PloS one*, Public Library of Science San Francisco, USA, v. 9, n. 7, p. e102755, 2014. Citado na página 66.

SHARMA, K. D.; MAHABIR, R. S.; CURTIN, K. M.; SUTHERLAND, J. M.; AGARD, J. B.; CHADEE, D. D. Exploratory space-time analysis of dengue incidence in trinidad: a retrospective study using travel hubs as dispersal points, 1998–2004. *Parasites & vectors*, BioMed Central, v. 7, n. 1, p. 1–11, 2014. Citado na página 67.

SIRDARI, M. Z.; ISLAM, M. A. Goodness of fit test for higher order binary markov chain models. *Cogent Mathematics & Statistics*, Taylor & Francis, v. 5, n. 1, p. 1421003, 2018. Citado na página 18.

SIREGAR, F. A.; MAKMUR, T.; SAPRIN, S. Forecasting dengue hemorrhagic fever cases using arima model: a case study in asahan district. In: IOP PUBLISHING. *IOP Conference Series: Materials Science and Engineering*. [S.l.], 2018. v. 300, n. 1, p. 012032. Citado na página 67.

SIRIYASATIEN, P.; CHADSUTHI, S.; JAMPACHAISRI, K.; KESORN, K. Dengue epidemics prediction: A survey of the state-of-the-art based on data science processes. *IEEE Access*, IEEE, v. 6, p. 53757–53795, 2018. Citado na página 68.

SIRIYASATIEN, P.; PHUMEE, A.; ONGRUK, P.; JAMPACHAISRI, K.; KESORN, K. Analysis of significant factors for dengue fever incidence prediction. *BMC bioinformatics*, BioMed Central, v. 17, n. 1, p. 1–9, 2016. Citado na página 66.

STANAWAY, J. D.; SHEPARD, D. S.; UNDURRAGA, E. A.; HALASA, Y. A.; COFFENG, L. E.; BRADY, O. J.; HAY, S. I.; BEDI, N.; BENSENOR, I. M.; CASTAÑEDA-ORJUELA, C. A. et al. The global burden of dengue: an analysis from the global burden of disease study 2013. *The Lancet infectious diseases*, Elsevier, v. 16, n. 6, p. 712–723, 2016. Citado 3 vezes nas páginas 64, 66, and 68.

TANNER, L.; SCHREIBER, M.; LOW, J. G.; ONG, A.; TOLFVENSTAM, T.; LAI, Y. L.; NG, L. C.; LEO, Y. S.; PUONG, L. T.; VASUDEVAN, S. G. et al. Decision tree algorithms predict the diagnosis and outcome of dengue fever in the early phase of illness. *PLoS neglected tropical diseases*, Public Library of Science San Francisco, USA, v. 2, n. 3, p. e196, 2008. Citado na página 66.

THAMMAPALO, S.; CHONGSUWIWATWONG, V.; MCNEIL, D.; GEATER, A. The climatic factors influencing the occurrence of dengue hemorrhagic fever in thailand. *Southeast Asian J Trop Med Public Health*, v. 36, n. 1, p. 191–196, 2005. Citado na página 64.

THIRUCHELVAM, L.; DASS, S. C.; ZAKI, R.; YAHYA, A.; ASIRVADAM, V. S. Correlation analysis of air pollutant index levels and dengue cases across five different zones in selangor, malaysia. *Geospatial health*, v. 13, n. 1, 2018. Citado na página 65.

THISYAKORN, U.; NIMMANNITYA, S. Nutritional status of children with dengue hemorrhagic fever. *Clinical Infectious Diseases*, The University of Chicago Press, v. 16, n. 2, p. 295–297, 1993. Citado na página 66.

VECCHIA, A. D.; BELTRAME, V.; D'AGOSTINI, F. M. Panorama da dengue na região sul do brasil de 2001 a 2017. *Cogitare Enfermagem*, Universidade Federal do Paraná, v. 23, n. 3, 2018. Citado na página 62.

VEERASEATAKUL, P.; SAOSATHAN, S.; CHUTIPONGVIVATE, S. Pattern of dengue serotypes in four provinces of northern thailand from 2003–2012. *1. Epidemiological importance of container pupal index (CPI), for vector surveillance and control of dengue in national capital territory (NCT)–Delhi*, v. 38, p. 11, 2014. Citado na página 65.

VENABLES, W.; RIPLEY, B. *Modern applied statistics with s fourth edition by, world.* [S.l.]: Springer, 2002. Citado 2 vezes nas páginas 37 and 41.

VERMUNT, J. K.; LANGEHEINE, R.; BOCKENHOLT, U. Discrete-time discrete-state latent markov models with time-constant and time-varying covariates. *Journal of Educational and Behavioral Statistics*, Sage Publications, v. 24, n. 2, p. 179–207, 1999. Citado na página 17.

WAIDAB, W.; SUPHAPEETIPORN, K. et al. Pathogenesis of dengue hemorrhagic fever: From immune to genetics. *Journal of Pediatric Infectious Diseases*, IOS Press, v. 3, n. 4, p. 221–227, 2008. Citado na página 66.

WEINBERGER, M. J.; RISSANEN, J. J.; FEDER, M. A universal finite memory source. *IEEE Transactions on Information theory*, IEEE, v. 41, n. 3, p. 643–652, 1995. Citado na página 30.

XU, H.-Y.; FU, X.; LEE, L. K. H.; MA, S.; GOH, K. T.; WONG, J.; HABIBULLAH, M. S.; LEE, G. K. K.; LIM, T. K.; TAMBYAH, P. A. et al. Statistical modeling reveals the effect of absolute humidity on dengue in singapore. *PLoS neglected tropical diseases*, Public Library of Science San Francisco, USA, v. 8, n. 5, p. e2805, 2014. Citado na página 66.

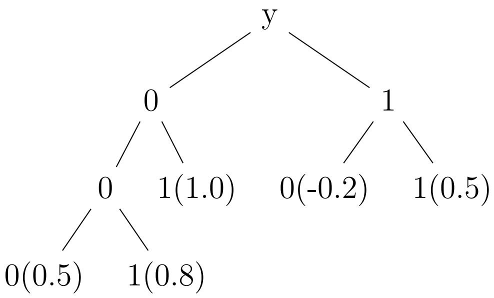
YU, H.-L.; ANGULO, J. M.; CHENG, M.-H.; WU, J.; CHRISTAKOS, G. An online spatiotemporal prediction model for dengue fever epidemic in k aohsiung (t aiwan). *Biometrical Journal*, Wiley Online Library, v. 56, n. 3, p. 428–440, 2014. Citado na página 65.

ZAMBOM, A. Z.; KIM, S.; GARCIA, N. L. Variable length markov chain with exogenous covariates. *Journal of Time Series Analysis*, Wiley Online Library, v. 43, n. 2, p. 312–328, 2022. Citado 10 vezes nas páginas 18, 30, 31, 34, 35, 36, 41, 42, 43, and 48.

Appendix

APPENDIX A – Extra simulations

Model 8 is similar to Model 2 but with all exogenous covariate parameters equal to zero ($\boldsymbol{\beta}^u = \mathbf{0}\ \forall u$), in order to evaluate the performance of the proposed algorithm in absence of additional information about the transition probability.



$$\boldsymbol{\beta}^{000} = (0,0,0)'$$
$$\boldsymbol{\beta}^{001} = (0,0,0)'$$
$$\boldsymbol{\beta}^{01} = (0,0)'$$
$$\boldsymbol{\beta}^{10} = (0,0)'$$
$$\boldsymbol{\beta}^{11} = (0,0)'$$

Figure 22 – Context tree τ and associated parameters $\boldsymbol{\theta}$ for Model 8 (numbers in parenthesis represent the values of α^u for each context u).

The performance metrics for the estimated context function in Model 8 are outlined in Table 37. To aid in the visualization of the estimated trees, Table 38 and Table 39 provide the count of estimated occurrences for each context. For both instances with $n = 1000$ and $n = 2000$, neither method achieved any identical tree τ. In the case of $n = 1000$, both methods exhibited suboptimal estimates, with a notable number of missing and extra nodes. Moreover, both methods consistently pruned nodes '000'and '001'and the original version struggled to prune other branches. The pruning of nodes '000'and '001'is attributed to the closely situated values of parameters α^{000} and α^{001}, necessitating a larger number of observations to discern the difference. Both methods failure to capture the significant nodes.

The average differences between real and estimated parameters for both the proposed modified beta-context algorithm and the beta-context algorithm are illustrated in Table 40. Larger differences are observed for nodes '000'and '001', which are consistently pruned in the modified version of the algorithm.

Model 9, outlined in Figure 23, is similar to Model 5, except that all exogenous covariate parameters are set to zero ($\boldsymbol{\beta}^u = \mathbf{0},\ \forall u$).

The performance metrics for the estimated context function in Model 9 are detailed in Table 41. To aid in the visualization of the estimated trees, Table 42 and Table 43 provide the count of estimated occurrences for each context. For both $n = 1000$ and $n = 2000$, neither of the algorithms achieved an identical τ tree. This is because the values of $\alpha^{10} = 1.00$ and $\alpha^{11} = 0.80$ are too close to each other, requiring a large number of

Table 37 – Simulation results for Model 8 with time-varying exogenous covariates generated from a N(0, 1) distribution (average over 100 simulations).

