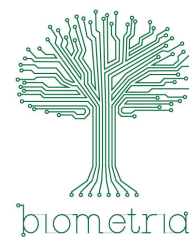




Universidade Estadual Paulista “Júlio de Mesquita Filho”
Instituto de Biociências – Câmpus de Botucatu
Programa de Pós-graduação em Biometria



Modelos espaço-temporais para mapeamento do risco de sífilis congênita e sua relação com as condições socioeconômicas e ambientais no município de São Paulo de 2010 a 2016.

Renato Ferreira da Cruz

Botucatu
2022

Renato Ferreira da Cruz

Modelos espaço-temporais para mapeamento do risco de sífilis congênita e sua relação com as condições socioeconômicas e ambientais no município de São Paulo de 2010 a 2016.

Tese de Doutorado apresentada ao Curso de Programa de Pós-graduação em Biometria da Universidade Estadual Paulista “Júlio de Mesquita Filho” como parte dos requisitos necessários para a obtenção do título de Doutor em Biometria.

Orientador: Profa. Dra. Liciane Vaz de Arruda Silveira

Coorientador: Prof. Dr. Thiago Santos Mota

Botucatu
2022

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. TRATAMENTO DA INFORM.
DIVISÃO TÉCNICA DE BIBLIOTECA E DOCUMENTAÇÃO - CÂMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: ROSEMEIRE APARECIDA VICENTE-CRB 8/5651

Cruz, Renato Ferreira da.

Modelos espaço-temporais para mapeamento do risco de sífilis congênita e sua relação com as condições socioeconômicas e ambientais no município de São Paulo de 2010 a 2016 / Renato Ferreira da Cruz. - Botucatu, 2022

Tese (doutorado) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências de Botucatu

Orientador: Liciane Vaz de Arruda Silveira

Coorientador: Thiago Santos Mota

Capes: 10203001

1. Sífilis congênita. 2. Fatores de riscos. 3. Análise espaço-temporal. 4. Biometria.

Palavras-chave: Distribuição espaço-temporal; GAM; INLA; Riscos relativos; Sífilis congênita.

RENATO FERREIRA DA CRUZ

**MODELOS ESPAÇO-TEMPORAIS PARA MAPEAMENTO DO RISCO DE
SÍFILIS CONGÊNITA E SUA RELAÇÃO COM AS CONDIÇÕES
SOCIOECONÔMICAS E AMBIENTAIS NO MUNICÍPIO DE SÃO PAULO
DE 2010 A 2016.**

Tese de Doutorado defendida e aprovada em Botucatu-SP, no dia 27 de Julho, pela banca examinadora constituída pelos professores:

Profa. Dra. Liciana Vaz de Arruda Silveira
Orientador

Profa. Dra. Lidia Raquel de Carvalho
Examinador

Profa. Dra. Maria del Pilar Diaz
Examinador

Profa. Dra. Maria Rita Donalísio Cordeiro
Examinador

Prof. Dra. Miriam Harumi Tsunemi
Examinador

Dedicatória

Dedico este trabalho aos meus pais Ilidia Gomes Ferreira e Francisco Ferreira da Cruz (*in memoriam*), que sempre foram meu exemplo máximo de superação, me dando a oportunidade de estudar, algo que eles jamais puderam ter. E à minha esposa e meus filhos.

Agradecimentos

Agradeço primeiramente a Deus pela sua infinita sabedoria e por me dar saúde e força para vencer todas as dificuldades da vida, me permitindo não desanimar frente às adversidades no caminho. À todos os meus familiares, especialmente meus pais Ilidia Gomes Ferreira e Francisco Ferreira da Cruz (*in memoriam*), que souberam entender a minha ausência não só ao longo desses quatro anos, mais desde a época de graduação e mestrado. À minha esposa e filha pela compreensão nos momentos sem minha presença.

À minha orientadora professora Dra. Lician Vaz de Arruda Silveira, pelo seu incentivo, por sua disponibilidade, paciência e apoio que sempre demonstrou ao longo desse quatro anos.

Ào meu coorientador professor Dr. Thiago Santos Mota, por todas as contribuições e sugestões que ajudaram na finalização do meu trabalho.

Às professoras da banca examinadora: Dra. Lidia Raquel de Carvalho, Dra. Maria del Pilar Diaz, Dra. Maria Rita Donalísio Cordeiro e Dra. Miriam Harumi Tsunemi por ter aceito prontamente o convite e dedicarem parte de seu tempo nas correções, sugestões e contribuições fundamentais para o enriquecimento deste trabalho.

Ao professor Dr. Francisco Chiaravalloti Neto da Faculdade de Saúde Pública da USP, por fornecer o banco de dados utilizado no trabalho, pela colaboração na leitura do trabalho e pelo suporte em sanar dúvidas do banco de dados.

Aos colegas de pós-graduação, pelo companheirismo e amizade, em especial à Agda Jessica de Freitas Galletti, Elizabete de Jesus Pinto, Lucas Ragiotto e Tiago Pereira marques por todas as ajudas com os conceitos de estatística e uso do software R.

A todos os professores do curso de doutorado em Biometria e demais funcionários do Departamento de Bioestatística, Biologia Vegetal, Parasitologia e Zoologia, que direta ou indiretamente contribuíram nesta etapa da minha formação.

À todos que, com um pensamento positivo, uma palavra amiga, alimentaram meus sonhos e contribuíram para esta grande conquista da minha vida.

“Deus é uno. Ele não está jamais, como pensam alguns, fora do mundo, mas sim, totalmente no mundo inteiro. Deus está no universo e o Universo está em Deus. O mundo e Deus não são mais que uma unidade”.

Pitágoras

Resumo

A sífilis congênita é uma doença de transmissão vertical causada pela espiroqueta *Treponema pallidum*. Apesar de curável, exclusiva do ser humano e com tratamento de baixo custo, ainda é considerada um dos maiores problemas de Saúde Pública do Mundo. O objetivo principal desta tese foi analisar a distribuição espaço-temporal do risco de sífilis congênita e sua associação com fatores socioeconômicos e ambientais no município de São Paulo-SP, no período de 2010 a 2016. A justificativa desse estudo se deve em virtude do aumento do número de casos diagnosticados de sífilis congênita nos últimos anos e das consequências na vida das crianças que nascem com essa doença. Os casos notificados de sífilis congênita foram obtidos do Sistema de Informação de Agravos de Notificação (SINAN), os de nascimento, no Sistema de Informações sobre Nascidos Vivos (SINASC) e as variáveis socioeconômicas e ambientais foram do Censo Demográfico de 2010 do Instituto Brasileiro de Geografia e Estatística (IBGE). Na modelagem foram utilizadas duas abordagens para estimar os riscos relativos espaço-temporais agregando as localizações das residências dos indivíduos que pertenciam à mesma área de ponderação. Na primeira abordagem, bayesiana, ajustou-se um modelo bayesiano empregando o método de Aproximação de Laplace Aninhada e Integrada (INLA) agregando os casos individuais que pertenciam à mesma área de ponderação. Na segunda abordagem, frequentista, utilizamos Modelos Aditivos Generalizados (GAM), considerando uma função de suavização com os centroides das áreas de ponderação (longitude e latitude) e o tempo. Como resultado, obtiveram-se os mapas de risco que apontaram as áreas que apresentaram maior e menor risco. Além disso, obtiveram-se os fatores de risco associados à sífilis congênita. No modelo INLA mostraram-se associadas positivamente à ocorrência desse agravo nas áreas de ponderação, o índice de Gini e a proporção de mulheres de 18 a 24 anos sem instrução ou com ensino fundamental incompleto e mostraram-se associadas negativamente, a proporção de mulheres em idade fértil e renda média per capita. No modelo GAM, associaram-se positivamente ao agravo a proporção de mulheres de 18 a 24 anos sem instrução ou com ensino fundamental incompleto e o índice de Gini e associou-se negativamente, apenas a renda média per capita. Os mapas e os fatores de risco estimados auxiliam possíveis ações por parte dos gestores de saúde e monitoramento da doença no município. Os resultados evidenciaram áreas com maior risco de sífilis congênita que persistem ao longo do tempo. O avanço de políticas sociais e proteções nessas áreas poderá contribuir para atingir a meta da OMS de eliminar a sífilis congênita.

Palavras-chave: Sífilis congênita, distribuição espaço-temporal, riscos relativos, INLA, GAM.

Abstract

Congenital syphilis is a vertically transmitted disease caused by the spirochete *Treponema pallidum*. Despite being curable, exclusive to humans and with low cost treatment, it is still considered one of the biggest public health problems in the world. The main objective of this thesis was to analyze the spatio-temporal distribution of the risk of congenital syphilis and its association with socioeconomic and environmental factors in the city of São Paulo-SP, from 2010 to 2016. The justification for this study is due to the increase in the number of diagnosed cases of congenital syphilis in recent years and the consequences in the lives of children born with this disease. The notified cases of congenital syphilis were obtained from the Information System of Notifiable Diseases (SINAN), those of birth from the Information System on Live Births (SINASC) and the socioeconomic and environmental variables were obtained from the 2010 Demographic Census of the Brazilian Institute of Geography and Statistics (IBGE). In the modeling, two approaches were used to estimate the relative spatio-temporal risks by aggregating the locations of the residences of individuals who belonged to the same weighting area. In the first, Bayesian approach, a Bayesian model was fitted using the Nested and Integrated Laplace Approximation (INLA) method, aggregating the individual cases that belonged to the same weighting area. In the second, frequentist approach, Generalized Additive Models (GAM) we use, considering a smoothing function with the centroids of the weighting areas (longitude and latitude) and time. As a result, risk maps were obtained that indicated the areas that presented higher and lower risk. In addition, the risk factors associated with congenital syphilis were obtained. In the INLA model, showed themselves positively associated with the occurrence of this condition in the weighted areas, the Gini index and the proportion of women aged 18 to 24 years without education or with incomplete primary education were positively associated with the occurrence of this condition, and the proportion of women in childbearing age and average per capita income. In the GAM model, the proportion of women aged 18 to 24 years without education or with incomplete primary education and the Gini index were positively associated with the disease, and only the average per capita income was negatively associated. The maps and estimated risk factors obtained help possible actions by health managers and disease monitoring in the municipality. The results showed areas with a higher risk of congenital syphilis that persist over time. The advancement of social policies and protections in these areas could contribute to achieving the WHO goal of eliminating congenital syphilis.

Keywords: Congenital Syphilis, spatio-temporal distribution, relative risks, INLA, GAM.

Lista de figuras

Figura 1 – Índices para os 96 distritos do município de São Paulo, Brasil.	28
Figura 2 – Representação simbólica do modelo de efeito principal. Os círculos representam a independência a priori, os arredondados representam a dependência a priori. As observações no espaço-tempo são indicadas por pontos pequenos e azuis.	38
Figura 3 – Representação gráfica para diferentes estruturas de interação espaço-temporal (círculos representam independência, retângulos arredondados representam dependência).	39
Figura 4 – Matriz de estrutura R_u , baseada no movimento de contiguidade espacial queen sobre um mapa de cinco áreas, e matriz de estrutura R_γ , associada ao modelo AR1 em cinco pontos de tempo. Fonte: Baseado em Haining e Li (2020).	42
Figura 5 – Localização geográfica do município São Paulo-SP, Brasil.	49
Figura 6 – Casos de sífilis congênita no período de 2010 a 2016.	58
Figura 7 – (a) Distribuição de frequência dos casos de sífilis congênita no período de 2010 a 2016. (b) Distribuição geográfica dos casos notificados de sífilis congênita no município de São Paulo de 2010 a 2016.	59
Figura 8 – Boxplots de casos observados de sífilis congênita (a) e número de nascidos vivos (b) no município de São Paulo de 2010 a 2016 por AP.	59
Figura 9 – Distribuição da incidência de sífilis congênita (número de casos por 1000 nascidos vivos) por área de ponderação no município de São Paulo de 2010 a 2016.	60
Figura 10 – Gráfico de dispersão exibindo as relações entre cada covariável e o SIR. Obtido com a função ggpairs do pacote GGally (SCHLOERKE; CROWLEY; COOK, 2018)	62
Figura 11 – Razão de incidência padronizada (y_{it}/E_{it}) para SC no município de São Paulo de 2010 a 2016.	63
Figura 12 – Critérios DIC e WAIC para escolha dos melhores modelos.	67
Figura 13 – Gráficos de diagnóstico para o modelo de Poisson-P29.	68
Figura 14 – Gráficos de diagnóstico para o modelo Binomial Negativa-NB24.	68
Figura 15 – Tendência temporal a posteriori de casos de sífilis congênita no município de São Paulo (2010-2016): efeito temporal não estruturado $\exp(\phi_t)$, efeito temporal estruturado $\exp(\gamma_t)$ e efeito temporal total $\exp(\phi_t + \gamma_t)$	71
Figura 16 – Mapa das oito áreas de ponderação selecionadas para a observação da tendência temporal entre diferentes áreas de acordo com os bairros do Município de São Paulo.	72

Figura 17 – Tendências temporais a posteriori de casos de sífilis congênita no município de São Paulo (2010-2016) específicas para oito áreas de ponderação selecionadas: efeito temporal $\exp(\delta_{it})$	73
Figura 18 – Médias a posterioris dos RRs espaço-temporais para SC no período de 2010 a 2016 no município de São Paulo-SP.	74
Figura 19 – Probabilidades a posteriori de que os RRs sejam maiores que 1 ($P(\Psi_{it} > 1 \mathbf{y})$) no período de 2010 a 2016 no município de São Paulo.	75
Figura 20 – Distribuição do Índice Socioeconômico do Contexto Geográfico para Saúde e Estudos Sociais (GeoSES) segundo a área de ponderação, município de São Paulo, 2010. Para a obtenção do mapa usamos os dados obtidos em https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0232074#sec015	76
Figura 21 – (a) Valores ajustados versus valores observados dos casos notificados de sífilis congênita no município de São Paulo de 2010 a 2016. (b) Histograma dos valores de PIT ajustados.	77
Figura 22 – Distribuição gerada por meio de 10000 simulações para os resíduos do modelo final.	79
Figura 23 – Gráficos de dispersão de Moran mostrando a relação entre os resíduos e o valor médio de seus vizinhos no período de 2010 à 2016.	80
Figura 24 – Resíduos padronizados por áreas de ponderação do município de São Paulo no período de 2010 a 2016.	81
Figura 25 – Gráfico dos resíduos padronizados versus valores ajustados.	82
Figura 26 – Gráficos da densidade a posteriori das estimativas de efeito fixo em médias logarítmicas do modelo final.	83
Figura 27 – Verificação dos pressupostos básicos do modelo P6 e análise de resíduos usando a metodologia GAM.	101
Figura 28 – Verificação dos pressupostos básicos do modelo NB6 e análise de resíduos usando a metodologia GAM.	102
Figura 29 – Gráficos de diagnóstico para o modelo final após seleção de covariáveis usando a metodologia GAM.	104
Figura 30 – Estimativas do efeito temporal para SC no período de 2010 a 2016 no município de São Paulo-SP. A linha contínua é o valor predito da média em função do tempo. As linhas tracejadas superior e inferior são o intervalo de confiança de 95%.	105
Figura 31 – Estimativas dos riscos relativos espaço-temporais para SC no período de 2010 a 2016 no município de São Paulo-SP.	106
Figura 32 – Gráfico dos resíduos de deviance versus valores ajustados.	107
Figura 33 – Distribuição gerada por meio de 10000 simulações para os resíduos do modelo final	108

Figura 34 – Resíduos de deviance por áreas de ponderação do município de São Paulo no período de 2010 a 2016.	109
Figura 35 – Intervalos de confiança de 95% para estimativas dos efeitos fixos usando a abordagem GAM.	116
Figura 36 – Intervalos de credibilidade de 95% para estimativas dos efeitos fixos usando a abordagem INLA.	116
Figura 37 – Estimativas do efeito temporal para SC no período de 2010 a 2016 no município de São Paulo-SP. A linha contínua é o valor predito da média em função do tempo. As faixas azuis superior e inferior são o intervalo de confiança de 95%.	118
Figura 38 – Estimativas do efeito temporal para SC no período de 2010 a 2016 no município de São Paulo-SP. A linha contínua é o valor predito da média em função do tempo. As faixas azuis superior e inferior são o intervalo de credibilidade de 95%.	118
Figura 39 – (a) Valores ajustados versus valores observados dos casos notificados de sífilis congênita no município de São Paulo de 2010 a 2016 segundo o modelo INLA. (b) Valores ajustados versus valores segundo o modelo GAM.	118
Figura 40 – Diagrama de dispersão entre o efeito espaço-temporal obtido pelas abordagens INLA e GAM.	119
Figura 41 – As quatro áreas selecionadas (esquerda) e a matriz de adjacência associada (direita).	143

Lista de tabelas

Tabela 1	– Número de casos de sífilis congênita em menores de um ano de idade e taxa de incidência (por 1.000 nascidos vivos) por ano de diagnóstico. Brasil, 2010-2019.	6
Tabela 2	– Restrições de identificabilidade para os modelos espaço-temporais na Equação 2.36	43
Tabela 3	– Estatísticas descritivas para o número de casos de sífilis congênita e taxa de incidência por 1000 nascidos vivos em nível de áreas de ponderação no município de São Paulo-2010 a 2016	58
Tabela 4	– Resumo da taxa de incidência por 1000 NV de SC específica do período analisado	60
Tabela 5	– Avaliação da multicolinearidade para as variáveis socioeconômica e ambientais	61
Tabela 6	– Comparação dos 36 modelos espaço-temporais ajustados para os casos de SC no município de São Paulo considerando a distribuição de Poisson. . . .	65
Tabela 7	– Comparação dos 36 modelos espaço-temporais ajustados para os casos de SC no município de São Paulo considerando a distribuição Binomial Negativa.	66
Tabela 8	– Critério para escolha do melhor modelo.	67
Tabela 9	– Estimativas a posterioris dos riscos relativos dos efeitos fixos e aleatórios e intervalos de credibilidade de 95% para ocorrência de sífilis congênita no modelo de Poisson (P29) espaço-temporal com covariáveis no município de São Paulo (2010-2016).	69
Tabela 10	– Estimativas a posterioris dos riscos relativos dos efeitos fixos e aleatórios e intervalos de credibilidade de 95% para ocorrência de sífilis congênita no modelo de Poisson espaço-temporal (P29) após seleção de covariáveis, município de São Paulo (2010-2016).	70
Tabela 11	– Análise de autocorrelação espacial residual anual para o modelo selecionado.	78
Tabela 12	– Variáveis socioeconômicas e ambientais utilizadas como potenciais variáveis explicativas no modelo GAM para a distribuição espaço-temporal de sífilis congênita.	94
Tabela 13	– Modelos ajustados, critério de informação de Akaike (AIC), critério de informação Bayesiano (BIC), logaritmo da verossimilhança marginal restrita (REML), raiz quadrada do erro quadrático médio (RMSE), erro absoluto médio (MAE), coeficiente de determinação (R^2) e deviance explicada (DE). Os melhores modelos são destacados em negrito.	99
Tabela 14	– Critério para escolha do melhor modelo.	100
Tabela 15	– Resumo das estimativas do modelo selecionado NB6 com todas as covariáveis para os efeitos fixos e termos suavizados.	100

Tabela 16 – Estimativas dos efeitos fixos e termos suavizados com intervalos de confiança de 95% para ocorrência de sífilis congênita no modelo Binomial Negativa espaço-temporal após seleção de covariáveis, município de São Paulo (2010-2016).	103
Tabela 17 – Resultados da execução da função <code>k.check()</code> no modelo final.	104
Tabela 18 – Análise de autocorrelação espacial residual anual para o modelo selecionado.	105
Tabela 19 – Estimativas dos modelos para cada covariável na presença dos efeitos espaciais e temporais, considerando as abordagens INLA e GAM.	115
Tabela 20 – Medidas de concordância entre os modelos obtidos pelas abordagens INLA e GAM e os dados observados	119
Tabela 21 – Disposição dos dados para análise com o INLA	139

Lista de abreviaturas e siglas

AIC	<i>Akaike Information Criterion</i>
AP	Áreas de Ponderação
AR(p)	<i>Autoregressive Model Of Order p</i>
BIC	<i>Bayesian Information Criterion</i>
BYM	Besag, York e Mollié
CAR	<i>Conditional Auto-Regressive</i>
CPO	<i>Conditional Predictive Ordinate</i>
DAG	<i>Directed Acyclic Graph</i>
DIC	<i>Deviance Information Criterion</i>
GAM	<i>Generalized Additive Models</i>
GLM	<i>Generalized Linear Models</i>
GRMF	<i>Gauss-Markov Random Field</i>
HIV	<i>Human Immunodeficiency Virus</i>
IBGE	Instituto Brasileiro de Geografia e Estatística
ICAR	<i>Intrinsic Conditional Auto-Regressive</i>
INLA	<i>Integrated Nested Laplace Approximations</i>
IST	Infecções Sexualmente Transmissíveis
LMPL	<i>Logarithm Of The Pseudomarginal Likelihood</i>
MAE	<i>Mean Absolute Error</i>
MAUP	<i>Modifiable Areal Unit Problem</i>
MCMC	<i>Markov Chain Monte Carlo</i>
NV	Nascidos Vivos
OMS	Organização Mundial da Saúde
OPAS	Organização Pan-Americana da Saúde

PIT	<i>Probability Integral Transformation</i>
REML	<i>Restricted Maximum Likelihood</i>
RMSE	<i>Root Mean Square Error</i>
RR	Risco Relativo
RW1	<i>Random Walk Model Of Order 1</i>
RW2	<i>Random Walk Model Of Order 2</i>
SC	Sífilis Congênita
SINAN	Sistema de Informação de Agravos de Notificação
SINASC	Sistema de Informações Sobre Nascidos Vivos
SIR	<i>Standardized Incidence Ratio</i>
UNICEF	<i>United Nations International Children's Emergency Fund</i>
VIF	<i>Variance Inflation Factor</i>
WAIC	<i>Watanabe-Akaike Information Criterion</i>

Sumário

1	INTRODUÇÃO	1
2	REVISÃO DE LITERATURA	11
2.1	Mapeamento de Doenças	11
2.2	Modelos lineares generalizados	12
2.2.1	A Família Exponencial de Distribuições	13
2.2.2	Modelo regressão de Poisson	14
2.2.3	Modelo binomial negativa	14
2.3	Modelos aditivos generalizados	15
2.3.1	Modelos de regressão spline de suavização	15
2.4	Estimação bruta de risco de sífilis congênita	17
2.4.1	Modelos espaciais para mapeamento de do- enças	17
2.4.2	Estimação para dados espaciais	18
2.5	Inferência Bayesiana	19
2.6	Dados espaciais pela metodologia INLA	21
2.7	Modelos autorregressivos de ordem 1	25
2.8	Modelos para dados de área	28
2.9	Modelos autorregressivos condicionais (CAR)	29
2.9.1	O modelo ICAR usando uma matriz de pe- sos espaciais com entradas binárias	31
2.10	Modelos espaço-temporais	32
2.10.1	Estimação para dados espaço-temporais	33
2.10.2	Estruturas espaciais e temporal	33
2.10.3	Modelo de Bernardinelli	36
2.10.4	Modelos de interação espaço-tempo	37
2.11	Prioris PC	43
3	UM ESTUDO ESPAÇO-TEMPORAL DA SÍFILIS CONGÊNITA NO MUNICÍPIO DE SÃO PAULO NO PERÍODO DE 2010 A 2016 USANDO O INLA	45
3.1	Introdução	46
3.2	Materiais e Métodos	49
3.2.1	Desenho e área do estudo	49
3.2.2	Fonte dos dados e variáveis do estudo	50
3.2.3	Multicolinearidade	52

	3.2.4	Estimação do risco relativo para dados agregados	52
	3.2.5	Métodos de avaliação e qualidade de ajuste do modelo	53
	3.2.5.1	Complexidade e adequação do modelo	54
	3.2.5.2	Qualidade preditiva do modelo . . .	54
	3.2.5.3	Precisão e qualidade de ajuste do modelo	55
	3.2.5.4	Transformação integral de probabilidade	56
	3.2.6	Escolha das prioris	56
	3.2.7	Software utilizado	57
3.3	Resultados		57
	3.3.1	Análise exploratória	57
	3.3.2	Análise espaço-temporal	63
	3.3.3	Distribuição espaço-temporal dos riscos relativos de SC	70
	3.3.4	Desempenho preditivo	76
3.4	Discussão		83
3.5	Conclusão		89
4	MODELAGEM DE DADOS ESPAÇO-TEMPORAIS DA SÍFILIS CONGÊNITA NO MUNICÍPIO DE SÃO PAULO DE 2010 A 2016 USANDO MODELOS ADITIVOS GENERALIZADOS		91
	4.1	Introdução	92
	4.2	Materiais e Métodos	93
	4.2.1	Multicolinearidade	93
	4.2.2	Análises espaciais e temporais	95
	4.2.3	Avaliação e qualidade de ajuste	96
	4.2.3.1	Critério de informação de Akaike .	96
	4.2.3.2	Critério de informação Bayesiano .	96
	4.2.3.3	Precisão e qualidade de ajuste do modelo	96
	4.2.4	Análise de resíduos	97
	4.3	Resultados	98
	4.4	Discussão	109
	4.5	Conclusão	113
5	COMPARAÇÃO ENTRE AS ABORDAGENS INLA E GAM		115

6	CONSIDERAÇÕES FINAIS	120
	Referências	122
	Apêndices	138
	APÊNDICE A – DISPOSIÇÃO DOS DADOS	139
	APÊNDICE B – PRODUTO DE KRONECKER	140
	APÊNDICE C – CAMPOS ALEATÓRIOS	141
	C.1 Campos aleatórios gaussianos	141
	C.2 Campos aleatórios gaussianos Markovianos	142
	APÊNDICE D – <i>SCRIPT</i> EM LINGUAGEM R PARA OBTENÇÃO DAS VARIÁVEIS DO IBGE	144
	APÊNDICE E – <i>SCRIPT</i> EM LINGUAGEM R PARA OBTER O VIF	150
	APÊNDICE F – <i>SCRIPT</i> EM LINGUAGEM R UTILIZADO NAS ANÁLISES DO CAPÍTULO 3	152
	F.1 Modelos com estrutura espacial BYM	152
	F.2 Modelos com estrutura espacial BYM2	158
	F.3 Modelos com estrutura espacial de Leroux	158
	F.4 Critérios de escolha do melhor modelo, estimativas, resíduos e mapas de risco	158
	APÊNDICE G – <i>SCRIPT</i> EM LINGUAGEM R UTILIZADO NAS ANÁLISES DO CAPÍTULO 4	172

1 Introdução

O mapeamento espacial e espaço-temporal de doenças são abordagens amplamente usadas na análise de dados de vigilância de doenças. O controle de doenças infecciosas requer respostas eficazes e imediatas a aumentos imprevistos na carga de doenças, ou seja, surtos/epidemias. A incapacidade de produzir estimativas oportunas e precisas da carga de doenças infecciosas pode afetar significativamente os programas de controle de saúde pública.

Nesse tipo de estudo, geralmente as unidades de análise consideradas são as menores possíveis, podendo levar aos chamados problemas de estimativa de pequenas áreas. Como consequência, várias propostas de modelagem foram formuladas para lidarem com esse problema e, assim, obter estimativas de riscos confiáveis. A maioria desses modelos considera a dependência entre unidades espaciais próximas, assumindo que elas apresentam riscos semelhantes. Portanto, a hipótese de dependência espacial é a principal chave para melhorar as estimativas de riscos em estudos de mapeamento de doenças (CORPAS-BURGOS; MARTINEZ-BENEITO, 2020). A autocorrelação espacial pode surgir devido aos efeitos de vizinhança, em que os comportamentos dos indivíduos em uma determinada área são influenciados por indivíduos que vivem em áreas adjacentes, e aos efeitos de agrupamento, onde grupos de pessoas com comportamentos semelhantes escolhem viver próximos (NAEEIM; RAHMAN; FAHIMI, 2020).

O modelo mais simples de mapeamento espacial de doenças, e por isso o mais popular, foi proposto por Besag et al (BESAG; YORK; MOLLIÉ, 1991) e desenvolvido posteriormente por vários outros pesquisadores (BERNARDINELLI et al., 1995; ELLIOT; CUZICK, 1992). Ao adicionar termos para uma tendência linear (SCHRÖDLE; HELD, 2011a; ASSUNCAO; REIS; OLIVEIRA, 2001) ou não paramétrica nas interações tempo e espaço-tempo (SCHRÖDLE; HELD, 2011a; KNORR-HELD, 2000; SCHMID; HELD, 2004), o modelo base foi estendido para uso no caso espaço-temporal (SCHRÖDLE; HELD, 2011a). Modelos de regressão ecológica têm sido frequentemente usados em epidemiologia para capturar a associação de risco entre doença e covariáveis relevantes (WAKEFIELD, 2007).

Normalmente, os modelos de mapeamento de doenças incluem modelos autorregressivos condicionais (CAR, do inglês *conditional autorregressive*) para os efeitos espaciais e prioris RW1 (passeio aleatório de ordem 1), RW2 (passeio aleatório de ordem 2) ou AR1 (autorregressivo de ordem 1) para os efeitos temporais. De modo geral, efeitos de interação também são incluídos e podem ser completamente desestruturados, ou espacialmente e/ou temporalmente estruturados

(KNORR-HELD, 2000).

Os efeitos espaciais têm sido amplamente utilizados, pois áreas contíguas tendem a apresentar valores semelhantes em um determinado período de tempo. Analogamente, os efeitos temporais são reconhecidos para uma determinada área, pois períodos de tempo anteriores tendem a apresentar valores semelhantes. Seguindo uma estrutura clássica de Knorr-Held (2000), os efeitos espaciais/temporais são divididos em duas partes: o efeito estruturado e o efeito não estruturado. O efeito estruturado indica que uma estrutura de variância-covariância predefinida de efeitos espaciais/temporais é usada para restringir a estimativa dos hiperparâmetros, enquanto o efeito não estruturado significa que a variância e a covariância dos efeitos espaciais/temporais são todas estimadas livremente a partir dos dados.

Para efeito de comparação consideramos uma segunda abordagem, frequentista, os modelos lineares aditivos generalizados (GAM, do inglês *generalized additive models*), no qual a variável resposta depende linearmente de funções suavizadas desconhecidas de algumas variáveis preditoras, e o interesse se concentra na inferência sobre essas funções suavizadas. As relações entre as preditoras individuais e a variável dependente seguem padrões suavizados que podem ser lineares ou não lineares.

Uma característica importante dos modelos GAM é a capacidade de controlar a suavidade das funções das preditoras, podendo evitar funções preditivas sem sentido, simplesmente ajustando o nível de suavidade. Em outras palavras, podemos impor a crença a priori de que as relações preditivas são inerentemente suaves por natureza, mesmo que o conjunto de dados possa sugerir uma relação mais ruidosa. Isso desempenha um papel importante na interpretação do modelo, bem como na credibilidade dos resultados. Aqui utilizamos uma função de suavização nas localizações espaciais dos centroides das áreas de ponderação (longitude e latitude) e no tempo.

Como aplicação desses modelos, consideramos os dados de Sífilis Congênita (SC) para o município de São Paulo. O sistema de vigilância da SC é uma importante fonte de dados para monitorar e indicar riscos aos quais as pessoas estão sujeitas, contribuindo assim, para a identificação da realidade epidemiológica de determinada área geográfica. No Brasil, dois grandes sistemas de vigilância, o Sistema de Informações sobre Nascidos Vivos (SINASC) e o Sistema de Informação de Agravos de Notificação (SINAN), têm sido amplamente utilizados na pesquisa de SC e auxiliam no planejamento da saúde, definição de prioridades de intervenção, além de permitir que seja avaliado o impacto das intervenções (BRASIL, 2020c; BRASIL, 2020d).

As Infecções Sexualmente Transmissíveis (IST) são caracterizadas por infecções causadas por mais de 30 diferentes agentes etiológicos (bactérias, vírus, fungos e protozoários), sendo transmitidas prioritariamente por contato sexual sem uso de preservativos, embora, eventualmente, também possam ser transmitidas por contato sanguíneo, e da mãe para a criança durante a gestação, no parto ou na amamentação (BRASIL, 2015b). Segundo estimativas da Organização Mundial da Saúde (OMS), em todo o mundo mais de 1 milhão de pessoas adquirem diariamente IST e em 2016 foi lançada uma estratégia global do setor de saúde sobre doenças sexualmente

transmissíveis 2016–2021, com o objetivo de eliminar as epidemias de IST como uma grande preocupação de saúde pública. A estratégia prevê que, até 2030, as taxas de sífilis congênita serão reduzidas para menos de 50 casos por 100.000 nascidos vivos em 80% dos países; e a incidência de infecções por *Treponema pallidum* (sífilis) e *Neisseria gonorrhoeae* (gonorréia) terá caído 90% mundialmente entre 2018 e 2030 (WHO, 2018).

A sífilis é uma infecção bacteriana crônica, curável e exclusiva do ser humano, causada pelo espiroqueta *Treponema pallidum*, subespécie *pallidum*, que é endêmica em países de baixa renda e ocorre em taxas mais baixas em países de renda média e alta, sendo reconhecida por profissionais de saúde há centenas de anos (HOOK, 2017; PEREIRA et al., 2021). A infecção por sífilis aumenta o risco de contrair HIV em populações de alto risco. O diagnóstico imediato, o tratamento oportuno e as intervenções preventivas contra a infecção por sífilis seriam um investimento valioso para a redução da incidência do HIV (WU et al., 2020). Segundo Karp et al. (2009), a infecção resulta na formação de lesões que ocorrem principalmente nos órgãos genitais, podendo facilitar a transmissão do vírus da imunodeficiência humana (HIV) e que o risco de contrair o HIV é aumentado em até 4 vezes mais.

A presença de sinais e sintomas da sífilis é muito variável e complexa, e quando não tratada, pode evoluir para formas mais graves, podendo comprometer o sistema nervoso, o aparelho cardiovascular, o aparelho respiratório e o aparelho gastrointestinal.

O primeiro registro de sífilis ocorreu na Europa em meados da década de 1490 (BARNETT, 2018). É uma das infecções sexualmente transmissíveis mais comuns em todo o mundo, com aproximadamente 6 milhões de novos casos a cada ano e altas taxas de incidência em países da América Latina, África e Ásia (LAFETÁ et al., 2016). A sífilis pode ser denominada de sífilis adquirida, sífilis gestacional ou sífilis congênita. A sífilis adquirida é transmitida através de contato sexual ou transfusão sanguínea. A gestacional ocorre quando a bactéria é detectada na gestante, enquanto a congênita é quando a bactéria é transmitida da mãe para o feto (BRASIL, 2020b).

A sífilis congênita pode ser transmitida aos filhos durante o desenvolvimento fetal ou no nascimento por mulheres com diagnóstico de sífilis, e pode causar aborto espontâneo, natimorto ou morte logo após o parto se não receber um tratamento precoce e eficaz. Até 40% dos bebês nascidos de mulheres com sífilis não tratada podem ser natimortos ou morrer de infecção quando recém-nascidos (CONTROL; PREVENTION, 2021). Os bebês nascidos com sífilis congênita podem apresentar:

- Ossos deformados
- Anemia grave (baixa contagem de hemácias sangue)
- Linfadenopatia (aumento dos gânglios)
- Hepatomegalia (aumento do fígado)

- Esplenomegalia (aumento do baço)
- Icterícia (amarelecimento da pele ou dos olhos)
- Problemas cerebrais e nervosos, como cegueira ou surdez
- Meningite e rinite
- Erupções cutâneas

A triagem precoce e o tratamento de mulheres grávidas com sífilis são abordagens preventivas e apresentam redução comprovada no número de casos de sífilis congênita, mas essa doença continua sendo um grande problema de saúde pública, e com grande impacto social, a qual os profissionais de saúde devem estar preparados para enfrentar. Importantes estudos sobre a sífilis congênita têm desempenhado papel fundamental na implementação de estratégias para sua prevenção (LAGO; VACCARI; FIORI, 2013).

A sífilis congênita ocorre devido à disseminação hematogênica do *Treponema pallidum* de uma gestante infectada ou tratada de forma inadequada para o feto e, quando as mães apresentam lesões genitais no canal do parto, pode ocorrer a transmissão direta da bactéria de mãe para o recém-nascido. Cerca de 50% das crianças infectadas são assintomáticas ao nascimento, com surgimento dos primeiros sintomas, geralmente, nos primeiros 3 meses de vida. Assim, é de suma importância a triagem sorológica no início da gravidez (primeiro trimestre), no início do terceiro trimestre e da mãe na maternidade durante a internação hospitalar por parto ou aborto, e o seguimento clínico e ambulatorial de todas as crianças até 18 meses de idade expostas à sífilis ou com sífilis congênita. (AVELLEIRA; BOTTINO, 2006; BRUNNER; SUDDARTH, 1987; BRASIL, 2005; FELIZ et al., 2016; DOMINGUES et al., 2021).

Segundo estimativa da OMS, houve mais de meio milhão (cerca de 661.000) de casos totais de sífilis congênita em 2016, resultando em mais de 200.000 natimortos e mortes neonatais. A sífilis congênita é a segunda principal causa de natimortalidade evitável em todo o mundo, precedida apenas pela malária (WHO, 2019). Na maioria dos países, é necessário a obtenção de estimativas precisas para reflexão do gravame da doença, pois seu diagnóstico é muito mais difícil do que da sífilis em gestantes (WIJESOORIYA et al., 2016).

No Brasil, a situação não é diferente de outros países. Os números de casos de sífilis adquirida, sífilis em gestantes e sífilis congênita são preocupantes e a infecção precisa ser controlada. De acordo com Cooper e Sánchez (2018), cada caso de sífilis congênita deve ser visto como uma falha no sistema de saúde pública em fornecer atendimento pré-natal ideal para mulheres grávidas, já que a sífilis congênita pode ser prevenida por exames sorológicos pré-natais nas gestantes e tratamento com penicilina de mulheres infectadas, seus parceiros sexuais e bebês recém-nascidos.

Em 2006, a Organização Pan-Americana da Saúde (OPAS) e o Fundo das Nações Unidas para a Infância (UNICEF, do inglês *United Nations International Children's Emergency Fund*)

apresentaram estratégias e ferramentas para apoiar na eliminação da transmissão materno infantil do vírus da imunodeficiência humana (HIV, do inglês *Human Immunodeficiency Virus*) e da sífilis congênita e colocaram como meta redução do HIV para 2% ou menos e redução da incidência de sífilis congênita (incluindo natimortos) para 0,5 casos por 1.000 nascidos vivos ou menos. Para isso, foram propostos cinco objetivos: melhorar o atendimento dos cuidados pré-natais e assistência obstétrica qualificada no momento do parto para 95% ou mais; aumentar para mais de 95% a cobertura de detecção de sífilis e HIV em mulheres grávidas, bem como a cobertura da profilaxia do HIV e tratamento da sífilis em mulheres grávidas e crianças; aumentar para mais de 95% os centros de atenção primária que prestam serviços de prevenção e diagnóstico de HIV e ISTs de forma integrada com outros serviços de saúde, incluindo centros de atenção pré-natal, serviços de saúde reprodutiva, serviços para adolescentes e de violência de gênero; aumentar para mais de 95% o monitoramento de países participantes, para avaliar os avanços na eliminação da transmissão materno infantil e da sífilis congênita e apoio na tomada de decisões (OPAS, 2010; OPAS, 2017).

No Brasil, estudos realizados no período entre 2010 e 2014 mostraram que, apesar das melhorias na saúde pública, ainda é alta a incidência de sífilis congênita, e assim distante das metas propostas pelo Ministério da Saúde, tornando-se um grande desafio a redução no número de casos, apesar do tratamento ser de baixo custo. Campanhas de conscientização podem ser instrumentos capazes de reduzir bruscamente o número de casos, pois tem a missão de informar e conscientizar a população sobre essa doença, que é responsável por elevadas taxas de mortalidade (CARMO et al., 2017; SOUSA et al., 2014).

Em uma pesquisa realizada entre 2014 e 2018, Marreiros et al. (2020) sugerem a inclusão do parceiro no tratamento durante a assistência pré-natal, evitando a reinfeção da gestante, atuando como ponto chave para o controle da doença, juntamente com diagnóstico e tratamento das grávidas, e melhor notificação do número de casos para controle epidemiológico adequado da infecção. Outro estudo, realizado no período de 2007 a 2012 em seis unidades federativas do Brasil, mostrou que a incidência de sífilis congênita variou de 35,6% no Distrito Federal a 63,9% no Rio Grande do Sul, sendo que entre as gestantes com sífilis, 43% tiveram desfecho notificado de sífilis congênita (SARACENI et al., 2017). No período de 2010 a 2018, a taxa de incidência de sífilis congênita aumentou 3,8 vezes, passando de 2,4 para 9,0 casos por mil nascidos vivos (Brasil, 2019).

Os dados expressos no Boletim Epidemiológico de sífilis do ano de 2020 mostram que, de 1998 a junho de 2020, foram notificados no SINAN 236.355 casos de sífilis congênita em menores de um ano de idade, das quais 105.084 (44,5%) residiam na região Sudeste, 70.478 (29,8%) no Nordeste, 27.269 (11,5%) no Sul, 20.159 (8,5%) no Norte e 13.365 (5,7%) no Centro-Oeste. Isso demonstra o insucesso em atingir a meta da Rede Cegonha proposta pelo Ministério da Saúde em 2011 para melhorar a qualidade da atenção ao pré-natal e diminuir o número de casos de sífilis congênita (BRASIL, 2020a). Com o objetivo de ampliação do diagnóstico, nos anos de

2012 a 2015 houve um aumento no número testes rápidos oferecidos pelo Ministério da Saúde passando de 1,2 milhões para mais de 6 milhões de testes, mas esse aumento da testagem não significa acesso ao tratamento. Apesar de a penicilina benzatina ser reconhecida como medicação essencial no tratamento da sífilis, em 2014 sua falta foi sentida mundialmente. No Brasil o desabastecimento foi solucionado por meio da integração entre vários departamentos de saúde tais como: Assistência Farmacêutica, Agência Nacional de Vigilância Sanitária (ANVISA) e Conselho Nacional de Secretários de Saúde (RUBERTI, 2020) .

A taxa de incidência de sífilis congênita chegou a alcançar, no ano de 2018, 9,0 casos por mil nascidos vivos, diminuindo para 8,2 casos por mil nascidos vivos em 2019, resultando em uma redução de 8,9% (Tabela 1). Apesar da diminuição dos casos de sífilis em quase todo o país, parte dessa redução pode estar relacionada à divergência na transferência de dados entre as esferas de gestão do SUS, o que pode ocasionar diferença no total de casos de sífilis entre as bases de dados municipal, estadual e federal. Além disso, o declínio no número de casos também pode decorrer de uma demora na notificação e alimentação das bases de dados do Sistema de Informação de Agravos de Notificação (SINAN), por parte dos profissionais de saúde ocasionada pela pandemia de COVID-19 (BRASIL, 2020a).

Tabela 1 – Número de casos de sífilis congênita em menores de um ano de idade e taxa de incidência (por 1.000 nascidos vivos) por ano de diagnóstico. Brasil, 2010-2019.

Sífilis congênita	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	Total
Número de casos	6.946	9.487	11.634	13.972	16.310	19.640	21.254	24.962	26.441	24.130	236.355
Taxa de detecção	2,4	3,3	4,0	4,8	5,5	6,5	7,4	8,5	9,0	8,2	

Fonte: <http://indicadoressifilis.aids.gov.br/>.

No Estado de São Paulo, ações articuladas do Programa Estadual de Infecções Sexualmente Transmissíveis e Aids de São Paulo (PE-IST/AIDS-SP) com Atenção Básica, Programas da Área Técnica Materno-Infantil, Humanização, Comitês de Mortalidade Materno-Infantil, municípios, regionais estaduais e sociedade civil têm contribuído para melhorar a resposta paulista no enfrentamento da sífilis. No período de 1986 a junho de 2020, foram notificados 43.713 casos de sífilis. A taxa de incidência de sífilis congênita após elevação contínua desde 2010, tem apresentado estabilidade em 2017, 2018 e 2019, com índices de 6,8, 6,7 e 6,9 casos por 1000 nascidos vivos-ano, respectivamente, ainda distante do que preconiza a OMS/OPAS, que para atingir a eliminação da sífilis congênita a taxa de incidência deverá ser menor ou igual a 0,5 casos por 1000 nascidos vivos (SÃO PAULO, 2020).

O conhecimento dos fatores que comprometem a eliminação da sífilis é de suma importância para o seu controle. Com isso, torna-se essencial investigar os fatores que implicam na sua transmissão, para o desenvolvimento de medidas que contribuam na redução da incidência da doença na gravidez e minimizem os riscos de transmissão da sífilis congênita (OLIVEIRA et al., 2020).

A justificativa para a realização do estudo se deu em virtude do aumento do número de casos diagnosticados de sífilis e sífilis congênita nos últimos anos e das consequências na vida

das crianças que nascem com essa doença, tendo em vista, ainda as consequências relacionadas ao agravo. Segundo a Organização Pan-Americana da Saúde (OPAS, 2017) a sífilis afeta um milhão de gestantes por ano em todo o mundo, ocasionando mais de 300 mil mortes fetais e neonatais, colocando mais de 200 mil crianças em risco de morte prematura. No Brasil, a taxa de mortalidade infantil por sífilis congênita passou de 2,3/100 mil nascidos vivos em 2006 para 6,7/100 mil nascidos vivos em 2016.

Ressalta-se a importância de rever e estudar fatores que influenciam tal aumento. Na literatura mundial há poucos estudos abordando a sífilis congênita sob o ponto de vista espaço-temporal (MEDEIROS et al., 2022).

Neste trabalho, investigamos o desempenho de modelos espaço-temporais Bayesianos e usamos uma verossimilhança de Poisson e Binomial Negativa, considerando três tipos de prioris CAR, como o modelo Besag, York & Mollié (BYM) (BESAG; YORK; MOLLÍE, 1991), o modelo BYM modificado (SIMPSON et al., 2017) e o modelo de Leroux (LEROUX; LEI; BRESLOW, 2000). Até onde sabemos, a comparação entre esses modelos espaciais Bayesianos ainda não foi explorada na análise do risco relativo de sífilis congênita. Neste cenário, a contribuição do presente trabalho está na utilização da estatística espaço-temporal para auxiliar gestores de saúde na tomada de decisões, proporcionando um melhor entendimento sobre a doença, auxiliando no monitoramento, e contribuindo para um melhor conhecimento dos riscos de transmissão da doença. Os resultados oferecem grande potencial para prevenção e controle da sífilis congênita em nível de áreas de ponderação (AP).

Visto a importância do estudo mais aprofundado da doença, propõe-se uma abordagem espaço-temporal hierárquica Bayesiana para a modelagem da incidência de sífilis congênita no município de São Paulo, Brasil. Usando dados de notificação de casos para o período de 2010–2016, foram estimados os parâmetros dos modelos e determinados os padrões espaço-temporais. Estes modelos incluíram interação de efeitos espaciais e temporais, conjuntamente, para estimar os riscos relativos de sífilis congênita que fornecerão informações adicionais para o monitoramento da doença e esforços de controle.

Os parâmetros de efeitos aleatórios podem ser interpretados como substitutos para processos espaciais, temporais ou espaço-temporais latentes. Os efeitos aleatórios espaciais, por exemplo, representam covariáveis espacialmente autocorrelacionadas e são frequentemente especificados por meio de modelos autorregressivos condicionais (CAR) (BESAG; YORK; MOLLÍE, 1991). Os modelos CAR assumem que a média do parâmetro de efeitos aleatórios espaciais em uma determinada área é condicional às estimativas de efeitos aleatórios espaciais em áreas geograficamente adjacentes e tem sido bastante utilizados em estudos ecológicos na análise de resultados relacionados à saúde (CONGDON, 2000; GRUENEWALD et al., 2013; LAWSON; ROTEJANAPRASERT, 2014; NETO et al., 2020; OTIENDE; ACHIA; MWAMBI, 2020; NGWIRA et al., 2021).

A abordagem hierárquica Bayesiana torna-se uma estrutura apropriada para resolver

complexidades nas estruturas espaço-temporais (RICHARDSON; ABELLAN; BEST, 2006; OLESON; SMITH; KIM, 2008). A aplicação de modelos Bayesianos na análise de conjuntos de dados espaço-temporais é computacionalmente desafiador usando algoritmos convencionais de Monte Carlo via cadeias de Markov (MCMC, do inglês *Markov chain Monte Carlo*) (BEST; RICHARDSON; THOMSON, 2005; BLANGIARDO; CAMELETTI, 2015) pois o MCMC requer avaliação iterativa de várias funções de densidade condicional completa e a convergência das distribuições de probabilidade a posteriori requer extensivas simulações (BANERJEE; CARLIN; GELFAND, 2015; CARROLL et al., 2015). Recentemente, Rue, Martino e Chopin (2009) propuseram o método Integrated Nested Laplace Approximation (INLA), que aproxima distribuições de probabilidade a posterioris por meio de integração numérica, ao invés de um processo iterativo baseado em MCMC. Pesquisas anteriores comparando abordagens MCMC e INLA tem mostrado que o INLA reduz o tempo computacional e obtém estimativas de parâmetros confiáveis (RUIZ-CÁRDENAS; KRAINSKI; RUE, 2012; BLANGIARDO et al., 2013). A importância da eficiência computacional ao analisar dados espaço-temporais está no fato de que, várias distribuições e suposições a priori para os parâmetros de efeitos aleatórios espaciais, temporais e espaço-temporais podem ser testados e comparados (KNORR-HELD, 2000).

Esta pesquisa analisa os padrões espaço-temporais de sífilis congênita em pequenas áreas no Município de São Paulo, Brasil. Os casos de sífilis congênita foram agregados temporalmente a 7 períodos de tempo, e geograficamente agregadas a 310 áreas de ponderação, resultando em um total de 2170 unidades espaço-temporais. Neste contexto, investigamos os padrões espaço-temporal da sífilis congênita, usando dados notificados de casos e caracterizando as áreas com riscos relativos excepcionalmente altos. Os mapas gerados a partir desses modelos descrevem associações que garantem mais investigações epidemiológicas sobre os desafios e oportunidades existentes para vigilância da doença.

Este estudo teve como objetivo estimar no espaço e no tempo o risco de sífilis congênita e avaliar sua relação com as variáveis socioeconômicas e ambientais (entre elas, índice de Gini¹, renda média per capita, proporção de mulheres de 18 a 24 anos sem instrução ou com ensino fundamental incompleto, proporção da população que vive em domicílios com banheiro e água encanada, etc.) por meio de modelos espaço-temporais utilizando abordagem Bayesiana com o algoritmo INLA, e uma abordagem frequentista GAM.

Outro objetivo foi descrever um padrão espaço-temporal da SC em nível de áreas de ponderação no município de São Paulo de 2010 a 2016 e explorar os fatores de risco associados às covariáveis escolhidas utilizando ferramentas de mapeamento espaço-temporal da doença e dados do SINAN e IBGE.

As contribuições deste trabalho são resumidas da seguinte forma:

¹ O índice de Gini é uma medida de desigualdade econômica, ou seja, mostra quão concentrada é a renda de uma determinada população. Quando o índice está mais próximo de 1, maiores são as desigualdades, maior é a concentração de renda. <<http://infodf.codeplan.df.gov.br/indice-de-gini-brasil-e-df/>>. Acesso em 11 de maio de 2021.

(a) A maioria dos estudos existentes se concentra apenas nos efeitos espaciais e/ou temporais dos dados, mas até onde sabemos os efeitos de interação espaço-temporal não são abordados em estudos sobre dados de sífilis congênita. Assim, este estudo contribui para a literatura ao explorar plenamente os efeitos de interação espaço-temporal dos dados.

(b) Espera-se que a maioria dos modelos forneça precisão de predição satisfatória, mas são frequentemente criticados devido à falta de interpretabilidade. Por outro lado, sem comprometer a precisão da predição, o método proposto permite avaliar a significância das variáveis explicativas (proporção de pessoas brancas, proporção de pessoas pardas, proporção de pessoas pretas, proporção de mulheres em idade fértil, proporção de nascidos vivos de mães adolescentes, índice de Gini, renda média per capita, proporção de mulheres de 18 a 24 anos sem instrução ou com ensino fundamental incompleto, proporção da população que vive em domicílios com banheiro e água encanada), efeitos espaciais, temporais e interação espaço-temporal, podendo inferir em que medida cada componente do modelo contribui para a variação da variável resposta.

(c) Modelos estatísticos Bayesianos são criticados pela grande carga computacional devido ao método MCMC geralmente usado para obter as estimativas dos parâmetros, o que limita os potenciais de predição. Propomos a utilização da técnica de aproximação de Laplace aninhada e integrada (INLA) para agilizar o processo de estimação dos modelos Bayesianos considerados.

Além disso, analisamos minuciosamente todos os quatro tipos de interação espaço-temporal Bayesianos utilizando três tipos de estruturas espaciais com aplicação de um método computacional recente para estimação dos modelos, o método INLA. Este estudo enfoca as potencialidades do modelo de interação espaço-temporal proposto na literatura para analisar o risco da sífilis congênita no município de São Paulo.

A investigação da variação geográfica da incidência de sífilis congênita conjuntamente com indicadores socioeconômicos e ambientais pode revelar diferenças locais importantes na ocorrência da doença.

Em epidemiologia da sífilis congênita, os estudos que têm adotado os modelos autorregressivos se restringem a um pequeno número de possibilidades para incorporar interações espaciais. Geralmente não testam a grande variedade de opções de especificação dos modelos autorregressivos (como diferentes vizinhos, matrizes de pesos ou, mesmo, tipos de modelo) e não investigam sistematicamente a qualidade do ajuste obtido.

A avaliação das frequências e distribuição geográfica dos casos de sífilis congênita de modo especializado e atualizado auxilia na compreensão da doença e pode ser importante na determinação de novas hipóteses etiológicas e no desenvolvimento de políticas (estratégias) de prevenção e controle.

O presente trabalho está organizado em capítulos, sendo eles: **2. Revisão de literatura**, no qual é apresentado um resumo da teoria estatística necessária para a modelagem dos dados,

os principais modelos espaço-temporal para dados de área, descrição de algumas estruturas condicionais autoregressivas - (CAR), tais como o modelo de Besag, York e Mollié (1991), Leroux, Lei e Breslow (2000) e de Besag, York e Mollié modificado sugerido por Simpson et al. (2017); **3. Um estudo espaço-temporal da sífilis congênita no Município de São Paulo no período de 2010 a 2016 usando o INLA**, em que são ajustados modelos para determinar os principais fatores associados à sífilis congênita, e mapas de risco que mostram as áreas que apresentam maior e menor risco da doença em cada ano no nível de áreas de ponderação para o município de São Paulo utilizando-se a metodologia INLA; **4. Modelagem de dados espaço-temporais da sífilis congênita no município de São Paulo de 2010 a 2016 usando modelos aditivos generalizados**, em que utilizamos os modelos GAM para analisar o risco de sífilis congênita com os efeitos espacial e temporal e variáveis associadas. Os modelos GAM constituem um método flexível para capturar relações não lineares; **5. Comparação entre as abordagens INLA e GAM**, que por fim são apresentados as comparações entre as duas abordagens consideradas, quando as estruturas espaciais e temporais estão presentes.

2 Revisão de literatura

2.1 Mapeamento de Doenças

O mapeamento espacial ou espaço-temporal do risco de doenças apresenta uma longa história na vigilância em saúde pública e tem sido cada vez mais utilizado em trabalhos científicos, tornando-se uma ferramenta popular para descrever o padrão na contagem de doenças e identificação de regiões com taxas de incidência incomuns (SCHRÖDLE; HELD, 2011b). Distribuições de doenças podem ser retratadas por meio de várias técnicas cartográficas, como pontos, linhas e polígonos (BHUNIA; SHIT, 2019). Segundo Elliott e Wartenberg (2004), os mapas de doenças fornecem um resumo gráfico das informações espaciais e permitem identificar formas elusivas nos dados que podem não ser perceptíveis em tabelas. Eles são empregados de várias maneiras: na formulação de hipóteses quanto à etiologia da doença, na verificação de áreas com suspeita de alto risco e no apoio do desenvolvimento de estratégias e distribuição de recursos de saúde (BHUNIA; SHIT, 2019; PFEIFFER et al., 2008). De acordo com Rezaeian et al. (2004), a produção de mapas atraentes e informativos complementa a análise formal de dados epidemiológicos espaciais e, como resultado do impacto visual que eles apresentam, é provável que influenciem o destinatário muito mais do que as estatísticas que os acompanham.

As estimativas de risco de uma determinada doença são baseadas em informações dos casos observados, o número de indivíduos em risco e, possivelmente, também em informações obtidas de covariáveis, como fatores sociodemográficos, de saúde e ambientais. Modelos hierárquicos Bayesianos são usados para descrever a variabilidade na variável resposta como uma função das covariáveis e efeitos aleatórios que buscam capturar a variação inexplicada. O uso da modelagem Bayesiana fornece uma abordagem flexível e robusta que permite levar em consideração os efeitos das variáveis explicativas e acomodar a correlação espacial e espaço-temporal, e fornece uma expressão formal de incerteza nas estimativas de risco (MORAGA, 2019).

Os dados de saúde pública são frequentemente obtidos pela agregação de dados pontuais em subáreas da região de estudo, como municípios, áreas de ponderação, setores censitários, etc., devido a vários motivos, como, por exemplo, sigilo do paciente. Na escolha de qual nível de agregação se dará o estudo, as unidades de análise são definidas segundo o problema apontado em uma investigação: as variáveis (como serão formuladas); os indicadores; a natureza das amostras e como mensurá-las; os procedimentos de análise; a interpretação dos resultados, e as

possíveis inferências (RESENDES et al., 2006). Frequentemente, os modelos de risco de doença visam obter estimativas de baixa variância nas mesmas áreas em que os dados estão disponíveis. Uma limitação dessa abordagem é que os mapas de risco de doenças obtidos nesta resolução são incapazes de mostrar como o risco varia dentro das áreas, o que dificultam direcionar intervenções de saúde e direcionar recursos onde eles são mais necessários.

2.2 Modelos lineares generalizados

Os modelos lineares generalizados (GLM, do inglês *generalized linear models*) são uma generalização e extensão útil dos modelos de regressão linear para acomodar distribuições de dados não normais e transformações para linearidade de maneira direta. O desenvolvimento da teoria do GLM é baseado na família exponencial de distribuições e surgiu a partir da teoria desenvolvida por Nelder e Wedderburn (1972) e McCullagh e Nelder (1989). Eles tiveram a ideia de reunir vários modelos que já haviam sido desenvolvidos, tais como modelo linear normal, modelo de regressão logística, modelo de regressão de Poisson, entre outros, e mostraram que os parâmetros poderiam ser estimados por meio de um algoritmo simples que produzia as estimativas de máxima verossimilhança. Além disso, o modelo GLM permite obter uma maior flexibilidade para a relação funcional entre a média da variável resposta e o preditor linear. Desde então, com os avanços dos *softwares* estatísticos, esses modelos tornaram-se uma ferramenta básica para a maioria dos pesquisadores.

Suponha que os valores $Y_i, i = 1, \dots, n$ são realizações de uma variável aleatória que é observada independentemente. Juntamente com os valores de $p < n$ variáveis explicativas $x_{i1}, \dots, x_{ip}$, um GLM contém os seguintes componentes (FOX, 2008):

1. **Componente aleatório:** A distribuição $f(\cdot)$ de Y_i em x_{ij} é um membro da família exponencial com média μ_i e parâmetro de escala ϕ .
2. **Componente sistemático:** As variáveis explicativas determinam a média de Y_i por meio de uma função linear de preditores:

$$\eta_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip},$$

ou em forma matricial, $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$, em que $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_p)$ é uma matriz $n \times p$ e $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)$ é um vetor p -dimensional de coeficientes de regressão.

3. **Função de ligação:** Uma das dificuldades no entendimento dos GLMs é a relação entre os valores da variável resposta (como medido nos dados e predito pelo modelo em valores ajustados) e o preditor linear. O aspecto a se lembrar é que a função de ligação relaciona o valor valor esperado da variável Y ao seu preditor linear. A função linear de preditores η_i está relacionada com a média μ_i da distribuição, e com a função densidade de probabilidade

$f(\cdot)$ por meio de uma função invertível suave chamada função de ligação $g(\cdot)$ tal que $g(\mu_i) = \eta_i$ ou em forma matricial,

$$g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} \Rightarrow g^{-1}(\mathbf{X}\boldsymbol{\beta}) = \boldsymbol{\mu} = E(\mathbf{Y})$$

em que $\mathbf{X}\boldsymbol{\beta}$ é um vetor n -dimensional das funções lineares do preditor $\boldsymbol{\eta}$. Por exemplo, a função de ligação canônica é *log* para Poisson e *logit* para binomial. Observamos que o preditor linear emerge do modelo linear como uma soma dos termos para cada um dos p parâmetros. Este não é um valor de Y (exceto no caso especial da ligação identidade). O valor de $\boldsymbol{\eta}$ é obtido pela transformação do valor de Y por meio da função de ligação, e o valor predito de Y é obtido pela aplicação da função de ligação inversa de $\boldsymbol{\eta}$ (CRAWLEY, 2012).

2.2.1 A Família Exponencial de Distribuições

A família exponencial compreende um conjunto flexível de distribuições que abrange variáveis aleatórias contínuas e discretas, muito rica quando se trata de modelagem estatística de conjuntos de dados na prática, entre os modelos estatísticos mais importantes e incluem muitos exemplos clássicos. As distribuições pertencentes a esta família gozam de certas propriedades matemáticas interessantes e importantes que são necessárias para grande parte da teoria e métodos inferenciais tradicionais, que muitas vezes atraem pesquisadores para membros específicos desta família a fim de realizar estudos estatísticos (MUKHOPADHYAY, 2020). Este conceito, assim como muitas outras noções básicas de estatística, foram introduzidas por Fisher (1934). Em homenagem a ele e a alguns outros precursores da teoria, a família exponencial de distribuições é às vezes chamada de família do tipo Fisher Darmois-Koopman-Pitman (DAUM, 1986; BARANKIN; MAITRA, 1963). O termo “família exponencial” vem da convenção de que as subfunções estão contidas no expoente da função exponencial natural (ou seja, o número irracional $e = 2,718281$ elevada a alguma potência especificada) (GILL et al., 2019).

Para entender como caracterizar uma variável pertencente à família exponencial e a flexibilidade mencionada do modelo, considere $Y_1, \dots, Y_n$ variáveis aleatórias independentes. Dizemos que uma variável aleatória pertence a família exponencial se sua função de densidade ou probabilidade, puder ser escrita da seguinte forma:

$$f(y_i; \theta_i; \phi) = \exp \left\{ \frac{1}{a(\phi)} [y\theta_i - b(\theta_i)] + c(y_i; \phi) \right\}$$

em que a , b e c são funções que assumem valores em subconjuntos dos reais, o parâmetro ϕ é conhecido como parâmetro de escala, de dispersão ou de perturbação e θ_i é denominado como parâmetro canônico da distribuição. A função $a(\phi)$ é geralmente da forma $a(\phi) = \phi \cdot \omega$, em que ω é uma constante conhecida (DOBSON; BARNETT, 2018; MYERS et al., 2012). A média e a

variância são expressas por:

$$E(Y_i) = \mu_i = \frac{db(\theta_i)}{d\theta_i}$$

$$\text{Var}(Y_i) = \frac{d^2b(\theta_i)}{d\theta_i^2} \cdot a(\phi).$$

2.2.2 Modelo regressão de Poisson

A regressão de Poisson é um método tradicional para modelar variáveis com resposta de contagem e a utilizamos como modelo para os casos de sífilis congênita. Este modelo é apropriado quando há uma contagem relativamente baixa de doenças e a população é relativamente grande em cada pequena área. A distribuição de Poisson é definida como:

$$p(Y_{it} = y_{it} | \mu_{it}) = \frac{\mu_{it}^{y_{it}} \exp(-\mu_{it})}{y_{it}!}, \quad (2.1)$$

em que $P(Y_{it})$ é a probabilidade do número de eventos (y_{it}) que ocorreram na i -ésima área de ponderação ($i = 1, 2, \dots, n$) no ano t , e μ_{it} é o parâmetro que representa o valor esperado de Y_{it} . Prova-se que a média é $E(Y_{it}) = \mu_{it}$ e a variância é $\text{Var}(Y_{it}) = \mu_{it}$. O conjunto de variáveis preditoras $\mathbf{X}$ está relacionado a μ_{it} por meio da função:

$$\log(\mu_{it}) = \mathbf{X}'\boldsymbol{\beta},$$

em que $\boldsymbol{\beta}$ é o vetor de parâmetros de regressão a serem estimados a partir dos dados, que descreve o efeito fixo das condições socioeconômicas e ambientais que descreve os efeitos das covariáveis no valor esperado da quantidade de casos de SC.

2.2.3 Modelo binomial negativa

Uma alternativa mais comum ao modelo de Poisson é o modelo binomial negativa, que aborda a sobredispersão incluindo um parâmetro de dispersão para acomodar a heterogeneidade não observada nos dados de contagem, sendo definida por:

$$P(Y_{it} = y_{it} | \mu_{it}, \kappa) = \frac{\Gamma(y_{it} + \frac{1}{\kappa})}{\Gamma(y_{it} + 1) \cdot \Gamma(\frac{1}{\kappa})} \left(\frac{1}{1 + \kappa\mu_{it}} \right)^{\frac{1}{\kappa}} \left(\frac{\kappa\mu_{it}}{1 + \kappa\mu_{it}} \right)^{y_{it}},$$

com a média $E(Y_{it}) = \mu_{it}$ e função de variância $\text{Var}(Y_{it}) = \mu_{it} + \kappa\mu_{it}^2$, em que κ é conhecido como parâmetro de dispersão. O modelo de Poisson é o modelo limite do modelo binomial negativa quando $\kappa \rightarrow 0$. A função de ligação do modelo binomial negativa também é:

$$\log(\mu_{it}) = \mathbf{X}'\boldsymbol{\beta}$$

Obs.: Na teoria da probabilidade e estatística, a distribuição binomial negativa é uma distribuição de probabilidade discreta que modela o número de sucessos em uma sequência de ensaios de Bernoulli independentes e identicamente distribuídos antes que um número especificado (não aleatório) de falhas ocorra.

2.3 Modelos aditivos generalizados

Um modelo aditivo generalizado (GAM, do inglês *generalized additive model*) é uma extensão do modelo GLM com um preditor linear contendo uma soma de funções suavizadas de covariáveis. Especificamente, seja Y_i uma variável aleatória independente com média $\mu_i = E(Y_i)$, $i = 1, 2, \dots, n$ e com uma distribuição pertencente à família exponencial. Muitos modelos estatísticos amplamente utilizados na literatura pertencem a essa classe geral, incluindo modelos aditivos para dados gaussianos, modelos logísticos não paramétricos para dados binários e modelos log-lineares não paramétricos para dados de Poisson (TERZI; CENGIZ, 2009). Na falta dessa condição, utiliza-se uma relação de variância média conhecida adequada para uso de métodos de quase-verossimilhança. O preditor linear de um modelo GAM tem uma estrutura da forma:

$$g(\mu_i) = \alpha + \mathbf{X}_i\beta + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}) + f_4(x_{4i}) + \dots \quad (2.2)$$

em que g é uma função de ligação monótona suavizada conhecida, α é o intercepto, $\mathbf{X}_i\beta$ é a parte paramétrica do preditor linear, x_{ji} são variáveis do preditor e f_j são funções suavizadas.

Uma das muitas vantagens de trabalhar com suavizações é a possibilidade de combinar várias covariáveis em uma única função suavizada. Isso estenderia a formulação na Equação (2.2) do seguinte modo:

$$g(\mu_i) = \alpha + \mathbf{X}_i\beta + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + \dots \quad (2.3)$$

A flexibilidade proporcionada pela incorporação de tais suavizações bivariadas em um modelo GAM será essencial para nosso estudo. Assumindo que a variável resposta, número de casos de SC por área de ponderação e por ano, é dependente das covariáveis socioeconômicas e ambientais e dos termos suaves das coordenadas dos centroides e do tempo.

Observe um ponto importante. Para que o modelo (2.3) seja identificável, as funções suavizadas geralmente precisam ser restritas a ter soma zero para não haver confundimento com o intercepto (geralmente tomada sobre o conjunto de valores das covariáveis, ou seja, $\sum_j f_j(x_{ij}) = 0$). Para uma discussão sobre esse assunto sugerimos consultar Wood, Scheipl e Faraway (2013). A função `gam` do pacote `mgcv` do *software* R sempre aplica tais restrições.

2.3.1 Modelos de regressão spline de suavização

O objetivo das técnicas de regressão não paramétrica é estender a regressão linear para formas mais flexíveis e suaves. Um modelo de regressão *spline* de suavização assume que:

$$y_i = f(x_i) + \epsilon_i, \quad i = 1, \dots, n \quad (2.4)$$

em que y_i são observações da função contínua f avaliada nos pontos x_i , e ϵ_i são erros aleatórios independentes com média zero e variância comum σ^2 .

