UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE ESTATÍSTICA

MARIA LUISA GOMES DOS REIS

MODELAGEM LINEAR GENERALIZADA COM EFEITO ESPACIAL: UM ESTUDO COM DADOS EDUCACIONAIS DO ESTADO DE MINAS GERAIS

BELO HORIZONTE

2022

MARIA LUISA GOMES DOS REIS

MODELAGEM LINEAR GENERALIZADA COM EFEITO ESPACIAL: UM ESTUDO COM DADOS EDUCACIONAIS DO ESTADO DE MINAS GERAIS

Monografia apresentada ao Departamento de Estatística da Universidade Federal de Minas Gerais (UFMG), como parte dos requisitos necessários à obtenção do título de bacharel em Estatística.

Orientador: Vinícius Diniz Mayrink

BELO HORIZONTE

2022

Agradecimentos

Agradeço ao meu orientador, professor Vinícius, pela orientação desde o projeto de Iniciação Científica que serviu de base para esta monografia e pela paciência durante todo esse processo.

Da Estatística, agradeço aos meus amigos que me ajudaram a aguentar todas as disciplinas sem colapsar. Agradeço a todos os professores com os quais tive aulas, aos que me orientaram nas atividades que realizei e àqueles com os quais participei da Câmara. Agradeço especialmente ao professor Cristiano, pela orientação no projeto de Iniciação Científica no qual construí um aplicativo Shiny; não fosse pela experiência adquirida nesse projeto, eu não teria conseguido o estágio que me proporcionou o emprego que tenho hoje.

Por fim, agradeço aos meus pais por terem me dado uma ótima educação e muito tempo de folga, o que me permitiu estudar e vir parar aqui. Agradeço ao Alexander por ter me ajudado a estudar para inúmeras matérias. Faço menção especial ao Lucas, por ter me distraído em várias das aulas online enquanto brincava no meu quarto. Finalmente, dedico esta monografia à Shasta, que passou dois anos do curso deitada na mesa ao lado do computador, estudando comigo.

Resumo

Palavras-chave: Modelos Lineares Generalizados; Estatística Bayesiana; Estatística Espacial; Taxa de abandono escolar.

Em casos em que os dados sob estudo não se adequam à distribuição Normal e a transformação dos valores traz complexidade para a interpretação dos resultados, é adequado o uso de Modelos Lineares Generalizados (MLG). Os dados estudados neste trabalho dizem respeito às taxas de abandono escolar para os municípios do estado de Minas Gerais nos anos de 2010, 2015 e 2020. O objetivo principal foi avaliar quais fatores são capazes de explicar a taxa de abandono escolar neste estado, além de verificar se estes fatores mudaram ao longo dos anos. A análise descritiva e visual dos dados indicou forte assimetria à esquerda, com muitos valores próximos de zero. O uso do MLG Gama se mostrou propício, mas, por se tratar de uma taxa que varia de 0% a 100%, o MLG Beta seria mais apropriado. Como as informações das taxas de abandono são por município, foi possível trabalhar com a Estatística Espacial, adicionando ao modelo uma associação espacial entre municípios vizinhos. Uma vez que a estrutura espacial foi vastamente explorada neste estudo, optou-se por trabalhar com a Estatística Bayesiana. Antes de modelar os dados reais das taxas foi conduzido um estudo simulado para avaliar qual o melhor modelo para ajustá-las. Após este estudo, concluiu-se que as taxas de abandono para os municípios de Minas Gerais poderiam ser ajustadas assumindo efeito espacial tanto na estrutura da média quanto na do parâmetro de dispersão. Dentre os blocos de ensino, o Ensino Médio foi o que apresentou menor concentração de valores próximos de zero e por isso analisou-se apenas as taxas para este bloco. Apesar disso, a frequência de valores iguais a zero na variável resposta poderia impactar negativamente no ajuste e por isso foram obtidas mais 3 amostras, uma para cada ano sob estudo, que excluíam os municípios com taxa igual a zero, a fim de avaliar se haveria grandes diferenças nos resultados de um grupo ou outro de amostras. No geral, os fatores capazes de explicar significativamente as taxas de abandono do Ensino Médio dos municípios de Minas Gerais foram a taxa de escolas na área urbana que, quanto maior, tende a diminuir a taxa de abandono; e a taxa de escolas públicas no município que, quanto maior, tende a aumentar a taxa de abandono escolar. Os resultados se apresentaram consistentes ao longo dos anos e para as amostras com e sem as taxas iguais a zero, indicando que as causas por trás das taxas de abandono do Ensino Médio não tenham mudado ao longo da última década.

Abstract

Key-words: Generalized Linear Models; Bayesian Statistics; Spatial Statistics; School abandonment rate.

In cases where the data under study does not correspond to a Normal distribution and a transformation upon the values brings complexity to the interpretation of the results, the use of Generalized Linear Models (GLM) is more adequate. The data studied in this work refers to the school abandonment rate from the cities of the brazilian state of Minas Gerais in the years 2010, 2015, and 2020. This work's main goal was to evaluate which factors were capable of explaining the school abandonment rate in this state, besides checking whether these factors changed through the years. The visual and descriptive analysis of the data showed strong left asymmetry, with many values close to zero. The use of a GLM Gamma shows itself favourable, but, as the values refer to a rate that varies between 0% and 100%, the GLM Beta would suit the data best. As each rate corresponds to a city, it was possible to work with Spatial Statistics, adding to the model an spatial association between neighbour cities. Since this work explored vastly the spatial structure, we chose to work with Bayesian Statistics. Before modelling the real abandonment rate data it was conducted a simulated study to evaluate what was the best model to fit them. After this study, we concluded that the abandonment rate for the cities of Minas Gerais could be fitted assuming spatial effects both in the structure of the mean as well as in the structure of the dispersion parameter. Among the teaching blocks, the High School was the one with the smallest concentration of values close to zero, and therefore only this block was considered for the analysis. Despite that, the frequency of values equal to zero in the response variable could bring problems to the fitting. For that matter, 3 more samples were obtained, one for each year under study, that excluded the cities with rate equal to zero, in order to evaluate if there would be great differences in the results from one or another group of samples. In general, the factors capable of significantly explaining the High School abandonment rates in the cities of Minas Gerais were the rate of schools in the urban area which, the larger, tends to decrease the abandonment rate; and the rate of public schools in the city, which, the larger, tends to increase the school abandonment rate. The results showed to be consistent through the years and for the samples that included and that excluded the rates equal to zero, indicating that causes behind the High School abandonment rate have not changed over the last decade.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

Uma das técnicas mais usuais para avaliar quais aspectos influenciam em uma variável de interesse é a Regressão Linear Múltipla (Draper, Smith; 1981). No entanto, modelos de regressão linear só se aplicam sob a suposição de normalidade dos dados. Uma alternativa para situações em que a variável de interesse não apresenta comportamento Normal é a aplicação de transformações nesta variável, na expectativa de que ela se adeque à distribuição Gaussiana. Outra alternativa mais apropriada é o uso de Modelos Lineares Generalizados (MLG) (Dobson, Barnett; 2008). Tais modelos permitem fazer a análise de regressão com base em distribuições que se adequem aos dados sem necessidade de modificá-los de alguma forma. Para dados assimétricos à esquerda, por exemplo, MLG's Gama fornecerão resultados melhores se comparados à aplicação da Regressão Linear Gaussiana. Outro modelo propício para dados assimétricos é o modelo Beta (Ferrari, Cribari-Neto; 2004), adequado a casos em que a variável resposta apresenta valores limitados ao intervalo de 0 a 1, seja por meio da padronização dos dados ou por apresentar estes limites naturalmente. Um caso como esse ocorre quando os valores de interesse são taxas ou percentuais.

Existem casos em que os dados sob estudo apresentam não somente covariáveis explicativas, mas também a informação geográfica das observações coletadas. Observações provenientes de regiões vizinhas podem estar correlacionadas, e ignorar esta correlação pode acarretar problemas como a invalidade das inferências obtidas. Sendo assim, nos casos em que a informação geográfica dos dados está disponível é interessante incorporar esta informação na modelagem. Modelos Lineares Generalizados permitem a incorporação do efeito espacial na modelagem de um ou mais parâmetros que descrevem a distribuição utilizada. Assim, a informação geográfica pode ser facilmente levada em conta no estudo de dados que são coletados em regiões próximas que compõem uma área limitada.

Um dos tipos de dados que são comumente coletados por região, como para estados ou municípios, são os dados educacionais. A forma mais conhecida de avaliação do ensino básico no Brasil é o Exame Nacional do Ensino Médio (ENEM). No entanto, por seu resultado ser utilizado pelos estudantes como forma de ingresso em universidades públicas, ele acaba sendo uma forma de avaliação enviesada, uma vez que os alunos deliberadamente se preparam para

prestar esse exame. Uma forma mais acurada de avaliação do ensino e das escolas brasileiras pode ser feita por meio do Censo Escolar, que é pouco conhecido, apesar de ser a pesquisa estatística educacional brasileira de maior importância.

O Censo escolar é realizado anualmente e tem como objetivo coletar dados sobre as instituições de ensino e sobre a situação acadêmica de seus alunos matriculados. Com essas informações é possível calcular os chamados Indicadores Educacionais, disponibilizados pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). Dentre esses indicadores, encontram-se as Taxas de Rendimento escolar, que se dividem entre as taxas de Aprovação, Reprovação e Abandono. Cada uma dessas taxas representa o total de alunos na situação em questão dividido pelo total de matrículas no ano do censo. O Inep considera o número total de matrículas como a soma de todos os alunos aprovados, reprovados e que abandonaram as aulas. Assim, essas taxas são compostas por valores percentuais, variando de 0 a 100. Para o cálculo da taxa de abandono, considera-se que o aluno matriculado abandonou a escola quando deixou de frequentar as aulas, executando-se os motivos de falecimento ou transferência de instituição de ensino. O portal do Inep disponibiliza planilhas com os valores anuais dessas taxas para Brasil, Grandes Regiões, Unidades Federativas, municípios e escolas. Diante desses indicadores, propôs-se um estudo sobre o abandono escolar nos municípios do estado de Minas Gerais.

As planilhas de taxas para escolas contemplam os dados para cada escola de todo o país, fornecendo o ano da coleta da informação, a Região, Unidade Federativa, código e nome do município no qual a escola se encontra, código e nome da escola e classificações quanto à localização e à dependência administrativa. As taxas de abandono estavam disponíveis com relação ao total do Ensino Fundamental, ao total dos anos iniciais (Ensino Fundamental I, que corresponde do 1º ao 5º ano do ensino fundamental), ao total dos anos finais (Ensino Fundamental II, que corresponde do 6º ao 9º ano do ensino fundamental) e com relação ao Ensino Médio (1ª à 4ª série, adicionalmente com a categoria não-seriado). Além disso, esses indicadores também estavam disponíveis para cada um desses anos escolares separadamente.

A fim de avaliar quais fatores podem influenciar na taxa de abandono de determinado município, os dados foram complementados com as informações do IDHM Renda, população e PIB *per capita* dos municípios. Os dados sobre o Índice de Desenvolvimento Humano Municipal Renda (IDHM Renda) foram obtidos no portal Atlas Brasil, referente ao censo do

ano de 2010. Os dados da população para o ano de 2010 foram obtidos no portal de dados abertos do Instituto de Pesquisa Econômica Aplicada (Ipea), enquanto as populações referentes a 2015 e 2020 foram oriundas das estimativas populacionais disponibilizadas no site do Instituto Brasileiro de Geografia e Estatística (IBGE). Por fim, o Produto Interno Bruto (PIB) *per capita* para os municípios foi obtido no portal IBGE com referência aos respectivos anos para 2010 e 2015. Para 2020 a referência mais recente disponibilizada para essa informação foi de 2018.

Uma vez que o estudo restringe-se aos municípios de Minas Gerais, com uma taxa por município, torna-se propício o uso do efeito espacial. As coordenadas geográficas utilizadas foram as referentes aos centróides de cada município (Prado, 2021).

De posse desses dados, o objetivo principal foi avaliar, por meio do uso de Modelos Lineares Generalizados com efeito espacial, quais fatores são capazes de explicar a taxa de abandono escolar nos municípios de Minas Gerais, inserindo, também, a informação geográfica na modelagem. Para a aplicação do efeito espacial as análises serão feitas sob a visão de inferência Bayesiana (Bolstad, 2007), investigando quais distribuições e quais tipos de modelagem melhor se adequam aos dados. Os dados utilizados neste estudo são referentes aos anos de 2020, 2015 e 2010 para os três blocos de ensino principais em cada ano. Para este estudo, considerou-se do 1º ao 5º anos do ensino fundamental como Ensino Fundamental I (EFI) e do 6º ao 9º como Ensino Fundamental II (EFII). Para o ensino médio, apesar dos dados incluírem informações sobre a 4ª série e sobre o ensino médio não-seriado, considerou-se para o Ensino Médio (EM) apenas da 1ª a 3ª série.

A Seção 2 deste relatório apresenta a análise descritiva e gráfica feita sobre os dados. A Seção 3 aborda a metodologia utilizada para as análises e validação dos resultados. A Seção 4 se divide em duas subseções de discussão e apresentação dos resultados, sendo que seção 4.1 diz respeito aos resultados do estudo simulado e a 4.2 aos resultados dos dados reais das taxas de abandono escolar em Minas Gerais. Por fim, a Seção 5 apresenta as conclusões da pesquisa.

2 Análise descritiva

As planilhas das taxas de abandono por escola eram as que apresentavam mais informações. Além de fornecer o nome e o código da instituição de ensino, bem como seu município e suas respectivas taxas de abandono, estas planilhas também continham algumas informações de caracterização da escola, sendo estas a dependência administrativa e a localização. A dependência administrativa poderia ser uma dentre quatro possíveis: Municipal, Estadual, Federal e Privada. Esta variável foi recategorizada para assumir apenas os valores Pública (referente às três primeiras dependências possíveis) e Privada. Com esta nova variável dicotômica para cada escola, foi possível calcular o percentual de escolas públicas para cada cidade, percentual este que foi denominado "Taxa públicas" e utilizado como covariável explicativa para as modelagens da taxa de abandono. Já a localização da escola se refere à região do município na qual a instituição se encontra, podendo ser na área Urbana ou na área Rural da cidade. Assim como foi feito para a dependência administrativa, esta variável dicotômica foi utilizada para a criação da covariável "Taxa urbana" para cada município.

Uma vez que os dados sobre o abandono estavam disponíveis por escola, foi necessário calcular a taxa de abandono para cada bloco de ensino para cada município. Optou-se por calcular a taxa média de abandono do respectivo bloco para cada escola, e então calcular a média sobre estes valores para obter a taxa de abandono para este bloco para o município no qual estas escolas se encontram. A fim de obter valores acurados para cada escola, foram consideradas válidas para o cálculo da média municipal apenas as que apresentavam pelo menos metade dos dados. Desta forma, para o bloco EFI, que contempla do 1º ao 5º ano escolar, foram consideradas as escolas com dados para 3 ou mais desses anos escolares; para o bloco EFII, entre os anos 6º e 9º, foram aceitas escolas com dados para 2 ou mais anos escolares; e para o bloco EM, contemplando da 1ª a 3ª série do Ensino Médio, foram aceitas apenas escolas com informações para 2 ou mais séries. Vale notar que as escolas consideradas para o cálculo das taxas dos diferentes blocos são independentes. Por exemplo, se uma escola apresenta apenas 2 taxas referentes a cada bloco de ensino, esta escola terá calculadas apenas as suas taxas para os blocos EFII e EM, e entrará no cálculo das taxas de abandono municipais para estes blocos apenas.

Após o cálculo das taxas por escola, foi calculada a taxa média de abandono para cada município. Antes de partir para a análise dos dados, no entanto, foram selecionados quais municípios fariam parte das amostras. Estabeleceu-se que o número mínimo de escolas válidas, ou seja, escolas que apresentavam a maioria da informação para o bloco de ensino em questão, por cidade, seria 3. Decidiu-se por este limiar, pois, nos casos de municípios com apenas 1 ou 2 escolas válidas, o cálculo da taxa de abandono média destes municípios seria fortemente influenciado por poucas escolas, podendo resultar em um valor não verdadeiramente representativo daquele município.

Este método foi utilizado para a definição das amostras para os blocos de ensino EFI, EFII e EM para os três anos considerados, 2010, 2015 e 2020. A Tabela 1, a seguir, apresenta o número de municípios considerados para cada amostra, bem como o número médio de escolas por cidade, o número médio de escolas válidas por cidade, o número de escolas válidas em todo o estado de Minas Gerais e quantas destas foram consideradas válidas para a respectiva amostra.

Tabela 1 - Descrição das amostras.

Ano	Ensino	Nº de municípios considerados	Nº médio de escolas por município	Nº médio de escolas válidas por município	Nº total de escolas	Nº total de escolas válidas
2020	EFI	617	12,4	9,1	10.540	7.760
	EFII	419	12,4	5,4	10.540	4.647
	EM	272	12,4	2,8	10.540	2.417
2015	EFI	652	13,2	10,2	11.222	8.664
	EFII	425	13,2	5,5	11.222	4.714
	EM	256	13,2	2,7	11.222	2.271
2010	EFI	692	14,8	11,9	12.608	10.186
	EFII	419	14,8	5,5	12.608	4.663
	EM	243	14,8	2,5	12.608	2.101

Além das covariáveis obtidas a partir dos bancos de dados sobre as taxas de abandono, taxa de escolas na área urbana e taxa de escolas públicas, os dados foram complementados com algumas variáveis referentes aos municípios. O IDHM Renda se restringe à escala de 0 a 1, assim como as taxas já mencionadas. Já a população e o PIB *per capita* apresentam valores muito grandes que podem afetar negativamente o desempenho dos modelos. Por este motivo optou-se por trabalhar com a transformação logarítmica destas duas variáveis.

A Tabela 2 a seguir apresenta as medidas resumo para as taxas de abandono e suas covariáveis para as amostras de 2020. Pode-se observar que para os três blocos de ensino a maior parte das taxas são iguais ou próximas de zero. Isso sugere que a distribuição dos dados

é assimétrica com cauda pesada à direita. Apesar dos três blocos apresentarem mais taxas próximas de zero, observa-se que do bloco de ensino EFI para o EFII, assim como do EFII para o EM, a taxa de abandono média aumenta cerca de um ponto percentual.

Tabela 2 - Medidas resumo para as amostras de 2020.

Ensino	Variável	Média	D.P.	Mín.	1ºQ	Mediana	3ºQ	Máx.	N
EFI	Taxa abandono (%)	0,53	1,70	0,00	0,00	0,00	0,19	20,00	617
	Taxa públicas (%)	83,24	27,14	2,70	80,00	100,00	100,00	100,00	
	Taxa urbana (%)	37,01	24,21	2,50	20,51	33,33	42,86	100,00	
	IDHM Renda	0,82	0,03	0,72	0,80	0,82	0,85	0,89	
	População	32.724,32	118.078,2	2.056	6.860	12.182	24.029	2.521.564	
	log(População)	9,57	1,03	7,63	8,83	9,41	10,09	14,74	
	PIB *per capita*	2,27	0,06	2,17	2,22	2,26	2,31	2,54	
	log(PIB *per capita*)	9,69	0,60	8,75	9,22	9,60	10,03	12,73	
EFII	Taxa abandono (%)	1,92	3,84	0,00	0,00	0,27	2,50	50,00	419
	Taxa públicas (%)	69,10	35,46	2,70	26,79	85,71	100,00	100,00	
	Taxa urbana (%)	47,61	29,39	2,50	25,00	38,46	75,00	100,00	
	IDHM Renda	0,83	0,03	0,74	0,80	0,83	0,85	0,89	
	População	44.782,83	141.730,4	2.646	10.765,50	18.193	34.485	2.521.564	
	log(População)	9,95	1,01	7,88	9,28	9,81	10,45	14,74	
	PIB *per capita*	2,27	0,06	2,17	2,22	2,27	2,31	2,53	
	log(PIB *per capita*)	9,74	0,61	8,75	9,24	9,71	10,09	12,5	
EM	Taxa abandono (%)	2,84	5,33	0,00	0,00	0,40	4,12	50	272
	Taxa públicas (%)	55,42	37,08	3,08	20	38,07	95,24	100	
	Taxa urbana (%)	60,02	28,86	6,99	33,33	63,64	86,9	100	
	IDHM Renda	0,83	0,03	0,74	0,81	0,84	0,86	0,89	
	População	62.880,92	173.285,9	2.646	16.121,75	26.316,5	52.494,5	2.521.564	
	log(População)	10,35	0,99	7,88	9,69	10,18	10,87	14,74	
	PIB *per capita*	2,28	0,06	2,17	2,23	2,29	2,32	2,53	
	log(PIB *per capita*)	9,82	0,61	8,75	9,31	9,84	10,18	12,5	

Na Tabela 3 podem ser observadas as medidas resumo para as amostras de 2015. Assim como ocorreu para o ano de 2020, as taxas de abandono tendem a estar mais próximas de zero e a taxa média tende a aumentar ao longo dos blocos de ensino, partindo do EFI até o EM.

Por fim, a Tabela 4 apresenta as descritivas para as amostras de 2010. Observamos que o comportamento destes dados é bem similar ao das amostras para os demais anos. Vale notar que para este ano as taxas médias de abandono, bem com suas medianas, são ligeiramente maiores que as observadas nos anos anteriores.

Tabela 3 - Medidas resumo para as amostras de 2015.

Ensino	Variável	Média	D.P.	Mín.	1ºQ	Mediana	3ºQ	Máx.	N
EFI	Taxa abandono (%)	0,30	0,93	0,00	0,00	0,00	0,20	11,10	652
	Taxa públicas (%)	83,35	28,24	1,96	80	100	100	100	
	Taxa urbana (%)	36,29	23,07	1,89	20,77	33,33	42,86	100	
	IDHM Renda	0,82	0,03	0,72	0,80	0,82	0,85	0,89	
	População	30.609,8	113.128,6	2.213	6.495,75	11.242,5	22.337	2.502.557	
	log(População)	9,51	1,02	7,70	8,78	9,33	10,01	14,73	
	PIB *per capita*	2,25	0,06	2,14	2,20	2,24	2,28	2,49	
	log(PIB *per capita*)	9,50	0,56	8,52	9,07	9,42	9,81	12,04	
EFII	Taxa abandono (%)	1,67	2,28	0,00	0,00	0,84	2,55	16,15	425
	Taxa públicas (%)	69,10	36,41	1,96	25	87,10	100	100	
	Taxa urbana (%)	46,24	28,79	2,44	23,08	38,89	66,67	100	
	IDHM Renda	0,83	0,03	0,74	0,80	0,83	0,85	0,89	
	População	43.309	138.493,1	3.487	10.620	18.014	33.082	2.502.557	
	log(População)	9,93	0,99	8,16	9,27	9,80	10,41	14,73	
	PIB *per capita*	2,26	0,06	2,14	2,21	2,26	2,29	2,48	
	log(PIB *per capita*)	9,57	0,57	8,52	9,09	9,54	9,91	11,94	
EM	Taxa abandono (%)	2,66	3,66	0,00	0,00	0,51	4,77	19,5	256
	Taxa públicas (%)	49,79	37,56	2,22	15,38	29,6	93,08	100	
	Taxa urbana (%)	58,97	29,11	5,00	32,03	62,02	85,32	100	
	IDHM Renda	0,83	0,03	0,74	0,81	0,84	0,86	0,89	
	População	64.316,1	175.372,7	4.983	17.816,75	27.123,5	54.614,25	2.502.557	
	log(População)	10,40	0,96	8,51	9,79	10,21	10,91	14,73	
	PIB *per capita*	2,27	0,06	2,14	2,22	2,27	2,31	2,48	
	log(PIB *per capita*)	9,67	0,59	8,52	9,23	9,67	10,05	11,94	

Tabela 4 - Medidas resumo para as amostras de 2010.

