Universidade do Minho
Escola de Ciências

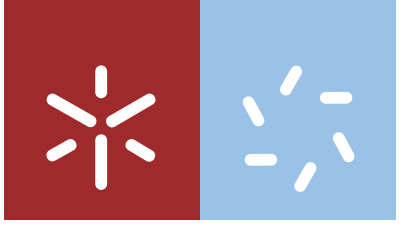
Antônio Neco de Oliveira

Modelação por regressão incorporando dependência espacial e temporal

julho de 2018

Antônio Neco de Oliveira **Modelação por regressão incorporando dependência espacial e temporal**

UMinho|2018



Universidade do Minho
Escola de Ciências

Antônio Neco de Oliveira

Modelação por regressão incorporando dependência espacial e temporal

Tese de Doutoramento em Ciências
Especialidade em Matemática

Trabalho efetuado sob a orientação da
Professora Doutora Raquel Menezes da Mota Leite
e da
Professora Doutora Susana Margarida Ferreira Sá Faria

julho de 2018

Declaração

Nome: Antônio Neco de Oliveira

Endereço electrónico: anecoo@gmail.com

Telefone: +55 (64) 99224-0102 / +55 (64) 3413-2994

Título da tese:

Modelação por regressão incorporando dependência espacial e temporal.

Orientadoras:

Professora Doutora Raquel Menezes da Mota Leite

Professora Doutora Susana Margarida Ferreira Sá Faria

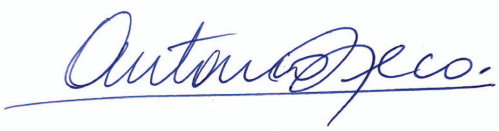
Ano de conclusão: 2018

Designação do Doutoramento:

Doutoramento em Ciências - Especialidade de Matemática

É AUTORIZADA A REPRODUÇÃO PARCIAL DESTA TESE APENAS PARA EFEITOS DE INVESTIGAÇÃO, MEDIANTE DECLARAÇÃO ESCRITA DO INTERESSADO, QUE A TAL SE COMPROMETE.

Universidade do Minho, 30 de julho de 2018.

Assinatura: 

Assinado de forma digital por ANTONIO NECO DE OLIVEIRA:33548315100
DN: c=BR, o=ICP-Brasil, ou=Secretaria da Receita Federal do Brasil - RFB, ou=RFB e-CPF A1, ou=(EM BRANCO), ou=Autenticado por AR Diggitare, cn=ANTONIO NECO DE OLIVEIRA:33548315100
Dados: 2018.07.30 14:29:15 +01'00'

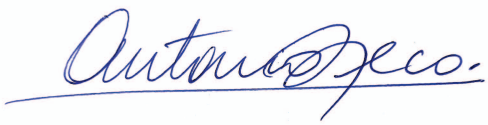
Declaração de Integridade

Declaro ter atuado com integridade na elaboração da presente tese. Confirmo que em todo o trabalho conducente à sua elaboração não recorri à prática de plágio ou a qualquer forma de falsificação de resultados.

Mais declaro que tomei conhecimento integral do Código de Conduta Ética da Universidade do Minho.

Universidade do Minho, 30 de julho de 2018.

Nome completo: Antônio Neco de Oliveira

Assinatura: 

Assinado de forma digital por ANTONIO NECO DE OLIVEIRA:33548315100
DN: c=BR, o=ICP-Brasil, ou=Secretaria da Receita Federal do Brasil - RFB, ou=RFB e-CPF A1, ou=(EM BRANCO), ou=Autenticado por AR Diggitare, cn=ANTONIO NECO DE OLIVEIRA:33548315100
Dados: 2018.07.30 14:28:23 +01'00'

Dedico este trabalho à minha família que sempre
me incentivou a enfrentar os desafios da vida.
Ao meu pai (in memoriam);
À minha mãe pela lucidez;
Aos meus irmãos pelo orgulho;
À minha esposa pela compreensão;
E aos meus filhos pelo amor.