As técnicas de regressão não paramétrica ou simplesmente técnicas de suavização são muito amplas na literatura (Wang (2011), Gu e Gu (2013), Wood (2017) apresentam uma revisão detalhada sobre essas técnicas). Há vários diferentes tipos de suavizações, que podem ser aplicadas a uma única variável ou termos de interação com até três variáveis. No caso espacial, a forma geral da função de suavização $f(\cdot)$ em (2.4) pode ser considerada como uma função bivariada das covariáveis espaciais $\mathbf{s} = (x_1, x_2)$, em que x_1 e x_2 representam as coordenadas geográficas (longitude e latitude).

Splines são funções polinomiais por partes restritas a se unirem em certos pontos chamados nós, que são espaçados uniformemente pelo intervalo de valores observados de x . Os elementos básicos a serem considerados para o ajuste são: (i) grau do polinômio; (ii) número de nós; e (iii) localização dos nós. A escolha mais popular é a *spline* cúbica.

Para entender melhor como funciona uma *spline* cúbica, consideremos os pontos x_1 e x_2 no intervalo $[a, b]$, tais que:

$$a < x_1 < x_2 < b.$$

A função $f(\cdot)$, definida no intervalo $[a, b]$ é uma *spline* cúbica, se satisfaz as seguintes condições:

1. Para cada um dos intervalos (a, x_1) , (x_1, x_2) , (x_2, b) , f é um polinômio cúbico.
2. Cada uma das partes do polinômio se une nos pontos x_i , de tal modo que a função f e suas derivadas de primeira e segunda ordem são contínuas em cada ponto x_i e, portanto, contínua no intervalo $[a, b]$. Os pontos x_i são os nós.

Temos as seguintes funções de base: $b_1(x) = 1$, $b_2(x) = x$, $b_3(x) = x^2$ e $b_4(x) = x^3$. Assim, uma *spline* cúbica ajustará um polinômio de grau 3 em cada intervalo $[a, x_1]$, $(x_1, x_2]$ e $(x_2, b]$. Como não há garantia de continuidade da função de suavização nos nós, é necessário criar mais duas outras funções para compor a base: $b_5(x) = (x - x_1)_+^3$ e $b_6(x) = (x - x_2)_+^3$. Para $k = 1, 2$, tem-se:

$$(x - x_k)_+^3 = \begin{cases} (x - x_k)^3, & \text{se } x > x_k \\ 0, & \text{c.c.} \end{cases}$$

A partir dessas considerações, a *spline* cúbica pode escrita como uma combinação linear das funções de base:

$$f(x) = \sum_{j=1}^6 \beta_j b_j(x),$$

em que os polinômios são descritos da seguinte forma:

$$f(x) = \begin{cases} \beta_1 + \beta_2 x + \beta_3 x^2 + \beta_4 x^3, & \text{se } a \leq x \leq x_1 \\ \beta_1 + \beta_2 x + \beta_3 x^2 + \beta_4 x^3 + \beta_5 (x - x_1)^3, & \text{se } x_1 < x \leq x_2 \\ \beta_1 + \beta_2 x + \beta_3 x^2 + \beta_4 x^3 + \beta_5 (x - x_1)^3 + \beta_6 (x - x_2)^3, & \text{se } x_2 < x \leq b \end{cases}$$

A solução do problema de estimação dos coeficientes β_j das funções de *splines*, consiste na minimização da soma de quadrados penalizados dada pela Equação (2.5) apresentada a seguir.

A regressão não paramétrica com *splines* é comumente conhecida como *splines* de suavização. Considere o problema de regressão em (2.4). As *splines* de suavização são a solução para o problema de minimizar a soma dos quadrados dos resíduos:

$$\text{SQR}(f, \lambda) = \sum_{i=1}^n \{y_i - f(x_i)\}^2 + \lambda \int \left(\frac{df(x_i)}{dx_i^2} \right)^2 dx, \quad (2.5)$$

em que o primeiro termo mede a proximidade do ajuste aos dados, o segundo termo penaliza a curvatura da função f , e λ é um parâmetro de suavização que controla o ajuste dos dados e a suavidade de f .

Para cada x único, existem funções de base, $b_j(x)$, para $j = 1, \dots, k$, tais que,

$$f_\lambda(x) = \sum_{j=1}^k b_j(x) \beta_j.$$

Outro tipo de técnica de suavização baseada em *spline*, são as *splines* de placa fina propostas por Duchon (1977). As *splines* de placa fina são o análogo natural das *splines* cúbicas em várias dimensões e o termo de penalidade neste caso é:

$$J_{m,d}(f) = \int_{\mathbb{R}^d} \sum \frac{m!}{\alpha_1! \dots \alpha_d!} \left(\frac{\partial^m f}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}} \right)^2 dx_1 \dots dx_d \quad (2.6)$$

em que a expressão (2.6), é calculada sobre todos os α 's não nulos, tais que $\alpha_1 + \dots + \alpha_d = m$ e $2m > d$. Neste caso, as *splines* de suavização são a solução para o problema de minimizar a soma dos quadrados dos resíduos:

$$\text{SQR}(f, \lambda) = \|\mathbf{y} - \mathbf{f}\|^2 + \lambda J_{m,k}(f), \quad (2.7)$$

em que $\mathbf{f} = [f(\mathbf{x}_1), f(\mathbf{x}_2), \dots, f(\mathbf{x}_n)]^T$, $\mathbf{x}_i$ ($i = 1, 2, \dots, n$) são vetores de k covariáveis observadas, m é a ordem de diferenciação dada em (2.6).

Para o caso bivariado, $J_{2,2}(f) = \iint \left(\frac{\partial^2 f}{\partial x_1^2} \right)^2 + 2 \left(\frac{\partial^2 f}{\partial x_1 \partial x_2} \right)^2 + \left(\frac{\partial^2 f}{\partial x_2^2} \right)^2 dx_1 dx_2$, que produz a penalidade para uma *spline* cúbica.

2.4 Estimação bruta de risco de sífilis congênita

2.4.1 Modelos espaciais para mapeamento de doenças

Suponhamos que o domínio espacial de interesse seja dividido em n áreas contíguas. Para uma dada área i , com $i = 1, \dots, n$, seja y_i o número de casos observados da doença e N_i

a população em risco. O indicador mais simples de mortalidade/incidência é a taxa bruta, que geralmente é definida como o número de casos por 100.000 pessoas, ou seja,

$$TX_i = \frac{y_i}{N_i} \times 100000.$$

Quando o objetivo do estudo é detectar quais áreas apresentam maior ou menor risco da doença, o número de casos esperados em cada área geralmente é calculado. Por exemplo, se a população é dividida em grupos de idade, o método de padronização indireta é comumente usado para calcular o número esperado de casos, e é obtido por:

$$E_i = \sum_{j=1}^J y_{ij} \frac{y_j}{N_{ij}}, \text{ para } i = 1, 2, \dots, n,$$

em que y_{ij} e N_{ij} são o número de casos e a população de risco na i -ésima área e j -ésima faixa etária, respectivamente e y_j é a população na j -ésima faixa etária. Observe que E_i representa o número de casos que esperamos observar nesta área se ela se comportar de modo semelhante em toda a região de estudo. Usando essas quantidades, a razão de mortalidade/incidência padronizada (SMR ou SIR) é definida como a razão de casos observados e esperados para a unidade de área correspondente.

A escolha da população padrão de referência é fundamental, pois diferentes escolhas podem levar a diferentes conclusões sobre o risco. Na maioria dos estudos, a taxa ou contagem esperada é considerada fixa depois de calculada. Observe também que uma padronização mais sofisticada pode ser usada, por meio da qual os substratos na população podem ser ajustados. Por exemplo, a distribuição de idade e sexo pode ser incluída como parte do ajuste da taxa esperada. Isso pode ser importante quando há uma forte diferenciação espacial na distribuição de substratos em uma área de estudo (LAWSON, 2018).

2.4.2 Estimação para dados espaciais

Suponha que o risco desconhecido de SC na área i seja dado como $\psi_i, i = 1, 2, \dots, n$. Sejam y_i e NV_i o número de casos de SC e o número de nascidos vivos, respectivamente, na área i . O número esperado de casos de SC na área i pode ser escrito como $E_i = rNV_i$, em que $r = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n NV_i}$ é a taxa bruta global da doença nos nascidos vivos do estudo. Estimamos ψ_i segundo a abordagem frequentista. Primeiramente, assumimos que $y_i \sim \text{Poisson}(E_i\psi_i)$. Com base na amostra $\mathbf{y} = y_1, \dots, y_n$, a função de verossimilhança e a função de log-verossimilhança correspondente são expressas como:

$$\ell(\psi_i) = \prod_{i=1}^n \frac{\exp(-E_i\psi_i) (E_i\psi_i)^{y_i}}{y_i!} = P(\mathbf{y}, \mathbf{E} | \boldsymbol{\psi}) \quad (2.8)$$

e

$$\ln \ell(\psi_i) = - \sum_{i=1}^n E_i\psi_i + \sum_{i=1}^n y_i \ln(E_i\psi_i) - \sum_{i=1}^n \ln(y_i!) \quad (2.9)$$

O estimador de máxima verossimilhança $\hat{\psi}_i$ de ψ é obtido por meio de $\frac{\partial \ln(l(\psi_i))}{\partial \psi_i}$ e dado por $\hat{\psi}_i = \frac{y_i}{E_i}$ (BOLFARINE; SANDOVAL, 2001; DEGROOT; SCHERVISH, 2012). Este estimador, $\hat{\psi}_i$, é referido como a razão de incidência padronizada na área i . O erro padrão de $\hat{\psi}_i$ é dado por $SE(\hat{\psi}_i) = \frac{\sqrt{y_i}}{E_i}$ e o correspondente intervalo de confiança (IC) $100(1 - \alpha)\%$ para ψ_i é dado por $\hat{\psi}_i \pm z_{\frac{\alpha}{2}} SE(\hat{\psi}_i)$, para $i = 1, 2, \dots, n$ e $t = 1, 2, \dots, T$.

Interpretação: $\hat{\psi}_i = 1$ significa ter o mesmo risco em relação ao risco médio em toda região; $\hat{\psi}_i > 1$ ($\hat{\psi}_i < 1$) indica risco maior (menor) que a média esperada.

2.5 Inferência Bayesiana

A abordagem estatística Bayesiana é baseada na atualização de informações usando a chamada Lei de Bayes (sinônimo de Teorema de Bayes) de seu famoso ensaio de 1763. O reverendo Thomas Bayes foi um matemático amador cuja principal contribuição foi um ensaio encontrado e publicado dois anos após sua morte por seu amigo Richard Price (BELLHOUSE, 2004). Bayes foi o primeiro a desenvolver explicitamente essa famosa lei, mas foi Laplace (LAPLACE; ÉTRANGES, 1774) quem apresentou uma análise mais detalhada que talvez seja mais relevante para a prática da estatística Bayesiana atualmente. Stigler (1986) apresenta uma discussão histórica interessante e, (SHEYNNIN, 1977) apresenta uma análise técnica detalhada sobre o assunto. Como Bayes, Laplace assumiu uma distribuição uniforme para o parâmetro desconhecido, mas se preocupou muito menos do que Bayes com as consequências dessa suposição. Distribuições a priori uniformes são simplesmente distribuições “flat” que atribuem probabilidades iguais para todos os resultados possíveis (GILL, 2015).

Na abordagem Bayesiana, consideramos que os parâmetros desconhecidos, θ , são variáveis aleatórias e introduzimos uma crença a priori sobre eles, permitindo a incerteza dos parâmetros nas predições. Em teoria, essa crença a priori deve ser declarada antes de olhar para os dados. Os métodos Bayesianos combinam explicitamente quatro componentes do conhecimento. Um conhecimento prévio e novos dados, que são combinados, utilizando um modelo, para produzir um conhecimento a posteriori (MCCARTHY, 2007). Esses quatro componentes podem ser representados como:

$$\text{priori} + \text{dados} \xrightarrow{\text{modelo}} \text{posteriori}.$$

A priori e a posteriori referem-se a antes e depois de considerar os dados coletados. Pelo teorema de Bayes (BAYES, 1763), a distribuição conjunta a posteriori dos parâmetros e das variáveis latentes é dada por:

$$p(\theta, \mathbf{x} | \mathbf{y}) = \frac{p(\mathbf{y} | \theta, \mathbf{x})p(\theta, \mathbf{x})}{p(\mathbf{y})}, \quad (2.10)$$

em que $\mathbf{x}$ são as variáveis latentes, $p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{x})$ é a verossimilhança dos dados, $p(\boldsymbol{\theta}, \mathbf{x})$ é a priori, $p(\mathbf{y})$ é a probabilidade marginal do vetor de resposta $\mathbf{y}$.

A probabilidade marginal $p(\mathbf{y})$ pode ser obtida integrando o numerador no lado direito da Equação (2.10) sobre o suporte para $\boldsymbol{\theta}$, ou seja,

$$p(\mathbf{y}) = \int p(\mathbf{y} | \boldsymbol{\theta}, \mathbf{x}) p(\boldsymbol{\theta}, \mathbf{x}) d\boldsymbol{\theta}. \quad (2.11)$$

A complexidade computacional do cálculo da posteriori (2.10) depende da complexidade do modelo. Usando um modelo com uma distribuição a priori que é conjugada à verossimilhança, uma fórmula analítica explícita para a distribuição a posteriori pode ser obtida e usada para a inferência. Em muitas aplicações práticas, tais modelos não são satisfatórios, e o cálculo ou aproximação da posteriori é feito numericamente. No entanto, em modelos relativamente complexos, o cálculo de $p(\mathbf{y})$ é inviável por integração numérica. Observemos que $p(\mathbf{y})$ é apenas uma constante em relação aos parâmetros de interesse, e existem métodos para estimar a distribuição a posteriori sem avaliar $p(\mathbf{y})$, escrevendo:

$$p(\boldsymbol{\theta}, \mathbf{x} | \mathbf{y}) \propto p(\mathbf{y} | \boldsymbol{\theta}, \mathbf{x}) p(\boldsymbol{\theta}, \mathbf{x}), \quad (2.12)$$

ou seja, a densidade a posteriori (evidência atualizada) é proporcional à verossimilhança (evidência dos dados) vezes a priori (evidência histórica ou suposições do modelo gerado). De modo mais geral, a constante de normalização requer que uma integral multidimensional seja avaliada, uma vez que o cálculo envolve a integração de todos os parâmetros no modelo (WOODWARD, 2019). Quando a dimensão de $\boldsymbol{\theta}$ é pequena, podemos usar a integração numérica aproximada, como a aproximação de Laplace (RAFTERY, 1996). Em aplicações mais complexas, tais aproximações não são viáveis e a integração para a obtenção de $p(\mathbf{y})$ é difícil, de forma que a amostragem direta de $p(\boldsymbol{\theta}, \mathbf{x} | \mathbf{y})$ não é viável. Neste caso, uma maneira de amostrar $p(\boldsymbol{\theta}, \mathbf{x} | \mathbf{y})$ é o uso dos métodos de Monte Carlo Via Cadeias de Markov (MCMC, do inglês *Markov chain Monte Carlo*), que não necessitam do conhecimento de sua forma analítica (CONGDON, 2014).

O principal problema prático com a abordagem Bayesiana vem do fato de que o cálculo da constante de normalização requer uma integral multidimensional. Até pouco tempo, os métodos numéricos disponíveis não eram capazes de lidar com esses problemas nem com complexidade moderada. Apenas nas duas últimas décadas do século XX que a abordagem Bayesiana se tornou uma opção viável para muitos problemas com a disponibilidade de algoritmos MCMC (WOODWARD, 2019). No contexto de uma análise Bayesiana, o método MCMC é um processo iterativo que gera uma série de valores para cada parâmetro em um modelo especificado, conhecido como cadeia. A teoria por trás do método afirma que, desde que o processo MCMC seja executado por um tempo suficientemente longo, o mesmo convergirá de tal forma que os valores por ele gerados se comportarão como se viessem da distribuição conjunta a posteriori dos parâmetros. Uma vez que o processo MCMC tenha convergido, a distribuição a posteriori pode ser caracterizada com precisão arbitrária simplesmente gerando um número suficientemente

grande de valores simulados da distribuição a posteriori e usando estatísticas de resumo padrão para estimar as quantidades de interesse, por exemplo, a média a posteriori, desvio padrão, ou percentis. Como este é um processo iterativo no qual cada geração dos valores simulados depende da geração anterior, conforme definido pelo algoritmo da cadeia de Markov, é necessário fornecer os valores iniciais com os quais iniciar a cadeia. Normalmente, esses valores iniciais não estarão próximos do estado convergente final e, portanto, é necessário incluir o que é denominado fase de “*burn-in*” durante a qual o processo MCMC deve convergir. Avaliar a convergência é uma parte extremamente importante do uso desse método. Somente após a convergência e, por definição, a fase de *burn-in* concluída, os valores simulados podem ser usados para resumir a distribuição a posteriori, ou seja, todos os valores gerados durante a fase de *burn-in* são descartados. Atualmente existem softwares disponíveis gratuitamente como o *OpenBugs*, que implementam a abordagem MCMC para inferência Bayesiana.

2.6 Dados espaciais pela metodologia INLA

Pesquisadores de diversas áreas têm-se confrontados cada vez mais com a tarefa de analisar dados que são:

- com muitas variáveis preditoras e de resposta;
- geograficamente referenciados e frequentemente apresentados como mapas;
- correlacionados temporalmente, como em estruturas longitudinais ou outras estruturas de séries temporais.

Por exemplo, em um estudo epidemiológico, podemos estar interessados na análise de taxas de câncer de pulmão, mama, colorretal e cervical por município e ano em um determinado estado ou região, incluindo as variáveis preditoras tais como informações sobre tabagismo, realização de mamografia entre outras. Para atender ao objetivo de estudos desta natureza, há uma necessidade de se fazer inferência estatística, como modelagem de tendências e estruturas de correlação, estimativa de parâmetros do modelo subjacente, teste de hipóteses (ou comparação de modelos) e predição de observações em horários ou locais não observados (BANERJEE; CARLIN; GELFAND, 2015).

Os métodos de estatística espacial existem há algum tempo, iniciado com o trabalho de Cressie (1993), que apresenta uma grande variedade de problemas que pode ser resolvidos usando métodos geoestatísticos. No entanto, na última década, avanços recentes na computação por meio de métodos como MCMC, permitem análises totalmente Bayesianas de modelos multiníveis mais sofisticados para dados complexos geograficamente referenciados (BANERJEE; CARLIN; GELFAND, 2015).

Métodos aplicados a dados com distribuição espacial que levam em conta a dependência espacial, como deste trabalho, estão cada vez mais avançados. Entre as ferramentas estatísticas disponíveis, uma abordagem recentemente desenvolvida por [Rue e Martino \(2007\)](#) e [Rue, Martino e Chopin \(2009\)](#), que utiliza estatística Bayesiana e é denominada por Aproximação Integrada de Laplace Aninhada (INLA, do inglês *Integrated Nested Laplace Approximations*), permite abordar a heterogeneidade espacial e espaço-temporal com eficiência computacional e levar em conta tanto a vizinhança (a dependência espacial) como a existência de estruturas hierárquicas dos dados ([CHIARAVALLOTI-NETO, 2017](#)). Modelos complexos que consideram estruturas espaciais e espaço-temporais com grandes conjuntos de dados podem levar várias horas ou dias de tempo computacional para a execução de inferência Bayesiana via MCMC ([BLANGIARDO; CAMELETTI, 2015](#)). A abordagem INLA fornece um método que permite melhorias substanciais neste tempo para inferência Bayesiana e na precisão das aproximações para as densidades marginais a posteriori $\pi(\theta_i|\mathbf{y})$, $\pi(\Psi|\mathbf{y})$ e $\pi(\Psi_j|\mathbf{y})$, de modelos aditivos estruturados, que são escritos como um modelo hierárquico Bayesiano do seguinte modo:

$$\pi(\boldsymbol{\theta}, \Psi|\mathbf{y}) \propto \pi(\Psi)\pi(\boldsymbol{\theta}|\Psi)\pi(\mathbf{y}|\boldsymbol{\theta}, \Psi). \quad (2.13)$$

Por ser um método determinístico, o mesmo não requer uso de simulação de Monte Carlo, ou seja, não é necessário verificar convergência.

O método INLA possui aplicações em modelos hierárquicos Bayesianos em que um vetor $\boldsymbol{\theta}$ Gaussiano multivariado de dimensão n é um campo latente (isto é, não-observado) e possui sua densidade multivariada controlada por um vetor de hiperparâmetros (parâmetros da priori gaussiana do campo latente) de dimensão m , em que $\Psi = (\Psi_1, \dots, \Psi_m)$. Alguns elementos do vetor aleatório $\boldsymbol{\theta}$ são indiretamente observados através dos dados $\mathbf{y}$, de dimensão n . Os dados são assumidos condicionalmente independentes dado o campo latente $\boldsymbol{\theta}$. No caso mais simples em que cada observação está conectada com apenas um elemento do campo latente, a independência condicional é escrita como:

$$\pi(\mathbf{y}|\boldsymbol{\theta}, \Psi) = \prod_{i=1}^n \pi(y_i|\theta_i, \Psi), \quad (2.14)$$

em que cada dado y_i está associado a único um elemento θ_i no campo latente $\boldsymbol{\theta}$.

A abordagem INLA tem como objetivo principal a aproximação analítica das marginais a posteriori tanto das variáveis latentes quanto dos hiperparâmetros do modelo Gaussiano latente condicionada aos dados observados ([BLANGIARDO; CAMELETTI, 2015](#)), que podem ser escritas como:

$$\pi(\theta_i|\mathbf{y}) = \int \pi(\theta_i|\Psi, \mathbf{y}) \pi(\Psi|\mathbf{y}) d\Psi \quad \text{e} \quad \pi(\Psi_j|\mathbf{y}) = \int \pi(\Psi|\mathbf{y}) d\Psi_{-j}. \quad (2.15)$$

em que Ψ_{-j} corresponde ao vetor de hiperparâmetros Ψ excluindo a j -ésima componente. Os índices i e j variam em uma faixa diferente dependendo do número de variáveis latentes e hiperparâmetros.

As aproximações das marginais a posterioris $\tilde{\pi}(\theta_i|\mathbf{y})$ considerando as variáveis latentes são então calculadas por integração numérica, do seguinte modo:

$$\tilde{\pi}(\theta_i|\mathbf{y}) = \int \tilde{\pi}(\theta_i|\Psi, \mathbf{y}) \tilde{\pi}(\Psi|\mathbf{y}) d\Psi \approx \sum_{k=1}^h \tilde{\pi}(\theta_i|\Psi_k, \mathbf{y}) \tilde{\pi}(\Psi_k|\mathbf{y}) \Delta_k, \quad (2.16)$$

em que Δ_k são pesos apropriados da área correspondente à exploração da grade da distribuição a posteriori dos hiperparâmetros Ψ , que seriam iguais a 1, por exemplo, se todos os pontos fossem escolhidos de forma equidistante.

A aproximação na Equação (2.16) usando a abordagem INLA é obtida pelo seguinte algoritmo:

1. Selecione um conjunto de $\Psi = (\psi_1, \dots, \psi_k)$;
2. Para $k = 1$ até h faça:
3. Calcule $\tilde{\pi}(\Psi_k|\mathbf{y})$;
4. Calcule $\tilde{\pi}(\theta_i|\Psi_k, \mathbf{y})$ como uma função de θ_i ;
5. Fim para;
6. Calcule $\tilde{\pi}(\theta_i|\mathbf{y}) = \sum_{k=1}^h \tilde{\pi}(\theta_i|\Psi_k, \mathbf{y}) \tilde{\pi}(\Psi_k|\mathbf{y}) \Delta_k$, para $i = 1, 2, \dots, n$.

Mais detalhes são apresentados por [Rue, Martino e Chopin \(2009\)](#), [Bonat \(2010\)](#).

As aproximações das marginais a posterioris $\tilde{\pi}(\Psi_j|\mathbf{y})$ para os hiperparâmetros, com $j = 1, \dots, h$ são calculadas de modo similar. [Rue, Martino e Chopin \(2009\)](#) mostram duas estratégias para a escolha dos pontos de integração θ_i e, também apresentam todos os detalhes do esquema de aproximação. Além disso, eles consideram que a integração numérica é possível quando a dimensão de Ψ é pequena (em geral menor ou igual a 6) e, a abordagem aninhada torna esta aproximação muito acurada quando aplicada em modelagens espacial e espaço-temporal. A rotina INLA é implementada no programa R ([TEAM, 2021](#)) com várias opções de verossimilhanças e estruturas de correlação no pacote R-INLA ([RUE; MARTINO; CHOPIN, 2009](#); [MARTINS et al., 2013](#)).

Para o ajuste dos modelos vamos usar a abordagem INLA que pode ser facilmente aplicada usando o pacote R-INLA ([RUE et al., 2013](#)), escrito em linguagem C e permite a especificação e ajuste do modelo em R. Pode ser baixado em [<http://www.r-inla.org/>](http://www.r-inla.org/), onde também estão disponíveis a documentação do pacote, vários exemplos e um fórum de discussão ([MARTINO; RUE, 2009](#)).

Geralmente, nas abordagens Bayesianas são utilizados métodos MCMC, principalmente o algoritmo de Metropolis-Hasting e o amostrador de Gibbs ([MORAGA et al., 2017](#)). Apesar desses

métodos terem tido grande impacto na estatística Bayesiana, sendo possível trabalhar com modelos mais complexos, eles apresentam uma gama de problemas em termos de convergência e eficiência computacional. Sua validade inferencial depende da convergência de uma cadeia de Markov para sua distribuição de equilíbrio, o que pode ser difícil de verificar empiricamente (TAYLOR; DIGGLE, 2014). Uma alternativa seria usar a abordagem INLA proposta por Rue, Martino e Chopin (2009), cuja grande vantagem é o tempo computacional significativamente menor quando comparados ao MCMC, além de não haver a necessidade de verificar a convergência.

O INLA usa uma combinação de aproximação analítica e integração numérica para fazer inferência Bayesiana aproximada em modelos gaussianos latentes, uma subclasse especial de modelos de regressão aditiva estruturada, abordado por Rue, Martino e Chopin (2009), que inclui uma grande classe de modelos que variam de modelos lineares mistos generalizados a modelos espaciais e espaço-temporais (MORAGA et al., 2017). Esta classe é caracterizada por uma variável resposta Y_i com distribuição pertencente à família exponencial, com valor esperado μ_i relacionado com um preditor linear aditivo η_i por meio de uma função de ligação $g(\cdot)$, tal que $g(\mu_i) = \eta_i$. O preditor linear η_i explica o efeito das várias covariáveis na resposta média de maneira aditiva:

$$\eta_i = \alpha + \sum_{j=1}^{n_f} f^{(j)}(u_{ji}) + \sum_{k=1}^{n_\beta} \beta_k x_{ki}, \quad (2.17)$$

em que α é uma constante que representa o intercepto, $f^{(j)}(\cdot)$ são funções que representam os efeitos aleatórios das covariáveis u , β_k são os efeitos fixos e lineares da covariável x . Os modelos gaussianos latentes são um subconjunto de todos os modelos aditivos Bayesianos com um preditor aditivo estruturado (2.17), ou seja, aqueles que atribuem priori gaussiana para α , $f^{(j)}(\cdot)$ e β_k .

Geralmente, na inferência Bayesiana, o interesse está na distribuição a posteriori de (θ, Ψ) , dado os valores observados y e, essa distribuição a posteriori pode ser obtida por meio do teorema de Bayes, que permite escrevê-la da seguinte forma:

$$\pi(\theta, \Psi | y) \propto \pi(\Psi) \pi(\theta | \Psi) \prod_{i=1}^n \pi(y_i | \theta_i, \Psi), \quad (2.18)$$

e que no caso dos modelos Gaussianos latentes é dado por:

$$\pi(\theta, \Psi | y) \propto \pi(\Psi) |Q(\Psi)|^{\frac{1}{2}} \exp \left\{ -\frac{1}{2} \theta^\top Q(\Psi) \theta + \sum_{i=1}^n \log [\pi(y_i | \theta_i, \Psi)] \right\}, \quad (2.19)$$

em que $Q(\Psi)$ é a matriz de precisão.

Essa classe de modelo possui diversas aplicações, devido às diferentes formas que as funções desconhecidas $f^{(j)}(\cdot)$, como exemplificamos abaixo:

- Modelos lineares generalizados: $f^{(j)}(\cdot) = 0$
- Modelos lineares generalizados mistos: $f^{(j)}(\cdot) \sim N(0, Q^{-1})$, onde Q é a matriz de precisão.

- Modelos hierárquicos: $f^{(j)}(\cdot) \sim N(\mathbf{0}, \sigma_f^2)$; $\sigma_f^2 \sim$ distribuição qualquer
- Modelos espaciais e espaço-temporais:

$$\begin{aligned} f^{(1)}(\cdot) &\sim \text{CAR} \quad (\text{efeitos espaciais estruturados}); \\ f^{(2)}(\cdot) &\sim \mathcal{N}\left(0, \sigma_{f^{(2)}}^2\right) \quad (\text{efeitos não espacialmente estruturados}); \\ f^{(3)}(\cdot) &\sim \text{AR}\left(\phi, \sigma_{f^{(3)}}^2\right) \quad (\text{efeitos temporais}); \\ f^{(4)}(\cdot) &\sim \mathcal{N}\left(0, \sigma_{f^{(4)}}^2\right) \quad (\text{efeitos não temporalmente estruturados}). \end{aligned}$$

em que CAR (do inglês *conditional autoregressive models*) significa modelo autoregressivo condicional, (JIN; CARLIN; BANERJEE, 2005) e AR (do inglês *autoregressive models*) significa modelo autorregressivo (BANERJEE; CARLIN; GELFAND, 2015; WALLER; GOTWAY, 2004).

2.7 Modelos autorregressivos de ordem 1

O modelo AR de primeira ordem é intensamente usado em séries temporais por conta de sua estrutura simples e sua ampla gama de aplicações. Segundo JUDY L. KLEIN, o modelo AR originou-se na década de 1920 e foi aplicado pela primeira vez em 1927 por UDN Y YULE na análise do comportamento de séries temporais de manchas solares (KLEIN; KLEIN, 1997).

Em modelos de regressão múltipla, a variável de interesse é modelada usando-se uma combinação linear de preditoras. Já, em modelos de autorregressão, modelamos a variável de interesse usando uma combinação linear de valores anteriores da variável. O termo auto regressão indica que é uma regressão da variável a partir dela mesma (AMARASINGHE et al., 2018).

Assim, um modelo autorregressivo de ordem p , ou simplesmente modelo AR(p), pode ser escrito como:

$$y_t = \rho_1 y_{t-1} + \rho_2 y_{t-2} + \dots + \rho_p y_{t-p} + \epsilon_t,$$

em que ϵ_t é o ruído branco (ou erro residual), ou seja, $\{\epsilon_t\}$ é uma sequência de variáveis aleatórias independentes e identicamente distribuídas com média e variância finitas. Se ϵ_t é normalmente distribuído com média zero e precisão $\tau_\epsilon = 1/\sigma_\epsilon^2$, então ϵ_t é chamado de “ruído branco gaussiano” (HAMILTON, 2020). Os pesos ρ_i da combinação linear são os parâmetros do modelo e são considerados constantes. Além disso, os $y_{t-1}, y_{t-2}, \dots, y_{t-p}$ são independentes de ϵ_t . A auto regressão é como uma regressão múltipla, mas com valores defasados como preditores (ULLRICH, 2021). Dinamicamente, o modelo diz que o valor no momento atual é igual a um “fator de propagação” (ou “fator de transição”) multiplicado pelo valor no momento anterior, mais um “erro de inovação” independente (CRESSIE, 1993).

Geralmente restringimos modelos autorregressivos a dados estacionários, caso em que algumas restrições sobre os valores dos parâmetros são necessárias (BOX et al., 2015):

- Para um modelo AR(1): $|\rho| < 1$.

- Para um modelo AR(2): $|\rho_2| < 1$, $\rho_1 + \rho_2 < 1$ e $\rho_2 - \rho_1 < 1$.

Quando $p \geq 3$, as restrições são muito abstratas, mas o software R lida com essas restrições ao estimar um modelo.

O modelo auto-regressivo de primeira ordem, denominado modelo AR(1), pode ser definido a partir da seguinte equação:

$$y_t - \rho y_{t-1} = \epsilon_t, \quad t = 1, 2, \dots, T, \quad (2.20)$$

em que $y_1 = \epsilon_1$, $\epsilon_t \sim N(0, \tau_\epsilon^{-1})$.

Neste caso, temos que $E(y_t) = 0$ para todo t e a variância é dada por (BOX et al., 2015):

$$\text{Var}(y_t) = \text{Var}(\rho y_{t-1} + \epsilon_t) = \rho^2 \text{Var}(y_{t-1}) + \tau_\epsilon^{-1} = \rho^2 \text{Var}(y_t) + \tau_\epsilon^{-1}.$$

Daí,

$$\text{Var}(y_t) = \frac{\tau_\epsilon^{-1}}{1 - \rho^2}, \quad |\rho| < 1.$$

Podemos obter a distribuição marginal para este processo em um conjunto finito de pontos no tempo, $\mathbf{y} = \{y_1, \dots, y_T\}$. Considerando que $y_1 \sim N(0, \tau_\epsilon^{-1}/(1 - \rho^2))$, a distribuição é estacionária, e $y_t|y_{t-1}, \tau_\epsilon \sim N(\rho y_{t-1}, \tau_\epsilon^{-1})$, para $t = 2, \dots, T$. Daí, a densidade conjunta de $\mathbf{y}$ é dada por:

$$\begin{aligned} \pi(\mathbf{y}) &= \pi(y_1) \cdot \pi(y_1|y_2) \cdot \pi(y_2|y_1, y_2) \cdot \dots \cdot \pi(y_T|y_1, y_2, \dots, y_{T-1}) \\ &= \pi(y_1) \cdot \pi(y_1|y_2) \cdot \pi(y_2|y_1) \cdot \dots \cdot \pi(y_T|y_{T-1}) \\ &= (2\pi\tau_\epsilon^{-1})^{-1/2} (1 - \rho^2)^{1/2} \exp\left\{-\frac{\tau_\epsilon}{2}(1 - \rho^2)y_1^2\right\} \cdot (2\pi\tau_\epsilon^{-1})^{-1/2} \exp\left\{-\frac{\tau_\epsilon}{2}(y_2 - \rho y_1)^2\right\} \cdot \\ &\quad \cdot (2\pi\tau_\epsilon^{-1})^{-1/2} \exp\left\{-\frac{\tau_\epsilon}{2}(y_3 - \rho y_2)^2\right\} \cdot \dots \cdot (2\pi\tau_\epsilon^{-1})^{-1/2} \exp\left\{-\frac{\tau_\epsilon}{2}(y_T - \rho y_{T-1})^2\right\} \\ &= (2\pi\tau_\epsilon^{-1})^{-T/2} \sqrt{1 - \rho^2} \exp\left\{-\frac{\tau_\epsilon}{2}[(1 - \rho^2)y_1^2 + (y_2 - \rho y_1)^2 + \dots + (y_T - \rho y_{T-1})^2]\right\}. \end{aligned}$$

A soma entre colchetes na exponencial pode ser escrita na forma matricial como:

$$\begin{aligned} &y_1^2 - \rho^2 y_1^2 + y_2^2 - 2\rho y_1 y_2 + \rho^2 y_1^2 + \dots + y_T^2 - 2\rho y_{T-1} y_T + \rho^2 y_{T-1}^2 \\ &= \sum_{i=1}^T y_i^2 - 2\rho \sum_{i=2}^T y_i y_{i-1} + \rho^2 \sum_{i=2}^{T-1} y_i^2 = \mathbf{y}^T \left[\mathcal{D}_T^{(1,1+\rho^2,1)} + \mathcal{T}_T^{(0,-\rho)} \right] \mathbf{y}, \end{aligned} \quad (2.21)$$

em que

$$\mathcal{D}_T^{(1,1+\rho^2,1)} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 + \rho^2 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 1 + \rho^2 & 0 \\ 0 & \dots & 0 & 0 & 1 \end{pmatrix}_{T \times T} \quad \text{e} \quad \mathcal{T}_T^{(0,-\rho)} = \begin{pmatrix} 0 & -\rho & 0 & \dots & 0 \\ -\rho & 0 & -\rho & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & 0 & -\rho & 0 & -\rho \\ 0 & \dots & 0 & -\rho & 0 \end{pmatrix}_{T \times T}.$$

Um dos benefícios ao lidar com matrizes esparsas é na fatoração de Cholesky ao calcular o determinante (KRAINSKI, 2018). Um fato interessante nesta parametrização para o modelo AR1 é que ele nem mesmo precisa da fatoração de Cholesky, pois $\left| \mathcal{D}_T^{(1,1+\rho^2,1)} + \mathcal{T}_T^{(0,-\rho)} \right| = 1 - \rho^2$ (Veja a Eq. (2.21)).

A matriz de covariância para uma série de tamanho T é (LINDSEY, 2004; WANG; FU; PAUL, 2022):

$$\Sigma = \frac{\tau_\epsilon^{-1}}{1 - \rho^2} \begin{pmatrix} 1 & \rho & \cdots & \rho^{T-2} & \rho^{T-1} \\ \rho & 1 & \cdots & \rho^{T-3} & \rho^{T-2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho^{T-2} & \rho^{T-3} & \cdots & 1 & \rho \\ \rho^{T-1} & \rho^{T-2} & \cdots & \rho & 1 \end{pmatrix}.$$

A inversa dessa matriz, denominada matriz de precisão, tem uma forma tridiagonal (WADE; QUAAS, 1993; HODGES, 2013):

$$\mathbf{Q} = \Sigma^{-1} = \tau_\epsilon \begin{pmatrix} 1 & -\rho & 0 & 0 & 0 & \cdots & 0 \\ -\rho & 1 + \rho^2 & -\rho & 0 & 0 & \cdots & 0 \\ 0 & -\rho & 1 + \rho^2 & -\rho & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & -\rho & 1 + \rho^2 & -\rho & 0 \\ 0 & \cdots & 0 & 0 & -\rho & 1 + \rho^2 & -\rho \\ 0 & \cdots & 0 & 0 & 0 & -\rho & 1 \end{pmatrix} = \mathcal{D}_T^{(1,1+\rho^2,1)} + \mathcal{T}_T^{(0,-\rho)}.$$

Além disso, $\mathbf{Q}$ possui uma decomposição simples de Cholesky (RUE, 2001), $\Sigma^{-1} = \mathbf{A}^\top \mathbf{A}$, em que:

$$\mathbf{A} = \frac{1}{\sqrt{\tau_\epsilon}} \begin{pmatrix} \sqrt{1 - \rho^2} & 0 & \cdots & 0 & 0 \\ -\rho & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & -\rho & 1 \end{pmatrix}.$$

A propriedade chave que faz com que $\mathbf{Q}$ seja esparsa está nas condicionais completas:

$$y_t \mid \mathbf{y}_{-t} \sim N \left(\frac{\rho}{1 - \rho^2} (y_{t-1} + y_{t+1}), \frac{\tau_\epsilon^{-1}}{1 + \rho^2} \right).$$

Cada ponto do tempo é apenas condicionalmente dependente dos dois pontos de tempo mais próximos, que é a razão para a estrutura tridiagonal de $\mathbf{Q}$.

Rue e Held (2005) mostra que a esparsidade da matriz de precisão $\mathbf{Q}$ é transportada para sua decomposição de Cholesky, permitindo uma avaliação da densidade de forma rápida (linear em T) e geração de números aleatórios. Em comparação, uma decomposição de Cholesky ou autodecomposição da matriz de covariância (densa) tem uma complexidade computacional de ordem T^3 (ALLÉVIUS, 2018).

2.8 Modelos para dados de área

Geralmente, parece razoável supor que áreas próximas no espaço apresentam uma carga de doenças mais semelhante do que as áreas que não estão próximas. Para definir o que significa “perto”, precisamos estabelecer uma estrutura de vizinhança. Uma suposição comum é considerar as áreas i e j como vizinhas se elas compartilham uma fronteira comum, que denotaremos por $i \sim j$. Isso parece razoável se as regiões forem igualmente dimensionadas e dispostas em um padrão regular (WAKEFIELD, 2007).

Cada elemento do conjunto D representa a i -região geográfica (área), $i = 1, \dots, n$, em que cada índice representa um conjunto de pares de coordenadas que delimitam uma área. Como ilustração, na Figura 1 são exibidos os 96 distritos do município de São Paulo, e para cada um deles está associado um código único. O conjunto de índices que denotam cada área D pode ser definido como $D = \{1, \dots, 96\}$ e a partir disso pode-se construir uma estrutura de dependência espacial para as observações, considerando regiões vizinhas cujas fronteiras físicas são comuns.

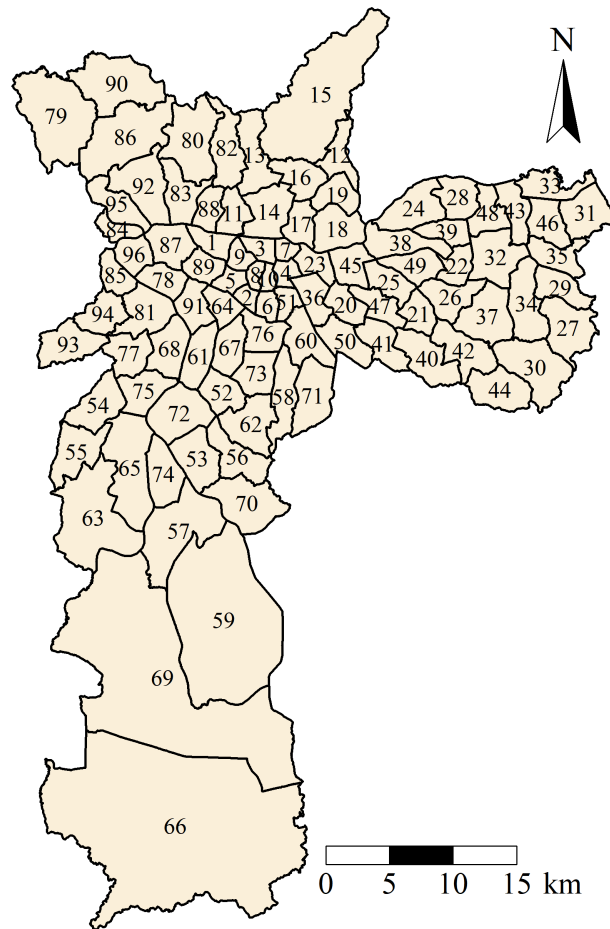


Figura 1 – Índices para os 96 distritos do município de São Paulo, Brasil.

Há várias maneiras de representar relações espaciais entre áreas (ou pontos fixos, como a localização dos casos ou os centroides de municípios, setores censitários, áreas de ponderação, etc.) em um mapa usando uma matriz de pesos espaciais (matriz de vizinhança), denotada por $\mathbf{W}$. Ao modelar dados espaciais, a matriz $\mathbf{W}$ desempenha um papel importante, pois fornece uma maneira de resumir as relações espaciais entre as áreas (ou pontos) às quais os dados observados são georreferenciados. As relações espaciais entre as áreas constituem um importante elemento na análise dos dados, e a construção da matriz $\mathbf{W}$ nos permite transformar essas relações em números. No entanto, diferente dos dados convencionais (como os números observados de casos de doenças), essa forma de dados geralmente é baseada em nossas suposições sobre como as áreas estão conectadas umas às outras.

Frequentemente, os elementos de $\mathbf{W}$ assumem valor 0 ou 1. Usamos w_{ij} para denotar o elemento de $\mathbf{W}$ na linha i e na coluna j ($i = 1, \dots, n$ e $j = 1, \dots, n$). Se $w_{ij} = 1$, então dizemos que as duas áreas i e j são vizinhas uma da outra. Se $w_{ij} = 0$, então as duas áreas i e j não são vizinhas. Cada elemento diagonal w_{ii} em $\mathbf{W}$ é definido como 0 (para todo i), o que significa que uma área não pode ser vizinha de si mesma. Por exemplo, ao construir uma matriz de vizinhança $\mathbf{W}$ de dimensão $n \times n$ dada pelo número de áreas (setores censitários, áreas de ponderação, distritos, bairros, etc), os elementos desta matriz denotam a dependência espacial entre as regiões. Observando a Figura 1, a linha 24 da matriz $\mathbf{W}$ que representa a relação do distrito 24 terá apenas os elementos $w_{24,28} = 1$, $w_{24,38} = 1$ e $w_{24,39} = 1$ com valores diferentes de zero. Desta forma, a representação do conjunto de vizinhos do distrito 24 pode ser expressa por $M_{24} = \{28, 38, 39\}$. E de modo geral para o exemplo acima, temos:

$$M_i = \{k; k \text{ é vizinho de } i\}, i = 1, \dots, 96.$$

2.9 Modelos autorregressivos condicionais (CAR)

Os modelos autorregressivos condicionais (CAR) têm sido usados extensivamente para identificar e detectar agrupamentos em risco de doenças. Nesses modelos, os riscos de doença em qualquer área são afetados pelo risco nas áreas vizinhas. Foram introduzidos há aproximadamente 30 anos por Besag (1974), e tiveram um aumento considerável no uso apenas na última década. Este ressurgimento acontece a partir de seu uso no contexto de amostragem de Gibbs e métodos mais gerais de Monte Carlo via cadeia de Markov (MCMC) para ajustar certas classes de modelos espaciais hierárquicos.

Para um vetor aleatório $\mathbf{u} = (u_1, \dots, u_n)$, um modelo CAR (Gaussiano) é definido como o seguinte conjunto de distribuições condicionais completas:

$$u_i \mid \mathbf{u}_{-i}, \boldsymbol{\mu}, \boldsymbol{\sigma} \sim N \left(\mu_i + \sum_{i \sim j} b_{ij} (u_j - \mu_j), \sigma_i^2 \right), i = 1, \dots, n, \quad (2.22)$$

em que $i \sim j$ indica que as regiões i e j são vizinhas (normalmente definida em termos de adjacência espacial), u_i indica a variável espacialmente autocorrelacionada da área i , o termo $\mathbf{u}_{-i}$

se refere a todos os elementos em $\mathbf{u}$, exceto seu i -ésimo componente e denota tais variáveis em locais vizinhos, μ_i é o valor de resposta esperada (tal que $E(u_i) = \mu_i$), σ_i^2 é a variância condicional, e b_{ij} são pesos conhecidos ou desconhecidos (com $b_{ii} = 0$), descrevendo a proximidade ou vizinhança entre os locais i e j .

Uma diferença notável entre os processos autorregressivos temporais e espaciais é que, para o primeiro, cada componente u_i depende apenas das observações anteriores $u_1, \dots, u_{i-1}$ na série temporal, enquanto que, para o caso espacial, não há ordenação semelhante a ser usada na definição das distribuições condicionais em (2.22). Portanto, a distribuição de u_i depende dos elementos restantes em $\mathbf{u}_{-i}$. Isso faz com que, para quaisquer dois componentes i e j com $i \neq j$, a distribuição condicional de u_i 's dependa de u_j , enquanto a distribuição condicional de u_j 's também depende de u_i .

Uma distribuição que tem sido a escolha mais popular para modelar riscos ao lidar com pequenas áreas é a distribuição autorregressiva condicional intrínseca (ICAR, do inglês *intrinsic conditional auto-regressive*) proposta por Besag, York e Mollié (1991). A distribuição ICAR é um caso especial da distribuição CAR quando $\rho = 1$. Embora isso leve a uma distribuição imprópria para $\mathbf{u}$ por causa da singularidade de sua matriz de covariância, a distribuição condicional associada de $\mathbf{u}$ dado Y é adequada.

Besag (1974) e Besag (1975) mostram com base no teorema da fatoração, que as distribuições condicionais completas em (2.22) sob a suposição de campo aleatório markoviano gaussiano multivariado (GMRF), determinam exclusivamente a seguinte distribuição conjunta:

$$\mathbf{u} \mid \boldsymbol{\mu}, \boldsymbol{\sigma} \sim NMV \left(\boldsymbol{\mu}, (\mathbf{I}_n - \mathbf{B})^{-1} \mathbf{M} \right), \quad (2.23)$$

em que $\mathbf{B} = [b_{ij}]$ é uma matriz $n \times n$ definida por contiguidade ou distância, $\mathbf{I}$ é uma matriz identidade, e $\mathbf{M}$ é uma matriz diagonal definida por $M_{ii} = \sigma_i^2$. Esta distribuição conjunta é usada, juntamente com a função de verossimilhança do conjunto de dados para estimar as distribuições a posterioris de todos os parâmetros desconhecidos.

A validade da distribuição na Equação (2.23) requer que sua matriz de covariância seja simétrica e positiva definida, condições que podem ser satisfeitas pela imposição de certas restrições nas formas de $\mathbf{B}$ e $\mathbf{M}$. Geralmente assumimos que $\mathbf{B} = \rho \mathbf{W}$ e $\sigma_i^2 = (\sigma^2 / w_{i+})$, em que ρ é referido como o coeficiente (ou parâmetro) de autocorrelação espacial ou dependência espacial, $\mathbf{W}$ é uma matriz de peso padronizada por linha, ou seja, $\mathbf{W} = [w_{ij}^*]$, com $w_{ij}^* = (w_{ij} / w_{i+})$ e w_{i+} é a soma da i -ésima linha de $\mathbf{W}$. Por construção, os elementos diagonais de $\mathbf{W}$ são todos zeros (CRESSIE, 1993). Essa estrutura de variância diz que, na presença de forte correlação espacial, quanto mais vizinhos uma área possui, mais informações existem nos dados sobre o valor de seu efeito aleatório, enquanto o parâmetro de variância σ^2 controla a variação entre os efeitos aleatórios estruturados (BANERJEE; CARLIN; GELFAND, 2015). As especificações CAR permitem matrizes de peso com base em contiguidade e distância, mas excluem o método de ponderação do k -ésimo vizinho mais próximo, pois tais pesos violam a condição de simetria.

A abordagem mais comum é definir vizinhos de primeira ordem, em que os pesos são definidos de modo que $w_{ij} = 1$ se i e j compartilham fronteiras comuns, $w_{ij} = 0$ caso contrário, e os elementos diagonais de $\mathbf{W}$ (w_{ii}) são todos zeros por construção. Segundo Cressie (1993), $(\mathbf{I}_n - \mathbf{B})^{-1} \rho \mathbf{W}$ é uma matriz de covariância válida se $\xi_{\max}^{-1} < \rho < \xi_{\min}^{-1}$, onde $\xi_{\min}$ e $\xi_{\max}$ são respectivamente, os autovalores mínimo e máximo de $\mathbf{W}$.

Na área da saúde pública, uma aplicação comum de tais modelos é o estudo de padrões regionais de determinada doença. Em geral, a localização precisa dos casos de doenças deve ser restrita em dados públicos devido a regulamentos de sigilo. Mas, resumos estatísticos de doenças em nível regional, entretanto, costumam ser relativamente mais acessíveis (CARLIN; LOUIS, 2008).

2.9.1 O modelo ICAR usando uma matriz de pesos espaciais com entradas binárias

A escolha do modelo ICAR e a respectiva matriz de pesos espaciais $\mathbf{W}$, em conjunto, terão forte influência na geografia da suavização espacial. Escolher uma matriz de pesos espaciais diferente, por exemplo, usando a definição de contiguidade tipo “queen” (rainha)¹, resultará em uma suavização espacial (ligeiramente) diferente (EARNEST et al., 2007). Dois indivíduos são espacialmente relacionados se compartilham pelo menos uma parte de suas fronteiras. No caso de unidades espaciais irregulares, pode haver muito poucas instâncias de unidades espaciais unidas em um vértice. Então, na prática, essas duas definições da matriz de pesos normalmente levam a suavizações muito semelhantes, se não idênticas. Outras definições de vizinhança baseadas na distância geográfica, por exemplo, $w_{ij} = 1$ se a distância entre os centroides das áreas i e j for menor que um limiar predefinido e $w_{ij} = 0$ caso contrário, produzirá novamente uma matriz de pesos binários, mas produzirá ainda outro padrão de suavização local (HAINING; LI, 2020).

O modelo ICAR define uma distribuição normal condicional para o parâmetro de cada área i de modo que a especificação desta distribuição dependa dos parâmetros dos vizinhos de i . Com base em uma matriz de pesos binários, a distribuição condicional para cada u_i ($i = 1, \dots, n$) é dada por:

$$u_i \mid \mathbf{u}_{-i}, \tau, \mathbf{W} \sim N \left(\frac{\sum_{j \sim i} u_j}{m_i}, \frac{\tau^{-1}}{m_i} \right) \quad (2.24)$$

em que $\mathbf{u}_{-i}$ denota o conjunto de parâmetros específicos da área excluindo u_i , a saber, $\mathbf{u}_{-i} = (u_1, \dots, u_{i-1}, u_{i+1}, \dots, u_n)$, $\tau = \frac{1}{\sigma^2}$ é o parâmetro de precisão desconhecido (σ^2 é a variância) e m_i é o número de vizinhos da área i (BESAG; KOOPERBERG, 1995).

A especificação condicional dada na Equação (2.24) implica em uma estrutura de dependência local que leva a uma suavização local ou de vizinhança dos parâmetros específicos da

¹ Assim como no movimento de peças em um tabuleiro de xadrez a contiguidade é dita ser do tipo “queen” (rainha), caso, além das fronteiras com extensão diferente de zero, puderem ser considerados os vértices como contíguos na visualização de um mapa, ou seja, esta contiguidade se dá em qualquer direção, com vértices ou lados (ALMEIDA, 2012).

área. Dados os parâmetros de todas as outras áreas, a média condicional de u_i , $\frac{\sum_{j \sim i} u_j}{m_i}$, depende apenas do número de áreas vizinhas, implicando uma estrutura de dependência local ao invés de uma estrutura de dependência global (no último caso, a média condicional de u_i dependeria de todos os outros u_j com $j \neq i$). Além disso, a média condicional é considerada como a média dos parâmetros das áreas vizinhas para que os parâmetros sejam semelhantes localmente. Quando aplicado à modelagem de SC, o modelo ICAR definido na Equação (2.24) assume que as áreas de ponderação que estão próximas no espaço têm níveis médios de casos de SC semelhantes. A Equação (2.24) sugere que quanto mais vizinhos uma área tiver, mais suavização local será captada. Isso porque, em comparação com uma área que possui apenas um ou dois vizinhos, uma área com um grande número de vizinhos (m_i grande), a variância condicional, $\frac{\tau^{-1}}{m_i}$, é pequena. Então é mais provável que seu parâmetro, u_i , esteja próximo da média m_i dos vizinhos da área.

A Equação (2.24) define o modelo ICAR por meio de um conjunto de n distribuições condicionais completas, e cada uma delas define uma distribuição de probabilidade para cada parâmetro dado todos os outros parâmetros. Usando as condicionais completas, podemos obter a distribuição de probabilidade conjunta para todos os parâmetros específicos da área, ou seja,

$$\pi(u_1, \dots, u_n) = \pi(\mathbf{u}) \propto \left\{ -\frac{\tau_u}{2} \mathbf{u}'(\mathbf{D} - \mathbf{W})\mathbf{u} \right\}. \quad (2.25)$$

2.10 Modelos espaço-temporais

Os modelos espaciais e espaço-temporais Bayesianos são frequentemente hierárquicos com três níveis (BERLINER, 1996). Nesse contexto, hierárquico é usado para descrever a estrutura dos modelos Bayesianos: o primeiro nível modela os dados observados com uma função de verossimilhança (distribuição de probabilidade) e especifica um modelo de processo que inclui efeitos aleatórios espaciais, temporais e espaço-temporais; o segundo nível atribui distribuições a priori para os parâmetros do modelo; e o terceiro nível especifica hiperpriors para parâmetros de distribuições a priori de segundo nível (por exemplo, variâncias). Distribuições a priori quantificam a incerteza do pesquisador no parâmetro.

No primeiro nível, o número de casos de sífilis congênita na área i ($i = 1, \dots, 310$) no tempo t ($t = 1, \dots, 7$) é representada por Y_{it} e amostrada a partir de uma distribuição de Poisson:

$$Y_{it} | \mu_{it} \sim \text{Poisson}(\mu_{it}),$$

$$\mu_{it} = E_{it} \psi_{it},$$

em que E_{it} é o número esperado de casos de sífilis congênita e ψ_{it} o risco relativo na área i e no tempo t . Os modelos de Poisson são geralmente usados em análises espaciais e espaço-temporais com dados de contagem.

2.10.1 Estimação para dados espaço-temporais

Suponha que o risco desconhecido de SC na área i e tempo t seja dado como ψ_{it} , $i = 1, 2, \dots, n$, $t = 1, 2, \dots, T$. Sejam y_{it} e NV_{it} o número de casos de SC e o número de nascidos vivos, respectivamente, na área i e no tempo t . O número esperado de casos de SC na área i e tempo t pode ser escrito como $E_{it} = rNV_{it}$, em que $r = \frac{\sum_{i=1}^n \sum_{t=1}^T y_{it}}{\sum_{i=1}^n \sum_{t=1}^T NV_{it}}$ é a taxa bruta global da doença nos nascidos vivos do estudo. Estimamos ψ_{it} segundo a abordagem frequentista, assumindo que $y_{it} \sim \text{Poisson}(E_{it}\psi_{it})$. Com base na amostra $\mathbf{y} = \{y_{11}, \dots, y_{n1}, y_{12}, \dots, y_{nT}\}$, a função de verossimilhança e a função de log-verossimilhança correspondente são expressas como:

$$\ell(\psi_{it}) = \prod_{i=1}^n \prod_{t=1}^T \frac{\exp(-E_{it}\psi_{it}) (E_{it}\psi_{it})^{y_{it}}}{y_{it}!} = P(\mathbf{y}, \mathbf{E} | \psi) \quad (2.26)$$

e

$$\ln \ell(\psi_{it}) = - \sum_{i=1}^n \sum_{t=1}^T E_{it}\psi_{it} + \sum_{i=1}^n \sum_{t=1}^T y_{it} \ln(E_{it}\psi_{it}) - \sum_{i=1}^n \sum_{t=1}^T \ln(y_{it}!). \quad (2.27)$$

O estimador de máxima verossimilhança $\hat{\psi}_{it}$ de ψ é obtido por meio de $\frac{\partial \ln(\ell(\psi_{it}))}{\partial \psi_{it}}$ e dado por $\hat{\psi}_{it} = \frac{y_{it}}{E_{it}}$ (BOLFARINE; SANDOVAL, 2001; DEGROOT; SCHERVISH, 2012). Este estimador, $\hat{\psi}_{it}$, é referido como a razão de incidência padronizada na área i no tempo t . O erro padrão de $\hat{\psi}_{it}$ é dado por $SE(\hat{\psi}_{it}) = \frac{\sqrt{y_{it}}}{E_{it}}$ e o correspondente intervalo de confiança (IC) $100(1 - \alpha)\%$ para ψ_{it} é dado por $\hat{\psi}_{it} \pm z_{\frac{\alpha}{2}} SE(\hat{\psi}_{it})$, para $i = 1, 2, \dots, n$ e $t = 1, 2, \dots, T$.

Interpretação: $\hat{\psi}_{it} = 1$ significa ter o mesmo risco em relação ao risco médio em toda região; $\hat{\psi}_{it} > 1$ ($\hat{\psi}_{it} < 1$) indica risco maior (menor) que a média esperada.

2.10.2 Estruturas espaciais e temporal

A estrutura de um modelo espaço-temporal separável considera a variabilidade espaço-temporal observada em dois componentes de modelagem: o componente espacial global que descreve o padrão espacial, permanecendo relativamente estável ao longo do tempo e o componente temporal global que captura o padrão geral de tendência. Com base nessa estrutura separável no espaço-tempo, a Equação (2.28) especifica o modelo usando uma formulação genérica para o logaritmo das taxas de sífilis congênita.

Como as áreas e o tempo, aparentemente, têm características distintas que podem não ser explicadas pelas covariáveis disponíveis, e tendo como objetivo reduzir a heterogeneidade dos dados, é razoável incluir no modelo efeitos aleatórios latentes para cada área e tempo. Com a suposição de que a distribuição de dados é Poisson, o modelo espaço-temporal que aplicamos neste estudo é decomposto aditivamente em componentes relativos a espaço, tempo e às covariáveis socioeconômicas e ambientais:

$$\eta_{it} = \log(\psi_{it}) = \alpha + \beta_{it} \mathbf{x}_{it} + SP_i + TM_t. \quad (2.28)$$

O parâmetro α quantifica a taxa média de casos de sífilis congênita em todas as áreas e β é o vetor dos coeficientes de regressão $\beta_{it} = (\beta_{1it}, \dots, \beta_{pit})'$ associado ao vetor x_{it} das covariáveis para $i = 1, 2, \dots, n$ e $t = 1, 2, \dots, T$.

Nessa formulação, os dois conjuntos de termos, $SP = (SP_1, \dots, SP_n)$ e $TM = (TM_1, \dots, TM_T)$, representam respectivamente a componente espacial global e a componente temporal global. O termo SP também é referido como efeitos espaciais principais, que são efeitos do espaço (localização) que permanecem constantes ao longo do tempo. Da mesma forma, TM é referido como os efeitos principais temporais, que são efeitos do tempo que permanecem os mesmos entre as áreas. A formulação na Equação (2.28) é genérico pois a modelagem dos efeitos principais espaciais e temporais pode ser feita usando vários modelos somente de espaço e somente de tempo (HAINING; LI, 2020).

Para avaliar o desempenho na modelagem espaço-temporal, construímos 72 modelos para comparação. Esses modelos tinham os mesmos componentes da Equação (2.28), incluindo informações das covariáveis no processo espaço-temporal, mas assumindo três diferentes especificações para as componentes espaciais e uma para a componente temporal.

As especificações consideradas são as seguintes:

- **Modelo BYM:** O modelo de Besag-York-Mollie - BYM (BESAG; YORK; MOLLIE, 1991), também conhecido como modelo de convolução, decompõe o efeito espacial (SP) na soma de um componente espacial estruturado e um componente não estruturado, de modo que $SP_i = u_i + v_i$, em que u_i e v_i são considerados independentes, a priori. Adotamos um modelo intrínseco condicional autorregressivo (ICAR) (BESAG; KOOPERBERG, 1995) para u_i , que pode ser escrito como $u_i | \mathbf{u}_{-i} \sim N\left(\frac{1}{m_i} \sum_{i \sim j} u_j, \frac{1}{\tau_u m_i}\right)$, em que $i \sim j$ indica que as áreas i e j são vizinhas, m_i é o número de áreas que compartilham fronteira com a i -ésima área, $\tau_u = 1/\sigma_u^2$ é o parâmetro de precisão e σ_u^2 , a variância. A dependência espacial em u_i assume a priori ICAR que estende o conhecido modelo Besag com uma distribuição gaussiana e implica que cada u_i é condicional ao vizinho u_j com variância dependente do número de áreas vizinhas m_i da i -ésima unidade espacial (Ver Seção 2.9.1). No modelo ICAR pressupõe-se que o risco de incidência de determinada doença em uma área espacial é derivado de vizinhos geograficamente próximos, ou seja, quanto menores são as distâncias espaciais, mais semelhante é o risco de incidência da doença nessas áreas. Este caráter espacial estruturado é denominado autocorrelação espacial.

O efeito aleatório espacialmente não estruturado v_i é modelado por uma priori permutável:

$$v_i | \tau_v \sim N(0, \tau_v^{-1}),$$

em que τ_v é o parâmetro de precisão.

A matriz de covariância resultante de $\mathbf{u} + \mathbf{v}$ é:

$$\text{Var}(\mathbf{u} + \mathbf{v} \mid \tau_u, \tau_v) = \tau_u^{-1} \mathbf{I} + \tau_v^{-1} \mathbf{Q}^{-},$$

em que $\mathbf{Q}^-$ denota a matriz inversa generalizada de Moore-Penrose (BAKSALARY; TRENKLER, 2021).

Uma restrição de soma zero $\sum_{i=1}^n (u_i + v_i) = 0$ é imposta a $u_i + v_i$ para garantir a identificabilidade do intercepto α , pois os componentes de variância no modelo BYM não são identificáveis a partir dos dados. Mais detalhes são apresentados por Besag, York e Mollié (1991), Besag et al. (1995) e Rue e Held (2005).

- **Modelo BYM2:** Mais recentemente, Simpson et al. (2017) afirmam que os componentes $\mathbf{u}$ e $\mathbf{v}$ não devem ser considerados independentes, sugerindo que as prioris para (2.28) devem ser dependentes. Eles propuseram uma reparametrização para o modelo de Besag, York e Mollié, com apenas um parâmetro de precisão nos componentes (τ) e um parâmetro de mistura (λ) para mensurar a variação espacial e não-espacial. O modelo BYM2 usa um componente estruturado espacialmente $\mathbf{u}_*$ e um componente não estruturado $\mathbf{v}_*$:

$$\mathbf{b} = \frac{1}{\sqrt{\tau_b}} \left(\sqrt{1 - \lambda} \mathbf{v}_* + \sqrt{\lambda} \mathbf{u}_* \right). \quad (2.29)$$

O parâmetro de precisão $\tau_b > 0$ controla a contribuição da variância marginal da soma ponderada de $\mathbf{u}_*$ e $\mathbf{v}_*$. O parâmetro de mistura $0 \leq \lambda \leq 1$ mede a proporção da variância marginal explicada pelo efeito estruturado $\mathbf{u}_*$. Assim, o modelo BYM2 é igual a um único modelo espacial quando $\lambda = 1$, e um único ruído espacial não estruturado quando $\lambda = 0$ (RIEBLER et al., 2016).

A matriz de covariância é dada por:

$$\text{Var}(\mathbf{b} \mid \tau_b, \lambda) = \tau_b^{-1} \left((1 - \lambda) \mathbf{I} + \lambda \mathbf{Q}_*^- \right), \quad (2.30)$$

em que $\mathbf{Q}_*^-$ denota a matriz inversa generalizada de Moore-Penrose (BAKSALARY; TRENKLER, 2021). A Equação (2.30) é uma reparametrização do modelo original BYM, em que $\tau_u^{-1} = \tau_b^{-1} \lambda$ e $\tau_v^{-1} = \tau_b^{-1} (1 - \lambda)$ (MACNAB, 2011). Assim, como no modelo anterior, uma restrição de soma zero $\sum_{i=1}^n b_i = 0$ deve ser imposta para evitar problemas de identificabilidade.

- **Modelo de Leroux:** O modelo proposto por Leroux, Lei e Breslow (2000) é outra variação do modelo BYM discutido acima. Enquanto o modelo BYM contém dois efeitos aleatórios espaciais, o modelo Leroux possui um único efeito aleatório espacial, $SP_i = u_i$, que ajusta a força da autocorrelação espacial da vizinhança local por uma constante λ . O modelo é dado por (LEE, 2011):

$$u_i \mid \mathbf{u}_{-i} \sim N \left(\frac{\lambda \sum_{i \sim j} u_j}{\lambda m_i + 1 - \lambda}, \frac{\tau_u^{-1}}{\lambda m_i + 1 - \lambda} \right).$$

Neste caso, a distribuição conjunta a priori do vetor $\mathbf{u}$ segue uma distribuição normal com média zero e matriz de covariância:

$$\text{Var}(\mathbf{u} \mid \tau_u, \lambda) = \frac{1}{\tau_u} [\lambda \mathbf{Q} + (1 - \lambda) \mathbf{I}_n]^{-1}, \quad (2.31)$$

em que a diagonal de $\mathbf{Q}$, q_{ii} , contém o número de vizinhos da área i , e $q_{ij} = -1$ se $i \sim j$, e 0 caso contrário. $0 \leq \lambda \leq 1$ denota um parâmetro de dependência espacial, tal que $\lambda = 0$ define

um modelo não espacial (modelo independente) e o nível de dependência espacial aumenta com λ . Quando $\lambda = 1$ o modelo de Leroux reduz ao modelo ICAR. O parâmetro λ pondera o tipo de suavização: a suavização correlacionada dos efeitos aleatórios vizinhos (ponderada por λ), em comparação com a suavização não correlacionada (ponderado por $1 - \lambda$) (LEE; MITCHELL, 2012).

Assim como no modelo BYM, os efeitos aleatórios no modelo Leroux devem ser restritos para evitar um problema de identificabilidade com o intercepto. Se o modelo incluir o intercepto α como um termo separado no segundo estágio da hierarquia, então os efeitos aleatórios devem ser restringidos para somar zero, ou seja, $\sum_{i=1}^n u_i = 0$.

- **Modelo autorregressivo de ordem 1:** Para a componente temporal estruturada de TM_t em (2.28), foi adotado o modelo autorregressivo de ordem 1 (AR1), que pode ser escrito da seguinte forma:

$$\gamma_t = \rho\gamma_{t-1} + \epsilon_t, \quad |\rho| < 1 \text{ e } \epsilon_t \sim N(0, \tau_\gamma^{-1}),$$

em que ϵ_t é um processo de ruído branco com média zero e variância constante σ_γ^2 e ρ é o parâmetro de autocorrelação temporal, que introduz uma estrutura de correlação temporal em γ_t .