Ensino	Variável	Média	D.P.	Mín.	1ºQ	Mediana	3ºQ	Máx.	N
EFI	Taxa abandono (%)	0,65	1,38	0,00	0,00	0,14	0,86	18,34	692
	Taxa públicas (%)	83,42	29,03	1,89	82,22	100	100	100	
	Taxa urbana (%)	34,66	21,34	2,22	20,57	32,17	42,29	100	
	IDHM Renda	0,82	0,03	0,72	0,80	0,82	0,85	0,89	
	População	27.366,88	103.605,2	1.613	5.930,75	10.278	20.427	2.375.151	
	log(População)	9,41	1,01	7,39	8,69	9,24	9,92	14,68	
	PIB *per capita*	2,20	0,07	2,09	2,15	2,19	2,24	2,50	
	log(PIB *per capita*)	9,06	0,61	8,09	8,59	8,94	9,40	12,20	
EFII	Taxa abandono (%)	3,00	3,64	0,00	0,00	1,92	4,82	28,01	419
	Taxa públicas (%)	67,88	37,12	1,89	23,76	85,71	100	100	
	Taxa urbana (%)	42,96	27,22	2,22	22,32	35,29	57,14	100	
	IDHM Renda	0,83	0,03	0,74	0,80	0,83	0,85	0,89	
	População	41.020,24	131.399,4	2.962	10.260,5	17.345	31.237,5	2.375.151	
	log(População)	9,89	0,98	7,99	9,24	9,76	10,35	14,68	
	PIB *per capita*	2,21	0,07	2,09	2,15	2,21	2,25	2,44	
	log(PIB *per capita*)	9,13	0,63	8,09	8,60	9,07	9,52	11,49	
EM	Taxa abandono (%)	3,31	4,64	0,00	0,00	0,62	6,00	19,13	243
	Taxa públicas (%)	46,98	37,05	2,38	15,38	27,66	90,91	100	
	Taxa urbana (%)	56,73	28,91	6,99	29,59	57,14	81,92	100	
	IDHM Renda	0,83	0,03	0,74	0,81	0,84	0,86	0,89	
	População	62.694,75	169.344,6	4.804	17.656,5	26.922	53.648	2.375.151	
	log(População)	10,40	0,93	8,48	9,78	10,20	10,89	14,68	
	PIB *per capita*	2,22	0,07	2,09	2,17	2,23	2,27	2,41	
	log(PIB *per capita*)	9,27	0,65	8,09	8,72	9,29	9,69	11,13	

Os quartis para as taxas apresentados nas tabelas indicam a assimetria desses dados. Para confirmar esta suposição, foram gerados os histogramas e *box-plots* das taxas de abandono escolar para cada ano e ensino. Na Figura 1 podemos ver estes gráficos para o ano 2020.

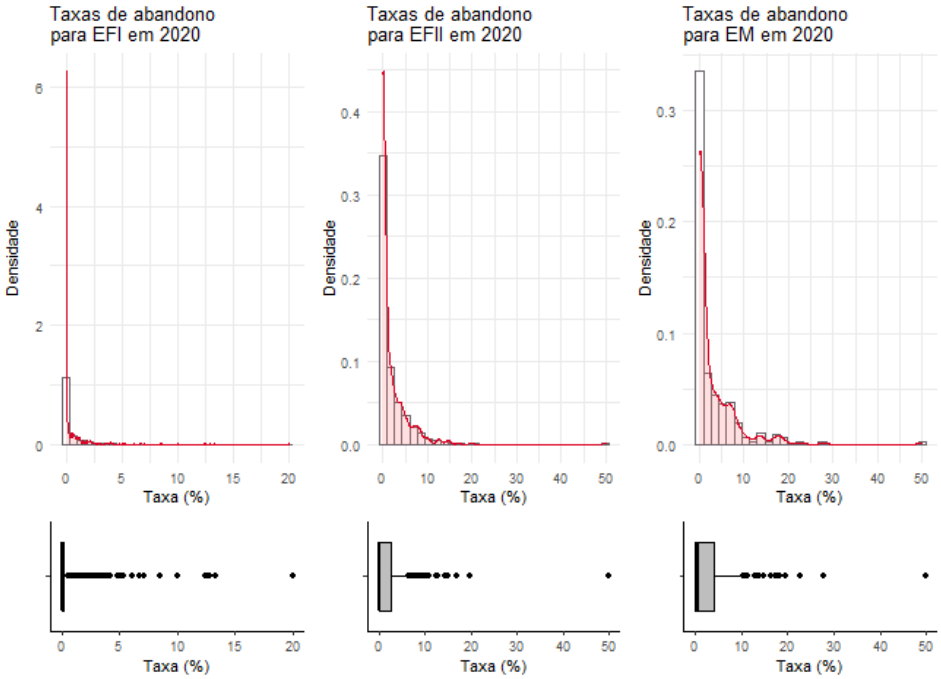

Figura 1 - Histogramas e *box-plots* das taxas de abandono em 2020.

Apesar das escalas serem diferentes, pode-se notar que o bloco EFI tem a maior concentração de valores próximos de 0, com seu *outlier* mais extremo sendo igual a 20%. As formas das distribuições para EFII e EM se assemelham, ambas apresentando a maior parte dos valores inferior a 5%, com a cauda mais pesada se estendendo até 20% e 30%. Vale notar que em ambos os casos existe um *outlier* mais extremo correspondente a uma taxa de abandono de 50%.

Pela Figura 2, referente às taxas de 2015, podemos ver que a forma dos dados para EFI se aproxima daquela para o bloco EFI em 2020. Os blocos EFII e EM neste ano também apresentam cauda pesada à direita, mas neste caso vemos a forma da distribuição mais claramente, devido à ausência de *outliers* extremos. Para EFII os valores de maior frequência são iguais a zero, enquanto para o bloco EM as taxas de maior frequência estão próximas de 5%. Vale notar que para o EM, apesar de em 2015 haver menos *outliers* em comparação com 2020, a mediana da taxa de abandono é aproximadamente 5%, enquanto que para 2020 ela está mais próxima de 0%.

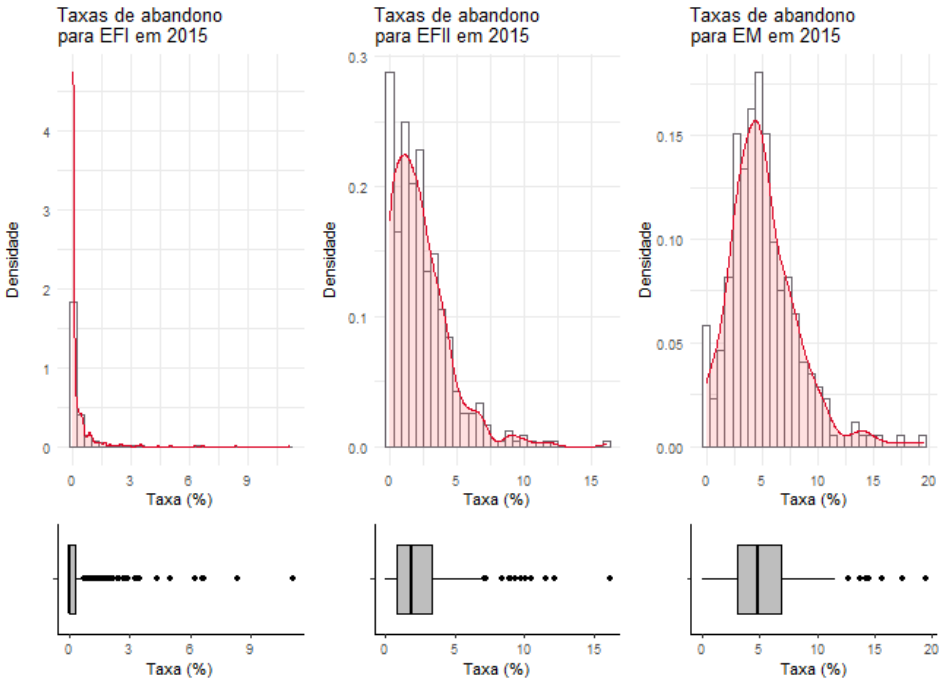

Figura 2 - Histogramas e *box-plots* das taxas de abandono em 2015.

Por fim, a Figura 3 apresenta as distribuições das taxas para os blocos de ensino em 2010. Assim como ocorreu para 2015, não há *outliers* muito extremos que interfiram na visualização da forma da distribuição. Neste ano, a taxa de abandono para EFII de maior frequência é aproximadamente 3%, enquanto a taxa de maior frequência para EM está próxima de 6%.

A Figura 4 apresenta as curvas para as taxas dos três blocos de ensino referentes ao ano de 2010, já que para este ano a diferença entre as escalas das densidades não era tão grande. Sob a mesma escala, pode-se ver a discrepância entre as taxas para os três blocos, discrepância esta que foi minimizada pelas escalas diferentes apresentadas em cada um dos três gráficos das Figuras 1, 2 e 3. Destaca-se que esta diferença entre as distribuições para bloco ilustra a necessidade dos anos escolares serem analisados em blocos distintos.

Uma vez que a presença de muitos valores iguais a zero na amostra pode afetar negativamente as modelagens que se pretende fazer, optou-se por seguir o estudo com apenas as amostras referentes ao EM. Apesar destas serem as menores amostras, elas apresentam maior massa de dados distante de zero, ao menos em comparação com as taxas para os demais blocos.

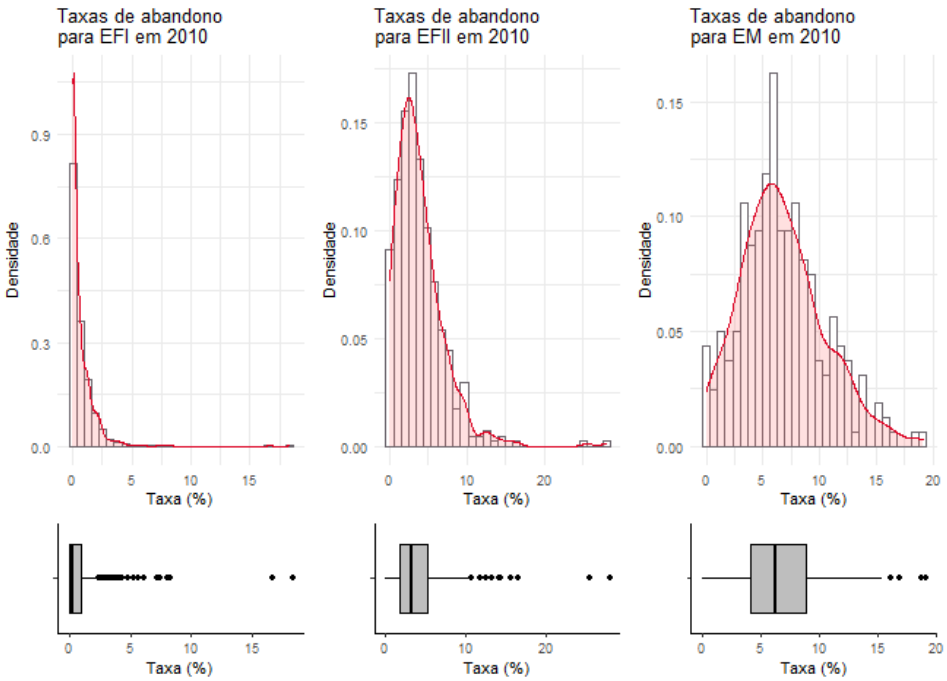

Figura 3 - Histogramas e *box-plots* das taxas de abandono em 2010.

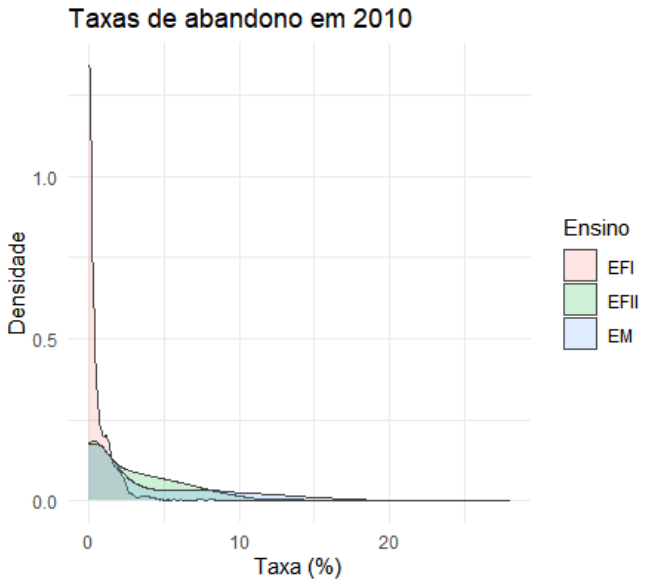

Figura 4 - Densidades das taxas de abandono em 2010.

Mesmo optando-se por utilizar as amostras com menor ocorrência de zeros, a frequência com que esse valor aparece ainda é alta, podendo causar imprecisões nas modelagens. A fim de testar quais seriam os resultados das modelagens, caso não houvesse uma frequência tão grande de zeros, foram criadas mais três amostras que são, na verdade, as três amostras para o EM já existentes, mas eliminando-se as cidades com taxa zero. Estas novas amostras são descritas na Tabela 5 a seguir.

Tabela 5 - Medidas resumo para as amostras do EM sem as taxas zero.

Ano	Variável	Média	D.P.	Mín.	1ºQ	Mediana	3ºQ	Máx.	N
2020	Taxa abandono	5,18	6,31	0,08	1,12	3,23	6,99	50	149
	Taxa públicas	75,63	31,65	7,69	40,10	92,00	100	100	
	Taxa urbana	51,26	28,29	10	28,57	42,86	76,92	100	
	IDHM Renda	0,82	0,03	0,74	0,80	0,82	0,85	0,89	
	População	69.506,03	222.085,7	2.646	12.739	20.545	47.825	2.521.564	
	log(População)	10,23	1,10	7,88	9,45	9,93	10,78	14,74	
	PIB *per capita*	2,26	0,06	2,17	2,21	2,26	2,31	2,39	
	log(PIB *per capita*)	9,63	0,57	8,75	9,16	9,55	10,08	10,95	
2015	Taxa abandono	4,57	3,78	0,07	1,32	4,03	6,92	19,50	149
	Taxa públicas	70,16	34,18	3,66	29,55	90,00	100	100	
	Taxa urbana	51,21	29,86	5,00	23,53	42,86	80,49	100	
	IDHM Renda	0,83	0,03	0,74	0,80	0,83	0,85	0,89	
	População	80.129,23	226.677,3	4.983	14.869	21.459	74.171	2.502.557	
	log(População)	10,37	1,14	8,51	9,61	9,97	11,21	14,73	
	PIB *per capita*	2,26	0,06	2,14	2,2	2,25	2,30	2,43	
	log(PIB *per capita*)	9,57	0,60	8,52	9,04	9,49	10,02	11,30	
2010	Taxa abandono	5,39	4,88	0,05	0,93	3,75	8,70	19,13	149
	Taxa públicas	64,89	35,57	4,88	25,0	82,61	100	100	
	Taxa urbana	50,95	30,09	6,99	25,0	43,75	80,0	100	
	IDHM Renda	0,83	0,03	0,74	0,80	0,83	0,86	0,88	
	População	74.176,68	211.717	4.804	15.024	25.311	72.220	2.375.151	
	log(População)	10,39	1,07	8,48	9,62	10,14	11,19	14,68	
	PIB *per capita*	2,21	0,07	2,09	2,14	2,21	2,26	2,41	
	log(PIB *per capita*)	9,13	0,62	8,09	8,52	9,11	9,58	11,13	

Observa-se que as taxas de abandono médias continuam baixas, com concentração de valores não iguais a, mas próximos de 0. As amostras originais para o bloco EM eram as de menor tamanho e coincidentemente apresentaram o mesmo número de municípios na remoção das taxas zero. Vale notar que, apesar dos tamanhos amostrais serem iguais, os municípios pertencentes a uma amostra não necessariamente pertencem às outras, apesar da maioria deles estar presente nas três.

A Figura 5 apresenta os histogramas e *box-plots* para estas amostras modificadas. Observa-se que, mesmo os valores iguais a zero tendo sido removidos, os valores próximos de zero são os de maiores frequências. As taxas parecem ser menores, no geral, para os anos de

2015 e 2020, apesar da amostra deste ano apresentar uma taxa extrema de 50%, que distorce a visão da estrutura destes dados.

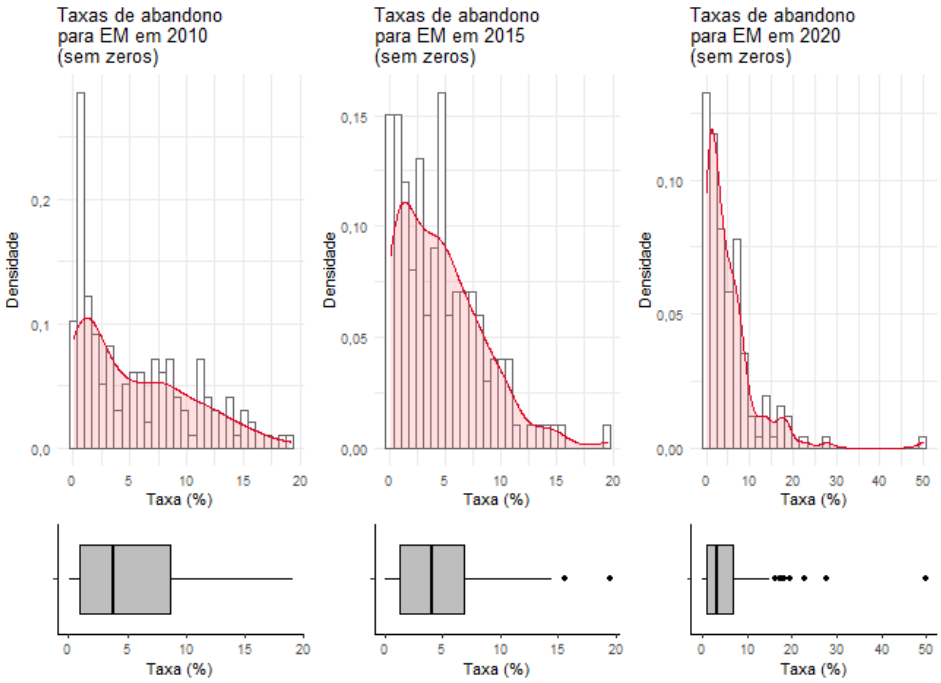

Figura 5 - Histogramas e *box-plots* das taxas de abandono sem zeros para EM.

3 Metodologia

Como observado pelas Figuras 1, 2 e 3, e das Tabelas 2, 3 e 4, os dados apresentam distribuição assimétrica à esquerda, com maior densidade de valores próximos de 0. Além disso, a variável sob estudo é uma taxa, e portanto varia de 0% a 100% ou, se for dividida por 100, varia de 0 a 1. Neste cenário em que os dados apresentam distribuição claramente não Normal, por ser assimétrica e com limites fixos, uma possibilidade seria a aplicação de alguma transformação na variável a fim de possibilitar o uso da Regressão Linear Múltipla. Uma vez que essa abordagem traria dificuldades para a interpretação dos resultados e poderia não incorporar bem à modelagem o fato destes dados serem restritos a um intervalo fixo, optou-se pelo uso de Modelos Lineares Generalizados. Os Modelos Lineares Generalizados permitem ajustar modelos de regressão linear a dados que violam a suposição de normalidade.

Assim como nas regressões lineares múltiplas, o objetivo é modelar a média da distribuição. No entanto, muitas distribuições não apresentam sua média como um de seus parâmetros, como ocorre naturalmente com a distribuição Normal. Nestes casos, basta aplicar alguma transformação sobre os parâmetros para se obter uma reparametrização que apresente a média da distribuição como um de seus parâmetros. Apesar do parâmetro a ser modelado ter relação direta com a média da distribuição, ainda é necessário estabelecer uma relação entre as covariáveis e a média por meio do preditor linear definido por

$$\eta_i = X_i^T \beta.$$

Em MLG, esta relação é generalizada pela função de ligação $g(.)$, de forma que para uma amostra aleatória $Y_1, Y_2, \dots, Y_n$, tem-se

$$E(Y_i) = \theta_i = g(\eta_i).$$

Quando acredita-se que os dados sob estudo possuem correlação espacial, pode-se inserir o efeito espacial na modelagem. A Estatística Espacial (Banerjee, Carlin, and Gelfand; 2014) se baseia no fato de que observações oriundas de espaços geográficos próximos, ou seja, observações que sejam vizinhas, podem ter seus valores correlacionados. Assim, a cada observação amostral corresponde um efeito aleatório espacial Δ_i, que é inserido na modelagem da estrutura sob estudo ao ser somado ao preditor linear, de forma que se tem

$$E(Y_i) = \theta_i = g(\eta_i + \Delta_i).$$

Estes efeitos espaciais aleatórios seguem uma distribuição Normal n-variada, sendo n o tamanho amostral, com sua média sendo um vetor de zeros de comprimento n e com uma matriz de variâncias e covariâncias $\boldsymbol{S_\Delta}$ de dimensão $n \times n$. Esta matriz é definida como

$$\boldsymbol{S_\Delta} = \tau \cdot [D - \rho W]^{-1},$$

em que $\tau > 0$ é a variância do efeito aleatório espacial, $0 \leq \rho \leq 1$ é a correlação admitida entre as observações vizinhas, W é uma matriz quadrada $n \times n$ das distâncias entre os elementos amostrais e D é uma matriz diagonal $n \times n$ com o número de vizinhos de cada observação como sua diagonal principal. A matriz W é calculada com alguma medida de distância e a matriz D é estimada determinando-se um limiar para decidir se uma observação é ou não vizinha da outra. Se a distância for menor que o limiar escolhido, as observações são consideradas vizinhas, e caso contrário são consideradas como não vizinhas. É ideal que o menor valor presente na matriz D seja 1, ou seja, o número mínimo de vizinhos dentre as observações deve ser 1. Desta forma, o limiar para definição das vizinhanças deve ser escolhido de modo a satisfazer esta condição para a amostra em estudo.