	Model 8 (n = 1000, N(0, 1) distr.)		Model 8 (n = 2000, N(0, 1) distr.)	
	modified-beta-VLMC	beta-VLMC	modified-beta-VLMC	beta-VLMC
BIC	1313.51	1329.89	2612.51	2615.62
AIC	1297.95	1294.45	2589.15	2589.19
logLik	-645.81	-640.01	-1290.41	-1289.87
# par. $\hat{\alpha}^u$	2.96	5.16	3.55	3.99
# par. $\hat{\boldsymbol{\beta}}^u$	0.21	2.06	0.62	0.73
order $\hat{\tau}$	1.93	3.56	2.44	2.81
order-Cov	0.21	1.69	0.60	0.69
# Missing $\hat{\tau}$	3.53	3.11	2.91	2.92
# Extra $\hat{\tau}$	0.40	3.62	0.90	1.57
Identical τ	0	0	0	0
Identical $\tau_{\boldsymbol{\theta}}$	0	0	0	0

Table 38 – Estimated τ trees for Model 8, with n = 1000 (frequency of occurrences over 100 simulations).

Contexts	# Contexts	modified-beta-VLMC	beta-VLMC
0 - 1	2	27	13
0 - 1 0 - 1 1	3	62	38
0 - 1 0 0 - 1 0 1 - 1 1	4	1	0
0 0 - 0 1 - 1 0 - 1 1	4	3	1
0 - 1 0 - 1 1 0 - 1 1 1 0 - 1 1 1 1	5	1	1
0 - 1 0 - 1 1 0 0 - 1 1 0 1 - 1 1 1	5	1	1
0 - 1 0 0 0 - 1 0 0 1 - 1 0 1 - 1 1	5	1	1
0 0 - 0 1 0 - 0 1 1 0 - 0 1 1 1 - 1	5	1	1
0 0 - 0 1 0 0 - 0 1 0 1 - 0 1 1 - 1	5	1	1
0 0 0 - 0 0 0 1 - 0 0 1 - 0 1 - 1	5	0	1
0 0 0 - 0 0 1 0 - 0 0 1 1 - 0 1 - 1	5	0	1
0 - 1 0 - 1 1 0 0 - 1 1 0 0 1 - 1 1 0 1 - 1 1 1	6	0	1
0 - 1 0 0 - 1 0 1 0 - 1 0 1 1 0 - 1 0 1 1 1 - 1 1	6	1	1
0 - 1 0 0 - 1 0 1 0 0 - 1 0 1 0 1 - 1 0 1 1 - 1 1	6	0	1
0 0 - 0 1 0 - 0 1 1 0 - 0 1 1 0 1 - 0 1 1 1 - 1	6	0	1
0 0 - 0 1 0 - 0 1 1 0 - 0 1 1 1 0 - 0 1 1 1 1 - 1	6	0	1
0 0 0 0 - 0 0 0 1 - 0 0 1 - 0 1 - 1 0 - 1 1	6	0	1
>= 7 contexts	7	1	35

observations to distinguish these nodes. Besides that, the modified beta-context algorithm seems to have better performance, with lower values of AIC and BIC and fewer extra nodes estimated. In general, the original beta-context algorithm results in larger trees.

The differences between the real and estimated parameters, when the context were correctly identified, are shown in Table 44 for both the proposed modified beta-context algorithm and the original beta-context algorithm. Both methods generally provide similar results for most parameters, except for α^{00}. This discrepancy arises for the same reason as explained before for Model 5. For $n = 2000$, the modified version of the algorithm did not estimate trees with nodes ‘10’and ‘11’due to the proximity of parameters α^{10} and α^{11}.

Table 39 – Estimated τ trees for Model 8, with n = 2000 (frequency of occurrences over 100 simulations).

Contexts	# Contexts	modified-beta-VLMC	beta-VLMC
0 - 1	2	1	5
0 - 1 0 - 1 1	3	78	72
0 0 - 0 1 - 1 0 - 1 1	4	7	4
0 - 1 0 - 1 1 0 0 - 1 1 0 1 - 1 1 1	5	1	0
0 - 1 0 0 - 1 0 1 0 - 1 0 1 1 - 1 1	5	1	0
0 - 1 0 - 1 1 0 - 1 1 1 0 0 - 1 1 1 0 1 - 1 1 1 1	6	1	0
0 - 1 0 - 1 1 0 0 - 1 1 0 1 0 - 1 1 0 1 1 - 1 1 1	6	1	0
0 - 1 0 0 - 1 0 1 0 - 1 0 1 1 0 - 1 0 1 1 1 - 1 1	6	1	0
0 - 1 0 0 - 1 0 1 0 0 - 1 0 1 0 1 - 1 0 1 1 - 1 1	6	1	0
0 - 1 0 0 0 - 1 0 0 1 0 - 1 0 0 1 1 - 1 0 1 - 1 1	6	1	0
0 0 - 0 1 - 1 0 0 0 - 1 0 0 1 - 1 0 1 - 1 1	6	1	0
0 0 - 0 1 0 0 - 0 1 0 1 - 0 1 1 - 1 0 - 1 1	6	1	0
>= 7 contexts	7	5	19

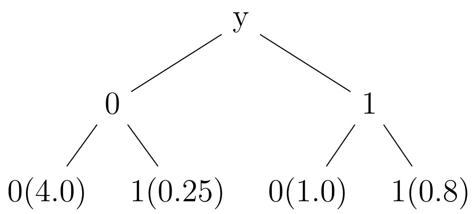
Table 40 – Differences between real and estimated values for Model 8 (average over 100 simulations).

	Model 8 (n = 1000, N(0, 1) distr.)		Model 8 (n = 2000, N(0,1) distr.)	
	modified-beta-VLMC	beta-VLMC	modified-beta-VLMC	beta-VLMC
α^{000}	-	0.12	-	0.37
α^{001}	-	0.16	-	-
α^{01}	0.06	0.15	0.12	0.09
α^{10}	0.09	0.09	0.06	0.06
α^{11}	0.08	0.09	0.07	0.06
$\boldsymbol{\beta}^{000}$	-	(0, 0, 0)	-	(0, 0, 0)
$\boldsymbol{\beta}^{001}$	-	(0, 0, 0)	-	-
$\boldsymbol{\beta}^{01}$	(0, 0)	(0, 0)	(0, 0)	(0, 0)
$\boldsymbol{\beta}^{10}$	(0, 0)	(0, 0)	(0, 0)	(0, 0)
$\boldsymbol{\beta}^{11}$	(0,0)	(0,0)	(0,0)	(0,0)

Model 10, outlined in Figure 24, is similar to Model 6, except that all exogenous covariate parameters are set to zero ($\boldsymbol{\beta}^u = \mathbf{0} \forall u$).

The performance metrics for the estimated context function in Model 10 are detailed in Table 45. To aid in the visualization of the estimated trees, Table 46 and Table 47 provide the count of estimated occurrences for each context. For both $n = 4000$ and $n = 8000$, neither of the algorithms achieved more than 30% identical τ trees. This is due to the values of $\alpha^{01} = (-0.5, -0.5)$ and $\alpha^{02} = (-0.35, -1)$ being too close to each other, requiring a large number of observations to distinguish these nodes. Additionally, for $n = 8000$, the modified beta-context algorithm seems to have better performance, with fewer missing and extra nodes estimated.

The disparities between the real and estimated parameters, when the context were correctly identified, are displayed in Table 48 for both the proposed modified



$$\boldsymbol{\beta}^{00} = (0,0)'$$
$$\boldsymbol{\beta}^{01} = (0,0)'$$
$$\boldsymbol{\beta}^{10} = (0,0)'$$
$$\boldsymbol{\beta}^{11} = (0,0)'$$

Figure 23 – Context tree τ and associated parameters $\boldsymbol{\theta}$ for Model 9 (numbers in parenthesis represent the values of α^u for each context u).

Table 41 – Simulation results for Model 9 with time-varying exogenous covariates generated from a N(0, 1) distribution (average over 100 simulations).

	Model 9 (n = 1000, N(0, 1) distr.)		Model 9 (n = 2000, N(0, 1) distr.)	
	modified-beta-VLMC	beta-VLMC	modified-beta-VLMC	beta-VLMC
BIC	1170.71	1201.43	2331.50	2372.92
AIC	1155.30	1174.34	2314.70	2325.48
logLik	-574.51	-581.65	-1154.35	-1154.30
# par. $\hat{\alpha}^u$	3.05	4.24	3.00	5.72
# par. $\hat{\boldsymbol{\beta}}^u$	0.09	1.28	0	2.75
order $\hat{\tau}$	2.04	3.00	2.00	4.04
order-Cov	0.09	1.08	0	2.31
# Missing $\hat{\tau}$	1.98	2.00	2.00	1.64
# Extra $\hat{\tau}$	0.08	2.26	0	4.51
Identical τ	0	0	0	0
Identical $\tau_{\boldsymbol{\theta}}$	0	0	0	0

Table 42 – Estimated τ trees for Model 9, with n = 1000 (frequency of occurrences over 100 simulations).