2.10.3 Modelo de Bernardinelli

Trata-se de um modelo paramétrico no qual o logarítmo do risco de uma determinada área é considerada uma função linear (ou quadrática) do tempo. Os coeficientes na função são específicos da área e espacialmente estruturados de modo que as áreas vizinhas tenham evoluções semelhantes (LÓPEZ-QUILEZ; MUNOZ, 2009). Bernardinelli et al. (1995) sugerem um modelo no qual tanto o intercepto específico da área quanto a tendência temporal são modelados como efeitos aleatórios. Esta formulação permite interações espaço-temporais onde a tendência temporal do risco pode ser diferente para diferentes localizações espaciais e pode até ter estrutura espacial (LAWSON; BROWNE; RODEIRO, 2003).

Para investigar o aumento linear no risco de sífilis congênita no município de São Paulo de 2010 a 2016, usamos o modelo proposto por Bernardinelli et al. (1995). Neste caso, assumindo que $y_{it} \sim \text{Poisson}(E_{it}\psi_{it})$, o preditor linear para o componente temporal é expresso por:

$$\log(\psi_{it}) = \alpha + u_i + v_i + (\beta + \delta_i)t, \quad (2.32)$$

em que u_i e v_i são respectivamente, efeitos aleatórios espaciais estruturados e não estruturados, βt é o efeito de tendência temporal linear global e $\delta_i t$ representa a interação entre espaço e tempo que define a diferença entre β e a tendência temporal específica da área, referida como a tendência diferencial da i -ésima área. O intercepto α quantifica a taxa média de casos em todas as 310 áreas de ponderação. Cada unidade espacial tem sua própria tendência temporal com um intercepto espacial $\alpha + u_i + v_i$ e uma inclinação $\beta + \delta_i$. Portanto, o risco relativo da doença para a i -ésima área e tempo t é dado por:

$$\psi_{it} = \exp(\alpha + u_i + v_i + (\beta + \delta_i)t), \quad (2.33)$$

para $i = 1, 2, \dots, n$ e $t = 1, 2, \dots, T$.

Para a estimação dos parâmetros, consideramos que a função de verossimilhança é:

$$L(\mathbf{y} \mid \mathbf{u}, \mathbf{v}, \boldsymbol{\delta}, \beta, \tau_u, \tau_v, \tau_\delta). \quad (2.34)$$

Portanto, a distribuição a posteriori é dada por:

$$p(\mathbf{u}, \mathbf{v}, \boldsymbol{\delta}, \beta, \tau_u, \tau_v, \tau_\delta \mid \mathbf{y}) \propto L(\mathbf{y} \mid \mathbf{u}, \mathbf{v}, \boldsymbol{\delta}, \beta, \tau_u, \tau_v, \tau_\delta) p(\mathbf{u})p(\mathbf{v})p(\boldsymbol{\delta})p(\beta). \quad (2.35)$$

A estimação dos parâmetros da Equação (2.35) foi realizada usando o pacote R-INLA.

Uma limitação do modelo de [Bernardinelli et al. \(1995\)](#) é a suposição de uma tendência temporal linear em cada unidade espacial, mas esta limitação foi resolvida pelo modelo de [Knorr-Held \(2000\)](#), apresentado na Seção 2.10.4.

2.10.4 Modelos de interação espaço-tempo

Este é um tipo de modelo de evolução temporal suave onde a evolução do risco estimado em cada área é uma função suave do tempo. [Knorr-Held \(2000\)](#) propôs este modelo para superar a limitação de [Bernardinelli et al. \(1995\)](#), relaxando a restrição de linearidade na tendência temporal diferencial presente no modelo paramétrico (2.33), e especificando uma tendência temporal dinâmica não paramétrica. Para explicar melhor as diferenças na tendência temporal das doenças para as diferentes áreas de ponderação, seria necessário expandir o modelo não paramétrico anterior para permitir uma interação flexível entre espaço e tempo, no qual o logaritmo do risco é modelado como:

$$\log(\psi_{it}) = \alpha + u_i + v_i + \phi_t + \gamma_t + \delta_{it}, \quad (2.36)$$

em que α quantifica o logaritmo do risco médio para todas as áreas, u_i e v_i denotam respectivamente o efeito espacial estruturado e o efeito espacial não estruturado para a área i . Do mesmo modo, ϕ_t e γ_t são respectivamente o efeito temporal estruturado e não estruturado no ano t . Esses componentes levam em consideração os efeitos espaciais e temporais separadamente. O componente δ_{it} é a tendência diferencial que identifica a interação espaço-tempo ([BLANGI-ARDO; CAMELETTI, 2015](#)). Essa interação explica as relações entre as tendências temporais em diferentes locais. Por conveniência computacional, os componentes espaciais/temporais u_i , v_i , ϕ_t , γ_t e δ_{it} seguem uma distribuição gaussiana multivariada com média zero e matriz de precisão $\tau \mathbf{R}$, em que τ é um escalar a ser estimado e $\mathbf{R}$ é uma matriz de estrutura conhecida, e diferente para cada componente ([GILKS; RICHARDSON; SPIEGELHALTER, 1995; KNORR-HELD, 2000](#)).

Para o efeito espacial estruturado u_i , a matriz $\mathbf{R}_u$ possui elementos não diagonais iguais a -1 para áreas vizinhas e elementos diagonais iguais ao número de áreas adjacentes. Todos os outros elementos em $\mathbf{R}_u$ são zero ([KNORR-HELD, 2000](#)). Seguindo essa regra, por exemplo, se

considerarmos 4 áreas numeradas de 1 a 4 e apenas as áreas 1 e 4 como sendo vizinhas, a matriz de estrutura espacial deste exemplo é dada por:

$$\mathbf{R}_u = \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 \end{bmatrix} \quad (2.37)$$

Para o efeito aleatório temporal estruturado γ_t , considerando um processo autorregressivo de primeira ordem e quatro pontos no tempo, a matriz de estrutura é dada por:

$$\mathbf{R}_\gamma = \begin{bmatrix} 1 & -\rho & 0 & 0 \\ -\rho & 1 + \rho^2 & -\rho & 0 \\ 0 & -\rho & 1 + \rho^2 & -\rho \\ 0 & 0 & -\rho & 1 \end{bmatrix} \quad (2.38)$$

As matrizes de estrutura para os efeitos aleatórios espacial e temporal não estruturados v_i e ϕ_t são matrizes identidades I , devido à suposição i.i.d. Uma representação simbólica do modelo de efeitos principais é mostrada na Figura 2.

Como existem dois componentes tanto no efeito espacial quanto no temporal, é fácil notar que há quatro tipos diferentes de interações (KNORR-HELD, 2000) que serão detalhadas a seguir. As representações gráficas desses quatro tipos de interação são mostradas na Figura 3. Gilks, Richardson e Spiegelhalter (1995) sugere que $\mathbf{R}_\delta$ seja especificado como o produto de Kronecker (indicado por $\otimes$ e apresentado no Apêndice B) das matrizes de estrutura dos efeitos espaciais e temporais que são assumidos interagir. Mais detalhes sobre esta especificação podem ser encontrados em Gilks, Richardson e Spiegelhalter (1995).

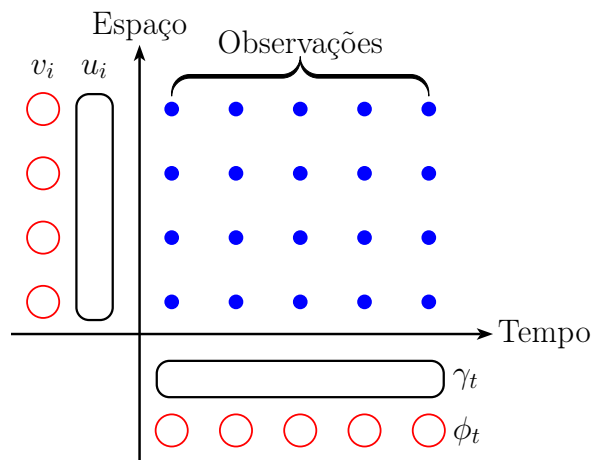


Figura 2 – Representação simbólica do modelo de efeito principal. Os círculos representam a independência a priori, os arredondados representam a dependência a priori. As observações no espaço-tempo são indicadas por pontos pequenos e azuis.

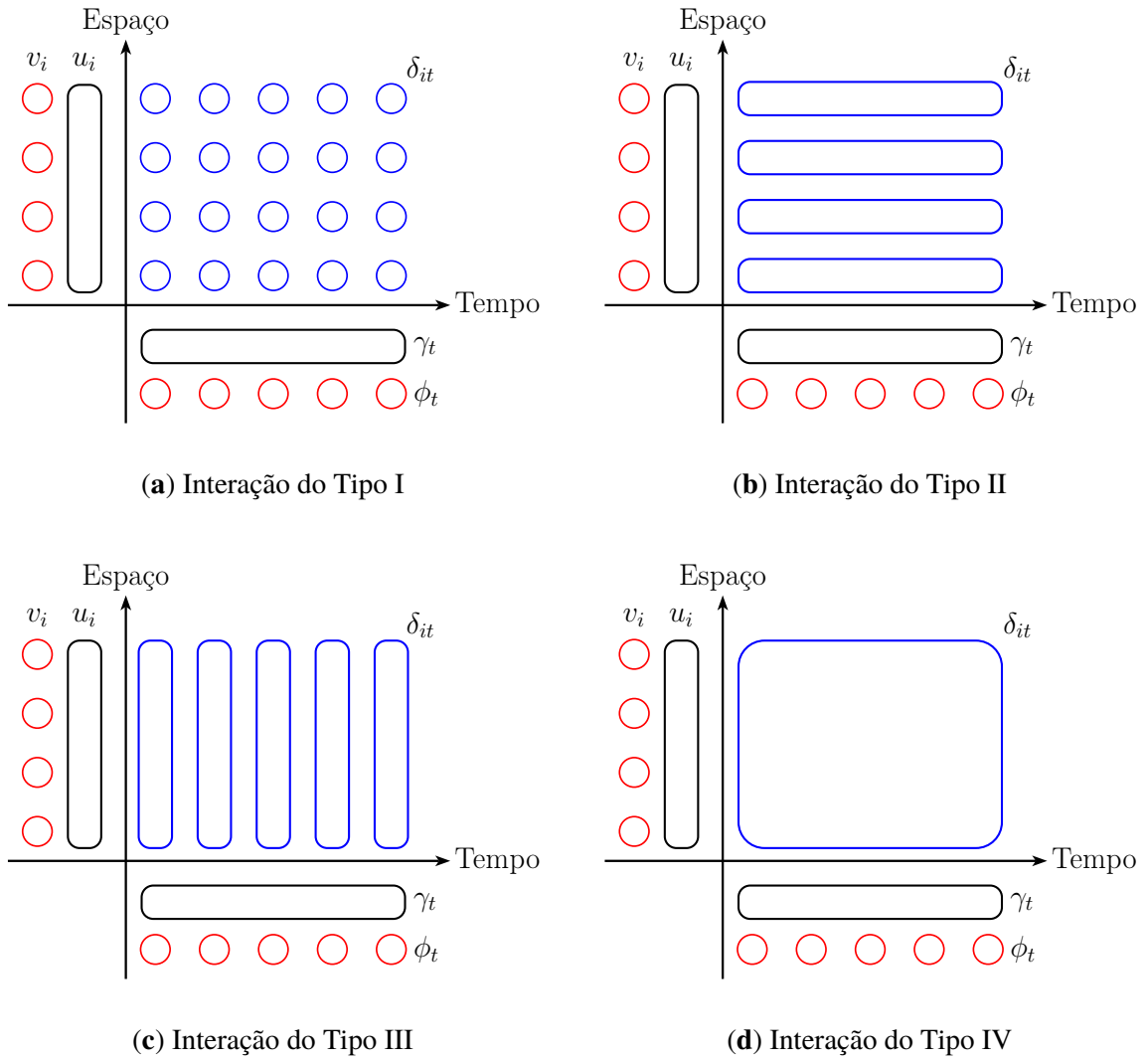


Figura 3 – Representação gráfica para diferentes estruturas de interação espaço-temporal (círculos representam independência, retângulos arredondados representam dependência).

Interação tipo I

A interação do Tipo I assume que os dois efeitos principais não estruturados espacial e temporal (v_i e γ_i) interagem, mas não resultam em nenhuma estrutura no espaço e/ou no tempo (Figura 3d(a)). Na interação do Tipo I, todos os δ_{it} em (2.36) são identicamente independentes com distribuição a priori gaussiana de média zero e precisão τ_δ . A matriz de estrutura ou matriz de precisão pode ser escrita como:

$$\mathbf{R}_\delta = \mathbf{R}_v \otimes \mathbf{R}_\phi = \mathbf{I}_n \otimes \mathbf{I}_T = \mathbf{I}_{nT},$$

em que $\otimes$ denota o produto de Kronecker. A densidade do componente de interação δ é então:

$$p(\delta \mid \tau_\delta) \propto \exp\left(-\frac{\tau_\delta}{2} \sum_{i=1}^n \sum_{t=1}^T \delta_{it}^2\right) = \exp\left(-\frac{1}{2} \delta' \mathbf{Q}_\delta \delta\right),$$

em que $\delta = (\delta_{11}, \dots, \delta_{nT})'$ e $\mathbf{Q}_\delta = \tau_\delta \mathbf{R}_\delta$ é a matriz de precisão. Podemos pensar na interação como covariáveis não observadas para cada observação que não possuem nenhuma estrutura no espaço e no tempo. As interações do tipo I respondem pelo ruído não observado de cada dado nas dimensões espacial e temporal.

Interação tipo II

A interação do tipo II assume que existem interações entre o efeito temporal estruturado γ_t e o efeito espacial não estruturado v_i . Como γ_t segue uma distribuição AR(1), δ_i também segue esta distribuição que é independente de todos os outros. Então, $\delta_i = (\delta_{i1}, \dots, \delta_{iT})'$, $i = 1, 2, \dots, n$ segue um AR(1) independente e δ é i.i.d com média zero.

A matriz de estrutura da interação tipo II é $\mathbf{R}_\delta = \mathbf{R}_v \otimes \mathbf{R}_\gamma$, em que $\mathbf{R}_v = \mathbf{I}_n$ e $\mathbf{R}_\gamma$ é a matriz de estrutura temporal de AR(1). O modelo de interação do tipo II indica que cada região tem uma tendência temporal, e as tendências temporais não apresentam nenhuma estrutura espacial. Como γ_t segue uma distribuição AR(1), δ_{it} segue uma distribuição AR1:

$$\delta_{it} | \delta_{i(t-1)} \sim N(\rho_1 \delta_{i(t-1)}, \tau_\delta^{-1}),$$

para $t = 1, 2, \dots, T$.

A matriz de estrutura $\mathbf{R}_\delta$ tem posto $n(T - 1)$, e a priori para δ é:

$$p(\delta | \tau_\delta) \propto \exp \left(-\frac{\tau_\delta}{2} \sum_{i=1}^n \sum_{t=1}^T (\delta_{it} - \rho_1 \delta_{i(t-1)})^2 \right) = \exp \left(-\frac{1}{2} \delta' \mathbf{Q}_\delta \delta \right),$$

em que $\delta = (\delta'_1, \dots, \delta'_n)'$, e $\mathbf{Q}_\delta = \tau_\delta \mathbf{R}_\delta$ é a matriz de precisão $\mathbf{R}_\delta$ é a matriz de estrutura de tendência temporal.

Interação tipo III

A interação do tipo III assume que existem interações entre o efeito espacial estruturado u_i e o efeito temporal não estruturado ϕ_t . Como u_i segue uma distribuição ICAR, δ_i segue uma distribuição ICAR dada por:

$$\delta_i | \delta_{-i} \sim \text{Normal} \left(\frac{1}{m_i} \sum_{i \sim j} \delta_j, \frac{1}{\tau_\delta m_i} \right). \quad (2.39)$$

A matriz de estrutura da interação do tipo III é $\mathbf{R}_\delta = \mathbf{R}_u \otimes \mathbf{R}_\phi$, em que $\mathbf{R}_\phi = \mathbf{I}_T$ e $\mathbf{R}_u$ é a matriz de estrutura espacial definida através da especificação a priori ICAR. O modelo de interação do tipo III indica que existe uma estrutura espacial diferente para cada ano, mas independentes entre anos distintos. A matriz estruturada $\mathbf{R}_\delta$ tem posto $(n - 1)T$, e a priori para δ é dada por:

$$p(\delta | \tau_\delta) \propto \exp \left(-\frac{\tau_\delta}{2} \sum_{t=2}^T \sum_{i \sim j} (\delta_{it} - \delta_{jt})^2 \right) = \exp \left(-\frac{\tau_\delta}{2} \delta' \mathbf{Q}_\delta \delta \right).$$

Interação tipo IV

A interação do tipo IV assume que existem interações entre o efeito espacial estruturado u_i e o efeito temporal estruturado γ_t . Como u_i produz uma estrutura ICAR e γ_t produz uma distribuição AR(1) e $\delta = (\delta_{11}, \dots, \delta_{nt})'$, $i = 1, \dots, n$ e $t = 1, \dots, T$ depende do espaço e do tempo, a estrutura de dependência temporal para cada região depende, portanto, da estrutura temporal das regiões vizinhas. Como u_i segue uma distribuição ICAR e γ_t segue uma distribuição AR(1), δ_{it} é completamente dependente do espaço e do tempo e, portanto, não pode ser parametrizado como nenhum desses dois efeitos estruturados, ou seja, δ_{it} pode ser considerado agora como uma combinação parcial dos modelos do Tipo II e Tipo III.

A matriz de estrutura da interação do tipo IV é $\mathbf{R}_\delta = \mathbf{R}_u \otimes \mathbf{R}_\gamma$, em que $\mathbf{R}_u$ denota a matriz de estrutura espacial definida pela priori ICAR e com $\mathbf{R}_\gamma$ sendo a matriz de estrutura de tendência temporal definida por meio de uma especificação AR(1). A matriz estruturada $\mathbf{R}_\delta$ tem posto $(T - 1)(n - 1)$. A densidade conjunta a priori para $\mathbf{R}_\delta$ é dada por:

$$p(\boldsymbol{\delta} \mid \tau_\delta) \propto \exp \left(-\frac{\tau_\delta}{2} \sum_{t=1}^T \sum_{i \sim j}^n (\delta_{it} - \delta_{jt} - \rho_1 \delta_{i(t-1)} + \delta_{j(t-1)})^2 \right) = \exp \left(-\frac{\tau_\delta}{2} \boldsymbol{\delta}' \mathbf{Q}_\delta \boldsymbol{\delta} \right),$$

com $\boldsymbol{\delta} = (\delta_{11}, \dots, \delta_{nT})'$, em que $\mathbf{Q}_\delta = \tau_\delta \mathbf{R}_\delta$ é a matriz de precisão e $\mathbf{R}_\delta$ é a matriz de estrutura de tendência temporal. Esta priori assume que existe uma estrutura espacial entre as regiões para um determinado ano, e essa estrutura espacial também é dependente em diferentes anos. De forma equivalente, pode-se enfatizar que existe uma estrutura temporal entre os anos para uma determinada região, e essa estrutura temporal também é dependente entre as regiões vizinhas.

Para efeito de ilustração, voltando agora à modelagem de $\boldsymbol{\delta}$ para interação do Tipo IV, formamos $\mathbf{R}_\delta$, uma matriz de estrutura para $\boldsymbol{\delta}$, tomando o produto de Kronecker das matrizes de estrutura dos efeitos principais espaciais $\mathbf{R}_u$ e temporais $\mathbf{R}_\gamma$. Para ilustrar o cálculo, a Figura 4 apresenta $\mathbf{R}_u$ associado a um mapa de quatro áreas e $\mathbf{R}_\gamma$ associado ao modelo AR(1) em cinco pontos de tempo. Então:

$$\mathbf{R}_u \otimes \mathbf{R}_\gamma = \begin{pmatrix} 2 \cdot \mathbf{R}_\gamma & -\mathbf{R}_\gamma & -\mathbf{R}_\gamma & 0 \cdot \mathbf{R}_\gamma \\ -\mathbf{R}_\gamma & 2 \cdot \mathbf{R}_\gamma & -\mathbf{R}_\gamma & 0 \cdot \mathbf{R}_\gamma \\ -\mathbf{R}_\gamma & -\mathbf{R}_\gamma & 3 \cdot \mathbf{R}_\gamma & -\mathbf{R}_\gamma \\ 0 \cdot \mathbf{R}_\gamma & 0 \cdot \mathbf{R}_\gamma & -\mathbf{R}_\gamma & \mathbf{R}_\gamma \end{pmatrix}_{20 \times 20}. \quad (2.40)$$

A matriz de estrutura resultante é de tamanho 20×20 , que combina com os 20 parâmetros que temos em $\boldsymbol{\delta}$. A matriz resultante, $\mathbf{R}_u \otimes \mathbf{R}_\gamma$, é uma matriz simétrica, e tanto a soma das linhas quanto das colunas é zero. Cada bloco assume a forma de uma entrada escalar em $\mathbf{R}_u$, multiplicando a matriz de estrutura $\mathbf{R}_\gamma$. Um modelo a priori para $\boldsymbol{\delta}$ é então dado por:

$$\pi(\boldsymbol{\delta}) \propto \exp \left\{ -\frac{\tau_\delta}{2} \boldsymbol{\delta}' \mathbf{R}_u \otimes \mathbf{R}_\gamma \boldsymbol{\delta} \right\}. \quad (2.41)$$

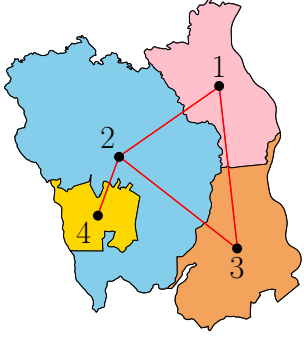
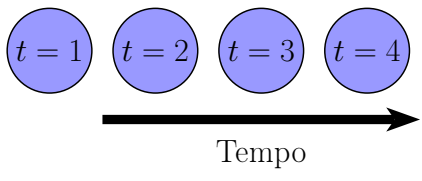
	Configuração	Matriz de estrutura
Espacial		$\mathbf{R}_u = \begin{pmatrix} 2 & -1 & -1 & 0 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 2 & 0 \\ 0 & -1 & 0 & 1 \end{pmatrix}$
Temporal		$\mathbf{R}_\gamma = \begin{pmatrix} 1 & -\rho & 0 & 0 & 0 \\ -\rho & 1 + \rho^2 & -\rho & 0 & 0 \\ 0 & -\rho & 1 + \rho^2 & -\rho & 0 \\ 0 & 0 & -\rho & 1 + \rho^2 & -\rho \\ 0 & 0 & 0 & -\rho & 1 \end{pmatrix}$

Figura 4 – Matriz de estrutura $\mathbf{R}_u$, baseada no movimento de contiguidade espacial queen sobre um mapa de cinco áreas, e matriz de estrutura $\mathbf{R}_\gamma$, associada ao modelo AR1 em cinco pontos de tempo. Fonte: Baseado em [Haining e Li \(2020\)](#).

É importante combinar corretamente os parâmetros em δ com as linhas na matriz de estrutura. Ao tomar o produto Kronecker $\mathbf{R}_u \otimes \mathbf{R}_\gamma$, como fizemos na Equação (2.40), as primeiras T linhas (equivalentemente as primeiras T colunas) em $\mathbf{R}_u \otimes \mathbf{R}_\gamma$ correspondem aos parâmetros temporais da primeira área, $\delta_{11}, \dots, \delta_{1T}$, as T linhas subsequentes (T colunas) correspondem aos parâmetros temporais da segunda área, $\delta_{21}, \dots, \delta_{2T}$, e assim por diante. No entanto, se o produto Kronecker for tomado como $\mathbf{R}_\gamma \otimes \mathbf{R}_u$ (ou seja, trocando as posições das duas matrizes de estrutura), então a rotulagem das linhas será diferente, em que as n primeiras linhas (equivalentemente as n primeiras colunas) em $\mathbf{R}_\gamma \otimes \mathbf{R}_u$ correspondem aos parâmetros específicos de área do primeiro ponto de tempo, $\delta_{11}, \dots, \delta_{n1}$, as n linhas subsequentes (n colunas) correspondem aos parâmetros específicos de área do segundo ponto de tempo, $\delta_{12}, \dots, \delta_{n2}$ e assim por diante ([HAINING; LI, 2020](#)).

De modo similar ao caso espacial, problemas de identificabilidade surgem no modelo (2.36), pois o nível geral pode ser absorvido tanto pelos efeitos espaciais quanto temporais. Além disso, o termo de interação é confundido com os efeitos principais. Para garantir a identificabilidade do modelo, geralmente são impostas restrições de soma zero sobre os efeitos aleatórios do modelo (Veja a Tabela 2). [Goicoa et al. \(2018\)](#) apresentam os detalhes sobre a derivação dessas restrições.

Tabela 2 – Restrições de identificabilidade para os modelos espaço-temporais na Equação 2.36

Interação	Espacial		Temporal
Tipo I	BYM	$\sum_{i=1}^n (u_i + v_i) = 0$	$\sum_{t=1}^T \gamma_t = 0, \sum_{i=1}^n \sum_{t=1}^T \delta_{it} = 0$
	BYM2	$\sum_{i=1}^n b_i = 0$	
	LEROUX	$\sum_{i=1}^n u_i = 0$	
Tipo II	BYM	$\sum_{i=1}^n (u_i + v_i) = 0$	$\sum_{t=1}^T \gamma_t = 0, \sum_{t=1}^T \delta_{it} = 0, \text{ para } i = 1, \dots, n$
	BYM2	$\sum_{i=1}^n b_i = 0$	
	LEROUX	$\sum_{i=1}^n u_i = 0$	
Tipo III	BYM	$\sum_{i=1}^n (u_i + v_i) = 0$	$\sum_{t=1}^T \gamma_t = 0, \sum_{i=1}^n \delta_{it} = 0, \text{ para } t = 1, \dots, T$
	BYM2	$\sum_{i=1}^n b_i = 0$	
	LEROUX	$\sum_{i=1}^n u_i = 0$	
Tipo IV	BYM	$\sum_{i=1}^n (u_i + v_i) = 0$	$\sum_{t=1}^T \gamma_t = 0,$ $\sum_{t=1}^T \delta_{it} = 0, \text{ para } i = 1, \dots, n$ $\sum_{i=1}^n \delta_{it} = 0, \text{ para } t = 1, \dots, T$
	BYM2	$\sum_{i=1}^n b_i = 0$	
	LEROUX	$\sum_{i=1}^n u_i = 0$	

2.11 Prioris PC

Recentemente, as priors PC (priors de penalidade complexa) foram propostas como uma forma de prevenir sobreajuste. Elas são baseadas em quatro princípios, que resumimos e ilustramos abaixo para o parâmetro de precisão τ de um efeito aleatório gaussiano. Para mais detalhes, consultar [Rue et al. \(2017\)](#).

Seja π_1 a densidade de um componente do modelo w com parâmetro de precisão τ . Este componente do modelo pode ser visto como uma extensão flexível de um modelo base com densidade π_0 e $\tau = \infty$ (ou seja, ausência de efeitos aleatórios). Os quatro princípios são:

1. Parcimônia: A priori para τ deve fornecer encolhimento adequado para $\tau = \infty$ e decair com complexidade crescente de π_1 , de modo que o modelo mais simples seja favorecido, a menos que haja evidência de um mais flexível.
2. A complexidade aumentada de π_1 em relação a π_0 é medida usando a divergência de Kullback-Leibler (KLD, Kullback e Leibler, 1951),

$$\text{KLD}(\pi_1 \| \pi_0) = \int \pi_1(w) \log \left(\frac{\pi_1(w)}{\pi_0(w)} \right) dw$$

Para facilitar a interpretação, o KLD é transformado em uma medida de distância unidirecional:

$$d(\tau) = d(\pi_1 \| \pi_0) = \sqrt{2 \text{KLD}(\pi_1 \| \pi_0)}$$

que pode ser interpretada como a distância do modelo flexível π_1 ao modelo base π_0 .

3. A priori PC é definida como uma distribuição exponencial na distância,

$$\pi(d(\tau)) = \lambda \exp(-\lambda d(\tau))$$

com taxa $\lambda > 0$. O parâmetro λ determina a magnitude da penalidade devido ao afastamento do modelo base e, valores mais altos aumentam esta penalidade. A priori PC para τ segue por uma mudança de transformação de variável, levando neste caso a uma distribuição de Gumbel tipo 2 com parâmetros $(1/2, \lambda)$:

$$\pi(\tau) = \frac{\lambda}{2} \tau^{-3/2} \exp(-\lambda \tau^{-1/2}), \quad \tau > 0, \lambda > 0 \quad (2.42)$$

4. O parâmetro λ em (2.42) pode ser selecionado pelo usuário com base em seu conhecimento a priori de τ (ou uma transformação interpretável dele, como o desvio padrão). Isso pode ser expresso de forma intuitiva com uma probabilidade, por exemplo, definindo U e α tal que $P(1/\sqrt{\tau} > U) = \alpha$ de modo que $\lambda = -\log(\alpha)/U$. O conhecimento do desvio padrão marginal pode auxiliar na escolha de um valor sensível para U e Rue et al. (2017) fornecem uma regra prática: uma vez que a precisão τ é integrada, o desvio padrão marginal do efeito aleatório para $\alpha = 0,01$ é de cerca de $0,31U$.

3 Um estudo espaço-temporal da sífilis congênita no Município de São Paulo no período de 2010 a 2016 usando o INLA

Resumo

O estudo teve como objetivo analisar o risco espaço-temporal de sífilis congênita (SC) nas áreas de ponderação do município de São Paulo-SP, considerado de grande porte e avaliar sua relação com variáveis socioeconômicas, demográficas e ambientais. Desenvolvemos um estudo ecológico baseado em dados secundários de SC com componentes espaço-temporais de 310 áreas incluindo dados de 2010 a 2016. Modelamos os dados em um contexto Bayesiano, considerando efeitos aleatórios espaciais e temporais, usando distribuições de probabilidade de Poisson e Binomial Negativa empregando o método de Aproximação de Laplace Aninhada e Integrada (INLA). Três diferentes efeitos aleatórios espaciais foram aplicados e comparados. O primeiro é baseado no modelo espacial BYM (Besag, York and Mollié), o segundo no modelo BYM2, enquanto o terceiro é baseado no modelo espacial de Leroux. Para os efeitos temporais, consideramos um modelo autorregressivo de primeira ordem. Os mapas de risco apontaram tendência de aumento da SC e evidenciaram as áreas que apresentaram maior e menor risco em cada ano. O modelo mostrou que, os fatores associados positivamente ao maior risco de ocorrência desse agravo foram índice de Gini e proporção de mulheres de 18 a 24 anos sem instrução ou com ensino fundamental incompleto, enquanto que proporção de mulheres em idade fértil, e renda média per capita apresentaram associação negativa. Embora ações de controle dessa doença sejam necessárias em todas as áreas do município, a identificação de áreas de alto risco aponta para regiões prioritárias que necessitam do desenvolvimento de possíveis ações e monitoramento da doença.

Abstract

The study aimed to analyze the spatio-temporal risk of congenital syphilis (CS) in the weighted areas of the city of São Paulo-SP, considered to be large, and to assess its relationship with socioeconomic, demographic and environmental variables. We developed an ecological study based on secondary data of SC with spatiotemporal components from 310 areas including data from 2010 to 2016. We modeled the data in a Bayesian context, considering spatial and temporal random effects, using Poisson and Negative Binomial probability distributions applying the Nested and Integrated Laplace Approximation (INLA) method. Three different spatial random effects were applied and compared. The first is based on the BYM spatial model (Besag, York and Mollié), the second on the BYM2 model, while the third is based on the Leroux spatial model. For temporal effects, we consider a first-order autoregressive model. The risk maps showed a tendency to increase CS and showed the areas that presented the highest and lowest risk in each year. The model showed that the factors positively associated with a higher risk of occurrence of this condition were the Gini index and the proportion of women aged 18 to 24 years without education or with incomplete elementary education, while the proportion of women of childbearing age and average income per capita showed a negative association. Although actions to control this disease are necessary in all areas of the municipality, the identification of high-risk areas points to priority regions that need the development of possible actions and monitoring of the disease.

3.1 Introdução

A redução da sífilis congênita (SC) ainda representa um desafio para gestores e profissionais da área de saúde em vários países, especialmente aqueles em desenvolvimento, e ainda distante das metas propostas pelo Ministério da Saúde, apesar do tratamento ser de baixo custo, de fácil acesso e simples de ser realizado. Essa doença ocorre devido à disseminação hematogênica do *Treponema pallidum* de uma gestante infectada ou tratada de forma inadequada para o feto e, quando as mães apresentam lesões genitais no canal parto, pode ocorrer a transmissão direta da bactéria de mãe para o recém-nascido (AVELLEIRA; BOTTINO, 2006; BRUNNER; SUDDARTH, 1987; BRASIL, 2005).

A SC é subdividida em dois tipos. O primeiro, denominada precoce, quando as manifestações clínicas da doença ocorrem entre 0 e 2 anos de idade, tendo como principais características: baixo peso ao nascer, lesões de pele, desconforto respiratório, anemia, icterícia e pseudoparalisia dos membros¹. No segundo tipo, referido como tardio, a criança com mais de 2 anos pode

¹ Também conhecida como pseudoparalisia de Parrot, é caracterizada por periostite com dor à movimentação ativa ou passiva dos membros, principalmente superiores, havendo irritabilidade e imobilidade.

apresentar arco palatino elevado, surdez neurológica, articulações de Clutton², dificuldades de aprendizagem, entre outros problemas de saúde (SANDES et al., 2019).

Segundo estimativa da Organização Mundial de Saúde (OMS), houve mais de meio milhão (cerca de 661.000) de casos totais de SC em 2016, resultando em mais de 200.000 natimortos e mortes neonatais. A SC é a segunda principal causa de natimortalidade evitável em todo o mundo, precedida apenas pela malária (WHO, 2019). Na maioria dos países, é necessário a obtenção de estimativas precisas para reflexão do gravame da doença, pois o seu diagnóstico é muito mais difícil do que da sífilis em gestantes (WIJESOORIYA et al., 2016). No Brasil, a situação não é diferente de outros países. Os números de casos de sífilis adquirida, sífilis em gestantes e sífilis congênita são preocupantes e a infecção precisa ser controlada. De acordo com Cooper e Sánchez (2018), cada caso de SC deve ser visto como uma falha no sistema de saúde pública em fornecer atendimento pré-natal ideal para mulheres grávidas, já que a SC pode ser prevenida por exames sorológicos, pré-natais nas gestantes e tratamento com penicilina das mulheres infectadas, seus parceiros sexuais e bebês recém-nascidos.

O estudo da evolução da SC pode ser abordado usando o mapeamento de doenças, o que pode ajudar na avaliação do desempenho das medidas de controle e contenção. Em particular, a quantificação da interação espaço-tempo, descrevendo como o padrão espacial muda ao longo do tempo, tem sido proposta como forma de aprofundar nossa compreensão sobre a evolução da doença (DIGGLE et al., 1995).

Os modelos de mapeamento de doenças (ELLIOTT; WARTENBERG, 2004; LAWSON, 2018; MORAGA, 2019) visam descrever as variações no risco de uma determinada doença ao longo do espaço e do tempo. Os dados geralmente estão disponíveis na forma de contagens agregadas em algum nível espacial, como, municípios, áreas de ponderação, setores censitários, etc., por vários motivos, como por exemplo, a confiabilidade do paciente. Modelos de tempo e espaço aditivos têm sido usados há muito tempo para a modelagem das taxas de doenças (KNORR-HELD; BESAG, 1998). Mais recentemente, com a disponibilidade e a complexidade dos dados houve um interesse em modelagens mais complexas que incluem um termo de interação para capturar adequadamente as relações espaço-temporais nos dados (consultar, por exemplo, Knorr-Held (2000) e Bernardinelli et al. (1995)). Compreender a distribuição espacial do risco de doença ou como ela evolui ao longo do tempo pode ser útil para os órgãos de saúde pública no planejamento e manejo de recursos e identificação de áreas a serem priorizadas.

As taxas brutas podem ser irreais devido à variabilidade da amostragem, de modo que a suavização é usada para obter informações entre áreas vizinhas e pontos do tempo. Assim, o mapeamento de doenças tem-se desenvolvido principalmente em uma estrutura hierárquica Bayesiana, na qual as suavizações em uma ou mais dimensões podem ser modeladas usando campos aleatórios gaussianos de Markov intrínsecos (IGMRF, do inglês, *intrinsic gaussian*

² Sinovite que acomete geralmente os joelhos (bilateral) com um aumento de volume, indolor e sem alterações radiográficas.

Markov random fields) , tais como o passeio aleatório de primeira e segunda ordem (LINDGREN; RUE; LINDSTRÖM, 2011), modelos autorregressivos de primeira ordem e segunda ordem ou os modelos ICAR (BESAG, 1974), entre outros. Na modelagem de interações, pesquisadores têm usado suavizadores de produtos tensoriais, em que, numa estrutura Bayesiana, a penalidade é baseada em produtos de Kronecker (RUE; HELD, 2005).

Uma das vantagens da estrutura Bayesiana é que sempre que a informação sobre o processo da doença estiver disponível, ela pode ser especificada na priori e, muitas vezes o epidemiologista pode ter uma intuição sobre a importância do termo de interação para explicar a variação espaço-temporal de uma determinada doença. No entanto, traduzir essas informações em termos de um parâmetro de precisão não é nada trivial (FRANCO-VILLORIA; VENTRUCCI; RUE, 2021).

Neste estudo, consideramos os modelos espaço-temporais propostos por Knorr-Held (2000), em que o termo de interação espaço-tempo pode ser de quatro diferentes tipos, dependendo do grau de dependência assumido entre tempo e espaço. Esses quatro tipos são caracterizados a priori por diferentes pressupostos, expressos em termos de um produto Kronecker.

A aplicação de modelos Bayesianos na análise de conjuntos de dados espaço-temporais é computacionalmente desafiador usando algoritmos MCMC (BEST; RICHARDSON; THOMSON, 2005; BLANGIARDO; CAMELETTI, 2015) pois requer avaliação iterativa de várias funções de densidade condicional completa e a convergência das distribuições de probabilidade a posteriori requer extensivas simulações (BANERJEE; CARLIN; GELFAND, 2015; CARROLL et al., 2015). Recentemente, Rue, Martino e Chopin (2009) propuseram o método Integrated Nested Laplace Approximation (INLA), que aproxima distribuições de probabilidade posterioris por meio de integração numérica, ao invés de um processo iterativo baseado em MCMC. Pesquisas anteriores comparando abordagens MCMC e INLA tem mostrado que o INLA reduz o tempo computacional e obtém estimativas de parâmetros confiáveis (RUIZ-CÁRDENAS; KRAINSKI; RUE, 2012; BLANGIARDO et al., 2013). A importância da eficiência computacional ao analisar dados espaço-temporais está no fato de que, várias distribuições e suposições a priori para os parâmetros de efeitos aleatórios espaciais, temporais e espaço-temporais podem ser testados e comparados (KNORR-HELD, 2000). O INLA tem resolvido problemas inferenciais em muitos estudos com aplicações espaço-temporais. Por exemplo, é aplicado a dados climáticos globais (LINDGREN; RUE; LINDSTRÖM, 2011), epidemiologia (BISANZIO et al., 2011; MEDEIROS et al., 2022), mapeamento e disseminação de doenças (SCHRÖDLE; HELD, 2011b; SCHRÖDLE; HELD; RUE, 2012), mapeamento do risco de poluição do ar (CAMELETTI et al., 2013) e econometria (GOMÉZ-RUBIO; BIVAND; RUE, 2015).

Esta pesquisa analisa os padrões espaço-temporais de sífilis congênita em pequenas áreas no Município de São Paulo, Brasil. Os casos de sífilis congênita foram agregados temporalmente a 7 períodos de tempo, e geograficamente agregadas a 310 áreas de ponderação, resultando em um total de 2170 unidades espaço-temporais. Neste contexto, investigamos os padrões espaço-

temporal da sífilis congênita, usando dados notificados de casos e caracterizando as áreas com riscos relativos excepcionalmente altos. Os mapas gerados a partir desses modelos descrevem associações que garantem mais investigações epidemiológicas sobre os desafios e oportunidades existentes para vigilância da doença.

Assim, este estudo teve como objetivo estimar no espaço e no tempo o risco de sífilis congênita e avaliar sua relação com as variáveis socioeconômicas e ambientais por meio de modelos espaço-temporais utilizando abordagem Bayesiana com o algoritmo INLA, e descrever um padrão espaço-temporal da SC em nível de áreas de ponderação no município de São Paulo de 2010 a 2016 explorando os fatores de risco por meio de ferramentas de mapeamento espaço-temporal e dados do SINAN e IBGE.

3.2 Materiais e Métodos

3.2.1 Desenho e área do estudo

Foi realizado um estudo ecológico espaço-temporal no município de São Paulo, região sudeste do Brasil (Figura 5), com os casos agregados de sífilis congênita, utilizando-se as áreas de ponderação (AP) como unidades de análise espacial e a unidade temporal foi o ano, variando de 2010 a 2016.

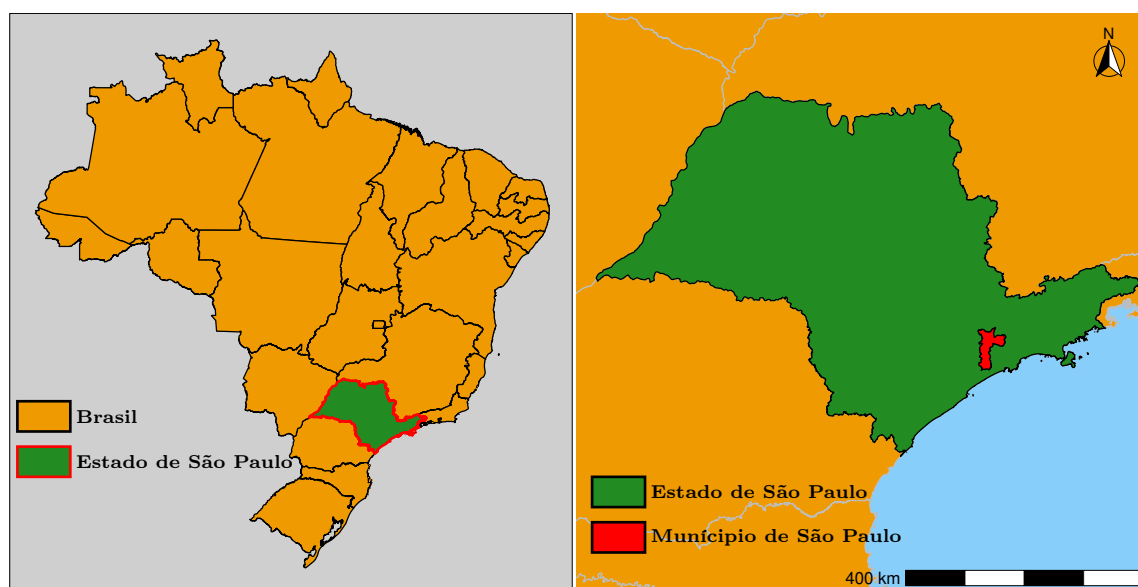


Figura 5 – Localização geográfica do município São Paulo-SP, Brasil.

O município de São Paulo é composto por 310 APs, ocupa uma área de 1521,11 km², tem uma população de aproximadamente 12,5 milhões de habitantes (IBGE, 2020), e densidade

demográfica de 7398,26 hab/km². Cada AP, por sua vez, é uma unidade geográfica formada por um agrupamento de setores censitários, com tamanho mínimo de 400 domicílios ocupados, para aplicação de procedimentos de calibração das estimativas obtidas na amostra com as informações conhecidas pela população como um todo, para evitar perda de precisão nas estimativas (IBGE, 2010).

As variáveis socioeconômicas e ambientais foram coletadas do IBGE (Instituto Brasileiro de Geografia e Estatística), que contém variáveis sobre tamanho da população, educação, moradia, saúde, trabalho, renda e vulnerabilidade.

3.2.2 Fonte dos dados e variáveis do estudo

Conduzimos este estudo por meio de uma extensa análise dos dados de notificação de casos de SC do banco de dados do SINAN (Sistema de Informação de Agravos de Notificação) para o município de São Paulo. O banco de dados é um sistema de vigilância robusto que captura os dados de notificação de casos das unidades de saúde de cada município e atualiza os registros na rede nacional. Para o processo de obtenção dos dados no sistema de vigilância, são utilizadas as fichas de notificação individual de cada agravo e os dados epidemiológicos são publicados periodicamente nos Boletins Epidemiológicos específicos (disponíveis em <http://www.aids.gov.br/pagina/publicacoes>).

O estudo utilizou três fontes de informação: bases de dados com os casos notificados de sífilis congênita do SINAN no período de 2010 a 2016, os dados de nascimentos coletados no Sistema de Informações sobre Nascidos Vivos (SINASC) e dados do censo demográfico de 2010 do IBGE.

O mapa das áreas de ponderação foi obtido do IBGE (<https://www.ibge.gov.br/>). Acesso em: 25 de out. de 2021). A união do *shapefile* de casos com o *shapefile* das áreas de ponderação foi realizado por meio do *software* QGIS versão 3.10, permitindo identificar, além da distribuição dos casos no município, as regiões em que estavam inseridos.

Para determinar a relação espaço-temporal entre os casos de sífilis congênita e variáveis explicativas, usamos uma variedade de modelos para mapeamento espaço-temporal da doença.

A variável resposta foi o número de casos de sífilis congênita em cada área de ponderação. Para identificar o melhor modelo para a variável resposta de contagem neste estudo, comparamos vários modelos como potenciais candidatos.

No processo de modelagem espaço-temporal, considerou-se como variáveis independentes condições socioeconômicas e ambientais a partir de variáveis do Censo 2010, já utilizadas em estudos anteriores (MELO; FILHO; FERREIRA, 2011; OLIVEIRA et al., 2014; SERAFIM et al., 2014; DOMINGUES; LEAL, 2016; REIS et al., 2018; RUBERTI, 2020) que demonstraram alguma relação com a incidência de sífilis congênita. Foram selecionadas as seguintes variáveis explicativas como possíveis preditoras:

- **Proporção de mulheres em idade fértil (PFERTIL):** Razão entre o número de mulheres de 15 a 49 anos e o total de mulheres, multiplicada por 100.
- **Proporção de nascidos vivos de mães adolescentes (PNVMA):** Razão entre o número de nascidos vivos de mulheres de 12 a 17 anos e o total de nascidos vivos, multiplicada por 100.
- **Percentual de mulheres de 18 a 24 anos de idade sem instrução ou com ensino fundamental incompleto (MSIF1824):** Razão entre a população de mulheres de 18 a 24 anos sem instrução ou com ensino fundamental incompleto e o total de mulheres nesta faixa etária, multiplicada por 100.
- **Índice de Gini (GINI):** Mede o grau de desigualdade existente na distribuição de indivíduos segundo a renda domiciliar per capita. Seu valor é 0 quando não há desigualdade (a renda domiciliar per capita de todos os indivíduos tem o mesmo valor) e tende a 1 à medida que a desigualdade aumenta. O universo de indivíduos é limitado àqueles que vivem em domicílios particulares permanentes.
- **Renda per capita média (RDPC):** Razão entre o somatório da renda de todos os indivíduos residentes em domicílios particulares permanentes e o número total desses indivíduos. Valores em reais de 01/agosto de 2010.
- **Percentual da população que vive em domicílios com banheiro e água encanada (PBA):** Razão entre a população que vive em domicílios particulares permanentes com água encanada em pelo menos um de seus cômodos e com banheiro exclusivo e a população total residente em domicílios particulares permanentes, multiplicada por 100. A água pode ser proveniente de rede geral, de poço, de nascente ou de reservatório abastecido por água das chuvas ou carro-pipa. Banheiro exclusivo é definido como cômodo que dispõe de chuveiro ou banheira e aparelho sanitário.
- **Proporção de pessoas de cor parda (PPARDA):** Percentual de residentes que se autodeclararam pardos, multiplicado por 100.
- **Proporção de pessoas de cor preta (PPRETA):** Percentual de residentes que se autodeclararam pretos, multiplicado por 100.
- **Proporção de pessoas de cor branca (PBRANCA):** Percentual de residentes que se autodeclararam brancos, multiplicado por 100.

3.2.3 Multicolinearidade

Antes da modelagem, uma etapa importante é a seleção de variáveis representativas de uma variedade de variáveis potenciais. Para a seleção das variáveis socioeconômicas e ambientais neste estudo, o fator de inflação da variância (VIF, do inglês *variance inflation factor*) foi calculado para avaliar a multicolinearidade, que é um problema potencial em todas as análises de regressão. Quanto maior o seu valor, mais severa é a multicolinearidade. Aqui consideramos $VIF < 5$ como o padrão de triagem (VATCHEVA et al., 2016). Geralmente, diferentes fatores representativos do nível socioeconômico possuem multicolinearidade, e uma maneira conveniente de lidar com essa questão é escolher aqueles que são mais representativos (com o menor VIF) ao invés de selecionar todos eles. O mesmo se aplica às variáveis ambientais.

O VIF é definido por

$$VIF_j = \frac{1}{1 - R_j^2}$$

em que R_j^2 é o R^2 da regressão linear da j -ésima variável explicativa sobre as demais variáveis explicativas do modelo.

A presença de multicolinearidade pode ser particularmente problemática para estudos que usam métodos que dependem de estimativas dos parâmetros por intervalos. Uma consequência comum é o aumento nos erros padrão das estimativas, podendo levar a intervalos de confiança maiores, ocasionando impacto no julgamento da importância de um preditor. Entre pesquisadores aplicados e teóricos, não há consenso sobre quanta ênfase colocar em um valor VIF, nem qual ponto de corte é mais apropriado. Alguns pesquisadores afirmam que um VIF maior que 10 indica a presença de multicolinearidade (O'BRIEN, 2007; GARCÍA et al., 2016; THOMPSON et al., 2017).

A função em linguagem R utilizada para o cálculo do VIF denominado `vif_func`, encontra-se no Apêndice E (WANG, 2018). Vários pacotes em R também fornecem funções para calcular o VIF, tais como: `HH` (HEIBERGER; HEIBERGER, 2022), `car` (FOX et al., 2007), `fmsb` (NAKAZAWA; NAKAZAWA, 2019) e `faraway` (FARAWAY, 2016).

3.2.4 Estimação do risco relativo para dados agregados

Geralmente, em uma abordagem Bayesiana hierárquica, os modelos lineares mistos generalizados (GLMM) são usados para a análise de dados de contagem. Condicionada ao risco relativo ψ_i , o número de casos observados na i -ésima área é assumido como tendo distribuição de Poisson com média $\mu_i = E_i\psi_i$, ou seja,

$$y_i|\psi_i \sim \text{Poisson}(E_i\psi_i), \quad i = 1, 2, \dots, n$$

$$\log(\psi_i) = \eta_i$$

Nas análises realizadas no decorrer da tese, para o mapeamento dos casos de sífilis congênita, o principal interesse foi utilizar as informações espaciais e temporais presentes no

banco dados. Os dados utilizados estão disponíveis como contagens para cada AP do município de São Paulo.

Neste estudo, acrescentando a questão temporal para unidade de análise (AP), y_{it} e E_{it} são respectivamente o número de casos e o número esperado de casos dentro da i -ésima área e ano t , $i = 1, \dots, n$ e $t = 1, 2, \dots, T$. As taxas esperadas são calculadas como $E_{it} = R_t \cdot NV_{it}$, em que $R_t = \sum_{i=1}^n y_{it} / \sum_{i=1}^n NV_{it}$ e NV_{it} é o número de nascidos vivos na i -ésima AP e ano t no município de São Paulo durante o período de 2010-2016. O número de casos esperados de SC, em cada área de ponderação e ano, foi calculado com base na taxa de incidência de SC global para todo o município e em todo o período de estudo. Geralmente supõe-se que as contagens de casos são variáveis independentes com distribuição Poisson, ou seja, $Y_{it} \sim \text{Poisson}(\mu_{it})$, $\mu_{it} = E_{it}\psi_{it}$ em que ψ_{it} é o risco relativo para a i -ésima área no tempo t . A estimativa do risco relativo da área i é o estimador de máxima verossimilhança e é conhecida como razão de incidência padronizada (SIR, do inglês *standardized incidence ratio*), que é uma razão bruta definida por $SIR_{it} = y_{it}/E_{it}$. Apesar dessa estimativa ser bastante conhecida e fácil de calcular, ela pode ser enganosa, especialmente quando aplicada a doenças raras, já que apresenta instabilidade frente a pequenos valores observados e/ou para populações pequenas. No entanto, o SIR pode ser utilizado em uma análise exploratória para visualizar a forma como as taxas da doença ou *clusters* se distribuem na área de estudo. Estimativas mais confiáveis dos riscos relativos de doenças raras em pequenas áreas podem ser obtidas usando modelagem hierárquica Bayesiana por meio de informações de áreas vizinhas (LAWSON; ROTEJANAPRASERT, 2014). Uma maneira de obter a relação entre áreas vizinhas é usando modelagem hierárquica que inclui informações de vizinhança e essa configuração pode ser tratada naturalmente em uma estrutura Bayesiana.

Antes da modelagem, procedeu-se a análise exploratória para avaliar a existência de *outliers* (valores discrepantes). Além disso, todas covariáveis foram padronizadas (subtração da média e divisão pelo desvio padrão), evitando que o ajuste fique enviesado para as variáveis com maior ordem de grandeza. Já na modelagem, os casos esperados de SC foram considerados como *offsets*³, e portanto, os resultados obtidos foram interpretados na escala natural como risco relativo (RR) para a incidência de SC. Inicialmente, foram ajustados apenas os modelos com intercepto (fixo) e efeitos aleatórios (espaciais e temporais) e, posteriormente, foram considerados os modelos com as covariáveis.

3.2.5 Métodos de avaliação e qualidade de ajuste do modelo

Um aspecto importante da modelagem Bayesiana diz respeito à avaliação de sua plausibilidade e ajuste. A primeira, também chamada de crítica do modelo, pode ser definida como a avaliação de quais variáveis incluir, quais suposições fazer sobre os parâmetros (por exemplo,

³ Como o número de nascidos vivos por região é diferente, não podemos analisar o número de casos de sífilis congênita isoladamente. Desse modo, para o ajuste, devemos incluir uma componente no modelo, chamada "offset", que é responsável por controlar o número de casos de SC pelo número de nascidos vivos. O *offset* é incluído em escala logarítmica, para estar de acordo com a função de ligação utilizada.

permutabilidade, independência, etc.) e sobre a verossimilhança (Normal, Distribuição t de Student, etc.), cuja distribuição a priori deve ser atribuída a parâmetros e hiperparâmetros.

3.2.5.1 Complexidade e adequação do modelo

O critério de informação de deviance (DIC, do inglês *deviance information criterion*) é um método de avaliação no processo de modelagem bastante conhecido na literatura para comparar ajuste e complexidade dos modelos em uma abordagem Bayesiana. De acordo com Spiegelhalter et al. (2002), o DIC é expresso por:

$$DIC = \bar{D} + P_D, \quad (3.1)$$

em que $\bar{D}$ é a média a posteriori da deviance (medida para o ajuste do modelo) e P_D é o número efetivo de parâmetros (medida para a complexidade do modelo) e, frequentemente é menor que o número total de parâmetros do modelo. A quantidade P_D é definida como a média a posteriori da deviance menos a deviance da média a posteriori dos parâmetros de interesse. Menores valores do DIC e P_D indicam o modelo com melhor ajuste. Modelos com menor DIC indicam uma melhor compensação entre ajuste e complexidade do modelo. Spiegelhalter et al. (2003) sugerem que diferenças entre DIC's maiores que 10 podem definitivamente descartar o modelo com o DIC mais alto, diferenças entre 5 e 10 são substanciais, mas se a diferença de DIC for, digamos, menor que 5, e os modelos apresentarem muitas diferenças nas inferências, então pode ser enganoso apenas considerar o modelo com o menor DIC.

Outra alternativa é o critério de informação de Watanabe-Akaike (WAIC, do inglês *Watanabe-Akaike information criterion*), que pode ser visto como uma melhoria no DIC para modelos Bayesianos (WATANABE; OPPER, 2010). Ao contrário do DIC, o WAIC é invariável à parametrização e também funciona para modelos singulares. O WAIC é interpretado como uma aproximação computacionalmente conveniente para validação cruzada e é definido por (VEHTARI; GELMAN; GABRY, 2017):

$$WAIC = LDP + P_W \quad (3.2)$$

em que LDP é o logaritmo da densidade preditiva pontual esperada e P_W é o número efetivo estimado de parâmetros. A explicação do WAIC é semelhante ao DIC.

3.2.5.2 Qualidade preditiva do modelo

A ordenada preditiva condicional (CPO, do inglês *conditional predictive ordinate*) foi definida pela primeira vez por Geisser e Eddy (1979) como um diagnóstico para detectar observações discrepantes de um determinado modelo, e é definida como uma densidade preditiva de validação cruzada em que o valor na unidade é previsto a partir dos dados restantes e comparado com os dados observados na unidade (BLANGIARDO; CAMELETTI, 2015). Para distribuições contínuas, é definida da seguinte forma:

$$CPO_{it} = p(y_{it}^* | \mathbf{y}_{-it}), \quad (3.3)$$

em que y_{it}^* é o valor observado na unidade espaço-temporal (i, t) e $\mathbf{y}_{-it}$ é o vetor que contém todas as observações, exceto aquela correspondente à unidade (i, t) . Observe que CPO_{it} pode ser usado como medida local de ajuste do modelo em que $0 < \text{CPO}_{it} < 1$ e valores mais próximos de 1 sugerem bom ajuste, enquanto valores próximos de zero sugerem dados mal ajustados. Um pequeno valor de CPO_{it} sugere que a observação em (i, t) é um *outlier* com base no resto das observações (PETTIT, 1990). Além disso, para a escolha do modelo pode ser usado o log da Pseudo Verossimilhança Marginal (LPML, do inglês *logarithm of the pseudomarginal likelihood*) calculado como $\text{LPML} = \sum_{i=1}^n \sum_{t=1}^T \log(\text{CPO}_{it})$ (HELD; SCHRÖDLE; RUE, 2010). Os valores de LPML fornecidos por diferentes modelos podem ser comparados para identificar qual modelo possui maior capacidade preditiva. Também neste caso seguimos a regra de que o modelo com o menor valor de LPML será o melhor.

3.2.5.3 Precisão e qualidade de ajuste do modelo

Para comparar a qualidade do ajuste e desempenho dos modelos, apresentamos algumas funções matemáticas, referidas como métricas estatísticas comumente usadas em ciências ambientais e sociais, a saber, raiz quadrada do erro quadrático médio (RMSE, do inglês *root mean square error*), o erro absoluto médio (MAE, do inglês *mean absolute error*) e o coeficiente de determinação (R^2). As métricas de erro são descritas pelas Equações (3.4) e (3.5) e , cujos detalhes podem ser vistos em Isaaks e Srivastava (1989), Willmott e Matsuura (2005) e Chai e Draxler (2014).

$$\text{RMSE} = \sqrt{\frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T (y_{it} - \hat{y}_{it})^2} \quad (3.4)$$

$$\text{MAE} = \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T |y_{it} - \hat{y}_{it}| \quad (3.5)$$

em que y_{it} é o valor observado e $\hat{y}_{it}$ é o valor predito na área i e tempo t , e nT é o tamanho da amostra. A partir das equações acima notamos que essas métricas medem o desempenho em termos do desvio de suas predições e dos valores observados. Assim, o modelo que apresenta métricas com valores menores, normalmente indica o modelo com melhor ajuste aos dados.

Por fim, o coeficiente de determinação (R^2) é comumente usado em modelos de regressão para avaliar quanto da variação total dos dados é explicada pelo modelo. Um valor mais alto de R^2 indica uma variação maior que o modelo pode explicar e mede o quão bem o modelo se ajusta aos dados, sendo dado pela Equação (3.6):

$$R^2 = 1 - \frac{\sum_{i=1}^n \sum_{t=1}^T (y_{it} - \hat{y}_{it})^2}{\sum_{i=1}^n \sum_{t=1}^T (y_{it} - \bar{y})^2} \quad (3.6)$$

em que $\bar{y}$ é o valor médio global de $\mathbf{y}$ e y_{it} , $\hat{y}_{it}$, n conforme definidos acima. O R^2 mede a proporção ou percentual da variação total de $\mathbf{y}$ explicada pelo modelo de regressão, e seu valor

varia de 0 a 1. Quando $R^2 = 1$ temos um ajustamento perfeito, ou seja, $\hat{y}_{it} = y_{it}$ para cada i e t . Por outro lado, se $R^2 = 0$ não há qualquer relação entre a resposta e as variáveis explicativas. Neste caso, $\hat{y}_{it} = \bar{y}$, e a melhor predição para qualquer valor de y é seu valor médio (GUJARATI; PORTER, 2011).

3.2.5.4 Transformação integral de probabilidade

Rosenblatt (1952) propôs uma medida de qualidade de ajuste baseado na transformação integral de probabilidade (PIT, do inglês *probability integral transform*). O PIT é um conceito fundamental em estatística que conecta a função de distribuição cumulativa, a função quantílica e a distribuição uniforme (DAWID, 1984). Os PITs são baseados em um teorema básico de probabilidade que afirma que uma variável aleatória $Y = F_X(X)$ é uniformemente distribuída em $[0, 1]$ para qualquer variável aleatória contínua X com função de distribuição cumulativa F_X . Isso decorre de:

$$P(Y \leq y) = P(F_X(X) \leq y) = P(X \leq F_X^{-1}(y)) = F_X(F_X^{-1}(y)) = y, \text{ para } 0 \leq y \leq 1.$$

Se os valores $\text{PIT} = p(y_{it}^* \leq y_{it} \mid \mathbf{y}_{-it})$ são uniformemente distribuídos, o modelo está bem coerente, enquanto a divergência da uniformidade indica que a distribuição preditiva do modelo é subdispersa (distribuição em forma de U), superdispersa (distribuição em forma de U inverso), e histogramas distorcidos ocorrem quando tendências centrais são tendenciosas (CZADO; GNEITING; HELD, 2009).

No caso de dados de contagem, como no presente trabalho, a distribuição preditiva não é discreta, e os PITs não são mais uniformes sob a hipótese de uma predição ideal. Assim, é necessário um ajuste como, por exemplo, o sugerido por Czado, Gneiting e Held (2009):

$$\text{PIT}_{it}^{\text{aj}} = \text{PIT}_{it} - 0,5\text{CPO}_{it}$$

O histograma resultante pode ser interpretado da mesma forma que um histograma PIT para dados contínuos.

3.2.6 Escolha das prioris

Na abordagem Bayesiana uma questão fundamental é a escolha da distribuição a priori dos parâmetros e hiperparâmetros para o mapeamento de doenças. Ao aplicar a formulação original do modelo BYM, as prioris geralmente são especificadas separadamente para τ_u e τ_v e as escolhas são frequentemente bastante específicas (BERNARDINELLI; CLAYTON; MONTOMOLI, 1995). Os parâmetros de precisão τ_u e τ_v são reparametrizados por $\nu_1 = \log(\tau_u)$ e $\nu_2 = \log(\tau_v)$. A priori adotada em ν_1 e ν_2 foi uma distribuição loggama não informativa, descrita por $\nu_1, \nu_2 \sim \text{loggama}(1; 5 \cdot 10^{-5})$. Na especificação BYM2 consideramos as prioris de penalidade complexa (PC) propostas por Simpson et al. (2017) com $U = 0.3/0.31$ e $\alpha = 0.01$ para os

os parâmetros de precisão dos efeitos aleatórios temporais (τ_ϕ , τ_γ e τ_σ) e τ_b . Para o parâmetro de mistura λ consideramos $U = 0.5$ e $\alpha = 2/3$. Na especificação de Leroux consideramos $\log(\tau_u) \sim \text{loggamma}(1; 5 \cdot 10^{-5})$, $\lambda \sim \text{Uniforme}(0, 1)$ e priors PC com $U = 0.3/0.31$ e $\alpha = 0.01$ para os parâmetros de precisão dos efeitos aleatórios temporais (τ_ϕ , τ_γ e τ_σ).

O parâmetro de autocorrelação temporal, ρ foi reparametrizado por $\theta_1 = \text{logit}\left(\frac{1+\rho}{1-\rho}\right)$ e adotamos a priori $\theta_1 \sim N(0; 0, 15)$. Devido à facilidade de processamento, estas reparametrizações são sugeridas para implementação no INLA (RUE; MARTINO; CHOPIN, 2009). No nível dos parâmetros de efeitos fixos, foram especificadas priors com distribuições Gaussianas não informativas $\alpha \sim N(0; \infty)$, $\beta_{it} \sim N(0; 10^{-3})$, para $i = 1, 2, \dots, n$ e $t = 1, 2, \dots, T$.

3.2.7 Software utilizado

Todos os modelos foram implementados no software estatístico R versão 4.1.2 (TEAM, 2021) usando o pacote R-INLA (RUE; MARTINO; CHOPIN, 2009; BLANGIARDO; CAMELETTI, 2015), que utiliza a abordagem INLA para estimação dos modelos Bayesianos e os mapas foram gerados a partir do pacote tmap (TENNEKES, 2018) usando *shapefile* obtido no portal do IBGE (<<https://portaldemapas.ibge.gov.br/portal.php#homepage>>). As análises foram conduzidas em um microcomputador com processador Intel-core i5 com 2,71 GHz e 16,0 GB de RAM.

3.3 Resultados

3.3.1 Análise exploratória

Analizamos um total de 5654 casos de sífilis congênita relatados no período 2010–2016 em São Paulo (média de 807,71 casos por ano). De maneira geral, houve um aumento substancial no número de casos notificados de sífilis congênita desde 2010, sendo que o maior número foi notificado em 2016 (1164 casos) (Tabela 3 e Figura 6).

A Figura 7 (a e b) mostra a frequência e o mapa da distribuição geográfica dos casos notificados de sífilis congênita no município de São Paulo ao longo do período de estudo, em que é pequeno o número de áreas apresentando “zero” ocorrências, mostrando não ter problemas de inflação de zeros. Podemos observar que a maioria das áreas possui um número menor de casos e poucas áreas possuem um número elevado de casos. Algumas áreas apresentam casos relativamente altos, principalmente grande parte do norte. O sul e oeste da zona central têm contagens relativamente baixas ou nulas.

Tabela 3 – Estatísticas descritivas para o número de casos de sífilis congênita e taxa de incidência por 1000 nascidos vivos em nível de áreas de ponderação no município de São Paulo-2010 a 2016

Ano	Nº de casos	Média	Mínimo	Máximo	DP	Taxa de Incidência por 1000 NV
2010	472	1,52	0,00	13,00	1,94	3,31
2011	577	1,86	0,00	9,00	1,97	3,43
2012	660	2,13	0,00	12,00	2,32	3,76
2013	810	2,61	0,00	16,00	2,64	4,70
2014	932	3,01	0,00	21,00	3,08	5,31
2015	1039	3,35	0,00	21,00	3,32	6,02
2016	1164	3,75	0,00	19,00	3,67	6,99

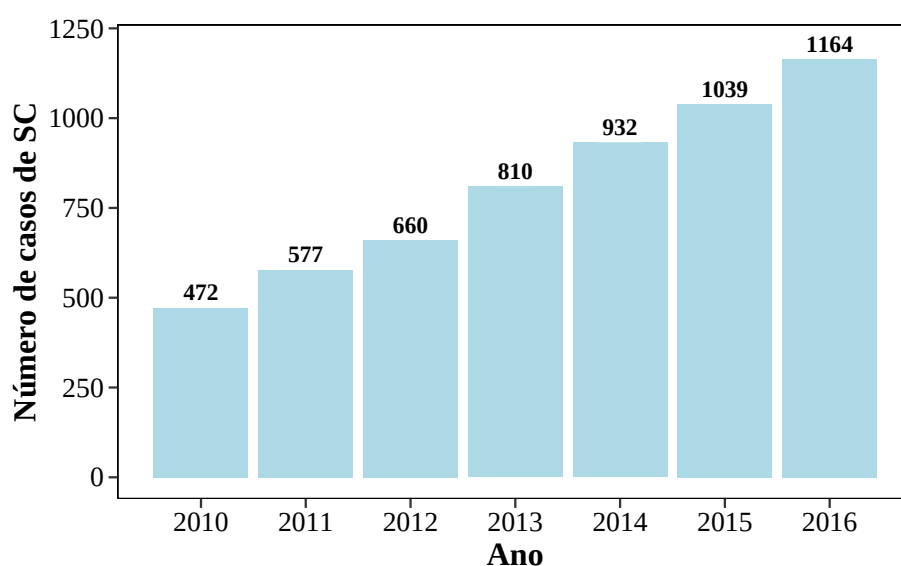


Figura 6 – Casos de sífilis congênita no período de 2010 a 2016.