Nos dados deste estudo, a distribuição assimétrica à esquerda com valores iguais ou maiores que zero sugere o uso do MLG Gama. Já com a transformação que faz com que os dados fiquem entre 0 e 1, o uso do MLG Beta aparenta ser mais adequado, uma vez que a distribuição Beta apresenta estes limites e pode ser assimétrica dependendo da sua configuração de parâmetros. Antes de partir para a modelagem dos dados reais, no entanto, foram conduzidos estudos simulados para avaliar qual a distribuição mais adequada para modelar os dados. Como um dos propósitos deste estudo é incluir o efeito espacial na modelagem, os estudos simulados também exploraram as possibilidades de inclusão deste efeito nas modelagens e quais os impactos de modelar efeitos espaciais em dados que não os apresentam e vice-versa.

A primeira hipótese foi a de que os dados pudessem seguir uma distribuição Gama. Para uma variável aleatória Y que segue a distribuição Gama, sua função de densidade probabilidade é dada por

$$f(y, \delta) = \frac{\delta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\delta y}, \qquad y \geq 0,$$

em que $\alpha > 0$ é o parâmetro de forma e $\delta > 0$ é o parâmetro de taxa. A esperança e variância de Y são dadas por

$$E(Y) = \frac{\alpha}{\delta} \qquad Var(Y) = \frac{\alpha}{\delta^2}.$$

Como a distribuição Gama não apresenta sua média como um de seus parâmetros, esta distribuição foi reparametrizada para que um de seus parâmetros apresentasse uma relação direta com a média. O novo parâmetro criado foi, então,

$$\theta = \frac{\alpha}{\delta}.$$

Dessa forma, a f.d.p. com a nova parametrização é

$$f(y, \theta) = \frac{\left(\frac{\alpha}{\theta}\right)^{\alpha}}{\Gamma(\alpha)} y^{\alpha-1} e^{-\frac{\alpha}{\theta}y}, \qquad y \geq 0,$$

e seu valor esperado e variância são dados por

$$E(Y) = \theta \qquad Var(Y) = \frac{\theta^2}{\alpha}.$$

Por fim, a função de ligação considerada para os MLG's Gama, a partir daqui denominados modelos **gama**, com a inclusão do efeito espacial, foi

$$\theta_i = g(\eta_i + \Delta_i) = e^{\eta_i + \Delta_i} = e^{X_i^T \beta + \Delta_i}.$$

Para as taxas de abandono transformadas, o MLG Beta se mostrou bastante adequado. A função densidade probabilidade para uma variável Y que segue a distribuição Beta é, comumente, dada por

$$f(y, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}, \qquad 0 \leq y \leq 1,$$

em que $\alpha, \beta > 0$ são os parâmetros de forma. Suas média e variância são

$$E(Y) = \frac{\alpha}{\alpha + \beta} \qquad Var(Y) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta-1)}$$

Assim como feito para a distribuição Gama, os parâmetros da Beta foram modificados para que um deles apresentasse relação direta com a esperança da distribuição. Assim, os novos parâmetros são

$$\theta = \frac{\alpha}{\alpha + \beta} \qquad \zeta = \alpha + \beta,$$

de forma que $0 \leq \theta \leq 1$ e $\zeta > 0$ é conhecido como parâmetro de ruído. Com essa parametrização, a função densidade probabilidade é dada por

$$f(y, \theta) = \frac{\Gamma(\zeta)}{\Gamma(\theta\zeta)\Gamma[(1-\theta)\zeta]} y^{\theta\zeta - 1}(1-y)^{(1-\theta)\zeta - 1}, \qquad 0 \leq y \leq 1.$$

Assim, tem-se

$$E(Y) = \theta \qquad Var(Y) = \frac{\theta(1-\theta)}{1+\zeta},$$

o que permite a modelagem da média da variável.

Como os dados sob estudo parecem ser melhor descritos por uma distribuição Beta do que por uma Gama, decidiu-se por testar mais variações desse MLG. A primeira variação foi como a descrita para o modelo Gama, com a inserção do efeito espacial no parâmetro da média. A partir daqui, este modelo será referido como modelo **beta**. A função de ligação utilizada neste caso foi a *logit* que, com a adição do efeito espacial, é dada por:

$$\theta_i = g(\eta_i + \Delta_i) = \frac{e^{\eta_i + \Delta_i}}{1 + e^{\eta_i + \Delta_i}} = \frac{e^{X_i^T \beta + \Delta_i}}{1 + e^{X_i^T \beta + \Delta_i}}.$$

A segunda variação modela adicionalmente o parâmetro de ruído ζ de forma similar à modelagem do parâmetro θ_i, ou seja, ele é descrito por uma matriz de covariáveis Z e um vetor de coeficientes $\boldsymbol{\gamma}$. Este modelo foi denominado **beta_zeta** e sua função de ligação para ζ_i foi a *log*:

$$\zeta_i = g(\nu_i) = e^{\nu_i} = e^{Z_i^T \gamma}.$$

Por fim, a terceira variação incorpora o efeito espacial à modelagem de ζ_i. Este modelo será referido como **beta_zeta_delta**, apresentando função de ligação para ζ_i dada por

$$\zeta_i = g(\nu_i + \Delta_i) = e^{\nu_{i+\Delta_i}} = e^{Z_i^T \gamma + \Delta_i}.$$

Uma vez que os MLG's que se deseja modelar são modelos hierárquicos com efeito aleatório espacial, optou-se por trabalhar com a Inferência Bayesiana. Sob o ponto de vista da Inferência Clássica, para que todos os parâmetros de interesse fossem estimados deveriam ser feitas várias integrações para obter resultados analíticos muito complexos, que computacionalmente seriam executados com dificuldade. Já sob a visão Bayesiana estas modelagens se tornam mais simples do ponto de vista computacional.

A inferência Bayesiana se baseia na suposição de que os parâmetros de uma distribuição são variáveis aleatórias e possuem distribuição de probabilidade própria. As distribuições desses parâmetros são ditas distribuições *a priori* e refletem o nosso grau de incerteza sobre o comportamento da variável de interesse. Quanto menos certeza se tem sobre os dados sob estudo, mais vagamente devem ser especificadas as distribuições *a priori* levadas em conta. Em um MLG Bayesiano, a informação amostral vem da função de verossimilhança, que leva em consideração a estrutura da distribuição dos dados condicional à distribuição *a priori*. Por outro lado, a distribuição que se supõe descrever os dados amostrais, mas que não conta com alguma informação amostral, é chamada de função de verossimilhança marginal.

Através do Teorema de Bayes, a distribuição *a priori* é atualizada com a informação amostral para obtenção da chamada distribuição *a posteriori*. A Figura 6 ilustra como as informações *a priori* e amostral se unem para encontrar a distribuição que melhor descreve o parâmetro modelado; nota-se que o uso de distribuições *a priori* vagas dão maior peso à informação amostral na obtenção da *a posteriori*.

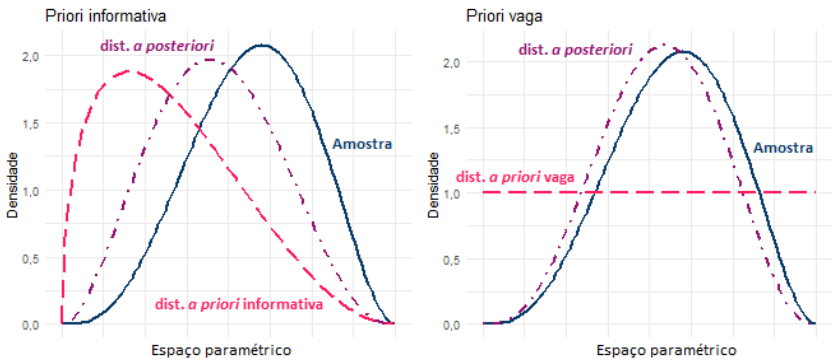

Figura 6 - Ilustração do efeito da distribuição *a priori* sobre a *a posteriori*.

As distribuições *a priori* para os parâmetros $\boldsymbol{\theta}$ a serem estimados podem ser escolhidas com base nas famílias conjugadas naturais da distribuição dos dados sob estudo. Quando a

distribuição *a priori* é definida com base na conjugada natural, tem-se que a função de verossimilhança, como função de θ_i, é proporcional à distribuição *a priori*. Além disso, a distribuição *a posteriori* obtida será aquela assumida pela *a priori*. Neste estudo, optou-se pela análise conjugada para definição das distribuições *a priori* utilizadas.

No caso dos MLG Bayesianos, é modelado o vetor de médias $\boldsymbol{\theta}$, bem como os demais parâmetros que descrevem a distribuição utilizada. Para o vetor de coeficientes, uma *a priori* adequada é a Normal *p*-variada, com *p* o número de coeficientes que modelam a média. Já para parâmetros de ruído e precisão, a distribuição Gama se mostra a mais adequada, por seu suporte ser maior que zero.

Assim, as distribuições *a priori* para o modelo Gama foram

$$\boldsymbol{\beta} \sim N_p(\boldsymbol{m_\beta}, \boldsymbol{S_\beta}) \qquad \alpha \sim Gamma(a_\alpha, b_\alpha)\ ,$$

em que $\boldsymbol{\beta}$ é o vetor de coeficientes para o modelo, de tamanho *p*, $\boldsymbol{m_\beta}$ é o vetor de médias de tamanho *p* para a *a priori* e $\boldsymbol{S_\beta}$ é a matriz de variâncias e covariâncias de $\boldsymbol{\beta}$, de dimensão $p \times p$.

Já para os modelos Beta, em todas suas variações foi utilizada para o vetor de coeficientes a mesma *a priori* utilizada para o caso da Gama. No modelo beta o parâmetro de ruído teve a distribuição Gama como sua *a priori*:

$$\zeta \sim Gamma(a_\zeta, b_\zeta).$$

Já para os modelos beta_zeta e beta_zeta_delta, que inserem covariáveis na modelagem de ζ, o vetor de coeficientes $\boldsymbol{\gamma}$ de tamanho *q* teve como *a priori* uma Normal *q*-variada:

$$\boldsymbol{\gamma} \sim N_q(\boldsymbol{m_\gamma}, \boldsymbol{S_\gamma}).$$

Para inserção do efeito espacial, são modelados os parâmetros Δ_i e τ. De forma semelhante ao caso do vetor de coeficientes, o vetor de efeito espacial $\boldsymbol{\Delta}$ tem como *a priori* uma Normal *n*-variada, em que *n* é o tamanho da amostra, e como nos casos dos parâmetros de precisão e ruído, τ tem como *a priori* uma Gama:

$$\boldsymbol{\Delta} \sim N_n(\boldsymbol{m_\Delta}, \boldsymbol{S_\Delta} \cdot \tau) \qquad \tau \sim Gamma(a_\tau, b_\tau).$$

As distribuições *a priori* apresentadas acima foram utilizadas para os modelos gama, beta e beta_zeta. Já no caso do modelo beta_zeta_delta, em que existe efeito espacial tanto em θ_i quanto em ζ_i, foram modelados efeitos aleatórios separados para cada um desses parâmetros. Assim, tem-se

$$\boldsymbol{\Delta}_\theta \sim N_n(\boldsymbol{m}_{\boldsymbol{\Delta},\boldsymbol{\theta}}, \boldsymbol{S}_{\boldsymbol{\Delta},\boldsymbol{\theta}} \cdot \tau_{\boldsymbol{\theta}}); \quad \tau_\theta \sim Gamma(a_{\tau,\theta}, b_{\tau,\theta})$$

$$\boldsymbol{\Delta}_\zeta \sim N_n(\boldsymbol{m}_{\boldsymbol{\Delta},\boldsymbol{\zeta}}, \boldsymbol{S}_{\boldsymbol{\Delta},\boldsymbol{\zeta}} \cdot \tau_{\boldsymbol{\zeta}}); \quad \tau_\zeta \sim Gamma(a_{\tau,\zeta}, b_{\tau,\zeta}).$$

Optou-se por fornecer distribuições *a priori* vagas para todas as modelagens, a fim de permitir que a *a posteriori* tivesse maior influência dos dados amostrais. Para as Normais multivariadas de $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$, foi fornecido um vetor de zeros como média, uma vez que os coeficientes de regressão tendem a assumir valores em torno de zero. Apesar das covariáveis que descrevem os parâmetros θ e ζ poderem ser diferentes, optou-se por utilizar a mesma matriz de covariáveis para ambas modelagens, ou seja, X=Z. Dessa forma, ambos parâmetros tiveram como *a priori* Normais p-variadas. Para a matriz de variâncias e covariâncias dessas *a priori*, foram fornecidas matrizes diagonais $p \times p$ com o valor 10 por toda a diagonal principal. A diagonalidade dessa matriz garante que não exista, *a priori*, correlação entre os coeficientes, enquanto o valor grande para as variâncias descreve uma *a priori* vaga. Em resumo, tem-se

$$\boldsymbol{m}_{\boldsymbol{\beta}} = \boldsymbol{m}_{\boldsymbol{\gamma}} = \mathbf{0} \qquad \boldsymbol{S}_{\boldsymbol{\beta}} = \boldsymbol{S}_{\boldsymbol{\gamma}} = \begin{pmatrix} 10 & 0 & \cdots & 0 \\ 0 & 10 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 10 \end{pmatrix}.$$

Já para os parâmetros de forma α, para a distribuição Gama, e ruído ζ, para a distribuição Beta, foram definidos os hiperparâmetros de suas *a priori* como $a_\alpha = a_\zeta = 0{,}1$ e $b_\alpha = b_\zeta = 0{,}1$. Estas especificações também configuram *a priori* vagas, uma vez que a esperança é 1 enquanto a variância é 10. Para o efeito espacial, os hiperparâmetros para a *a priori* de τ também foram $a_\tau = a_{\tau,\theta} = a_{\tau,\zeta} = b_\tau = b_{\tau,\theta} = b_{\tau,\zeta} = 0{,}1$. Os vetores de médias foram definidos como $\boldsymbol{m}_{\boldsymbol{\Delta}} = \boldsymbol{m}_{\boldsymbol{\Delta},\boldsymbol{\theta}} = \boldsymbol{m}_{\boldsymbol{\Delta},\boldsymbol{\zeta}} = \mathbf{0}$, assim como para as demais Normais multivariadas. Já as matrizes $\boldsymbol{S}_{\boldsymbol{\Delta}}$, $\boldsymbol{S}_{\boldsymbol{\Delta},\boldsymbol{\theta}}$ e $\boldsymbol{S}_{\boldsymbol{\Delta},\boldsymbol{\zeta}}$ devem carregar a informação espacial dos dados e são definidas pelas respectivas matrizes de vizinhanças amostrais. Para o cálculo dessas matrizes, considerou-se $\rho = 0{,}95$ e a matriz W foi calculada com a distância Euclidiana. Como já tratado anteriormente, a vizinhança entre duas observações deve ser determinada através de um limiar pré-estabelecido. A Tabela 6 a seguir apresenta os limiares escolhidos para que o

número mínimo de vizinhos fosse igual a 1, bem como o número máximo de vizinhos em cada matriz, para cada amostra das taxas de abandono.

Tabela 6 - Limiares e máximo de vizinhos para as matrizes de vizinhança amostrais.

Amostra	Limiar	Nº municípios	Nº máximo de vizinhos
EM 2020	0,90	272	39
EM 2015	0,87	256	36
EM 2010	0,87	243	35
EM 2020 (sem zeros)	1,25	149	32
EM 2015 (sem zeros)	1,30	149	31
EM 2010 (sem zeros)	1,50	149	38

Na Inferência Bayesiana, as estimativas são obtidas a partir de alguma medida de centralidade da distribuição *a posteriori*; neste estudo, todas as estimativas foram obtidas a partir da média *a posteriori*. No entanto, estimativas pontuais não são evidência suficiente de que o valor real do parâmetro realmente seja aquele encontrado. O processo de inferência pode ser complementado pelo uso de Intervalos de Credibilidade.

Uma região do espaço paramétrico é uma região com credibilidade α se a probabilidade do valor real do parâmetro pertencer a essa região, condicional à amostra, for maior ou igual à α. A forma mais usual de definir o intervalo de credibilidade é encontrando a região de mais alta densidade *a posteriori*, mais conhecida como região HPD, do inglês *Highest Posterior Density*. Estes intervalos são únicos e os de menor amplitude, ou seja, concentram a maior densidade de probabilidade do valor real do parâmetro estar dentre as estimativas *a posteriori* nele contidas. Os intervalos de credibilidade possuem interpretação probabilística direta, diferentemente dos Intervalos de Confiança. Em outras palavras, a credibilidade α é de fato a probabilidade do valor real estar dentro do intervalo. No caso dos modelos deste estudo, os intervalos HPD permitem verificar a significância dos coeficientes estimados. Se o intervalo para um dado coeficiente não contiver o zero, considera-se que este coeficiente é significativo, e vice versa. Em todos os intervalos calculados considerou-se credibilidade de 95%.

Como já discutido anteriormente, foram conduzidos estudos simulados para decidir qual modelo aplicar aos dados reais. Foi explorado o MLG Gama com efeito espacial na estrutura da média (modelo gama), o MLG Beta com efeito espacial apenas na média (modelo beta), uma variação deste modelo beta que incluía a modelagem do parâmetro de ruído ζ com o auxílio de covariáveis (modelo beta_zeta) e, por fim, uma variação do modelo beta_zeta que incluía efeito espacial na estrutura do parâmetro de ruído (modelo beta_zeta_delta). Para cada um destes

modelos foram gerados bancos de dados artificiais seguindo a mesma estrutura a ser modelada. Para o modelo gama, por exemplo, a variável resposta foi gerada com o auxílio de covariáveis e inserção de efeito espacial. Esta mesma lógica foi utilizada para a geração de três bancos de dados artificiais para cada estrutura de modelo, nos tamanhos amostrais 50, 100 e 200. Todos os bancos apresentaram um covariável binária, gerada a partir da distribuição Bernoulli, e uma contínua, gerada a partir de uma distribuição *Uniforme*(-1, 1). O efeito espacial foi inserido tanto na estrutura das médias quanto na estrutura do ruído a partir de matrizes de vizinhança diagonais com até 4 vizinhos por elemento amostral, com $\rho = 0{,}95$. A Tabela 7 a seguir apresenta os parâmetros reais considerados para a geração de cada conjunto de dados. Os coeficientes de índices 1 são referentes às covariáveis binárias, enquanto os de índice 2 são referentes às contínuas, e os de índice 0 correspondem aos interceptos.

Tabela 7 - Parâmetros das distribuições dos dados artificiais.

Modelo	β_0	β_1	β_2	γ_0	γ_1	γ_2	α	ζ	τ /τ_θ	τ_ζ
gama	1,5	0,5	0,5	-	-	-	10,0	-	2,0	-
beta	-0,75	0,5	-1,5	-	-	-	-	5,0	2,0	-
beta_zeta	-0,75	0,5	-1,5	1,75	1,0	2,0	-	-	2,0	-
beta_zeta_delta	-0,75	0,5	-1,5	1,75	1,0	2,0	-	-	2,0	1,0

Os dados foram então ajustados com os modelos correspondentes à sua estrutura. Além disso, avaliou-se as consequências de não modelar a estrutura espacial em dados que a apresentam e de modelar essa estrutura em dados que não a apresentam. Assim, foram feitos outros dois ajustes denominados **beta_zeta_trocado** e **beta_zeta_delta_trocado**. O ajuste beta_zeta_trocado modela os dados beta_zeta, que não contêm efeito espacial na estrutura de $\boldsymbol{\zeta}$, com o modelo que inclui efeito espacial nesta estrutura. Analogamente, o modelo beta_zeta_delta_trocado modela os dados com efeito espacial em $\boldsymbol{\zeta}$ utilizando o modelo que não modela esta estrutura espacial.

A fim de avaliar mais extensamente o desempenho dos modelos considerados, foi conduzido um estudo Monte Carlo para obtenção dos vícios relativos para os parâmetros estimados. Para cada conjunto de dados artificiais gerado, foram gerados mais 9, totalizando 10 bancos de dados para cada estrutura e tamanho amostral, a serem utilizados no estudo Monte Carlo.

Os cenários das modelagens beta_zeta_trocado e beta_zeta_delta_trocado permitem avaliar o impacto de não modelar a estrutura espacial existente nos dados e vice-versa. No entanto, para estes modelos, avaliou-se apenas o impacto sobre o efeito aleatório espacial na

estrutura do ruído, mantendo-se sempre a modelagem da estrutura espacial da média. A fim de avaliar o que ocorre na modelagem de dados espaciais sem modelar esta estrutura, ajustou-se um modelo sem qualquer modelagem de efeito espacial aos dados que apresentavam efeito espacial tanto na média quanto no parâmetro de ruído. O objetivo foi, então, verificar se a correlação espacial, que não foi tratada pelo modelo, foi para os resíduos. Esta hipótese foi verificada através do teste I de Moran (Getis, Ord; 2010), que avalia a existência de associação espacial entre os dados dada a matriz quadrada W de pesos, ou distâncias, entre as observações. Sob a hipótese nula deste teste, não existe associação espacial, para este caso, nos resíduos. Uma vez que o ajuste para este modelo não incluía qualquer estrutura adicional à estrutura básica de um MLG Beta, optou-se por fazê-lo por meio da abordagem de Inferência Clássica. A modelagem foi então feita através da função *betareg* do pacote homônimo (Cribari-Neto, Zeileis; 2010). Além desta função retornar os resíduos do modelo, o pacote fornece uma função para transformação dos resíduos crus naqueles que se deseja avaliar. Neste caso optou-se por utilizar os resíduos de Pearson, de forma que o uso deste pacote facilitou os cálculos.

As demais modelagens para o estudo simulado, bem como a modelagem dos dados reais, foram feitas utilizando o software *Stan* (Stan Modeling Language Users Guide and Reference Manual, 2.18) através do R (R Core Team, 2021), por meio do pacote *rstan* (Stan Development Team, 2020). O *Stan* é uma linguagem de programação própria para se trabalhar com modelagem, inferência e predição estatísticas, permitindo a construção de algoritmos além de oferecer funções próprias que facilitam o aspecto estatístico e probabilístico das análises. O pacote *rstan* possibilita o ajuste de modelos Bayesianos que são rodados no *Stan*, mas permitindo que o trabalho seja feito na interface do R. A linguagem *Stan* só precisa ser utilizada na implementação de um *script* com as especificações dos parâmetros a serem estimados e suas distribuições *a priori*, além da função de verossimilhança.

Uma vez que as modelagens feitas pelo *Stan*, especialmente as com efeitos espaciais, demandam certo tempo computacional, as modelagens para o Monte Carlo precisaram ser feitas em um servidor. Para isto foi utilizado o *Colab* ou "*Colaboratory*" (Google, 2018), plataforma gratuita do *Google* que permite escrever e executar códigos em algumas linguagens, inclusive R, a partir do navegador. Os comandos executados pelo *Colab* são rodados em um servidor, melhorando consideravelmente o tempo computacional em comparação com um computador de uso pessoal. Os pacotes mais utilizados do R já estão disponíveis no ambiente do *Colab*, enquanto pacotes mais específicos precisam ser instalados manualmente. O pacote *rstan*

utilizado para as modelagens no computador pessoal não funciona nos *notebooks* do *Colab*, de forma que o pacote alternativo utilizado foi o *cmdstanr* (Gabry, Cesnovar; 2021). Ele funciona de forma similar ao *rstan*, sendo necessário fornecer apenas um *script* escrito em *Stan* com as especificações do modelo, enquanto o restante do código é escrito em R.