Agradecimentos

Agradeço a Deus por ter me dado força para prosseguir nos momentos difíceis e a todas as pessoas que, direta ou indiretamente, contribuíram para a realização deste trabalho.

Agradecimento especial às orientadoras, Professora Doutora Raquel Menezes e Professora Doutora Susana Faria, pelas palavras de incentivo, pela paciência e pela sábia orientação.

Agradeço aos meus amigos, do IF Goiano e da UMINHO, que sempre acreditaram e me apoiaram nessa jornada; Em especial, aos Professores da equipe de Informática: Ana Maria, Felipe Gaia, Fernando Matos, Hiury Luiz, Jesmmer Alves, José Pereira, Luciano Carlos, Leila Rolling, Marcel Melo, Norton Coelho, Odilon Neto, Rodrigo Elias; e às amigas Débora Cristina e Patrícia Barea.

“... antes de tudo, um forte.”

(Euclides da Cunha)

Resumo

Modelação por regressão incorporando dependência espacial e temporal

A dengue é uma doença viral transmitida pelo mosquito *Aedes aegypti*, devendo ser tratada e o vetor transmissor eliminado para evitar epidemias e mortes. Nesta tese, a modelação por regressão incorporando dependência espacial e temporal é utilizada para estimar o número de notificações de casos de dengue no estado de Goiás, Brasil, onde o clima tropical favorece a proliferação do principal vetor transmissor dessa doença. Para a cidade de Goiânia, capital e principal cidade do Estado, aplica-se a metodologia das Equações de Estimação Generalizadas, e numa análise englobando 20 cidades desse Estado recorre-se aos Modelos Lineares Generalizados Mistos. Nos modelos mistos, o número de notificações de casos de dengue é estimado utilizando as variáveis meteorológicas e as estações do ano como efeitos fixos e os fatores *cidade* e *ano* como efeitos aleatórios de modo a caracterizar a dependência espacial e temporal. Para analisar o nível de significância das componentes de variância, associadas aos efeitos aleatórios, aplicam-se métodos *bootstrap* em dados simulados e em dados reais, obtendo-se probabilidades de cobertura superiores a 90 % quando o processo de reamostragem no nível do fator em estudo envolve cerca de 50 % dos dados originais. Os resultados indicam uma associação direta entre as variáveis meteorológicas e o número de notificações de casos de dengue, sendo o verão e o outono as estações do ano com maior número de casos registados. A precipitação, a temperatura mínima e a humidade relativa do ar são as variáveis que mais contribuem para o aumento no número de casos de dengue. O ano e a localização da cidade são os fatores determinantes na incidência de casos de dengue. A partir dos resultados obtidos, tem-se a indicação da necessidade de políticas públicas, juntamente com ações conjuntas da população local, para combater o vetor transmissor da dengue e evitar períodos epidémicos.

Palavras-chave: Binomial Negativa, *Bootstrap*, Clima, Dengue, Equações de Estimação Generalizadas, Modelos Lineares Generalizados Mistos, Poisson, Regressão.

Abstract

Regression modeling incorporating spatial and temporal dependence

Dengue is a viral disease transmitted by the mosquito Aedes aegypti, that requires treatment and the transmitter vector needs to be eliminated to avoid epidemics and deaths. In this thesis, regression modeling incorporating spatial and temporal dependence is used to model the number of reports of dengue cases in the state of Goiás, Brazil, where the tropical climate favors the proliferation of the main transmitting vector of this disease. For the city of Goiânia, capital and main city of the State, the methodology of Generalized Estimation Equations is applied, and in an analysis encompassing 20 cities of that State the Mixed Generalized Linear Models are used. In the later case, the number of notifications of dengue cases is estimated using the meteorological variables and years seasons, as fixed effects. The information on the city and the year is included in to the model as random effects, aiming to characterize spatial and temporal dependence. In order to analyze the significance level of the variance components, associated to random effects, bootstrap methods are applied to both simulated and real data, obtaining coverage probabilities above 90 %, when the resampling process at the factor's level analysis under involves about 50 % of the original data. The results indicate a direct association between the meteorological variables and the number of reports of dengue cases, being the summer and the fall the seasons of the year with greater number of registered cases. Precipitation, minimum temperature and relative air humidity are the variables that most contribute to the increase in the number of dengue cases. The year and the location of the city are the determining factors in the incidence of dengue cases. Based on the results obtained, the need for public policies, together with joint actions involving local population, are confirmed to be important to combat the vector transmitting dengue and avoid epidemic periods.
Keywords: Negative Binomial, Bootstrap, Climate, Dengue, Generalized Estimating Equations, Generalized Linear Mixed Models, Poisson, Regression.