Contexts	# Contexts	modified-beta-VLMC	beta-VLMC
0 - 1	2	0	14
0 0 - 0 1 - 1	3	98	64
0 0 - 0 1 0 0 - 0 1 0 1 - 0 1 1 - 1	5	1	1
0 0 - 0 1 - 1 0 - 1 1 0 0 - 1 1 0 1 - 1 1 1	6	1	0
>= 7 contexts	7	0	21

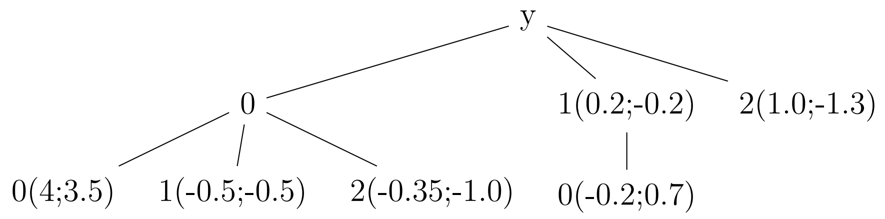
Table 43 – Estimated τ trees for Model 9, with n = 2000 (frequency of occurrences over 100 simulations).

Contexts	# Contexts	modified-beta-VLMC	beta-VLMC
0 - 1	2	0	4
0 0 - 0 1 - 1	3	100	57
0 0 - 0 0 1 0 - 0 0 1 1 0 - 0 0 1 1 1 - 0 1 - 1	6	0	2
>= 7 contexts	7	0	37

beta-context algorithm and the original beta-context algorithm. Both methods exhibit comparable results for all parameters.

Table 44 – Differences between real and estimated values for Model 9 (average over 100 simulations).

	Model 9 (n = 1000, N(0, 1) distr.)		Model 9 (n = 2000, N(0,1) distr.)	
	modified-beta-VLMC	beta-VLMC	modified-beta-VLMC	beta-VLMC
α^{00}	8.34	0.54	5.57	0.53
α^{01}	0.10	0.10	0.08	0.08
α^{10}	0.09	0.12	-	0.09
α^{11}	-	0.04	-	0.07
$\boldsymbol{\beta}^{00}$	(0, 0)	(0, 0)	(0, 0)	(0, 0)
$\boldsymbol{\beta}^{01}$	(0, 0)	(0, 0)	(0, 0)	(0, 0)
$\boldsymbol{\beta}^{10}$	(0, 0)	(0, 0)	(0, -)	(0, 0)
$\boldsymbol{\beta}^{11}$	-	(0, 0)	-	(0, 0)



$$\begin{array}{ll}
\boldsymbol{\beta}_1^{00} = (0,0)' & \boldsymbol{\beta}_2^{00} = (0,0)' \\
\boldsymbol{\beta}_1^{01} = (0,0)' & \boldsymbol{\beta}_2^{01} = (0,0)' \\
\boldsymbol{\beta}_1^{02} = (0,0)' & \boldsymbol{\beta}_2^{02} = (0,0)' \\
\boldsymbol{\beta}_1^{10} = (0,0)' & \boldsymbol{\beta}_2^{10} = (0,0)' \\
\boldsymbol{\beta}_1^{1*} = (0)' & \boldsymbol{\beta}_2^{1*} = (0)' \\
\boldsymbol{\beta}_1^{2} = (0)' & \boldsymbol{\beta}_2^{2} = (0)'
\end{array}$$

Figure 24 – Context tree τ and associated parameters $\boldsymbol{\theta}$ for Model 10. Numbers in parenthesis represent the values of (α_1^u, α_2^u) from each context u and $\boldsymbol{\beta}_1^u$ and $\boldsymbol{\beta}_2^u$ are the coefficient vectors of the covariates. $\boldsymbol{\beta}^{1*}$ means the context is 1 preceded by other state that is not 0.

Table 45 – Simulation results for Model 10 with time-varying exogenous covariates generated from a N(0, 1) distribution (average over 100 simulations).

	Model 10 (n = 4000, N(0, 1) distr.)		Model 10 (n = 8000, N(0, 1) distr.)	
	modified-beta-VLMC	beta-VLMC	modified-beta-VLMC	beta-VLMC
BIC	7793.40	7798.41	15491.46	15496.14
AIC	7730.21	7709.54	15416.28	15414.25
logLik	-3855.06	-3840.65	-7697.38	-7695.40
# par. $\hat{\alpha}^u$	10.04	12.40	10.66	12.92
# par. $\hat{\boldsymbol{\beta}}^u$	0	1.72	0.10	0.80
order $\hat{\tau}$	2.00	2.60	2.03	2.23
order-Cov	0	0.83	0.05	0.36
# Missing $\hat{\tau}$	1.96	1.44	1.42	1.60
# Extra $\hat{\tau}$	0	1.01	0.05	0.26
Identical τ	0.02	0.16	0.29	0.18
Identical $\tau_{\boldsymbol{\theta}}$	0.02	0.15	0.29	0.18

Table 46 – Estimated τ trees for Model 10, with n = 4000 (frequency of occurrences over 100 simulations).

Contexts	# Contexts	modified-beta-VLMC	beta-VLMC
0 - 0 0 - 1 - 1 0 - 2	5	92	53
0 - 0 1 - 1 - 1 0 - 2	5	6	0
0 0 - 0 1 - 0 2 - 1 - 1 0 - 2	**6**	**2**	**16**
>= 7 contexts	7	0	31

Table 47 – Estimated τ trees for Model 10, with n = 8000 (frequency of occurrences over 100 simulations).

Contexts	# Contexts	modified-beta-VLMC	beta-VLMC
0 - 0 0 - 1 - 1 0 - 2	5	69	75
0 - 0 1 - 1 - 1 0 - 2	5	1	0
0 0 - 0 1 - 0 2 - 1 - 1 0 - 2	**6**	**29**	**18**
>= 8 contexts	8	1	7

Table 48 – Differences between real and estimated values for Model 10 (average over 100 simulations).

	Model 10 (n = 4000, N(0, 1) distr.)		Model 10 (n = 8000, N(0,1) distr.)	
	modified-beta-VLMC	beta-VLMC	modified-beta-VLMC	beta-VLMC
$(\alpha_1^{00}, \alpha_2^{00})$	(0.31, 0.32)	(0.37, 0.38)	(0.28, 0.27)	(0.34, 0.33)
$(\alpha_1^{01}, \alpha_2^{01})$	(0.06, 0.06)	(0.09, 0.09)	(0.05, 0.06)	(0.05, 0.06)
$(\alpha_1^{02}, \alpha_2^{02})$	(0.19, 0.26)	(0.12, 0.16)	(0.08, 0.11)	(0.08, 0.10)
$(\alpha_1^{10}, \alpha_2^{10})$	(0.05, 0.05)	(0.05, 0.05)	(0.04, 0.04)	(0.04, 0.04)
$(\alpha_1^{1*}, \alpha_2^{1*})$	(0.10, 0.09)	(0.11, 0.09)	(0.08, 0.06)	(0.08, 0.06)
$(\alpha_1^{2}, \alpha_2^{2})$	(0.05, 0.11)	(0.06, 0.11)	(0.04, 0.08)	(0.04, 0.08)
$\boldsymbol{\beta}_1^{00}$	(0, 0)	(0, 0)	(0, 0)	(0, 0)
$\boldsymbol{\beta}_2^{00}$	(0, 0)	(0, 0)	(0, 0)	(0, 0)
$\boldsymbol{\beta}_1^{01}$	(0, 0)	(0, 0)	(0, 0)	(0, 0)
$\boldsymbol{\beta}_2^{01}$	(0, 0)	(0, 0)	(0, 0)	(0, 0)
$\boldsymbol{\beta}_1^{02}$	(0, 0)	(0.01, 0)	(0, 0)	(0, 0)
$\boldsymbol{\beta}_2^{02}$	(0, 0)	(0.02, 0)	(0, 0)	(0, 0)
$\boldsymbol{\beta}_1^{10}$	(0, 0)	(0, 0)	(0, 0)	(0, 0)
$\boldsymbol{\beta}_2^{10}$	(0, 0)	(0, 0)	(0, 0)	(0, 0)
$\boldsymbol{\beta}_1^{1*}$	(0)	(0)	(0)	(0)
$\boldsymbol{\beta}_2^{1*}$	(0)	(0)	(0)	(0)
$\boldsymbol{\beta}_1^{2}$	(0)	(0)	(0)	(0)
$\boldsymbol{\beta}_2^{2}$	(0)	(0)	(0)	(0)

APPENDIX B – Descriptive analysis - dengue incidence in Brazilian municipalities

To decide which exogenous variable to include in the model and understand the behavior of dengue cases in Brazil over the months, we conducted some descriptive analyses, as outlined below:

Dengue incidence

Figure 25 presents the average monthly dengue cases per 100,000 inhabitants from January 2008 to July 2023 for selected municipalities. These municipalities are categorized based on the Brazilian Ministry of Health's epidemic definition, where a year is considered epidemic if it exceeds 100 cases per 100,000 inhabitants. The graph reveals a pattern of high dengue incidence in the first semester of the year, followed by a decrease in the subsequent months. Notably, the years 2013, 2015, 2016, 2019, 2022, and 2023 stand out with elevated dengue incidence. It's worth considering that reported dengue cases may have been underreported during the COVID-19 pandemic (2020 and 2021) due to social isolation and healthcare systems grappling with COVID-19 cases.

To provide information about the representation of the sample, Table 49 presents the distribution of municipalities in each category of epidemic and non-epidemic situations.