A Figura 8(a) apresenta visualmente a distribuição e a presença de *outliers* nos casos observados de SC por AP, e confirma um aumento no número casos de 2010 a 2016. A Figura 8(b) mostra que o número de nascidos vivos no município São Paulo parece estar aumentando de 2010 a 2012, apresentado uma diminuição em 2013 e mantendo-se aparentemente constante até 2016.

A Figura 9 apresenta a distribuição espaço-temporal das taxas de incidência de SC no município de São Paulo para o período 2010–2016. A distribuição espaço-temporal das taxas de incidência de SC para cada área de ponderação mostra uma grande variação na ocorrência de SC nas 310 áreas. Foram observadas áreas com altas taxas de incidência de SC durante o período de estudo.

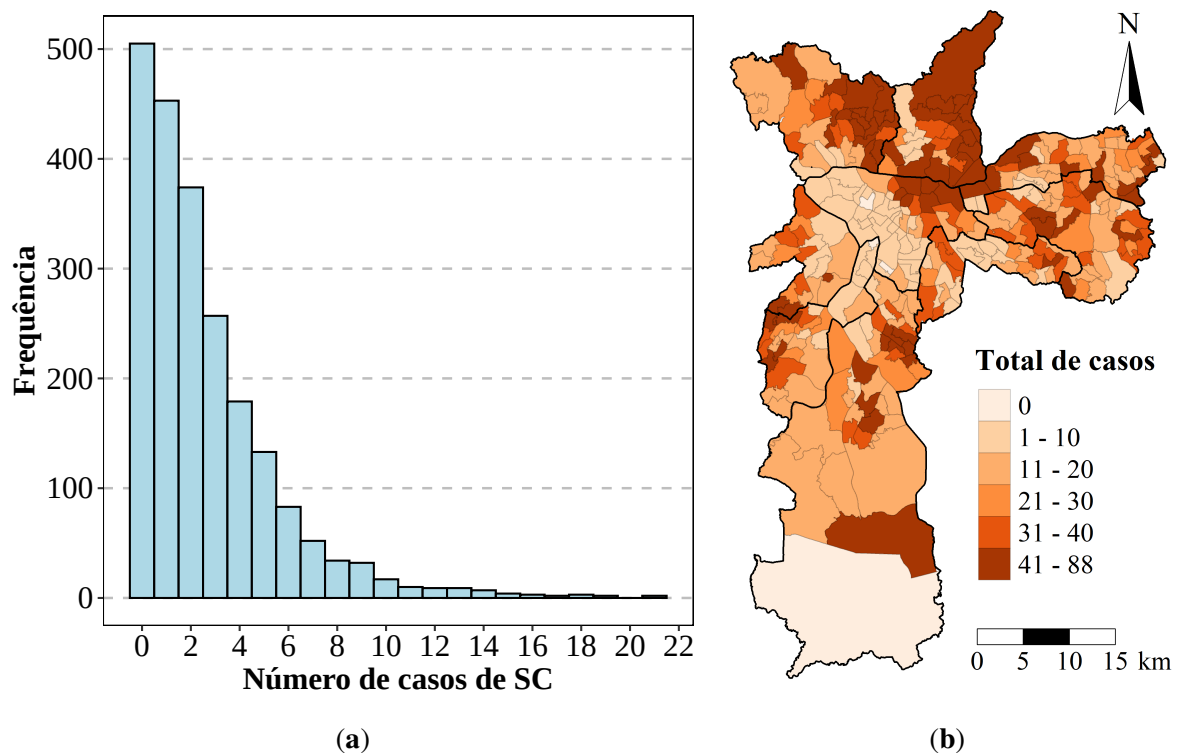


Figura 7 – (a) Distribuição de frequência dos casos de sífilis congênita no período de 2010 a 2016. (b) Distribuição geográfica dos casos notificados de sífilis congênita no município de São Paulo de 2010 a 2016.

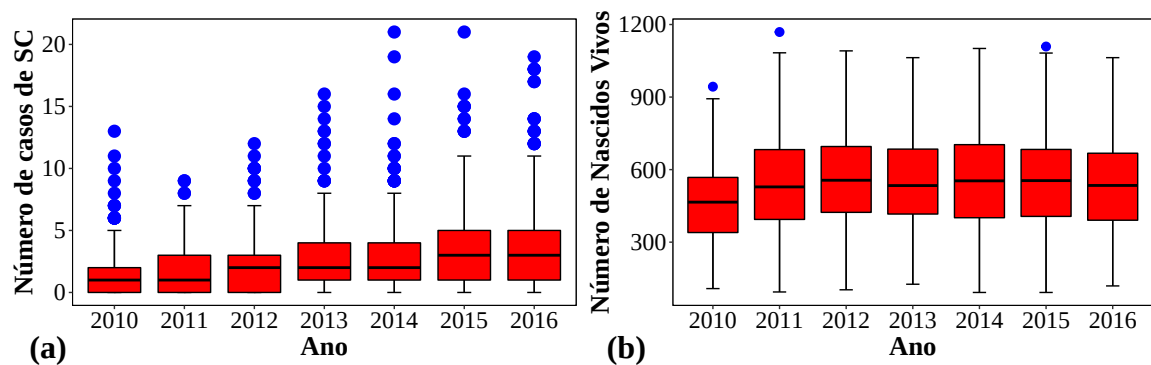


Figura 8 – Boxplots de casos observados de sífilis congênita (a) e número de nascidos vivos (b) no município de São Paulo de 2010 a 2016 por AP.

A taxa média geral para o município foi de 4,55 casos por 1000 nascidos vivos, e verificou-se que os casos em cada ano do estudo se distribuíram de forma heterogênea, e que as taxas variaram entre 0 e 40,4 casos por 1000 nascidos vivos.

A Tabela 4 apresenta estatísticas resumidas da taxa de incidência por 1000 NV para cada ano, em que Q1, Q2 e Q3 representam o primeiro quartil, a mediana e o terceiro quartil, respectivamente.

O menor risco de incidência de SC é observado em 2010, mas o risco aumentou gradativamente de 2010-2016.

Tabela 4 – Resumo da taxa de incidência por 1000 NV de SC específica do período analisado

Ano	Média	SD	Q1	Q2	Q3	Min	Máx
2010	3,36	4,61	0,00	2,14	4,87	0,00	36,31
2011	3,41	3,71	0,00	2,46	5,34	0,00	19,07
2012	3,54	3,73	0,00	2,73	4,83	0,00	25,88
2013	4,49	4,05	1,80	3,76	5,86	0,00	20,38
2014	4,92	4,66	1,87	3,98	7,00	0,00	34,57
2016	6,51	5,63	2,75	5,53	9,03	0,00	40,36

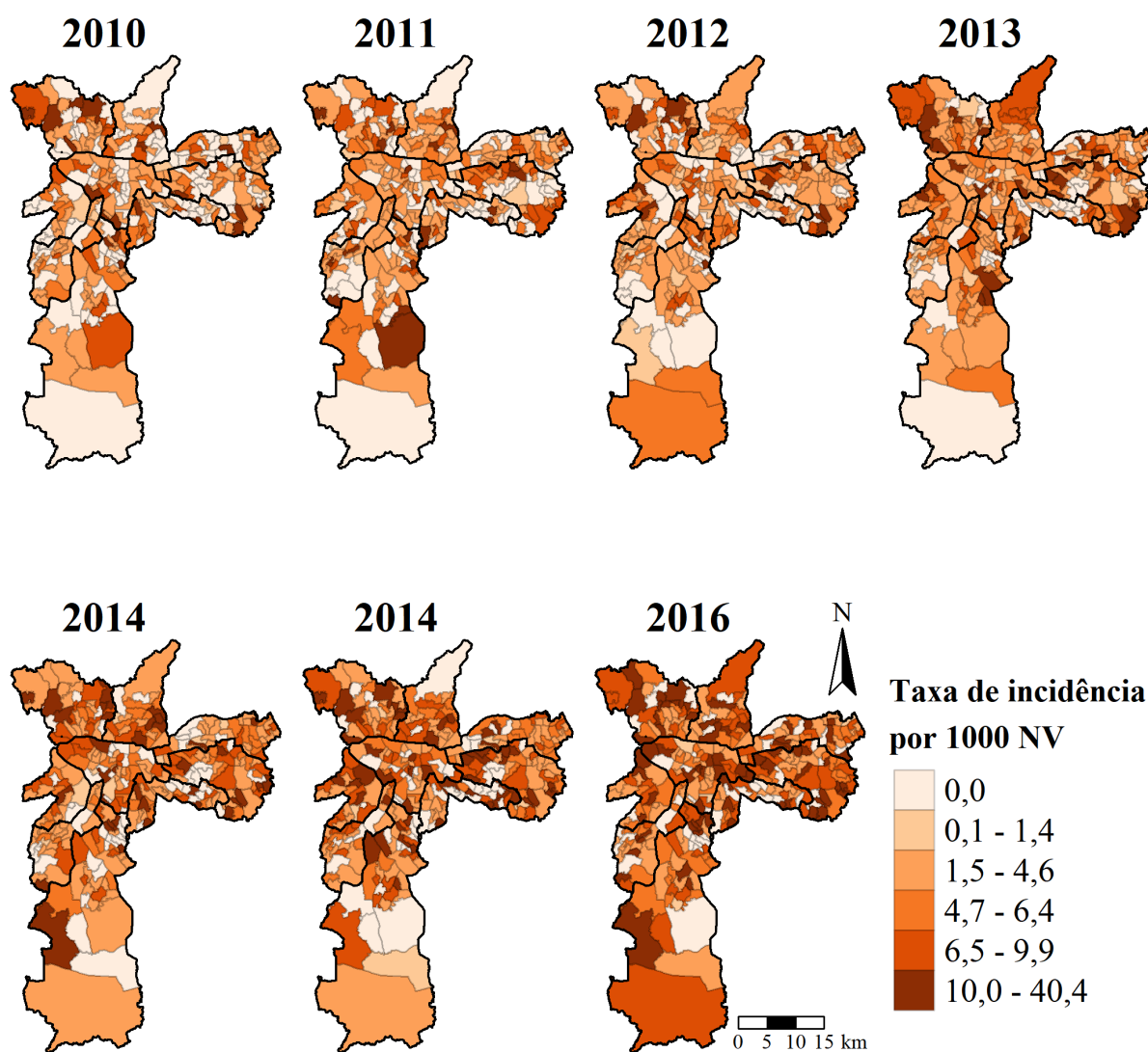


Figura 9 – Distribuição da incidência de sífilis congênita (número de casos por 1000 nascidos vivos) por área de ponderação no município de São Paulo de 2010 a 2016.

Os valores VIF foram calculados para as variáveis com o objetivo de avaliar a presença de multicolinearidade, e os resultados estão na Tabela (5a) constatando-se a existência de multicolinearidade entre as variáveis. A variável PBRANCA foi retirada do conjunto de variáveis explicativas por apresentar maior VIF. O procedimento é realizado iterativamente retirando-se uma a uma a variável de maior VIF acima de 5 até que todas tenham VIF menor que 5 (Veja a Tabela (5b)).

Portanto, as variáveis explicativas selecionadas para a modelagem são: PPRETA, PNVMA, MSIF1824, PFERTIL, PBA, RDPC e GINI.

Tabela 5 – Avaliação da multicolinearidade para as variáveis socioeconômica e ambientais

(a)		(b)	
Variáveis	VIF	Variáveis	VIF
PBRANCA	41,27	PPRETA	2,27
PPRETA	3,77	PNVMA	1,32
PPARDA	34,66	MSIF1824	2,07
PNVMA	1,35	PFERTIL	1,76
MSIF1824	2,35	PBA	1,20
PFERTIL	2,87	RDPC	2,73
PBA	1,22	GINI	1,85
RDPC	2,95		
GINI	2,01		

Outra etapa importante na análise exploratória desses dados é a razão de incidência padronizada (SIR, do inglês *standardized incidence ratio*) para cada área de ponderação com a combinação de ano para medir o risco da doença, que é calculado dividindo o número observado de casos (y) pelo número esperado de casos (E). Se o SIR for igual a 1, a contagem de casos observada e esperada são iguais, o que representa uma área de risco média em relação a toda a região de estudo, enquanto um SIR maior que 1 denota uma área de alto risco. Por exemplo, um SIR de 1,2 corresponde a um risco 20% maior em comparação com a média de todo o município de São Paulo durante o período de estudo de 7 anos. Da mesma forma, um SIR de 0,9 corresponde a uma diminuição de 10% do risco em comparação com a média municipal. A relação entre o SIR e cada covariável pode ser visualizada por meio de um gráfico de dispersão (Figura 10).

Observamos que há correlações (de Pearson) baixas a médias entre o SIR e as sete covariáveis, sugerindo que elas podem ser preditoras significativas de risco da doença. Olhando para a última coluna do gráfico, vemos que o SIR dos casos de SC está positivamente associado com PPRETA, PNVMA, PFERTIL e MSIF1824, mas mostra uma associação negativa com PBA, RDPC e GINI.

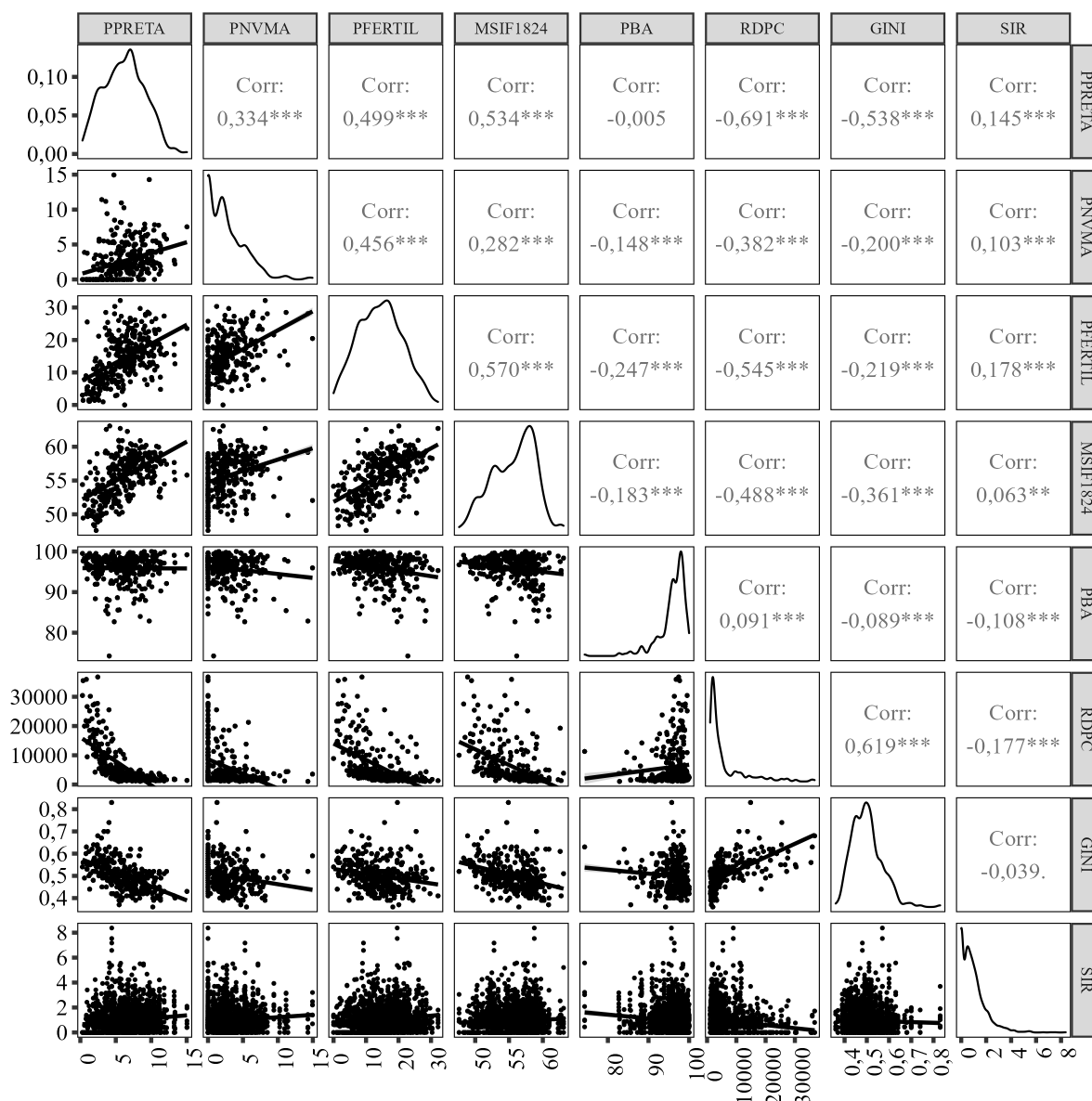


Figura 10 – Gráfico de dispersão exibindo as relações entre cada covariável e o SIR. Obtido com a função ggpairs do pacote GGally (SCHLOERKE; CROWLEY; COOK, 2018)

Na Figura 7 (a), apresentamos a distribuição de frequência das contagens dos casos de sífilis congênita. O histograma mostra alta variação na distribuição de frequência, caracterizada por uma assimetria positiva. A média e o desvio padrão do número de casos de sífilis congênita e da razão de incidência padronizada (SIR), por AP, foram de 2,61 ($\pm 2,88$) casos e 0,94 ($\pm 0,97$), respectivamente. Esta análise descritiva, na qual a dispersão das contagens dos casos é praticamente a mesma que a sua média, juntamente com a distribuição de frequência mostrado no gráfico de barras da Figura 7 (b), sugerem que aparentemente não há presença de superdispersão nos dados de contagens de sífilis congênita, e uma distribuição de Poisson seria adequada.

O mapa dos SIRs para cada área de ponderação é apresentado na Figura 11.

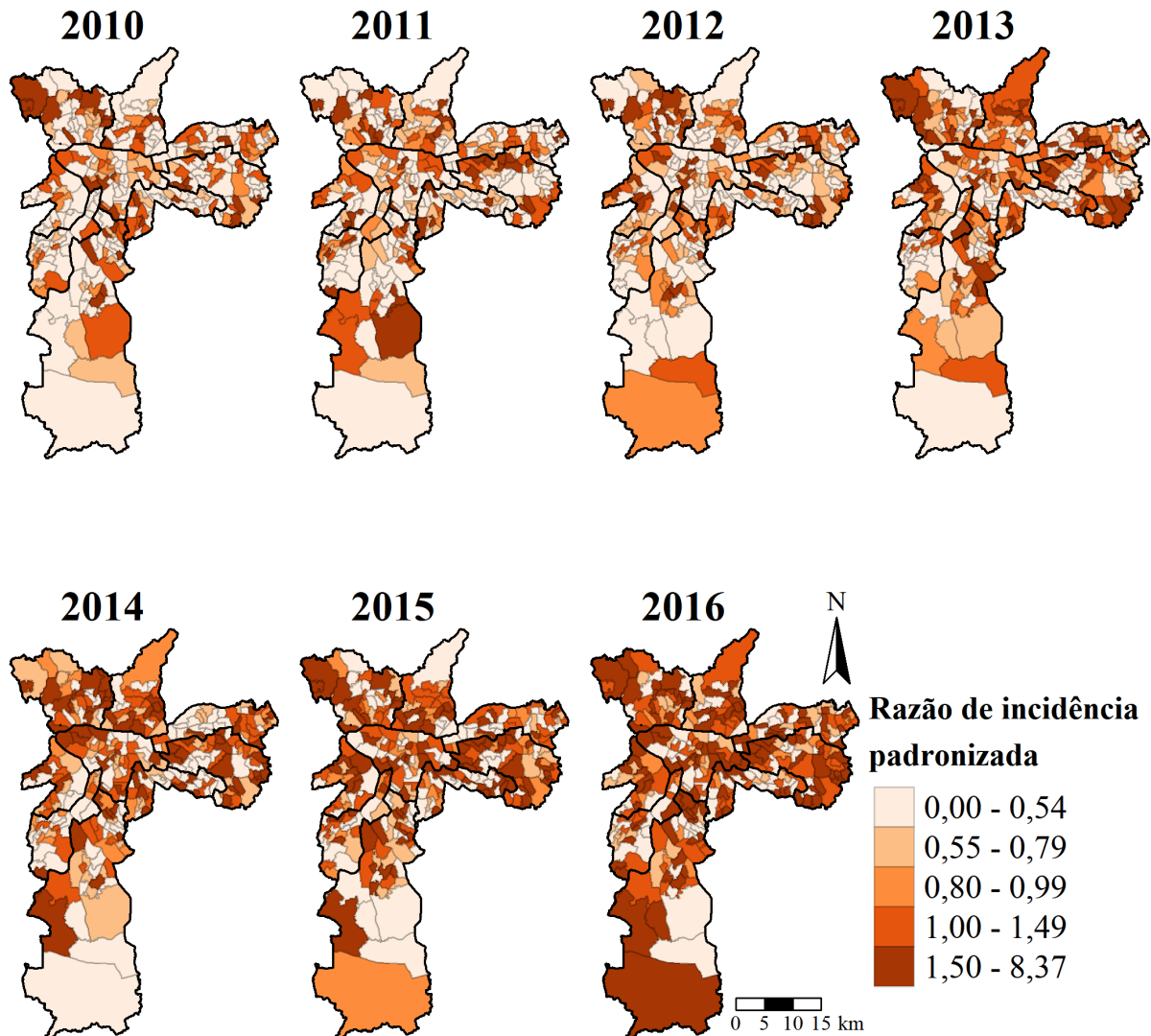


Figura 11 – Razão de incidência padronizada (y_{it}/E_{it}) para SC no município de São Paulo de 2010 a 2016.

3.3.2 Análise espaço-temporal

Para a estimação do risco relativo de sífilis congênita, aplicamos modelos hierárquicos Bayesianos e comparamos o resultado com o SIR bruto. Um total de 36 modelos foram estimados considerando a distribuição de Poisson, com três diferentes prioris espaciais (BYM, BYM2 e LEROUX)(Tabela 2) para os efeitos espaciais estruturados e uma priori AR(1) para o efeito temporal estruturado. Em seguida, uma distribuição Binomial Negativa foi usada para observar

se houve melhorias nos ajustes. As comparações dos modelos baseados em DIC, WAIC, LMPL, RMSE, MAE e R^2 são apresentadas nas Tabelas 6 e 7.

Com base nos valores de DIC, os modelos não paramétricos parecem mais adequados em comparação aos modelos paramétricos. O Modelo P29, que incluiu todas as covariáveis, efeitos aleatórios não estruturados e estruturados e termo de interação tempo-área (Tipo II) com estrutura BYM2, apresenta-se com melhor ajuste entre todos os modelos ajustados para esses dados.

As Tabelas 6 e 7 mostram os resultados da avaliação dos setenta e dois modelos (trinta e seis usando distribuição de Poisson, e trinta e seis usando Binomial Negativa) hierárquicos Bayesianos espaço-temporais considerados na análise obtidos com R-INLA usando a estratégia de aproximação simplificada de Laplace.

Com os menores valores avaliados, os modelos espaço-temporais P24, P29, P36, NB24, NB29 e NB34 mostraram-se os melhores em relação à aptidão (DIC e WAIC), em comparação com os outros modelos (Figura 12). Em seguida consideramos os modelos P29 e NB24 como sendo os melhores. Como os valores DIC e WAIC eram muito próximos para esses modelos, a escolha foi baseada naqueles que apresentaram menores valores de LMPL, RMSE, MAE e maior valor de R^2 . Finalmente, o modelo final escolhido foi o P24, por apresentar menores valores de DIC, WAIC, RMSE, MAE e maior valor R^2 (Tabela 8) e resíduos mais distribuídos em torno de -2 e 2 (Figuras 13 e 14). Neste caso, a melhor especificação da distribuição a priori do componente espacial é o BYM2 com interação espaço-temporal do tipo II.

Tabela 6 – Comparação dos 36 modelos espaço-temporais ajustados para os casos de SC no município de São Paulo considerando a distribuição de Poisson.

Modelos paramétricos								
Modelo			DIC	WAIC	LMPL ¹	RMSE	MAE	R ²
Bernardinelli	BYM	P1	7810,58	7877,15	-3932,17	1,68	1,25	0,66
	BYM2	P2	7811,68	7878,03	-3925,14	1,68	1,25	0,66
	LEROUX	P3	7812,82	7879,77	-4076,09	1,68	1,25	0,66
Bernardinelli+cov	BYM	P4	7789,59	7856,22	-3940,71	1,69	1,25	0,65
	BYM2	P5	7790,85	7857,04	-3934,90	1,70	1,25	0,65
	LEROUX	P6	7792,31	7859,25	-4087,21	1,69	1,25	0,65
Modelos não-paramétricos								
Modelo			DIC	WAIC	LMPL	RMSE	MAE	R2
BYM	SI ²	P7	7839,41	7904,41	-3942,18	1,73	1,28	0,64
	TIPO I	P8	7724,92	7741,65	-3933,84	1,30	1,01	0,80
	TIPO II	P9	7715,78	7736,60	-3928,33	1,31	1,02	0,79
	TIPO III	P10	7810,84	7885,47	-4872,29	1,61	1,21	0,69
	TIPO IV	P11	7797,76	7868,35	-5534,27	1,56	1,18	0,71
BYM2	SI	P12	7840,12	7904,57	-3938,71	1,73	1,28	0,64
	TIPO I	P13	7727,45	7734,92	-3924,73	1,28	1,00	0,80
	TIPO II	P14	7714,87	7726,45	-3904,66	1,29	1,00	0,80
	TIPO III	P15	7817,75	7891,93	-4881,23	1,64	1,23	0,68
	TIPO IV	P16	7802,27	7875,60	-5528,25	1,58	1,19	0,70
LEROUX	SI	P17	7840,50	7906,08	-4091,02	1,73	1,28	0,64
	TIPO I	P18	7727,04	7747,87	-4067,36	1,31	1,01	0,79
	TIPO II	P19	7717,09	7734,00	-4074,54	1,30	1,01	0,79
	TIPO III	P20	7808,02	7881,07	-5020,47	1,56	1,18	0,71
	TIPO IV	P21	7808,18	7881,78	-5680,74	1,58	1,19	0,70
BYM+cov	SI	P22	7808,30	7873,39	-3952,21	1,73	1,27	0,64
	TIPO I	P23	7696,29	7716,31	-3939,72	1,31	1,01	0,79
	TIPO II	P24	7688,31	7714,24	-3938,91	1,33	1,02	0,79
	TIPO III	P25	7787,79	7860,25	-4881,09	1,61	1,21	0,69
	TIPO IV	P26	7772,45	7845,93	-5542,69	1,58	1,19	0,70
BYM2+cov	SI	P27	7808,32	7872,93	-3944,36	1,73	1,27	0,64
	TIPO I	P28	7694,87	7716,81	-3930,30	1,31	1,01	0,79
	TIPO II	P29	7688,35	7711,34	-3930,68	1,32	1,02	0,79
	TIPO III	P30	7790,03	7862,72	-4874,58	1,62	1,22	0,68
	TIPO IV	P31	7774,27	7849,31	-5534,23	1,59	1,19	0,69
LEROUX+cov	SI	P32	7811,40	7876,59	-4099,42	1,73	1,27	0,64
	TIPO I	P33	7699,66	7716,55	-4086,99	1,30	1,00	0,80
	TIPO II	P34	7692,14	7716,55	-4085,34	1,32	1,02	0,79
	TIPO III	P35	7794,45	7867,69	-5028,26	1,63	1,22	0,68
	TIPO IV	P36	7778,98	7855,98	-5690,81	1,60	1,20	0,69

¹ LMPL-Logaritmo da pseudo-máxima verossimilhança.

² SI-Sem interação.

Tabela 7 – Comparação dos 36 modelos espaço-temporais ajustados para os casos de SC no município de São Paulo considerando a distribuição Binomial Negativa.

Modelos paramétricos								
Modelo			DIC	WAIC	LMPL ¹	RMSE	MAE	R ²
Bernardinelli	BYM	NB1	7831,20	7838,72	-3913,90	1,73	1,27	0,64
	BYM2	NB2	7833,88	7840,36	-3905,49	1,73	1,27	0,64
	LEROUX	NB3	7835,74	7842,13	-4055,53	1,73	1,27	0,64
Bernardinelli+cov	BYM	NB4	7803,39	7811,55	-3916,33	1,74	1,28	0,63
	BYM2	NB5	7803,12	7811,70	-3908,21	1,74	1,28	0,63
	LEROUX	NB6	7807,78	7816,39	-4065,65	1,74	1,28	0,63
Modelos não-paramétricos								
Modelo			DIC	WAIC	LMPL	RMSE	MAE	R2
BYM	SI ²	NB7	7842,36	7847,41	-3919,88	1,75	1,29	0,63
	TIPO I	NB8	7807,45	7826,05	-3932,83	1,59	1,19	0,69
	TIPO II	NB9	7779,98	7782,25	-3927,36	1,46	1,11	0,74
	TIPO III	NB10	7862,63	7846,51	-4878,85	1,75	1,28	0,63
	TIPO IV	NB11	7837,17	7847,61	-5515,26	1,74	1,28	0,63
BYM2	SI	NB12	7843,23	7847,16	-3914,44	1,75	1,29	0,63
	TIPO I	NB13	7835,94	7843,05	-3921,25	1,71	1,26	0,65
	TIPO II	NB14	7780,73	7800,84	-3916,52	1,54	1,16	0,71
	TIPO III	NB15	7834,52	7846,28	-4882,54	1,74	1,28	0,63
	TIPO IV	NB16	7844,76	7847,83	-5507,40	1,74	1,28	0,63
LEROUX	SI	NB17	7845,42	7849,30	-4066,62	1,75	1,28	0,63
	TIPO I	NB18	7833,13	7839,79	-4077,41	1,68	1,24	0,66
	TIPO II	NB19	7775,77	7797,28	-4072,29	1,52	1,14	0,72
	TIPO III	NB20	7848,44	7853,58	-4994,51	1,75	1,29	0,63
	TIPO IV	NB21	7842,30	7849,83	-5656,36	1,75	1,28	0,63
BYM+cov	SI	NB22	7808,69	7816,89	-3929,16	1,75	1,28	0,63
	TIPO I	NB23	7783,38	7800,50	-3936,52	1,63	1,21	0,68
	TIPO II	NB24	7743,45	7756,08	-3935,99	1,46	1,11	0,74
	TIPO III	NB25	7812,90	7819,84	-4860,85	1,75	1,29	0,63
	TIPO IV	NB26	7801,48	7814,94	-5523,99	1,74	1,28	0,63
BYM2+cov	SI	NB27	7810,91	7817,36	-3925,35	1,75	1,28	0,63
	TIPO I	NB28	7796,07	7805,62	-3928,82	1,66	1,23	0,66
	TIPO II	NB29	7760,25	7779,47	-3924,84	1,57	1,17	0,70
	TIPO III	NB30	7812,65	7819,27	-4854,76	1,75	1,29	0,63
	TIPO IV	NB31	7808,86	7816,68	-5510,83	1,75	1,28	0,63
LEROUX+cov	SI	NB32	7811,85	7820,57	-4075,76	1,75	1,28	0,63
	TIPO I	NB33	7788,52	7812,49	-4084,25	1,68	1,24	0,66
	TIPO II	NB34	7762,82	7784,36	-4084,53	1,57	1,17	0,70
	TIPO III	NB35	7817,11	7823,74	-5005,39	1,75	1,28	0,63
	TIPO IV	NB36	7814,85	7820,40	-5666,96	1,74	1,28	0,63

¹ LMPL-Logaritmo da pseudo-máxima verossimilhança.

² SI-Sem interação.

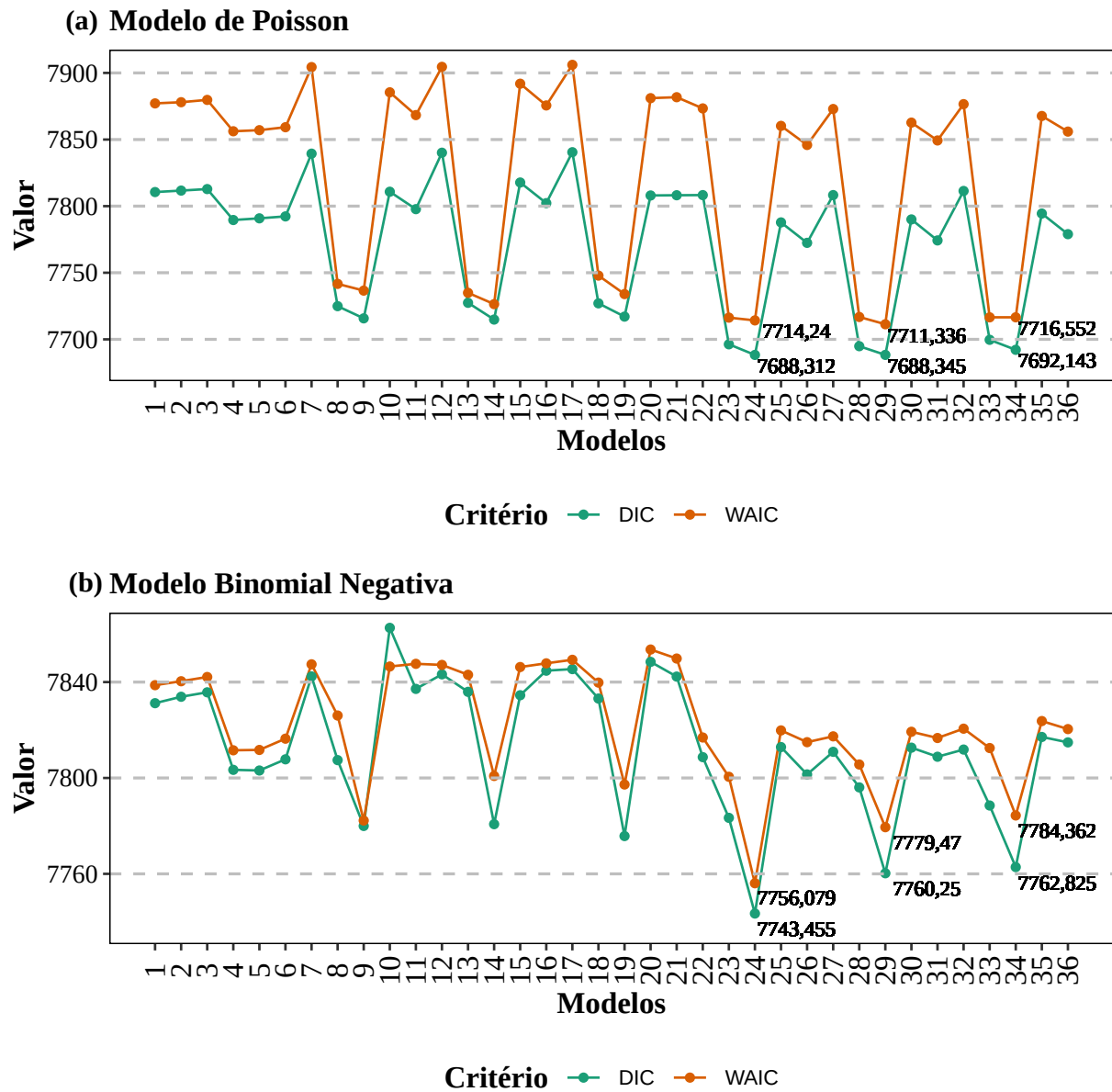


Figura 12 – Critérios DIC e WAIC para escolha dos melhores modelos.

Tabela 8 – Critério para escolha do melhor modelo.

Modelo	DIC	WAIC	Deviance	RMSE	MAE	R ²
P29	7688,35	7711,34	6655,17	1,32	1,02	0,79
NB24	7743,45	7756,08	6886,10	1,46	1,11	0,74

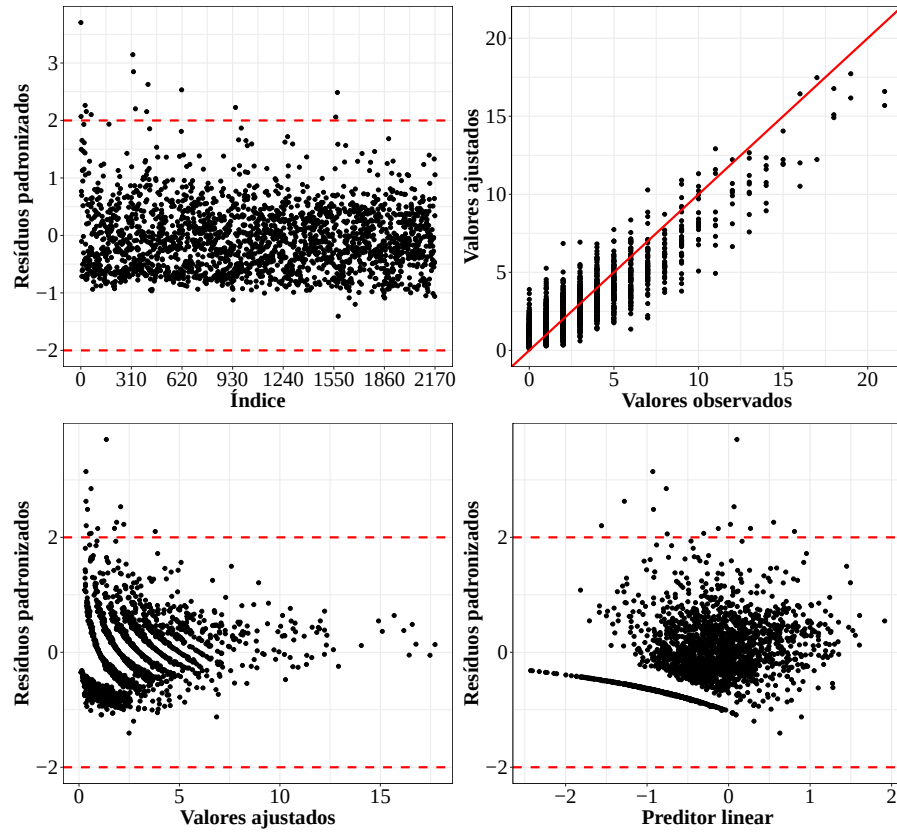


Figura 13 – Gráficos de diagnóstico para o modelo de Poisson-P29.

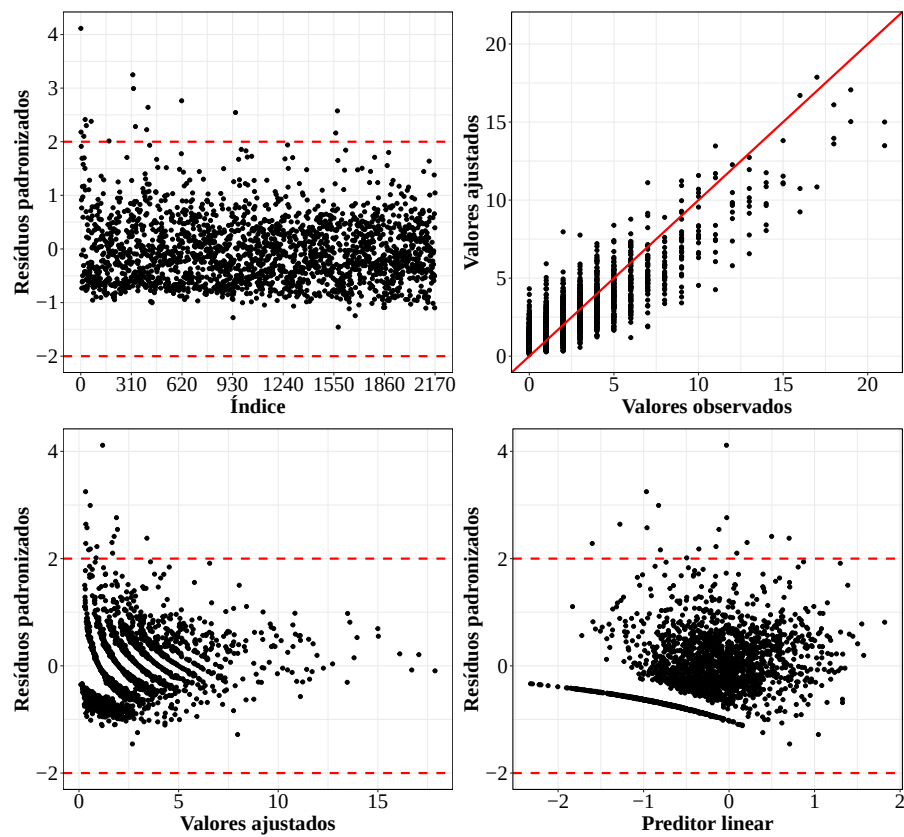


Figura 14 – Gráficos de diagnóstico para o modelo Binomial Negativa-NB24.

Os efeitos fixos auxiliam na identificação de quais covariáveis são significativamente associadas a um aumento ou diminuição do risco de sífilis congênita. Consequentemente, os esforços da saúde pública podem ser direcionados para melhorar e/ou reforçar áreas em que os cuidados de saúde são mais necessários.

A Tabela 9 resume as estimativas a posterioris para os parâmetros dos efeitos fixos, valores do RR das covariáveis selecionadas no modelo P29 e intervalos de credibilidade 95%. Os efeitos das variáveis proporção de NV de mães adolescentes e proporção de pessoas que vivem em domicílios com banheiro e água encanada, não foram significativas, pois a média a posteriori igual a zero se encontra no intervalo de credibilidade de 95%. Os fatores significativos no resultado da regressão foram usados para explicar o risco relativo de covariáveis para toda a área de estudo. Entre as variáveis, o risco de incidência de sífilis congênita aumentou com o aumento de um desvio padrão no índice de Gini ($RR = 1,131$), na proporção de pretos ($RR = 1,102$) e na proporção de mulheres de 18 a 24 anos sem instrução ou com ensino fundamental incompleto ($RR = 1,110$). Por outro lado, o risco de incidência de sífilis congênita diminuiu com o aumento de um desvio padrão na proporção de mulheres em idade fértil ($RR = 0,891$) e na renda média per capita ($RR = 0,739$).

Tabela 9 – Estimativas a posterioris dos riscos relativos dos efeitos fixos e aleatórios e intervalos de credibilidade de 95% para ocorrência de sífilis congênita no modelo de Poisson (P29) espaço-temporal com covariáveis no município de São Paulo (2010-2016).

Efeitos Fixos	Média	Q _{2,5%}	Q _{97,5%}	RR
Intercepto	-0,277	-1,386	0,846	0,988
Proporção de pretos	0,096	0,002	0,190	1,102
Proporção de NV de mães adolescentes	-0,007	-0,076	0,063	0,994
Proporção de mulheres de 18 a 24 anos sem instrução ou com ensino fundamental incompleto	0,103	0,013	0,193	1,110
Proporção de mulheres em idade fértil	-0,117	-0,203	-0,030	0,891
Índice de Gini	0,122	0,036	0,209	1,131
Proporção de pessoas que vivem em domicílios com banheiro e água encanada	-0,055	-0,121	0,011	0,947
Renda média per capita	-0,305	-0,432	-0,180	0,739
Efeitos aleatórios				
ρ_1	0,920	0,739	0,997	-
ρ_2	0,416	0,061	0,720	-
λ	0,140	0,035	0,384	-
σ_b	0,459	0,400	0,523	-
σ_ϕ	0,681	0,222	2,249	-
σ_γ	0,039	0,002	0,134	-
σ_δ	0,320	0,261	0,388	-

A contribuição relativa de cada variável no modelo de melhor ajuste (apresentado na Tabela 9) foi avaliada retirando-se uma a uma as variáveis individuais não significantes e reajustando novamente o modelo. O modelo final é apresentado na Tabela 10.

O risco de incidência de sífilis congênita aumentou com o aumento de um desvio padrão no índice de Gini ($RR = 1,126$) e na proporção de mulheres de 18 a 24 anos sem instrução ou com ensino fundamental incompleto ($RR = 1,129$). Por outro lado, o risco de incidência de sífilis congênita diminuiu com o aumento de um desvio padrão na proporção de mulheres em idade fértil ($RR = 0,914$) e na renda média per capta ($RR = 0,704$).

Tabela 10 – Estimativas a posterioris dos riscos relativos dos efeitos fixos e aleatórios e intervalos de credibilidade de 95% para ocorrência de sífilis congênita no modelo de Poisson espaço-temporal (P29) após seleção de covariáveis, município de São Paulo (2010-2016).

Efeitos Fixos	Média	$Q_{2,5\%}$	$Q_{97,5\%}$	RR
Intercepto	-0,280	-0,863	0,305	0,791
Proporção de mulheres de 18 a 24 anos sem instrução ou com ensino fundamental incompleto	0,120	0,035	0,206	1,129
Proporção de mulheres em idade fértil	-0,091	-0,175	-0,007	0,914
Índice de Gini	0,118	0,034	0,202	1,126
Renda média per capta	-0,353	-0,472	-0,237	0,704
Efeitos aleatórios				
ρ_1	0,916	0,791	0,990	-
ρ_2	0,404	0,091	0,651	-
λ	0,088	0,025	0,160	-
σ_b	0,450	0,399	0,497	-
σ_ϕ	0,471	0,233	1,103	-
σ_γ	0,072	0,036	0,161	-
σ_δ	0,322	0,264	0,388	-

3.3.3 Distribuição espaço-temporal dos riscos relativos de SC

A Tabela 10 também apresenta as estatísticas resumidas para os desvios padrão dos parâmetros dos efeitos aleatórios do modelo selecionado. A média a posteriori do parâmetro de suavização espacial da distribuição BYM2 para o efeito aleatório da área de ponderação é $\lambda = 0,088$ com um intervalo de credibilidade (0,025; 0,160), o que é interpretado como uma pequena dependência espacial entre essas áreas, indicando que há baixa correlação espacial entre regiões vizinhas. Portanto cerca de 8% da variabilidade é capturado pelo componente espacial. A investigação da temporalidade foi realizada testando o coeficiente ρ do modelo autorregressivo de primeira ordem com a hipótese nula de que o processo não era estacionário, ou seja, $|\rho_1| > 1$. O modelo indicou presença significativa de estacionariedade de 0,916.

Os riscos relativos para cada um dos 7 anos foram calculados, mantendo as covariáveis e o risco espacial constantes, expondo o efeito temporal correlacionado à estrutura autorregressiva

de primeira ordem (γ_t) e o efeito não estruturado (ϕ_t) nos modelos espaço-temporais de risco de SC.

A Figura 15 mostra os resultados obtidos com o modelo final e destaca as médias a posterioris dos RRs temporais de casos de SC no período 2010–2016, ajustados por AP com covariáveis. Percebe-se uma tendência crescente para o efeito estruturado (AR(1)) ao longo dos anos, enquanto o termo não estruturado mostra algumas flutuações em torno de 1, com riscos acima de 1 após 2013. O efeito aleatório total foi praticamente igual ao efeito aleatório estruturado.

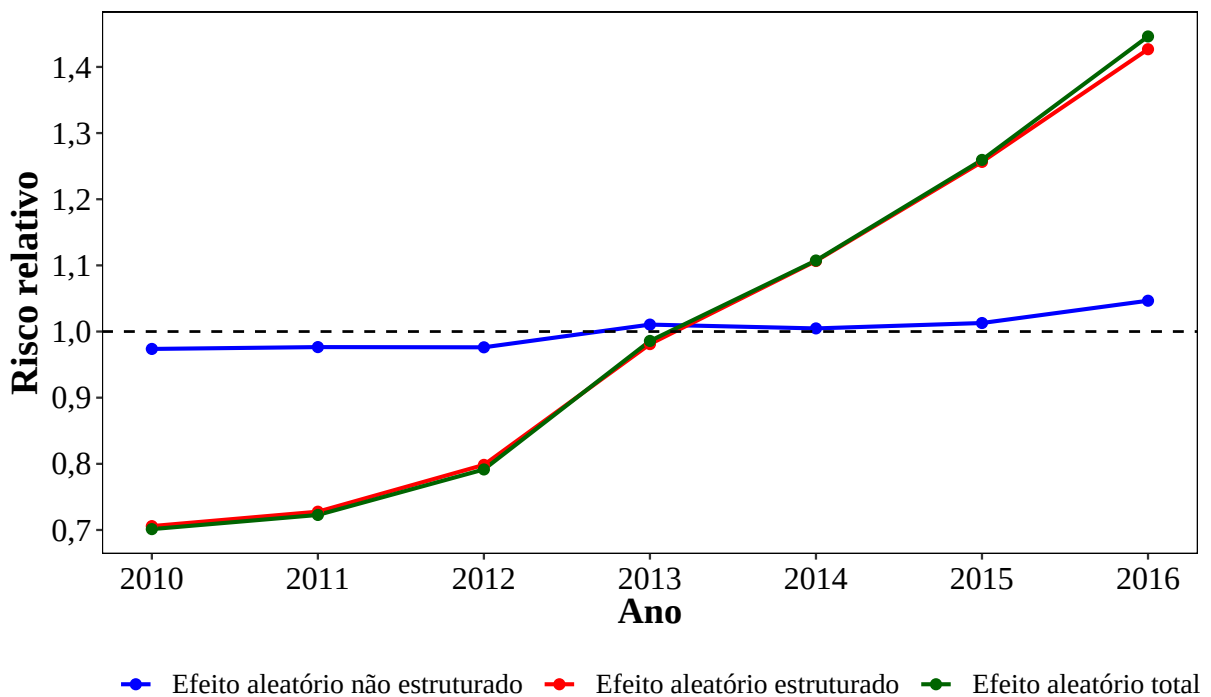


Figura 15 – Tendência temporal a posteriori de casos de sífilis congênita no município de São Paulo (2010-2016): efeito temporal não estruturado $\exp(\phi_t)$, efeito temporal estruturado $\exp(\gamma_t)$ e efeito temporal total $\exp(\phi_t + \gamma_t)$.

Para as tendências temporais específicas, selecionamos oito áreas de ponderação para representar as regiões norte, noroeste, leste, sul, oeste, sudoeste, nordeste e centro do município de São Paulo, a fim de observar a tendência temporal de diferentes áreas (Figura 16). Essas tendências temporais específicas estão apresentadas na Figura 17. A Área 303-Sul mostra uma aparente flutuação próxima de 1 ao longo dos anos. No entanto, Área 29-Centro, Área 135-Norte e Área 254-Nordeste apresentam tendências diferentes. Há uma tendência temporal crescente para Área 152-Leste até o ano 2014 e diminuindo até 2016. Enquanto isso, para Área 103-Sudoeste, o efeito do tempo está diminuindo no primeiro ano, com leve acréscimo de 2011 a 2012, seguido de uma queda até 2015, e novamente um leve acréscimo após 2015.

No geral, observa-se que a tendência da SC é diferente para cada área de ponderação ao longo dos anos. Isso está de acordo com a interação espaço-tempo do Tipo II, em que há uma tendência temporal diferente de região para região, mas não possui nenhuma estrutura no espaço. Algumas regiões apresentaram alto risco de SC nos primeiros anos, algumas passaram a ter alto risco a partir da metade do período, enquanto outras não apresentaram nenhuma tendência temporal significativa. Isso pode estar relacionado a diferentes fatores socioeconômicos e ambientais para cada área, por exemplo, a renda.

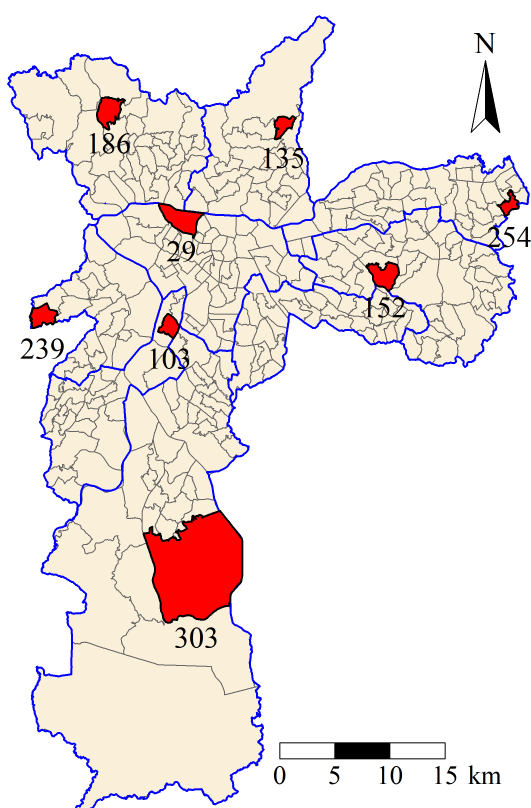


Figura 16 – Mapa das oito áreas de ponderação selecionadas para a observação da tendência temporal entre diferentes áreas de acordo com os bairros do Município de São Paulo.

O mapeamento das estimativas de risco relativo da SC é uma das principais saídas do processo de modelagem. A Figura 18 mostra os mapas com as médias a posterioris dos RRs espaço-temporais para SC no período de 2010 a 2016 segundo a AP. As cores mais escuras representam as áreas de maior risco. Nos anos de 2010 e 2011 há um baixo risco relativo geral na maioria das áreas de ponderação, mas posteriormente, o risco relativo se espalhou ao longo do tempo para o resto do município, possivelmente indicando a presença de fatores nesses locais que estão envolvidos na ocorrência de casos de SC.

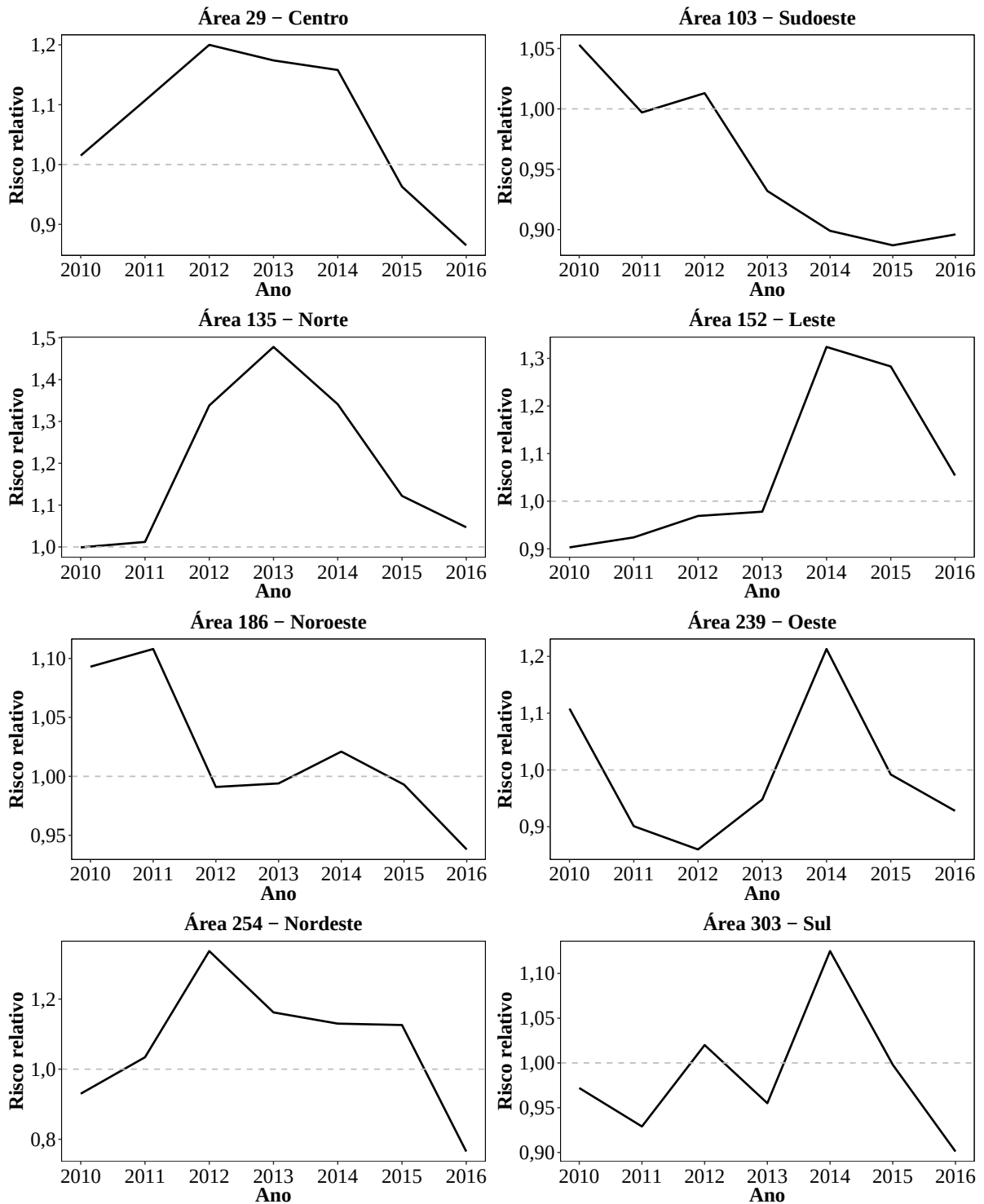


Figura 17 – Tendências temporais a posteriori de casos de sífilis congênita no município de São Paulo (2010-2016) específicas para oito áreas de ponderação selecionadas: efeito temporal $\exp(\delta_{it})$.

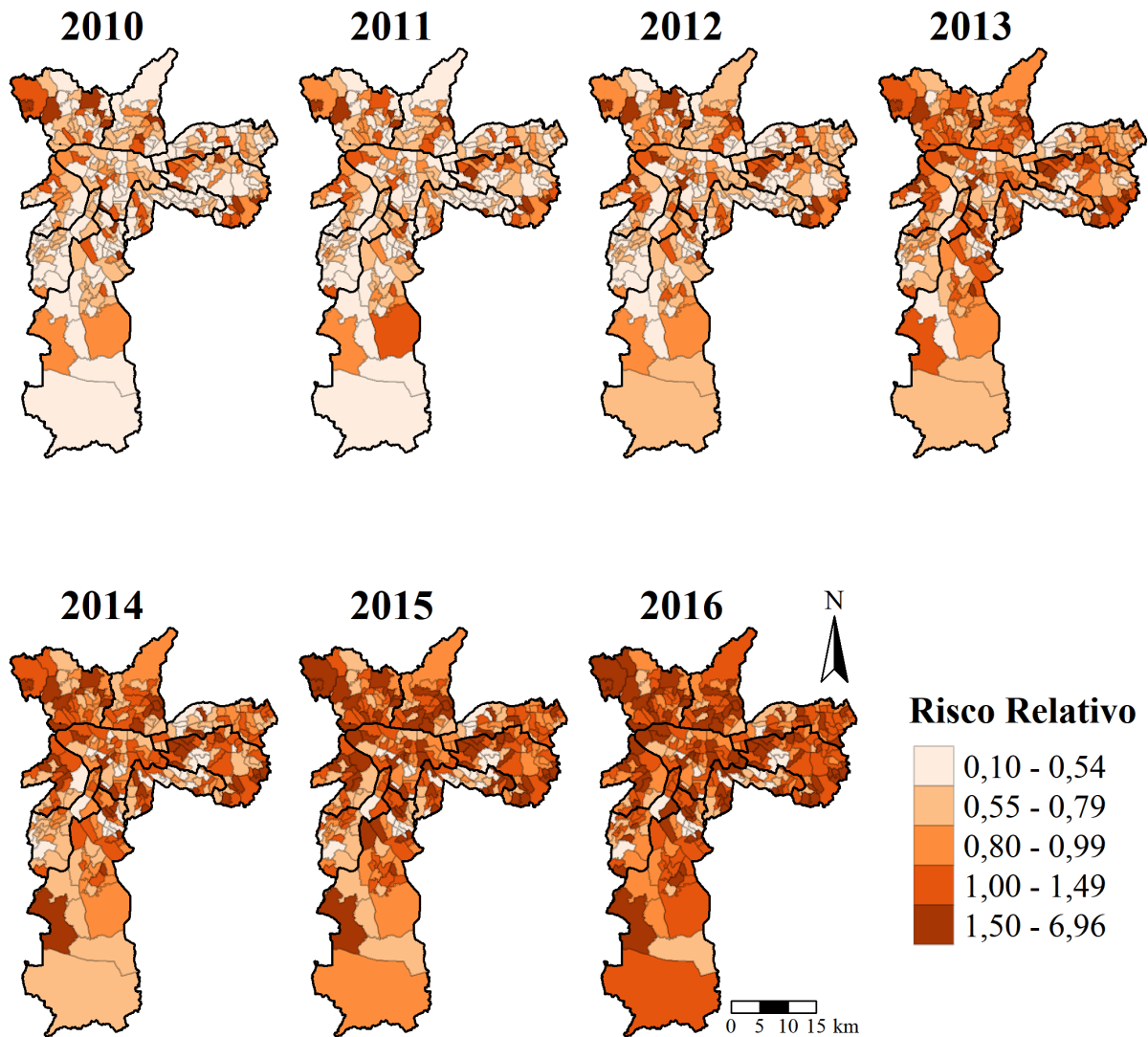


Figura 18 – Médias a posterioris dos RRs espaço-temporais para SC no período de 2010 a 2016 no município de São Paulo-SP.

A Figura 19 mostra as probabilidades a posterioris dos RRs espaço-temporais serem maiores do que a unidade ($P(e^{\psi_{it}} > 1)$) para cada ano. Esta distribuição de probabilidade a posteriori fornece uma espécie de valor p Bayesiano, que pode ser usado para detectar ou destacar áreas de alto risco com base na definição de um ponto de corte pelo analista. Geralmente nos estudos de mapeamento de doenças, as regiões com probabilidades acima de 0,8 e 0,9 são consideradas de alto risco, como sugerido por (RICHARDSON et al., 2004). Neste estudo, o valor de corte de 0,8 é utilizado para detectar regiões de alto risco em todos os períodos de tempo. De modo sutil, ao longo dos anos, a incerteza associada às estimativas a posterioris (risco de excedência) variou para as diferentes áreas, embora mais consistente nos anos de 2013 a 2016 nas regiões norte e noroeste, nas partes leste e oeste da zona central e grande parte da zona sul que se destacam por apresentarem a maior concentração de áreas de alto risco para casos de SC no período do estudo.

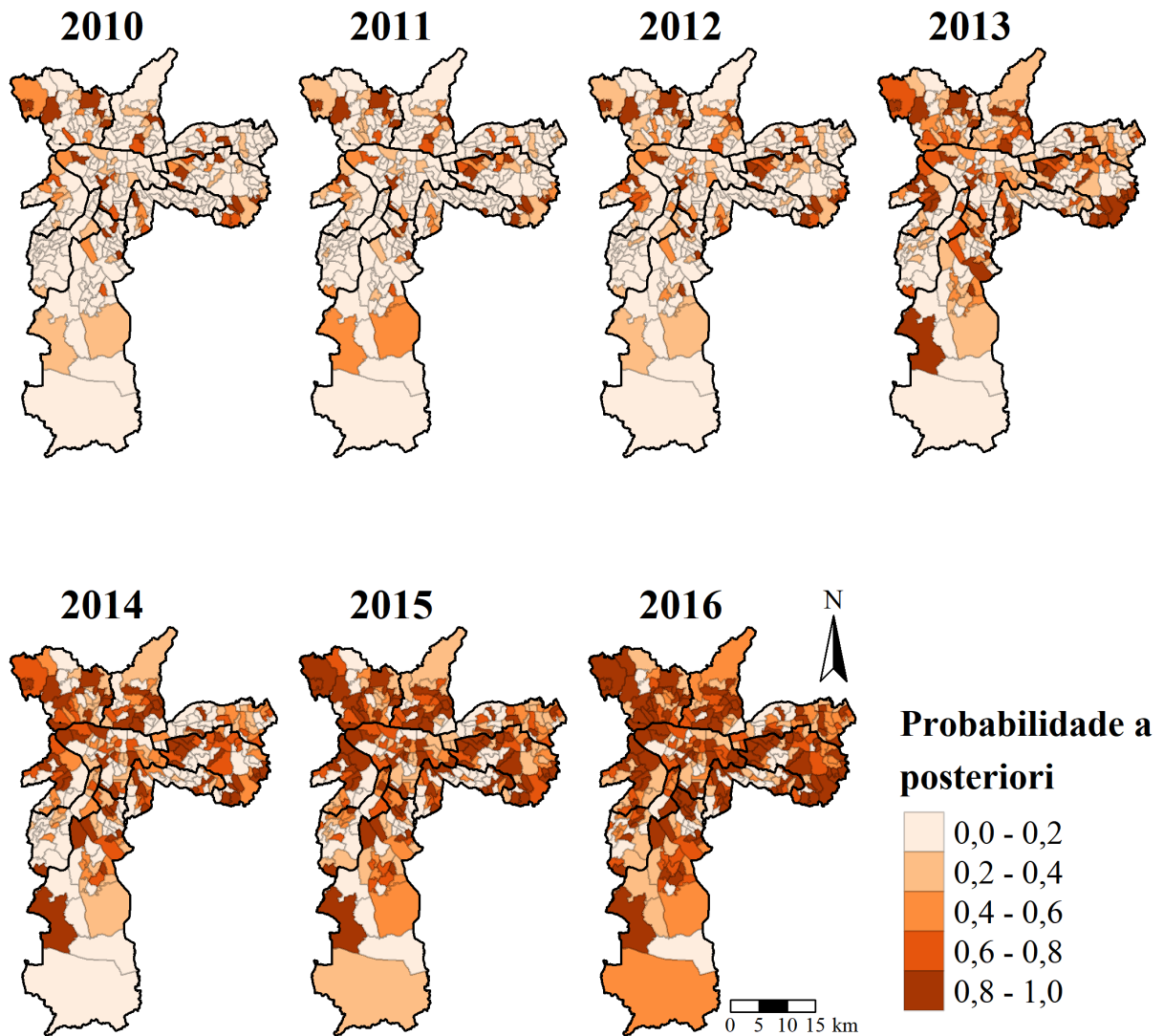


Figura 19 – Probabilidades a posteriori de que os RRs sejam maiores que 1 ($P(\Psi_{it} > 1|\mathbf{y})$) no período de 2010 a 2016 no município de São Paulo.

Segundo dados do Agrupamento Socioambiental dos territórios de saúde do município de São Paulo (ASATS), estas áreas estão em regiões com condições de renda, escolaridade, estrutura familiar e de acesso classificadas como ruins ou péssimas (PREFEITURA DE SÃO PAULO, 2015). Além, disso as mesmas estão de acordo com o Índice Socioeconômico do Contexto Geográfico para Saúde e Estudos Sociais (GeoSES) proposto por Barrozo et al. (2020), em que as áreas com melhores condições socioeconômicas (GeoSES igual ou próximo a 1) estão localizadas na parte central do município e as com piores condições socioeconômicas (GeoSES igual ou próximo de -1) se encontram na periferia (Figura 20).

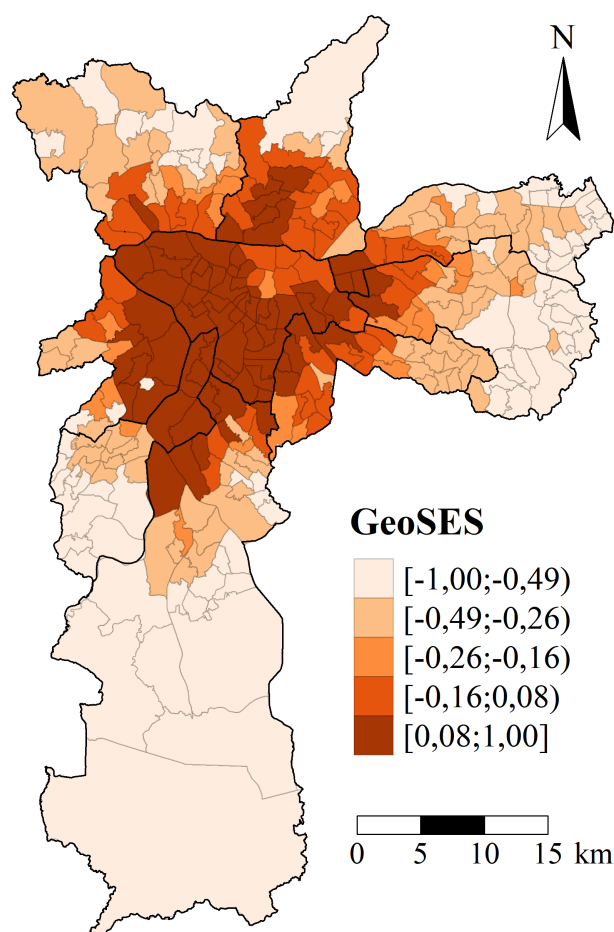


Figura 20 – Distribuição do Índice Socioeconômico do Contexto Geográfico para Saúde e Estudos Sociais (GeoSES) segundo a área de ponderação, município de São Paulo, 2010. Para a obtenção do mapa usamos os dados obtidos em <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0232074#sec015>.

3.3.4 Desempenho preditivo

Conforme mostrado na Tabela 9, o modelo Tipo II apresenta o melhor desempenho preditivo com os menores valores de MAE e RMSE e maior valor de R^2 . A Figura 21(a) mostra o gráfico de dispersão entre os valores observados de SC e os valores ajustados. Podemos ver que, em geral, os pontos estão muito próximos da linha diagonal, indicando que os valores ajustados são consistentes com os valores de casos observados. Isso mostra que em média a predição está muito próxima dos valores observados. Especificamente, para frequências de casos abaixo de 7, os pontos estão distribuídos ao redor da linha diagonal, enquanto que para as frequências de casos acima de 7, os pontos tendem a estar distribuídos abaixo da linha diagonal.