Índice

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

AIC	*Akaike Information Criterion*
BR	Brasil
CV	Coeficiente de Variação
DF	Distrito Federal (Unidade Federativa onde localiza-se a capital do Brasil)
EQM	Erro Quadrático Médio
GEE	*Generalized Estimating Equations*
GLM	*Generalized Linear Models*
GLMM	*Generalized Linear Mixed Models*
GO	Goiás (Estado do Brasil localizado na região Centro-Oeste)
IBGE	Instituto Brasileiro de Geografia e Estatística
IC	Intervalo de Confiança
IMB	Instituto Mauro Borgers
INMET	Instituto Nacional de Meteorologia
LIRAa	Levantamento rápido de índice de infestação por *Aedes aegypi*
LM	*Linear Models*
LMM	*Linear Mixed Models*
MAE	*Mean Absolute Error*
MCMC	*Markov Chain Monte Carlo*

MMQ	Método de Mínimos Quadrados
MSE	*Mean Square Error*
OLS	*Ordinary Least Squares*
OMS	Organização Mundial de Saúde
QIC	*Quasi-Likelihood Under Independence Model Criterion*
RMSE	*Root Mean Square Error*
SADMET	Seção de Armazenamento de Dados Meteorológicos
SE	*Standard Error*
Sinan	Sistema de Informação de Agravos de Notificação
SUVISA-GO	Superintendência de Vigilância em Saúde do Estado de Goiás

Capítulo 1

Introdução

Nos últimos anos, a saúde pública enfrenta novos riscos devido à proliferação dos focos do mosquito *Aedes aegypti*, vetor transmissor de doenças que podem levar a óbito as pessoas infectadas pelos distintos vírus que lhe estão associados. O Brasil tem vivenciado constantes períodos de epidemias de dengue, o que torna necessário a implantação de políticas públicas para combater o vetor transmissor da doença. No estado de Goiás, a proliferação do mosquito *Aedes aegypti* dá-se, principalmente, pelas condições climáticas favoráveis da região central do Brasil que, para o período em estudo, apresenta temperaturas médias que variam entre 16.7 °C e 32.4 °C.

O agravamento da situação da saúde pública brasileira, causado pelos repetidos períodos de epidemias de dengue, requer um combate contínuo e sistemático ao mosquito transmissor do vírus da dengue. Nesse sentido, as autoridades políticas vêm atuando com medidas que necessitam da colaboração de toda a comunidade local para acabar com os focos do mosquito *Aedes aegypti* e eliminar os seus criadouros.

Dessa forma, recorrer a técnicas estatísticas para modelar e compreender este fenómeno, que permitam inclusivamente prever possíveis períodos epidémicos, torna-se relevante na tomada de decisões. Com base em valores preditos poderão ser tomadas medidas para evitar o desenvolvimento do mosquito *Aedes aegypti*, procurando controlar o vetor transmissor de doenças como a dengue, a febre *chikungunya*, a febre amarela e do zika vírus (possível causador da microcefalia).

1.1 Motivação

A dengue é uma doença infecciosa de grande impacto epidemiológico, social e económico, que constitui um problema crescente para a saúde pública mundial e em particular no continente americano. Esta doença manifesta-se por febre alta (39 °C, 40 °C) com duração de até sete dias, acompanhada de dor de cabeça, dor nos olhos, dor no corpo e articulações, erupção e coceira na pele. Numa fase mais grave, a doença pode incluir sintomas como dor abdominal intensa e contínua, vómitos e sangramento de mucosas. Ao sentir os primeiros sintomas, deve-se procurar o atendimento médico nos serviços de saúde, pois o agravamento da doença pode levar à morte (Brasil et al., 2009).