Table 49 – Number of municipalities per year categorized as epidemic and non-epidemic

Municipality	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Epidemic	50	55	87	64	59	85	52	70	71	25	25	61	43	27	41	48
Non-epidemic	50	49	20	46	28	17	42	17	11	46	53	16	26	36	18	19

Temperature

Figure 26 presents the arithmetic mean of monthly average temperatures from January 2008 to July 2023 for selected municipalities, categorized into epidemic and non-epidemic years. The graph indicates that, in the majority of years, the average monthly temperatures tends to be higher for municipalities in epidemic years.

In Figure 27, we attempt to discern whether there is a correlation between higher yearly average temperatures and an increased number of dengue cases in the year. However, the graph suggests that dengue occurs in municipalities with an average temperature above 20ºC, and beyond this threshold, there is no apparent association between higher temperatures and increased case numbers.

Rainfall

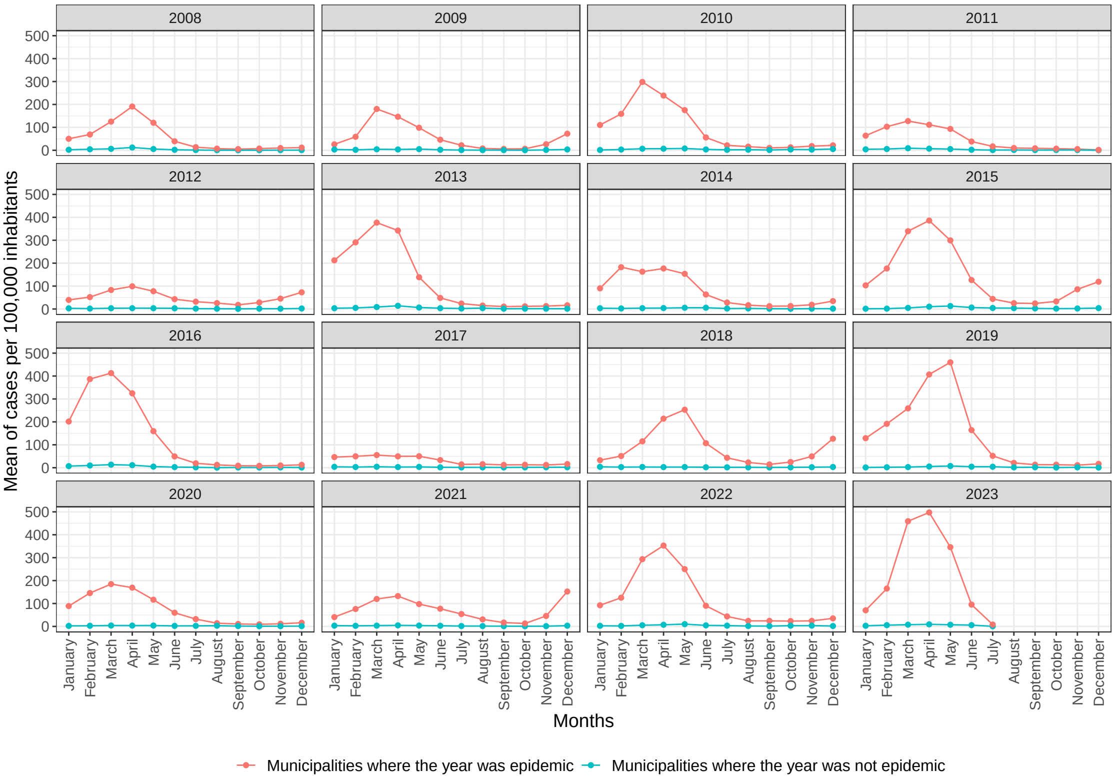


Figure 25 – Comparison of average monthly dengue cases per 100,000 inhabitants in municipalities during dengue epidemic and non-epidemic years (2008 to 2023)

Figure 28 presents a comparison of monthly rainfall averages (mm) among municipalities during dengue epidemic and non-epidemic years. The figure illustrates that, despite higher rainfall in months with increased dengue incidence (typically in the first months of the year), there is no big difference between municipalities in epidemic years and those in non-epidemic years. Interestingly, in some years, the monthly rainfall average (mm) is higher for municipalities in a non-epidemic situation. On the other hand, months with higher precipitation present higher number of cases.

Days of rainfall

When examining the days of rainfall instead of total rainfall (mm) (Figure 29), the results remain consistent with those observed in Figure 28.

Low-income population

Figure 30 illustrates the relationship between the percentage of the population in low-income situations and yearly dengue cases in municipalities during epidemic years. The analysis suggests no direct association between the two factors. This observation may be related to the potential underreporting of cases in areas with a low-income situation due to limited access to health centers or with the findings highlighted by Horstick, Tozan e Wilder-Smith (2015) - while dengue predominantly affects resource-limited countries, it

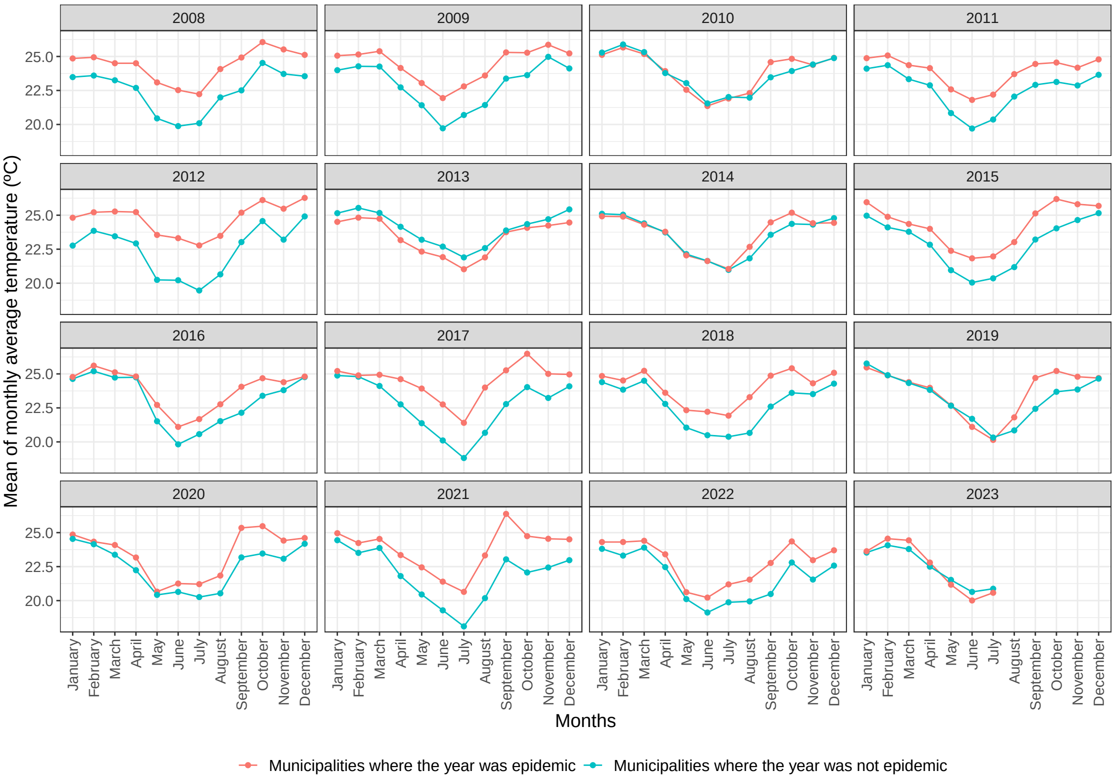


Figure 26 – Comparison of average monthly temperature (ºC) in municipalities during dengue epidemic and non-epidemic years (2008 to 2023)

does not exclusively target the poor.

Figure 31 shows the distribution of yearly dengue cases based on the percentage of the population in low-income situations in municipalities during epidemic years. It is interesting to note that the years with higher dengue incidence for municipalities with less than 20% of the population in low-income situations are the same years with high dengue incidence in the country, as observed in Figure 25. For other years, the incidence is higher in municipalities with more than 20% of the population in low-income situations.

To provide information about the representation of the sample used, Table 50 presents the distribution of the percentage of population in low-income situations for all the 126 municipalities.