Para avaliar a qualidade dos ajustes dos dados reais das taxas de abandono, foi conduzida uma breve análise de resíduos e do desempenho das modelagens. Foram analisados os *traceplots* das cadeias dos parâmetros de maior interesse, para verificar a convergência dos valores. Além disso, calculou-se o resíduo de Pearson para este tipo de MLG Beta, a fim de analisar os gráficos de dispersão desses resíduos contra os valores preditos e as covariáveis presentes na estrutura da média. Os resíduos foram calculados da seguinte maneira:

$$r_i^{Pearson} = \frac{y_i - \hat{\theta}_i}{\sqrt{g(\hat{\zeta}_i)\hat{\theta}_i(1-\hat{\theta}_i)}},$$

em que y_i é o valor real da observação i e $g(\hat{\zeta}_i) = \left(1 + \hat{\zeta}_i\right)^{-1}$. Vale notar, ainda, que na modelagem Bayesiana da estrutura da média, as estimativas *a posteriori* $\hat{\theta}_i$ da média são os próprios valores preditos.

As especificações para as modelagens tanto do estudo simulado quanto dos dados reais foram fixadas, a fim de tornar o desempenho dos modelos comparável. Determinou-se as modelagens com apenas uma cadeia com 5.000 iterações no total, sendo 2.500 de *burn in*. Os chutes iniciais para os vetores de coeficientes $\boldsymbol{\beta}$ e $\boldsymbol{\gamma}$ e para os efeitos espaciais $\boldsymbol{\Delta}$ foram determinados por distribuições *Uniforme*(-0.1, 0.1). No caso em que o parâmetro ζ foi modelado sem o uso de covariáveis, o chute inicial foi 1. O parâmetro de forma α do modelo gama também teve 1 como seu chute inicial. Já para as variâncias τ, os chutes iniciais foram iguais a 2.

Ademais, todas as análises descritivas e gráficos foram feitos no R. Para todos os testes de hipóteses feitos o nível de significância considerado foi de 5%.

4 Resultados e discussão

4.1 Estudo simulado

Os primeiros dados artificiais gerados, em três tamanhos amostrais para cada tipo de modelo, foram ajustados e tiveram seus resultados pontuais analisados. A qualidade dos ajustes foi verificada através de gráficos de pontos para os valores reais de θ_i contra os valores estimados, para todos os modelos, e através, também, dos gráficos de pontos para os valores reais e estimados de ζ_i, para os modelos que modelavam essa estrutura. O cenário esperado é aquele no qual os pontos formam uma reta, que os resultados preditos pelo modelo são fiéis à realidade. Vale notar que esses gráficos de ponto, apresentados nas figuras a seguir para todos os modelos ajustados, distinguem se a observação correspondente àquele ponto apresenta o valor 0 ou 1 da covariável binária. Optou-se por apresentar esta distinção pois, particularmente nas modelagens dos ruídos ζ_i, observou-se clara separação dos pontos entre os grupos dos dois valores desta variável binária.

A Figura 7 a seguir apresenta os gráficos de pontos para as três amostras dos dados gerados a partir da distribuição Gama. Observa-se que, para os três tamanhos amostrais, os pontos formam aproximadamente uma reta, indicando bom ajuste deste modelo aos dados artificiais gama. A distinção entre as categorias da variável binária mostra que, apesar das retas para ambas categorias apresentarem a mesma direção, a reta para a categoria 0 é menor que aquela para a categoria 1, em todos os casos.

A Figura 8 apresenta o gráfico de pontos para os resultados dos modelos beta. Para todos os tamanhos amostrais os resultados foram bons, não havendo, neste caso, grande distinção entre as retas para as duas categorias da variável binária. Apesar dos pontos estarem bem distribuídos ao longo das retas, indicando que os valores das médias estão distribuídos por todo o intervalo, pode-se perceber que para a amostra de tamanho 200 há uma maior concentração de valores mais próximos de zero. Uma vez que estas amostras artificiais foram provenientes de uma distribuição Beta assimétrica com cauda pesada à direita, ou seja, com maior

concentração de valores mais próximos de zero, vemos que tamanhos amostrais maiores são mais capazes de refletir a realidade da sua distribuição de origem.

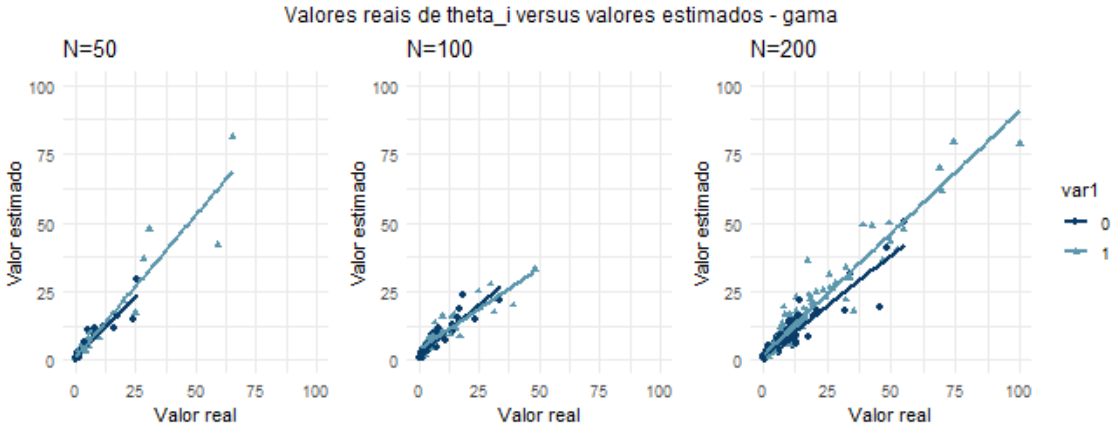

Figura 7 - Valores reais contra estimados de θ_i para modelo gama.

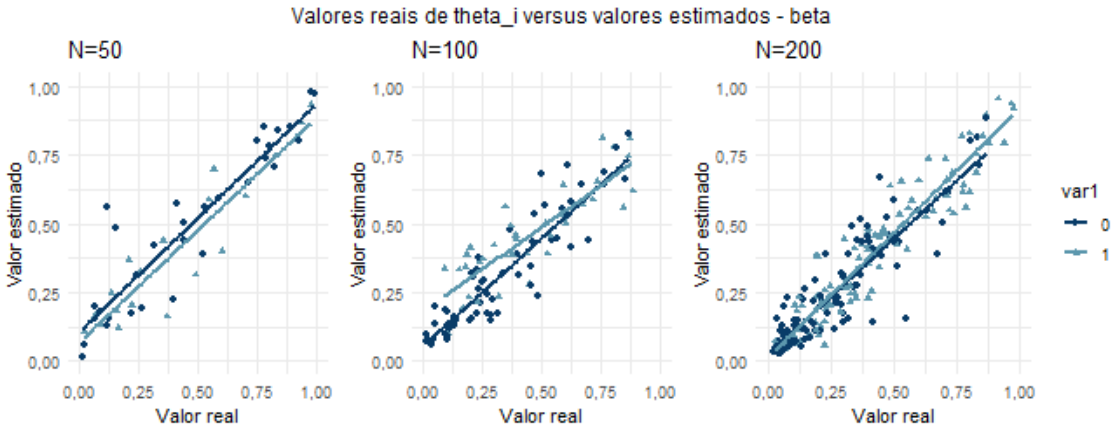

Figura 8 - Valores reais contra estimados de θ_i para o modelo beta.

A Figura 9 a seguir ilustra os resultados do ajuste beta_zeta, que apresenta coeficientes tanto na estrutura da média, θ_i, quanto do ruído, ζ_i. O ajuste para a estrutura da média foi bom, como no caso do modelo anterior. Já para a modelagem do vetor de ruídos $\boldsymbol{\zeta}$, há uma distinção evidente entre os valores assumidos pelos elementos amostrais pertencentes à categoria 0 e à categoria 1 da covariável binária. Para os tamanhos amostrais 50 e 100, percebe-se sobrestimação da média, enquanto que para a amostra de tamanho 200 os valores estimados estão bem mais próximos dos valores reais. Isto sugere que amostras grandes beneficiam o ajuste do modelo, especialmente para a estrutura do ruído no caso da distribuição Beta.

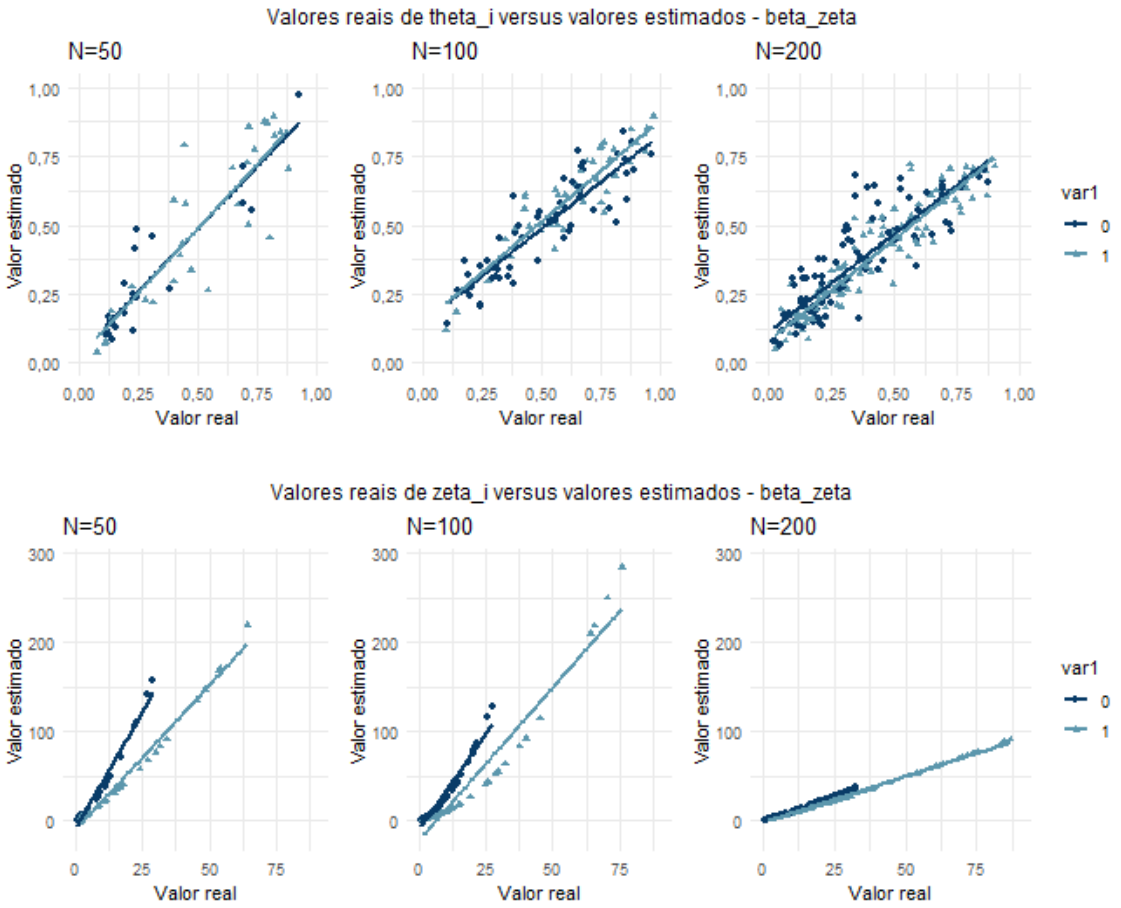

Figura 9 - Valores reais contra estimados de θ_i e ζ_i para o modelo beta_zeta.

No caso em que os dados apresentavam efeito espacial tanto na estrutura da média quanto na da variabilidade, os resultados foram um pouco diferentes. Pela Figura 10, pode-se observar que o ajuste para a média foi bom, como nos casos anteriores. Já para o ruído, a amostra de tamanho 100 teve seu ajuste prejudicado, com algumas das estimativas *a posteriori* sendo muito maiores que os valores reais. Vale notar que todos estes valores sobrestimados pertencem à categoria 1 da variável binária. Para as amostras de tamanhos 50 e 200 o ajuste foi consideravelmente melhor, ainda que apresentando certa sobrestimação. Observa-se também que, neste caso em que há efeito espacial aleatório na estrutura do ruído, os valores estimados não concordaram tanto com os reais quanto ocorreu no caso beta_zeta, em que não havia efeito espacial. Isto indica que a inserção do efeito espacial nesta estrutura aumenta a variabilidade da variável resposta, aumentando também a variabilidade das estimativas, ainda que de modo geral os resultados estejam condizentes.

O modelo beta_zeta_trocado consistiu em modelar os dados beta_zeta com o modelo para os dados beta_zeta_delta. Dessa forma, foi possível avaliar o impacto de modelar uma estrutura espacial em $\boldsymbol{\zeta}$ que em verdade não existe. A Figura 11 apresenta os resultados deste

ajuste com relação aos parâmetros de média e ruído. O ajuste da média não mudou nesta modelagem e, como esperado, os gráfico de pontos de θ_i real contra estimado apresentaram comportamento similar ao das modelagens anteriores. O ajuste do ruído, neste caso, foi prejudicado para as amostras menores, apresentando superestimação dos parâmetros. Já para a amostra de tamanho 200 as estimativas foram condizentes com os valores reais, indicando que tamanhos amostrais maiores favorecem o desempenho deste modelo.

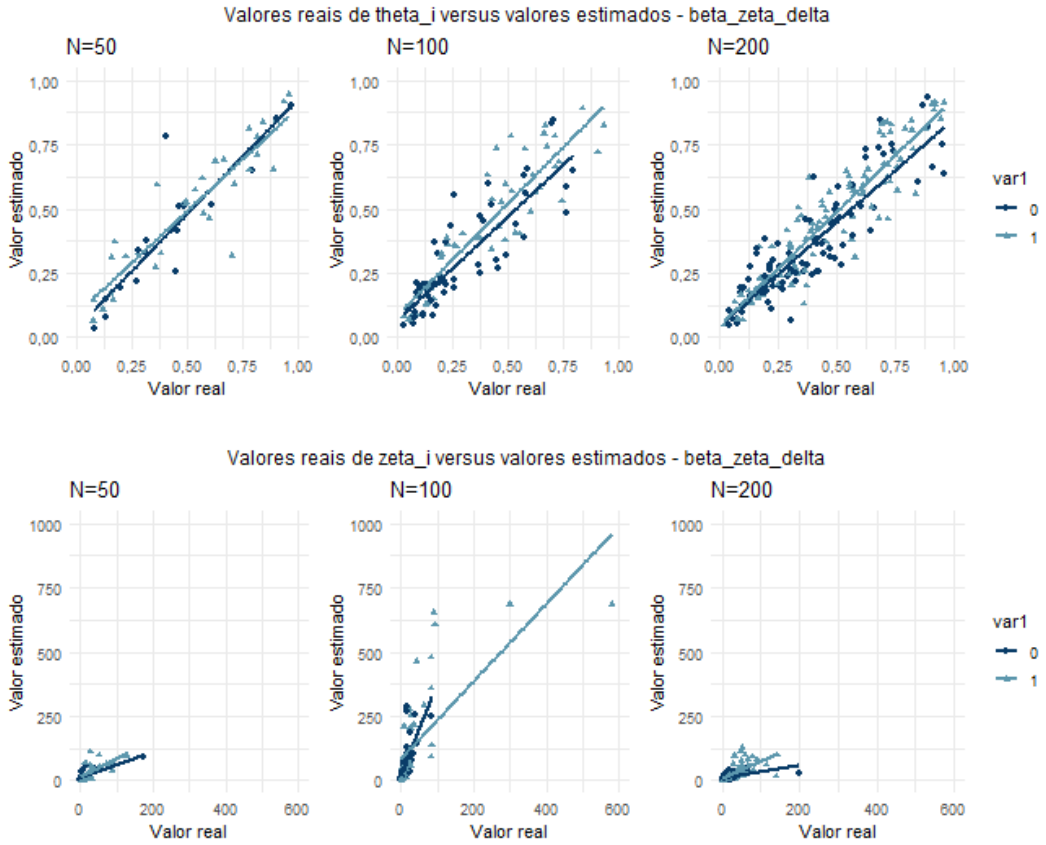

Figura 10 - Valores reais contra estimados de θ_i e ζ_i para o modelo beta_zeta_delta.

Por fim, ajustou-se os dados beta_zeta_delta com o modelo beta_zeta, a fim de avaliar o impacto de não modelar uma estrutura espacial existente no ruído. Como pode ser observado pela Figura 12, o comportamento dos pontos no gráfico é similar ao observado na Figura 10, mas com menor discrepância entre os valores reais e estimados para a amostra de tamanho 100. No geral, o desempenho deste modelo não foi ruim.

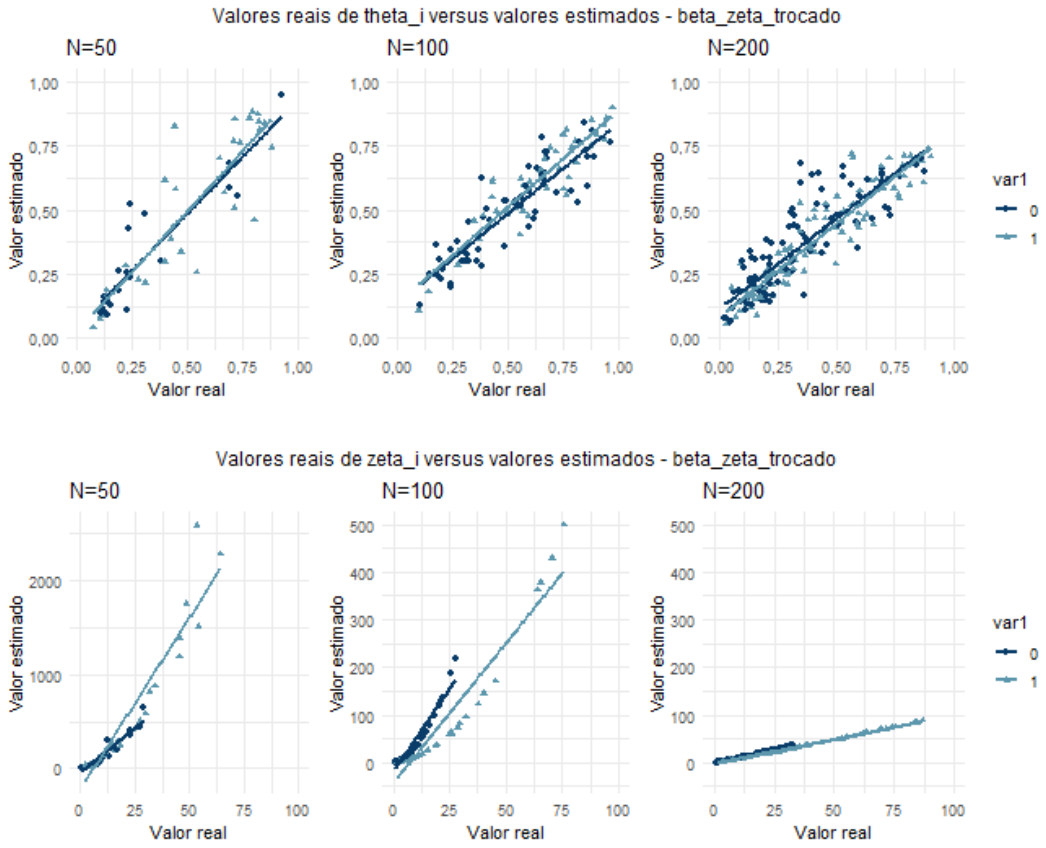

Figura 11 – Valores reais contra estimados de θ_i e ζ_i para o modelo beta_zeta_trocado.

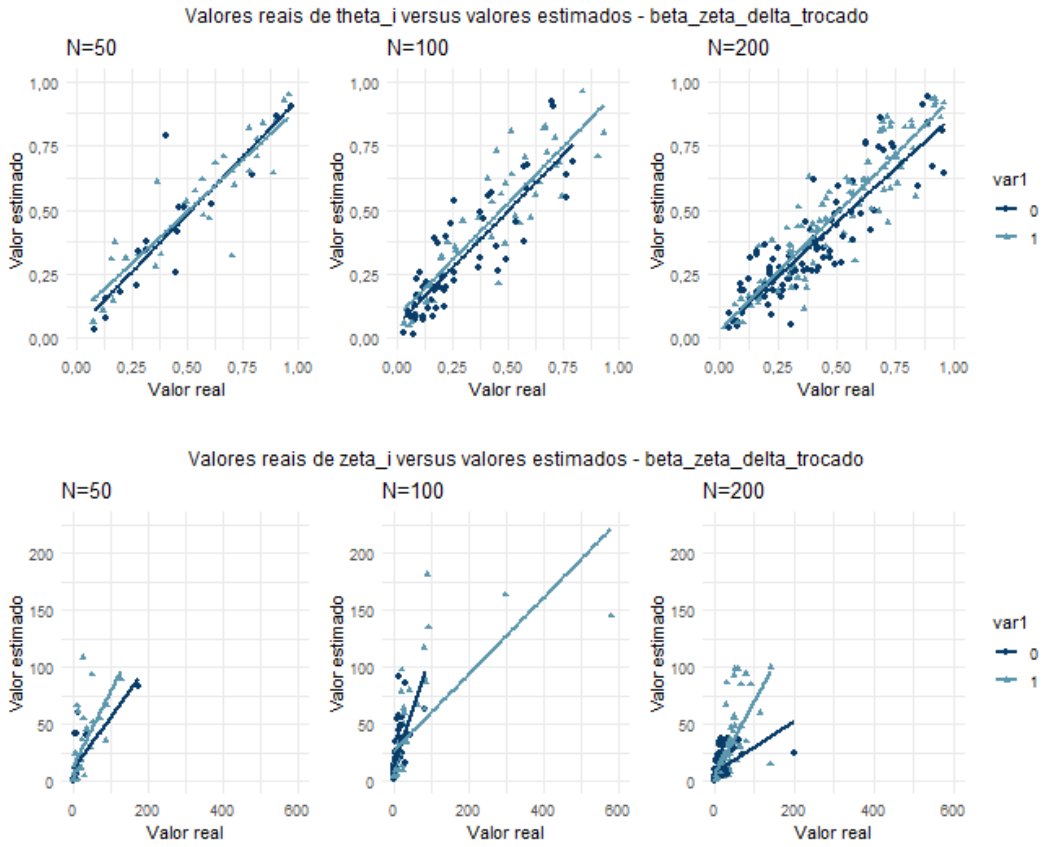

Figura 12 - Valores reais contra estimados de θ_i e ζ_i para o modelo beta_zeta_delta_trocado.