O vírus causador da febre dengue é transmitido pelo mosquito *Aedes aegypti*. Esse mosquito desenvolve-se em águas limpas e paradas, em criadouros formados após o início do período de chuva, ou em objetos residenciais que acumulam água, como vasos de plantas. O ciclo de desenvolvimento do *Aedes aegypti*, incluindo a transmissão da dengue e a manifestação da doença na pessoa infectada, dura aproximadamente 40 dias dependendo das variações climáticas. A partir de então, caso não haja ações conjuntas entre o poder público e a população, o número de casos de dengue poderá evoluir para índices epidémicos, sendo que o período mais crítico corresponde aos meses de janeiro a abril de cada ano.

No Brasil, as primeiras epidemias de dengue ocorreram em 1981-1982 em Boa Vista (Roraima), e em 1986 na cidade do Rio de Janeiro e em algumas capitais do Nordeste. Desde então, ocorrem epidemias associadas à circulação dos quatro sorotipos denominados DEN-1, DEN-2, DEN-3 e DEN-4, sendo todos causadores de febre clássica, com possíveis evoluções para dengue hemorrágica com padrões clínicos graves e fatais. Para acompanhar o padrão de transmissão da doença e possibilitar ações preventivas, todos os casos suspeitos de dengue devem ser notificados compulsoriamente à vigilância epidemiológica do município pela rede de saúde pública e privada (Brasil and Ministério da Saúde (MS), 2010).

Em 2014 foi identificado no Brasil a circulação do vírus causador da febre *chikungunya*, doença com sintomas semelhantes à febre dengue. Em 2015 foi identificado um novo vírus denominado “zika” que tem sido estudado por causar, dentre outras, a microcefalia em recém-nascidos. A transmissão de ambos os vírus é atribuída ao mosquito *Aedes aegypti* e o tratamento possibilita apenas o alívio dos sintomas.

Nesta tese estuda-se como as variações climáticas influenciam a incidência das notificações de casos de dengue nas cidades do estado de Goiás. Este Estado localiza-se na região Centro-Oeste do Brasil, com uma área de 340086 km^2 e uma população estimada, em 2015, de aproximadamente 7 milhões de habitantes. A temperatura média varia entre 18 °C e 26 °C e o verão húmido entre os meses de dezembro a março favorecem a formação de criadouros e a proliferação do mosquito *Aedes aegypti*, sendo que a primeira epidemia de dengue foi registada em 1994, com repetições nos anos de 2008, 2010, 2013 e 2015.

As relações entre as variáveis climatológicas e a incidência de casos de dengue foram analisadas por diversos autores, em diferentes locais e utilizando abordagens distintas. A Ilha de Barbados (Depradine and Lovell, 2004), cidade de Porto Rico (Jury, 2008), Guangzhou (na China) (Lu et al., 2009), Taiwan (Chen et al., 2010), (Yu et al., 2011), são exemplos de locais referenciados em estudos, nos quais se salienta que os casos de dengue estão fortemente relacionados com a sazonalidade inerente à variáveis meteorológicas, com poucos casos na estação seca e valores elevados na estação das chuvas.

O risco de dengue no Brasil, no período de 2001 a 2008, foi analisado recorrendo a modelos lineares generalizados mistos (*Generalized Linear Mixed Models* - GLMM), assumindo que a variável resposta seguia as distribuições de Poisson e Binomial Negativa, debaixo de uma abordagem *Bayesiana* apoiada em métodos de Monte Carlo via Cadeias de Markov (*Markov Chain Monte Carlo* - MCMC) (Lowe et al., 2011). Estes autores concluíram que a incidência de dengue está associada à região do país, às suas variações climáticas e às condições socioeconómicas locais.