Table 50 – Distribution of the percentage of population in low-income situations for all the 126 municipalities

Minimun	1st Quartile	Median	Mean	3st Quartile	Maximun
7.49	19.05	27.81	31.37	41.63	75.75

Gross Domestic Product (GDP) per capita

Figure 32 illustrates the correlation between Gross National Product (GDP)

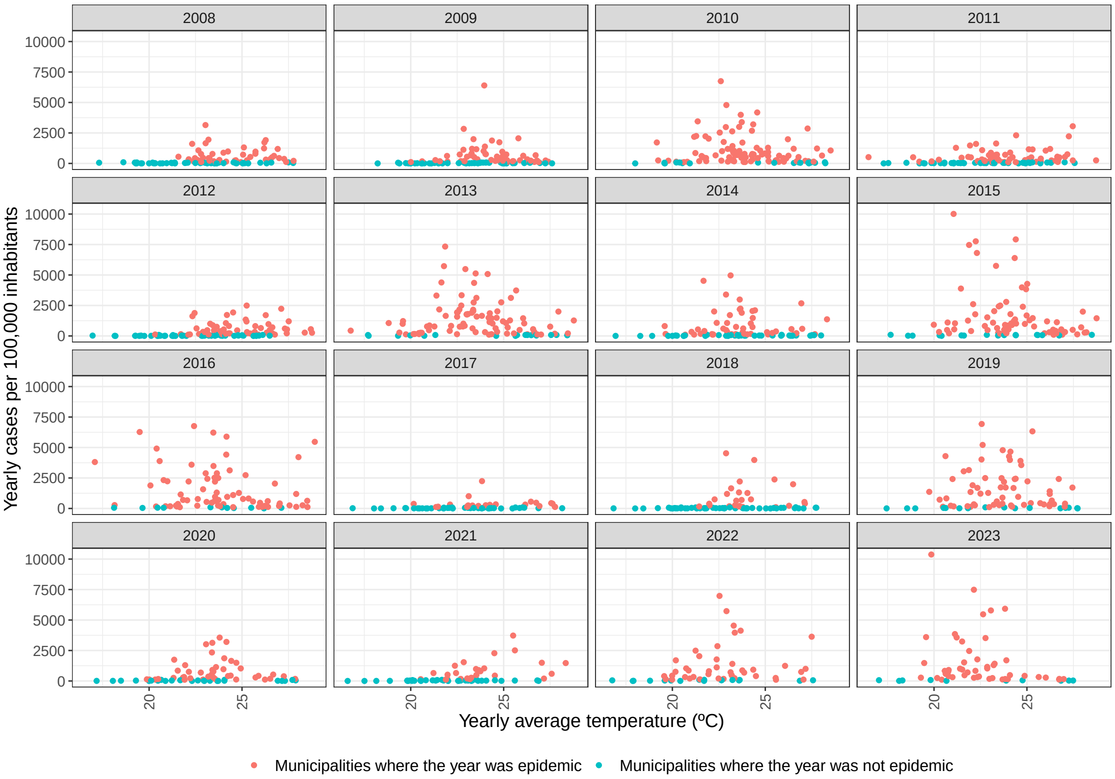


Figure 27 – Association between average yearly temperature (ºC) and yearly dengue cases per 100,000 inhabitants in municipalities during epidemic and non-epidemic years (2008 to 2023)

(R$) and yearly dengue cases per 100,000 inhabitants for municipalities during epidemic years. It indicates that, in general, municipalities with a high incidence of dengue tend to have a GDP below R$ 25,000. However, within this range, there is no clear direct association between the two factors.

Figure 33 shows the distribution of yearly dengue cases based on GDP per capita (R$) in municipalities during epidemic years. Similar to the observation for the low-income population factor, it is noteworthy that the years with higher dengue incidence for municipalities with GDP per capita higher than R$ 25,000 align with some of the years with high dengue incidence in the country, as observed in Figure 25. Conversely, for other years, the incidence is higher in municipalities with less than R$ 25,000 of GDP per capita.

To provide information about the representation of the sample used, Table 51 presents the distribution of GDP per capita (R$) for all the 126 municipalities.

Table 51 – Distribution of GDP per capita (R$) for all the 126 municipalities

Minimun	1st Quartile	Median	Mean	3st Quartile	Maximun
3643	10963	16035	20046	25004	83428

Population with sewage system

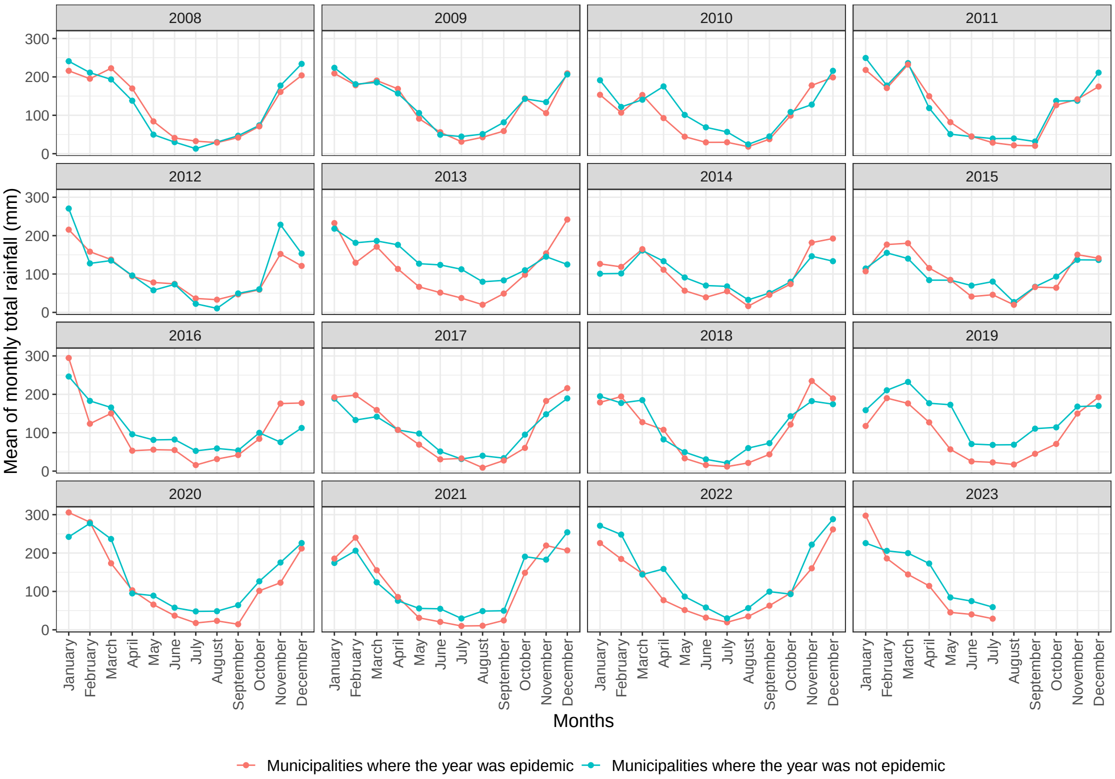


Figure 28 – Comparison of monthly rainfall (mm) averages in municipalities during dengue epidemic and non-epidemic years (2008 to 2023)

Figure 34 illustrates the relationship between the percentage of the population with sewage system and yearly dengue cases in municipalities during epidemic years. The analysis suggests no direct association between the two factors.

Figure 35 illustrates the distribution of yearly dengue cases based on the percentage of the population with a sewage system in municipalities during epidemic years. Similar conclusions can be drawn as those presented for the low-income population and GDP per capita factors.

To provide information about the representation of the sample used, Table 52 presents the distribution of the percentage of population with sewage system for all the 126 municipalities.

Table 52 – Distribution of the percentage of population with sewage system for all the 126 municipalities

Minimun	1st Quartile	Median	Mean	3st Quartile	Maximun
0.06	30.70	66.81	56.30	85.97	97.58

Population in urban area

Figure 36 illustrates the relationship between the percentage of the population

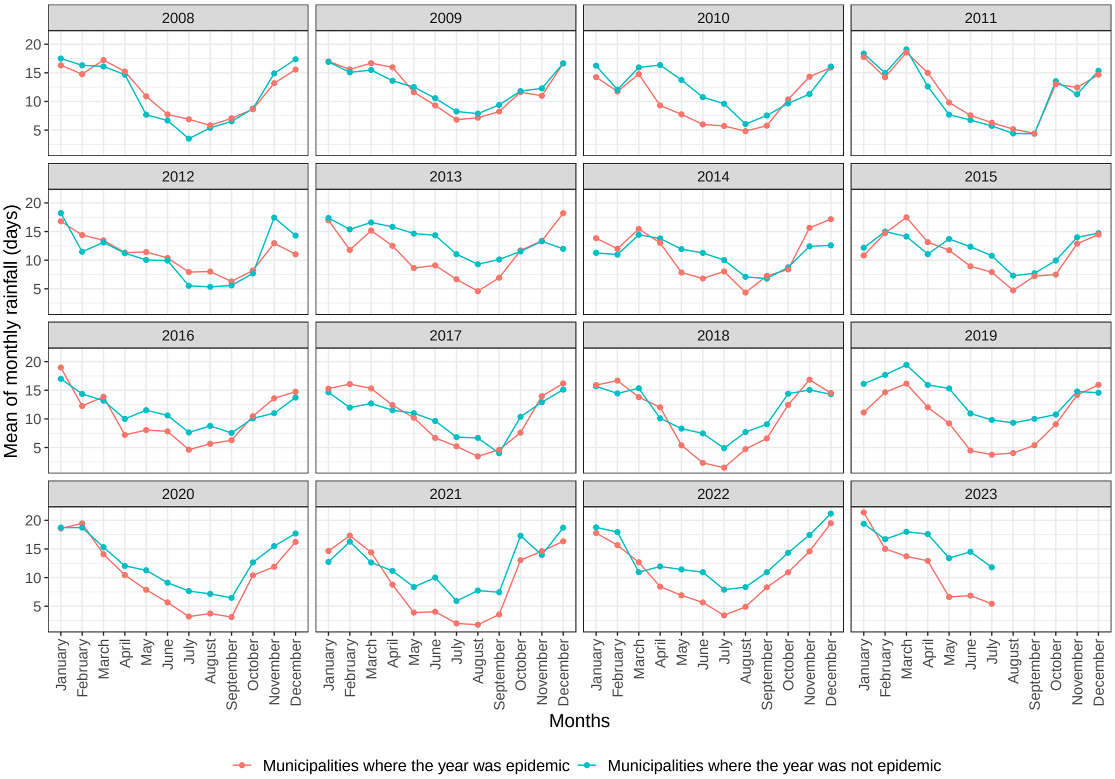


Figure 29 – Comparison of monthly rainfall (days) averages in municipalities during dengue epidemic and non-epidemic years (2008 to 2023)

living in urban areas and yearly dengue cases in municipalities during epidemic years. The graph suggests a correlation between the percentage of the population in urban areas and the number of dengue cases, with higher cases in municipalities where a larger proportion of the population lives in urban areas.