O histograma da transformação integral de probabilidade (PIT) adaptada está representado na Figura 21(b) e mostra uma distribuição que tende à uniformidade. Para o presente modelo, o

histograma PIT não apresenta nenhum sinal de calibração errada. Portanto, concluímos que a distribuição preditiva é consistente com nossos dados.

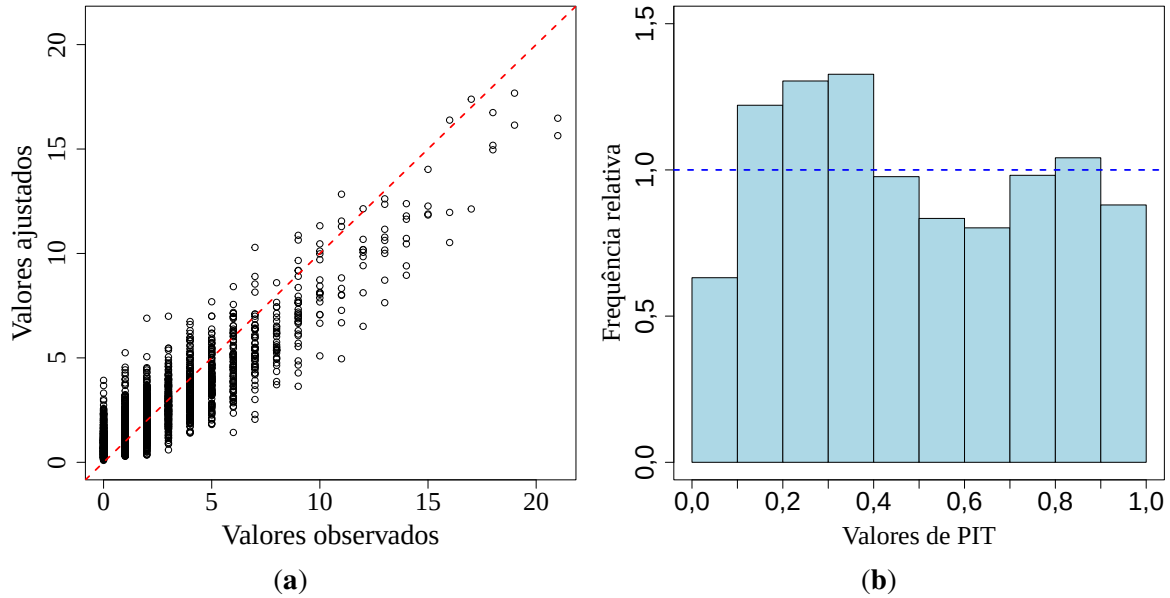


Figura 21 – (a) Valores ajustados versus valores observados dos casos notificados de sífilis congênita no município de São Paulo de 2010 a 2016. (b) Histograma dos valores de PIT ajustados.

Outro fator interessante na análise dos dados são os resíduos, pois espera-se que os mesmos se distribuam aleatoriamente entre as áreas, visto que as estimativas dos modelos autorregressivos espaciais são tendenciosas se a variância do erro não for homogênea no espaço (OTTO, 2018). Uma medida comumente usada de autocorrelação espacial é o índice I de Moran (MORAN, 1950), que é uma versão espacial do coeficiente de correlação de Pearson e mede quanto o valor do resíduo de determinada área está associado com o valor deste em áreas vizinhas, sendo calculado por:

$$I = \frac{K \sum_{k=1}^K \sum_{j=1}^K w_{kj} (r_k - \bar{r}) (r_j - \bar{r})}{\left(\sum_{k=1}^K \sum_{j=1}^K w_{kj} \right) \sum_{k=1}^K (r_k - \bar{r})^2}, \quad (3.7)$$

em que r_k o resíduo da k -ésima área. Além disso, $\mathbf{W} = (w_{kj})$ é uma matriz de vizinhança ou de adjacência $K \times K$, que indica se cada par de áreas estão próximas. Geralmente, uma especificação binária é considerada, em que $w_{kj} = 1$ se as áreas k e j compartilham uma fronteira comum e, $w_{kj} = 0$ caso contrário. Existem outras possibilidades na especificação de $\mathbf{W}$, e uma revisão é dada por Earnest et al. (2007). O valor da estatística I de Moran está entre -1 e 1 , com valor zero correspondendo à independência espacial, enquanto um valor 1 denota forte autocorrelação espacial positiva. Após o cálculo, é estabelecida uma significância estatística. Um teste de permutação com a hipótese nula de independência espacial (autocorrelação espacial nula) pode ser conduzido com base na estatística I de Moran, em que o valor p é baseado em N permutações aleatórias dos dados (ANSELIN, 2005). Analisamos a presença de autocorrelação

especial nos resíduos do modelo de Poisson para cada ano. Os resultados são apresentados na Tabela 11. Identificamos autocorrelações espaciais positivas e negativas não significativas para os resíduos no nível de AP de 2010 a 2016. O valor de I variou de -0,0443 a 0,0222. No entanto, o índice de autocorrelação espacial não apresentou tendência crescente ou decrescente, com o menor valor observado em 2016 ($I = -0,0443$; $p = 0,8793$) e maior valor em 2013 ($I = 0,0222$; $p = 0,2287$). Destacamos também que os anos de 2010 e 2013, apresentam os valores p mais baixos. Os resultados mostram que em todos os anos, a correlação espacial é não significativa nos resíduos, podendo assumir que se distribuem aleatoriamente na região do estudo (Figuras 22, 23 e 24). Em suma, e dado os resultados obtidos, há evidência para rejeitar a hipótese nula de que não existe auto-correlação espacial nos resíduos.

É importante ressaltar que a autocorrelação temporal residual poderia ser avaliada de forma semelhante, mas com apenas 7 pontos de tempo as estimativas resultantes não seriam confiáveis.

Tabela 11 – Análise de autocorrelação espacial residual anual para o modelo selecionado.

Ano	Índice de Moran	Valor p
2010	0,0188	0,2575
2011	-0,0221	0,6932
2012	-0,0368	0,8305
2013	0,0222	0,2287
2014	-0,0001	0,4518
2015	-0,0398	0,8513
2016	-0,0443	0,8793

A Figura 23 apresenta os gráficos de dispersão de Moran dos resíduos para os anos de 2010 a 2016, que ilustram a relação entre os valores dos resíduos em cada local e o valor médio dos resíduos em locais vizinhos. Nota-se claramente que os pontos estão bastante espalhados, não existindo nenhum aglomerado aparente em nenhum dos quadrantes sugerindo, assim, fraca associação espacial nos resíduos. Observamos que cada uma das retas apresentadas possui um declive igual aos respectivos índices de Moran: 0.0188, -0,0221, -0,0368, 0,0222, -0,0001, -0,0398 e -0,0443. Todos os gráficos mostram ainda alguns valores um pouco diferentes dos restantes, assinalados com um número. Esta é apenas outra maneira de ter uma percepção do comportamento do dados, sendo por isso um complemento ao que já tínhamos verificado anteriormente.

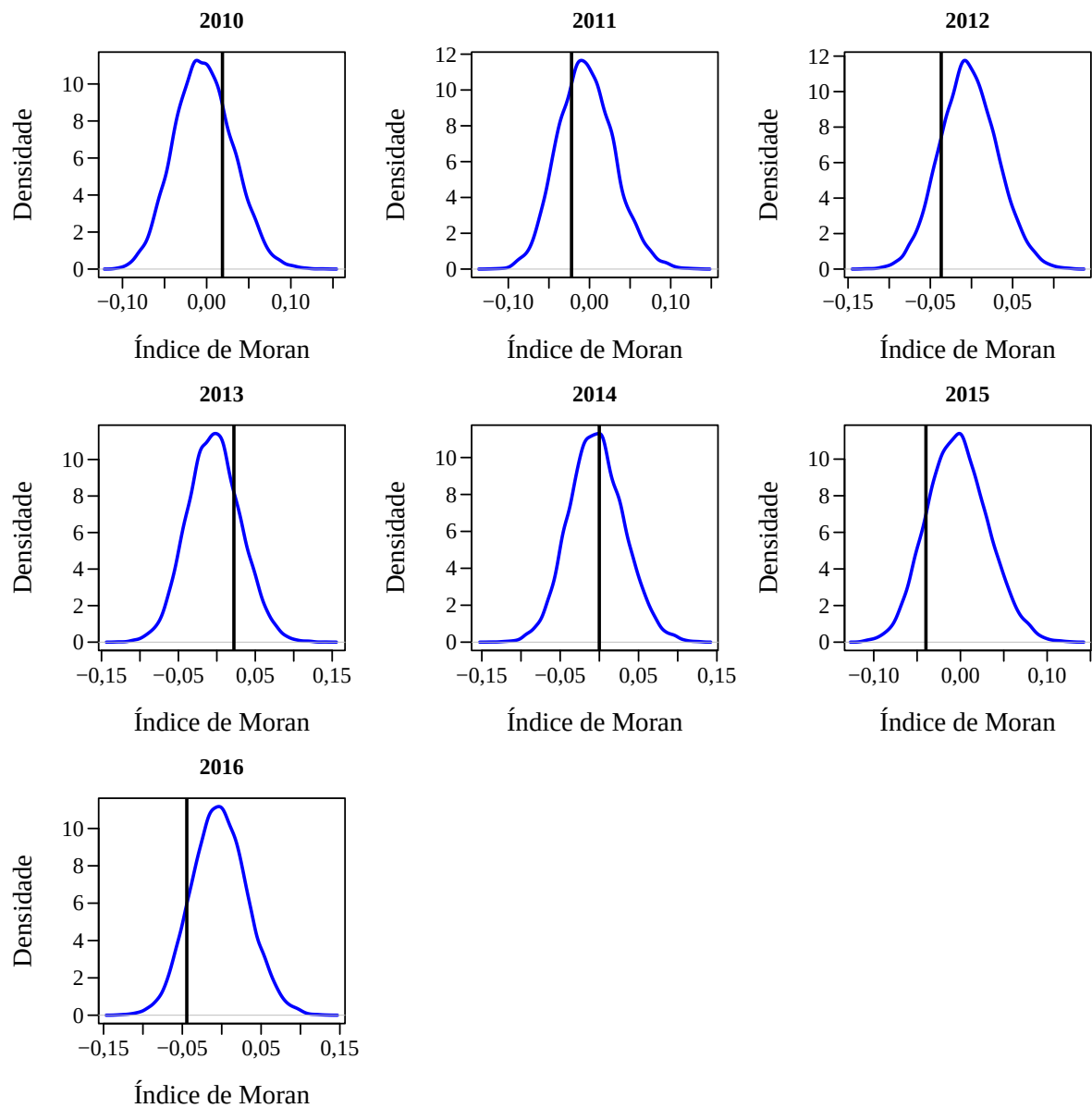


Figura 22 – Distribuição gerada por meio de 10000 simulações para os resíduos do modelo final.

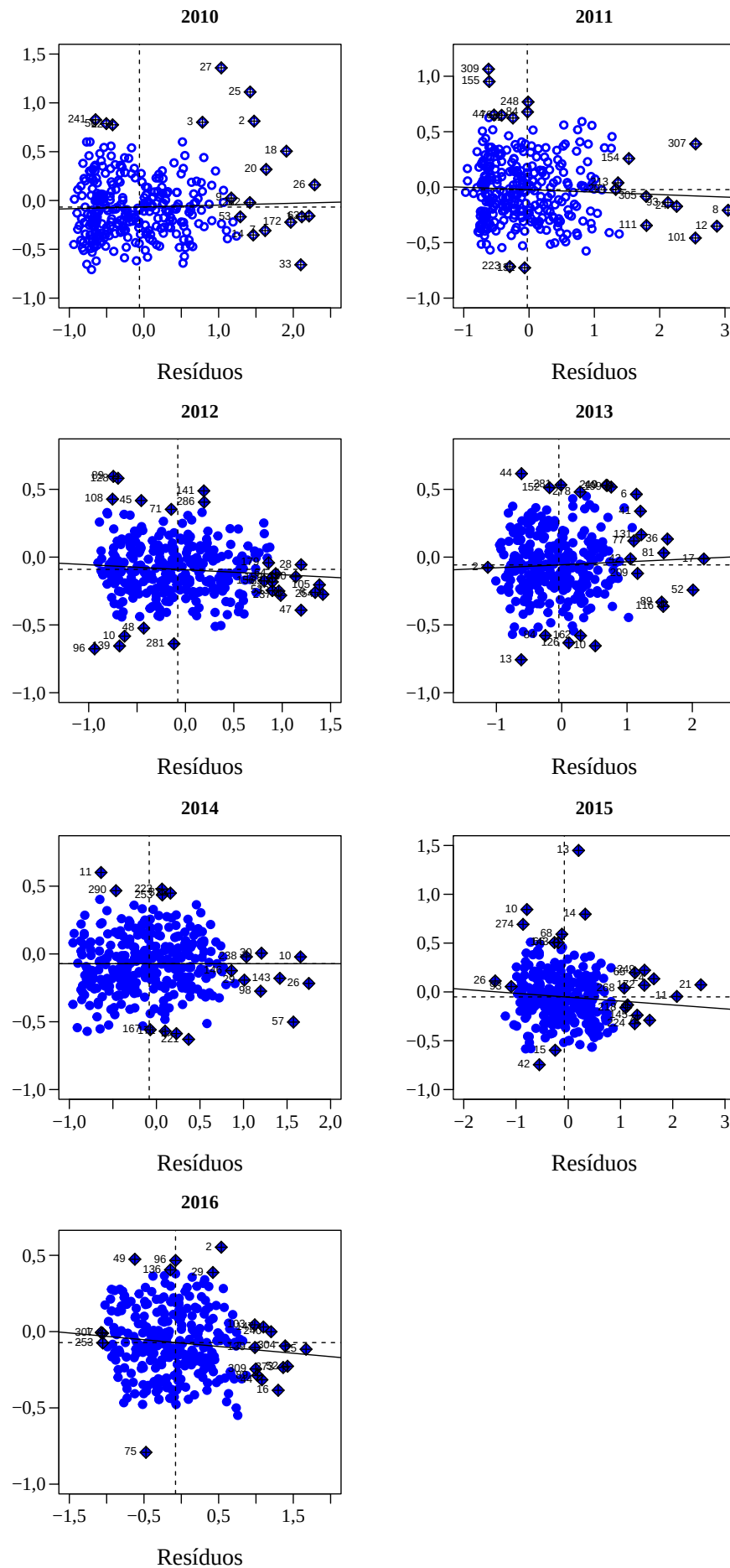


Figura 23 – Gráficos de dispersão de Moran mostrando a relação entre os resíduos e o valor médio de seus vizinhos no período de 2010 à 2016.

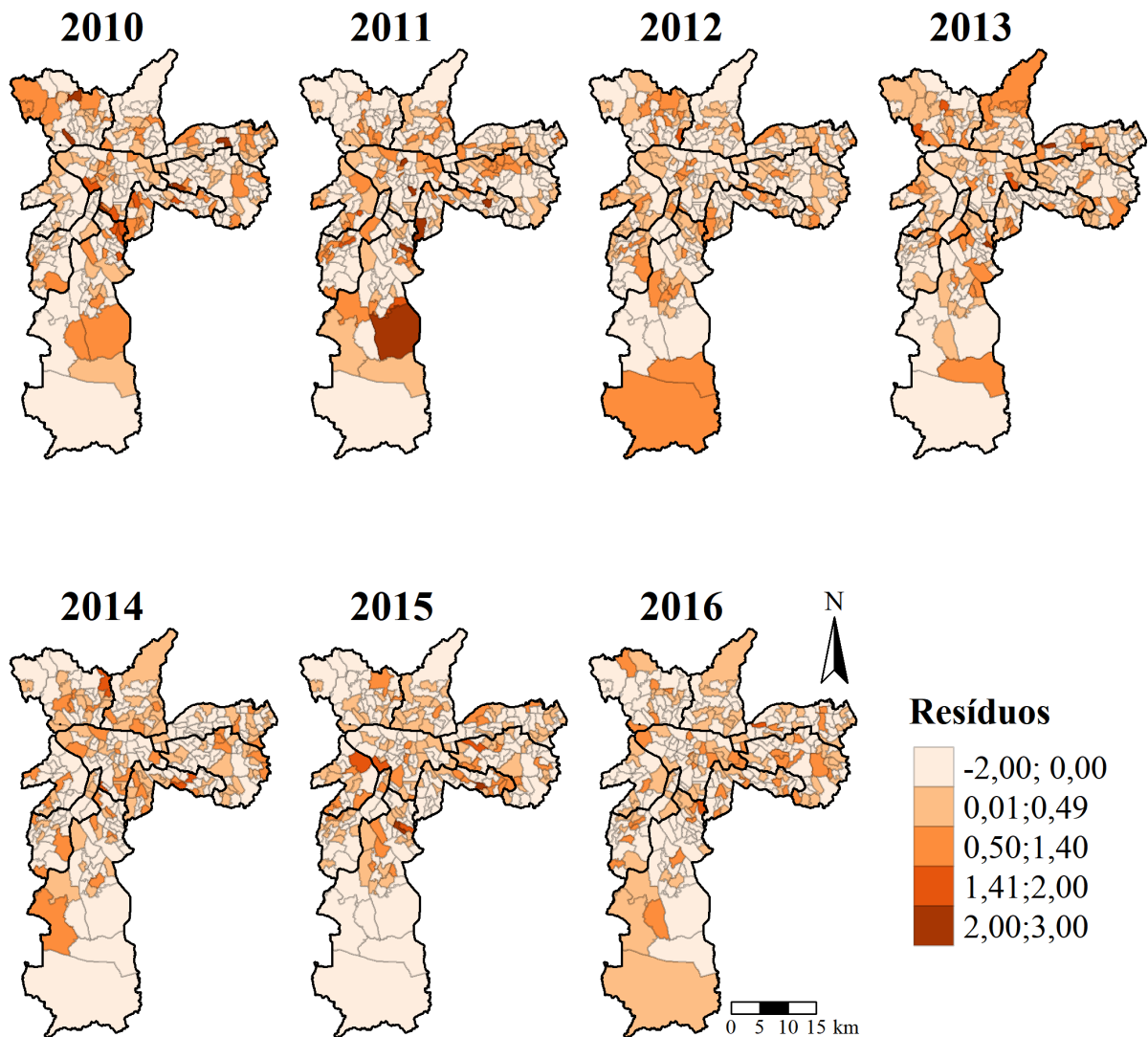


Figura 24 – Resíduos padronizados por áreas de ponderação do município de São Paulo no período de 2010 a 2016.

A Figura 25 mostra os gráficos dos resíduos padronizados de nosso modelo final versus os valores ajustados para cada ano, mostrando que nosso modelo foi bem ajustado, com valores concentrados entre -2 e 2 e poucos valores acima de 2.

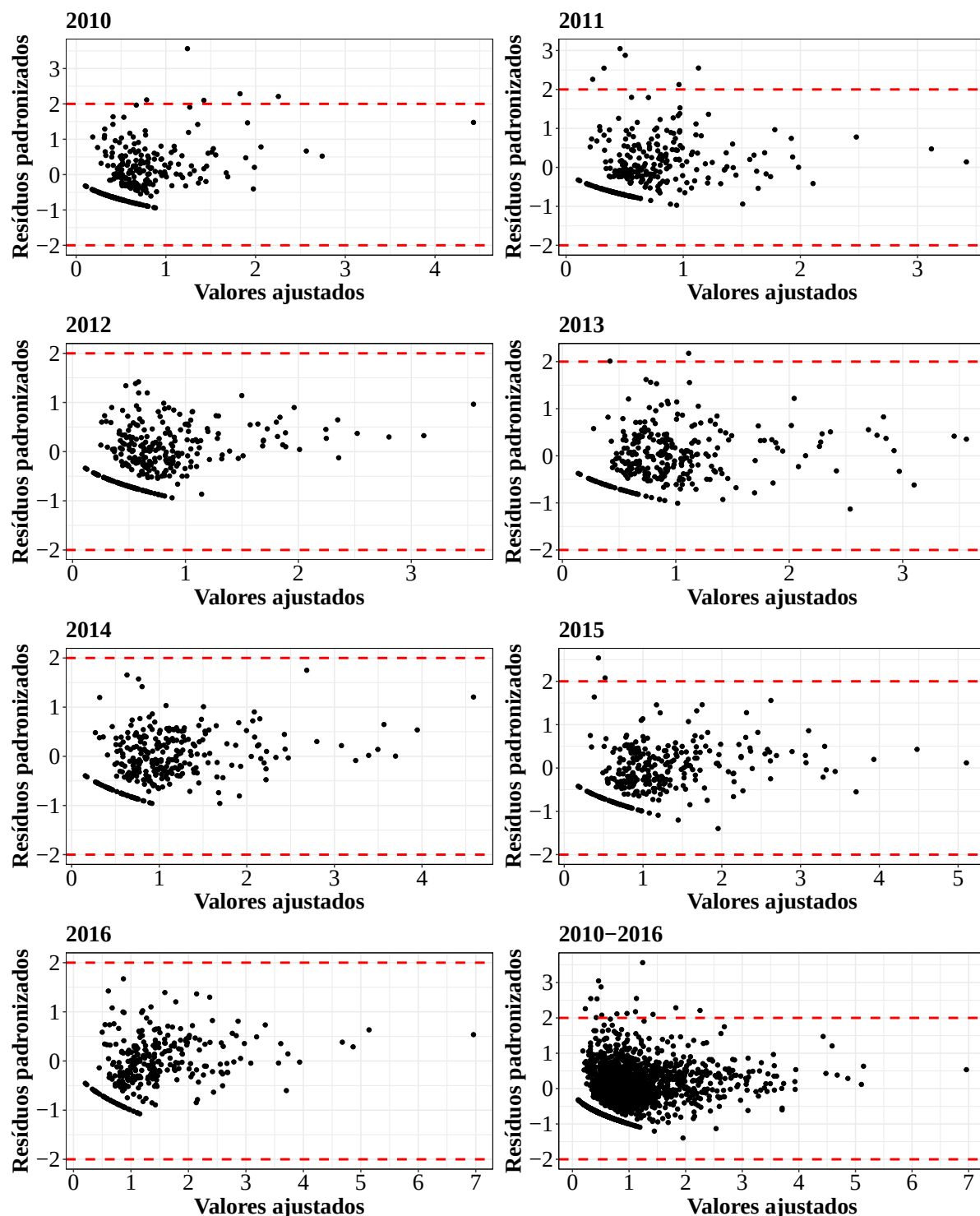


Figura 25 – Gráfico dos resíduos padronizados versus valores ajustados.

Como forma de apresentar graficamente a convergência para as estimativas dos parâmetros, mostramos os gráficos da densidade a posteriori para os estimadores de efeito fixo em médias logarítmicas do modelo de Poisson (Figura 26). Os gráficos mostram que os parâmetros se ajustaram bem como um campo aleatório gaussiano de Markov, dado que são simétricos em relação à média, e portanto, a média é igual à moda.

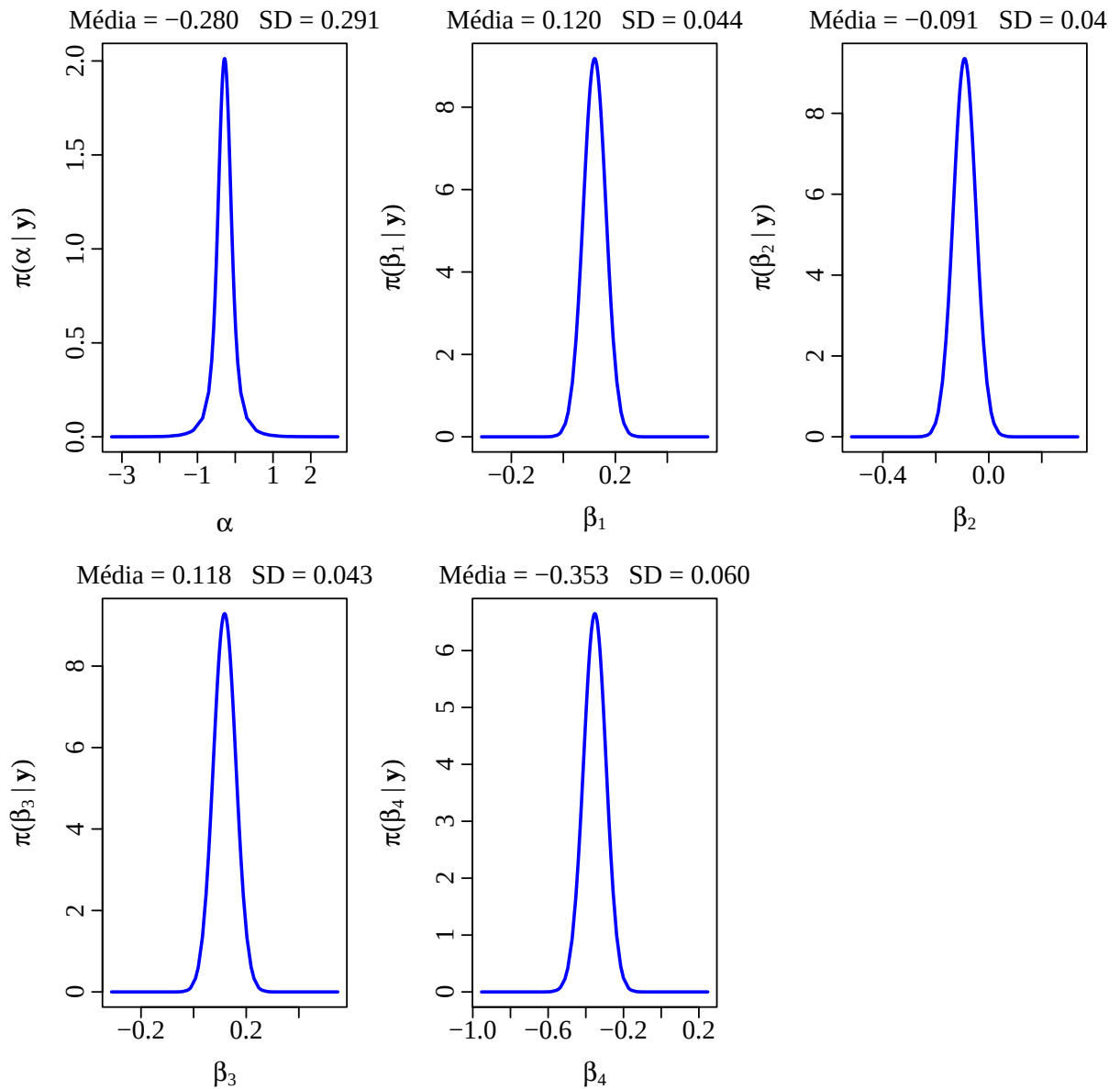


Figura 26 – Gráficos da densidade a posteriori das estimativas de efeito fixo em médias logarítmicas do modelo final.

Os casos mostraram tendências crescentes em todo o período de 2010 a 2016. Em pesquisas futuras, os dados podem ser analisados usando unidades de espaço menores e períodos maiores, o que produziria achados mais direcionados e práticos. Além disso, conforme mostrado na Tabela 11, as correlações espaciais foram diferentes em anos diferentes, ou seja, as correlações espaciais podem evoluir dinamicamente ao longo do tempo.

3.4 Discussão

Este estudo teve como objetivo estimar no espaço e no tempo o risco de SC no município de São Paulo, e avaliar sua associação com as variáveis socioeconômicas e ambientais por meio

de uma abordagem espaço-temporal Bayesiana para dados de contagem em nível de áreas de ponderação no período de 2010 a 2016.

Analizamos e mapeamos o risco relativo de SC considerando as variáveis socioeconômicas e ambientais como fatores de risco. Este estudo é uma aplicação de regressão espacial para dados epidemiológicos em pequenas áreas. Usando um modelo hierárquico espacial Bayesiano, obtivemos estimativas de risco relativo mais homogêneas do que razões de incidência padronizadas brutas, suavizando os dados por meio do modelo BYM2.

A análise Bayesiana aplicada a conjuntos de dados espaço-temporais por meio de algoritmos MCMC é computacionalmente dispendioso pois as distribuições a posterioris são alteradas de modo gradual, exigindo que os pesquisadores monitorem as estimativas a posterioris e verifiquem os diagnósticos de convergência. Por outro lado, o INLA calcula as distribuições a posterioris por meio de integração numérica e não requer análise de convergência (CARROLL et al., 2015). Embora neste trabalho tenhamos utilizado na análise 2170 unidades de espaço-tempo (ou seja, 310 APs $\times$ 7 anos), as vantagens da modelagem INLA podem ser mais plenamente percebidas em conjuntos de dados espaço-temporais maiores.

Os pontos espaço-temporais, identificados pela avaliação da probabilidade a posteriori de $\exp(\psi_{it}) > 1$, são áreas específicas que requerem recursos adicionais na saúde para lidar com um grande número de casos de SC. A probabilidade a posteriori varia de 0 a 1 e permite que as áreas de pontos ativos sejam classificadas com base na intensidade em que $\exp(\psi_{it}) > 1$.

As análises realizadas no município de São Paulo tiveram como propósito a identificação da distribuição da SC e as áreas com riscos mais elevados. Segundo Penna e Ferreira (2014) isso permite que estas áreas possam ser priorizadas pelo setor público e que os resultados obtidos tenham impacto na diminuição da ocorrência dos agravos, uma vez que se tratam de áreas de alto risco. Houve concentração de áreas de alto risco nas regiões norte e noroeste, nas partes leste e oeste da zona central e grande parte da zona sul.

Por meio dos modelos espaço-temporais, os achados mostraram que houve aumento no risco de sífilis congênita no período de 2010 a 2016. Os resultados dos modelos espaço-temporais forneceram evidências de algumas áreas com maior risco de sífilis congênita que persistiram ao longo do tempo e outras em que houve aumento no risco. Isso sugere que os fatores socioeconômicos e ambientais foram responsáveis por parte da doença nessas regiões, embora haja outros fatores para explicar sua incidência. De acordo com Teixeira et al. (2018), o conhecimento de áreas com maiores riscos pode auxiliar na tomada de decisões na saúde, com a implementação de ações já existentes e intervenções para a diminuição destes agravos por meio dos programas de prevenção. O aumento da taxa do risco de SC em algumas regiões do município pode estar associado a melhores condições de acesso e atendimento nas unidades básicas de saúde, consequentemente, melhor acesso aos métodos diagnósticos e maiores taxas de notificações (CONCEIÇÃO; CÂMARA; PEREIRA, 2020).

Este estudo destaca-se pelo uso de técnicas de análise espaço-temporal, com a integração entre os dados socioeconômicos e ambientais, disponíveis no Censo 2010 (IBGE, 2010), e a base de dados dos casos de sífilis congênita, registradas no SINAN. Estas técnicas permitem levantar hipóteses sobre possíveis fatores de risco da doença, identificando áreas de maior risco de forma oportuna para a intervenção.

Este estudo teve várias limitações. A primeira pode ser apontada como decorrente da utilização de dados secundários para a análise, o que não exclui a possibilidade dos dados estarem incompletos, prováveis sub-registros, subnotificações, erros de classificação e preenchimento, o que não isenta o presente estudo de possíveis problemas na qualidade da informação. No entanto, o Sistema Nacional de Agravos de Notificação é reconhecido por sua precisão entre os sistemas de informação brasileiros, está sempre em constante atualização. Além disso, as medidas de controle utilizadas no Brasil não têm sido suficientes para a diminuição da transmissão de sífilis, fazendo com que o país atinja anualmente um grande número de casos da doença. Outra limitação importante é a indisponibilidade de dados em nível de áreas de ponderação para variáveis altamente relacionados com casos de sífilis congênita tais como proporção de gestantes que fizeram consulta de pré-natal durante a gravidez, proporção de gestantes com pré-natal inadequado, proporção de parceiros tratados, proporção de gestantes com teste trepanêmico realizado, entre outros. Estudos de Galatoire, Rosso e Sakae (2012) e Benito e Souza (2016) mostram que a falta de tratamento ou o tratamento inadequado do parceiro poderiam explicar a transmissão vertical de sífilis. No estudo de Ruberti (2020), a proporção de mães com pré-natal inadequado, proporção de mães com menos de quatro consultas e proporção de nascidos vivos de mães adolescentes foram fatores associados à sífilis congênita. De acordo com Wijesooriya et al. (2016), pré-natais de qualidade, incluindo teste e tratamento de sífilis, são importantes para alcançar a eliminação da transmissão da sífilis de mãe para filho. O tratamento inadequado das gestantes e parceiros é o ponto chave na ocorrência da sífilis congênita, sendo uma realidade carente de foco dos órgãos públicos e que não foi aqui considerado. Segundo um estudo realizado na China por Wan et al. (2020), quando o tratamento é inadequado o risco de sífilis congênita é 3,63 vezes maior. Entretanto, por meio de diagnóstico e tratamento adequado durante o acompanhamento pré-natal, a transmissão vertical pode ser reduzida em até 97% (DOMINGUES; LEAL, 2016). Uma terceira limitação é que as covariáveis foram assumidas como constantes ao longo do tempo. Essa suposição foi imposta pela indisponibilidade de dados anuais para as covariáveis. No entanto, os modelos utilizados neste estudo podem ter reduzido o viés nas estimativas dos parâmetros.

Indicadores epidemiológicos de um município são fundamentais na execução de ações de vigilância e devem ser analisadas em vários contextos, pois um baixo número de casos de sífilis congênita não indica necessariamente que a transmissão vertical está sob controle, uma vez que a doença pode estar ocorrendo, mas não há notificação. A busca pela eliminação da sífilis congênita pode ser realizada pelos municípios por meio de projetos em consonância com as propostas desenvolvidas pela OMS (MASCHIO-LIMA et al., 2020).

O estudo conseguiu mostrar a associação que as variáveis tiveram para o aumento dos riscos da SC no Município. As covariáveis proporção de NV de mães adolescentes e a proporção de pessoas que vivem em domicílios com banheiro e água encanada não tiveram influências significativas nos casos de SC, enquanto proporção de mulheres de 18 a 24 anos sem instrução ou com ensino fundamental incompleto e índice de Gini mostraram influências significativamente positivas, e a proporção de mulheres em idade fértil e a renda média per capita mostraram influências significativamente negativas nos casos de SC.

As covariáveis renda per capita, índice de Gini, proporção de mulheres de 18 a 24 anos sem instrução ou com ensino fundamental incompleto e proporção de mulheres em idade fértil foram fatores importantes no risco de sífilis congênita.

A associação positiva da proporção de mulheres em idade fértil com a SC pareceu um pouco diferente do esperado. Nos trabalhos de [Domingues e Leal \(2016\)](#) e [Ruberti \(2020\)](#) para o estado de São Paulo, essa variável não foi significativa. Além do mais, seria de se esperar que onde tivesse maior proporção de mulheres em idade fértil, teríamos maior incidência de SC. Uma explicação para isso, talvez seja por conta de que essa proporção muda com a idade, que é uma variável individual, mas subjacente a uma variabilidade na agregação. Quanto maior a idade da mulher, mais ela busca cuidados com a saúde e assistência médica, pois já pode ter passado por uma gravidez. Quando os dados espaciais são analisados individualmente, vê-se que a idade é um fator que contrai todo o efeito, e os dados agregados podem mascarar a natureza desse efeito.

É importante destacar que o nível de escolaridade se configura como um fator importante na análise da população estudada, dado que se pode considerar uma associação entre a baixa escolaridade e a dificuldade no acesso aos serviços de saúde e prevenção de doenças. Corroborando com outros estudos, observou-se um aumento no risco em mulheres de 18 a 24 anos com baixa escolaridade, o que se assemelha a uma pesquisa realizada em Santa Cruz no Rio Grande do Norte e em outras regiões ([BOTTURA et al., 2019](#); [VASCONCELOS, 2019](#)). A baixa escolaridade materna é um fator importante que pode favorecer o aparecimento de potenciais situações de risco para a mãe e o recém-nascido, pois está associada ao baixo peso ao nascer, à mortalidade infantil e ao aumento do número de partos ([XIMENES et al., 2008](#); [HAIDAR; OLIVEIRA; NASCIMENTO, 2001](#)). O nível de escolaridade pode ser considerado um fator importante uma vez que está diretamente relacionado à falta de conhecimento e ao acesso a tratamento e métodos de prevenção, interferindo no entendimento sobre a importância do acompanhamento da saúde durante a gestação, principalmente nas medidas preventivas, dessa forma, prejudicando a interrupção da transmissão ([CONCEIÇÃO; CÂMARA; PEREIRA, 2020](#); [HERINGER et al., 2020](#)). Dessa forma, salienta-se que a educação em nosso país requer atenção especial, dado que a escolarização é uma variável importante, pois facilita a compreensão de informações, o acesso ao mercado de trabalho formal, aos serviços de saúde e, conseqüentemente, a melhoria das condições de vida e de saúde da população.

Foi possível evidenciar a questão da desigualdade mostrada por meio do Índice de Gini e

renda média per capita. Após a infecção, as gestantes com poder aquisitivo muito baixo podem ter menor chance de receber tratamento adequado, devido à falta de informação e dificuldades de acesso ao sistema de saúde (LAGO et al., 2004; MACÊDO et al., 2017; NOGUEIRA et al., 2022).

Assim, como no Brasil, na América Latina e Caribe, a sífilis apresenta impactos negativos e são amplamente descritos como fatores que contribuem para a persistência da doença na região, entre eles a epidemia de AIDS que coloca à prova a capacidade de resposta institucional dos governos às infecções sexualmente transmissíveis (as IST, sendo a AIDS uma delas) em muitos países da região. Mesmo em países desenvolvidos houve aumento da sífilis em mulheres em idade fértil e, consequentemente aumento da incidência de Sífilis Congênita. Nos Estados Unidos da América (EUA) foi registrado no período de 2004 a 2007 um aumento de 38% na prevalência da Sífilis primária e secundária entre gestantes (OLIVEIRA, 2012).

O diagnóstico precoce da infecção materna continua sendo a melhor maneira de prevenção da sífilis congênita, por meio dos testes sorológicos. As elevadas taxas de incidência de sífilis congênita no município em questão, como em outras localidades, nos levam a questionar sobre a qualidade da assistência prestada. Finalmente, vale destacar que diferentes formas de agregação espaço-temporal dos casos podem influenciar os efeitos da relação entre fatores de risco e a doença.

Por se tratar de um estudo ecológico, há dificuldade na generalização dos resultados e inferência causal, pois está sujeito a viés ecológico ou problemas da unidade de área modificável (MAUP - *Modifiable Areal Unit Problem*), isto é, pode-se obter resultados diferentes simplesmente alterando as fronteiras destas áreas (DIAS et al., 2000; OPENSHAW, 1984). Portanto, sugere-se estudar os efeitos das condições socioambientais considerando-se diferentes níveis de agregação geográfica e também a dimensão temporal.

Os achados apresentados pelo presente estudo evidenciam que a sífilis congênita ainda se encontra fora de controle em São Paulo-SP, apontando para uma baixa resolubilidade da rede de atenção básica com relação à saúde da mulher e, indicando a necessidade de revisão na qualidade da assistência, para reduzir a transmissão vertical da doença.

Ressalta-se a importância da vigilância epidemiológica da sífilis congênita em nível da atenção básica, dos distritos sanitários e da Secretaria Municipal de Saúde, coletando dados e gerando informações de qualidade para a programação das ações municipais de controle, onde cada caso novo deve ser interpretado como uma falha no processo construído para sua eliminação. A ênfase da sífilis congênita enquanto problema de Saúde Pública, junto aos gestores e profissionais responsáveis pela assistência à saúde, é de fundamental importância na condução das decisões a serem tomadas.

A eliminação da sífilis congênita vem sendo preconizado há décadas no Brasil. Apesar dos avanços alcançados, a complexidade dos fatores que interferem na cadeia de transmissão

continua desafiando os serviços de saúde (BRASIL, 2015a). Diversas estratégias de abrangência nacional tem sido executadas com auxílio do Ministério da Saúde no controle da sífilis no país, tais como: compra centralizada e distribuição de insumos de diagnóstico e tratamento (testes rápidos, penicilina benzatina e cristalina); desenvolvimento de instrumentos de disseminação de informação estratégica aos gestores, auxiliando a tomada de decisão; instrumentalização de salas de situação de saúde⁴ em todos os estados e no Distrito Federal; realização de Campanha Nacional de Prevenção; e desenvolvimento de estudos e pesquisas voltados para o enfrentamento da sífilis no SUS (BRASIL, 2020a). Nos últimos anos, houve um crescimento regular no quadro de casos de sífilis em gestantes, congênita e adquirida, que pode ser devido em parte, pela expansão da cobertura de testagem, com a extensão do uso de testes rápidos, diminuição do uso de preservativos, resistência dos profissionais de saúde à administração da penicilina na Atenção Básica e escassez em nível global de penicilina entre 2014 e 2016 (ESCOBAR et al., 2020).

Estudos mostram que a ampliação com qualidade do atendimento na Estratégia Saúde da Família e a adesão à Rede Cegonha são elementos que contribuíram para a melhoria da notificação e, conseqüentemente, o real aumento do número de casos de sífilis congênita no Brasil (SILVA et al., 2020). No entanto, alguns desafios, como a prática sexual desprotegida, são necessários para interromper a cadeia de transmissão da sífilis congênita (SOUZA et al., 2020).

Alguns países da América Latina e do Caribe mostraram que é possível a redução e até a eliminação da transmissão da sífilis congênita (ESPINOSA et al., 2018; KORENROMP et al., 2018). Em junho de 2015, Cuba tornou-se o primeiro país do mundo a receber a validação oficial da OMS de eliminação da transmissão materno-infantil de HIV e sífilis (RODRÍGUEZ et al., 2016). Em junho de 2016, a Tailândia, a Bielorrússia e a República da Moldávia também foram validadas, bem como, Anguila, Antígua e Barbuda, Bermudas, Ilhas Cayman, Montserrat, São Cristóvão e Nevis em 2017. Já em 2018 foi a vez da Malásia de compor os países que receberam essa validação pela OMS (WHO, 2019). De acordo com o relatório da OPAS (2017), 15 países notificaram em 2017 dados indicando a eliminação da sífilis congênita, mas como citado anteriormente, apenas sete receberam a validação da OMS (WHO, 2019). O Ministério da Saúde recomenda que sejam instituídos comitês de investigação para prevenção da transmissão vertical em locais que apresentem elevado número de casos de SC. Além disso, são necessárias estratégias de saúde pública voltadas à população com maior vulnerabilidade social, como fácil acesso a informações sobre a sífilis e sua prevenção e sobre a importância do tratamento, bem como capacitação dos profissionais de saúde para o melhor controle da sífilis gestacional e para o correto preenchimento das fichas de notificação (BRASIL, 2020a).

Embora o Brasil possua uma extensa gama de insumos e protocolos clínicos bem estabelecidos necessários para o diagnóstico e tratamento do agravo da SC, ainda é necessária a capacitação dos profissionais de saúde para o manejo com a sífilis gestacional, e a realização de

⁴ A sala de situação de saúde é um espaço físico e virtual onde a informação em saúde é analisada sistematicamente por uma equipe técnica, para caracterizar a situação de saúde de uma população (MINISTÉRIO DA SAÚDE, 2022).

trabalho educativo com usuários dos serviços de saúde, principalmente os adolescentes. Além disso, mecanismos de gestão e integração de políticas públicas sobre determinantes sociais são primordiais para atingir a meta de eliminação da SC preconizada pela OMS. Os serviços de saúde devem estar preparados para o atendimento adequado da gestante no pré-natal e para a promoção de atividades educativas sobre a sífilis e outras IST. Além disso, devem ser asseguradas a não discriminação da população e a confidencialidade dos dados dos pacientes.

Os resultados obtidos neste estudo podem ser úteis no desenho de novos estudos para avaliação da vigilância e controle da SC no município de São Paulo, tanto do ponto de vista espacial quanto temporal.

Por fim, é interessante considerar a importância de pesquisas futuras que possam analisar questões diretamente relacionadas às características individuais e que possam contribuir para a avaliação da persistência da SC, pois existem muitas outras estruturas e modelos espaço-temporais que podem ser usados nessa perspectiva.

3.5 Conclusão

Neste trabalho ajustamos modelos espaço-temporais Bayesianos, para analisar a distribuição espaço-temporal da incidência de sífilis congênita ocorrida no município de São Paulo no período de 2010 a 2016, e verificar as variáveis associadas ao desfecho. Foi possível evidenciar a existência de áreas de alto risco para casos de sífilis congênita, que persistem por vários anos.

Usamos modelos CAR aos dados de contagem de sífilis congênita distribuídos pelas áreas de ponderação do município. Uma vez que os dados disponíveis são contagens por unidade espacial, os modelos CAR parecem apropriados, pois podem lidar com a descrição de uma estrutura de vizinhança. Devido à estrutura hierárquica dos modelos considerados, há um desafio computacional, que pode ser diminuído utilizando-se as vantagens da abordagem INLA para inferência.

Atualmente na análise de estatística espacial e de mapeamento de doenças, os modelos Bayesianos hierárquicos são os mais utilizados, pois permitem a incorporação de diversas estruturas por meio de efeitos aleatórios, podendo assim acrescentar informações espaciais e temporais. No entanto, à medida que adiciona-se tais efeitos a complexidade do modelo aumenta. Mas o método INLA, por exemplo, que apresenta melhor desempenho computacional comparado ao MCMC, e que surgiu há pouco mais de 10 anos, ajuda nas análises em diferentes contextos e conjuntos de dados. O uso da abordagem INLA na estimação dos parâmetros mostrou-se uma alternativa atrativa e conveniente para a realização da inferência. O processo de estimação dos parâmetros não apresentou nenhum problema de convergência, pois os mesmos foram estimados maneira bem rápida, conforme já demonstrado em alguns estudos (RUE; MARTINO, 2007; RUE; MARTINO; CHOPIN, 2009). A abordagem estatística utilizada neste estudo é uma importante contribuição para a literatura e pode ser estendida para outros contextos onde há necessidade

de integrar dados de múltiplas fontes em grandes escalas espaciais e temporais para o ganho de mais poder estatístico.

As pesquisas e esforços dos órgãos públicos devem continuar a fim de controlar, tratar e, eventualmente, erradicar a sífilis congênita em todo o mundo. O impacto na saúde pública da sífilis na gravidez e na infância continua a ser significativa e a eliminação da transmissão vertical somente se tornará uma realidade por meio de serviços de saúde de alta qualidade.

Verificou-se que as covariáveis, renda per capita, índice de Gini, proporção de mulheres em idade fértil, proporção de mulheres de 18 a 24 anos sem instrução ou com ensino fundamental incompleto e proporção de pessoas pretas foram fatores importantes no risco de sífilis congênita. Sua eliminação representa um grande desafio para o município. Assim, são necessárias intervenções para a eliminação da sífilis congênita e estas devem ser direcionadas principalmente para áreas que apresentam um risco elevado e persistente de casos.

Em resumo, a sífilis congênita ainda apresenta impacto negativo, em que o surgimento de casos da doença manifestam deficiências de ordem tanto estrutural como técnico nos serviços de saúde. Assim mesmo, apesar de que a cobertura de atenção pré-natal ser aceitável, a falta de detecção e tratamento sistemáticos da doença e fatores socioeconômicos determina que a sífilis congênita continue sendo um problema de saúde pública. Trabalhos futuros precisam considerar o ajuste para outros fatores no nível agregado e no nível individual.

4 Modelagem de dados espaço-temporais da sífilis congênita no município de São Paulo de 2010 a 2016 usando modelos aditivos generalizados

Resumo

O mapeamento das distribuições espaciais de ocorrência e risco de doenças pode ser uma ferramenta útil para identificar exposições de interesse da saúde pública. Os dados do registro de doenças geralmente são mapeados por estado, município, áreas de ponderação ou setores censitários. Os modelos aditivos generalizados (GAMs) fornecem uma estrutura útil para mapear dados epidemiológicos baseados em coordenadas geográficas. Geramos mapas de riscos usando toda a área de estudo como referência. O objetivo desse estudo foi analisar o risco de sífilis congênita (SC) no município de São Paulo-SP no período de 2010 a 2016, em função de variáveis socioeconômicas e ambientais, considerando a localização espacial do centroide das áreas de ponderação (longitude e latitude) e o tempo. Os dados foram obtidos do Sistema de Informação de Agravos de Notificação (SINAN), do Sistema de Informação sobre Nascidos Vivos (SINASC) e Instituto Brasileiro de Geografia e Estatística (IBGE). Modelos de riscos com efeitos espaciais e temporais aditivos, incorporada nos modelos aditivos generalizados, foram ajustados, possibilitando obter os riscos espaço-temporais de SC no município. As análises mostram aumento do risco ao longo dos anos e diferentes variáveis associadas à SC no período analisado.

Abstract

Mapping the spatial distributions of disease occurrence and risk can be a useful tool to identify exposures of public health interest. Disease registry data are usually mapped by state, municipality, weighting areas, or census tracts. Generalized additive models (GAMs) provide a useful framework for mapping epidemiological data based on geographic coordinates. We generate risk maps using the entire study area as a reference. The objective of this study was to analyze the risk of

congenital syphilis (CS) in the municipality of São Paulo-SP from 2010 to 2016, depending on socioeconomic and environmental variables, considering the spatial location of the centroid of the weighting areas (longitude and latitude) and time. Data were obtained from the Notifiable Diseases Information System (SINAN), the Live Birth Information System (SINASC) and the Brazilian Institute of Geography and Statistics (IBGE). Risk models with additive spatial and temporal effects, incorporated in the generalized additive models, were adjusted, making it possible to obtain the spatio-temporal risks of SC in the municipality. The analyzes show an increase in risk over the years and different variables associated with CS in the analyzed period.

4.1 Introdução

Geralmente, os dados de incidência ou mortalidade de doenças raras têm sido modelados usando modelo de Poisson com uma distribuição autoregressiva condicional (CAR, do inglês *conditional autorregressive*) para os efeitos aleatórios da região que levam em conta a dependência espacial. Uma das possíveis extensões para analisar dados de contagem espaço-temporal tem sido a inclusão de um efeito aleatório temporal com uma distribuição do tipo CAR (KNORR-HELD, 2000). Uma abordagem alternativa é o uso de modelos P-spline, que são diferentes dos modelos CAR geralmente usados no mapeamento de doenças, pois suavizam as tendências espaço-temporais globais, enquanto os modelos CAR tradicionais levam em consideração os efeitos locais (UGARTE; GOICOA; MILITINO, 2010). O modelo Besag, York e Mollié (BYM) é um modelo CAR de regressão Bayesiana em nível de área com componentes para correlação espacial e heterogeneidade não espacial que tem sido bastante usado na literatura para a detecção de áreas de alto risco, em que distribuições a posterioris para os parâmetros de nível de área são usadas para identificar áreas de risco elevado.

Os modelos aditivos generalizados (GAM, do inglês *generalized additive models*) são comumente usados para determinar áreas de alto ou baixo risco para uma variável resposta ajustada por covariáveis, incluindo confundidores espaciais (WEBSTER et al., 2006). O desempenho desse tipo de modelo pode diferir devido a diferentes características das áreas de risco elevado em relação à forma, tamanho, localização e nível de risco ou probabilidade de doença (SIANGPHOE; WHEELER, 2015).

Desde a sua introdução, os modelos GAM têm sido usados extensivamente na área ecológica (YEE; MITCHELL, 1991; MARTÍN; GONZÁLEZ-ARIAS; VICENTE-VÍRSEDA, 2021), muitas vezes para modelar a distribuição de espécies de plantas e animais (FEWSTER et al., 2000; AUSTIN, 2002; LEATHWICK; ELITH; HASTIE, 2006; MOISEN et al., 2006; DREXLER; AINSWORTH, 2013), na área ambiental (NGANDO et al., 2020; KELLER et al., 2016; LU et al., 2020; LONG et al., 2021) e na área epidemiológica (CHE-HIM et al., 2018; GAUDART et al., 2013; ATEBA et al., 2020; UGWU; ZEWOTIR, 2020; FENG, 2021).

Na maioria dos modelos de regressão clássicos, assume-se que o efeito para cada cova-

riável seja linear ou polinomial. No entanto, isso muitas vezes não é suficientemente flexível para abranger o verdadeiro efeito subjacente, podendo resultar em estimativas tendenciosas (SPIEGEL; KNEIB; OTTO-SOBOTKA, 2020). Então Stone (1986) introduziram a classe de modelos GAM, em que o efeito por covariável é definido como uma curva suave. Esse método apresenta vantagens na análise de relações não lineares complexas para obter melhores previsões (FENG; TONG, 2017).

O objetivo deste estudo foi analisar o risco de sífilis congênita com os efeitos espacial e temporal e variáveis associadas, considerando-se o período de 2010 a 2016 no município de São Paulo. Os modelos GAM constituem um método flexível para capturar relações não lineares, e até onde sabemos, ainda não foram utilizados na modelagem do risco de SC.

4.2 Materiais e Métodos

O estudo utilizou três fontes de informação: bases de dados com os casos notificados de sífilis congênita do Sistema de Informação de Agravos de Notificação (SINAN) no período de 2010 a 2016, os dados de nascimentos coletados no Sistema de Informações sobre Nascidos Vivos (SINASC) e dados do Censo Demográfico de 2010 do IBGE (Instituto Brasileiro de Geografia e Estatística).

Após a checagem e organização do banco de dados, as variáveis explicativas candidatas com base em sua importância e disponibilidade de dados utilizadas no estudo foram: socioeconômicas, ambientais, as localizações dos centroides georreferenciadas (longitude e latitude) e o tempo (ano) descritas na Tabela 12.

Essas covariáveis foram selecionadas baseado em estudos anteriores como fatores relevantes associados à ocorrência de sífilis congênita (MELO; FILHO; FERREIRA, 2011; OLIVEIRA et al., 2014; SERAFIM et al., 2014; DOMINGUES; LEAL, 2016; REIS et al., 2018; RUBERTI, 2020), além dos efeitos espaciais e temporais considerados como funções suaves das coordenadas geográficas dos centroides das áreas de ponderação e dos anos em que foram obtidos os casos de SC. As variáveis longitude, latitude e ano foram incluídas em todos os modelos para explorar as tendências espaciais e temporais.

4.2.1 Multicolinearidade

Antes da modelagem, uma etapa importante é a seleção de variáveis representativas de uma variedade de variáveis potenciais. Para a seleção das variáveis socioeconômicas e ambientais neste estudo, o fator de inflação da variância (VIF, do inglês *variance inflation factor*) foi calculado para avaliar a multicolinearidade, que é um problema potencial em todas as análises de regressão. Quanto maior o seu valor, mais severa é a multicolinearidade. Aqui consideramos $VIF < 5$ como o padrão de triagem (VATCHEVA et al., 2016).

Tabela 12 – Variáveis socioeconômicas e ambientais utilizadas como potenciais variáveis explicativas no modelo GAM para a distribuição espaço-temporal de sífilis congênita.

Variável	Descrição
PBRANCA	Proporção de pessoas de cor branca: Percentual de residentes que se autodeclararam brancos, multiplicado por 100.
PPARDA	Proporção de pessoas de cor parda: Percentual de residentes que se autodeclararam pardos, multiplicado por 100.
PPRETA	Proporção de pessoas de cor preta: Percentual de residentes que se autodeclararam pretos, multiplicado por 100.
PNVMA	Proporção de nascidos vivos de mães adolescentes: Razão entre o número de nascidos vivos de mulheres de 12 a 17 anos e o total de nascidos vivos, multiplicada por 100.
MSIF1824	Percentual de mulheres de 18 a 24 anos de idade sem instrução ou com ensino fundamental incompleto: Razão entre a população de mulheres de 18 a 24 anos sem instrução ou com ensino fundamental incompleto e o total de mulheres nesta faixa etária, multiplicada por 100.
PFERTIL	Proporção de mulheres em idade fértil: Razão entre o número de mulheres de 15 a 49 anos e o total de mulheres, multiplicada por 100.
PBA	Percentual da população que vive em domicílios com banheiro e água encanada: Razão entre a população que vive em domicílios particulares permanentes com água encanada em pelo menos um de seus cômodos e com banheiro exclusivo e a população total residente em domicílios particulares permanentes, multiplicada por 100. A água pode ser proveniente de rede geral, de poço, de nascente ou de reservatório abastecido por água das chuvas ou carro-pipa. Banheiro exclusivo é definido como cômodo que dispõe de chuveiro ou banheira e aparelho sanitário.
RDPC	Renda per capita média: Razão entre o somatório da renda de todos os indivíduos residentes em domicílios particulares permanentes e o número total desses indivíduos. Valores em reais de 01/agosto de 2010.
GINI	Índice de Gini: Mede o grau de desigualdade existente na distribuição de indivíduos segundo a renda domiciliar per capita. Seu valor é 0 quando não há desigualdade (a renda domiciliar per capita de todos os indivíduos tem o mesmo valor) e tende a 1 à medida que a desigualdade aumenta. O universo de indivíduos é limitado àqueles que vivem em domicílios particulares permanentes
xcoord	coordenada geográfica cartesiana da longitude em UTM do centroide da área de ponderação. Zona 23 Sul, Datum SIRGAS 2000.
ycoord	coordenada geográfica cartesiana da latitude em UTM do centroide da área de ponderação. Zona 23 Sul, Datum SIRGAS 2000.
Ano	Tempo

Geralmente, diferentes fatores representativos do nível socioeconômico possuem multicolinearidade, e uma maneira conveniente de lidar com essa questão é escolher aqueles que são mais representativos (com o menor VIF) ao invés de selecionar todos eles. O mesmo se aplica às variáveis ambientais.

O VIF é definido por

$$VIF_j = \frac{1}{1 - R_j^2}$$

em que R_j^2 é o R^2 da regressão da j -ésima variável explicativa sobre as demais variáveis explicativas do modelo.

A presença de multicolinearidade pode ser particularmente problemática para estudos que usam métodos que dependem de estimativas dos parâmetros por intervalo. Uma consequência comum é o aumento nos erros padrão das estimativas, podendo levar a intervalos de confiança maiores, ocasionando um impacto no julgamento da importância de um preditor. Entre pesquisadores aplicados e metodológicos, não há consenso sobre quanta ênfase colocar em um valor VIF, nem qual ponto de corte é mais apropriado. Alguns pesquisadores afirmam que um VIF maior que 10 indica a presença de multicolinearidade (O'BRIEN, 2007; GARCÍA et al., 2016; THOMPSON et al., 2017).

4.2.2 Análises espaciais e temporais

As variáveis espaciais, correspondentes às coordenadas x e y dos centroides das áreas de ponderação e a variável temporal Ano, foram utilizadas para analisar os padrões espaço-temporais do risco.

Ajustamos modelos da forma apresentada na Equação (2.3) usando uma distribuição de Poisson para modelar a variável resposta e, em seguida, uma distribuição Binomial Negativa para observar se houve melhorias nos ajustes. Utilizaram-se os critérios de informação de Akaike (AIC) e o critério de informação Bayesiano (BIC) para selecionar o modelo mais adequado. A contribuição relativa de cada variável no modelo de melhor ajuste foi avaliada retirando-se as variáveis individuais com menor significância e calculando a mudança na deviance total explicada entre o modelo completo e o modelo com variáveis descartadas. A deviance explicada (DE) foi obtida para examinar a capacidade de explicação dos efeitos das variáveis, em que um valor de DE maior representa uma melhor capacidade explicativa.

Os modelos foram ajustados usando o método de máxima verossimilhança restrita (REML, do inglês *restricted maximum likelihood*) de Patterson e Thompson (1971) (ver Cressie (1993) e Wood, Pya e Säfken (2016) para mais detalhes).

Além disso, realizamos testes de verificação para selecionar a melhor *spline* para as variáveis espaço-temporais. O parâmetro de suavização foi escolhido controlando o equilíbrio entre a função de verossimilhança e o sobreajuste, fundamentando a convergência do algoritmo de iteração.

4.2.3 Avaliação e qualidade do ajuste

4.2.3.1 Critério de informação de Akaike

Uma medida comumente usada de adequação de modelos frequentistas é o critério de informação de Akaike (AIC, do inglês *Akaike information criterion*) proposto por Akaike (1973). A ideia é selecionar um modelo que minimize a verossimilhança negativa penalizada pelo número de parâmetros, sendo definida por:

$$\text{AIC} = -2\ell(\hat{\boldsymbol{\theta}} \mid \mathbf{x}) + 2p,$$

em que $\ell(\hat{\boldsymbol{\theta}} \mid \mathbf{x})$ é o valor do logaritmo de verossimilhança do modelo maximizado e p é o número de variáveis explicativas no modelo (incluindo o intercepto).

4.2.3.2 Critério de informação Bayesiano

O critério de informação Bayesiano (BIC, do inglês *bayesian information criterion*) é comumente utilizado como um critério de escolha entre modelos Bayesianos e hierárquicos, proposto por Schwarz (1978) e estreitamente relacionado ao AIC discutido acima, mas está fortemente ligado à teoria Bayesiana. O BIC introduz o termo de penalidade para o número de parâmetros nos modelos sempre que ocorrer sobreajuste. É quando a probabilidade é aumentada pela adição dos parâmetros. O BIC é definido como:

$$\text{BIC} = -2\ell(\hat{\boldsymbol{\theta}} \mid \mathbf{x}) + p \log(n),$$

em que $\ell(\hat{\boldsymbol{\theta}} \mid \mathbf{x})$ é o valor logarítmico de verossimilhança maximizado, p é o número de variáveis explicativas no modelo (incluindo o intercepto) e n é o tamanho da amostra.

Kass (1993) e Kass e Raftery (1995) sugerem o uso do critério de informação Bayesiano (BIC) como substituto para o cálculo do Fator Bayes quando tais cálculos são difíceis, pois o BIC pode ser calculado sem especificar a priori (GILL, 2015).

Modelos com valores mais baixos de AIC e BIC são considerados melhores. O número de parâmetros p no modelo determina a penalidade para a complexidade do modelo no cálculo do AIC e do BIC.

4.2.3.3 Precisão e qualidade de ajuste do modelo

Para comparar a qualidade do ajuste e desempenho dos modelos, apresentamos algumas funções matemáticas, referidas como métricas estatísticas comumente usadas em ciências ambientais e sociais, a saber, raiz quadrada do erro quadrático médio (RMSE, do inglês *root mean square error*), o erro absoluto médio (MAE, do inglês *mean absolute error*) e o coeficiente de determinação (R^2). As métricas de erro são descritas pelas Equações (4.1) e (4.2) e , cujos detalhes podem ser vistos em Isaaks e Srivastava (1989), Willmott e Matsuura (2005) e Chai e

Draxler (2014).

$$\text{RMSE} = \sqrt{\frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T (y_{it} - \hat{y}_{it})^2} \quad (4.1)$$

$$\text{MAE} = \frac{1}{nT} \sum_{i=1}^n \sum_{t=1}^T |y_{it} - \hat{y}_{it}| \quad (4.2)$$

em que y_{it} é o valor observado e $\hat{y}_{it}$ é o valor predito na área i e tempo t , e nT é o tamanho da amostra. A partir das equações acima notamos que essas métricas medem o desempenho em termos do desvio de suas predições e dos valores observados. Assim, o modelo que apresenta métricas com valores menores, normalmente indicam que ele é o melhor.

Por fim, o coeficiente de determinação (R^2) é comumente usado em modelos de regressão para avaliar a variação explicada ou variância, que é definida como o grau em que os valores preditos e reais são consistentes. Um valor mais alto de R^2 indica uma variação maior que o modelo pode explicar e mede o quão bem o modelo se ajusta aos dados, sendo dado pela Equação (4.3):

$$R^2 = 1 - \frac{\sum_{i=1}^n \sum_{t=1}^T (y_{it} - \hat{y}_{it})^2}{\sum_{i=1}^n \sum_{t=1}^T (y_{it} - \bar{y})^2} \quad (4.3)$$

em que $\bar{y}$ é o valor médio global de y e y_{it} , $\hat{y}_{it}$, n conforme definidos acima. O R^2 mede a proporção ou percentual da variação total de y explicada pelo modelo de regressão, e seu valor varia de 0 a 1. Quando $R^2 = 1$ temos um ajustamento perfeito, ou seja, $\hat{y}_{it} = y_{it}$ para cada i e t . Por outro lado, se $R^2 = 0$ não há qualquer relação entre a resposta e as variáveis explicativas. Neste caso, $\hat{y}_{it} = \bar{y}$, e a melhor predição para qualquer valor de y é seu valor médio (GUJARATI; PORTER, 2011).

4.2.4 Análise de resíduos

A verificação do modelo é a busca por evidências de que ele está substancialmente correto. Como em qualquer análise de regressão, uma análise de resíduos completa deve ser empregada. Para a verificação do ajuste, utilizamos os gráfico dos resíduos de deviance e QQ-plot, sendo que neste último espera-se que a maioria dos resíduos estejam dentro do envelope simulado. Os resíduos de deviance são baseados em duas vezes a diferença entre o valor máximo que a contribuição da verossimilhança para y_i poderia ter tomado e o valor (geralmente menor) que ela assumiu no modelo. A exibição gráfica de valores residuais versus valores ajustados ou preditores pode fornecer diagnósticos mais focados.

Além disso, estudamos a autocorrelação espacial nos resíduos dos modelos GAMs para cada ano, que foi estimada calculando-se os valores do índice I de Moran (MORAN, 1950) nos resíduos do modelo final por meio do pacote spde. Os valores I de Moran variam de -1 a 1 ,

com valores positivos altos indicando associação espacial positiva entre os locais (semelhança) e valores negativos altos indicando associação negativa (dissimilaridade). Valores em torno de 0 indicam distribuição aleatória sem associação espacial. Para verificar a hipótese nula de que todas as taxas dinâmicas são distribuídas aleatoriamente na área, realizamos um teste de permutação de Monte Carlo para a estatística I de Moran usando 10000 permutações aleatórias dos valores para o esquema de ponderação espacial dado.

Todos os modelos foram implementados no software estatístico R versão 4.1.2 (TEAM, 2021) usando o pacote mgcv (RUE; MARTINO; CHOPIN, 2009; BLANGIARDO; CAMELETTI, 2015), que utiliza a abordagem GAM para estimação dos modelos e os mapas foram gerados a partir do pacote tmap (TENNEKES, 2018) usando shapefile da malha de áreas de ponderação do ano 2010 obtido no portal do IBGE (<<https://portaldemapas.ibge.gov.br/portal.php#homepage>>). As análises foram conduzidas em um microcomputador com processador Intel-core i5 com 2,71 GHz e 16,0 GB de RAM. A união do *shapefile* de casos com o *shapefile* das áreas de ponderação foi realizado por meio do software QGIS versão 3.10, permitindo identificar, além da distribuição dos casos no município, as regiões em que estavam inseridos.

4.3 Resultados

Nesta seção, são apresentadas as análises dos GAMs que modelam a distribuição temporal e espacial do risco de sífilis congênita no município de São Paulo. A modelagem espaço-temporal do número de casos de SC precisa levar em conta a variabilidade interanual, o ano e as covariáveis espaciais, como longitude e latitude. Outros fatores, como a renda, podem afetar a distribuição temporal e espacial dos casos.

No processo de modelagem, ajustamos os modelos de Poisson e Binomial Negativa sem os efeitos espacial e temporal, com apenas o efeito espacial, com apenas o efeito temporal, com apenas o efeito de interação espaço-tempo, e com todos os efeitos juntamente com as covariáveis do estudo. Foram ajustados um total de 12 modelos e as comparações foram baseadas em AIC, BIC, REML, RMSE, MAE, R^2 e DE apresentadas na Tabela 13.

A aplicação do GAM não linear (Modelo P2) melhorou modestamente o modelo linear de base (Modelo P1), em que o R^2 aumentou de 0,23 para 0,51 e o RMSE diminuiu de 2,52 para 2,01. A inclusão de uma interação espaço-temporal (Modelo P5) melhorou ainda mais o ajuste, aumentando o R^2 de 0,51 para 0,62 e diminuindo o RMSE de 2,01 para 1,77 (Tabela 13). Além disso, examinar as contribuições separadas de tempo (Modelo P2) e espaço (Modelo P3) mostra a importância que a componente espacial desempenha na associação com o risco de SC. Mas nenhum tem um desempenho tão bom quanto os Modelos P5 e P6, sendo que o modelo com apenas a componente espacial (Modelo P2, $R^2 = 0,51$; RMSE = 2,01; MAE = 1,43) supera significativamente o modelo com apenas a componente temporal (Modelo P3, $R^2 = 0,31$; RMSE = 2,38; MAE = 1,66). As mesmas considerações são feitas para os modelos com distribuição Binomial Negativa.

Tabela 13 – Modelos ajustados, critério de informação de Akaike (AIC), critério de informação Bayesiano (BIC), logaritmo da verossimilhança marginal restrita (REML), raiz quadrada do erro quadrático médio (RMSE), erro absoluto médio (MAE), coeficiente de determinação (R^2) e deviance explicada (DE). Os melhores modelos são destacados em negrito.