Um estudo utilizando análise de componentes principais e modelos de regressão de Poisson (Pinto et al., 2011) identificou que os desfasamentos mais significativos, das variáveis meteorológicas, ocorrem até a décima sexta semana. Tal significa, por exemplo, que um pico de temperatura máxima poderá explicar um pico do número de casos de dengue até 16 semanas mais tarde. A análise mostrou uma correlação positiva entre a incidência de casos de dengue e as temperaturas máxima e mínima, seguida da precipitação, que são as variáveis com maior influência nos estágios do ciclo de vida do vetor transmissor da dengue.

Em Viana e Ignotti (2013), os autores relataram trabalhos que comprovaram o favorecimento da dinâmica sazonal do vetor da dengue pelas variações climáticas, as quais

proporcionam maior número de criadouros para o desenvolvimento do vetor transmissor. As infestações ocorreram principalmente entre os meses de maiores índices de precipitação, os quais favorecem o desenvolvimento do mosquito *Aedes aegypti*.

1.2 Objetivos

Nesta tese pretende-se:

- Aplicar a metodologia das equações de estimação generalizadas (*Generalized Estimating Equations* - GEE) para estudar as influências das variações climáticas no número de notificações de casos de dengue na cidade de Goiânia, considerando a correlação temporal inerente aos dados.

- Recorrer a modelos lineares generalizados mistos para se estudar o número de notificações de casos de dengue no estado de Goiás, incorporando efeitos fixos associados às variáveis meteorológicas e efeitos aleatórios para explicar as influências espaço-temporais caracterizadas pelos fatores *cidade*, *ano* e *semana*.

- Analisar, com base em estudos de simulação, a significância das estimativas das componentes de variância associadas aos efeitos aleatórios, recorrendo à metodologia *bootstrap* com reamostragem em dimensão inferior à amostra inicial para permitir estimativas com menor carga computacional.

1.3 Estrutura da Tese

Esta tese está organizada da seguinte maneira:

O Capítulo 2 apresenta uma análise exploratória da base de dados, com as especificidades que caracterizam a proliferação da dengue no estado de Goiás, representadas pelas informações das cidades analisadas. São consideradas as notificações de casos de dengue registadas pela Secretaria de Estado da Saúde, tendo em conta as informações das variáveis meteorológicas disponibilizadas pelo Instituto Nacional de Meteorologia. Também se considera o total de habitantes de acordo com os valores indicados pelo censo demográfico realizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE) nas cidades em estudo.

O Capítulo 3 descreve a análise realizada recorrendo-se ao método das equações de estimação generalizadas (GEE), para modelos lineares generalizados, aplicado aos dados

de notificações de casos de dengue na cidade de Goiânia para o período de 2008 a 2015, tendo em conta a dependência temporal presente nos dados.

No Capítulo 4 analisa-se o número de notificações de casos de dengue no estado de Goiás utilizando a abordagem dos modelos lineares generalizados mistos, incorporando efeitos fixos e aleatórios para considerar as especificidades de cada cidade, o período em análise e as variáveis meteorológicas num estudo espacial e temporal.

O Capítulo 5 propõe uma abordagem para a análise da significância da variância associada aos efeitos aleatórios em modelos mistos, assumindo-se a distribuição de Poisson para a variável resposta, para obtenção de estatísticas para o erro padrão e os intervalos de confiança, recorrendo-se à metodologia *bootstrap*.

No Capítulo 6, apresentam-se as conclusões desta tese e as indicações das expectativas para trabalhos futuros.

1.4 *Softwares* estatísticos utilizados

Neste trabalho foram utilizados os *softwares* estatísticos R (R Core Team, 2016) e IBM SPSS *Statistics* (Released, 2015). R é uma linguagem de programação e um ambiente de desenvolvimento integrado para cálculos estatísticos e gráficos, similar à linguagem de programação S desenvolvida originalmente pela *Bell Labs*. Esta linguagem de programação é desenvolvida em código fonte aberto e distribuída gratuitamente sob a licença GNU GPL (*General Public Licence*). Oferece diversas técnicas de análise de dados, inferência estatística e permite a visualização gráfica dos resultados. O IBM SPSS *Statistics* é um *software* estatístico proprietário, usado amplamente na solução de problemas de negócios e de pesquisas, com um variado conjunto de recursos para a gestão de dados como a seleção e a execução de análise, teste de hipótese, relatórios e a partilha dos resultados. Os seguintes pacotes do *software* R foram utilizados:

base pacote básico disponível na instalação do R;