Os ajustes trocados, com dados que apresentam estruturas diferentes daquelas modeladas, não indicaram grandes impactos de se ajustar erroneamente a estrutura do ruído. Pelos resultados aqui apresentados, estimar uma estrutura espacial não existente ou não estimar uma existente levou a resultados similares àqueles obtidos com a modelagem correta das estruturas, especialmente quando o tamanho amostral é grande.

Deve-se levar em consideração que estes casos apresentados são resultados pontuais para apenas 3 amostras de tamanhos diferentes para cada estrutura de dados. A fim de avaliar mais adequadamente o desempenho dos modelos trocados, e também dos demais modelos, foi realizado um estudo Monte Carlo com 10 iterações para cada cenário sob estudo. Como não seria possível avaliar os gráficos de pontos de valores reais contra estimados para a média e a dispersão para todos os modelos ajustados, optou-se por avaliar a qualidade do ajuste por meio do vício relativo das estimativas *a posteriori* dos coeficientes e demais parâmetros dos modelos. Valores grandes para o vício relativo, em módulo, indicam que as estimativas estão muito distantes da realidade; valores negativos indicam subestimação e valores positivos indicam superestimação. Assim, o ideal é que os vícios relativos obtidos nas simulações Monte Carlo tendam a assumir valores em torno do eixo zero e apresentando variabilidade pequena, não se distanciando tanto do vício relativo ideal de 0%.

Para avaliar a distribuição dos vícios relativos calculados para os parâmetros estimados em cada modelo, foram construídos *box-plots* comparando o desempenho das estimativas para os três tamanhos amostrais. A Figura 13 apresenta estes resultados para o modelo gama. Os gráficos para os coeficientes β_i foram construídos sob a mesma escala para facilitar a comparação dos valores. Pode-se observar que os vícios relativos mais próximos de zero foram os para o coeficiente β_2, referente à variável contínua. Para todos estes coeficientes estimados, seus vícios relativos tendem a se aproximar do zero e a diminuir sua dispersão à medida que o tamanho amostral aumenta. Já para o parâmetro de forma α, a variabilidade dos vícios diminuiu, mas a mediana não se aproximou do zero, indicando vício na modelagem deste parâmetro. Um comportamento parecido pode ser observado para os vícios relativos calculados para o parâmetro de variância do efeito espacial τ. Vale notar que neste caso, apesar da mediana não se aproximar mais do zero com o aumento da amostra, ela se estabiliza em um vício relativo não muito grande, entre 15% e 20%.

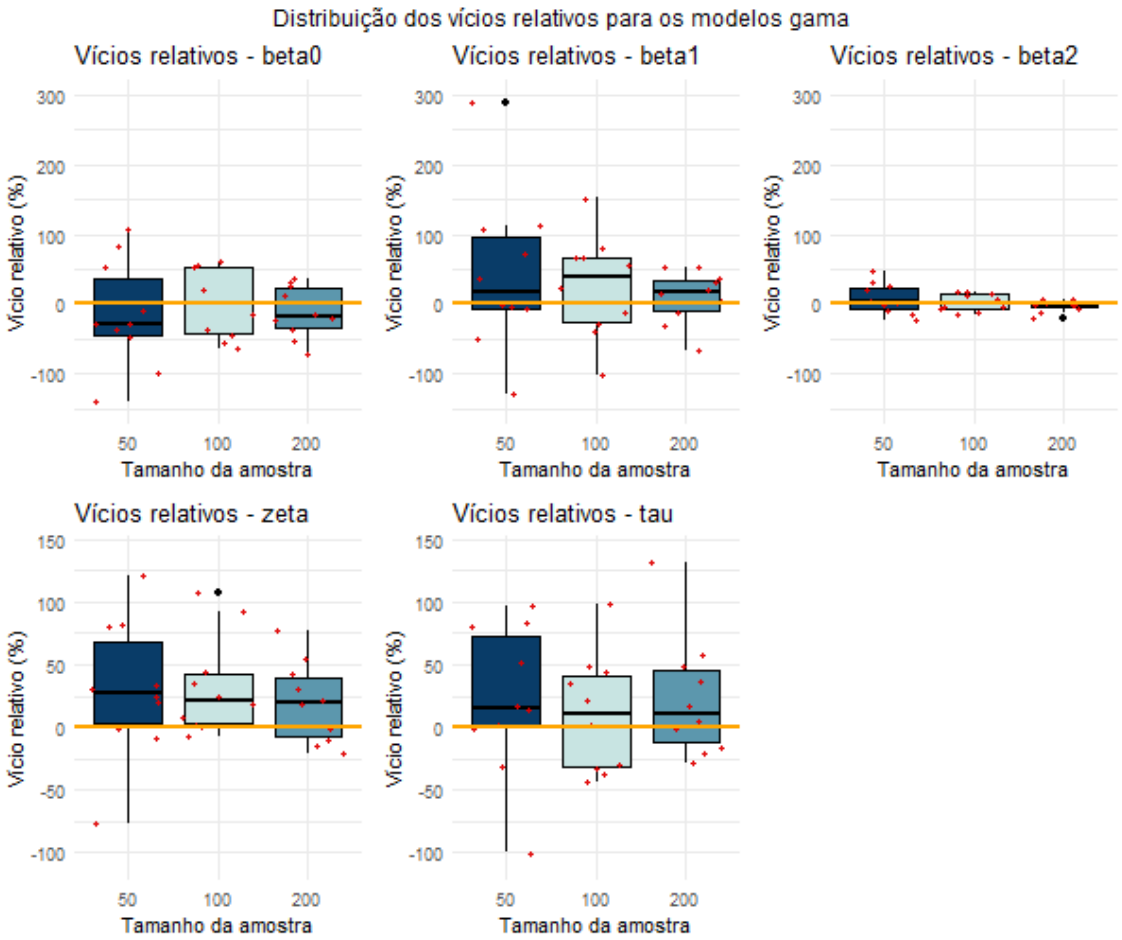

Figura 13 - Vícios relativos para os parâmetros das simulações do modelo gama.

A Figura 14 apresenta os vícios relativos para a simulação Monte Carlo do modelo beta. Pode-se perceber um comportamento similar ao observado nas simulações do modelo gama: novamente os vícios tendem a se aproximar de zero com o aumento do tamanho amostral e a estimação do coeficiente da variável contínua tem melhor desempenho que a dos demais coeficientes. Neste caso, o parâmetro de dispersão ζ apresentou sua mediana acima de 0% em todos os tamanhos amostrais, indicando vício nas estimativas deste parâmetro. Um comportamento parecido pode ser observado para o parâmetro τ.

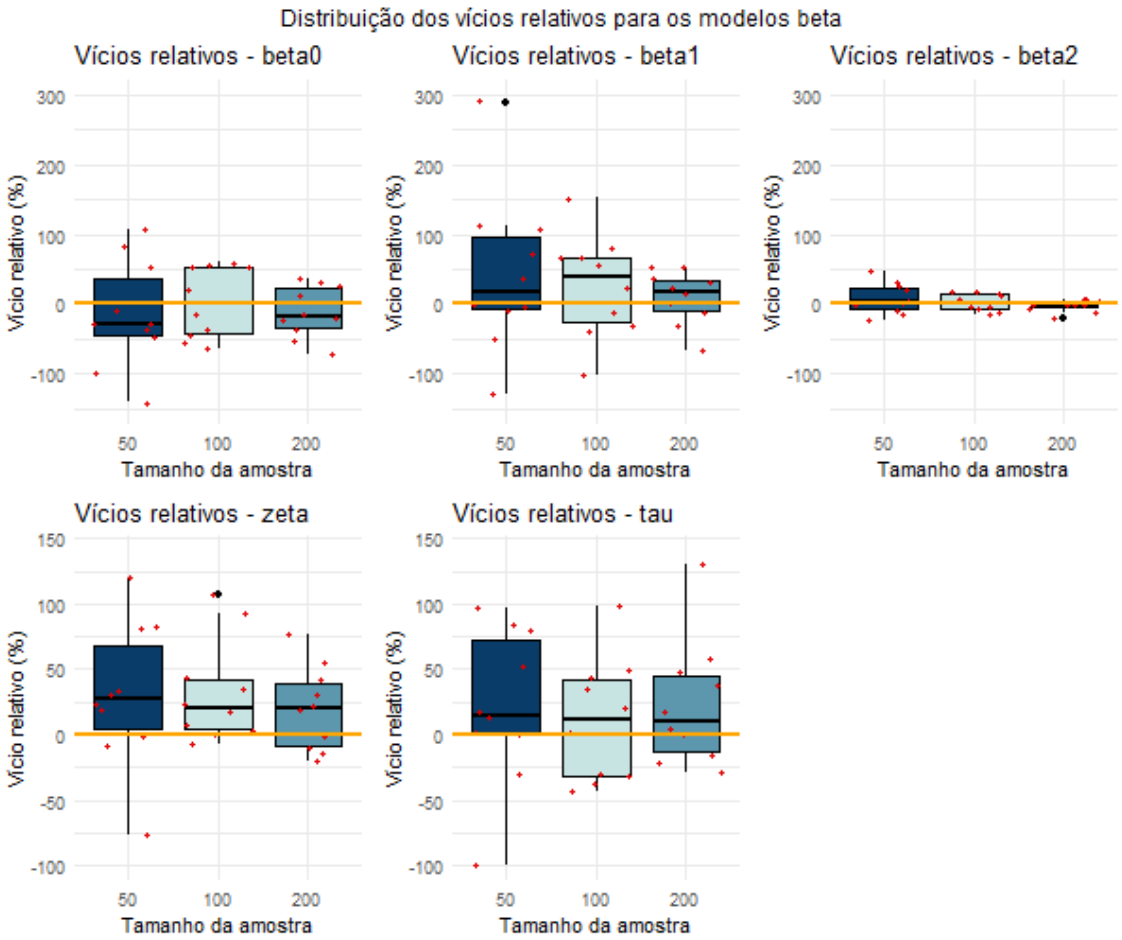

Figura 14 - Vícios relativos para os parâmetros das simulações do modelo beta.

Nos modelos que incluíram covariáveis explicativas na modelagem do parâmetro de dispersão, foram inseridos na modelagem os coeficientes γ_i . A Figura 15 apresenta os resultados da simulação para o modelo beta_zeta e pode-se observar que os vícios relativos para os coeficientes da dispersão se comportam como os coeficientes da média, com melhor desempenho para a covariável contínua e melhora das estimativas à medida que o tamanho amostral aumenta. Neste caso as medianas para os vícios relativos de τ não se mantiveram próximas do mesmo valor como ocorreu para as duas simulações apresentadas anteriormente. Além disso, elas estão mais próximas de zero e apresentaram tanto valores negativos quanto positivos, sugerindo que o ajuste deste parâmetro tenha tido melhor desempenho neste caso do que no caso do modelo beta, que não modelava a estrutura da dispersão se utilizando de covariáveis.

Pode-se perceber pela Figura 16 que, para o modelo beta_zeta_delta, os vícios relativos obtidos para os coeficientes β_i e γ_i apresentam distribuição similar às observadas para o modelo beta_zeta. A variância para o efeito espacial aleatório da média é descrita por τ_θ neste

modelo, enquanto que nos modelos anteriores esta variância era representada simplesmente por τ, uma vez que estes modelos apresentavam apenas uma estrutura espacial. O modelo beta_zeta_delta forneceu um bom ajuste para τ_θ, com diminuição da variabilidade à medida que a amostra aumentou em tamanho, além dos *box-plots* para os três tamanhos amostrais apresentarem mediana constantemente próxima de zero. A variância τ_ζ para a estrutura espacial do parâmetro de dispersão, no entanto, não foi bem ajustada, apresentando vício relativo grande e negativo, indicando subestimação deste parâmetro. Uma hipótese para esta má estimação é que neste caso modelou-se muitas estruturas ao mesmo tempo com poucas covariáveis explicativas, prejudicando as estimativas de algumas estruturas do modelo.

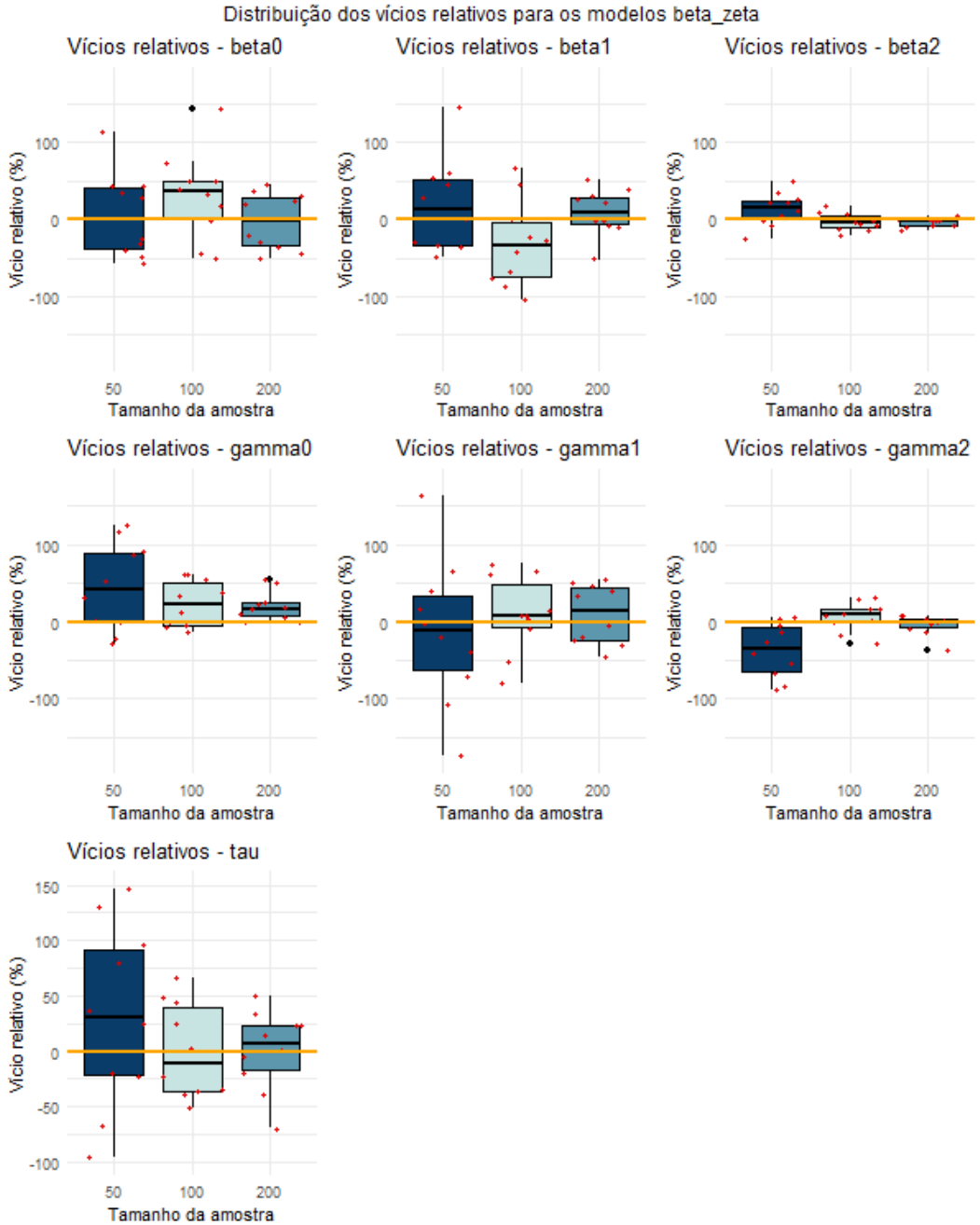

Figura 15 - Vícios relativos para os parâmetros das simulações do modelo beta_zeta.

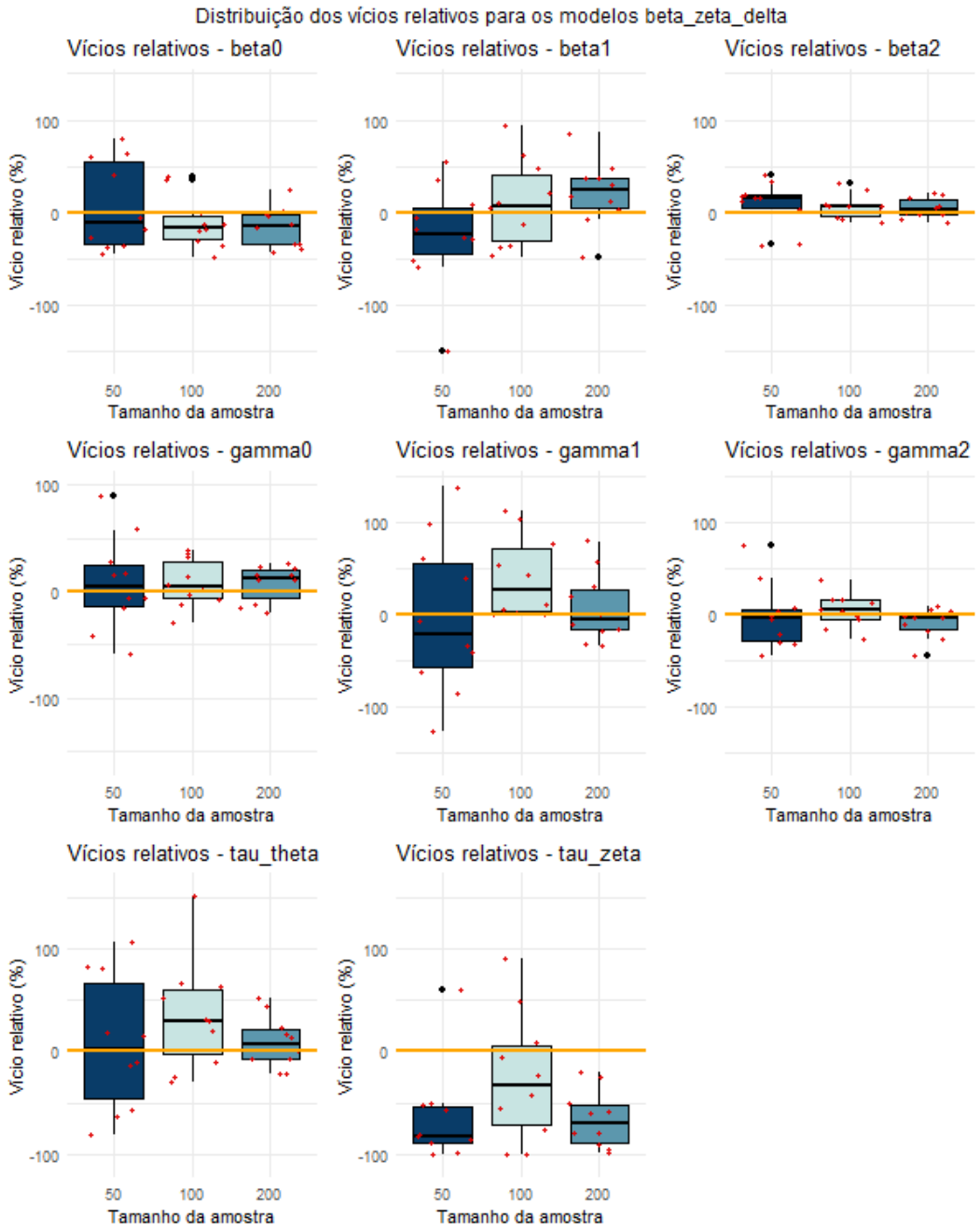

Figura 16 - Vícios relativos para os parâmetros das simulações do modelo beta_zeta_delta.

Os resultados para o estudo Monte Carlo do modelo beta_zeta_trocado, que considerou estrutura espacial no parâmetro de dispersão para os dados que não apresentavam esta estrutura, são apresentados na Figura 17. A estimação dos coeficientes foi satisfatória, como nos casos anteriores para os modelos da distribuição Beta. Os vícios relativos para a variância real do efeito aleatório espacial τ foram calculados utilizando as estimativas $\hat{\tau}_\theta$ fornecidas por este ajuste. Nota-se que incluir a modelagem de uma estrutura espacial não existente não prejudicou

o ajuste; as distribuições dos vícios para este parâmetro são similares às obtidas para o modelo beta_zeta, que modelava corretamente a estrutura dos dados beta_zeta.

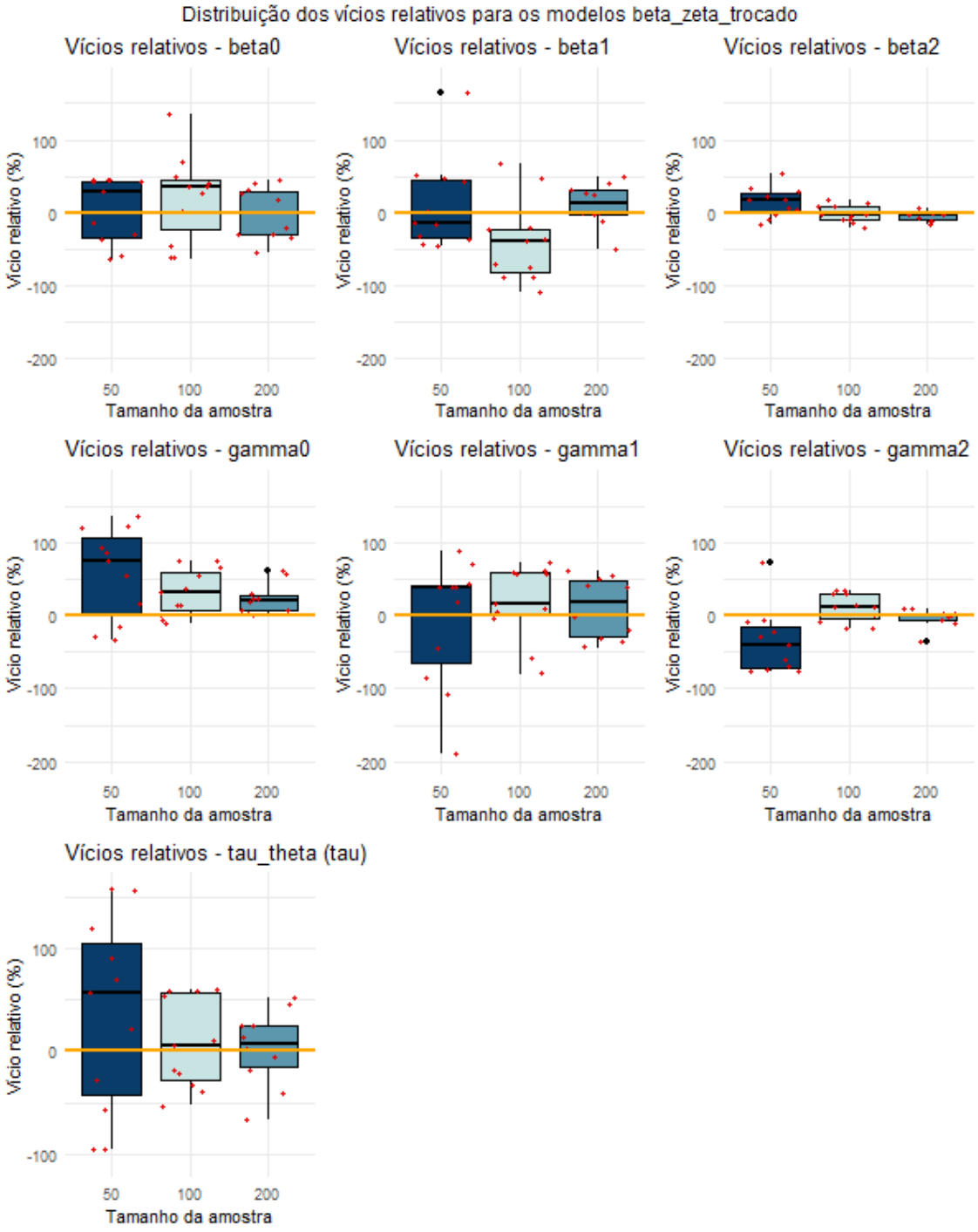

Figura 17 - Vícios relativos para os parâmetros das simulações do modelo beta_zeta_trocado.