Poisson							
Modelo ^a	AIC	BIC	REML	RMSE	MAE	R^2	DE
P1	9329,88	9375,34	4682,37	2,52	1,76	0,23	0,06
P2	8185,60	8958,21	4191,24	2,01	1,43	0,51	0,35
P3	8948,63	9001,61	4492,98	2,38	1,66	0,31	0,14
P4	7813,75	8616,62	4003,32	1,80	1,32	0,61	0,43
P5	7780,00	8706,36	3979,74	1,77	1,30	0,62	0,45
P6	7786,93	8738,32	3978,71	1,77	1,30	0,62	0,45
Binomial Negativa							
Modelo	AIC	BIC	REML	RMSE	MAE	R^2	DE
NB1	8544,61	8595,75	4285,65	2,53	1,76	0,23	0,06
NB2	8065,86	8527,14	4084,07	2,10	1,49	0,47	0,29
NB3	8380,16	8440,71	4204,55	2,40	1,67	0,30	0,13
NB4	7803,51	8390,27	3965,12	1,88	1,35	0,57	0,38
NB5	7768,26	8552,15	3949,83	1,81	1,32	0,60	0,41
NB6	7773,26	8574,00	3948,83	1,81	1,32	0,60	0,41

^aOs modelos possuem as seguintes estruturas:

$$\text{NB1: } \eta = \alpha + X\beta$$

$$\text{NB2: } \eta = \alpha + X\beta + s(x, y)$$

$$\text{NB3: } \eta = \alpha + X\beta + s(t)$$

$$\text{NB4: } \eta = \alpha + X\beta + s(x, y) + s(t),$$

$$\text{NB5: } \eta = \alpha + X\beta + s(x, y, t)$$

$$\text{NB6: } \eta = \alpha + X\beta + s(x, y) + s(t) + s(x, y, t)$$

$s(x, y)$ é a suavização espacial, $s(t)$ é a suavização temporal e $s(x, y, t)$ é a suavização para a interação espaço-tempo.

Em seguida consideramos os modelos P6 e NB6 como sendo os melhores. Como os valores AIC e BIC eram muito próximos para esses modelos, a escolha foi baseada naqueles que apresentaram menores valores de REML, RMSE, MAE e maior valor de R^2 . Finalmente, o modelo final escolhido foi o NB6, por apresentar menores valores de AIC e BIC e melhores ajustes com relação aos resíduos de deviance (Tabela 14, Figuras 27 e 28). Também são exibidos os gráficos QQ dos resíduos de deviance para os modelos de Poisson e Binomial Negativa com envelopes simulados. Conforme mostrado, para o modelo de Poisson, alguns resíduos em uma faixa mais alta ficam fora do envelope simulado. Por outro lado, os pontos nos gráficos QQ baseados no modelo Binomial Negativa se encontram principalmente dentro dos envelopes, indicando que o modelo se ajusta bem aos dados. Os envelopes simulados foram obtidos com base em 1.000 dados simulados do modelo ajustado. Para um modelo bem ajustado, espera-se que a maioria dos resíduos fiquem desse do envelope. As Figuras 27 e 28 mostram também os gráficos dos resíduos de deviance versus os valores ajustados, verificando-se que o modelo NB6 apresenta melhor ajuste, com valores concentrados entre -2 e 2 e poucos valores acima de 2 ou abaixo de -2.

A Tabela 15 resume as estimativas para os parâmetros dos efeitos fixos, valores do RR

Tabela 14 – Critério para escolha do melhor modelo.

Modelo	AIC	BIC	REML	RMSE	MAE	R ²
P6	7786,93	8738,32	3978,71	1,77	1,30	0,62
NB6	7773,26	8574,00	3948,83	1,81	1,32	0,60

das covariáveis selecionadas no modelo NB6 e intervalos de confiança de 95%. Os efeitos das variáveis proporção de pessoa pretas, proporção de NV de mães adolescentes, proporção de mulheres em idade fértil e proporção de pessoas que vivem em domicílios com banheiro e água encanada, não foram significativas, pois a média igual a zero se encontra no intervalo de confiança de 95%. Os fatores significativos no resultado da regressão foram usados para explicar o risco relativo de covariáveis para toda a área de estudo. Entre as variáveis do estudo, o risco de incidência de sífilis congênita aumentou com o aumento de um desvio padrão no índice de Gini ($RR = 1,085$) e na proporção de mulheres de 18 a 24 anos sem instrução ou com ensino fundamental incompleto ($RR = 1,079$). Por outro lado, o risco de incidência de sífilis congênita diminuiu com o aumento de um desvio padrão na renda média per capita ($RR = 0,754$).

Tabela 15 – Resumo das estimativas do modelo selecionado NB6 com todas as covariáveis para os efeitos fixos e termos suavizados.

Efeitos fixos	Média	SD	valor t	valor p	RR
(Intercept)	-0,2278	0,0188	-12,1297	< 0,0001	0,796
PPRETA	0,0662	0,0417	1,5854	0,1129	1,068
PNVMA	0,0078	0,0240	0,3239	0,7460	1,001
MSIF1824	0,0756	0,0322	2,3481	0,0189	1,079
PFERTIL	-0,0186	0,0396	-0,4707	0,6379	0,982
GINI	0,0818	0,0346	2,3598	0,0183	1,085
PBA	-0,0344	0,0249	-1,3824	0,1669	0,966
RDPC	-0,2820	0,0809	-3,4835	0,0005	0,754
Termos suavizados	EDF ¹	Ref.df ²	valor F	valor p	
s(xcoord,ycoord)	2,0016	2,0020	4,9313	0,0851	
s(Ano)	1,3937	1,6850	262,7285	< 0,0001	
s(xcoord,ycoord,Ano)	124,6175	306,0000	671,3190	< 0,0001	

¹ graus de liberdade efetivos (EDF, do inglês *effective degrees of freedom*).

² Graus de liberdade de referência (Ref.df, do inglês *reference degrees of freedom*) usados no cálculo da estatística de teste e dos valores p .

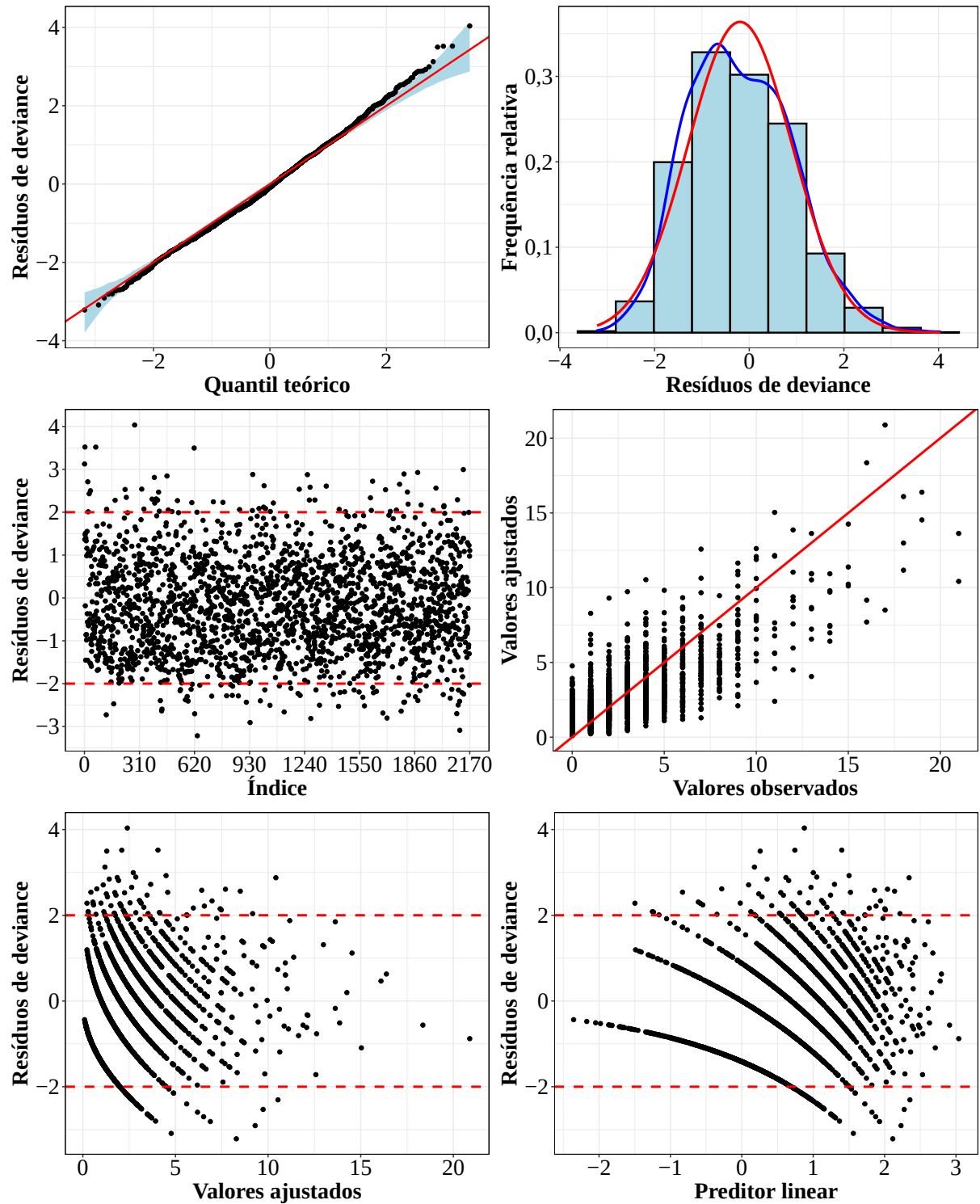


Figura 27 – Verificação dos pressupostos básicos do modelo P6 e análise de resíduos usando a metodologia GAM.

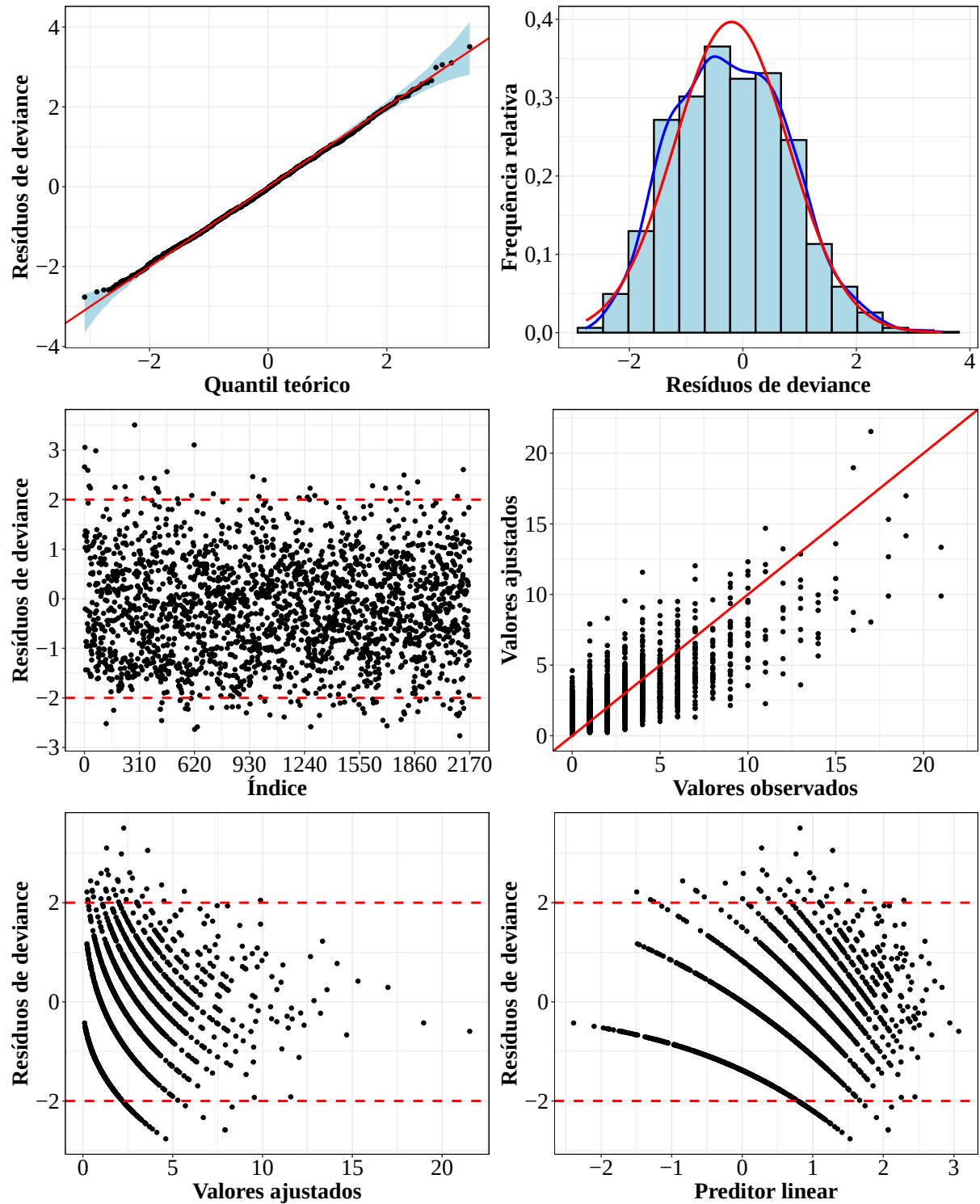


Figura 28 – Verificação dos pressupostos básicos do modelo NB6 e análise de resíduos usando a metodologia GAM.

A contribuição relativa de cada variável no modelo de melhor ajuste (apresentado na Tabela 15) foi avaliada retirando-se uma a uma as variáveis individuais não significantes e reajustando novamente o modelo. O modelo final é apresentado na Tabela 16.

Tabela 16 – Estimativas dos efeitos fixos e termos suavizados com intervalos de confiança de 95% para ocorrência de sífilis congênita no modelo Binomial Negativa espaço-temporal após seleção de covariáveis, município de São Paulo (2010-2016).

Efeitos fixos	Média	SD	valor t	valor p	RR
Intercepto	-0,2261	0,0188	-12,0432	< 0,0001	0,798
MSIF1824	0,0845	0,0295	2,8705	0,0041	1,088
GINI	0,0855	0,0341	2,5104	0,0121	1,089
RDPC	-0,3116	0,0776	-4,0137	0,0001	0,732
Termos suavizados	EDF ^a	Ref.df ^b	Valor F	Valor p	
s(xcoord,ycoord)	2,0146	2,0180	4,5124	0,1063	
s(Ano)	5,0003	5,0007	278,2887	< 0,0001	
s(xcoord,ycoord,Ano)	126,1439	301,0000	701,1158	< 0,0001	

^a Graus de liberdade efetivos (EDF, do inglês *effective degrees of freedom*).

^b Graus de liberdade de referência (Ref.df, do inglês *reference degrees of freedom*) usados no cálculo da estatística de teste e dos valores p .

Nota-se que o risco de incidência de sífilis congênita aumentou com o aumento de um desvio padrão no índice de Gini ($RR = 1,089$) e na proporção de mulheres de 18 a 24 anos sem instrução ou com ensino fundamental incompleto ($RR = 1,088$). Por outro lado, o risco de incidência de sífilis congênita diminuiu com o aumento de um desvio padrão na renda média per capita ($RR = 0,732$).

Para avaliar a qualidade do ajuste do modelo selecionado anteriormente, foram obtidos gráficos utilizando todos os valores preditos versus seus valores reais e o histograma da diferença entre esses valores considerando todas as 310 áreas de ponderação. A Figura 29(a) mostra uma relação linear considerável da forma $y = x$ entre os valores estimados e observados, que foi sustentada por um valor de R^2 de 0,60 indicando que 60% da variabilidade espaço-temporal do risco pode ser explicada pelo modelo. Observando a análise visual das diferenças entre esses valores, notamos que elas estão razoavelmente distribuídas simetricamente, ou seja, o gráfico de dispersão dos valores observados versus os estimados no período 2010-2016, bem como o histograma das diferenças entre esses valores (Figura 29(b)), mostram que as predições não estão significativamente enviesadas.

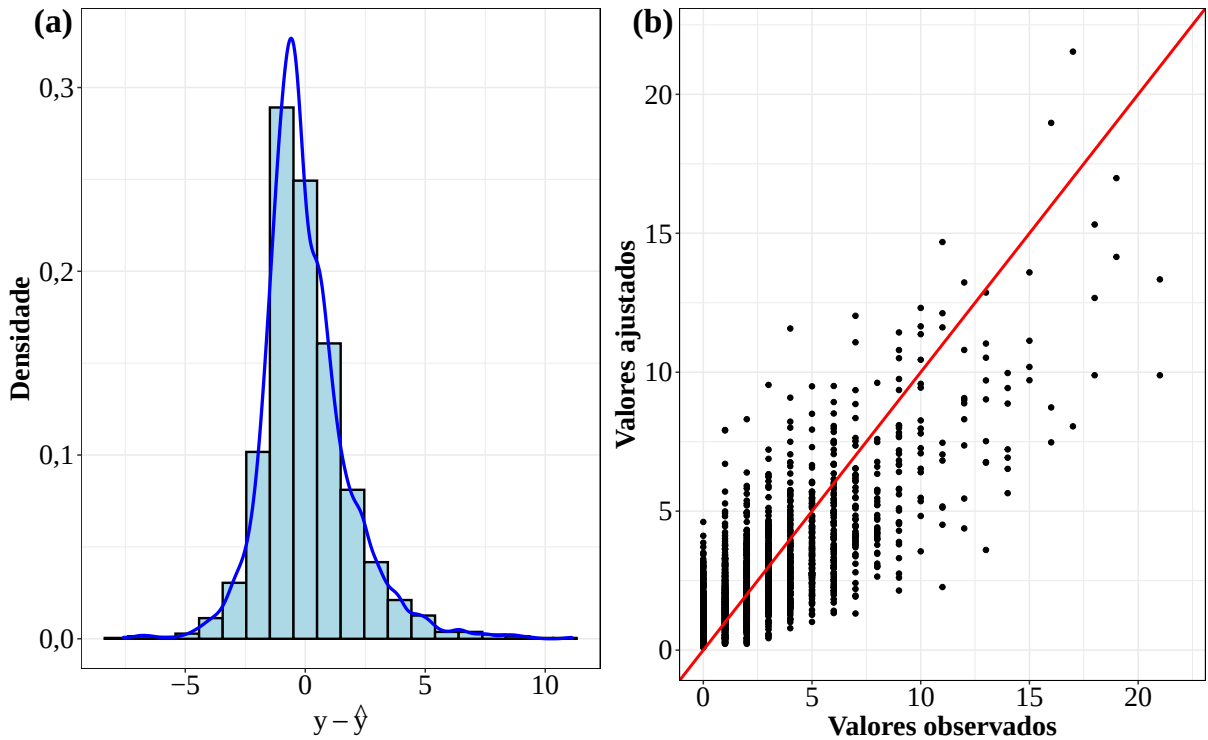


Figura 29 – Gráficos de diagnóstico para o modelo final após seleção de covariáveis usando a metodologia GAM.

O teste `k.check()` do pacote `mgcv` (Tabela 17) mostra que os graus de liberdade máximos padrão para as funções suavizadas usadas no modelo final são suficientes, pois os graus de liberdade efetivos (EDF, do inglês *effective degrees of freedom*) para todas as funções suavizadas estimadas está bem abaixo do seu valor máximo possível (k'), e o valor de p para o índice k observado (que mede o padrão nos resíduos) não é significativo (PEDERSEN et al., 2019), exceto para o termo temporal $s(\text{Ano})$, o que não representa grandes problemas. Nesta tabela, cada linha corresponde a um único termo suave, k' corresponde ao número de funções de base usadas para aquele suavizador no modelo ajustado (menor que o k especificado no próprio modelo, pois algumas funções de base são descartadas automaticamente para garantir que o modelo é identificável). A coluna EDF é o número de graus de liberdade efetivos estimado para aquele suavizador, o índice k é uma medida do padrão remanescente nos resíduos e o valor p é calculado com base na distribuição do índice k após randomizar a ordem dos resíduos.

Tabela 17 – Resultados da execução da função `k.check()` no modelo final.

	k'	EDF	índice k	valor p
$s(\text{xcoord}, \text{ycoord})$	304	2,0146	1,0016	0,9950
$s(\text{Ano})$	6	5,0003	0,9084	0,0075
$s(\text{xcoord}, \text{ycoord}, \text{Ano})$	301	126,1439	0,9627	0,5375

As mudanças temporais mostraram uma tendência crescente ao longo dos anos no período de 2010 a 2016 (Figura 30).

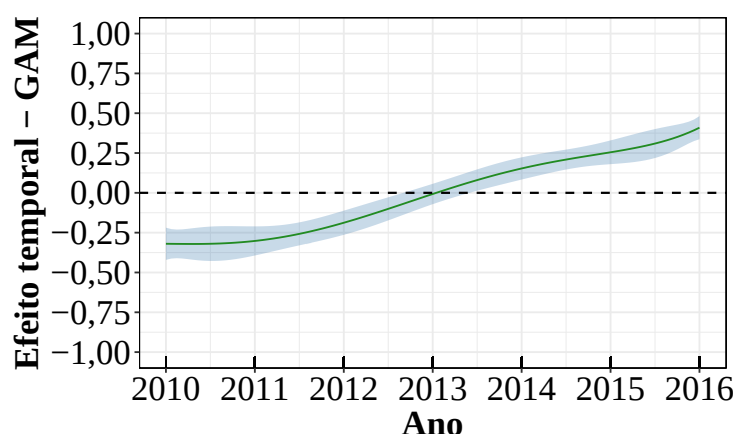


Figura 30 – Estimativas do efeito temporal para SC no período de 2010 a 2016 no município de São Paulo-SP. A linha contínua é o valor predito da média em função do tempo. As linhas tracejadas superior e inferior são o intervalo de confiança de 95%.

A Figura 31 mostra os mapas dos RRs espaço-temporais para SC no período de 2010 a 2016 segundo a AP. As cores mais escuras representam as áreas de maior risco. Nos anos de 2010 e 2011 há um baixo risco relativo geral na maioria das áreas de ponderação, mas posteriormente, o risco relativo se espalhou ao longo do tempo para o resto do município, possivelmente indicando a presença de fatores nesses locais que estão envolvidos na ocorrência de casos de SC.

A Figura 32 mostra os gráficos dos resíduos de deviance do modelo final versus os valores ajustados para cada ano, mostrando que o modelo foi bem ajustado, com valores concentrados entre -2 e 2 e poucos valores acima de 2.

Analizamos a presença de autocorrelação espacial nos resíduos para cada ano. Os resultados são apresentados na Tabela 18. Identificamos autocorrelações espaciais positivas e negativas não significativas para os resíduos no nível de AP de 2010 a 2016. O valor de I variou de $-0,0488$ a $0,0386$. No entanto, o índice de autocorrelação espacial não apresentou tendência crescente ou decrescente, com o menor valor observado em 2016 ($I = -0,0488$; $p = 0,9030$) e maior valor em 2010 ($I = 0,0386$; $p = 0,1168$). Os resultados mostram que em todos os anos, a correlação espacial é não significativa nos resíduos, podendo assumir-se que os mesmos se distribuem aleatoriamente na região do estudo (Figuras 33 e 34).

Tabela 18 – Análise de autocorrelação espacial residual anual para o modelo selecionado.

Ano	Índice de Moran	Valor p
2010	0,0386	0,1168
2011	0,0122	0,3143
2012	-0,0186	0,6680
2013	0,0106	0,3302
2014	0,0370	0,1259
2015	-0,0730	0,9819
2016	-0,0488	0,9030

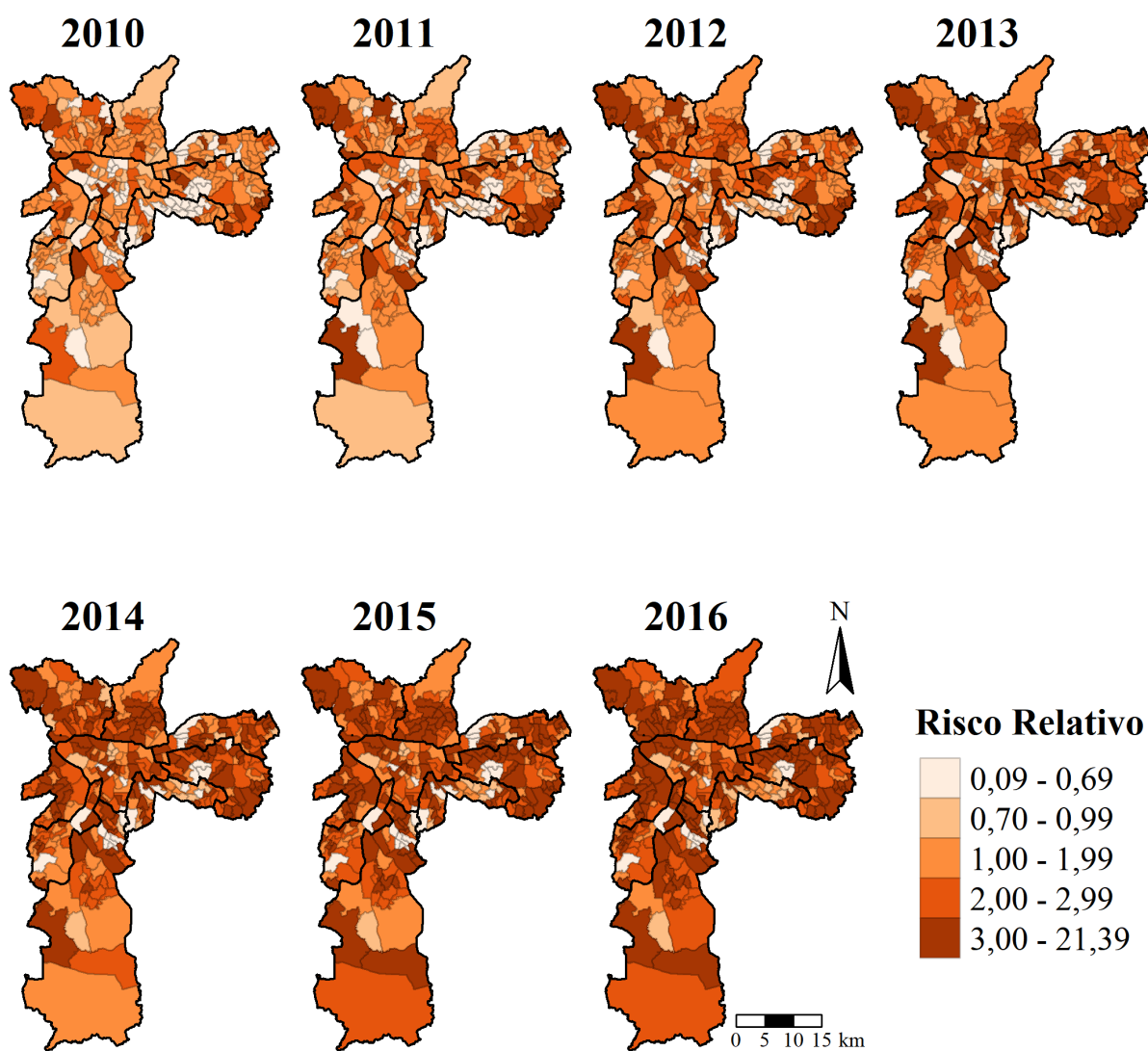


Figura 31 – Estimativas dos riscos relativos espaço-temporais para SC no período de 2010 a 2016 no município de São Paulo-SP.

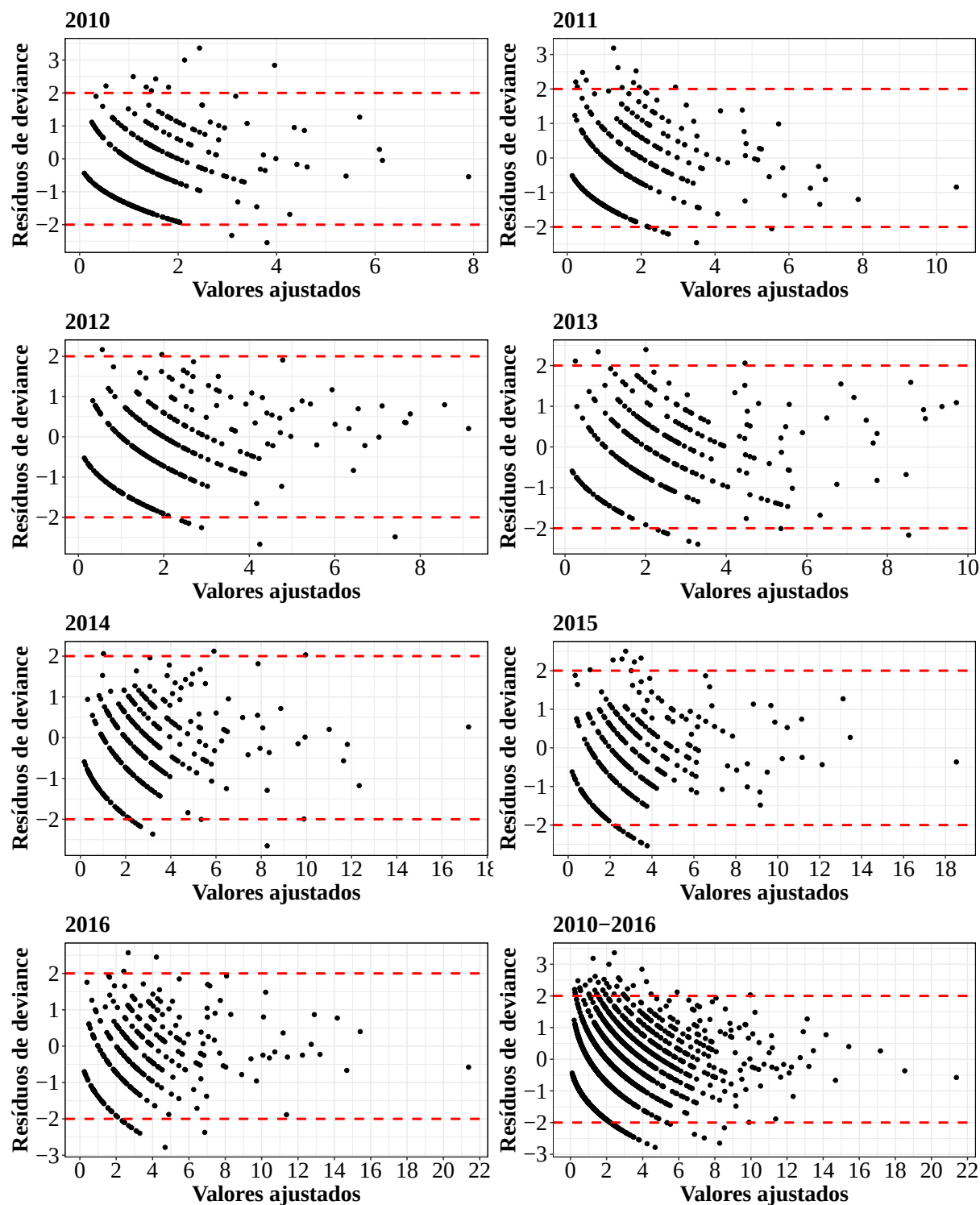


Figura 32 – Gráfico dos resíduos de deviance versus valores ajustados.

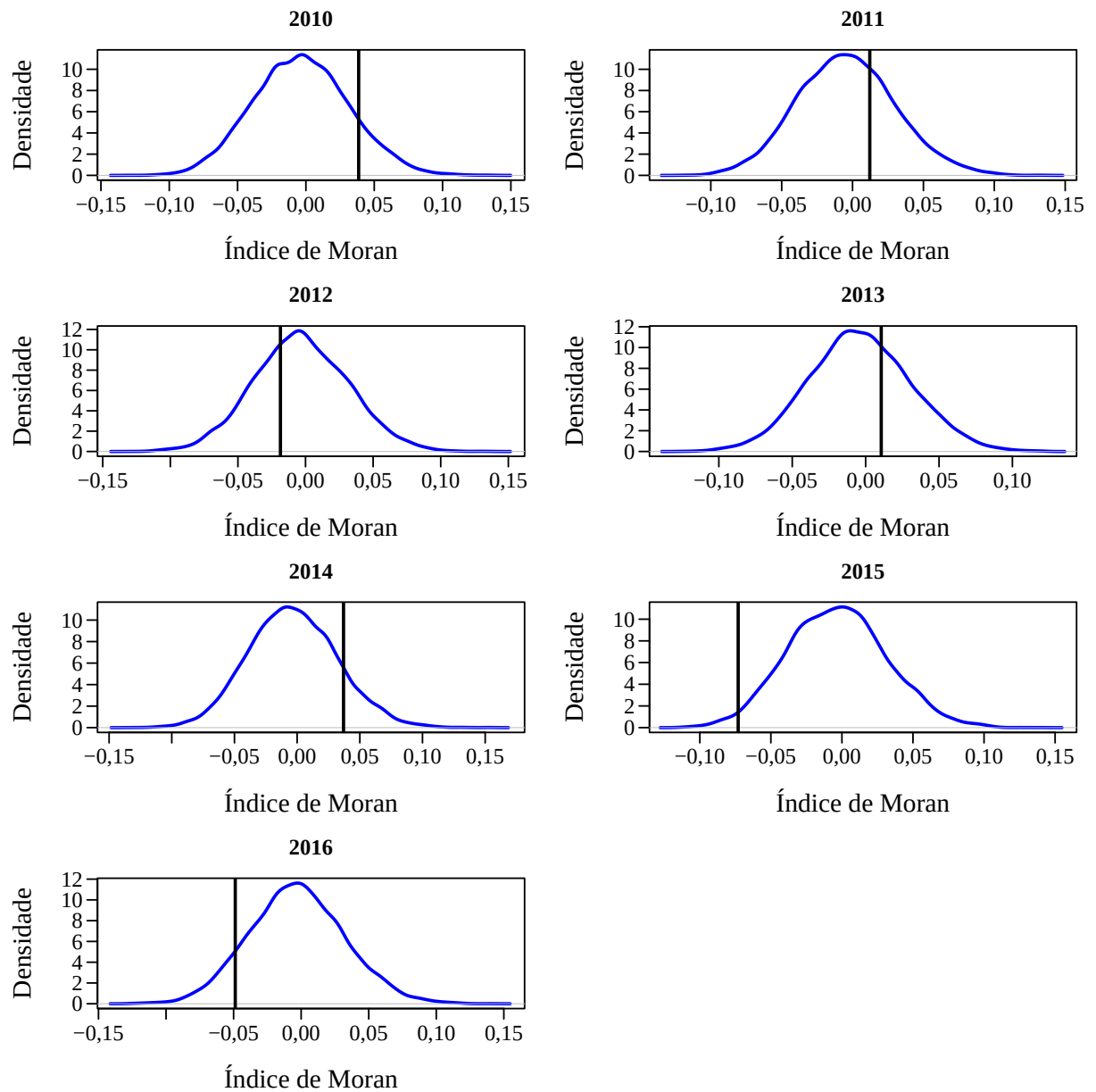


Figura 33 – Distribuição gerada por meio de 10000 simulações para os resíduos do modelo final

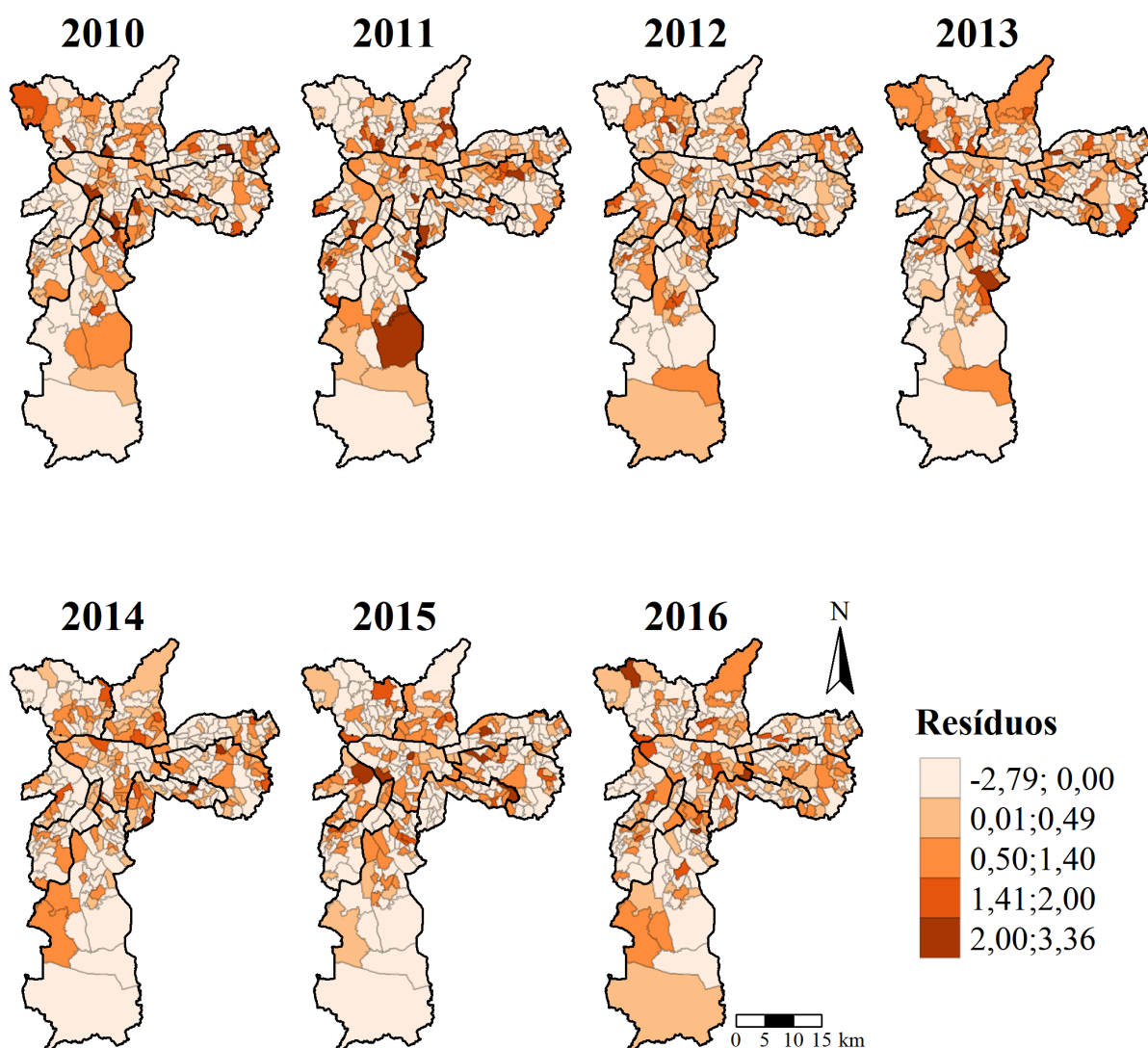


Figura 34 – Resíduos de deviance por áreas de ponderação do município de São Paulo no período de 2010 a 2016.

4.4 Discussão

Este estudo teve como objetivo estimar no espaço e no tempo o risco de SC no município de São Paulo, e avaliar sua associação com as variáveis socioeconômicas e ambientais por meio de uma abordagem espaço-temporal frequentista para dados de contagem em nível de áreas de ponderação no período de 2010 a 2016.

Analizamos e mapeamos o risco relativo de SC considerando as variáveis socioeconômicas e ambientais como fatores de risco. Obtivemos estimativas de risco relativo mais homogêneas do

que razões de incidência padronizadas brutas, suavizando os dados por meio do modelo GAM.

As análises realizadas no município de São Paulo tiveram como meta a identificação da distribuição da SC e as áreas com riscos mais elevados. Segundo [Penna e Ferreira \(2014\)](#) isso permite que estas áreas possam ser priorizadas pelo setor público e que os resultados obtidos tenham impacto na diminuição da ocorrência dos agravos, uma vez que se tratam de áreas de alto risco. Houve concentração de áreas de alto risco nas regiões norte, oeste e noroeste, nas partes leste e oeste da zona central, na periferia da zona leste e grande parte da zona sul.

Por meio dos modelos espaço-temporais, os achados mostraram que houve aumento no risco de sífilis congênita no período de 2010 a 2016. Os resultados dos modelos espaço-temporais forneceram evidências de algumas áreas com maior risco de sífilis congênita que persistiram ao longo do tempo e outras em que houve aumento no risco. Isso sugere que os fatores socioeconômicos e ambientais foram responsáveis por parte da doença nessas regiões, embora haja outros fatores para explicar a incidência da doença. De acordo com [Teixeira et al. \(2018\)](#), o conhecimento de áreas com maiores riscos pode auxiliar na tomada de decisões na saúde, com a implementação de ações já existentes e intervenções para a diminuição destes agravos por meio dos programas de prevenção. O aumento da taxa do risco de SC em algumas regiões do município pode estar associado a melhores condições de acesso e atendimento nas unidades básicas de saúde, consequentemente, melhor acesso aos métodos diagnósticos e maiores taxas de notificações ([CONCEIÇÃO; CÂMARA; PEREIRA, 2020](#)). Segundo [Domingues e Leal \(2016\)](#), a sífilis congênita está estreitamente relacionada com alguns grupos de maior risco, como mulheres muito pobres ou com estilos de vida vulneráveis, sendo que alguns fatores de risco específicos podem variar entre as diversas regiões e ao longo do tempo.

Existem várias limitações que devem ser consideradas em relação aos nossos resultados. Em particular, a modelagem com poucas variáveis que estão relacionadas à sífilis congênita conforme apresentado na literatura de estudos anteriores. Por motivos de confiabilidade, os dados espaciais e temporais para doenças geralmente não são desagregados em determinados níveis e, portanto, uma análise mais precisa dos dados não é possível. Outro fator importante é que não foi encontrado nenhum estudo no setor público de saúde utilizando modelos aditivos generalizados com dados epidemiológicos de sífilis congênita. Isso pode ser do ponto de vista prático, pois os modelos aditivos generalizados são executados mais lentamente quando comparados aos modelos lineares, levando a desvantagens com grandes conjuntos de dados ([VISLOCKY; FRITSCH, 1995](#)). Segundo esses autores, outra preocupação computacional das estruturas GAM é a quantidade de espaço de armazenamento necessária para a informação preditiva.

Neste estudo usamos uma metodologia estatística moderna para avaliar covariáveis para a distribuição espacial e temporal de casos de sífilis congênita no município de São Paulo. A modelagem GAM nos permitiu incluir a interação entre variáveis espaciais e temporais para explicar os efeitos espaço-temporais por meio de *splines de placa fina* ([WOOD, 2017](#)).

O modelo GAM teve um desempenho razoável (representando 60% da variação). Como

ocorre em todos os modelos, o GAM é uma representação matemática de um processo complexo em que as relações estatísticas (dentre e entre áreas ao longo do tempo) são calculadas em média. Os modelos são imperfeitos e suas imprecisões podem ter implicações para os resultados que resumimos neste estudo. No estudo atual, os GAM's explicaram uma porcentagem relativamente aceitável da variação do risco de SC, confirmando sua adequação em abordar tendências não lineares que não podem ser facilmente capturadas por meio de modelos lineares (VENABLES; RIPLEY, 2013)

Nossa análise indicou a presença de uma tendência crescente do risco de sífilis congênita ao longo dos anos e esta descoberta está de acordo com outros estudos que mostram uma tendência semelhante (MEDEIROS et al., 2022).

O estudo conseguiu mostrar a associação que as variáveis tiveram para o aumento dos riscos da SC no Município. As covariáveis proporção de NV de mães adolescentes, proporção de pretos, proporção de mulheres em idade fértil e a proporção de pessoas que vivem em domicílios com banheiro e água encanada não tiveram influências significativas nos casos de SC, enquanto proporção de mulheres de 18 a 24 anos sem instrução ou com ensino fundamental incompleto e índice de Gini mostraram influências significativamente positivas, e a renda média per capita mostrou influência significativamente negativa nos casos de SC.

As covariáveis renda per capita, índice de Gini e proporção de mulheres de 18 a 24 anos sem instrução ou com ensino fundamental incompleto foram fatores importantes no risco de sífilis congênita. Esses resultados corroboram com estudos realizados por (DOMINGUES; LEAL, 2016; RUBERTI, 2020).

É importante destacar que o nível de escolaridade se configura como um fator importante na análise da população estudada, dado que se pode considerar uma associação entre a baixa escolaridade e a dificuldade no acesso aos serviços de saúde e prevenção de doenças. Corroborando com outros estudos, observou-se um aumento no risco em mulheres de 18 a 24 anos com baixa escolaridade, o que se assemelha a uma pesquisa realizada em Santa Cruz no Rio Grande do Norte e em outras regiões (BOTTURA et al., 2019; VASCONCELOS, 2019). A baixa escolaridade materna é um fator importante que pode favorecer o aparecimento de potenciais situações de risco para a mãe e o recém-nascido, pois está associada ao baixo peso ao nascer, à mortalidade infantil e ao aumento do número de partos (XIMENES et al., 2008; HAIDAR; OLIVEIRA; NASCIMENTO, 2001). A associação da sífilis congênita com a escolaridade materna também tem sido observada em outras capitais brasileiras, onde se destaca a baixa escolaridade materna como um dos principais fatores relacionados a esse agravo (MELO; FILHO; FERREIRA, 2011). O nível de escolaridade pode ser considerado um fator importante uma vez que está diretamente relacionado à falta de conhecimento e ao acesso a tratamento e métodos de prevenção, interferindo no entendimento sobre a importância do acompanhamento da saúde durante a gestação, principalmente nas medidas preventivas, dessa forma, prejudicando a interrupção da transmissão (CONCEIÇÃO; CÂMARA; PEREIRA, 2020; HERINGER et al., 2020). A baixa escolaridade

também pode estar associada à taxa de adesão ao tratamento oferecido gratuitamente pelo sistema único de saúde (SUS), uma vez que a gestante pode não ter uma compreensão da gravidade dessa doença para si e para o bebê (GARBIN et al., 2021). Dessa forma, salienta-se que a educação em nosso país requer atenção especial, dado que a escolarização é uma variável importante, pois facilita a compreensão de informações, o acesso ao mercado de trabalho formal, aos serviços de saúde e, conseqüentemente, a melhoria das condições de vida e de saúde da população.

Foi possível evidenciar a questão da desigualdade mostrada por meio do Índice de Gini e renda média per capita. Após a infecção, as gestantes com poder aquisitivo muito baixo podem ter menor chance de receber tratamento adequado, devido à falta de informação e dificuldades de acesso ao sistema de saúde (LAGO et al., 2004; MACÊDO et al., 2017; NOGUEIRA et al., 2022).

É importante mencionar que a distribuição espaço-temporal de casos de sífilis congênita não é apenas controlado pelos fatores socioeconômicos e ambientais considerados neste trabalho. Estudos futuros devem incluir outros preditores que considerem número de consultas pré-natal, tratamento do parceiro, entre outras, para melhorar os resultados do modelo e fornecer maiores percepções sobre a sífilis congênita e os fatores ecológicos que influenciam a distribuição da doença no município. Vários estudos têm mostrado que o tratamento adequado e o tratamento de parceiros continuam sendo fatores que contribuem para a reinfecção e, conseqüentemente, a dificuldade de prevenção da transmissão e demais consequências da doença (LINO; SOUSA; BATISTA, 2021). Galatoire, Rosso e Sakae (2012) e Benito e Souza (2016) mostram que a falta de tratamento ou o tratamento inadequado do parceiro poderiam explicar a transmissão vertical de sífilis. No estudo de Ruberti (2020), a proporção de mães com pré-natal inadequado, proporção de mães com menos de quatro consultas e proporção de nascidos vivos de mães adolescentes foram fatores associados à sífilis congênita. De acordo com Wijesooriya et al. (2016), pré-natais de qualidade, incluindo teste e tratamento de sífilis, são importantes para alcançar a eliminação da transmissão da sífilis de mãe para filho. O tratamento inadequado das gestantes e parceiros é o ponto chave na ocorrência da sífilis congênita, sendo uma realidade carente de foco dos órgãos públicos e que não foi aqui considerado. Segundo um estudo realizado na China por Wan et al. (2020), quando o tratamento é inadequado o risco de sífilis congênita é 3,63 vezes maior. Entretanto, por meio de diagnóstico e tratamento adequado durante o acompanhamento pré-natal, a transmissão vertical pode ser reduzida em até 97% (DOMINGUES; LEAL, 2016).

O Brasil adotou uma estratégia para a redução das taxas de mortalidade materno-infantil denominada de Rede “Cegonha”, que tem como um de seus objetivos melhorar a qualidade da assistência prestada à mulher e à criança, facilitando o acesso ao diagnóstico e ao tratamento precoce e correto e, conseqüentemente, a redução dos casos de sífilis congênita no país (CARDOSO et al., 2018).

Além disso, embora reconhecidos os progressos quanto a implantação e adesão às condutas recomendadas pela ciência para diagnóstico e prevenção da transmissão vertical da sífilis,

ainda existem lacunas importantes na rede assistencial com a ausência de questões estruturais básicas que permitam às equipes de saúde a prestação de um cuidado adequado. Para tal, o fortalecimento nos municípios da rede de assistência à saúde materno-infantil, com aumento da testagem das gestantes na atenção primária e oferta e disponibilidade de tais ferramentas de cuidado contribuirá para ampliação do diagnóstico precoce da sífilis, permitindo o tratamento oportuno durante a gestação.

Embora um sistema integrado de vigilância da sífilis tenha sido proposto pela OMS como uma ferramenta útil em todo o mundo, ainda falta um conjunto de dados georreferenciados bem estabelecido para estudos futuros da evolução da doença no município de São Paulo. Finalmente, são necessárias mais publicações científicas sobre a sífilis congênita para estimular estudos e programas efetivos de controle e prevenção para o bem-estar da população. Segundo Conway (2007), somente por meio de um esforço unificado e bem financiado na política nacional de saúde que incorpore o suporte necessário para a erradicação da sífilis congênita, o fardo dessa doença será retirado de mulheres e crianças no mundo. Apenas quando isso ocorrer, começaremos a cumprir as recomendações da OMS, melhorando assim a saúde de mulheres e crianças.

4.5 Conclusão

Neste trabalho ajustamos modelos espaço-temporais usando a metodologia GAM para analisar a distribuição da incidência de sífilis congênita e mapear seus padrões espaço-temporais no município de São Paulo em sete pontos do tempo de 2010 a 2016. Observou-se a existência de áreas de alto risco para casos de sífilis congênita, que persistiram por vários anos. Os resultados fornecem uma visão para ajudar os governos locais a otimizar o planejamento assistencial à saúde e implementar regulamentos específicos melhorando assim os cuidados necessários para o controle da doença no município. As pesquisas e esforços dos órgãos públicos devem continuar a fim de controlar, tratar e, eventualmente, erradicar a sífilis congênita em todo o mundo. O impacto na saúde pública da sífilis na gravidez e na infância continua a ser significativo e a eliminação da transmissão vertical somente se tornará uma realidade por meio de serviços de saúde de alta qualidade.

Verificou-se que as covariáveis, renda per capita, índice de Gini e proporção de mulheres de 18 a 24 anos sem instrução ou com ensino fundamental incompleto foram fatores importantes no risco de sífilis congênita. Sua eliminação representa um grande desafio para o município. Assim, são necessárias intervenções para a eliminação da sífilis congênita e estas devem ser direcionadas principalmente para áreas que apresentam um risco elevado e persistente de casos.

Até onde sabemos, apesar de ser inexplorado para o mapeamento do risco da sífilis congênita, o GAM é uma alternativa promissora: é de fácil ajuste, computacionalmente eficiente e pode gerar previsões espaciais de qualidade semelhante às geradas pelo INLA. Devemos ser cautelosos no uso do INLA, pois a especificação incorreta da estrutura de covariância pode

produzir uma predição pobre.

O INLA usa a aproximação analítica de Laplace e esquemas de integração numérica eficientes para obter uma aproximação analítica altamente precisa de quantidades posteriores de interesse com tempos de computação relativamente curtos. No entanto, pode ser um desafio saber quando tais aproximações são adequadas. Por essas razões, o modelo aditivo generalizado espaço-temporal, especificando espaço-tempo como produto tensorial mais suavizado, oferece vantagem computacional e também flexibilidade significativa para modelar a interação espaço-tempo. Além disso, os procedimentos de ajuste para modelos GAM são mais diretos do que o INLA usando os pacotes em R.

Como trabalho futuro, uma análise em nível individual poderia fornecer melhores ajustes para os dados do que a análise baseada em dados agregados. No entanto, esses dados podem não ser fáceis de serem disponibilizados pelas unidades de saúde.

5 Comparação entre as abordagens INLA e GAM

Neste capítulo, comparamos a metodologia INLA (RUE; MARTINO; CHOPIN, 2009) com a metodologia GAM (HASTIE; TIBSHIRANI, 1995) por meio dos pacotes R-INLA (RUE et al., 2013) e mgcv (WOOD, 2001). Aqui, estamos particularmente interessados em comparar os modelos apresentados nos Capítulos 3 e 4 quando as estruturas espaciais e temporais estão presentes.

Na abordagem GAM, os efeitos espaciais e temporais foram controlados por funções suavizadas das coordenadas geográficas e dos anos pertencentes aos dados. Ajustamos o mesmo modelo pela abordagem INLA para possibilitar comparações dos resultados. O resumo desta análise é apresentado na Tabela 19 e nas Figuras 35 e 36 por meio de médias a posterioris e intervalos de credibilidade para a abordagem INLA e estimativas pontuais com intervalos de confiança para a abordagem GAM, ambos com confiança de 95%.

Tabela 19 – Estimativas dos modelos para cada covariável na presença dos efeitos espaciais e temporais, considerando as abordagens INLA e GAM.

Variável	Média Post.	Int. Cred.	Estimativa	Int. Conf.
Intercepto	-0,280	(-0,862; 0,306)	-0,226	(-0,263; -0,189)
MSIF1824	0,120	(0,035; 0,206)	0,087	(0,026; 0,148)
PFERTIL	-0,091	(-0,175; -0,007)	-0,011	(-0,087; 0,065)
GINI	0,118	(0,034; 0,203)	0,085	(0,018; 0,152)
RDPC	-0,353	(-0,472; -0,236)	-0,315	(-0,470; -0,160)

Pela abordagem GAM a interpretação dos intervalos de confiança indica que as covariáveis MSIF1824, GINI e RDPC foram significativas. Nota-se que destas quatro covariáveis, duas delas apresentam uma das extremidades do intervalo estimado muito próximo do zero, o que indicaria não significância do efeito para intervalos ligeiramente mais conservadores.

A Tabela 19 mostra também que pela abordagem INLA as covariáveis consideradas no modelo apresentam significância, dado que, os intervalos de credibilidade baseados de 95% não incluem o zero, contradizendo os resultados do modelo GAM, no qual a variável PFERTIL não foi significativa.

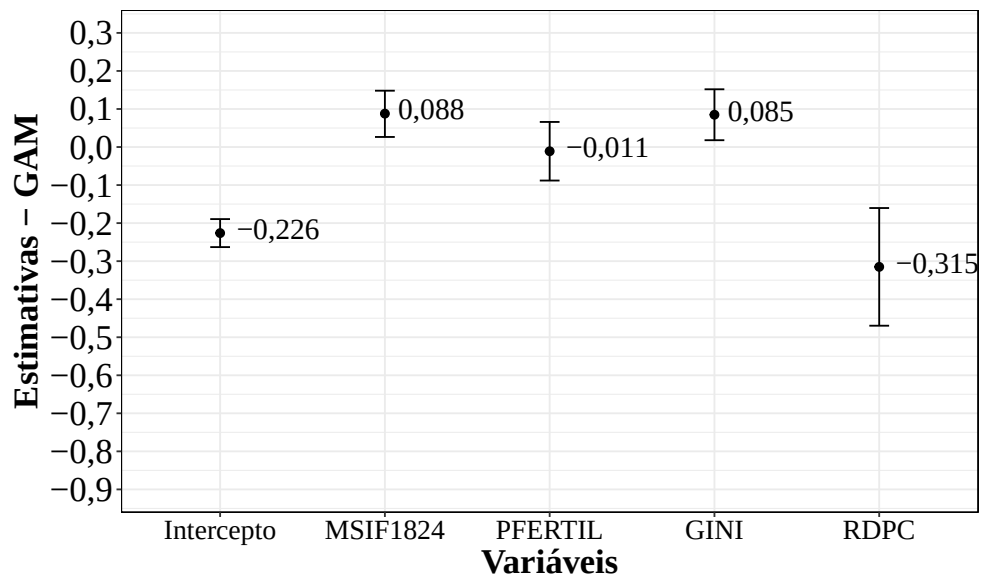


Figura 35 – Intervalos de confiança de 95% para estimativas dos efeitos fixos usando a abordagem GAM.

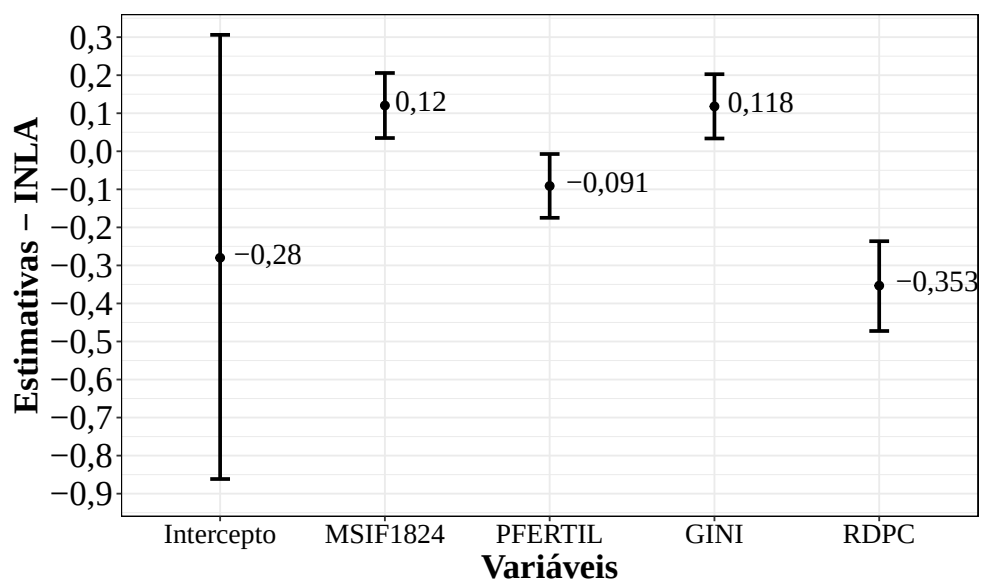


Figura 36 – Intervalos de credibilidade de 95% para estimativas dos efeitos fixos usando a abordagem INLA.

Pensando que, pela abordagem INLA tem-se uma distribuição a priori não informativa normal para os efeitos fixos com média zero e variância alta, à covariável tem que trazer muito mais informação para ser considerada significativa do que pela abordagem GAM, em que não se tem priori, e isto pode ser uma explicação para as diferenças entre os resultados. Outra

diferença importante entre os modelos é o modo como o efeito espaço-temporal é considerado no modelo. O INLA trabalha com propriedades de independência condicional, decodificada de forma conveniente por meio de uma matriz de precisão na distribuição a priori para os efeitos estruturados o que torna este um efeito aleatório. Já no GAM estes efeitos são suavizados por uma *spline*, o que impede que mudanças bruscas ocorram na função estimada dos efeitos espaciais e temporais. Isto faz com que uma porção menor da variabilidade da resposta possa ser atribuída à função suavizada do que a um processo estocástico subjacente, e consequentemente ao menos parcialmente captada pelo intercepto e possivelmente pelas covariáveis.

O modelo aditivo leva em consideração o efeito temporal como uma função suavizada dos anos da coleta, enquanto pela abordagem INLA tal efeito é considerado por meio de um modelo autorregressivo de primeira ordem.

Conforme mostrado nas Figuras 37 e 38, a diferença no crescimento do efeito temporal é muito pequena entre as abordagens INLA e GAM. Apesar de esperar que o GAM tende a suavizar mais o efeito, porém neste caso a suavização comparada com o efeito estimado pelo modelo autorregressivo de ordem 1 é muito pequena. Pelo modelo autorregressivo de ordem 1 o efeito temporal apresenta valores que variam de -0.311 a 0.342, enquanto que na *spline* ajustada pelo GAM este efeito varia de -0.375 a 0.560, uma diferença relativamente baixa. Nota-se também que as bandas de credibilidade obtidas pelo INLA são mais largas que as bandas de confiança obtidas pelo GAM. As bandas de credibilidade do INLA são calculadas por meio de quantis e o intervalo de confiança do GAM é baseado na teoria assintótica usando os quantis da distribuição normal.

Neste ponto, fazemos um alerta. Quando analisamos interações espaço-temporais, embora não haja estudos específicos na literatura, até onde sabemos, as diferenças entre as duas abordagens não deveria ser tão diferentes. A quantidade de informação no tempo é ainda escassa. Nos modelos autorregressivos, as estimativas estruturais se comportam bem quando o tempo é longo, ou seja, são mais confiáveis do ponto de vista da consistência estatística. Neste trabalho, a informação espacial é abundante, mas não do tempo, o que pode ter causado um impacto na precisão dos intervalos de credibilidade. Quanto menos informação se tem no tempo, mais o modelo capta a estimação da parte linear, e pode-se ter algum problema com a precisão dos fatores de risco, além da informação da tendência temporal. Por isso, é importante considerar um período maior, mas há uma limitação no fornecimento dos dados.

Pela figura 39 fica claro que o modelo que acompanha melhor as observações é o estimado pela abordagem INLA. Desta análise o que pode-se concluir é que modelar os efeitos espaciais e temporais por meio de *splines* e ainda olhar seu intervalo de confiança baseado na distribuição assintótica pode ser muito enganoso.

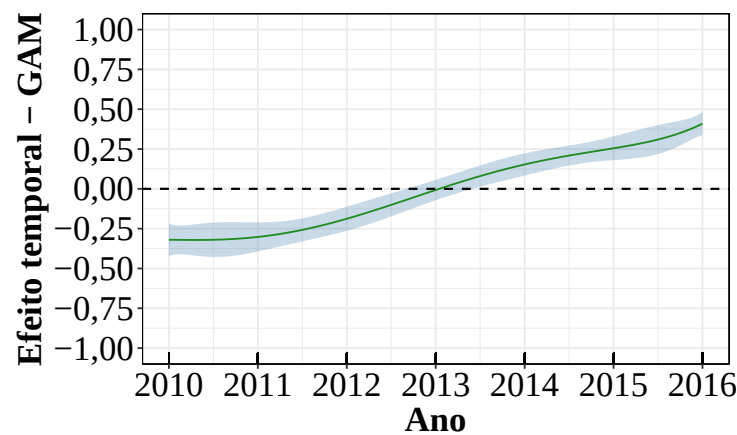


Figura 37 – Estimativas do efeito temporal para SC no período de 2010 a 2016 no município de São Paulo-SP. A linha contínua é o valor predito da média em função do tempo. As faixas azuis superior e inferior são o intervalo de confiança de 95%.

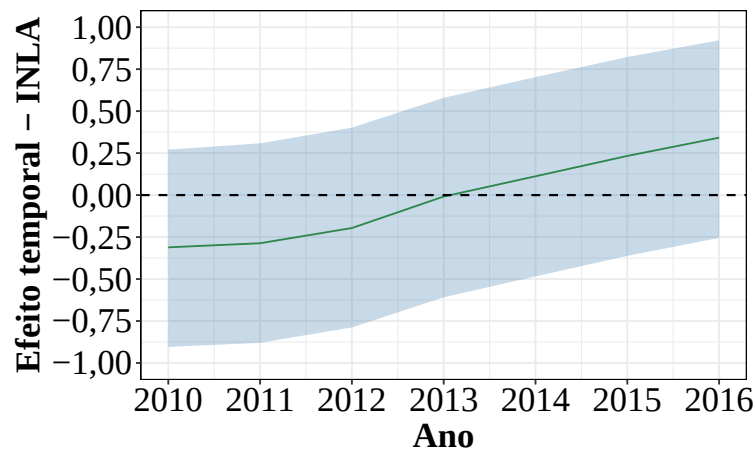


Figura 38 – Estimativas do efeito temporal para SC no período de 2010 a 2016 no município de São Paulo-SP. A linha contínua é o valor predito da média em função do tempo. As faixas azuis superior e inferior são o intervalo de credibilidade de 95%.

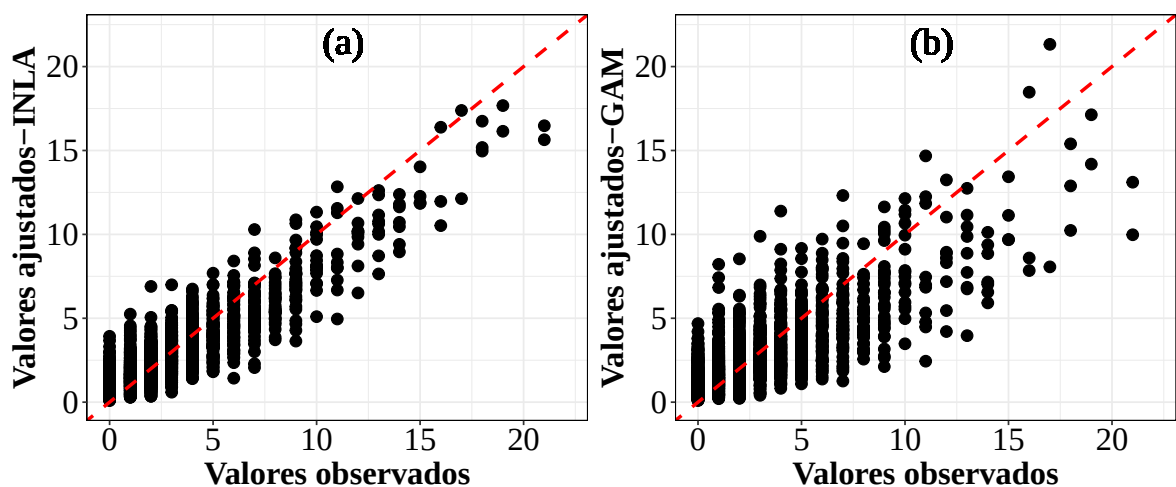


Figura 39 – (a) Valores ajustados versus valores observados dos casos notificados de sífilis congênita no município de São Paulo de 2010 a 2016 segundo o modelo INLA. (b) Valores ajustados versus valores segundo o modelo GAM.

A figura 40 apresenta um diagrama de dispersão entre o efeito espaço-temporal obtido pelas abordagens INLA e GAM.

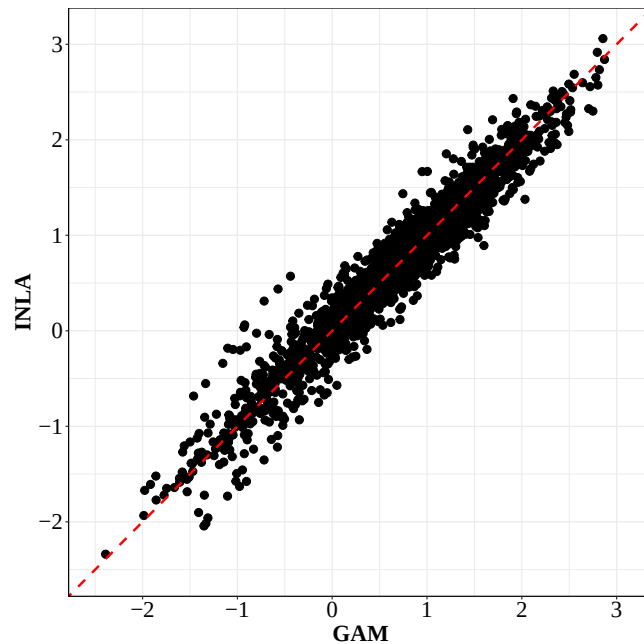


Figura 40 – Diagrama de dispersão entre o efeito espaço-temporal obtido pelas abordagens INLA e GAM.

Em geral, apresenta-se uma boa correspondência nas estimativas entre esses dois modelos, com coeficientes de correlação para efeitos estimados pelos modelos INLA e GAM sendo 0.790 e 0.720, respectivamente (Tabela 20).

Por fim comparamos a capacidade preditiva das duas abordagens no sentido de verificar qual acompanha melhor o conjunto de dados. Para isso foram calculadas algumas medidas de similaridade: erro quadrático médio, erro absoluto médio e correlação entre valores observados e preditos. Os resultados são apresentados na Tabela 20 e mostram que a abordagem INLA apresenta melhores resultados em todas as medidas.

Tabela 20 – Medidas de concordância entre os modelos obtidos pelas abordagens INLA e GAM e os dados observados

Abordagem	RMSE	MAE	Correlação
INLA	1,319834	1,019183	0,7897897
GAM	1,807202	1,316113	0,7202871

6 Considerações finais

O objetivo geral desta tese foi avaliar o desempenho de diferentes modelos espaciais e espaço-temporais Bayesianos e frequentista com o intuito de fornecer orientações sobre o uso dessas abordagens de modelagem para pequenas áreas, com foco na compreensão da sífilis congênita no município de São Paulo. A contribuição geral desta pesquisa foi no aprofundamento do conhecimento em modelagem espacial e espaço-temporal por meio de diferentes componentes de suavização espacial e temporal em termos de metodologia e aplicação. Espera-se que os resultados deste estudo sejam uma contribuição para pesquisadores em estatística Bayesiana e epidemiologia espacial, bem como outras áreas que envolvam modelagem espacial e espaço-temporal.