DHARMa funções para análise de diagnóstico em modelos de regressão mistos;

foreign funções para manipulação de dados de ficheiros SPSS em R;

gee funções para estudar modelos lineares generalizados utilizando a abordagem das equações de estimação generalizadas;

ggplot2	funções para visualização gráfica;
glmmADMB	funções para análise de modelos lineares generalizados mistos;
gplots	funções para visualização gráfica;
graphics	funções para visualização gráfica;
lme4	funções para ajustar modelos lineares generalizados mistos;
lmerTest	funções para testes em modelos lineares generalizados mistos;
maptools	funções para manipulação de objetos geo-referenciados;
simr	funções para análise de modelos lineares generalizados mistos por simulação;
xtable	funções para exportar tabelas do R para o Latex.

Capítulo 2

Base de dados

2.1 Introdução

A base de dados foi construída a partir das informações de notificações de casos de dengue fornecidas pela Superintendência de Vigilância em Saúde do Estado de Goiás (SUVISA-GO), em folhas de cálculo excel com informações diárias registadas pelo Sistema de Informação de Agravos de Notificação (Sinan), as quais foram contabilizadas para quantificar o número de notificações de casos de dengue semanais em 20 cidades do estado de Goiás.

As informações meteorológicas foram fornecidas pelo Instituto Nacional de Meteorologia (INMET), Seção de Armazenamento de Dados Meteorológicos (SADMET), Brasília - DF, em folhas de cálculo excel com registos diários. As informações sobre a população foram obtidas junto ao Instituto Mauro Borges (IMB), site http://www.imb.go.gov.br/bde/, base de dados estatísticos do estado de Goiás, sendo a população em 2010 definida pelo censo demográfico e, para os restantes anos, por estimativa.

2.2 Organização da base de dados

A base de dados é formada pelas informações semanais das notificações de casos de dengue, no período de janeiro de 2008 a março de 2015, para 20 cidades do estado de Goiás, apresentadas na Tabela 2.1 com o respetivo número total de habitantes estimado em 2015.

Na Figura 2.1 apresenta-se o estado de Goiás localizado na região central do Brasil, com a identificação das 20 cidades em estudo. Para cada cidade, existe um conjunto de

Tabela 2.1: Cidades do estado de Goiás com o respetivo número total de habitantes estimado em 2015.

Cidade	Habitantes	Cidade	Habitantes
01 Alto Paraíso	7391	11 Jataí	95998
02 Aragarças	19583	12 Luziânia	194039
03 Caiapônia	18148	13 Monte Alegre	8319
04 Cristalina	53300	14 Morrinhos	44607
05 Goianésia	65767	15 Niquelândia	45243
06 Goiânia	1430697	16 Pirenópolis	24444
07 Goiás	24439	17 Pires do Rio	30703
08 Ipameri	26373	18 Posse	34663
09 Itapaci	20945	19 Rio Verde	207296
10 Itumbiara	100548	20 São Simão	19110

Fonte: Instituto Mauro Borges (http://www.imb.go.gov.br/bde/)

dados com 377 registos (correspondente às 377 semanas) representando as notificações de casos de dengue, as variações meteorológicas e o número total de habitantes anuais, totalizando 7540 registos. A cidade de Goiânia, capital do estado de Goiás, apresenta um número total de habitantes superior às restantes cidades, bem como o número total de notificações de casos de dengue.