A Figura 18 apresenta os vícios relativos para o estudo Monte Carlo do modelo beta_zeta_delta_trocado. Neste caso, a estrutura espacial existente na dispersão não foi modelada. Observa-se que isto impactou negativamente na modelagem do parâmetro τ_θ, que neste caso foi estimado como a única variância do efeito espacial do modelo. Os vícios relativos

apresentaram maior variabilidade e as medianas, para os três tamanhos amostrais, estão bem maiores que zero, indicando superestimação deste parâmetro. Como no ajuste correto para estes dados, apresentado na Figura 16, os vícios relativos para este parâmetro foram melhor distribuídos, suspeita-se que ignorar a estrutura espacial presente na dispersão tenha causado o mau ajuste desse parâmetro.

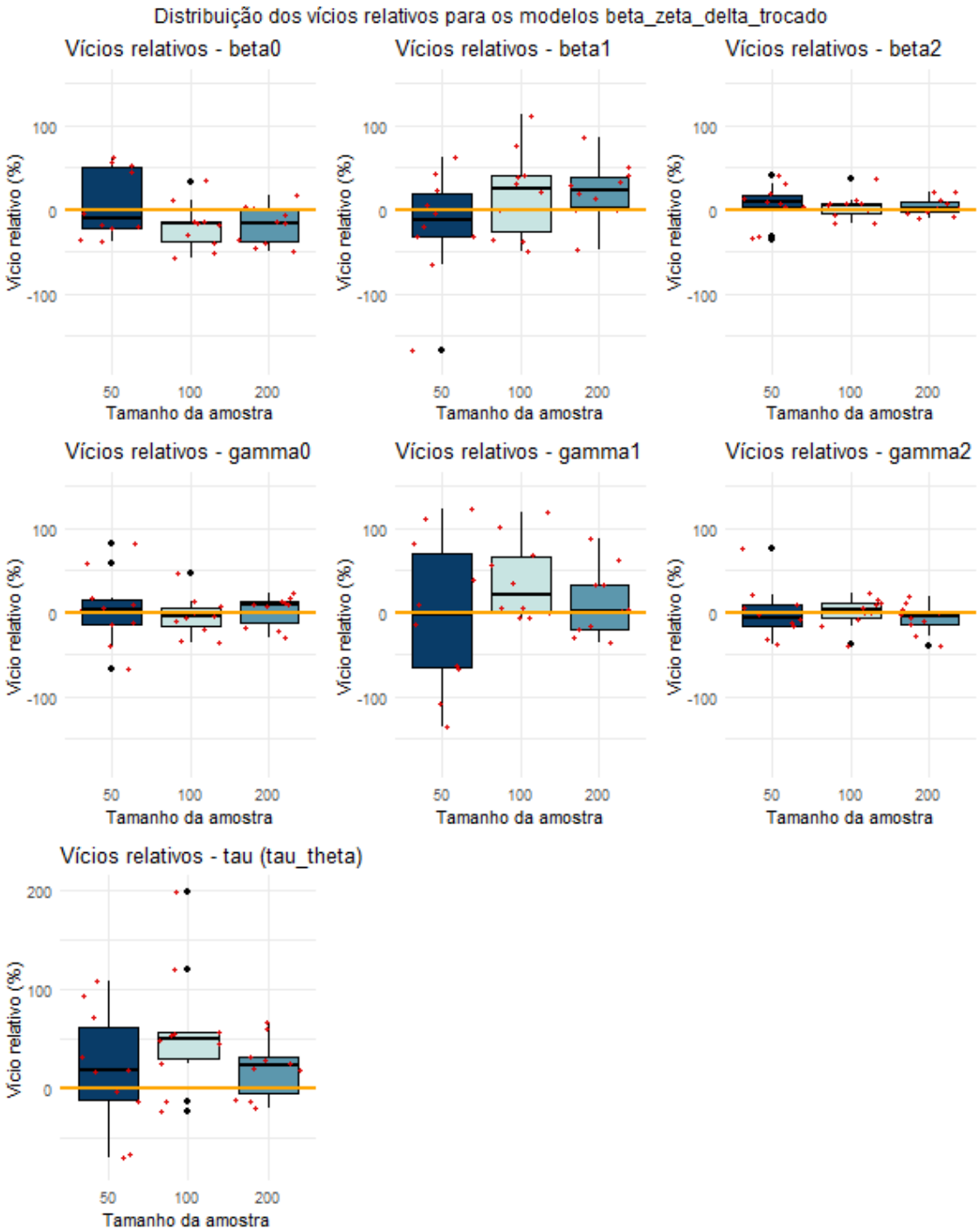

Figura 18 - Vícios relativos para os parâmetros das simulações do modelo beta_zeta_delta_trocado.

Os resultados dos estudos Monte Carlo para os cenários construídos indicam que ignorar uma estrutura espacial existente nos dados é prejudicial para a modelagem. Já o contrário, modelar uma estrutura não existente, não leva a impactos negativos na modelagem, apresentando resultados semelhantes aos obtidos com a modelagem correta da estrutura dos dados. No entanto, como ambos modelos trocados modificaram a modelagem apenas do efeito espacial do parâmetro de dispersão, modelou-se os dados com efeito espacial tanto na média quanto na dispersão utilizando um modelo que não considera qualquer efeito espacial. Assim, as três amostras dos dados beta_zeta_delta foram ajustadas com um MLG Beta e aplicou-se o teste I de Moran sobre seus resíduos de Pearson, para avaliar a existência de associação espacial nesses resíduos. Os resultados destes testes são apresentados na Tabela 8.

Tabela 8 - Valores-p para o teste I de Moran aplicado aos dados simulados.

Amostra	Valor-p
N=50	< 0,01
N=100	< 0,01
N=200	< 0,01

Como os valores-p para os testes foram inferiores a 5%, conclui-se que havia associação espacial nos resíduos desses modelos. Ou seja, o efeito espacial aleatório existente nesses dados não foi tratado pelo modelo e, consequentemente, foi para os resíduos.

As análises dos resultados pontuais dos cenários considerados, juntamente com os resultados dos estudos Monte Carlo, indicaram que ajustar um modelo com estrutura espacial não existente nos dados não prejudica a modelagem. Pelos resultados do estudo Monte Carlo e pelos testes de I de Moran, conclui-se que não modelar a estrutura espacial de dados que a apresentam, ou que suspeita-se que a apresentam, é prejudicial para o ajuste. Assim, optou-se por modelar os dados reais das taxas de abandono escolar do Ensino Médio com o modelo beta_zeta_delta, que considera efeito espacial tanto na estrutura da média quanto na da dispersão. Vale notar que, apesar do desempenho do modelo gama ter sido bom, optou-se por seguir para a análise dos dados reais com o MLG Beta pois as taxas de abandono se adequam bem à distribuição Beta, além de ter limites correspondentes ao suporte desta distribuição.

Adicionalmente, vale notar que nos estudos Monte Carlo, para todos os cenários considerados, os coeficientes das covariáveis contínuas foram bem estimados, com vícios relativos muito baixos em todos os tamanhos amostrais estudados. Como já discutido na Seção 2, todas as covariáveis explicativas consideradas para as taxas de abandono são contínuas, sendo que parte delas é limitada ao intervalo de 0 a 1. Baseando-se nos resultados simulados,

espera-se que as estimativas para os coeficientes β_i e γ_i sejam condizente com a realidade, levando a um bom ajuste do modelo.

4.1 Aplicação aos dados reais

Os dados reais das taxas de abandono escolar para o Ensino Médio foram modelados utilizando-se o modelo beta_zeta_delta. Dessa forma, modelou-se a estrutura da média, neste caso a taxa de abandono, com as covariáveis explicativas taxa de escolas na área urbana, taxa de escolas públicas, IDHM Renda, população e PIB *per capita*, sendo as duas últimas na escala logarítmica. Além disso, foi modelada a estrutura espacial para a média. A estrutura da dispersão foi modelada dessa mesma forma, com as mesmas covariáveis e com a inclusão do efeito aleatório espacial.

A Tabela 9 apresenta as médias *a posteriori* para os parâmetros estimados, bem como seus desvios padrões e os limites de seus intervalos HPD, para as amostras originais obtidas para o bloco de ensino EM, nos três anos sob estudo. De forma geral, pode-se perceber que os sinais dos coeficientes para a média β_i tendem a se manter os mesmos para os três anos. Estes coeficientes para IDHM Renda e para o logaritmo do PIB *per capita* apresentam sinais negativos em todos os anos, enquanto a taxa de escolas na área urbana apresenta coeficiente negativo apenas em 2020 e 2015. Isto sugere que estes fatores atuam em benefício da taxa de abandono, no sentido de que, quanto maiores os valores destas variáveis para um determinado município, menor será a taxa de abandono escolar para o EM. Já as demais covariáveis, taxa de escolas públicas e logaritmo da população, contribuem para o aumento da taxa de abandono nesses municípios.

Apesar dos sinais dos coeficientes coincidirem na maioria dos anos, as suas significâncias não se mantiveram as mesmas. Os coeficientes β_i para a taxa de escolas públicas do município foram considerados significativos para os três anos sob estudo, uma vez que seus intervalos HPD não incluíam o zero. Para 2015 a taxa de escolas na área urbana também foi considerada significativa, apresentando sinal oposto ao da taxa de escolas públicas. Vale notar que, para 2020, o limite superior do intervalo HPD para o coeficiente da taxa de escolas públicas está muito próximo de zero, com a maior porção do intervalo na área negativa; isto coincide

com o valor deste coeficiente estimado para o ano de 2015, indicando que este é um fator potencialmente influente sobre a taxa de abandono escolar.

Tabela 9 - Médias *a posteriori* para os coeficientes das amostras originais.

Ano	Coeficiente	Variável	Média	D.P.	Intervalo HPD
2020	β_i	Intercepto	-1,720	2,049	(-5,603; 2,492)
		Taxa urbana	-0,755	0,473	(-1,699; 0,136)
		Taxa pública	1,983	0,477	**(1,068; 2,900)**
		IDHM Renda	-2,110	2,426	(-6,733; 2,658)
		log(População)	0,052	0,144	(-0,212; 0,339)
		log(PIB *per capita*)	-0,167	0,258	(-0,659; 0,331)
	γ_i	Intercepto	1,654	2,141	(-2,297; 5,940)
		Taxa urbana	1,675	0,536	**(0,656; 2,766)**
		Taxa pública	0,347	0,500	(-0,618; 1,319)
		IDHM Renda	-1,435	2,579	(-6,314; 3,612)
		log(População)	0,096	0,164	(-0,242; 0,397)
		log(PIB *per capita*)	-0,035	0,289	(-0,577; 0,545)
2015	β_i	Intercepto	-4,265	1,696	(-7,405; -0,878)
		Taxa urbana	-0,642	0,321	**(-1,232; -0,004)**
		Taxa pública	2,745	0,391	**(2,030; 3,561)**
		IDHM Renda	-2,634	2,000	(-6,846; 1,028)
		log(População)	0,171	0,095	(-0,016; 0,355)
		log(PIB *per capita*)	-0,050	0,175	(-0,393; 0,297)
	γ_i	Intercepto	1,618	2,132	(-2,240; 5,988)
		Taxa urbana	1,582	0,470	**(0,634; 2,473)**
		Taxa pública	1,230	0,486	**(0,319; 2,211)**
		IDHM Renda	-0,049	2,470	(-4,624; 4,798)
		log(População)	-0,097	0,145	(-0,390; 0,179)
		log(PIB *per capita*)	0,101	0,238	(-0,371; 0,551)
2010	β_i	Intercepto	-1,978	1,598	(-5,317; 1,012)
		Taxa urbana	0,330	0,347	(-0,307; 1,027)
		Taxa pública	3,428	0,371	**(2,665; 4,139)**
		IDHM Renda	-3,121	2,027	(-7,557; 0,492)
		log(População)	0,085	0,113	(-0,131; 0,308)
		log(PIB *per capita*)	-0,244	0,190	(-0,615; 0,125)
	γ_i	Intercepto	5,958	1,918	**(2,282; 9,820)**
		Taxa urbana	0,669	0,443	(-0,179; 1,558)
		Taxa pública	0,333	0,448	(-0,617; 1,155)
		IDHM Renda	-0,126	2,381	(-4,784; 4,752)
		log(População)	0,071	0,153	(-0,222; 0,372)
		log(PIB *per capita*)	-0,430	0,226	(-0,897; -0,002)

Uma vez que a função de ligação utilizada para θ_i neste modelo é a *logit*, a interpretação direta dos coeficientes estimados recai sobre a razão de chances da taxa de abandono, ao invés de sobre a taxa em si. Ao adicionar uma unidade à qualquer covariável explicativa X_i, o

coeficiente β_i passa a ser multiplicado por $X_i + 1$ ao invés de apenas X_i, o que introduz o valor de β_i ao somatório exponenciado de $e^{X_i^T \beta_i}$. Dessa forma, para saber o impacto da adição de uma unidade ao valor da covariável X_i basta multiplicar a razão de chances original por e^{β_i}. Este raciocínio pode ser utilizado para a interpretação dos coeficientes β_i, referentes à estrutura da média, de todos os ajustes feitos. Para a amostra original de 2020, por exemplo, o coeficiente significativo foi $\beta_{Taxa\ de\ escolas\ públicas} = 1{,}983$, de forma que o aumento de uma unidade percentual na taxa de escolas públicas de um município acarreta em um aumento de $\left(e^{1{,}983/100} - 1\right) \times 100 = 2\%$ na razão de chances da taxa de abandono escolar neste município. Vale notar que, apesar do impacto do aumento ou da diminuição das covariáveis recair diretamente sobre a razão de chances, ele recai indiretamente sobre a taxa em si, uma vez que o aumento da taxa de abandono está relacionado ao aumento de sua razão de chances e vice versa.

Já os coeficientes γ_i se referem ao parâmetro de dispersão, de forma que seus valores representam o impacto da covariável na variabilidade da taxa de abandono. Para os três anos, este coeficiente para o IDHM Renda foi negativo, indicando que quanto maior o valor deste índice, menor é a variabilidade das taxas de abandono escolar no EM em Minas Gerais. Em 2015 e 2010 os coeficientes para o logaritmo do PIB *per capita* também foram negativos, indicando que para estes anos o aumento do PIB *per capita* está relacionado com menor variabilidade da taxa sob estudo. Os demais coeficientes foram positivos, indicando que suas respectivas covariáveis contribuem para o aumento da dispersão das taxas de abandono em Minas Gerais.

Os coeficientes γ_i significativos não coincidiram para os três anos. Em 2015 as taxas de escolas na área urbana e de escolas públicas impactam significativamente no aumento da variabilidade dos dados, enquanto que em 2020 apenas a taxa de escolas públicas contribui significativamente para este aumento. Já em 2010, apenas o intercepto foi considerado significativo, ao não incluir o zero em seu intervalo HPD. Vale notar, no entanto, que para este ano os limites inferiores dos intervalos HPD para as taxas de escolas urbanas e de escolas públicas estão muito próximos de zero, apesar de negativos, deixando a maior porção do intervalo na área positiva. Isto indica que estas covariáveis podem ter influência considerável sobre a dispersão dos dados, o que coincide com os resultados observados em 2020 e 2015.

A Tabela 10 apresenta as médias *a posteriori* e seus respectivos desvios padrão para os parâmetros de variância dos efeitos espaciais das estruturas da média (τ_θ) e da dispersão (τ_ζ).

Em todos os anos a variância para a estrutura espacial da média foi pequena, destacando-se que em 2010 ela foi ainda menor que nos demais anos. A interpretação deste parâmetro recai sobre a variabilidade dos efeitos aleatórios espaciais estimados para os municípios da amostra. Quando a estimativa de τ_θ é próxima de zero, há indícios de que os efeitos estimados para os elementos amostrais variaram pouco com relação à estrutura da média, enquanto valores maiores desta estimativa indicam que existem efeitos aleatórios espaciais na amostra com valores muito discrepantes entre si.

Já as variâncias do efeito espacial da dispersão apresentaram valores maiores e que variaram mais de ano para ano. Em 2015 esta estimativa foi a de maior valor, indicando que esta amostra apresentou grande variabilidade para os efeitos aleatórios espaciais estimados. No ano de 2010 o valor para τ_ζ foi o menor, indicando que esta amostra apresentou pequena variabilidade para os efeitos espaciais para a estrutura da dispersão estimados.

Tabela 10 - Médias *a posteriori* para τ_θ e τ_ζ.

Ano	Parâmetro	Média	D.P.
2020	τ_θ	0,140	0,191
	τ_ζ	1,019	0,964
2015	τ_θ	0,175	0,178
	τ_ζ	2,896	1,028
2010	τ_θ	0,060	0,084
	τ_ζ	0,174	0,296

A Figura 19 a seguir apresenta as distribuições das estimativas obtidas para os efeitos aleatórios espaciais, tanto para a estrutura da média quanto para a da dispersão. Pode-se observar que as formas assumidas pelos *box-plots* refletem as respectivas estimativas obtidas para as variâncias τ_θ e τ_ζ. O *box-plot* de maior amplitude é justamente aquele para os valores $\Delta_{i,\zeta}$, cuja estrutura de variância é, em parte, definida pela estimativa $\hat{\tau}_\zeta = 2{,}896$, que foi a maior variância estimada dentre as três amostras, considerando-se tanto a estrutura de variância quanto a da média. Nota-se, ainda, que as medianas dos *box-plots* estão próximas de zero, refletindo a distribuição Normal multivariada com o vetor de médias sendo um vetor de zeros.

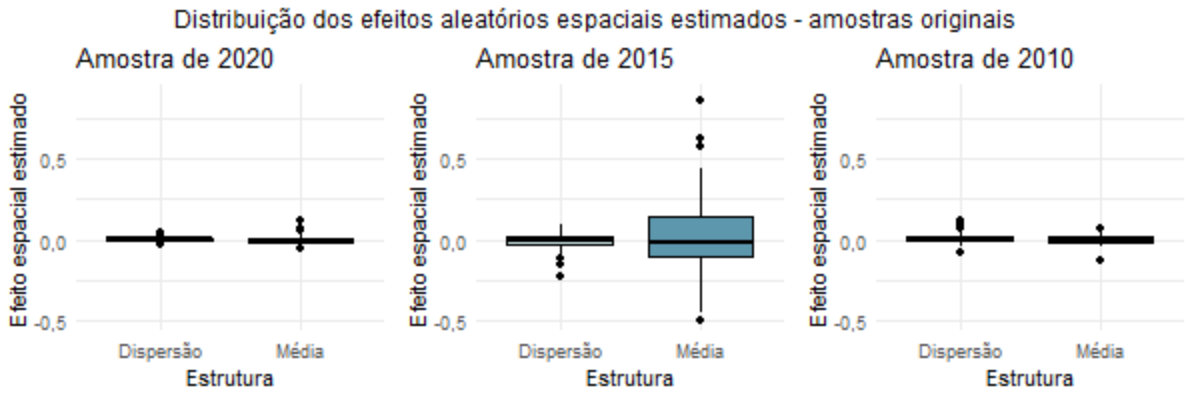

Figura 19 - Distribuições dos efeitos aleatórios espaciais estimados (amostras originais).

A Tabela 11 a seguir apresenta as estimativas dos coeficientes β_i e γ_i obtidas dos ajustes das amostras modificadas, sem as taxas iguais a zero. Com relação aos coeficientes da estrutura da média, nota-se que alguns resultados não coincidem com aqueles obtidos com as amostras originais. O sinal negativo nas estimativas dos coeficientes para a covariável taxa de escolas na área urbana e o sinal positivo para os coeficientes da covariável taxa de escolas públicas foram os mesmos para ambos grupos de amostras. Já o logaritmo da população apresentou sinal negativo, indicando que municípios com maior população tendem a apresentar menores taxas de abandono escolar, conclusão contrária à obtida com as amostras originais. O regressor linear estimado para o logaritmo do PIB *per capita* apresentou sinal positivo para os anos 2020 e 2015, e negativo para 2010, de forma que os resultados para esses dois anos não coincidem com os obtidos para a amostra original. Por fim, o coeficiente estimado para o IDHM Renda diferiu com relação ao caso original apenas no ano de 2020, em que apresentou sinal positivo. Vale notar, no entanto, que esta estimativa foi próxima de zero e os limites inferior e superior de seu intervalo HPD estão distantes de zero e quase à mesma distância de zero, indicando que este coeficiente não é significativo.

Levando-se em conta apenas os resultados significativos, ou seja, coeficientes estimados cujos intervalos HPD não contêm o zero, as conclusões obtidas são similares àquelas para a estrutura da média para as amostras originais. Em 2020 e 2015 a taxa de escolas na área urbana foi considerada um fator significativo na diminuição da taxa de abandono escolar, enquanto que para os anos de 2015 e 2010 a taxa de escolas públicas apresenta influência significativa sobre a taxa de abandono, de forma que quanto maior a proporção de escolas públicas no município, maior será sua taxa de abandono escolar para o EM.

Tabela 11 - Médias *a posteriori* para os coeficientes das amostras sem zeros.