Este trabalho foi sustentado por uma revisão da literatura dos principais tópicos abordados na tese e aplicação de abordagens espaciais e espaço-temporais Bayesianas para modelar a sífilis congênita. Foram considerados modelos Bayesianos com um efeito aleatório espacialmente estruturado usando prioris CAR com especificações BYM, BYM2 e Leroux.

Os modelos propostos foram nossa melhor opção ao lidar com dados agregados em nível de áreas de ponderação. A partir do conjunto de dados pouco detalhado e ausência de variáveis fortemente associadas à sífilis congênita presentes em vários estudos da literatura, nosso estudo ainda mostra as vantagens de aplicação dos modelos CAR propostos.

A família dos modelos gaussianos latentes é adaptável para uma grande quantidade de situações de análise de dados complexos, como as apresentadas no Capítulo 3. Destaca-se a situação de dados espaço-temporais que é possivelmente a mais complexa entre os atuais modelos estatísticos da literatura. A modelagem aditiva estruturada permite que diferentes tipos de covariáveis sejam adicionadas em modelos clássicos de maneira aditiva, emprestando força dos modelos paramétricos e não paramétricos. Em particular, investigamos efeitos espaço-temporais com aplicações em um conjunto de dados de sífilis congênita usando a aproximação integrada de Laplace aninhada (INLA) e a abordagem GAM. Ambas as abordagens foram comparadas para verificar a concordância entre elas, principalmente no que diz respeito ao modo como cada uma capta os efeitos espaciais e temporais.

Além disso, modelos Bayesianos para mapeamento de doenças conjuntas podem ser considerados em pesquisas futuras, com intuito investigar simultaneamente os determinantes

da sífilis congênita com outras doenças que podem influenciar na saúde, como HIV e sífilis em gestantes. Alguns tipos de comparações com a metodologia de inferência baseada em modelos mistos também podem ser investigadas. Os resultados podem ser usados no planejamento regional de saúde, vigilância e intervenção da sífilis congênita, e também na alocação de recursos para a saúde.

Referências

- AKAIKE, H. Theory and an extension of the maximum likelihood principal. In: *International symposium on information theory. Budapest, Hungary: Akademiai Kiado.* [S.l.: s.n.], 1973. 96
- ALLÉVIUS, B. On the precision matrix of an irregularly sampled AR(1) process. *arXiv preprint arXiv:1801.03791*, 2018. 27
- ALMEIDA, E. Econometria espacial. *Campinas–SP. Alínea*, 2012. 31
- AMARASINGHE, G. K. et al. Taxonomy of the order mononegavirales: update 2018. *Archives of virology*, Springer, v. 163, n. 8, p. 2283–2294, 2018. 25
- ANSELIN, L. Exploring spatial data with GeoDaTM: a workbook. *Center for spatially integrated social science*, 2005. 77
- ASSUNCAO, R. M.; REIS, I. A.; OLIVEIRA, C. D. L. Diffusion and prediction of Leishmaniasis in a large metropolitan area in Brazil with a Bayesian space–time model. *Statistics in Medicine*, Wiley Online Library, v. 20, n. 15, p. 2319–2335, 2001. 1
- ATEBA, F. F. et al. Spatio-temporal dynamic of malaria incidence: A comparison of two ecological zones in Mali. *International journal of environmental research and public health*, Multidisciplinary Digital Publishing Institute, v. 17, n. 13, p. 4698, 2020. 92
- AUSTIN, M. P. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological modelling*, Elsevier, v. 157, n. 2-3, p. 101–118, 2002. 92
- AVELLEIRA, J. C. R.; BOTTINO, G. Sífilis: diagnóstico, tratamento e controle. *Anais brasileiros de dermatologia*, SciELO Brasil, v. 81, p. 111–126, 2006. 4, 46
- BAKSALARY, O. M.; TRENKLER, G. The Moore–Penrose inverse: a hundred years on a frontline of physics research. *The European Physical Journal H*, Springer, v. 46, n. 1, p. 1–10, 2021. 35
- BANERJEE, S.; CARLIN, B. P.; GELFAND, A. E. *Hierarchical modeling and analysis for spatial data*. 2. ed. [S.l.]: Chapman and Hall/CRC, 2015. 8, 21, 25, 30, 48
- BANERJEE, S.; ROY, A. *Linear algebra and matrix analysis for statistics*. [S.l.]: Crc Press Boca Raton, FL, USA:, 2014. v. 181. 140
- BARANKIN, E. W.; MAITRA, A. P. Generalization of the Fisher-Darmois-Koopman-Pitman theorem on sufficient statistics. *Sankhyā: The Indian Journal of Statistics, Series A*, JSTOR, p. 217–244, 1963. 13
- BARNETT, R. Syphilis. *The Lancet*, v. 391, n. 10129, p. 1471, 2018. ISSN 0140-6736. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S014067361830833X>>. 3
- BARROZO, L. V. et al. Geoses: A socioeconomic index for health and social research in Brazil. *PloS one*, Public Library of Science San Francisco, CA USA, v. 15, n. 4, p. e0232074, 2020. 75

- BAYES, T. Lii. an essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, frs communicated by mr. price, in a letter to john canton, amfr s. *Philosophical transactions of the Royal Society of London*, The Royal Society London, n. 53, p. 370–418, 1763. [19](#)
- BELLHOUSE, D. R. The reverend Thomas Bayes, FRS: a biography to celebrate the tercentenary of his birth. *Statistical Science*, Institute of Mathematical Statistics, v. 19, n. 1, p. 3–43, 2004. [19](#)
- BENITO, L. A. O.; SOUZA, W. N. de. Perfil epidemiológico da sífilis congênita no Brasil no período de 2008 a 2014. *Universitas: Ciências da Saúde*, v. 14, n. 2, 2016. [85](#), [112](#)
- BERLINER, L. M. Hierarchical Bayesian time series models. In: *Maximum entropy and Bayesian methods*. [S.l.]: Springer, 1996. p. 15–22. [32](#)
- BERNARDINELLI, L.; CLAYTON, D.; MONTOMOLI, C. Bayesian estimates of disease maps: how important are priors? *Statistics in medicine*, Wiley Online Library, v. 14, n. 21-22, p. 2411–2431, 1995. [56](#)
- BERNARDINELLI, L. et al. Bayesian analysis of space—time variation in disease risk. *Statistics in medicine*, Wiley Online Library, v. 14, n. 21-22, p. 2433–2443, 1995. [1](#), [36](#), [37](#), [47](#)
- BESAG, J. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 36, n. 2, p. 192–225, 1974. [30](#), [48](#)
- BESAG, J. Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society: Series D (The Statistician)*, Wiley Online Library, v. 24, n. 3, p. 179–195, 1975. [30](#)
- BESAG, J. et al. Bayesian computation and stochastic systems. *Statistical science*, JSTOR, p. 3–41, 1995. [35](#)
- BESAG, J.; KOOPERBERG, C. On conditional and intrinsic autoregressions. *Biometrika*, Oxford University Press, v. 82, n. 4, p. 733–746, 1995. [31](#), [34](#)
- BESAG, J.; YORK, J.; MOLLIE, A. Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, Springer, v. 43, n. 1, p. 1–20, 1991. [1](#), [7](#), [10](#), [30](#), [34](#), [35](#)
- BEST, N.; RICHARDSON, S.; THOMSON, A. A comparison of Bayesian spatial models for disease mapping. *Statistical methods in medical research*, Sage Publications Sage CA: Thousand Oaks, CA, v. 14, n. 1, p. 35–59, 2005. [8](#), [48](#)
- BHUNIA, G. S.; SHIT, P. K. *Geospatial analysis of public health*. [S.l.]: Springer, 2019. [11](#)
- BISANZIO, D. et al. Spatio-temporal patterns of distribution of West Nile virus vectors in eastern Piedmont Region, Italy. *Parasites & vectors*, BioMed Central, v. 4, n. 1, p. 1–11, 2011. [48](#)
- BLANGIARDO, M.; CAMELETTI, M. *Spatial and spatio-temporal Bayesian models with R-INLA*. [S.l.]: John Wiley & Sons, 2015. [8](#), [22](#), [37](#), [48](#), [54](#), [57](#), [98](#)
- BLANGIARDO, M. et al. Spatial and spatio-temporal models with R-INLA. *Spatial and spatio-temporal epidemiology*, Elsevier, v. 4, p. 33–49, 2013. [8](#), [48](#)

- BOLFARINE, H.; SANDOVAL, M. C. *Introdução à inferência estatística*. [S.l.]: SBM, 2001. v. 2. 19, 33
- BONAT, W. H. Aplicações de inferência Bayesiana aproximada para modelos gaussianos latentes espaço temporais. 2010. 23
- BOTTURA, B. R. et al. Perfil epidemiológico da sífilis gestacional e congênita no Brasil—período de 2007 a 2016. *Arquivos Médicos dos Hospitais e da Faculdade de Ciências Médicas da Santa Casa de São Paulo*, p. 69–75, 2019. 86, 111
- BOX, G. E. et al. *Time series analysis: forecasting and control*. [S.l.]: John Wiley & Sons, 2015. 25, 26
- BOX, G. E.; TIAO, G. C. *Bayesian inference in statistical analysis*. [S.l.]: John Wiley & Sons, 2011. v. 40. 140
- BRASIL. *Diretrizes para o controle da sífilis congênita*. Brasília: Ministério da Saúde Brasília, 2005. Ministério da Saúde e Secretaria de Vigilância em Saúde do Programa Nacional de DST E AIDS. Disponível em: <https://bvsmms.saude.gov.br/bvs/publicacoes/diretrizes_controle_sifilis_congenita.pdf>. 4, 46
- BRASIL. *Caderno de boas práticas: o uso da penicilina na Atenção Básica para a prevenção da sífilis congênita no Brasil*. Brasília: Ministério da Saúde Brasília (DF), 2015. Brasil. Ministério da Saúde. Secretaria de Vigilância em Saúde. Departamento de DST, Aids e Hepatites Virais. 88
- BRASIL. *Protocolo Clínico e Diretrizes Terapêuticas para Atenção Integral às Pessoas com Infecções Sexualmente Transmissíveis*. Brasília: Ministério da Saúde, 2015. Secretaria de Vigilância em Saúde. Disponível em: <https://bvsmms.saude.gov.br/bvs/publicacoes/protocolo_clinico_diretrizes_terapeutica_atencao_integral_pessoas_infecoes_sexualmente_transmissiveis.pdf>. 2
- BRASIL. *Boletim Epidemiológico de Sífilis - Ano VI – nº 01*. Brasília: Ministério da Saúde, 2020. Secretaria de Vigilância em Saúde do Programa Nacional de DST E AIDS. Disponível em: <<http://www.aids.gov.br/pt-br/pub/2020/boletim-sifilis-2020>>. 5, 6, 88
- BRASIL. *Secretaria de Vigilância em Saúde. Departamento de Doenças de Condições Crônicas e Infecções Sexualmente Transmissíveis. Protocolo clínico e diretrizes terapêuticas para atenção integral às pessoas com infecções sexualmente transmissíveis (IST)*. [S.l.]: Ministério da Saúde Brasília, 2020. 3
- BRASIL. *SINAN: Sistema de Informação de Agravos de Notificação*. [S.l.]: Ministério da Saúde Brasília, 2020. 2
- BRASIL. *SINASC: Sistema de Informações sobre Nascidos Vivos*. [S.l.]: Ministério da Saúde Brasília, 2020. 2
- BRUNNER, L. S.; SUDDARTH, D. S. Tratado de enfermagem médico-cirúrgica. In: . [S.l.]: Guanabara Koogan: Rio de Janeiro, 1987. 4, 46
- CAMELETTI, M. et al. Spatio-temporal modeling of particulate matter concentration through the SPDE approach. *AStA Advances in Statistical Analysis*, Springer, v. 97, n. 2, p. 109–131, 2013. 48

- CARDOSO, A. R. P. et al. Analysis of cases of gestational and congenital syphilis between 2008 and 2010 in Fortaleza, State of Ceará, Brazil. *Ciência & Saúde Coletiva*, SciELO Brasil, v. 23, p. 563–574, 2018. 112
- CARLIN, B. P.; LOUIS, T. A. *Bayesian methods for data analysis*. [S.l.]: CRC Press, 2008. 31
- CARMO, M. P. Daniel do et al. A prevalência da sífilis congênita no BRASIL: uma breve revisão. *Saúde & Ciência em Ação*, v. 3, n. 1, p. 1–10, 2017. 5
- CARROLL, R. et al. Comparing INLA and OpenBUGS for hierarchical Poisson modeling in disease mapping. *Spatial and spatio-temporal epidemiology*, Elsevier, v. 14, p. 45–54, 2015. 8, 48, 84
- CHAI, T.; DRAXLER, R. R. Root mean square error (RMSE) or mean absolute error (MAE)?—arguments against avoiding rmse in the literature. *Geoscientific model development*, Copernicus GmbH, v. 7, n. 3, p. 1247–1250, 2014. 55, 97
- CHE-HIM, N. et al. Spatio-temporal modelling of dengue fever incidence in Malaysia. In: IOP PUBLISHING. *Journal of Physics: Conference Series*. [S.l.], 2018. v. 995, n. 1, p. 012003. 92
- CHIARAVALLOTI-NETO, F. O geoprocessamento e saúde pública. *Arquivos de Ciências da Saúde*, v. 23, n. 4, p. 01–02, 2017. ISSN 2318-3691. Disponível em: <<http://www.cienciasdasaude.famerp.br/index.php/racs/article/view/661>>. 22
- CONCEIÇÃO, H. N. d.; CÂMARA, J. T.; PEREIRA, B. M. Análise epidemiológica e espacial dos casos de sífilis gestacional e congênita. *Saúde em debate*, SciELO Brasil, v. 43, p. 1145–1158, 2020. 84, 86, 110, 111
- CONGDON, P. Monitoring suicide mortality: a Bayesian approach. *European Journal of Population/Revue européenne de Démographie*, Springer, v. 16, n. 3, p. 251–284, 2000. 7
- CONGDON, P. *Applied Bayesian modelling*. [S.l.]: John Wiley & Sons, 2014. v. 595. 20
- CONTROL, C. for D.; PREVENTION. *Congenital Syphilis - CDC Fact Sheet*. [S.l.], 2021. 3
- CONWAY, J. H. *Recognizing and reducing the global burden of congenital syphilis: the time is now*. [S.l.]: LWW, 2007. S2–S4 p. 113
- COOPER, J. M.; SÁNCHEZ, P. J. Congenital syphilis. In: ELSEVIER. *Seminars in perinatology*. [S.l.], 2018. v. 42, n. 3, p. 176–184. 4, 47
- CORPAS-BURGOS, F.; MARTINEZ-BENEITO, M. A. On the use of adaptive spatial weight matrices from disease mapping multivariate analyses. *Stochastic Environmental Research and Risk Assessment*, Springer, v. 34, n. 3, p. 531–544, 2020. 1
- CRAWLEY, M. J. *The R book*. [S.l.]: John Wiley & Sons, 2012. 13
- CRESSIE, N. A. *Statistics for spatial data revised edition*. [S.l.]: Wiley-Interscience, 1993. 21, 25, 30, 31, 95
- CZADO, C.; GNEITING, T.; HELD, L. Predictive model assessment for count data. *Biometrics*, Wiley Online Library, v. 65, n. 4, p. 1254–1261, 2009. 56
- DAUM, F. E. The Fisher-Farmonis-Koopman-Pitman theorem for random processes. In: IEEE. *1986 25th IEEE Conference on Decision and Control*. [S.l.], 1986. p. 1043–1044. 13

- DAWID, A. P. Present position and potential developments: Some personal views statistical theory the prequential approach. *Journal of the Royal Statistical Society: Series A (General)*, Wiley Online Library, v. 147, n. 2, p. 278–290, 1984. [56](#)
- DEGROOT, M. H.; SCHERVISH, M. J. *Probability and statistics*. [S.l.]: Pearson Education, 2012. [19](#), [33](#)
- DIAS, T. de L. et al. Problemas de escala e a relação área-indivíduo em análise espacial de dados censitários. 2000. [87](#)
- DIGGLE, P. J. et al. Second-order analysis of space-time clustering. *Statistical methods in medical research*, Sage Publications Sage CA: Thousand Oaks, CA, v. 4, n. 2, p. 124–136, 1995. [47](#)
- DOBSON, A. J.; BARNETT, A. G. *An introduction to generalized linear models*. [S.l.]: Chapman and Hall/CRC, 2018. [13](#)
- DOMINGUES, C. S. B. et al. Protocolo brasileiro para infecções sexualmente transmissíveis 2020: sífilis congênita e criança exposta à sífilis. *Epidemiologia e Serviços de Saúde*, SciELO Brasil, v. 30, 2021. [4](#)
- DOMINGUES, R. M. S. M.; LEAL, M. d. C. Incidência de sífilis congênita e fatores associados à transmissão vertical da sífilis: dados do estudo Nascir no Brasil. *Cadernos de Saúde Pública*, SciELO Brasil, v. 32, n. 6, 2016. [50](#), [85](#), [86](#), [93](#), [110](#), [111](#), [112](#)
- DREXLER, M.; AINSWORTH, C. H. Generalized additive models used to predict species abundance in the Gulf of Mexico: an ecosystem modeling tool. *PloS one*, Public Library of Science San Francisco, USA, v. 8, n. 5, p. e64458, 2013. [92](#)
- DUCHON, J. Splines minimizing rotation-invariant semi-norms in Sobolev spaces. In: *Constructive theory of functions of several variables*. [S.l.]: Springer, 1977. p. 85–100. [17](#)
- EARNEST, A. et al. Evaluating the effect of neighbourhood weight matrices on smoothing properties of conditional autoregressive (car) models. *International journal of health geographics*, BioMed Central, v. 6, n. 1, p. 1–12, 2007. [31](#), [77](#)
- ELLIOT, P.; CUZICK, J. *Geographical and environmental epidemiology: methods for Small area studies*. [S.l.]: World Health Organization [Geneva], 1992. [1](#)
- ELLIOTT, P.; WARTENBERG, D. Spatial epidemiology: current approaches and future challenges. *Environmental health perspectives*, National Institute of Environmental Health Sciences, v. 112, n. 9, p. 998–1006, 2004. [11](#), [47](#)
- ESCOBAR, N. D. et al. Perfil epidemiológico de sífilis adquirida nas regiões do Brasil no período de 2010 a 2019. *AMAZÔNIA: SCIENCE & HEALTH*, v. 8, n. 2, p. 51–63, 2020. [88](#)
- ESPINOSA, M. C. S. et al. Atención a la salud materno-infantil en Cuba: logros y desafíos. *Revista Panamericana de Salud Pública*, SciELO Public Health, v. 42, p. e27, 2018. [88](#)
- FARAWAY, J. J. *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. [S.l.]: Chapman and Hall/CRC, 2016. [52](#)
- FELIZ, M. C. et al. Aderência ao seguimento no cuidado ao recém-nascido exposto à sífilis e características associadas à interrupção do acompanhamento. *Revista Brasileira de Epidemiologia*, SciELO Brasil, v. 19, p. 727–739, 2016. [4](#)

- FENG, C. Spatial-temporal generalized additive model for modeling COVID-19 mortality risk in Toronto, Canada. *Spatial statistics*, Elsevier, p. 100526, 2021. [92](#)
- FENG, Y.; TONG, X. Calibrating nonparametric cellular automata with a generalized additive model to simulate dynamic urban growth. *Environmental Earth Sciences*, Springer, v. 76, n. 14, p. 1–15, 2017. [93](#)
- FEWSTER, R. M. et al. Analysis of population trends for farmland birds using generalized additive models. *Ecology*, Wiley Online Library, v. 81, n. 7, p. 1970–1984, 2000. [92](#)
- FISHER, R. A. Two new properties of mathematical likelihood. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, The Royal Society London, v. 144, n. 852, p. 285–307, 1934. [13](#)
- FOX, J. *Applied regression, generalized linear models, and related methods*. [S.l.]: Sage Publications, Beverly Hills, Calif, USA, 2008. [12](#)
- FOX, J. et al. The car package. *R Foundation for Statistical Computing*, v. 1109, 2007. [52](#)
- FRANCO-VILLORIA, M.; VENTRUCCI, M.; RUE, H. Variance partitioning in spatio-temporal disease mapping models. *arXiv preprint arXiv:2109.13374*, 2021. [48](#)
- GALATOIRE, P. S. A.; ROSSO, J. A.; SAKAE, T. M. Incidência de sífilis congênita nos estados do Brasil no período de 2007 a 2009. *Arq. Catarin. Med*, v. 41, n. 2, p. 26–32, 2012. [85](#), [112](#)
- GARBIN, C. A. S. et al. Syphilis in pregnancy: Profile and associated factors in the northwest region of São Paulo State. *Saúde e Pesquisa*, v. 14, n. 3, p. 1–10, 2021. [112](#)
- GARCÍA, J. et al. Standardization of variables and collinearity diagnostic in ridge regression. *International Statistical Review*, Wiley Online Library, v. 84, n. 2, p. 245–266, 2016. [52](#), [95](#)
- GAUDART, J. et al. Spatio-temporal dynamics of cholera during the first year of the epidemic in Haiti. *PLoS neglected tropical diseases*, Public Library of Science San Francisco, USA, v. 7, n. 4, p. e2145, 2013. [92](#)
- GEISSER, S.; EDDY, W. F. A predictive approach to model selection. *Journal of the American Statistical Association*, Taylor & Francis, v. 74, n. 365, p. 153–160, 1979. [54](#)
- GILKS, W. R.; RICHARDSON, S.; SPIEGELHALTER, D. *Markov chain Monte Carlo in practice*. [S.l.]: CRC press, 1995. [37](#), [38](#)
- GILL, J. *Bayesian methods: A social and behavioral sciences approach*. 3th. ed. [S.l.]: Chapman and Hall/CRC, 2015. [19](#), [96](#)
- GILL, J. et al. *Generalized linear models: a unified approach*. [S.l.]: Sage Publications, 2019. v. 134. [13](#)
- GOICOA, T. et al. In spatio-temporal disease mapping models, identifiability constraints affect PQL and INLA results. *Stochastic Environmental Research and Risk Assessment*, Springer, v. 32, n. 3, p. 749–770, 2018. [42](#)
- GOMÉZ-RUBIO, V.; BIVAND, R.; RUE, H. A new latent class to fit spatial econometrics models with integrated nested Laplace approximations. *Procedia Environmental Sciences*, Elsevier, v. 27, p. 116–118, 2015. [48](#)

- GRUENEWALD, P. J. et al. Mapping the spread of methamphetamine abuse in California from 1995 to 2008. *American journal of public health*, American Public Health Association, v. 103, n. 7, p. 1262–1270, 2013. 7
- GU, C.; GU, C. *Smoothing spline ANOVA models*. [S.l.]: Springer, 2013. v. 297. 16
- GUJARATI, D. N.; PORTER, D. C. *Econometria básica-5*. [S.l.]: Amgh Editora, 2011. 56, 97
- HAIDAR, F. H.; OLIVEIRA, U. F.; NASCIMENTO, L. F. C. Escolaridade materna: correlação com os indicadores obstétricos. *Cadernos de Saúde Pública*, SciELO Brasil, v. 17, n. 4, p. 1025–1029, 2001. 86, 111
- HAINING, R. P.; LI, G. *Modelling Spatial and Spatial-Temporal Data: A Bayesian Approach*. [S.l.]: CRC Press, 2020. ix, 31, 34, 42
- HAMILTON, J. D. *Time series analysis*. [S.l.]: Princeton university press, 2020. 25, 140
- HASTIE, T.; TIBSHIRANI, R. Generalized additive models for medical research. *Statistical methods in medical research*, Sage Publications Sage CA: Thousand Oaks, CA, v. 4, n. 3, p. 187–196, 1995. 115
- HEIBERGER, R. M.; HEIBERGER, M. R. M. Package ‘hh’. 2022. 52
- HELD, L.; SCHRÖDLE, B.; RUE, H. Posterior and cross-validators predictive checks: a comparison of MCMC and INLA. In: *Statistical modelling and regression structures*. [S.l.]: Springer, 2010. p. 91–110. 55
- HERINGER, A. L. D. S. et al. Inequalities in congenital syphilis trends in the city of Niterói, Brazil, 2007-2016. *Revista Panamericana de Salud Publica.*, v. 44, p. e3, 2020. 86, 111
- HODGES, J. S. *Richly parameterized linear models: additive, time series, and spatial models using random effects*. [S.l.]: CRC Press, 2013. 27
- HOOK, E. W. Syphilis. *The Lancet*, v. 389, n. 10078, p. 1550–1557, 2017. ISSN 0140-6736. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0140673616324114>>. 3
- IBGE. *Instituto Brasileiro de Geografia e Estatística - IBGE*. [S.l.]: Cd-Rom, 2010. 50, 85
- IBGE. *Instituto Brasileiro de Geografia e Estatística - IBGE*. [S.l.]: Cd-Rom, 2020. <<https://www.ibge.gov.br/cidades-e-estados/sp/sao-paulo.html>>. 49
- ISAACS, E. H.; SRIVASTAVA, M. R. *Applied geostatistics*. [S.l.: s.n.], 1989. 55, 96
- JIN, X.; CARLIN, B. P.; BANERJEE, S. Generalized hierarchical multivariate CAR models for areal data. *Biometrics*, Wiley Online Library, v. 61, n. 4, p. 950–961, 2005. 25
- KARP, G. et al. Syphilis and HIV co-infection. *European journal of internal medicine*, Elsevier, v. 20, n. 1, p. 9–13, 2009. 3
- KASS, R.; RAFTERY, A. Bayes factors, in journal of the American statistical association. 90, 773-795, 1995. 96
- KASS, R. E. Bayes factors in practice. *Journal of the Royal Statistical Society: Series D (The Statistician)*, Wiley Online Library, v. 42, n. 5, p. 551–560, 1993. 96

- KELLER, S. et al. Large-scale spatio-temporal patterns of Mediterranean cephalopod diversity. *PloS one*, Public Library of Science San Francisco, CA USA, v. 11, n. 1, p. e0146469, 2016. [92](#)
- KLEIN, J. L.; KLEIN, D. *Statistical visions in time: a history of time series analysis, 1662-1938*. [S.l.]: Cambridge University Press, 1997. [25](#)
- KNORR-HELD, L. Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in medicine*, Wiley Online Library, v. 19, n. 17-18, p. 2555–2567, 2000. [1](#), [2](#), [8](#), [37](#), [38](#), [47](#), [48](#), [92](#)
- KNORR-HELD, L.; BESAG, J. Modelling risk from a disease in time and space. *Statistics in medicine*, Wiley Online Library, v. 17, n. 18, p. 2045–2060, 1998. [47](#)
- KORENROMP, E. L. et al. Prevalence and incidence estimates for syphilis, chlamydia, gonorrhea, and congenital syphilis in Colombia, 1995–2016. *Revista Panamericana de Salud Pública*, SciELO Public Health, v. 42, p. e118, 2018. [88](#)
- KRAINSKI, E. T. Statistical analysis of space-time data: New models and applications. NTNU, 2018. [27](#)
- LAFETÁ, K. R. G. et al. Sífilis materna e congênita, subnotificação e difícil controle. *Revista Brasileira de Epidemiologia*, SciELO Brasil, v. 19, p. 63–74, 2016. [3](#)
- LAGO, E. G. et al. Congenital syphilis: identification of two distinct profiles of maternal characteristics associated with risk. *Sexually transmitted diseases*, JSTOR, p. 33–37, 2004. [87](#), [112](#)
- LAGO, E. G.; VACCARI, A.; FIORI, R. M. Clinical features and follow-up of congenital syphilis. *Sexually Transmitted Diseases*, LWW, v. 40, n. 2, p. 85–94, 2013. [4](#)
- LAPLACE, P.; ÉTRANGES, S. Mémoire sur la probabilité des causes par les événements [memory on the probability of the causes of the events]. *Savants Étranges*, v. 6, p. 621–656, 1774. [19](#)
- LAWSON, A. B. *Bayesian disease mapping: hierarchical modeling in spatial epidemiology*. [S.l.]: Chapman and Hall/CRC, 2018. [18](#), [47](#)
- LAWSON, A. B.; BROWNE, W. J.; RODEIRO, C. L. V. *Disease mapping with WinBUGS and MLwiN*. [S.l.]: John Wiley & Sons, 2003. v. 11. [36](#)
- LAWSON, A. B.; ROTEJANAPRASERT, C. Childhood brain cancer in Florida: a Bayesian clustering approach. *Statistics and Public Policy*, Taylor & Francis, v. 1, n. 1, p. 99–107, 2014. [7](#), [53](#)
- LEATHWICK, J.; ELITH, J.; HASTIE, T. Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. *Ecological modelling*, Elsevier, v. 199, n. 2, p. 188–196, 2006. [92](#)
- LEE, D. A comparison of conditional autoregressive models used in Bayesian disease mapping. *Spatial and spatio-temporal epidemiology*, Elsevier, v. 2, n. 2, p. 79–89, 2011. [35](#)
- LEE, D.; MITCHELL, R. Boundary detection in disease mapping studies. *Biostatistics*, Oxford University Press, v. 13, n. 3, p. 415–426, 2012. [36](#)

- LEROUX, B. G.; LEI, X.; BRESLOW, N. Estimation of disease rates in small areas: a new mixed model for spatial dependence. In: *Statistical models in epidemiology, the environment, and clinical trials*. [S.l.]: Springer, 2000. p. 179–191. [7](#), [10](#), [35](#)
- LINDGREN, F.; RUE, H.; LINDSTRÖM, J. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Wiley Online Library, v. 73, n. 4, p. 423–498, 2011. [48](#)
- LINDSEY, J. K. *Statistical analysis of stochastic processes in time*. [S.l.]: Cambridge University Press, 2004. v. 14. [27](#)
- LINO, C. M.; SOUSA, M. d. L. R. de; BATISTA, M. J. Epidemiological profile, spatial distribution, and syphilis time series: a cross-sectional study in a Brazilian municipality. *The Journal of Infection in Developing Countries*, v. 15, n. 10, p. 1462–1470, 2021. [112](#)
- LONG, X. et al. Spatio-temporal distribution of *Konosirus punctatus* spawning and nursing ground in the South Yellow Sea. *Acta Oceanologica Sinica*, Springer, v. 40, n. 8, p. 133–144, 2021. [92](#)
- LÓPEZ-QUÍLEZ, A.; MUNOZ, F. Review of spatio-temporal models for disease mapping. *Final Report for the EUROHEIS*, v. 2, 2009. [36](#)
- LU, Y. et al. Environmental influence on the spatiotemporal variability of spawning grounds in the western Guangdong waters, South China Sea. *Journal of Marine Science and Engineering*, Multidisciplinary Digital Publishing Institute, v. 8, n. 8, p. 607, 2020. [92](#)
- LÜTKEPOHL, H. *New introduction to multiple time series analysis*. [S.l.]: Springer Science & Business Media, 2005. [140](#)
- MACÊDO, V. C. d. et al. Risk factors for syphilis in women: case-control study. *Revista de Saúde Pública*, SciELO Brasil, v. 51, 2017. [87](#), [112](#)
- MACNAB, Y. C. On Gaussian Markov random fields and Bayesian disease mapping. *Statistical Methods in Medical Research*, SAGE Publications Sage UK: London, England, v. 20, n. 1, p. 49–68, 2011. [35](#)
- MARREIROS, A. P. et al. Sífilis congênita no Brasil: Quantificação do número de casos, realização do pré-natal e tratamento do parceiro. *Biota Amazônia (Biote Amazonie, Biota Amazonia, Amazonian Biota)*, v. 10, n. 3, p. 22–24, 2020. [5](#)
- MARTÍN, B.; GONZÁLEZ-ARIAS, J.; VICENTE-VÍRSEDA, J. Machine learning as a successful approach for predicting complex spatio-temporal patterns in animal species abundance. *Machine learning*, v. 44, p. 289–301, 2021. [92](#)
- MARTINO, S.; RUE, H. Implementing approximate Bayesian inference using integrated nested Laplace approximation: A manual for the inla program. *Department of Mathematical Sciences, NTNU, Norway*, Citeseer, 2009. [23](#)
- MARTINS, T. G. et al. Bayesian computing with INLA: new features. *Computational Statistics & Data Analysis*, Elsevier, v. 67, p. 68–83, 2013. [23](#)

- MASCHIO-LIMA, T. et al. Perfil epidemiológico de pacientes com sífilis congênita e gestacional em um município do Estado de São Paulo, Brasil. *Revista Brasileira de Saúde Materno Infantil*, SciELO Brasil, v. 19, p. 865–872, 2020. 85
- MCCARTHY, M. A. *Bayesian methods for ecology*. [S.l.]: Cambridge University Press, 2007. 19
- MCCULLAGH, P.; NELDER, J. A. *Generalized linear models*. [S.l.]: Chapman & Hall, New York, 1989. 12
- MEDEIROS, J. A. R. et al. Spatiotemporal dynamics of syphilis in pregnant women and congenital syphilis in the state of São Paulo, Brazil. *Scientific reports*, Nature Publishing Group, v. 12, n. 1, p. 1–13, 2022. 7, 48, 111
- MELO, N. G. D. O.; FILHO, D. A. de M.; FERREIRA, L. O. C. Intraurban differences related to inherited syphilis in Recife, State of Pernambuco, Brazil (2004-2006). *EPIDEMIOLOGIA E SERVICOS DE SAUDE*, MINISTERIO DA SAUDE ESPLANADA MINISTERIOS, v. 20, n. 2, p. 213–222, 2011. 50, 93, 111
- MINISTÉRIO DA SAÚDE. *Secretaria de Vigilância em Saúde - Governo Federal*. 2022. Resposta a Emergências. Disponível em: <<https://www.gov.br/saude/pt-br/composicao/svs/resposta-a-emergencias/sala-de-situacao-de-saude>>. Acesso em 10 de maio de 2022. 88
- MOISEN, G. G. et al. Predicting tree species presence and basal area in Utah: a comparison of stochastic gradient boosting, generalized additive models, and tree-based methods. *Ecological modelling*, Elsevier, v. 199, n. 2, p. 176–187, 2006. 92
- MORAGA, P. *Geospatial health data: Modeling and visualization with R-INLA and shiny*. [S.l.]: Chapman and Hall/CRC, 2019. 11, 47
- MORAGA, P. et al. A geostatistical model for combined analysis of point-level and area-level data using INLA and SPDE. *Spatial Statistics*, Elsevier, v. 21, p. 27–41, 2017. 23, 24
- MORAN, P. A. Notes on continuous stochastic phenomena. *Biometrika*, JSTOR, v. 37, n. 1/2, p. 17–23, 1950. 77, 97
- MUKHOPADHYAY, N. *Probability and statistical inference*. [S.l.]: CRC Press, 2020. 13
- MYERS, R. H. et al. *Generalized linear models: with applications in engineering and the sciences*. [S.l.]: John Wiley & Sons, 2012. 13
- NAEEM, N. S. A.; RAHMAN, N. A.; FAHIMI, F. A. M. A spatial–temporal study of dengue in Peninsular Malaysia for the year 2017 in two different space–time model. *Journal of Applied Statistics*, Taylor & Francis, v. 47, n. 4, p. 739–756, 2020. 1
- NAKAZAWA, M.; NAKAZAWA, M. M. Package ‘fmsb’. See <https://cran.r-project.org/web/packages/fmsb/fmsb.pdf>, 2019. 52
- NELDER, J. A.; WEDDERBURN, R. W. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, Wiley Online Library, v. 135, n. 3, p. 370–384, 1972. 12
- NETO, F. C. et al. Ocorrência de dengue e sua relação com medidas de controle e níveis de infestação de *Aedes aegypti* em uma cidade do sudeste brasileiro. *BEPA. Boletim Epidemiológico Paulista*, v. 17, n. 195, p. 3–19, 2020. 7

- NGANDO, N. E. et al. Relationship between the spatiotemporal distribution of dominant small pelagic fishes and environmental factors in mauritanian waters. *Journal of Ocean University of China*, Springer, v. 19, n. 2, p. 393–408, 2020. 92
- NGWIRA, A. et al. Spatial temporal distribution of COVID-19 risk during the early phase of the pandemic in Malawi. *PeerJ*, PeerJ Inc., v. 9, p. e11003, 2021. 7
- NOGUEIRA, W. P. et al. Syphilis in riverine communities: prevalence and associated factors. *Revista da Escola de Enfermagem da USP*, SciELO Brasil, v. 56, 2022. 87, 112
- OLESON, J. J.; SMITH, B. J.; KIM, H. Joint spatio-temporal modeling of low incidence cancers sharing common risk factors. *J Data Sci*, v. 6, p. 105–23, 2008. 8
- OLIVEIRA, L. R. d. Avaliação da efetividade das ações para prevenção e controle da transmissão vertical da sífilis em Cuiabá e Mato Grosso. Dissertação apresentada ao Programa de Pós-Graduação do Instituto de Saúde ..., 2012. 87
- OLIVEIRA, L. R. d. et al. Evaluation of preventative and control measures for congenital syphilis in State of Mato Grosso. *Revista da sociedade brasileira de medicina tropical*, SciELO Brasil, v. 47, n. 3, p. 334–340, 2014. 50, 93
- OLIVEIRA, S. I. M. d. et al. Syphilis notifications and the triggering processes for vertical transmission: a cross-sectional study. *International journal of environmental research and public health*, Multidisciplinary Digital Publishing Institute, v. 17, n. 3, p. 984, 2020. 6
- OPAS. Organización panamericana de la salud. iniciativa regional para la eliminación de la transmisión materno-infantil del VIH y de la sífilis congénita en América Latina y el Caribe: Estrategia de monitoreo regional. *Publicación Científica CLAP; 1574*, OPS, 2010. Disponível em <https://iris.paho.org/bitstream/handle/10665.2/9994/9789275330951_esp.pdf?sequence=1&isAllowed=y>;. 5
- OPAS. Organización panamericana de la salud. elimination of mother-to-child transmission of HIV and syphilis in the Americas. update 2016. *Publicación Científica CLAP; 1574*, OPS, 2017. Disponível em <https://iris.paho.org/bitstream/handle/10665.2/9994/9789275330951_esp.pdf?sequence=1&isAllowed=y>;. 5, 7, 88
- OPENSHAW, S. Ecological fallacies and the analysis of areal census data. *Environment and planning A*, SAGE Publications Sage UK: London, England, v. 16, n. 1, p. 17–31, 1984. 87
- OTIENDE, V. A.; ACHIA, T. N.; MWAMBI, H. G. Bayesian hierarchical modeling of joint spatiotemporal risk patterns for human immunodeficiency virus (HIV) and tuberculosis (TB) in Kenya. *PloS one*, Public Library of Science San Francisco, CA USA, v. 15, n. 7, p. e0234456, 2020. 7
- OTTO, P. spgarch: An R-package for spatial and spatiotemporal arch models. *arXiv preprint arXiv:1812.01871*, 2018. 77
- O'BRIEN, R. M. A caution regarding rules of thumb for variance inflation factors. *Quality & quantity*, Springer, v. 41, n. 5, p. 673–690, 2007. 52, 95
- PATTERSON, H. D.; THOMPSON, R. Recovery of inter-block information when block sizes are unequal. *Biometrika*, Oxford University Press, v. 58, n. 3, p. 545–554, 1971. 95

- PEDERSEN, E. J. et al. Hierarchical generalized additive models in ecology: an introduction with mgcv. *PeerJ*, PeerJ Inc., v. 7, p. e6876, 2019. 104
- PENNA, N. A.; FERREIRA, I. B. Desigualdades socioespaciais e áreas de vulnerabilidades nas cidades. *Mercator (Fortaleza)*, SciELO Brasil, v. 13, p. 25–36, 2014. 84, 110
- PEREIRA, T. A. C. et al. Aspectos epidemiológicos da sífilis congênita no Brasil. *Revista de Casos e Consultoria*, v. 12, n. 1, p. e24303–e24303, 2021. 3
- PETTIT, L. The conditional predictive ordinate for the normal distribution. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley Online Library, v. 52, n. 1, p. 175–184, 1990. 55
- PFEIFFER, D. et al. *Spatial analysis in epidemiology*. [S.l.]: Oxford University Press Oxford, 2008. v. 142. 11
- PREFEITURA DE SÃO PAULO. *Boletim CEINFO*. Ano XIV, nº 14, Junho/2015. 2015. 75
- RAFTERY, A. E. Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika*, Oxford University Press, v. 83, n. 2, p. 251–266, 1996. 20
- REIS, G. J. d. et al. Intraurban differentials in congenital syphilis: a predictive analysis by neighborhood in the city of Rio de Janeiro, Brazil. *Cadernos de Saúde Pública*, SciELO Brasil, v. 34, n. 9, 2018. 50, 93
- RESENDES, A. P. d. C. et al. Abordagens espaciais na saúde pública. 2006. 12
- REZAEIAN, M. et al. The production and interpretation of disease maps. *Social psychiatry and psychiatric epidemiology*, Springer, v. 39, n. 12, p. 947–954, 2004. 11
- RICHARDSON, S.; ABELLAN, J. J.; BEST, N. Bayesian spatio-temporal analysis of joint patterns of male and female lung cancer risks in Yorkshire (UK). *Statistical methods in medical research*, Sage Publications Sage CA: Thousand Oaks, CA, v. 15, n. 4, p. 385–407, 2006. 8
- RICHARDSON, S. et al. Interpreting posterior relative risk estimates in disease-mapping studies. *Environmental health perspectives*, National Institute of Environmental Health Sciences, v. 112, n. 9, p. 1016–1025, 2004. 74
- RIEBLER, A. et al. An intuitive Bayesian spatial model for disease mapping that accounts for scaling. *Statistical methods in medical research*, Sage Publications Sage UK: London, England, v. 25, n. 4, p. 1145–1165, 2016. 35
- RODRÍGUEZ, I. et al. The Cuban experience in the elimination of mother-to-child transmission of congenital syphilis. *American journal of public health*, American Public Health Association, v. 106, n. 11, p. 1975, 2016. 88
- ROSENBLATT, M. Remarks on a multivariate transformation. *The annals of mathematical statistics*, JSTOR, v. 23, n. 3, p. 470–472, 1952. 56
- RUBERTI, J. A. Análise espacial e espaço temporal dos casos de sífilis em gestante e sífilis congênita nos municípios do Estado de São Paulo, de 2010 a 2016. 2020. 6, 50, 85, 86, 93, 111, 112

- RUE, H. Fast sampling of Gaussian Markov random fields. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Wiley Online Library, v. 63, n. 2, p. 325–338, 2001. 27
- RUE, H.; HELD, L. *Gaussian Markov Random Fields: Theory and Applications (Monographs on Statistics & Applied Probability)*. 1. ed. [S.l.]: Chapman and Hall/CRC, 2005. 27, 35, 48, 142, 143
- RUE, H.; MARTINO, S. Approximate bayesian inference for hierarchical Gaussian Markov random field models. *Journal of statistical planning and inference*, Elsevier, v. 137, n. 10, p. 3177–3192, 2007. 22, 89
- RUE, H.; MARTINO, S.; CHOPIN, N. Approximate bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, Wiley Online Library, v. 71, n. 2, p. 319–392, 2009. 8, 22, 23, 24, 48, 57, 89, 98, 115
- RUE, H. et al. R-INLA: Approximate bayesian inference using integrated nested Laplace approximations. *Trondheim, Norway*. URL <http://www.r-inla.org>, 2013. 23, 115
- RUE, H. et al. Bayesian computing with INLA: a review. *Annual Review of Statistics and Its Application*, Annual Reviews, v. 4, p. 395–421, 2017. 43, 44
- RUIZ-CÁRDENAS, R.; KRAINSKI, E. T.; RUE, H. Direct fitting of dynamic models using integrated nested Laplace approximations—INLA. *Computational Statistics & Data Analysis*, Elsevier, v. 56, n. 6, p. 1808–1828, 2012. 8, 48
- SANDES, M. F. et al. Análise epidemiológica por sífilis congênita em menores de um ano no Estado de Sergipe. *Brazilian Journal of Health Review*, v. 2, n. 3, p. 1609–1615, 2019. 47
- SARACENI, V. et al. Vigilância epidemiológica da transmissão vertical da sífilis: dados de seis unidades federativas no Brasil. *Revista panamericana de salud publica*, SciELO Public Health, v. 41, p. e44, 2017. 5
- SCHLOERKE, B.; CROWLEY, J.; COOK, D. Package ‘ggally’. *Extension to ‘ggplot2’*. See, v. 713, 2018. ix, 62
- SCHMID, V.; HELD, L. Bayesian extrapolation of space–time trends in cancer registry data. *Biometrics*, Wiley Online Library, v. 60, n. 4, p. 1034–1042, 2004. 1
- SCHRÖDLE, B.; HELD, L. A primer on disease mapping and ecological regression using INLA. *Computational statistics*, Springer, v. 26, n. 2, p. 241–258, 2011. 1
- SCHRÖDLE, B.; HELD, L. Spatio-temporal disease mapping using INLA. *Environmetrics*, Wiley Online Library, v. 22, n. 6, p. 725–734, 2011. 11, 48
- SCHRÖDLE, B.; HELD, L.; RUE, H. Assessing the impact of a movement network on the spatiotemporal spread of infectious diseases. *Biometrics*, Wiley Online Library, v. 68, n. 3, p. 736–744, 2012. 48
- SCHWARZ, G. Estimating the dimension of a model. *The annals of statistics*, JSTOR, p. 461–464, 1978. 96

- SERAFIM, A. S. et al. Incidence of congenital syphilis in the South Region of Brazil. *Revista da Sociedade Brasileira de Medicina Tropical*, SciELO Brasil, v. 47, n. 2, p. 170–178, 2014. 50, 93
- SHEYNIN, O. B. Laplace's theory of errors. *Archive for history of exact sciences*, Springer, v. 17, n. 1, p. 1–61, 1977. 19
- SIANGPHOE, U.; WHEELER, D. C. Evaluation of the performance of smoothing functions in generalized additive models for spatial variation in disease. *Cancer informatics*, SAGE Publications Sage UK: London, England, v. 14, p. CIN–S17300, 2015. 92
- SILVA, M. J. N. d. et al. Distribuição da sífilis congênita no Estado do Tocantins, 2007-2015. *Epidemiologia e Serviços de Saúde*, SciELO Brasil, v. 29, 2020. 88
- SIMPSON, D. et al. Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical science*, JSTOR, p. 1–28, 2017. 7, 10, 35, 56
- SOUSA, D. M. d. N. et al. Sífilis congênita: reflexões sobre um agravamento sem controle na saúde mãe e filho. *Revista de Enfermagem UFPE On Line*, 2014. 5
- SOUZA, R. R. de et al. Perfil de casos notificados de sífilis congênita no Estado de Goiás entre 2015 a 2018. *Brazilian Journal of Development*, v. 6, n. 7, p. 48715–48725, 2020. 88
- SPIEGEL, E.; KNEIB, T.; OTTO-SOBOTKA, F. Spatio-temporal expectile regression models. *Statistical Modelling*, SAGE Publications Sage India: New Delhi, India, v. 20, n. 4, p. 386–409, 2020. 93
- SPIEGELHALTER, D. J. et al. Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)*, Wiley Online Library, v. 64, n. 4, p. 583–639, 2002. 54
- SPIEGELHALTER, D. J. et al. Winbugs version 1.4 user manual. *MRC Biostatistics Unit, Cambridge*. URL <http://www.mrc-bsu.cam.ac.uk/bugs>, 2003. 54
- STIGLER, S. M. *The history of statistics: The measurement of uncertainty before 1900*. [S.l.]: Harvard University Press, 1986. 19
- STONE, C. J. [generalized additive models]: comment. *Statistical Science*, JSTOR, v. 1, n. 3, p. 312–314, 1986. 93
- SÃO PAULO. *Boletim Epidemiológico de HIV-AIDS-IST*. [S.l.]: SES-SP – Secretaria de Estado da Saúde de São Paulo, 2020. Coordenadoria de Controle de Doenças. Centro de Referência e Treinamento em DST/AIDS. Programa Estadual de IST/AIDS de São Paulo. 36(1):1-291. 6
- TAYLOR, B. M.; DIGGLE, P. J. INLA or MCMC? a tutorial and comparative evaluation for spatial prediction in log-Gaussian Cox processes. *Journal of Statistical Computation and Simulation*, Taylor & Francis, v. 84, n. 10, p. 2266–2284, 2014. 24
- TEAM, R. C. R: A language and environment for statistical computing. Vienna, Austria, 2021. 23, 57, 98
- TEIXEIRA, L. O. et al. Tendência temporal e distribuição espacial da sífilis congênita no estado do Rio Grande do Sul entre 2001 e 2012. *Ciência & Saúde Coletiva*, SciELO Brasil, v. 23, p. 2587–2597, 2018. 84, 110

- TENNEKES, M. tmap: Thematic maps in R. *Journal of Statistical Software*, v. 84, p. 1–39, 2018. [57](#), [98](#)
- TERZI, Y.; CENGIZ, M. Using of generalized additive model for model selection in multiple poisson regression for air pollution data. *Scientific Research and Essays*, Academic Journals, v. 4, n. 9, p. 867–871, 2009. [15](#)
- THOMPSON, C. G. et al. Extracting the variance inflation factor and other multicollinearity diagnostics from typical regression results. *Basic and Applied Social Psychology*, Taylor & Francis, v. 39, n. 2, p. 81–90, 2017. [52](#), [95](#)
- UGARTE, M.; GOICOA, T.; MILITINO, A. Spatio-temporal modeling of mortality risks using penalized splines. *Environmetrics: The Official Journal of the International Environmetrics Society*, Wiley Online Library, v. 21, n. 3-4, p. 270–289, 2010. [92](#)
- UGWU, C. L. J.; ZEWOTIR, T. Evaluating the effects of climate and environmental factors on under-5 children malaria spatial distribution using generalized additive models (GAMs). *Journal of Epidemiology and Global Health*, Springer, v. 10, n. 4, p. 304, 2020. [92](#)
- ULLRICH, T. On the autoregressive time series model using real and complex analysis. *Forecasting*, Multidisciplinary Digital Publishing Institute, v. 3, n. 4, p. 716–728, 2021. [25](#)
- VASCONCELOS, M. N. d. Sífilis congênita em uma região de saúde do Ceará: um estudo ecológico. Biblioteca da Universidade Federal do Ceará, 2019. [86](#), [111](#)
- VATCHEVA, K. P. et al. Multicollinearity in regression analyses conducted in epidemiologic studies. *Epidemiology (Sunnyvale, Calif.)*, NIH Public Access, v. 6, n. 2, 2016. [52](#), [93](#)
- VEHTARI, A.; GELMAN, A.; GABRY, J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and computing*, Springer, v. 27, n. 5, p. 1413–1432, 2017. [54](#)
- VENABLES, W. N.; RIPLEY, B. D. *Modern applied statistics with S-PLUS*. [S.l.]: Springer Science & Business Media, 2013. [111](#)
- VISLOCKY, R. L.; FRITSCH, J. M. Generalized additive models versus linear regression in generating probabilistic MOS forecasts of aviation weather parameters. *Weather and Forecasting*, American Meteorological Society, v. 10, n. 4, p. 669–680, 1995. [110](#)
- WADE, K.; QUAAS, R. Solutions to a system of equations involving a first-order autoregressive process. *Journal of dairy science*, Elsevier, v. 76, n. 10, p. 3026–3032, 1993. [27](#)
- WAKEFIELD, J. Disease mapping and spatial regression with count data. *Biostatistics*, Oxford University Press, v. 8, n. 2, p. 158–183, 2007. [1](#), [28](#)
- WALLER, L. A.; GOTWAY, C. A. *Applied spatial statistics for public health data*. [S.l.]: John Wiley & Sons, 2004. v. 368. [25](#), [141](#)
- WAN, Z. et al. Maternal syphilis treatment and pregnancy outcomes: a retrospective study in Jiangxi Province, China. *BMC Pregnancy and Childbirth*, Springer, v. 20, n. 1, p. 1–8, 2020. [85](#), [112](#)
- WANG, P. D. M. *Spatio-temporal statistical modeling: climate impacts due to bioenergy crop expansion*. [S.l.], 2018. [52](#)

- WANG, Y. *Smoothing splines: methods and applications*. [S.l.]: CRC press, 2011. 16
- WANG, Y.-G.; FU, L.; PAUL, S. *Analysis of Longitudinal Data with Examples*. 1. ed. [S.l.]: Chapman and Hall/CRC, 2022. 27
- WATANABE, S.; OPPER, M. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of machine learning research*, v. 11, n. 12, 2010. 54
- WEBSTER, T. et al. Method for mapping population-based case-control studies: an application using generalized additive models. *International journal of health geographics*, BioMed Central, v. 5, n. 1, p. 1–10, 2006. 92
- WHO. Report on global sexually transmitted infection surveillance 2018. World Health Organization, 2018. World Health Organization. Disponível em: <<http://apps.who.int/iris/bitstream/handle/10665/277258/9789241565691-eng.pdf?ua=1>>. 3
- WHO. 2019. <<https://www.who.int/news/item/26-02-2019-who-publishes-new-estimates-on-congenital-syphilis>>. Acesso em 09 de junho de 2021. 4, 47, 88
- WIJESOORIYA, N. S. et al. Global burden of maternal and congenital syphilis in 2008 and 2012: a health systems modelling study. *The Lancet Global Health*, Elsevier, v. 4, n. 8, p. e525–e533, 2016. 4, 47, 85, 112
- WILLMOTT, C. J.; MATSUURA, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate research*, v. 30, n. 1, p. 79–82, 2005. 55, 96
- WOOD, S. N. mgcv: Gams and generalized ridge regression for r. *R news*, v. 1, n. 2, p. 20–25, 2001. 115
- WOOD, S. N. *Generalized additive models: an introduction with R*. 2th. ed. [S.l.]: chapman and hall/CRC, 2017. 16, 110
- WOOD, S. N.; PYA, N.; SÄFKEN, B. Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, Taylor & Francis, v. 111, n. 516, p. 1548–1563, 2016. 95
- WOOD, S. N.; SCHEIPL, F.; FARAWAY, J. J. Straightforward intermediate rank tensor product smoothing in mixed models. *Statistics and Computing*, Springer, v. 23, n. 3, p. 341–360, 2013. 15
- WOODWARD, P. *Bayesian analysis made simple: an Excel GUI for WinBUGS*. [S.l.]: Chapman and Hall/CRC, 2019. 20
- WU, M. Y. et al. Effect of syphilis infection on HIV acquisition: a systematic review and meta-analysis. *Sexually Transmitted Infections*, BMJ Publishing Group Ltd, 2020. 3
- XIMENES, I. P. E. et al. Incidência e controle da sífilis congênita no Ceará. *Rev Rene*, Universidade Federal do Ceará, v. 9, n. 3, p. 74–80, 2008. 86, 111
- YEE, T. W.; MITCHELL, N. D. Generalized additive models in plant ecology. *Journal of vegetation science*, Wiley Online Library, v. 2, n. 5, p. 587–602, 1991. 92

Apêndices

A Disposição dos dados

Tabela 21 – Disposição dos dados para análise com o INLA

idarea	idtime	idint	Ano	casos	NV	E	x1	x2	x3	x4
1	1	1	2010	6	239	1,152	-0,710	-0,184	0,845	2,374
2	1	2	2010	2	156	0,752	-0,042	-1,004	-1,407	0,470
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
309	1	309	2010	1	473	2,280	1,350	0,257	1,667	0,703
310	1	310	2010	1	261	1,258	-0,658	-0,343	-0,647	0,281
1	2	311	2011	3	549	2,646	-0,710	-0,184	0,845	2,374
2	2	312	2011	2	218	1,051	-0,042	-1,004	-1,407	0,470
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
309	2	619	2011	4	718	3,461	1,350	0,257	1,667	0,703
310	2	620	2011	0	294	1,417	-0,658	-0,343	-0,647	0,281
1	3	621	2012	0	576	2,776	-0,710	-0,184	0,845	2,374
2	3	622	2012	2	259	1,248	-0,042	-1,004	-1,407	0,470
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
309	3	929	2012	1	709	3,417	1,350	0,257	1,667	0,703
310	3	930	2012	0	257	1,239	-0,658	-0,343	-0,647	0,281
1	4	931	2013	2	607	2,926	-0,710	-0,184	0,845	2,374
2	4	932	2013	0	241	1,162	-0,042	-1,004	-1,407	0,470
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
309	4	1239	2013	3	639	3,080	1,350	0,257	1,667	0,703
310	4	1240	2013	1	265	1,277	-0,658	-0,343	-0,647	0,281
1	5	1241	2014	1	516	2,487	-0,710	-0,184	0,845	2,374
2	5	1242	2014	1	232	1,118	-0,042	-1,004	-1,407	0,470
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
309	5	1549	2014	10	668	3,220	1,350	0,257	1,667	0,703
310	5	1550	2014	0	270	1,301	-0,658	-0,343	-0,647	0,281
1	6	1551	2015	5	533	2,569	-0,710	-0,184	0,845	2,374
2	6	1552	2015	1	207	0,998	-0,042	-1,004	-1,407	0,470
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
309	6	1859	2015	6	635	3,061	1,350	0,257	1,667	0,703
310	6	1860	2015	0	228	1,099	-0,658	-0,343	-0,647	0,281
1	7	1861	2016	2	481	2,318	-0,710	-0,184	0,845	2,374
2	7	1862	2016	0	197	0,950	-0,042	-1,004	-1,407	0,470
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
309	7	2169	2016	10	656	3,162	1,350	0,257	1,667	0,703
310	7	2170	2016	2	232	1,118	-0,658	-0,343	-0,647	0,281

B Produto de Kronecker

O produto Kronecker (denotado por $\otimes$) consiste no produto de duas matrizes $\mathbf{A}$ e $\mathbf{B}$ de tamanho arbitrário e resulta em uma matriz em blocos. Por exemplo, se $\mathbf{A}$ é uma matriz 2×2 da forma $\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ e $\mathbf{B}$ é uma matriz de tamanho arbitrário, então o produto de Kronecker de $\mathbf{A}$ por $\mathbf{B}$ é:

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} \end{pmatrix},$$

em que $a_{11}\mathbf{B}$, por exemplo, corresponde à multiplicação de todos os elementos em $\mathbf{B}$ pelo escalar a_{11} . Em geral, se $\mathbf{A}$ é uma matriz $n \times m$ e $\mathbf{B}$ é uma matriz $p \times q$, então $\mathbf{A} \otimes \mathbf{B}$ é uma matriz em blocos de tamanho $np \times mq$. A definição e as propriedades do produto de Kronecker podem ser encontradas em [Lütkepohl \(2005\)](#), [Box e Tiao \(2011\)](#), [Banerjee e Roy \(2014\)](#), [Hamilton \(2020\)](#).

C Campos aleatórios

C.1 Campos aleatórios gaussianos

Um campo aleatório é um processo estocástico $\{X(\mathbf{s}); \mathbf{s} \in D\}$ definido num espaço $D \subset \mathbb{R}^k$, com k inteiro e D fixo e contínuo, em que $\mathbf{s}$ representa as localizações, $X(\mathbf{s})$ é a variável aleatória observada nas localizações $\mathbf{s} \in D \subset \mathbb{R}^k$ e k representa a dimensão do campo aleatório. No nosso caso, vamos restringir a duas dimensões ($k = 2$), de modo que cada local é um vetor bidimensional, $\mathbf{s} = (x, y)$, referenciando um ponto no plano, em que x e y são as coordenadas de longitude e latitude de uma localização geográfica.

Nas aplicações geoestatísticas, supõe-se que os dados sejam uma realização parcial (pois não podemos, por razões práticas, observar o processo em todos os pontos de D). Assim, para um local fixo $\mathbf{s}$, $X(\mathbf{s})$ é uma variável aleatória à qual as leis da probabilidade se aplicam (WALLER; GOTWAY, 2004).

Dado um campo aleatório, se

$$E(X(\mathbf{s})) = \boldsymbol{\mu} \text{ para todo } \mathbf{s} \in D, \quad (\text{C.1})$$

ou seja, a média do processo não depende da localização e

$$\text{Cov}(X(\mathbf{s}_i), X(\mathbf{s}_j)) = C(\mathbf{s}_i - \mathbf{s}_j) \quad \text{para todo } \mathbf{s}_i, \mathbf{s}_j \in D, \forall i \neq j \text{ e } i, j = 1, 2, \dots, n \quad (\text{C.2})$$

ou seja, a covariância depende apenas da diferença entre os locais $\mathbf{s}_i$ e $\mathbf{s}_j$, e não dos próprios locais, isto é, $X(\cdot)$ é considerado estacionário de segunda ordem. A função $C(\cdot)$ definida pela Equação (C.2) é chamada de função de covariância e é uma medida de autocorrelação espacial. Assim, por meio do pressuposto de estacionariedade, o processo se repete essencialmente no espaço, fornecendo a replicação necessária para estimação e inferência. Se, além disso, $C(\mathbf{s}_i - \mathbf{s}_j)$ é uma função apenas da distância entre $\mathbf{s}_i$ e $\mathbf{s}_j$ e não da direção, o processo é denominado isotrópico.

O campo aleatório $\{X(\mathbf{s}), \mathbf{s} \in D\}$ é dito campo aleatório Gaussiano se para todo $n \geq 1$ e quaisquer localizações $\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_n \in D$, o vetor aleatório $(X(\mathbf{s}_1), X(\mathbf{s}_2), \dots, X(\mathbf{s}_n))^T$ tem distribuição multivariada com vetor de médias $\boldsymbol{\mu} = (\mu_1(\mathbf{s}_1), \mu_2(\mathbf{s}_2), \dots, \mu_n(\mathbf{s}_n))^T$ e matriz de

covariância:

$$\Sigma = \begin{bmatrix} C(\mathbf{s}_1, \mathbf{s}_1) & C(\mathbf{s}_1, \mathbf{s}_2) & \cdots & C(\mathbf{s}_1, \mathbf{s}_n) \\ C(\mathbf{s}_2, \mathbf{s}_1) & C(\mathbf{s}_2, \mathbf{s}_2) & \cdots & C(\mathbf{s}_2, \mathbf{s}_n) \\ \vdots & \vdots & \ddots & \vdots \\ C(\mathbf{s}_n, \mathbf{s}_1) & C(\mathbf{s}_n, \mathbf{s}_2) & \cdots & C(\mathbf{s}_n, \mathbf{s}_n) \end{bmatrix},$$

em que $C(\mathbf{s}_i, \mathbf{s}_j) = \text{Cov}(X(\mathbf{s}_i), X(\mathbf{s}_j))$, para todo $i, j = 1, 2, \dots, n$ e $i \neq j$. Consequentemente, cada $X(\mathbf{s}_i)$ é uma variável aleatória Gaussiana univariada.

A função de covariância deve ser positiva definida, ou seja, para qualquer conjunto de locais, a matriz de covariância é definida positiva.

C.2 Campos aleatórios gaussianos Markovianos

Nesta seção é apresentada uma definição de campos aleatórios gaussianos de Markov e sua conexão com os modelos autorregressivos.

A definição de campo aleatório gaussiano de Markov (GMRF, do inglês *gaussian markov random field*) envolve alguma teoria básica de grafos. Grafos não direcionados são usados para representar a estrutura de independência condicional em um GMRF.

Um grafo $G = (V, E)$ é definido por um grupo de vértices (V), geralmente chamados nós, unidos entre si por um grupo de linhas chamado arestas (E). Se dois nós $i, j \in V$ são unidos por uma aresta, eles são ditos vizinhos e denotamos por $i \sim j$. Se todas as arestas não tiverem direção, este grafo será chamado de grafo não direcionado. Se todas as arestas tiverem direção, este grafo será chamado de grafo direcionado (DAG, do inglês *directed acyclic graph*).

Como exemplo, a Figura 41 mostra quatro áreas de ponderação do município de São Paulo ($V = 1, 2, 3, 4$) e a estrutura de vizinhança do tipo *queen* ($E = \{\{1, 2\}, \{1, 3\}, \{2, 3\}, \{2, 4\}\}$) e a matriz de adjacência associada $\mathbf{A}$, com $a_{ij} = a_{ji} = 1$ se $\{i, j\} \in E$, $\forall i \neq j \in V$ e $a_{ij} = a_{ji} = 0$ caso contrário.

Um campo aleatório de Markov é um campo aleatório que satisfaz as hipóteses de independência condicional em relação a um grafo, daí o termo Markov. Os campos aleatórios de Markov que também são gaussianos são chamados de campos aleatórios de Markov gaussianos e tiveram um enorme impacto na estatística espacial (RUE; HELD, 2005).

Um campo aleatório $\{X(\mathbf{s}); \mathbf{s} \in D\}$ é dito campo aleatório Markoviano Gaussiano (GMRF) em relação a um grafo $G = (V, E)$ com média $\boldsymbol{\mu}$ e matriz de precisão $\mathbf{Q} > 0$ (positiva definida) se, e somente se, para quaisquer localizações $\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_n \in D$, a distribuição conjunta de $\mathbf{X} = (X(\mathbf{s}_1), X(\mathbf{s}_2), \dots, X(\mathbf{s}_n))^T$ é dada por:

$$f_{\mathbf{X}}(\mathbf{x}) = (2\pi)^{-(n/2)} |\mathbf{Q}|^{1/2} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{Q}(\mathbf{x} - \boldsymbol{\mu})\right)$$

em que

$$Q_{ij} \neq 0 \iff i, j \in E, \text{ para todo } i \neq j.$$

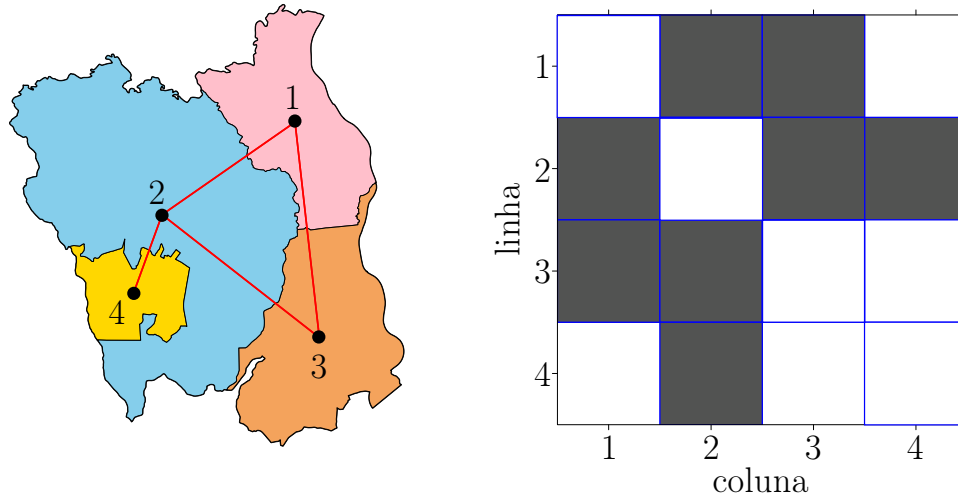


Figura 41 – As quatro áreas selecionadas (esquerda) e a matriz de adjacência associada (direita).

Assim, a matriz de precisão $\mathbf{Q}$ informa quais áreas são vizinhas, dado algum critério de vizinhança. Aqui, o conjunto de vértices V corresponde aos nós $\{s_1, s_2, \dots, s_n\}$ e o conjunto de arestas E especifica as dependências entre as variáveis aleatórias $X(s_1), X(s_2), \dots, X(s_n)$. Além disso, se $\mathbf{Q}$ é uma matriz $n \times n$ simétrica definida positiva, então Q_{ij} é igual a zero se e somente se, os nós i e j não estão conectados por uma aresta. Logo, para $i \neq j$, tem-se:

$$X(s_i) \perp X(s_j) | X(s_{-ij}) \iff Q_{ij} = 0.$$

Isso implica que $Y(s_i)$ e $Y(s_j)$ são condicionalmente independentes e significa que a distribuição condicional da variável observada em algum nó depende apenas de seus vizinhos. Essa definição nos diz que é possível saber se dois nós são condicionalmente independentes “lendo” a matriz de precisão $\mathbf{Q}$, em que $\mathbf{Q}$ determina o grafo G por seus valores não-nulos. Se $\mathbf{Q}$ é uma matriz completamente densa, então $\mathbf{Q}$ está totalmente conectado. Outra característica importante dos GMRF é que, devido às suas propriedades de Markov preservadas, a matriz de precisão $\mathbf{Q}$ é esparsa, isto é, terá alguns elementos não nulos. Portanto, trabalhar com uma matriz de precisão esparsa em vez de uma matriz de covariância densa nos permite obter inferência muito mais rápida. Assim, o benefício de usar um GMRF é puramente computacional e reside na escassez da matriz de precisão, porque existem muitos métodos numéricos que usam esse recurso para computação rápida. Para uma abordagem completa dos GMRFs, sugerimos consultar (RUE; HELD, 2005).