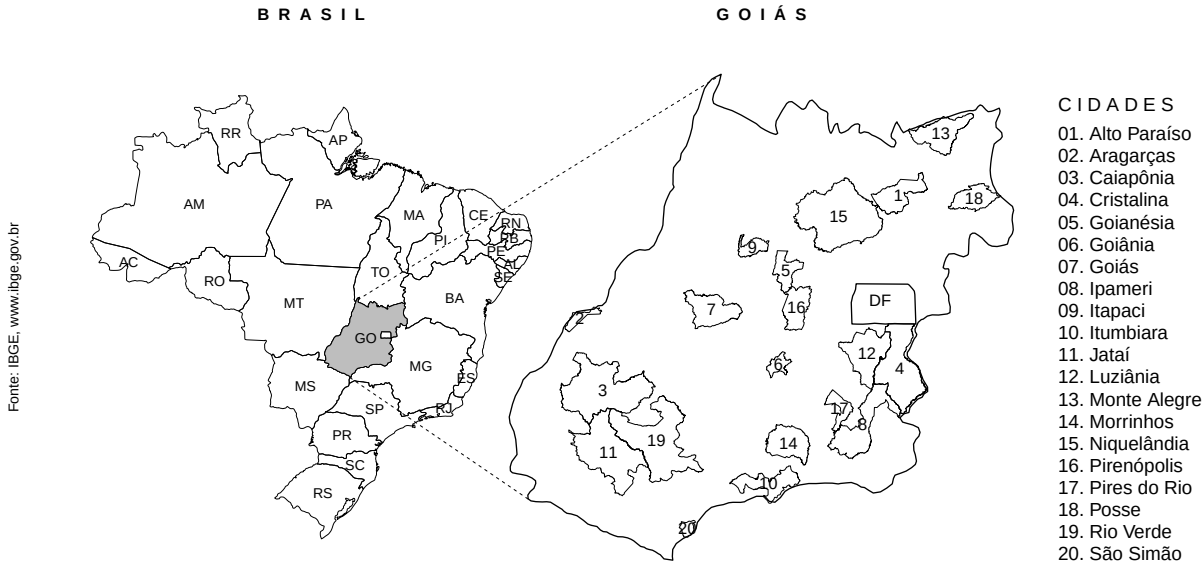


Figura 2.1: Mapas do Brasil e do estado de Goiás com as localizações das 20 cidades em estudo.

As notificações de casos de dengue contabilizadas semanalmente serão modeladas utilizando, como variáveis preditoras, a precipitação (*prec*), a temperatura mínima (*tmin*),

a temperatura máxima (*tmax*), a média da humidade relativa do ar (*hram*) e a velocidade do vento (*vvto*). Em resumo, na Tabela 2.2 apresenta-se uma descrição das variáveis da base de dados.

Tabela 2.2: Descrição das variáveis da base de dados.

Variável	**Definição**
cid	identifica a cidade do estado de Goiás.
ano	identifica o ano em estudo (de 2008 a 2015).
sem	identifica a semana de cada ano (de 1 a 53).
dengue	número de notificações de casos de dengue contabilizados semanalmente.
prec (mm)	precipitação acumulada em cada semana, obtida a partir da quantidade total diária, dada em milímetro.
tmin (°C)	temperatura mínima registada em cada semana, obtida a partir das temperaturas mínimas diárias, dada em graus Celsius.
tmax (°C)	temperatura máxima registada em cada semana, obtida a partir das temperaturas máximas diárias, dada em graus Celsius.
hram (%)	média da humidade relativa do ar em cada semana, obtida a partir das médias diárias da humidade relativa do ar, dada em percentagem.
vvto (m/s)	velocidade média do vento em cada semana, obtida a partir das médias diárias da velocidade do vento, dada em metros por segundo.
thab	número total de habitantes anual em cada cidade, obtida por censo no ano de 2010 e por estimativa para os restantes anos.

O logaritmo natural do número total de habitantes de cada cidade (*ln(thab)*), registado anualmente, será utilizado como variável *offset* nos modelos ajustados, importante para se considerar a frequência relativa do número de notificações de casos de dengue em cada cidade.