Ano	Coeficiente	Variável	Média	D.P.	Intervalo HPD
2020	β_i	Intercepto	-2,841	1,835	(-6,440; 0,711)
		Taxa urbana	-0,988	0,342	**(-1,732; -0,352)**
		Taxa pública	0,333	0,411	(-0,432; 1,137)
		IDHM Renda	0,672	2,321	(-4,003; 5,083)
		log(População)	-0,053	0,106	(-0,251; 0,151)
		log(PIB *per capita*)	0,015	0,206	(-0,368; 0,42)
	γ_i	Intercepto	0,870	2,315	(-3,784; 5,085)
		Taxa urbana	0,540	0,560	(-0,501; 1,669)
		Taxa pública	1,060	0,542	**(0,044; 2,129)**
		IDHM Renda	-0,195	2,857	(-5,653; 5,480)
		log(População)	0,165	0,169	(-0,158; 0,501)
		log(PIB *per capita*)	-0,056	0,312	(-0,694; 0,514)
2015	β_i	Intercepto	-3,296	1,733	**(-6,847; -0,171)**
		Taxa urbana	-0,958	0,289	**(-1,509; -0,355)**
		Taxa pública	1,334	0,297	**(0,756; 1,912)**
		IDHM Renda	-0,618	1,871	(-4,102; 3,250)
		log(População)	-0,014	0,088	(-0,188; 0,161)
		log(PIB *per capita*)	0,022	0,139	(-0,246; 0,289)
	γ_i	Intercepto	0,440	2,579	(-4,307; 5,646)
		Taxa urbana	1,045	0,717	(-0,395; 2,392)
		Taxa pública	0,765	0,645	(-0,427; 2,086)
		IDHM Renda	-0,796	2,672	(-5,839; 4,424)
		log(População)	-0,386	0,201	**(-0,787; -0,006)**
		log(PIB *per capita*)	0,790	0,390	**(0,059; 1,541)**
2010	β_i	Intercepto	-1,092	1,419	(-3,926; 1,591)
		Taxa urbana	-0,326	0,292	(-0,892; 0,243)
		Taxa pública	2,057	0,224	**(1,641; 2,509)**
		IDHM Renda	-2,202	1,951	(-6,122; 1,530)
		log(População)	-0,083	0,092	(-0,273; 0,088)
		log(PIB *per capita*)	-0,065	0,161	(-0,36; 0,273)
	γ_i	Intercepto	6,198	1,995	**(2,299; 10,114)**
		Taxa urbana	-0,002	0,534	(-1,030; 1,018)
		Taxa pública	-1,439	0,432	**(-2,22; -0,512)**
		IDHM Renda	2,207	2,584	(-2,789; 7,281)
		log(População)	-0,106	0,164	(-0,425; 0,214)
		log(PIB *per capita*)	-0,221	0,245	(-0,701; 0,244)

Com respeito aos coeficientes γ_i, referentes à estrutura da dispersão dos dados, observa-se que os resultados para 2020 e 2015 parecem ser, de forma geral, opostos àqueles para 2010. Nesses dois anos as taxas de escolas na área urbana e escolas públicas apresentaram coeficientes estimados com sinal positivo, indicando que à medida que os valores dessas covariáveis aumentam, dispersão da taxa de abandono escolar aumenta. Já o coeficiente da covariável IDHM Renda apresenta sinal negativo para 2020 e 2015, indicando que o aumento desta

covariável diminui a variabilidade das taxas de abandono. Para o ano de 2010 foram estimados coeficientes que indicam o contrário dessas interpretações sobre a dispersão.

Como já mencionado anteriormente, apesar dos sinais e valores dos coeficientes variarem muito de amostra para amostra, apenas aqueles que não incluem o zero em seu intervalo HPD podem ser considerados significativos. Em 2020 e 2010 apenas a taxa de escolas públicas impacta significativamente na dispersão da taxa de abandono escolar para o EM; no entanto, em 2020 a interpretação é a de que esta covariável colabora para o aumento da dispersão das taxas, enquanto que em 2010 ela colabora para a diminuição dessa dispersão. Em 2015 o logaritmo da população do município contribui para a diminuição da dispersão dos dados, enquanto o logaritmo do PIB *per capita* contribui para seu aumento. Vale notar que os resultados significativos obtidos para esta estrutura não coincidiram com aqueles obtidos com base nas amostras originais.

Em geral, as interpretações dos resultados referentes às estruturas da média, para as amostras originais e para as sem os zeros, coincidiram. Apenas para a estrutura do parâmetro de dispersão as conclusões divergiram. Isto pode ter ocorrido justamente pela remoção dos valores iguais a zero, o que interferiu, ainda que pouco, na forma da distribuição dos dados e, consequentemente, na sua dispersão.

A Tabela 12 a seguir apresenta as estimativas para as variâncias dos efeitos aleatórios espaciais para a estrutura da média e a da dispersão. As estimativas foram pequenas e similares entre as amostras de 2020 e 2010, indicando menor variabilidade dos efeitos aleatórios espaciais estimados para ambas estruturas. Já em 2015 ambas estimativas foram maiores que as dos outros anos, particularmente a estimativa para τ_θ, que apresentou o maior valor para a estrutura da média, inclusive em comparação com os valores estimados para as amostras originais. No geral, nota-se que para as amostras sem taxas iguais a zero as variâncias para as estruturas espaciais foram menores em comparação com aquelas obtidas com os dados completos.

Tabela 12 - Médias *a posteriori* para τ_θ e τ_ζ para as amostras sem zeros.

Ano	Parâmetro	Média	D.P.
2020	τ_θ	0,115	0,217
	τ_ζ	0,106	0,397
2015	τ_θ	1,151	0,852
	τ_ζ	0,602	1,118
2010	τ_θ	0,172	0,191
	τ_ζ	0,391	0,672

As distribuições dos efeitos aleatórios espaciais estimados $\widehat{\Delta}_{i,\theta}$ e $\widehat{\Delta}_{i,\zeta}$ são representadas na Figura 20. Assim como ocorreu para as amostras originais, essas estimativas giram em torno de zero, como é esperado pela distribuição que seguem, e as amplitudes dos *box-plots* refletem os valores estimados para as variâncias de suas respectivas estruturas apresentadas na Tabela 12.

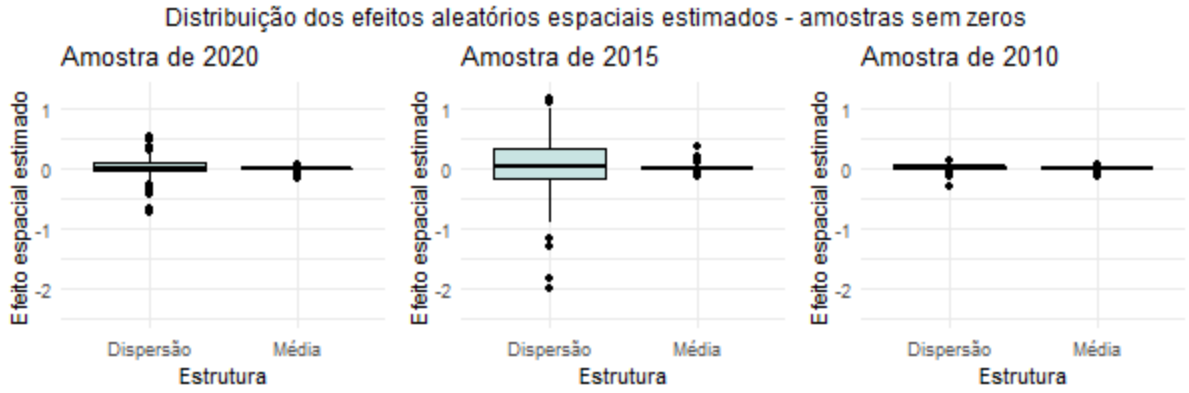

Figura 20 - Distribuições dos efeitos aleatórios espaciais estimados (amostras sem zeros).

As convergências dos valores nas modelagens foram verificadas pela análise dos *traceplots* das cadeias para os parâmetros de interesse. A Figura 21 apresenta os *traceplots* para os coeficientes β_i e γ_i, bem como para as variâncias τ_θ e τ_ζ, para o ajuste feito com a amostra original de 2020. As cadeias convergiram bem para todos os coeficientes; já para as variâncias, as cadeias não parecem convergir tão bem, mas pode-se observar que variaram sobre um intervalo pequeno de valores, indicando que a convergência dessas cadeias foi satisfatória. Comportamentos similares foram observados para as cadeias dos demais ajustes feitos.

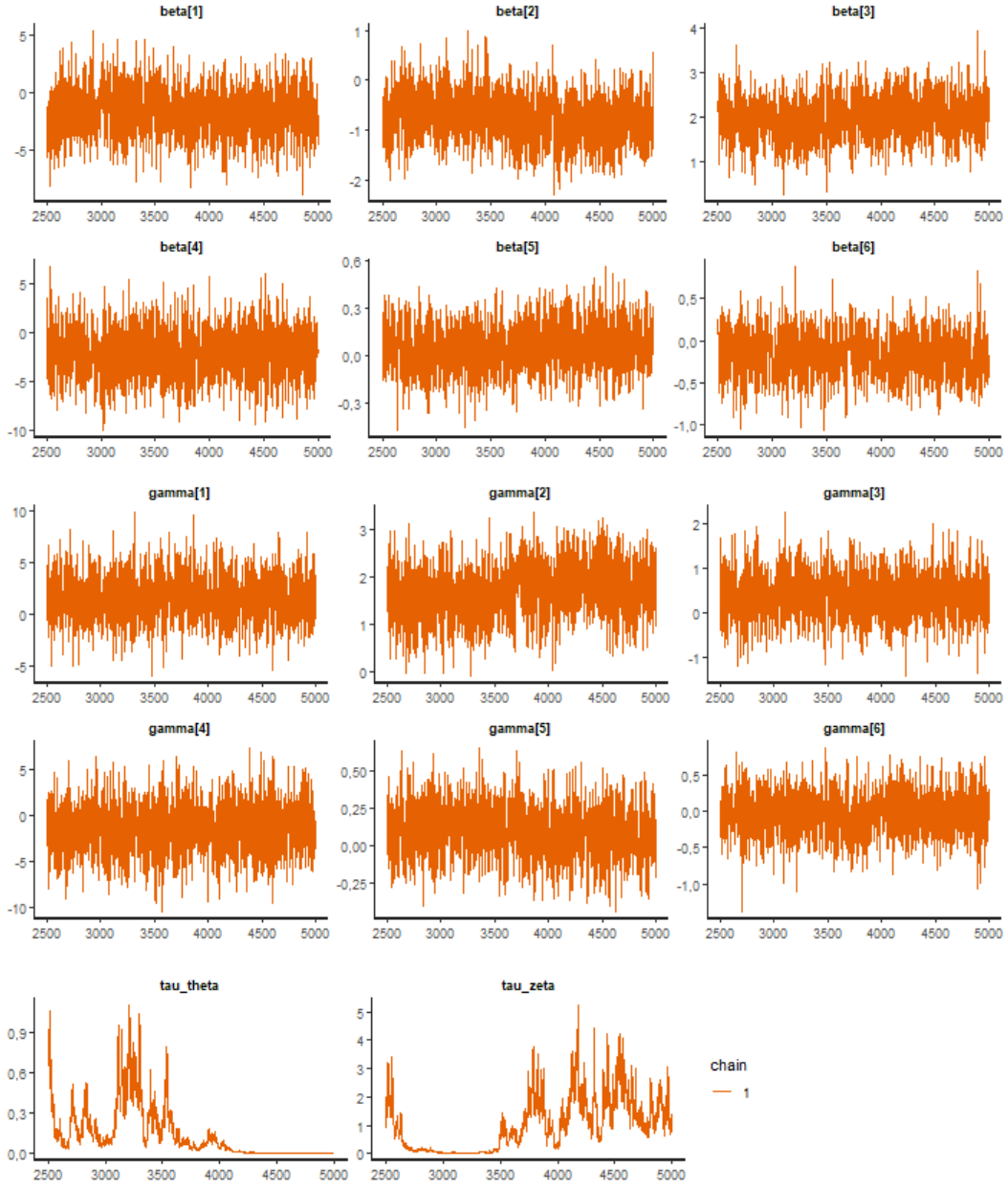

Figura 21 - *Traceplots* para as cadeias do ajuste da amostra original de 2020.

As Figuras 22, 23, 24, 25, 26 e 27 correspondem aos gráficos dos resíduos de Pearson para os modelos ajustados contra os valores preditos e as covariáveis presentes nos modelos. Nos gráficos das Figuras 22 e 25, referentes às amostras de 2020 com e sem as taxas iguais a zero, respectivamente, foi removido um *outlier* que estava alterando a escala dos gráficos e dificultando a avaliação da dispersão dos pontos. Estes resíduos são referentes ao município que apresentou a taxa de abandono escolar extrema de 50% em 2020.

No geral, em todos os gráficos os pontos estão dispersos em torno do eixo zero, sem formar qualquer configuração de pontos indesejada que caracterize um ajuste ruim. Nos gráficos para os resíduos contra a taxa de escolas públicas, pode-se observar uma concentração

de pontos na extremidade direita do eixo das abcissas, causada pela alta frequência de municípios com 100% de suas escolas de Ensino Médio fazendo parte da rede pública. Nas Figuras 22, 23 e 24, referentes às amostras com os zeros, os gráficos dos resíduos contra as covariáveis incluídas na modelagem da estrutura da média apresentam uma nuvem de pontos mais densa ao longo do eixo das abcissas, na região negativa com relação ao eixo das ordenadas. Já nos gráficos para as amostras sem zeros, apresentados nas Figuras 25, 26 e 27, estas concentrações de pontos não existem, indicando que elas eram formadas pelos resíduos das observações com taxa de abandono igual a zero.

Dados estes gráficos para os resíduos, pode-se concluir que o ajuste desses modelos foi satisfatório. A forma geral dos gráficos não indicou variação não constante dos dados nem a necessidade de alguma transformação sobre a variável resposta.

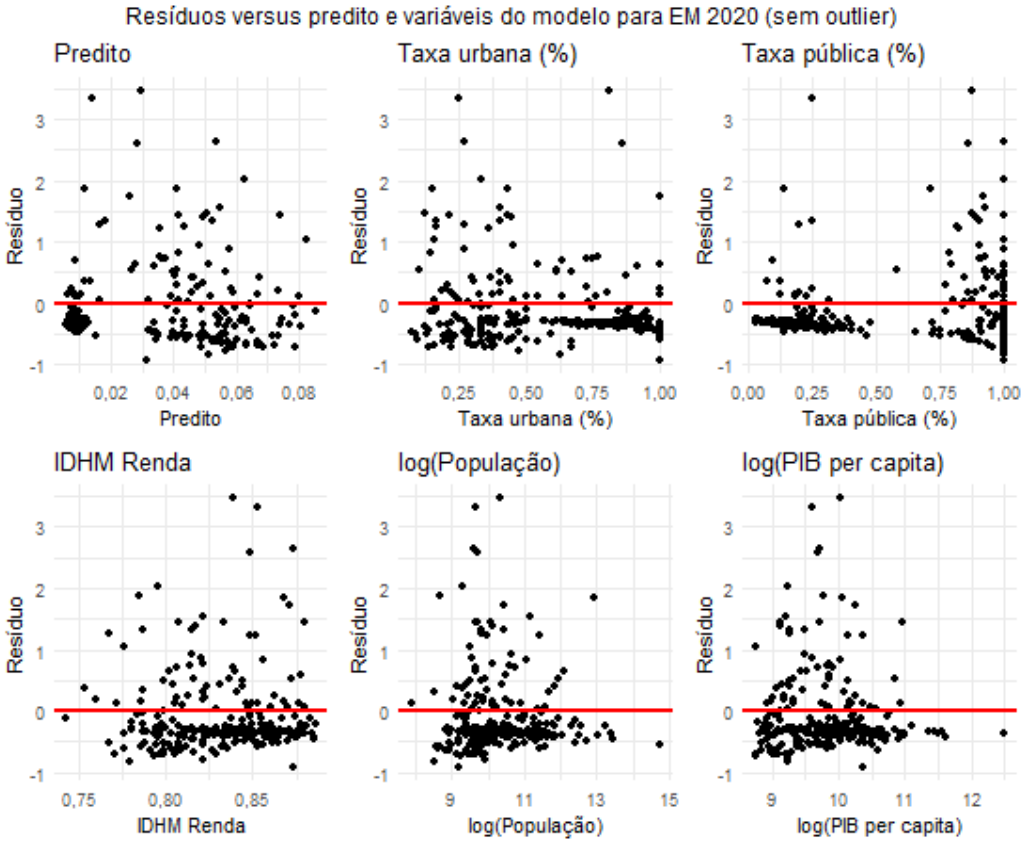

Figura 22 - Gráficos de dispersão dos resíduos para a amostra de 2020.

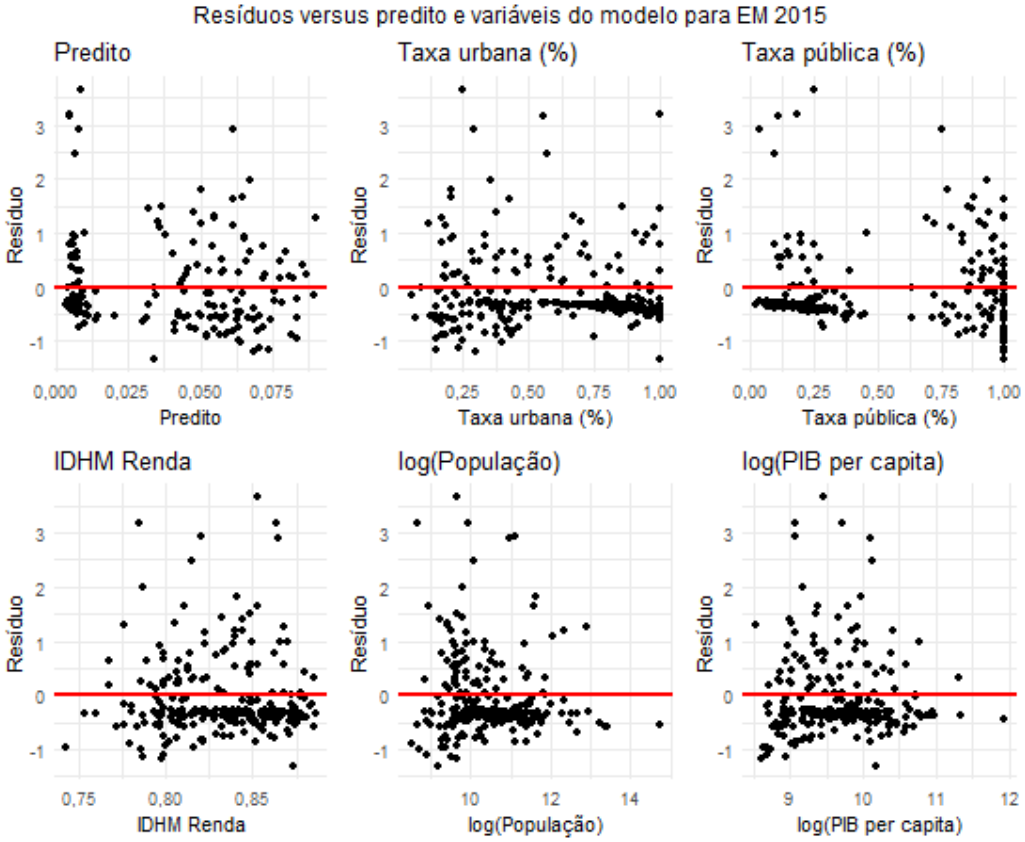

Figura 24 - Gráficos de dispersão dos resíduos para a amostra de 2015.

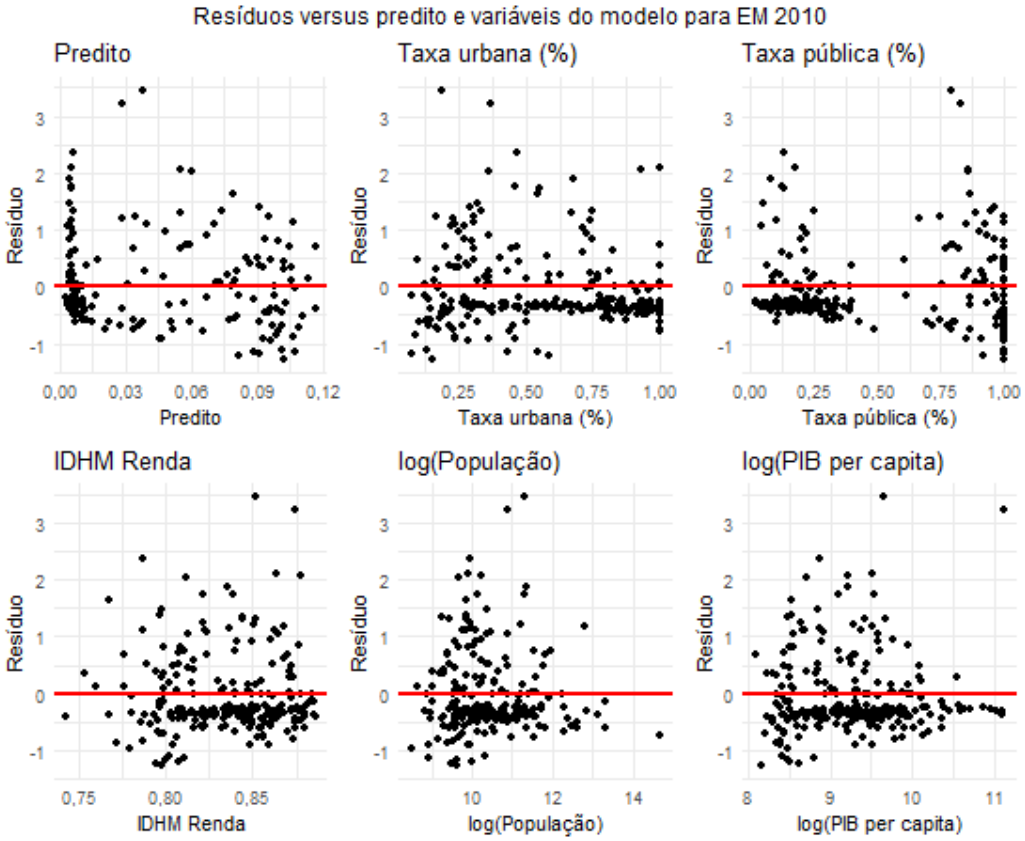

Figura 23 - Gráficos de dispersão dos resíduos para a amostra de 2010.