Figure 37 illustrates the distribution of yearly dengue cases based on the percentage of the population with a sewage system in municipalities during epidemic years. Except for the years 2021 and 2018, conclusions are similar to the ones presented for Figure 36.

To provide information about the representation of the sample used, Table 53 presents the distribution of the percentage of population living in urban areas for all the 126 municipalities.

Table 53 – Distribution of the percentage of population living in urban areas for all the 126 municipalities

Minimun	1st Quartile	Median	Mean	3st Quartile	Maximun
34.16	81.82	93.01	88.32	97.70	100.00

Demographic density

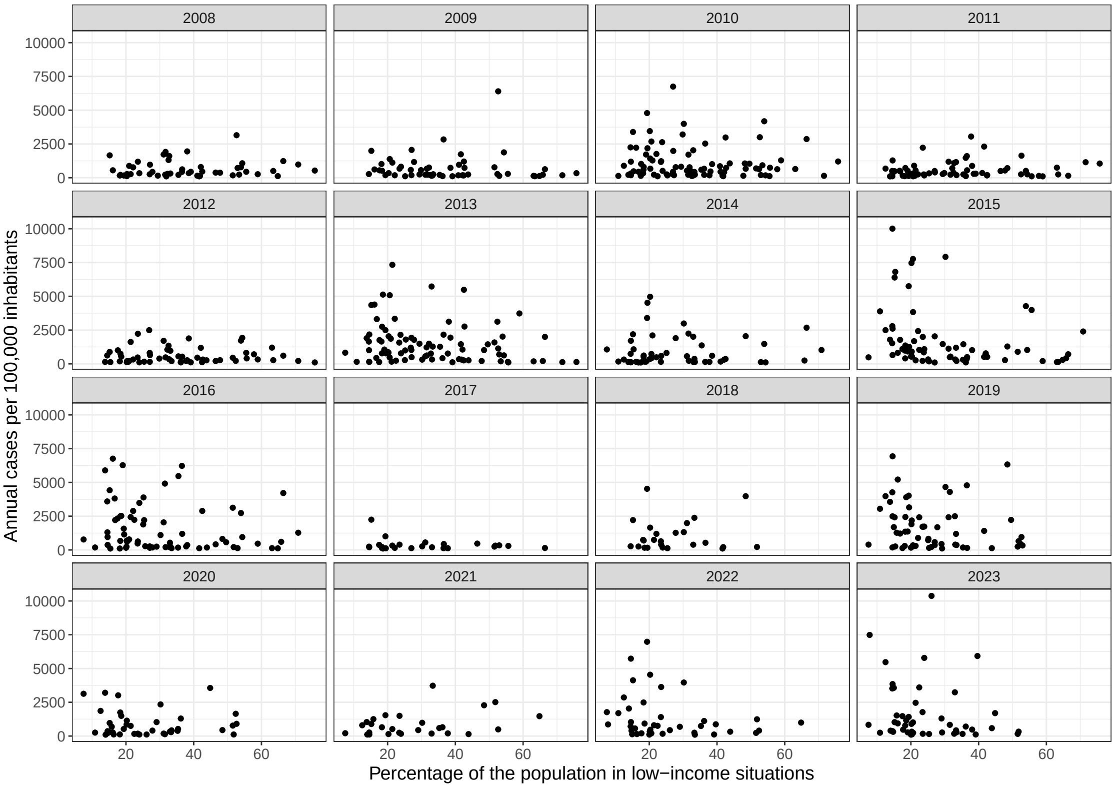


Figure 30 – Association between the percentage of population in low-income situations and yearly dengue cases per 100,000 inhabitants for municipalities during epidemic years (2008 to 2023)

To address the high variation in values for demographic density, as indicated in Table 54, we opted to present only the boxplot visualization, as shown in Figure 38. This format allows for a clearer visualization of the varying behaviors observed in different years.

Table 54 – Distribution of the demographic density for all the 126 municipalities

Minimun	1st Quartile	Median	Mean	3st Quartile	Maximun
0.19	16.98	77.32	588.82	299.15	7771.649

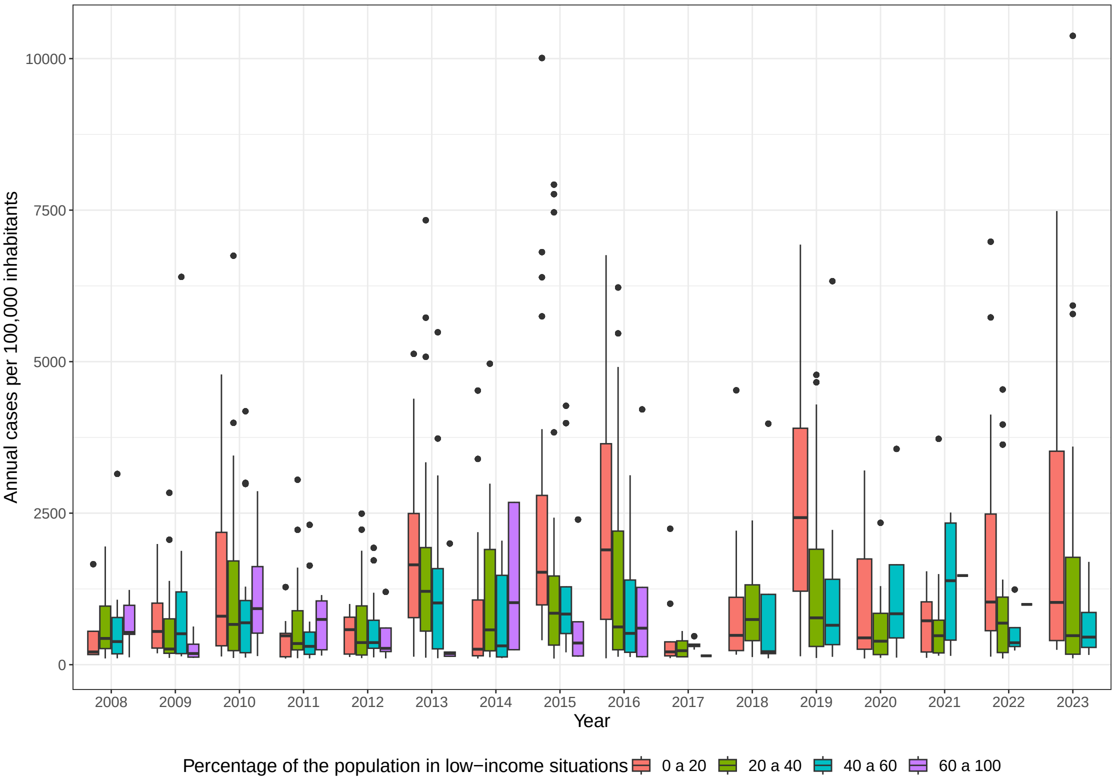


Figure 31 – Distribution of yearly dengue cases per 100,000 inhabitants based on percentage of population in low-income situations in municipalities during epidemic years (2008 to 2023)

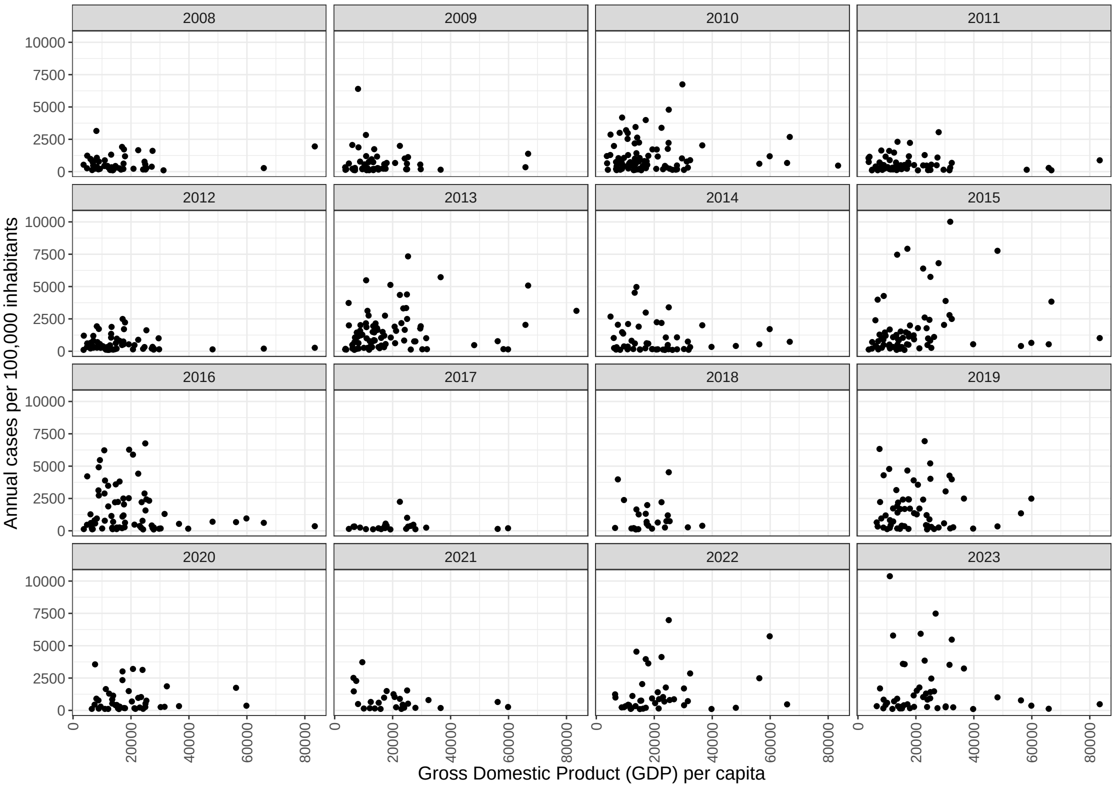


Figure 32 – Association between GDP per capita in Brazilian Real (R$) and yearly dengue cases per 100,000 inhabitants for municipalities in epidemic years (2008 to 2023)