D *Script* em linguagem R para obtenção das variáveis do IBGE

```
library(RSQLite)
library(data.table)
library(DBI)
library(survey)
library(convey)
library(dplyr)
library(sqldf)
library(rio)

#----- PESSOAS -----
con <- dbConnect(RSQLite::SQLite(), dbname = "PESSOAS")

dbWriteTable(conn = con, name = "PESSOAS", value = "Amostra_
  Pessoas_35_RMSP.txt", header = FALSE, row.names = FALSE,
  overwrite = TRUE, eol = '\n')

dados_sp_pessoa <- data.table(dbGetQuery(con,
"select substr(V1, 29, 16) 'V0010',
      substr(V1, 1, 2) 'V0001', substr(V1, 8, 13) 'V0011',
      substr(V1, 21, 8) 'V0300', substr(V1, 154, 2) 'V0633',
      substr(V1, 58, 1) 'V0601', substr(V1, 68, 1) 'V0606',
      substr(V1, 146, 1) 'V0627', substr(V1, 148, 2) 'V0629',
      substr(V1, 190, 1) 'V0637', substr(V1, 194, 1) 'V0640',
      substr(V1, 404, 2) 'V5060', substr(V1, 406, 8) 'V5070',
      substr(V1, 414, 9) 'V5080', substr(V1, 62, 3) 'V6036',
      substr(V1, 190, 1) 'V6037', substr(V1, 158, 1) 'V0635',
      substr(V1, 352, 1) 'V0663', substr(V1, 353, 2) 'V6631',
      substr(V1, 355, 2) 'V6632', substr(V1, 357, 2) 'V6633',
      substr(V1, 247, 7) 'V6525', substr(V1, 279, 7) 'V6529',
      substr(V1, 296, 8) 'V6531', substr(V1, 319, 1) 'V0657',
```

```
      substr(V1, 399, 1) 'V0604', substr(V1, 156, 1) 'V0634',
      substr(V1, 391, 1) 'V6900', substr(V1, 396, 3) 'V6121'
    from PESSOAS where substr(V1, 8, 7) = '3550308'))

save(dados_sp_pessoa, file="dados_sp_pessoa.RData")
load("dados_sp_pessoa.RData")

dados_sp_pessoa[, V0010 := as.numeric(V0010) / 10 ^ 13]
dados_sp_pessoa[, V0001 := as.factor(V0001)]
dados_sp_pessoa[, V0634 := as.factor(V0634)]
dados_sp_pessoa[, V0601 := as.factor(V0601)]
dados_sp_pessoa[, V0606 := as.factor(V0606)]
dados_sp_pessoa[, V0627 := as.factor(V0627)]
dados_sp_pessoa[, V0637 := as.factor(V0637)]
dados_sp_pessoa[, V0640 := as.factor(V0640)]
dados_sp_pessoa[, V0604 := as.factor(V0604)]
dados_sp_pessoa[, V0635 := as.factor(V0635)]
dados_sp_pessoa[, V0663 := as.factor(V0663)]
dados_sp_pessoa[, V6121 := as.factor(V6121)]
dados_sp_pessoa[, V5060 := as.integer(V5060)]
dados_sp_pessoa[, V6631 := as.integer(V6631)]
dados_sp_pessoa[, V6632 := as.integer(V6632)]
dados_sp_pessoa[, V6633 := as.integer(V6633)]
dados_sp_pessoa[, V6529 := as.numeric(V6529)]
dados_sp_pessoa[, V6531 := as.numeric(V6531)]
dados_sp_pessoa[, V5070 := as.numeric(V5070)]
dados_sp_pessoa[, V5080 := as.numeric(V5080)]
dados_sp_pessoa[, V0010 := as.numeric(V0010)]
dados_sp_pessoa[, V6036 := as.integer(V6036)]

dados_sp_pessoa[, FPC:=sum(V0010), by = V0011]

dados_sp_pessoa$one<-rep(1,552037)

# Usando amostragem complexa
options(survey.lonely.psu="adjust")
amostra_pessoa <- svydesign(ids = ~ V0300, strata = ~ V0011,
  weights = ~ V0010, fpc = ~ FPC, data = dados_sp_pessoa)

# Raca
raca<-svyby(~ V0606, ~ V0011, amostra_pessoa, svymean,
  na.rm = TRUE)[,c(1,2,3,5)]
```



```

raca<-data.frame(cbind(raca[,1] ,100*raca[,2:4]))
colnames(raca)<-c("AP","PBRANCA","PPRETA","PPARDA")

# Amostra de mulheres com 18 a 24 anos
amostra.mulheres18a24<-subset(amostra_pessoa , V0601 == 2& V6036
    >=18 & V6036<=24)
amostra.mulheres<-subset(amostra_pessoa , V0601 == 2)

# Amostra de mulheres adolescentes
amostra.mulheres12a17<-subset(amostra_pessoa , V0601 == 2& V6036
    >=12 & V6036<=17)

# Proporcao de nascidos vivos de maes adolescentes
PNVMA<-svyby(~V6633, ~ V0011, amostra.mulheres12a17, svymean, na.
    rm = TRUE)[,c(1,2)]
PNVMA<-data.frame(cbind(PNVMA[,1] ,100*PNVMA[,2]))
colnames(PNVMA)<-c("AP","PNVMA")

# Proporcao de Mulheres de 18 a 24 anos sem instrucao ou com
    fundamental incompleto
MSIF1824<-svyby(~V0635, ~ V0011, amostra.mulheres18a24, svymean,
    na.rm = TRUE)[,c(1,2)]
MSIF1824<-data.frame(cbind(MSIF1824[,1] ,100*MSIF1824[,2]))
colnames(MSIF1824)<-c("AP","MSIF1824")

# Proporcao de Mulheres em Idade Fertil
PFERTIL<-svyby(~V6036>=15&V6036<=49, ~ V0011, amostra.mulheres,
    svymean, na.rm = TRUE)[,c(1,3)]
PFERTIL<-data.frame(cbind(PFERTIL[,1] ,100*PFERTIL[,2]))
colnames(PFERTIL)<-c("AP","PFERTIL")

## Calculo do indice de Gini
amostra_pessoa <- convey_prep(amostra_pessoa)
GINI<-svyby(~ V6531, ~V0011, amostra_pessoa, svygini,na.rm=T)[,c
    (1,2)]
GINI<-data.frame(GINI)
colnames(GINI)<-c("AP","GINI")

#----- DOMICILIOS -----
con2 <- dbConnect(RSQLite::SQLite(), dbname = "DOMICILIO")

dbWriteTable(conn = con2, name = "DOMICILIO", value = "Amostra_

```

```

Domicilios_35_RMSP.txt", header = FALSE, row.names = FALSE,
overwrite = TRUE, eol = '\n')

dados_sp_domic <- data.table(dbGetQuery(con2, "select substr(V1,
  29, 16) 'V0010',
    substr(V1, 1, 2) 'V0001', substr(V1, 8, 13) 'V0011',
    substr(V1, 21, 8) 'V0300', substr(V1, 53, 1) 'V1006',
    substr(V1, 75, 2) 'V0203', substr(V1, 79, 3) 'V6203',
    substr(V1, 85, 1) 'V0205', substr(V1, 86, 1) 'V0206',
    substr(V1, 87, 1) 'V0207', substr(V1, 88, 2) 'V0208',
    substr(V1, 90, 1) 'V0209', substr(V1, 91, 1) 'V0210',
    substr(V1, 92, 1) 'V0211', substr(V1, 106, 2) 'V0401',
    substr(V1, 116, 10) 'V6530', substr(V1, 126, 8) 'V6531',
    substr(V1, 109, 7) 'V6529', substr(V1, 144, 1) 'V6210',
    substr(V1, 98, 1) 'V0217', substr(V1, 99, 1) 'V0218',
    substr(V1, 101, 1) 'V0220'
  from DOMICILIO where substr(V1, 8, 7) = '3550308'"))

save(dados_sp_domic, file="dados_sp_domic.RData")

load("dados_sp_domic.RData")

dados_sp_domic[, V0010 := as.numeric(V0010) / 10 ^ 13]
dados_sp_domic[, V0001 := as.factor(V0001)]
dados_sp_domic[, V1006 := as.factor(V1006)]
dados_sp_domic[, V0205 := as.factor(V0205)]
dados_sp_domic[, V0206 := as.factor(V0206)]
dados_sp_domic[, V0207 := as.factor(V0207)]
dados_sp_domic[, V0208 := as.factor(V0208)]
dados_sp_domic[, V0209 := as.factor(V0209)]
dados_sp_domic[, V0210 := as.factor(V0210)]
dados_sp_domic[, V0211 := as.factor(V0211)]
dados_sp_domic[, V0217 := as.factor(V0217)]
dados_sp_domic[, V0218 := as.factor(V0218)]
dados_sp_domic[, V6210 := as.factor(V6210)]
dados_sp_domic[, V0001 := as.factor(V0001)]
dados_sp_domic[, V0203 := as.integer(V0203)]
dados_sp_domic[, V6203 := as.numeric(V6203)]
dados_sp_domic[, V0401 := as.integer(V0401)]
dados_sp_domic[, V6530 := as.numeric(V6530)]
dados_sp_domic[, V6531 := as.numeric(V6531)]
dados_sp_domic[, V6529 := as.numeric(V6529)]

```

```

dados_sp_domic$abast_agua[dados_sp_domic$V0208=="01"]<-1

dados_sp_domic$abast_agua[(
  dados_sp_domic$V0208=="02"|dados_sp_domic$V0208=="03"|
  dados_sp_domic$V0208=="04"|dados_sp_domic$V0208=="05"|
  dados_sp_domic$V0208=="06"|dados_sp_domic$V0208=="07"|
  dados_sp_domic$V0208=="08"|dados_sp_domic$V0208=="09"|
  dados_sp_domic$V0208=="10"|dados_sp_domic$V0208==" ")<-2

dados_sp_domic[, abast_agua:= as.factor(abast_agua)]
dados_sp_domic$agua.esgoto[dados_sp_domic$abast_agua==2 & dados_sp_domic$esgot_sanitario==2]<-1
dados_sp_domic$agua.esgoto[dados_sp_domic$abast_agua!=2 | dados_sp_domic$esgot_sanitario!=2]<-2

dados_sp_domic[, agua.esgoto:= as.factor(agua.esgoto)]

dados_sp_domic[, FPC:=sum(V0010), by = V0011]

dados_sp_domic$one<-rep(1,174162)

amostra_domic <- svydesign(ids = ~ V0300, strata = ~ V0011,
  weights = ~ V0010, fpc = ~ FPC, data = dados_sp_domic)

# Percentual da populacao que vive em domicilios com banheiro e
  agua encanada
PBA<-svyby(~ agua_ban, ~ V0011, amostra_domic, svymean, na.rm =
  TRUE)[,c(1,2)]
PBA<-data.frame(cbind(PBA[,1] ,100*PBA[,2]))
colnames(PBA)<-c("AP","PBA")

# Renda media per capta
RDPC<-merge(sqldf("SELECT V0011, sum(V6531) from dados_sp_domic
  group by V0011"), sqldf("SELECT V0011, sum(V0401) from dados_sp_domic
  group by V0011"),by="V0011")
RDPC<-data.frame(cbind(RDPC[,1],RDPC[,2]/RDPC[,3]))
colnames(RDPC)<-c("AP","RDPC")

variaveis<-raca %>% left_join(mulh_mae_viva, by = c('AP'))%>%
  left_join(NV_MAE_ADOLES,by = c('AP'))%>%
  left_join(M_analfa_18a24,by = c('AP'))%>%

```

```
left_join(M_sem_inst_fund18a24, by = c('AP'))%>%
left_join(fertil, by = c('AP'))%>%
left_join(gini, by = c('AP'))%>%
left_join(T_BANAGUA, by = c('AP'))%>%
left_join(T_LIXO, by = c('AP'))%>%
left_join(T_LUZ, by = c('AP'))%>%
left_join(AGUA_ESGOTO, by = c('AP'))%>%
left_join(moradia_adeq, by = c('AP'))%>%
left_join(RDPC, by = c('AP'))%>%
left_join(religiao, by = c('AP'))%>%
left_join(INTERNET, by = c('AP'))%>%
left_join(FONE, by = c('AP'))

export(variaveis, "variaveis.xlsx")
```

E Script em linguagem R para obter o VIF

```
vif_func<-function(in_frame,thresh=10,trace=T,...){

  library(fmsb)

  if(any(!'data.frame' %in% class(in_frame))) in_frame<-data.frame(in_frame)

  # get initial vif value for all comparisons of variables
  vif_init<-NULL
  var_names <- names(in_frame)
  for(val in var_names){
    regressors <- var_names[-which(var_names == val)]
    form <- paste(regressors, collapse = '+')
    form_in <- formula(paste(val, '~', form))
    vif_init<-rbind(vif_init, c(val, VIF(lm(form_in, data = in_frame, ...))))
  }
  vif_max<-max(as.numeric(vif_init[,2]), na.rm = TRUE)

  if(vif_max < thresh){
    if(trace==T){ #print output of each iteration
      prmatrix(vif_init,collab=c('Variáveis','VIF'),rowlab=rep('',nrow(vif_init)),quote=F)
      cat('\n')
      cat(paste('Todas as variáveis tem VIF < ', thresh,', max VIF ',round(vif_max,2), sep=''),'\n\n')
    }
    return(var_names)
  }
  else{

    in_dat<-in_frame
```

```

#backwards selection of explanatory variables, stops when all
  VIF values are below 'thresh'
while(vif_max >= thresh){

  vif_vals<-NULL
  var_names <- names(in_dat)

  for(val in var_names){
    regressors <- var_names[-which(var_names == val)]
    form <- paste(regressors, collapse = '+')
    form_in <- formula(paste(val, '~', form))
    vif_add<-VIF(lm(form_in, data = in_dat, ...))
    vif_vals<-rbind(vif_vals,c(val,vif_add))
  }
  max_row<-which(vif_vals[,2] == max(as.numeric(vif_vals[,2])
    , na.rm = TRUE))[1]

  vif_max<-as.numeric(vif_vals[max_row,2])

  if(vif_max<thresh) break

  if(trace==T){ #print output of each iteration
    prmatrix(vif_vals,collab=c('Variáveis','VIF'),rowlab=rep(
      '',nrow(vif_vals)),quote=F)
    cat('\n')
    cat('Variável Removida: ',vif_vals[max_row,1],vif_max,'\n
      \n')
    flush.console()
  }

  in_dat<-in_dat[,!names(in_dat) %in% vif_vals[max_row,1]]

}

prmatrix(vif_vals,collab=c('Variáveis','VIF'),rowlab=rep('',
  nrow(vif_vals)),quote=F)

return(names(in_dat))

}
}

```

F *Script* em linguagem R utilizado nas análises do Capítulo 3

```
# Pacotes necessários
library(ggplot2)
library(GGally)
library(cowplot)
library(INLA)
library(brinla)
library(spdep)
library(gridExtra)
library(raster)
library(tmap)
library(RColorBrewer)
library(INLAOutputs)
library(Metrics)
library(tidyverse)
library(rio)
```

F.1 Modelos com estrutura espacial BYM

A seguir apresentamos os códigos em R utilizados no ajuste dos modelos considerando-se a distribuição de Poisson. Para a distribuição Binomial Negativa, o procedimento é análogo, substituindo-se `family="poisson"` por `family="nbinomial"`.

```
temp <- poly2nb(area_pond, queen = T)
nb2INLA("adjmat", temp)
adjmat <- paste(getwd(), "/adjmat", sep="")

# Plotando a Matriz de adjacência
H <- inla.read.graph(filename="adjmat")
image(inla.graph2matrix(H), xlab="", ylab="")
```

```

dados<-read.csv("dados_sifilis_AP.csv")

# Padronizando as variáveis
dados[,c(6:12)] <- scale(dados[,c(6:12)], scale = T)

# Obtendo o valor esperado e o SIR bruto

S=310 # número de áreas
T=7   # número de anos

d<-d%>%data.frame(idtime=rep(1:T,each=S),
                  idtime2=rep(1:T,each=S),
                  idtime3=rep(1:T,each=S),
                  idarea=rep(1:S,T),
                  idarea2=rep(1:S,T),
                  idint=seq(1,T*S),
                  idareaint=rep(1:S,T),
                  idtimeint=rep(1:T,each=S))

# Definindo as priors
priors <-list(prec = list(prior = "loggmamma", param = c
  (1,0.00005)))

priors_bym<-list(phi = list(prior = "loggmamma", param = c
  (1,0.00005)),prec = list(prior = "loggmamma", param = c
  (1,0.00005)))

##### Modelos sem covariáveis #####

# 1 - Modelo espaço-temporal assumindo tendência linear para o
  efeito temporal

formula1_bym_bernardinelli_sem_cov_sc<- y_sc ~ f(idarea, model =
  "bym", param=priors_bym, graph = adjmat,scale.model = T) + f(
  idarea2, idtime, model = "iid",param=priors) + idtime

mod1_bym_bernardinelli_sem_cov_sc<- inla(formula1_bym_
  bernardinelli_sem_cov_sc,family="poisson",data=d,E=E,control.
  predictor=list(compute=TRUE),control.compute=list(dic=TRUE,cpo=
  TRUE,waic=T,config = TRUE),control.inla=list(strategy="
  simplified.laplace",int.strategy="auto"),verbose=T)

```


2 - Modelo espaço-temporal sem interação espaço-temporal

```
formula2_bym_ar1_sem_cov_sc<- y_sc ~ f(idarea,model="bym", param=
  priors_bym,graph=adjmat,scale.model = T) + f(idtime,model="ar1"
  , param=priors) + f(idtime2,model="iid",param=priors)

mod2_bym_ar1_sem_cov_sc<- inla(formula2_bym_ar1_sem_cov_sc,family
  ="poisson",data=d,E=E,control.predictor=list(compute=TRUE),
  control.compute=list(dic=TRUE,cpo=TRUE,waic=TRUE,config = TRUE)
  ,control.inla=list(strategy="simplified.laplace",int.strategy="
  auto"),verbose=T)
```

3 - Modelos espaço-temporal com interação espaço-temporal

3.1 - Modelo espaço-temporal com interação do tipo I

```
formula31_bym_ar1_sem_cov_sc<- y_sc ~ f(idarea,model="bym", param
  =priors_bym,graph=adjmat,scale.model = T) + f(idtime,model="ar1
  ", param=priors) + f(idtime2,model="iid",param=priors) + f(
  idint,model="iid",param=priors)

mod31_bym_ar1_sem_cov_sc<- inla(formula31_bym_ar1_sem_cov_sc,
  family="poisson",data=d,E=E,control.predictor=list(compute=TRUE
  ),control.compute=list(dic=TRUE,cpo=TRUE,waic=TRUE,config =
  TRUE),control.inla=list(strategy="simplified.laplace",int.
  strategy="auto"),verbose=T)
```

3.2 - Modelo espaço-temporal com interação do tipo II

```
formula32_bym_ar1_sem_cov_sc<- y_sc ~ f(idarea,model="bym",param=
  priors_bym, graph=adjmat,adjust.for.con.comp = TRUE,scale.model
  = T) + f(idtime,model="ar1",param=priors) + f(idtime2,model="
  iid",param=priors)+ f(idareaint,model = "iid",param=priors,
  group = idtimeint,control.group = list(model = "ar1",param=
  priors,scale.model = T))

mod32_bym_ar1_sem_cov_sc<- inla(formula32_bym_ar1_sem_cov_sc,
  family="poisson",data=d,E=E,control.predictor=list(compute=TRUE
  ),control.compute=list(dic=TRUE,cpo=TRUE,waic=TRUE,config =
  TRUE),control.inla=list(strategy="simplified.laplace",int.
  strategy="auto"),verbose=T)
```

```

# 3.3 - Modelo espaço-temporal com interação do tipo III
formula33_bym_ar1_sem_cov_sc<- y_sc ~ f(idarea,model="bym", param
  =priors_bym,graph=adjmat,scale.model = T) + f(idtime,model="ar1
  ",param=priors) + f(idtime2,model="iid",param=priors)+ f(
  idtimeint, model = "iid",param=priors, group = idareaint,
  control.group = list(model = "besag",param=priors,graph=
    adjmat))

mod33_bym_ar1_sem_cov_sc<- inla(formula33_bym_ar1_sem_cov_sc,
  family="poisson",data=d,E=E,control.predictor=list(compute=TRUE
  ),control.compute=list(dic=TRUE,cpo=TRUE,waic=TRUE,config =
  TRUE),control.inla=list(strategy="simplified.laplace",int.
  strategy="auto"),verbose=T)

# 3.4 - Modelo espaço-temporal com interação do tipo IV

mod34_bym_ar1_sem_cov_sc<- y_sc ~ f(idarea,model="bym", param=
  priors_bym,graph=adjmat,scale.model = T) + f(idtime,model="ar1"
  ,param=priors) + f(idtime2,model="iid",param=priors) + f(
  idareaint, model = "besag",param=priors, graph = adjmat,group =
  idtimeint,control.group = list(model = "ar1",param=priors,
  scale.model = T))

mod34_bym_ar1_sem_cov_sc<- inla(mod34_bym_ar1_sem_cov_sc,family="
  poisson",data=d,E=E,control.predictor=list(compute=TRUE),
  control.compute=list(dic=TRUE,cpo=TRUE,waic=TRUE,config = TRUE)
  ,control.inla=list(strategy="simplified.laplace",int.strategy="
  auto"),verbose=T)

##### Modelos com covariáveis #####

# 4 - Modelo espaço-temporal assumindo tendência linear para
  efeito temporal

formula4_bym_bernardinelli_com_cov_sc<- y_sc ~ prop_preta + PROP_
  NV_MAE_ADOLES + prop_m_sem_inst_fund18a24 + PROP_FERTIL + GINI
  + T_BANAGUA + RDPC + f(idarea, model = "bym", param=priors_bym,
  graph = adjmat,scale.model = T) + f(idarea2, idtime, model = "
  iid",param=priors) + idtime

mod4_bym_bernardinelli_com_cov_sc<- inla(formula4_bym_
  bernardinelli_com_cov_sc,family="poisson",data=d,E=E,control.

```

```

    predictor=list(compute=TRUE),control.compute=list(dic=TRUE,cpo=
    TRUE,waic=T,config = TRUE),control.inla=list(strategy="
    simplified.laplace",int.strategy="auto"),verbose=T)

# 5 - Modelo espaço-temporal sem interação espaço-temporal

formula5_bym_ar1_com_cov_sc<- y_sc ~ prop_preta+PROP_NV_MAE_
    ADOLAS+prop_m_sem_inst_fund18a24 + PROP_FERTIL + GINI + T_
    BANAGUA + RDPC + f(idarea,model="bym", param=priors_bym,graph=
    adjmat,scale.model = T) + f(idtime,model="ar1", param=priors) +
    f(idtime2,model="iid",param=priors)

mod5_bym_ar1_com_cov_sc <- inla(formula5_bym_ar1_com_cov_sc,
    family="poisson",data=d,E=E,control.predictor=list(compute=TRUE
    ),control.compute=list(dic=TRUE,cpo=TRUE,waic=TRUE,config =
    TRUE),control.inla=list(strategy="simplified.laplace",int.
    strategy="auto"),verbose=T)

# 6 - Modelo espaço-temporal com interação espaço-temporal

# 6.1 - Modelo espaço-temporal com interação do tipo I

formula61_bym_ar1_com_cov_sc<- y_sc ~ prop_preta + PROP_NV_MAE_
    ADOLAS + prop_m_sem_inst_fund18a24 + PROP_FERTIL+ GINI+T_
    BANAGUA + RDPC + f(idarea,model="bym",param=priors_bym,graph=
    adjmat,scale.model = T) + f(idtime,model="ar1", param=priors) +
    f(idtime2,model="iid",param=priors) + f(idint,model="iid",
    param=priors)

mod61_bym_ar1_com_cov_sc <- inla(formula61_bym_ar1_com_cov_sc,
    family="poisson",data=d,E=E,control.predictor=list(compute=TRUE
    ),control.compute=list(dic=TRUE,cpo=TRUE,waic=TRUE,config =
    TRUE),control.inla=list(strategy="simplified.laplace",int.
    strategy="auto"),verbose=T)

# 6.2 - Modelo espaço-temporal com interação do tipo II

formula62_bym_ar1_com_cov_sc<- y_sc ~ prop_preta+PROP_NV_MAE_
    ADOLAS+prop_m_sem_inst_fund18a24 + PROP_FERTIL+ GINI+T_BANAGUA
    + RDPC + f(idarea,model="bym",param=priors_bym, graph=adjmat,
    adjust.for.con.comp = TRUE,scale.model = T) + f(idtime,model="
    ar1",param=priors) + f(idtime2,model="iid",param=priors) + f(

```

```

idareaaint,model = "iid",param=priors,group = idtimeint,control.
group = list(model = "ar1",param=priors,scale.model = T))

mod62_bym_ar1_com_cov_sc<- inla(formula62_bym_ar1_com_cov_sc,
family="poisson",data=d,E=E,control.predictor=list(compute=TRUE
),control.compute=list(dic=TRUE,cpo=TRUE,waic=TRUE,config =
TRUE),control.inla=list(strategy="simplified.laplace",int.
strategy="auto"),verbose=T)

# 6.3 - Modelo espaço-temporal com interação do tipo III

formula63_bym_ar1_com_cov_sc<- y_sc ~ prop_preta +PROP_NV_MAE_
ADOLES + prop_m_sem_inst_fund18a24 + PROP_FERTIL + GINI + T_
BANAGUA + RDPC + f(idarea,model="bym",param=priors_bym,graph=
adjmat,scale.model = T) + f(idtime,model="ar1",param=priors) +
f(idtime2,model="iid",param=priors) + f(idtimeint,model = "iid"
,param=priors, group = idareaaint,control.group = list(model = "
besag",param=priors,graph=adjmat))

mod63_bym_ar1_com_cov_sc <- inla(formula63_bym_ar1_com_cov_sc,
family="poisson",data=d,E=E,control.predictor=list(compute=TRUE
),control.compute=list(dic=TRUE,cpo=TRUE,waic=TRUE,config =
TRUE),control.inla=list(strategy="simplified.laplace",int.
strategy="auto"),verbose=T)

# 6.4 - Modelo espaço-temporal com interação do tipo IV

mod64_bym_ar1_com_cov_sc<- y_sc ~ prop_preta + PROP_NV_MAE_ADOLES
+ prop_m_sem_inst_fund18a24 + PROP_FERTIL+ GINI + T_BANAGUA +
RDPC + f(idarea,model="bym",param=priors_bym,graph=adjmat,scale
.model = T) + f(idtime,model="ar1",param=priors) + f(idtime2,
model="iid",param=priors) + f(idareaaint, model = "besag",param=
priors, graph = adjmat,group = idtimeint,control.group = list(
model = "ar1",param=priors,scale.model = T))

mod64_bym_ar1_com_cov_sc<- inla(mod64_bym_ar1_com_cov_sc,family="
poisson",data=d,E=E,control.predictor=list(compute=TRUE),
control.compute=list(dic=TRUE,cpo=TRUE,waic=TRUE,config = TRUE)
,control.inla=list(strategy="simplified.laplace",int.strategy="
auto"),verbose=T)

```

F.2 Modelos com estrutura espacial BYM2

Os códigos para ajuste dos modelos são os mesmos utilizados para a estrutura BYM com `model="bym2"` e `param=priors_bym2`.

```
# Definindo as prioris

priors_bym2 <-list(phi = list(prior = "pc", param = c(.5, 2/3)),
  prec = list(prior = "pc.prec", param = c(.3/.31, .01)))
```

F.3 Modelos com estrutura espacial de Leroux

Os códigos para ajuste dos modelos são os mesmos utilizados para a estrutura BYM com `model="besagproper2"` e `param=priors_leroux`.

```
# Definindo as prioris
# - Unif(0,Inf) para desvios padrão
# - Unif(0,1) para o parâmetro de suavização espacial

sdunif="expression:
  logdens=-log_precision/2;
  return(logdens)"

lunif = "expression:
  a = 1;
  b = 1;
  beta = exp(theta)/(1+exp(theta));
  logdens = lgamma(a+b)-lgamma(a)-lgamma(b)+(a-1)*log(beta)+(b
    -1)*log(1-beta);
  log_jacobian = log(beta*(1-beta));
  return(logdens+log_jacobian)"

priors_leroux<-list(prec =list(prior = sdunif),lambda = list(
  prior = lunif))
```

F.4 Critérios de escolha do melhor modelo, estimativas, resíduos e mapas de risco

```
# Função para o cálculo do coeficiente de determinação
R2<-function(x)
{
```

```

1-sum((d$y_sc- x$summary.fitted.values$mean*d$E)^2)/sum((d$y_sc
  -mean(d$y_sc))^2)
}

# Critérios e métricas para escolha do melhor modelo
resumos_poisson_bernardinelli<-cbind(
  rbind(mod1_bym_bernardinelli_sem_cov_sc$dic$dic,
    mod1_bym2_bernardinelli_sem_cov_sc$dic$dic,
    mod1_leroux_bernardinelli_sem_cov_sc$dic$dic,
    mod4_bym_bernardinelli_com_cov_sc$dic$dic,
    mod4_bym2_bernardinelli_com_cov_sc$dic$dic,
    mod4_leroux_bernardinelli_com_cov_sc$dic$dic),

  rbind(mod1_bym_bernardinelli_sem_cov_sc$waic$waic,
    mod1_bym2_bernardinelli_sem_cov_sc$waic$waic,
    mod1_leroux_bernardinelli_sem_cov_sc$waic$waic,
    mod4_bym_bernardinelli_com_cov_sc$waic$waic,
    mod4_bym2_bernardinelli_com_cov_sc$waic$waic,
    mod4_leroux_bernardinelli_com_cov_sc$waic$waic),

  rbind(mod1_bym_bernardinelli_sem_cov_sc$mlik[1],
    mod1_bym2_bernardinelli_sem_cov_sc$mlik[1],
    mod1_leroux_bernardinelli_sem_cov_sc$mlik[1],
    mod4_bym_bernardinelli_com_cov_sc$mlik[1],
    mod4_bym2_bernardinelli_com_cov_sc$mlik[1],
    mod4_leroux_bernardinelli_com_cov_sc$mlik[1]),

  rbind(rmse(mod1_bym_bernardinelli_sem_cov_sc$summary.fitted.
    values$mean*d$E,d$y_sc),
    rmse(mod1_bym2_bernardinelli_sem_cov_sc$summary.fitted.
      values$mean*d$E,d$y_sc),
    rmse(mod1_leroux_bernardinelli_sem_cov_sc$summary.fitted.
      values$mean*d$E,d$y_sc),
    rmse(mod4_bym_bernardinelli_com_cov_sc$summary.fitted.
      values$mean*d$E,d$y_sc),
    rmse(mod4_bym2_bernardinelli_com_cov_sc$summary.fitted.
      values$mean*d$E,d$y_sc),
    rmse(mod4_leroux_bernardinelli_com_cov_sc$summary.fitted.
      values$mean*d$E,d$y_sc)),

  rbind(mae(mod1_bym_bernardinelli_sem_cov_sc$summary.fitted.
    values$mean*d$E,d$y_sc),

```

```

    mae(mod1_bym2_bernardinelli_sem_cov_sc$summary.fitted.
        values$mean*d$E,d$y_sc),
    mae(mod1_leroux_bernardinelli_sem_cov_sc$summary.fitted.
        values$mean*d$E,d$y_sc),
    mae(mod4_bym_bernardinelli_com_cov_sc$summary.fitted.
        values$mean*d$E,d$y_sc),
    mae(mod4_bym2_bernardinelli_com_cov_sc$summary.fitted.
        values$mean*d$E,d$y_sc),
    mae(mod4_leroux_bernardinelli_com_cov_sc$summary.fitted.
        values$mean*d$E,d$y_sc)),

rbind(
  R2(mod1_bym_bernardinelli_sem_cov_sc),
  R2(mod1_bym2_bernardinelli_sem_cov_sc),
  R2(mod1_leroux_bernardinelli_sem_cov_sc),
  R2(mod4_bym_bernardinelli_com_cov_sc),
  R2(mod4_bym2_bernardinelli_com_cov_sc),
  R2(mod4_leroux_bernardinelli_com_cov_sc)))

colnames(resumos_poisson_bernardinelli)<-c("DIC","WAIC","P_W","P_
  D","LMP","RMSE","MAE","R2","LS")
rownames(resumos_poisson_bernardinelli)<-c("Bern BYM","Bern BYM2"
  ,"Bern LEROUX","Bern BYM+cov","Bern BYM2+cov","Bern LEROUX+cov"
  )

resumos_poisson_bym<-cbind(
rbind(
mod2_bym_ar1_sem_cov_sc$dic$dic,
mod31_bym_ar1_sem_cov_sc$dic$dic,
mod32_bym_ar1_sem_cov_sc$dic$dic,
mod33_bym_ar1_sem_cov_sc$dic$dic,
mod34_bym_ar1_sem_cov_sc$dic$dic),

rbind(
  mod2_bym_ar1_sem_cov_sc$waic$waic,
  mod31_bym_ar1_sem_cov_sc$waic$waic,
  mod32_bym_ar1_sem_cov_sc$waic$waic,
  mod33_bym_ar1_sem_cov_sc$waic$waic,
  mod34_bym_ar1_sem_cov_sc$waic$waic),

rbind(
  mod2_bym_ar1_sem_cov_sc$mlik[1],

```

```

mod31_bym_ar1_sem_cov_sc$mlik[1],
mod32_bym_ar1_sem_cov_sc$mlik[1],
mod33_bym_ar1_sem_cov_sc$mlik[1],
mod34_bym_ar1_sem_cov_sc$mlik[1]),

rbind(
  rmse(mod2_bym_ar1_sem_cov_sc$summary.fitted.values$mean*d$E,d$y_
    _sc),
  rmse(mod31_bym_ar1_sem_cov_sc$summary.fitted.values$mean*d$E,d$y_
    y_sc),
  rmse(mod32_bym_ar1_sem_cov_sc$summary.fitted.values$mean*d$E,d$y_
    y_sc),
  rmse(mod33_bym_ar1_sem_cov_sc$summary.fitted.values$mean*d$E,d$y_
    y_sc),
  rmse(mod34_bym_ar1_sem_cov_sc$summary.fitted.values$mean*d$E,d$y_
    y_sc)),

rbind(
  mae(mod2_bym_ar1_sem_cov_sc$summary.fitted.values$mean*d$E,d$y_
    _sc),
  mae(mod31_bym_ar1_sem_cov_sc$summary.fitted.values$mean*d$E,d$y_
    _sc),
  mae(mod32_bym_ar1_sem_cov_sc$summary.fitted.values$mean*d$E,d$y_
    _sc),
  mae(mod33_bym_ar1_sem_cov_sc$summary.fitted.values$mean*d$E,d$y_
    _sc),
  mae(mod34_bym_ar1_sem_cov_sc$summary.fitted.values$mean*d$E,d$y_
    _sc)),

rbind(
  R2(mod2_bym_ar1_sem_cov_sc),
  R2(mod31_bym_ar1_sem_cov_sc),
  R2(mod32_bym_ar1_sem_cov_sc),
  R2(mod33_bym_ar1_sem_cov_sc),
  R2(mod34_bym_ar1_sem_cov_sc)))

colnames(resumos_poisson_bym)<-c("DIC","WAIC","P_W","P_D","LMP","
  RMSE","MAE","R2","LS")
rownames(resumos_poisson_bym)<-c("BYM SI","BYM TIPO I","BYM TIPO
  II","BYM TIPO III","BYM TIPO IV")

numer<-data.frame(cbind(seq(1:36)))

```



```

colnames(numer)<-c("Modelo")

resumo_geral_p<-cbind(numer, rbind(resumos_poisson_bernardinelli,
  resumos_poisson_bym, resumos_poisson_bym2, resumos_poisson_leroux,
  resumos_poisson_bym_cov, resumos_poisson_bym2_cov, resumos_
  poisson_leroux_cov))

# Seleção de covariáveis significativas
formula1<- y_sc ~ prop_preta+prop_m_sem_inst_fund18a24+
  PROP_FERTIL+ GINI+T_BANAGUA +RDPC+
  f(idarea,model="bym2",param=priors_bym2, graph=adjmat,adjust.
    for.con.comp = TRUE,scale.model = T)+
  f(idtime,model="ar1",param=priors) +
  f(idtime2,model="iid",param=priors)+
  f(idareaint,model = "iid",param=priors,group = idtimeint,
    control.group = list(model = "ar1",param=priors,scale.model =
      T))

mod1 <- inla(formula1,family="poisson",data=d,E=E,control.
  predictor=list(compute=TRUE),control.compute=list(dic=TRUE,cpo=
  TRUE,waic=TRUE,config = TRUE),control.inla=list(strategy="
  simplified.laplace",int.strategy="auto"),verbose=T)

formula2<- y_sc ~ prop_preta+prop_m_sem_inst_fund18a24+
  PROP_FERTIL+ GINI +RDPC+
  f(idarea,model="bym2",param=priors_bym2, graph=adjmat,adjust.
    for.con.comp = TRUE,scale.model = T)+
  f(idtime,model="ar1",param=priors) +
  f(idtime2,model="iid",param=priors)+
  f(idareaint,model = "iid",param=priors,group = idtimeint,
    control.group = list(model = "ar1",param=priors,scale.model =
      T))

mod2 <- inla(formula2,family="poisson",data=d,E=E,control.
  predictor=list(compute=TRUE),control.compute=list(dic=TRUE,cpo=
  TRUE,waic=TRUE,config = TRUE),control.inla=list(strategy="
  simplified.laplace",int.strategy="auto"),verbose=T)

summary(mod2)

formula3<- y_sc ~ prop_m_sem_inst_fund18a24+

```

```

PROP_FERTIL+ GINI      +RDPC+
f(idarea,model="bym2",param=priors_bym2, graph=adjmat,adjust.
  for.con.comp = TRUE,scale.model = T)+
f(idtime,model="ar1",param=priors) +
f(idtime2,model="iid",param=priors)+
f(idareaint,model = "iid",param=priors,group = idtimeint,
  control.group = list(model = "ar1",param=priors,scale.model =
    T))

mod3 <- inla(formula3,family="poisson",data=d,E=E,control.
  predictor=list(compute=TRUE),control.compute=list(dic=TRUE,cpo=
    TRUE,waic=TRUE,config = TRUE),control.inla=list(strategy="
    simplified.laplace",int.strategy="auto"),verbose=T)
summary(mod3)

save(list=c("mod1","mod2","mod3"),file="Modelos/modelos_selecao_
  INLA_poisson.Rdata")

# Tendência temporal a posteriori
formula.ST2<-y_sc ~ prop_m_sem_inst_fund18a24+PROP_FERTIL+GINI+
  RDPC+ f(idarea,model="bym",param=priors_bym2, graph=adjmat,
  adjust.for.con.comp = TRUE,scale.model = T) + f(idtime,model="
  ar1",param=priors) + f(idtime2,model="iid",param=priors) + f(
  idareaint,model = "iid",param=priors,group = idtimeint,control.
  group = list(model = "ar1",param=priors,scale.model = T))

lcs <- inla.make.lincombs(idtime = diag(7), idtime2 = diag(7))

model.ST2 <- inla(formula.ST2,family="poisson",data=d,E=E,control
  .predictor=list(compute=TRUE),lincomb=lcs,verbose=T)

temporal.IID<-unlist(exp(data.frame(lapply(model.ST2$marginals.
  random$idtime2,function(x){marg <- inla.tmarginal(function(x) x
    , x) inla.emarginal(mean, marg)}}))))

temporal.CAR <-unlist(exp(data.frame(lapply(model.ST2$marginals.
  random$idtime,function(x){marg <- inla.tmarginal(function(x) x
    , x) inla.emarginal(mean, marg)}}))))

temporal.total<-unlist(exp(data.frame(lapply(model.ST2$marginals.
  lincomb.derived,function(x){marg<-inla.tmarginal(function(x) x,
    x) inla.emarginal(mean,marg)}}))))

```

```

efeito<-data.frame(cbind(cbind(rep(seq(2010,2016,1),3)),rbind(
  cbind(temporal.IID), cbind(temporal.CAR),cbind(temporal.total))
,rbind(cbind(rep("T1",7)),cbind(rep("T2",7)),cbind(rep("T3",7))
)))
colnames(efeito)<-c("Ano","RR","Efeito")
row.names(efeito)<-c(seq(1,21,1))
efeito$Ano<-as.numeric(efeito$Ano)
efeito$RR<-as.numeric(efeito$RR)
efeito_temp<-efeito

ggplot(efeito_temp, aes(x =Ano, y = RR,color = Efeito)) +
  geom_line(lty =1,lwd=0.9) + geom_point(size = 2) +
  scale_y_continuous(breaks=seq(0.7,1.6 ,0.1)) +
  scale_x_continuous(name = "Ano", breaks=c(2010, 2011, 2012,
    2013, 2014, 2015,2016)) +
  scale_colour_manual(name="",labels = c("Efeito aleatório não
    estruturado","Efeito aleatório estruturado","Efeito aleatório
    total"),values = c("blue","red","darkgreen")) +
  geom_hline(yintercept=1, linetype="dashed", color = "black",
    size=0.6) +
  theme_bw() + labs(y="Risco relativo",x="Ano",color = "Efeitos",
    family = "Times") +
  theme(panel.grid = element_blank(),panel.spacing = unit(2, "cm"
    ),legend.position = "bottom",
  text=element_text(color = "black", size = 16,family = "Times"),
    panel.border = element_rect(colour = "black", fill =NA, size
    = 0.5),
    legend.title = element_text(color = "black", size = 16,face="
    bold",family = "Times"),
    axis.text.x = element_text(color = "black", size = 14,family
    = "Times",angle = 0,vjust=0.5,hjust=0.5),
    axis.text.y = element_text(color = "black", size = 14,family
    = "Times"),
    axis.title.x = element_text(color = "black", size = 18,face="
    bold",family = "Times"),
    axis.title.y = element_text(color = "black", size = 18,face="
    bold",family = "Times"),
    plot.title = element_text(color = "black", size = 16,face="
    bold",family = "Times"))

ggsave("efeito_temporal_SC.pdf",width=20,height=12,units = "cm")

```

```

# Mapas do risco relativo e probabilidade a posteriori
d$RR_inla_f<-mod3$summary.fitted.values$mean*d$E
RR_inla_f<-data.frame(d$RR_inla_f)
export(RR_inla_f,"RR_inla_f.csv")
d$PP<-FittedExcess(mod3,cutoff = 1) # P(RR>1)

RR<-data.frame(cbind(d[1:310,1], matrix(d$RR,byrow=F,ncol=7)))
colnames(RR)<-colnames(RR)<-c("ID_AP","RR2010","RR2011","RR2012",
  "RR2013","RR2014","RR2015","RR2016")

PP<-data.frame(cbind(d[1:310,1], matrix(d$PP,byrow=F,ncol=7)))
colnames(PP)<-colnames(PP)<-c("ID_AP","PP2010","PP2011","PP2012",
  "PP2013","PP2014","PP2015","PP2016")

export(RR,"RR_inla_congenita.csv")
export(PP,"PP_congenita.csv")

data<-read.csv2("RR_inla_congenita.csv",h=F,sep="," ,dec="," )[-c
  (1),]
colnames(data)<-c("ID_AP","RR2010","RR2011","RR2012","RR2013","
  RR2014","RR2015","RR2016")

area_pond@data$RR2010<-as.numeric(data$RR2010)
area_pond@data$RR2011<-as.numeric(data$RR2011)
area_pond@data$RR2012<-as.numeric(data$RR2012)
area_pond@data$RR2013<-as.numeric(data$RR2013)
area_pond@data$RR2014<-as.numeric(data$RR2014)
area_pond@data$RR2015<-as.numeric(data$RR2015)
area_pond@data$RR2016<-as.numeric(data$RR2016)

my.color=brewer.pal(5, "Oranges")

labels.RRSC=c("0,10 - 0,54", "0,55 - 0,79", "0,80 - 0,99", "1,00
  - 1,49", "1,50 - 6,96")
breaks.RRSC=c(0,0.55,0.8,1 ,1.5,Inf)

RR_plot2010<-tm_shape(area_pond)+tm_fill(rbind("RR2010"),labels =
  labels.RRSC,title="Risco Relativo",palette=my.color,breaks=
  breaks.RRSC) + tm_style("white",frame.lwd=0,alpha=0) +
  tm_borders(alpha=.2,col="black") +
  tm_layout(legend.text.color='red',legend.text.size =0.8,frame =

```

```

    FALSE, legend.title.color="blue", compass.type="arrow", legend.
    show=F, legend.only=F, legend.outside=T, legend.outside.position
    ="bottom", legend.title.fontface ="bold", legend.position=c
    (0.75,0.7), title = c("2010"), title.fontface = "bold", title.
    size = 1.5, title.position=c(0.2,0.91), fontfamily = "serif") +
    tm_shape(bairros) + tm_borders("black", lwd = 1.5)

# Os mapas de 2011 a 2015 são obtidos de modo análogo

RR_plot2016<-tm_shape(area_pond)+tm_fill("RR2016", labels = labels
    .RRSC, title="Risco Relativo", palette=my.color, breaks=breaks.
    RRSC, between.margin=c(0,0,0,1)) + tm_style("white", frame.lwd=1,
    alpha=1) + tm_legend(position=c(0.5,0.2)) +
    tm_compass(position=c(0.9,0.77), text.size =1, lwd = 1, show.
    labels = 1, just="center") +
    tm_scale_bar(position=c(0.5 ,0.11), text.size =3, lwd = 1, breaks=
    c(0,5,10,15)) + tm_borders(alpha=.2, col="black", lwd=1) +
    tm_layout(legend.text.color='black', legend.text.size =1.2, frame
    = F, legend.width = 3, legend.height = 5, legend.title.color="
    black", legend.title.size=1.6, compass.type="arrow", legend.show
    =T, legend.only=F, legend.outside=F, legend.outside.position="
    bottom", legend.title.fontface ="bold", legend.position=c
    (1.15,0.28), title = "2016", title.fontface = "bold", title.size
    = 1.5, title.position=c(0.2,0.91), fontfamily = "serif") +tm_
    shape(bairros) + tm_borders("black", lwd = 1.5)

RR_plot<-tmap_arrange(RR_plot2010, RR_plot2011, RR_plot2012, RR_
    plot2013, RR_plot2014, RR_plot2015, RR_plot2016, nrow=2, ncol=4)

tmap_save(RR_plot, "Figuras/mape_risco_relativo_inla_SC.png")

data<-read.csv2("PP_congenita.csv", h=F, sep=",", dec=",")[-c(1),]
colnames(data)<-c("ID_AP", "PP2010", "PP2011", "PP2012", "PP2013", "
    PP2014", "PP2015", "PP2016")

area_pond@data$PP2010<-as.numeric(data$PP2010)
area_pond@data$PP2011<-as.numeric(data$PP2011)
area_pond@data$PP2012<-as.numeric(data$PP2012)
area_pond@data$PP2013<-as.numeric(data$PP2013)
area_pond@data$PP2014<-as.numeric(data$PP2014)
area_pond@data$PP2015<-as.numeric(data$PP2015)
area_pond@data$PP2016<-as.numeric(data$PP2016)

```

```

labels.PPSC=c("0,0 - 0,2","0,2 - 0,4","0,4 - 0,6", "0,6 - 0,8", "
0,8 - 1,0")
breaks.PPSC=c(0,0.2,0.4,0.6,0.8,1)

PP_plot2010<-tm_shape(area_pond)+tm_fill(rbind("PP2010"),labels =
labels.PPSC,title=" ",palette=my.color,breaks=breaks.PPSC)+
tm_style("white",frame.lwd=0,alpha=0)+
tm_borders(alpha=.2,col="black")+
tm_layout(legend.text.color='red',legend.text.size =0.8,frame =
FALSE,legend.title.color="blue",compass.type="arrow",legend.
show=F,legend.only=F,legend.outside=T,legend.outside.position
="bottom",legend.title.fontface ="bold",legend.position=c
(0.75,0.7),title = c("2010"),title.fontface = "bold",title.
size = 1.5,title.position=c(0.2,0.91),fontfamily = "serif") +
tm_shape(bairros) + tm_borders("black", lwd = 1.5)

PP_plot2016<-tm_shape(area_pond)+tm_fill("PP2016",labels = labels
.PPSC,title="Probabilidade a \nposteriori",palette=my.color,
breaks=breaks.PPSC,between.margin=c(0,0,0,1)
)+tm_style("white",frame.lwd=0,alpha=0) +
tm_legend(position=c(0.5,0.2)) +
tm_compass(position=c(0.9,0.77),text.size = 1,lwd = 1,show.
labels = 1,just="center")+
tm_scale_bar(position=c(0.5,0.11),text.size =3,lwd = 1,breaks=c
(0,5,10,15)) + tm_borders(alpha=.2,col="black",lwd=1) +
tm_layout(legend.text.color='black',legend.text.size =1.2,frame
= F,legend.width = 3,legend.height = 5,legend.title.color="
black",legend.title.size=1.6,compass.type="arrow",legend.show
=T,legend.only=F,legend.outside=F,legend.outside.position="
bottom",legend.title.fontface ="bold",legend.position=c
(1.12,0.16),title = "2016",title.fontface = "bold",title.size
= 1.5,title.position=c(0.2,0.91),fontfamily = "serif") + tm_
shape(bairros) + tm_borders("black", lwd = 1.5)

PP_plot<-tmap_arrange(PP_plot2010,PP_plot2011,PP_plot2012,PP_
plot2013,PP_plot2014,PP_plot2015,PP_plot2016,nrow=2,ncol=4)

tmap_save(PP_plot,"Figuras/mapa_probabilidade_posteriori_SC.png")

# Distribuição dos resíduos e Índice de Moran
sifilis.pred<- mod3$summary.fitted.values$mean

```

```

standardized.residuals<- (d$SIR - sifilis.pred)/sqrt(sifilis.pred
)
d$resid<-standardized.residuals

resid<-data.frame(cbind(d[1:310,1], matrix(d$resid,byrow=F,ncol
=7)))
colnames(resid)<-colnames(resid)<-c("ID_AP","resid2010","
resid2011","resid2012","resid2013","resid2014","resid2015","
resid2016")

export(resid,"resid_poisson_congenita.csv")

data<-read.csv2("resid_poisson_congenita.csv",h=F,sep=" ",dec=","
)[-c(1),]
colnames(data)<-c("ID_AP","resid2010","resid2011","resid2012","
resid2013","resid2014","resid2015","resid2016")

area_pond@data$resid2010<-as.numeric(data$resid2010)
area_pond@data$resid2011<-as.numeric(data$resid2011)
area_pond@data$resid2012<-as.numeric(data$resid2012)
area_pond@data$resid2013<-as.numeric(data$resid2013)
area_pond@data$resid2014<-as.numeric(data$resid2014)
area_pond@data$resid2015<-as.numeric(data$resid2015)
area_pond@data$resid2016<-as.numeric(data$resid2016)

set.seed(987653)
mc1<-morán.mc(cbind(area_pond$resid2010) , nb2listw(temp, style =
"W"), nsim = 9999)
mc2<-morán.mc(cbind(area_pond$resid2011) , nb2listw(temp, style =
"W"), nsim = 9999)
mc3<-morán.mc(cbind(area_pond$resid2012) , nb2listw(temp, style =
"W"), nsim = 9999)
mc4<-morán.mc(cbind(area_pond$resid2013) , nb2listw(temp, style =
"W"), nsim = 9999)
mc5<-morán.mc(cbind(area_pond$resid2014) , nb2listw(temp, style =
"W"), nsim = 9999)
mc6<-morán.mc(cbind(area_pond$resid2015) , nb2listw(temp, style =
"W"), nsim = 9999)
mc7<-morán.mc(cbind(area_pond$resid2016) , nb2listw(temp, style =
"W"), nsim = 9999)

ANO<-c(2010:2016)

```

```

indice_moran_residuos_SC<-cbind(rbind(mc1$statistic,mc2$statistic
  ,mc3$statistic,mc4$statistic,mc5$statistic,mc6$statistic,mc7$
  statistic),rbind(mc1$p.value,mc2$p.value,mc3$p.value,mc4$p.
  value,mc5$p.value,mc6$p.value,mc7$p.value))

indice_moran_residuos_SC<-cbind(ANO,indice_moran_residuos_SC)
colnames(indice_moran_residuos_SC)<-c("Ano","Índice de Moran","
  Valor p")

xtable(indice_moran_residuos_SC,digits=4)

# Distribuição para os Resíduos do modelo Final
pdf("Figuras/graficos_moran_inla.pdf")
par(mfrow=c(3,3),mar=c(4,5,3,1),mgp=c(3, 0.7, 0),bg="white")
plot(mc1,cex.lab=1.5,cex.axis=1.2, main="2010",las=1,lwd=2,col="
  blue",xlab="Índice de Moran",ylab="Densidade",family="Times")
plot(mc2,cex.lab=1.5,cex.axis=1.2, main="2011",las=1,lwd=2,col="
  blue",xlab="Índice de Moran",ylab="Densidade",family="Times")
plot(mc3,cex.lab=1.5,cex.axis=1.2, main="2012",las=1,lwd=2,col="
  blue",xlab="Índice de Moran",ylab="Densidade",family="Times")
plot(mc4,cex.lab=1.5,cex.axis=1.2, main="2013",las=1,lwd=2,col="
  blue",xlab="Índice de Moran",ylab="Densidade",family="Times")
plot(mc5,cex.lab=1.5,cex.axis=1.2, main="2014",las=1,lwd=2,col="
  blue",xlab="Índice de Moran",ylab="Densidade",family="Times")
plot(mc6,cex.lab=1.5,cex.axis=1.2, main="2015",las=1,lwd=2,col="
  blue",xlab="Índice de Moran",ylab="Densidade",family="Times")
plot(mc7,cex.lab=1.5,cex.axis=1.2, main="2016",las=1,lwd=2,col="
  blue",xlab="Índice de Moran",ylab="Densidade",family="Times")
dev.off()

# Gráficos de dispersão de Moran para os resíduos
pdf("Figuras/scaterplot_moran_inla1.pdf")
par(mfrow=c(2,2),mar=c(4,5,3,1),mgp=c(3, 0.7, 0),bg="white")
moran.plot(area_pond$resid2010, nb2listw(temp, style = "W"),
  labels=as.character(afcon$name),pch=1,ylim=c(-1,1.5),xlim=c
  (-1,2.5),cex.lab=1.5,cex.axis=1.2,main="2010",las=1,lwd=2,col=
  "blue",xlab="Resíduos",ylab="",family="Times")

moran.plot(area_pond$resid2011, nb2listw(temp, style = "W"),
  labels=as.character(afcon$name),pch=1,ylim=c(-1,1.2),xlim=c
  (-1,3),cex.lab=1.5,cex.axis=1.2,main="2011",las=1,lwd=2,col=
  "blue",xlab="Resíduos",ylab="",family="Times")

```



```

moran.plot(area_pond$resid2012, nb2listw(temp, style = "W"),
  labels=as.character(afcon$name), pch=19, ylim=c(-1,0.8), xlim=c
(-1.2,1.5), cex.lab=1.5, cex.axis=1.2, main="2012", las=1, lwd=2,
col="blue", xlab="Resíduos", ylab="", family="Times")

moran.plot(area_pond$resid2013, nb2listw(temp, style = "W"),
  labels=as.character(afcon$name), pch=19, ylim=c(-1,0.8), xlim=c
(-1.5,2.5), cex.lab=1.5, cex.axis=1.2, main="2013", las=1, lwd=2,
col="blue", xlab="Resíduos", ylab="", family="Times")
dev.off()

pdf("Figuras/scaterplot_moran_inla2.pdf")
par(mfrow=c(2,2), mar=c(4,5,3,1), mgp=c(3, 0.7, 0), bg="white")

moran.plot(area_pond$resid2014, nb2listw(temp, style = "W"),
  labels=as.character(afcon$name), pch=19, ylim=c(-1,0.8), xlim=c
(-1,2), cex.lab=1.5, cex.axis=1.2, main="2014", las=1, lwd=2, col="
blue", xlab="Resíduos", ylab="", family="Times")

moran.plot(area_pond$resid2015, nb2listw(temp, style = "W"),
  labels=as.character(afcon$name), pch=19, ylim=c(-1,1.5), xlim=c
(-2,3), cex.lab=1.5, cex.axis=1.2, main="2015", las=1, lwd=2, col="
blue", xlab="Resíduos", ylab="", family="Times")

moran.plot(area_pond$resid2016, nb2listw(temp, style = "W"),
  labels=as.character(afcon$name), pch=19, ylim=c(-1,0.6), xlim=c
(-1.5,2), cex.lab=1.5, cex.axis=1.2, main="2016", las=1, lwd=2, col="
blue", xlab="Resíduos", ylab="", family="Times")
dev.off()

# Gráficos das médias a posteriores dos efeitos fixos
pdf("grafico_densidade.pdf")
par(mfrow=c(2,3), mar=c(4,5,3,1), mgp=c(3, 0.7, 0), bg="white")
plot(mod3$marginals.fixed[[1]], type = "l", col="blue", lwd=2,
  family="Times",
  xlab = expression(alpha), cex.lab=1.5, cex.axis=1.5,
  ylab = expression(paste(pi, "(", alpha, " | ", bold(y), ")")
))
mtext(sprintf("Média = %.3f    SD = %.3f", mod3$summary.fixed[1,1]
, mod3$summary.fixed[1,2] ), side=3, cex = 1, line = 0.5, family="
Times")

```

```

plot(mod3$marginals.fixed[[2]], type = "l", col="blue",lwd=2,
     family="Times",
     xlab = expression(beta[1]) , cex.lab=1.5,cex.axis=1.5,
     ylab = expression(paste(pi, "(", beta[1], " | ", bold(y), "
       )))
mtext(sprintf("Média = %.3f    SD = %.3f", mod3$summary.fixed
  [2,1], mod3$summary.fixed[2,2]),side=3,cex =1,line = 0.5,family
  ="Times")

plot(mod3$marginals.fixed[[3]], type = "l", col="blue",lwd=2,
     family="Times",
     xlab = expression(beta[2]), cex.lab=1.5,cex.axis=1.5,
     ylab = expression(paste(pi, "(", beta[2], " | ", bold(y), "
       )))
mtext(sprintf("Média = %.3f    SD = %.3f", mod3$summary.fixed[3,1]
  ,mod3$summary.fixed[3,2] ),side=3,cex =1,line = 0.5,family="
  Times")

plot(mod3$marginals.fixed[[4]], type = "l", col="blue",lwd=2,
     family="Times",
     xlab = expression(beta[3]), cex.lab=1.5,cex.axis=1.5,
     ylab = expression(paste(pi, "(", beta[3], " | ", bold(y), "
       )))
mtext(sprintf("Média = %.3f    SD = %.3f", mod3$summary.fixed[4,1]
  ,mod3$summary.fixed[4,2] ),side=3,cex =1,line = 0.5,family="
  Times")

plot(mod3$marginals.fixed[[5]], type = "l", col="blue",lwd=2,
     family="Times",
     xlab = expression(beta[4]), cex.lab=1.5,cex.axis=1.5,
     ylab = expression(paste(pi, "(", beta[4], " | ", bold(y), "
       )))
mtext(sprintf("Média = %.3f    SD = %.3f", mod3$summary.fixed[5,1]
  ,mod3$summary.fixed[5,2] ),side=3,cex =1,line = 0.5,family="
  Times")
dev.off()

```

G *Script* em linguagem R utilizado nas análises do Capítulo 4

A seguir apresentamos os códigos em R utilizados no ajuste dos modelos considerando-se a distribuição de Poisson. Para a distribuição Binomial Negativa, o procedimento é análogo, substituindo-se `family="poisson"` por `family="nb"`.

```
# Pacotes necessários
library(rsq)
library(ggplot2)
library(cowplot)
library(spdep)
library(gridExtra)
library(Metrics)
library(tidyverse)
library(rio)
library(xtable)

area_pond<-shapefile("areas_ponderacao_SP.shp")
bairros <- shapefile("bairros_SP.shp")

d<-read.csv("dados_sifilis_AP.csv")
d[,c(6:13)] <- scale(d[,c(6:13)],scale =T)

# Obtendo o valor esperado
d$E <- d$NV*sum(d$y_sc)/sum(d$NV)

#----- Modelos GAM - Poisson -----
gam_mrf_p0 <- gam(y_sc ~ offset(log(E))+prop_preta+PROP_NV_MAE_
  ADOLES+prop_m_sem_inst_fund18a24+PROP_FERTIL+ GINI+T_BANAGUA+
  RDPC, data = d, method = 'REML', family = poisson)

gam_mrf_p0_spat <- gam(y_sc ~ offset(log(E))+prop_preta+PROP_NV_
  MAE_ADOLES+prop_m_sem_inst_fund18a24+PROP_FERTIL+GINI+T_BANAGUA
```

```

+RDPC+s(xcoord,ycoord,bs="tp",k=310), data = d, method = 'REML',
, family = poisson)

gam_mrf_p0_temp <- gam(y_sc ~ offset(log(E))+prop_preta+PROP_NV_
MAE_ADOLES+prop_m_sem_inst_fund18a24+PROP_FERTIL+GINI+T_BANAGUA
+RDPC+s(Ano,bs="cr", k=7,m=6), data = d, method = 'REML',
family = poisson)

gam_mrf_p0_spat_temp_s <- gam(y_sc ~ offset(log(E))+prop_preta+
PROP_NV_MAE_ADOLES+prop_m_sem_inst_fund18a24+PROP_FERTIL+GINI+T
_BANAGUA+RDPC+s(xcoord,ycoord,bs="tp",k=310)+s(Ano,bs="cr", k
=7,m=6), data = d, method = 'REML', family = poisson )

gam_mrf_p0_spat_temp<-gam(y_sc ~ offset(log(E))+prop_preta+PROP_
NV_MAE_ADOLES+prop_m_sem_inst_fund18a24+PROP_FERTIL+GINI+T_
BANAGUA+RDPC+s(xcoord,ycoord,Ano, bs = 'tp',k=310, m=2), data =
d, method = 'REML', family = poisson)

gam_mrf_p1 <- gam(y_sc ~ offset(log(E))+prop_preta+PROP_NV_MAE_
ADOLES+prop_m_sem_inst_fund18a24+PROP_FERTIL+GINI+T_BANAGUA+
RDPC+s(xcoord,ycoord,bs="tp",k=310)+s(Ano,bs="cr", k=7,m=6)+s(
xcoord,ycoord,Ano, bs = 'tp',k=310, m=2), data = d, method = '
REML', family = poisson)

gam_mrf_p2 <- gam(y_sc ~ offset(log(E))+prop_preta+prop_m_sem_
inst_fund18a24+PROP_FERTIL+GINI+T_BANAGUA+RDPC+s(xcoord,ycoord,
bs="tp",k=310)+s(Ano,bs="cr", k=7,m=6)+s(xcoord,ycoord,Ano, bs
= 'tp',k=310, m=2), data = d, method = 'REML', family = poisson
)

gam_mrf_p3 <- gam(y_sc ~offset(log(E))+prop_preta+prop_m_sem_inst
_fund18a24+GINI+T_BANAGUA+RDPC+s(xcoord,ycoord,bs="tp",k=310)+s
(Ano,bs="cr", k=7,m=6)+s(xcoord,ycoord,Ano, bs = 'tp',k=310, m
=2), data = d, method = 'REML', family = poisson)

gam_mrf_p4 <- gam(y_sc ~ offset(log(E))+prop_preta+prop_m_sem_
inst_fund18a24+GINI+RDPC+s(xcoord,ycoord,bs="tp",k=310)+s(Ano,
bs="cr", k=7,m=6)+s(xcoord,ycoord,Ano, bs = 'tp',k=310, m=2),
data = d, method = 'REML', family = poisson)

gam_mrf_p5 <- gam(y_sc ~ offset(log(E))+prop_m_sem_inst_fund18a24
+GINI+RDPC+s(xcoord,ycoord,bs="tp",k=310)+s(Ano,bs="cr", k=7,m

```

```

=6)+s(xcoord,ycoord,Ano, bs = 'tp',k=310, m=2), data = d,
method = 'REML', family = poisson)

#----- AIC e BIC -----
AIC_poi<-AIC(gam_mrf_p0,gam_mrf_p0_spat,gam_mrf_p0_temp,
gam_mrf_p0_spat_temp,gam_mrf_p0_spat_temp_s,gam_mrf_p1,
gam_mrf_p2,gam_mrf_p3,gam_mrf_p4,gam_mrf_p5)[2]

BIC_poi<-BIC(gam_mrf_p0,gam_mrf_p0_spat,gam_mrf_p0_temp,
gam_mrf_p0_spat_temp,gam_mrf_p0_spat_temp_s,gam_mrf_p1,
gam_mrf_p2,gam_mrf_p3,gam_mrf_p4,gam_mrf_p5)[2]

AIC_nb<-AIC(gam_mrf_nb0,gam_mrf_nb0_spat,gam_mrf_nb0_temp,
gam_mrf_nb0_spat_temp,gam_mrf_nb0_spat_temp_s,gam_mrf_nb1,
gam_mrf_nb2,gam_mrf_nb3,gam_mrf_nb4,gam_mrf_nb5)[2]

BIC_nb<-BIC(gam_mrf_nb0,gam_mrf_nb0_spat,gam_mrf_nb0_temp,
gam_mrf_nb0_spat_temp,gam_mrf_nb0_spat_temp_s, gam_mrf_nb1,
gam_mrf_nb2, gam_mrf_nb3,gam_mrf_nb4,gam_mrf_nb5)[2]

# ---- R2
R2_poi<-rbind(rsq(gam_mrf_p0,adj=FALSE,type="sse"),
rsq(gam_mrf_p0_spat),rsq(gam_mrf_p0_temp,adj=FALSE,type="sse"),
rsq(gam_mrf_p0_spat_temp,adj=FALSE,type="sse"),
rsq(gam_mrf_p0_spat_temp_s,adj=FALSE,type="sse"),
rsq(gam_mrf_p1,adj=FALSE,type="sse"),rsq(gam_mrf_p2,adj=FALSE,
type="sse"),
rsq(gam_mrf_p3,adj=FALSE,type="sse"),rsq(gam_mrf_p4,adj=FALSE,
type="sse"),
rsq(gam_mrf_p5,adj=FALSE,type="sse"))

#--- REML
REML_poi<-rbind(summary(gam_mrf_p0)$sp.criterion,summary(gam_mrf_
p0_spat)$sp.criterion,
summary(gam_mrf_p0_temp)$sp.criterion,summary(gam_mrf_p0_spat_
temp)$sp.criterion,
summary(gam_mrf_p0_spat_temp_s)$sp.criterion,summary(gam_mrf_p1)$
sp.criterion,
summary(gam_mrf_p2)$sp.criterion,summary(gam_mrf_p3)$sp.criterion
,

```

```

summary(gam_mrf_p4)$sp.criterion,summary(gam_mrf_p5)$sp.criterion
)

# Deviance explicada
dev.expl_poi<-rbind(summary(gam_mrf_p0)$dev.expl,summary(gam_mrf_
p0_spat)$dev.expl,
summary(gam_mrf_p0_temp)$dev.expl,summary(gam_mrf_p0_spat_temp)$
dev.expl,
summary(gam_mrf_p0_spat_temp_s)$dev.expl,summary(gam_mrf_p1)$dev.
expl,
summary(gam_mrf_p2)$dev.expl,summary(gam_mrf_p3)$dev.expl,
summary(gam_mrf_p4)$dev.expl,summary(gam_mrf_p5)$dev.expl)

# --- RMSE
RMSE_poi<-rbind(
rmse(d$y_sc,fitted(gam_mrf_p0)),rmse(d$y_sc,fitted(gam_mrf_p0_
spat)),
rmse(d$y_sc,fitted(gam_mrf_p0_temp)),rmse(d$y_sc,fitted(gam_mrf_
p0_spat_temp)),
rmse(d$y_sc,fitted(gam_mrf_p0_spat_temp_s)),rmse(d$y_sc,fitted(
gam_mrf_p1)),
rmse(d$y_sc,fitted(gam_mrf_p2)),rmse(d$y_sc,fitted(gam_mrf_p3)),
rmse(d$y_sc,fitted(gam_mrf_p4)),rmse(d$y_sc,fitted(gam_mrf_p5)))

# --- MAE
MAE_poi<-rbind(mae(d$y_sc,fitted(gam_mrf_p0)),mae(d$y_sc,fitted(
gam_mrf_p0_spat)),
mae(d$y_sc,fitted(gam_mrf_p0_temp)),mae(d$y_sc,fitted(gam_mrf_p0_
spat_temp)),
mae(d$y_sc,fitted(gam_mrf_p0_spat_temp_s)),mae(d$y_sc,fitted(gam_
mrf_p1)),
mae(d$y_sc,fitted(gam_mrf_p2)),mae(d$y_sc,fitted(gam_mrf_p3)),
mae(d$y_sc,fitted(gam_mrf_p4)),mae(d$y_sc,fitted(gam_mrf_p5)))

resumo_gam_poisson<-round(cbind(AIC_poi,BIC_poi,REML_poi,RMSE_poi
,MAE_poi,R2_poi,dev.expl_poi),3)

```