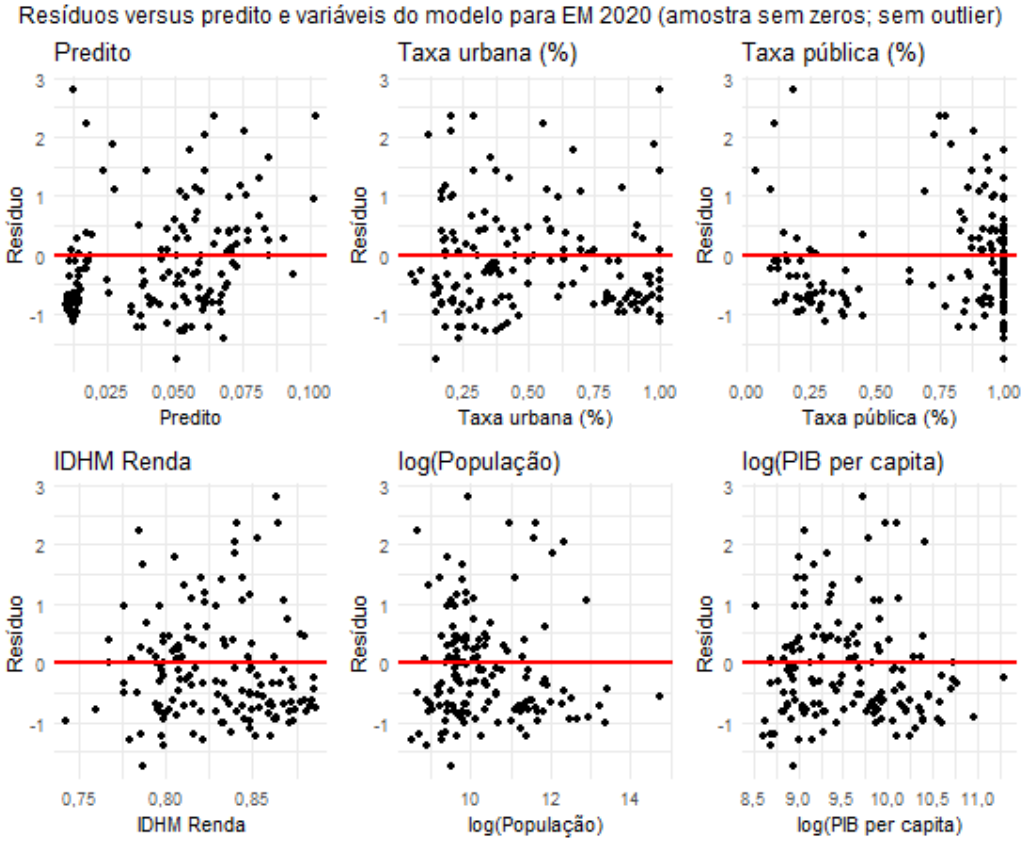

Figura 25 - Gráficos de dispersão dos resíduos para a amostra de 2020 sem zeros.

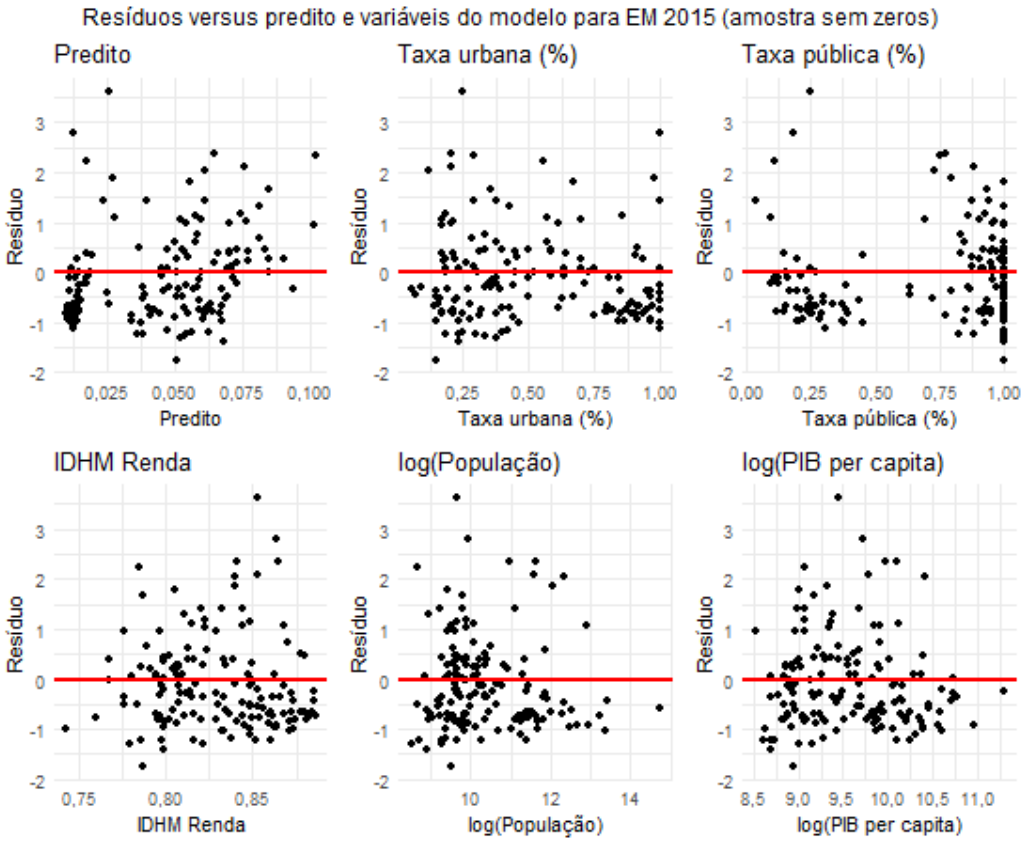

Figura 26 - Gráficos de dispersão dos resíduos para a amostra de 2015 sem zeros.

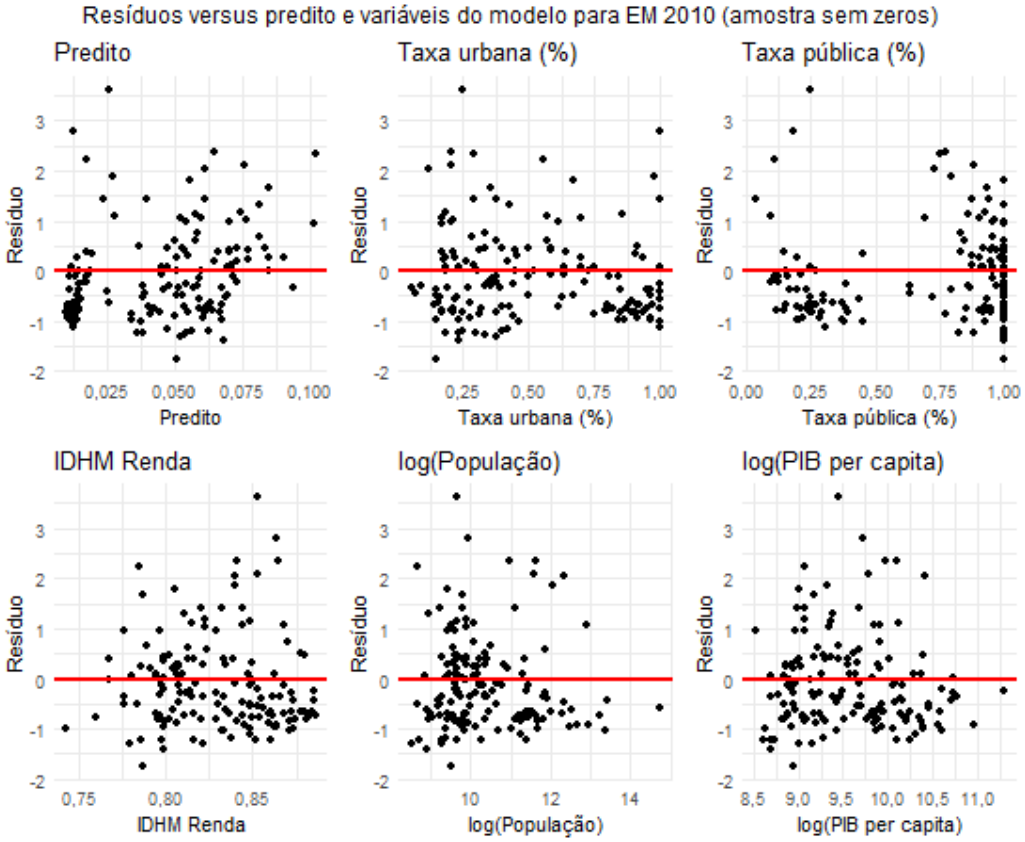

Figura 27 - Gráficos de dispersão dos resíduos para a amostra de 2010 sem zeros.

Por fim, aplicou-se o teste I de Moran aos resíduos calculados para os modelos, a fim de verificar se as estruturas espaciais foram devidamente tratadas pelo modelo. A Tabela 13 a seguir apresenta os valores-p obtidos para os testes aplicados aos resíduos de cada uma das amostras. Como em todos os casos os valores-p foram superiores a 5%, há evidência estatística de que não existe associação espacial entre os resíduos. Assim, confirma-se que o modelo foi capaz de captar a informação espacial dos dados e incluí-la na estimação dos parâmetros de interesse.

Tabela 13 - Valores-p para o teste I de Moran para os resíduos dos modelos das taxas.

Amostra	Valor-p
EM 2020	0,877
EM 2015	0,316
EM 2010	0,202
EM 2020 (sem zeros)	0,736
EM 2015 (sem zeros)	0,382
EM 2010 (sem zeros)	0,598

5 Conclusão

Pelos resultados obtidos com os ajustes tanto para as amostras originais quanto para as amostras sem as taxas zero, os fatores que apresentaram influência significativa sobre a taxa de abandono escolar nos municípios do estado de Minas Gerais na maioria dos anos foram as taxas de escolas públicas e de escolas na área urbana. A taxa de escolas na rede pública está relacionada com taxas de abandono escolar do ensino médio mais altas, ou seja, quanto maior a proporção de escolas públicas em um município maior tenderá a ser a sua taxa de abandono. Já a taxa de escolas na área urbana está relacionada com a diminuição da taxa de abandono.

Além disso, como estes dois fatores permaneceram significativos na maioria das amostras, tanto as originais quanto as sem taxas iguais a zero, considerando os três anos sob estudo, conclui-se que os principais fatores capazes de explicar a taxa de abandono escolar do ensino médio, em Minas Gerais, não mudaram ao longo dos anos da última década.

Referências

Draper, N.R., Smith, H. (1981) **Applied regression analysis**. 2 ed. New York: John Wiley.

Dobson, A. J., Barnett, A. G. (2008) **An Introduction to Generalized Linear Models**, 3rd ed., Boca Raton: Chapman & Hall/CRC.

Ferrari, S., & Cribari-Neto, F. (2004). **Beta regression for modelling rates and proportions**. *Journal of applied statistics*, *31*(7), 799-815.

Prado, K. S. (2021). **Municipios-Brasileiros**. Disponível em: https://github.com/kelvins/Municipios-Brasileiros. Acessado em 27 de julho de 2021.

Bolstad, W. M. (2007) **Introduction to Bayesian Statistics**, 2ed. John Wiley and Sons.

Banerjee, S., Carlin, B.P., Gelfand, A.E. (2014) **Hierarchical Modeling and Analysis for Spatial Data**. 2 ed. Chapman and Hall/CRC.

Getis, A., Ord, K. (1992). **The Analysis of Spatial Association by Use of Distance Statistics**. Geographical Analysis. 24. 189 - 206.

Cribari-Neto F., Zeileis A. (2010). **Beta Regression in R**. *Journal of Statistical Software*, **34** (2), 1-24.

Stan Development Team (2021). **Stan Modeling Language Users Guide and Reference Manual**, 2.18. https://mc-stan.org/.

R Core Team (2021). **R: A language and environment for statistical computing.** R Foundation for Statistical Computing, Vienna, Austria. URL https://www.r-project.org/.

Stan Development Team (2020). **RStan: the R interface to Stan**. R package version 2.21.2. https://mc-stan.org/.

Google (2018). **Google Colaboratory**. https://colab.research.google.com/notebooks/intro.ipynb (19 de janeiro de 2022).

Gabry J., Cesnovar R. (2021). **cmdstanr: R Interface to 'CmdStan'**. https://mc-stan.org/cmdstanr, https://discourse.mc-stan.org.

Anexo

ANEXO I

Municípios de Minas Gerais que fazem parte das amostras estudadas.

Município	2010	2015	2020	2010 (sem zero)	2015 (sem zero)	2020 (sem zero)
Açucena	x	x	x	x	x	x
Água Boa	x	x	x	x	x	x
Águas Formosas			x			
Águas Vermelhas	x	x	x	x	x	x
Aimorés	x	x	x		x	x
Além Paraíba	x	x	x			
Alfenas	x	x	x	x	x	x
Almenara	x	x	x			x
Alpinópolis	x	x	x			
Andradas	x	x	x	x		
Antônio Carlos		x	x		x	x
Antônio Dias		x	x		x	x
Araçuaí	x	x	x			x
Araguari	x	x	x	x	x	x
Araponga	x	x	x	x	x	x
Araxá	x	x	x			
Arcos	x	x	x			
Arinos	x	x	x			
Ataléia	x	x	x	x	x	x
Baependi	x	x	x	x	x	
Bambuí	x	x	x	x	x	
Barão de Cocais	x	x	x			
Barbacena	x	x	x	x	x	x
Barroso	x	x	x	x	x	
Belo Horizonte	x	x	x	x	x	x
Berilo	x	x	x	x	x	x
Betim	x	x	x		x	
Bicas	x	x			x	
Boa Esperança	x	x	x	x		
Bocaiúva	x	x	x		x	
Bom Despacho	x	x	x			
Bom Jesus do Galho	x	x	x	x	x	x
Bonito de Minas		x	x		x	x
Borda da Mata		x				
Brasilândia de Minas	x	x	x	x	x	x
Brasília de Minas	x	x	x	x	x	
Brazópolis	x	x	x	x	x	x
Brumadinho	x	x	x			
Buritis	x	x	x		x	
Buritizeiro	x	x	x	x	x	x
Cachoeira de Pajeú			x			x
Caeté	x	x	x			
Caldas			x			x
Camanducaia	x	x	x			
Cambuí	x	x	x			
Campanha	x	x	x		x	
Campina Verde	x	x	x		x	x
Campo Belo	x	x	x	x		

Município	2010	2015	2020	2010 (sem zero)	2015 (sem zero)	2020 (sem zero)
Campos Gerais	x	x	x			
Capelinha	x	x	x	x		
Capitão Enéas	x	x	x	x	x	x
Caraí	x	x	x	x	x	x
Carandaí	x	x	x			x
Carangola	x	x	x	x		
Caratinga	x	x	x	x		
Carlos Chagas	x	x	x			
Carmésia			x			x
Carmo do Cajuru	x	x	x	x	x	
Carmo do Paranaíba	x	x	x			
Carmo do Rio Claro	x	x	x			
Carneirinho	x	x	x	x	x	x
Cataguases	x	x	x	x	x	x
Caxambu	x	x	x	x		
Chapada do Norte	x	x	x	x	x	x
Chapada Gaúcha	x	x	x	x	x	x
Cláudio			x			
Comercinho	x	x	x	x	x	x
Conceição das Alagoas			x			
Conceição do Mato Dentro		x	x		x	
Cônego Marinho	x	x	x	x	x	x
Congonhas	x	x	x			x
Conselheiro Lafaiete	x	x	x			x
Conselheiro Pena	x	x	x	x		
Contagem	x	x	x	x	x	x
Coração de Jesus	x	x	x	x	x	x
Corinto	x	x	x			
Coroaci	x	x	x	x	x	x
Coromandel	x	x	x	x	x	x
Coronel Fabriciano	x	x	x	x		
Cruzília		x	x		x	x
Curvelo	x	x	x	x		
Delfim Moreira	x	x	x			
Diamantina	x	x	x			x
Divino	x	x	x	x	x	x
Divinópolis	x	x	x	x	x	x
Dores de Guanhães			x			x
Dores do Indaiá	x	x		x		
Entre Rios de Minas	x	x	x	x	x	x
Ervália			x			
Esmeraldas	x	x	x			x
Espera Feliz		x	x			x
Espinosa	x	x	x	x		x
Extrema	x	x	x			
Ferros		x	x		x	x
Fervedouro			x			x
Formiga	x	x	x		x	
Francisco Badaró			x			x
Francisco Sá	x	x	x			x
Frutal	x	x	x	x	x	x
Governador Valadares	x	x	x	x	x	x
Grão Mogol	x	x	x	x	x	x
Guanhães	x	x	x			x
Guaranésia			x			
Guaxupé	x	x	x			
Ibiá	x	x	x	x		
Ibiracatu		x	x		x	x

Município	2010	2015	2020	2010 (sem zero)	2015 (sem zero)	2020 (sem zero)
Ibirité	x	x	x	x	x	x
Icaraí de Minas	x	x	x	x	x	
Igarapé	x	x	x			
Indaiabira	x	x	x	x	x	x
Inhapim	x	x	x		x	x
Ipaba		x	x		x	
Ipatinga	x	x	x	x	x	x
Itabira	x	x	x	x	x	
Itabirito	x	x	x			
Itacarambi	x	x	x	x	x	x
Itajubá	x	x	x	x	x	x
Itamarandiba	x	x	x	x		x
Itamonte	x	x	x		x	x
Itanhandu			x			
Itanhomi			x			x
Itaobim	x	x	x	x		x
Itapecerica	x	x	x			
Itaú de Minas		x	x		x	x
Itaúna	x	x	x		x	
Itinga	x	x	x	x	x	x
Ituiutaba	x	x	x	x		
Iturama	x	x	x			
Jaboticatubas		x	x		x	x
Jacinto	x	x	x	x	x	x
Jacutinga	x	x	x			
Jaíba	x	x	x	x	x	
Janaúba	x	x	x		x	
Januária	x	x	x		x	
Jequeri	x	x	x	x	x	x
Jequitinhonha	x	x	x			x
João Monlevade	x	x	x	x	x	
João Pinheiro	x	x	x			
Juiz de Fora	x	x	x	x	x	x
Juvenília	x	x	x	x	x	x
Lagoa da Prata	x	x	x	x		
Lagoa Dourada	x					
Lagoa Formosa	x	x	x	x	x	x
Lagoa Santa	x	x	x			
Lajinha	x	x				
Lavras	x	x	x		x	x
Leme do Prado	x	x	x	x	x	x
Leopoldina	x	x	x	x		x
Lima Duarte	x	x	x	x	x	
Luz			x			
Machado	x	x	x	x	x	x
Malacacheta	x	x	x	x	x	x
Manga	x	x	x	x	x	x
Manhuaçu	x	x	x	x	x	
Manhumirim	x	x	x			
Mantena	x	x	x			
Maria da Fé	x	x	x			
Mariana	x	x	x	x		x
Martinho Campos	x	x	x	x	x	x
Mateus Leme	x	x	x	x		x
Matias Cardoso	x	x	x	x	x	x
Matipó	x	x	x			
Matozinhos	x	x	x			
Medina	x	x	x	x	x	x

Município	2010	2015	2020	2010 (sem zero)	2015 (sem zero)	2020 (sem zero)
Minas Novas	x	x	x	x	x	x
Montalvânia	x	x	x	x	x	x
Monte Alegre de Minas	x		x			x
Monte Azul	x	x	x			
Monte Carmelo	x	x	x		x	
Monte Santo de Minas	x	x	x			
Montes Claros	x	x	x	x	x	x
Muriaé	x	x	x	x	x	
Mutum	x	x	x			
Muzambinho	x	x	x	x	x	x
Nanuque	x	x	x			
Nepomuceno	x	x	x	x		
Ninheira		x	x		x	x
Nova Era	x	x	x			
Nova Lima	x	x	x		x	
Nova Porteirinha	x	x	x	x	x	x
Nova Serrana	x	x	x			
Novo Cruzeiro	x	x	x	x	x	x
Novo Oriente de Minas	x	x	x	x		x
Oliveira	x	x	x			
Ouro Branco	x	x	x			
Ouro Fino	x	x	x	x		
Ouro Preto	x	x	x			
Padre Paraíso	x	x	x	x	x	
Papagaios	x	x	x			x
Pará de Minas	x	x	x		x	
Paracatu	x	x	x	x	x	x
Paraisópolis	x	x	x	x		
Paraopeba	x	x	x			
Passa Quatro			x			
Passos	x	x	x		x	x
Patos de Minas	x	x	x	x	x	x
Patrocínio	x	x	x	x	x	x
Pedras de Maria da Cruz			x			x
Pedro Leopoldo	x	x	x			x
Perdizes	x	x	x	x	x	x
Perdões	x	x	x			
Pintópolis			x			
Piranga	x	x	x	x	x	x
Pirapora	x	x	x			
Pitangui	x	x	x	x		x
Piumhi	x	x	x		x	x
Poços de Caldas	x	x	x	x	x	
Pompéu	x	x	x		x	
Ponte Nova	x	x	x		x	x
Ponto dos Volantes			x			x
Porteirinha	x	x	x	x		x
Poté	x	x	x	x	x	x
Pouso Alegre	x	x	x	x	x	x
Prata	x			x		
Raul Soares	x	x	x			
Resplendor		x			x	
Riachinho	x	x	x	x	x	
Ribeirão das Neves	x	x	x			
Rio Pardo de Minas	x	x	x	x	x	x
Rio Pomba	x	x	x	x		
Rio Vermelho	x	x	x	x	x	x
Rubelita	x	x	x	x	x	x

Município	2010	2015	2020	2010 (sem zero)	2015 (sem zero)	2020 (sem zero)
Sabará	x	x	x	x	x	
Sabinópolis	x	x	x	x	x	
Sacramento	x	x	x	x		
Salinas	x	x	x	x		x
Santa Bárbara	x	x	x	x		
Santa Luzia	x	x	x	x	x	x
Santa Maria do Suaçuí	x	x	x	x	x	x
Santa Rita do Sapucaí	x	x	x	x	x	
Santa Vitória	x	x				
Santana do Paraíso	x	x	x	x	x	x
Santo Antônio do Amparo	x	x	x			
Santo Antônio do Jacinto	x	x	x	x		x
Santo Antônio do Monte	x	x	x		x	
Santos Dumont	x	x	x	x		
São Domingos do Prata	x					
São Francisco	x	x	x	x	x	x
São Gonçalo do Sapucaí	x	x	x	x		
São Gotardo	x	x	x			
São João da Ponte	x	x	x	x	x	x
São João das Missões	x	x	x	x	x	x
São João del Rei	x	x	x	x	x	x
São João do Paraíso	x	x	x	x	x	x
São João Evangelista	x	x	x	x	x	x
São João Nepomuceno	x	x	x	x		x
São Joaquim de Bicas	x	x	x			
São Lourenço	x	x	x	x	x	x
São Roque de Minas	x	x	x			
São Sebastião do Maranhão	x	x	x	x	x	x
São Sebastião do Paraíso	x	x	x	x		
Sarzedo			x			x
Serro	x	x	x			
Sete Lagoas	x	x	x		x	
Setubinha	x	x	x	x	x	x
Simonésia	x	x	x	x	x	x
Taiobeiras		x	x		x	x
Tarumirim	x	x	x	x	x	x
Teófilo Otoni	x	x	x	x	x	x
Timóteo	x	x	x	x		x
Três Corações	x	x	x	x		x
Três Marias	x	x	x		x	
Três Pontas	x	x	x	x		
Tumiritinga			x			
Tupaciguara	x	x	x	x	x	
Turmalina	x	x	x	x	x	x
Ubá	x	x	x	x	x	x
Ubaí			x			x
Ubaporanga		x	x		x	x
Uberaba	x	x	x	x	x	x
Uberlândia	x	x	x	x	x	
Unaí	x	x	x	x	x	x
Varginha	x	x	x	x		
Várzea da Palma	x	x	x		x	x
Varzelândia	x	x	x	x	x	x
Vazante	x	x	x			
Veredinha		x	x		x	x
Vespasiano	x	x	x	x		
Viçosa	x	x	x	x	x	x
Virgem da Lapa	x	x	x	x	x	x
Visconde do Rio Branco	x	x	x		x	x