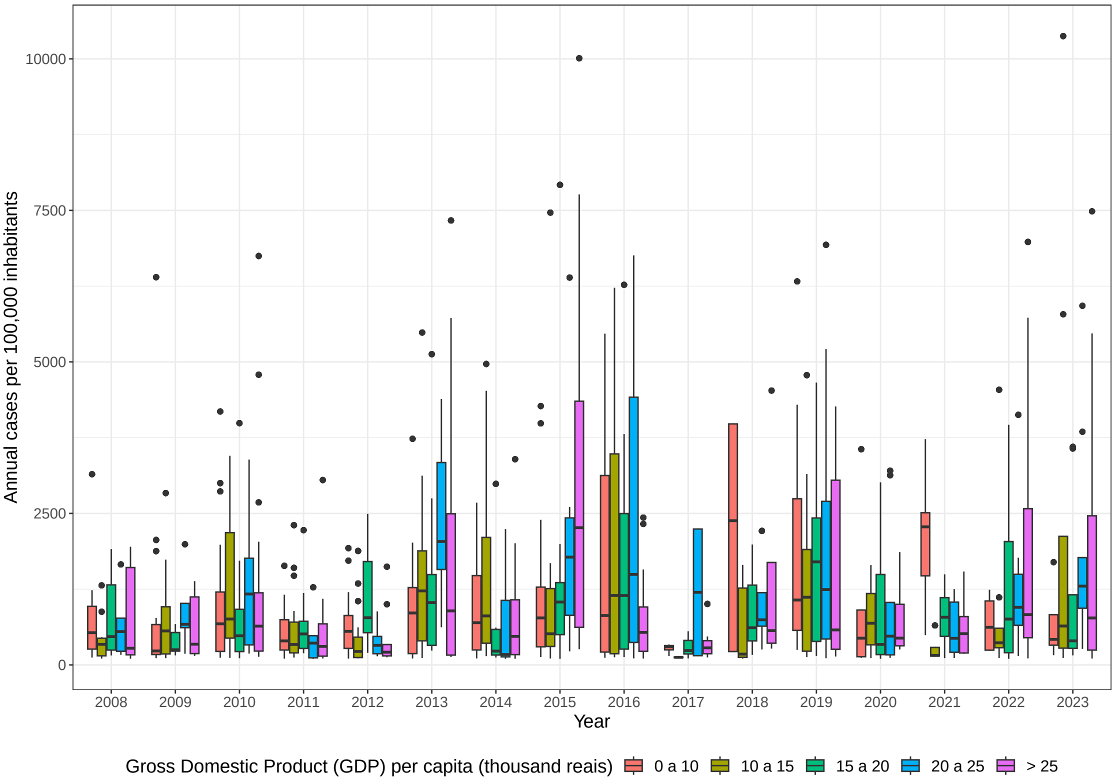


Figure 33 – Distribution of yearly dengue cases per 100,000 inhabitants based on GDP per capita (R$) in municipalities during epidemic years (2008 to 2023)

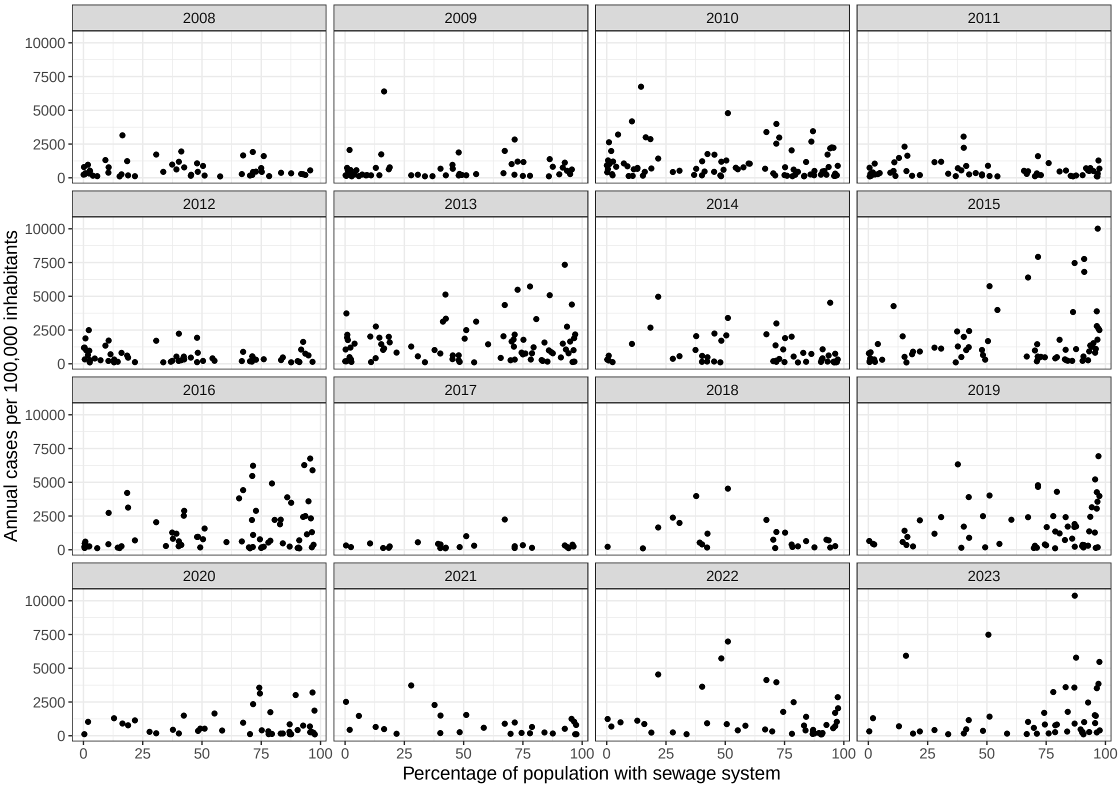


Figure 34 – Association between the percentage of population with sewage system and yearly dengue cases per 100,000 inhabitants for municipalities during epidemic years (2008 to 2023)

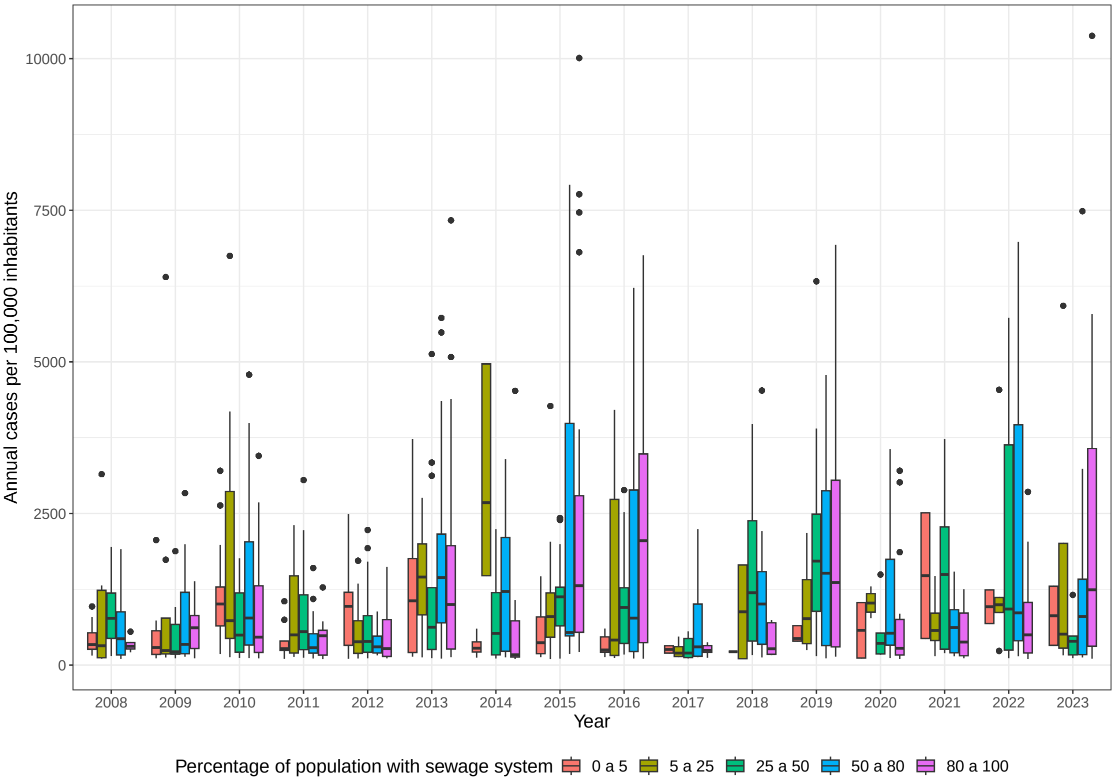


Figure 35 – Distribution of yearly dengue cases per 100,000 inhabitants based on percentage of population with sewage system in municipalities during epidemic years (2008 to 2023)

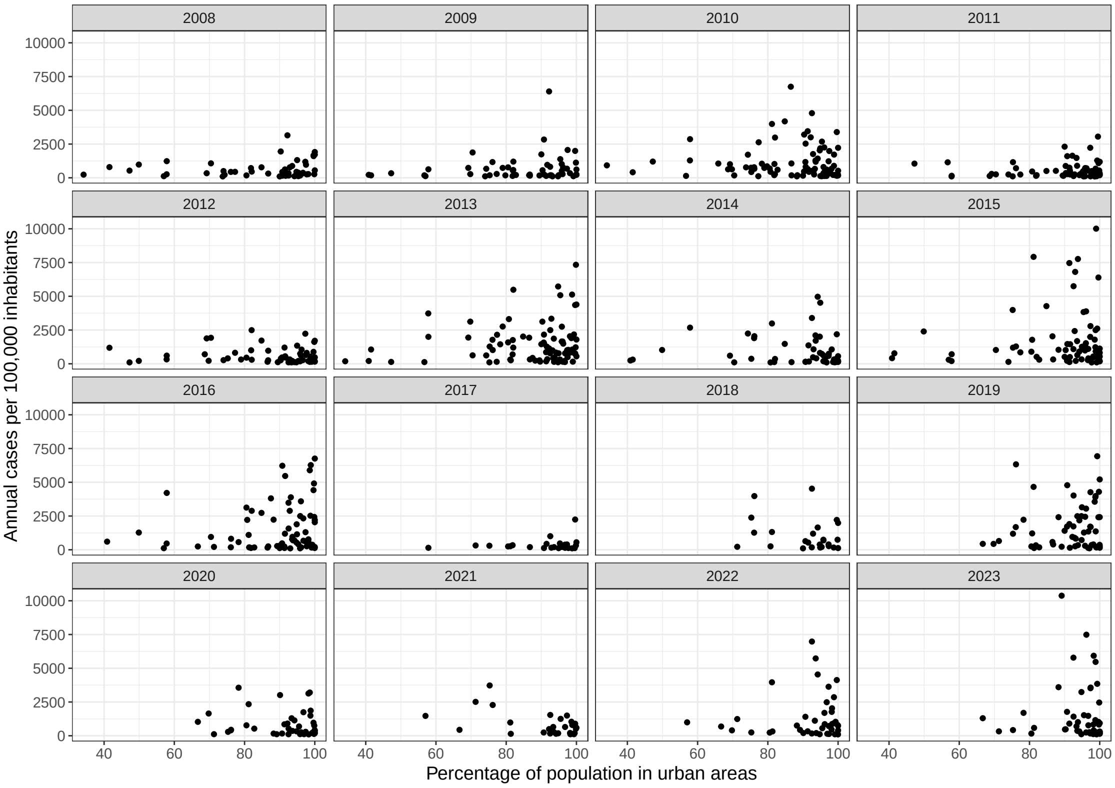


Figure 36 – Association between the percentage of population living in urban areas and yearly dengue cases per 100,000 inhabitants for municipalities during epidemic years (2008 to 2023)

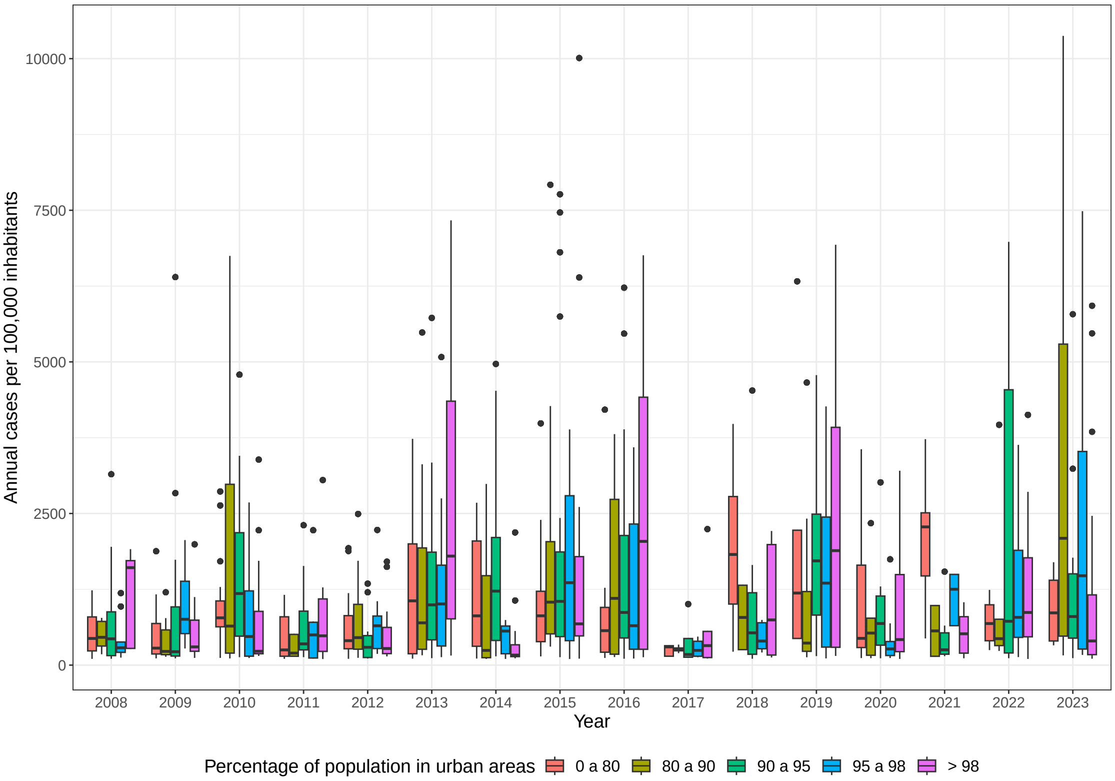


Figure 37 – Distribution of yearly dengue cases per 100,000 inhabitants based on percentage of population living in urban areas in municipalities during epidemic years (2008 to 2023)

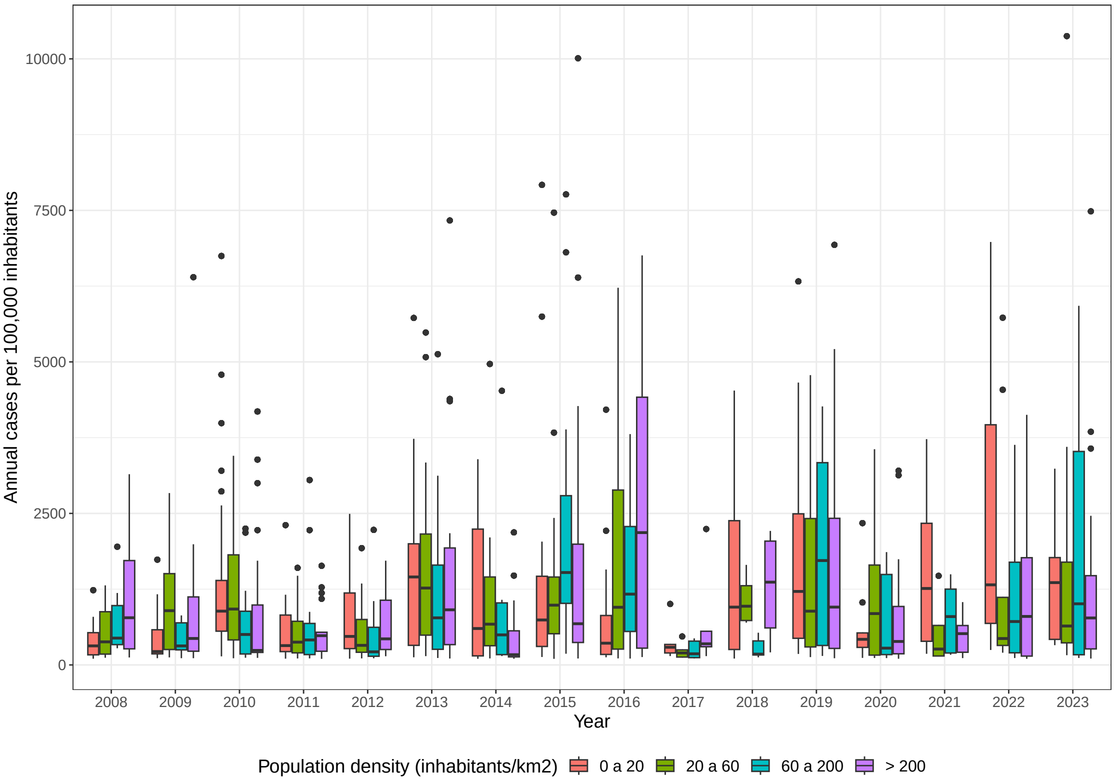


Figure 38 – Distribution of yearly dengue cases per 100,000 inhabitants based on demographic density in municipalities during epidemic years (2008 to 2023)