2.3 Análise descritiva dos dados

O comportamento das variáveis, para cada cidade, está representado nas Figuras 2.2 a 2.7. Na Figura 2.2 observa-se o número médio de notificações de casos de dengue por 100 mil habitantes, contabilizados semanalmente, considerando o período de janeiro de 2008 a março de 2015. Nota-se que as cidades de Goiânia (6) e Jataí (11) apresentam uma maior incidência da doença, com média semanal superior a 40 casos de dengue por 100 mil habitantes. Segundo a Organização Mundial de Saúde (OMS), mais de 300 casos por 100 mil habitantes/ano, já é considerado epidemia.

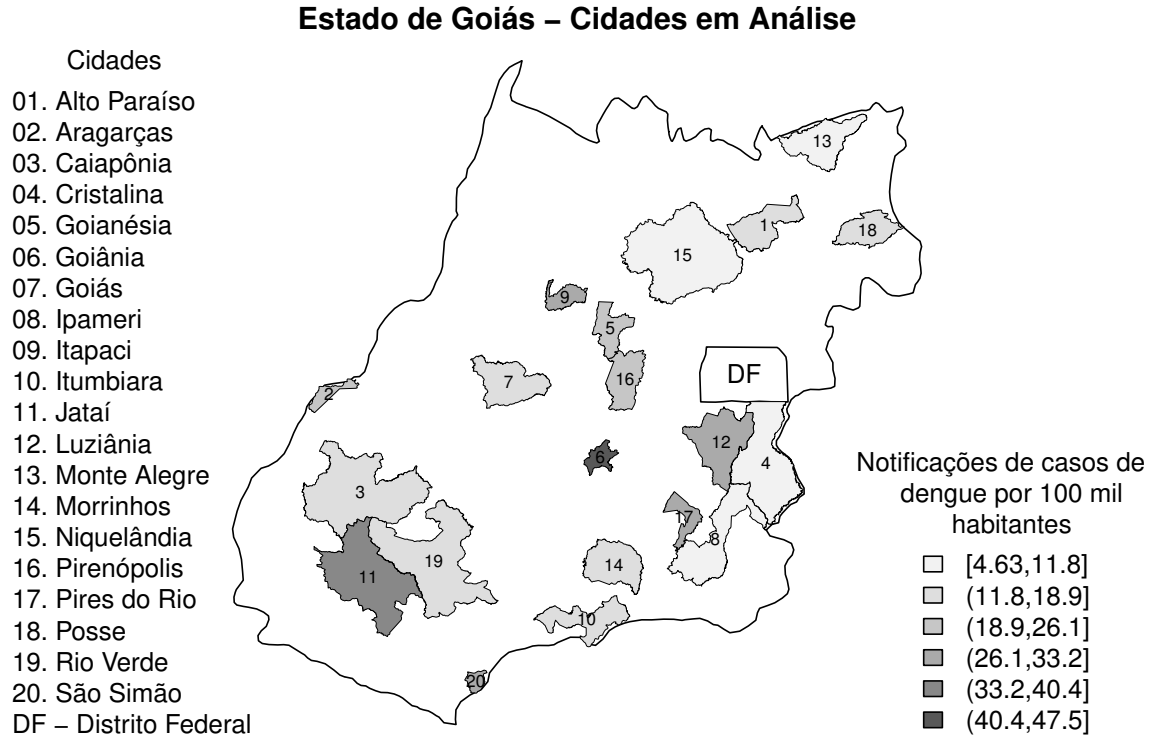


Figura 2.2: Média semanal do número de notificações de casos de dengue por 100 mil habitantes, considerando o período de janeiro de 2008 a março de 2015.

Na Figura 2.3, a precipitação acumulada para o período em estudo não parece apresentar um padrão espacial. Observa-se que a cidade de Pirenópolis (16) tem uma maior quantidade de precipitação média semanal (34.5 mm), correspondendo a 1794 mm anuais, enquanto que a cidade de Niquelândia (15) registou a menor quantidade de precipitação média semanal (18.3 mm), correspondendo a 951 mm anuais. Considerando as 20 cidades em estudo, temos uma quantidade de precipitação média semanal de 25.06 mm, equivalente a 1303 mm anuais. A chuva predomina no verão, com mais de 70 % entre os meses de novembro a março.

Na Figura 2.4 apresentam-se os valores médios semanais da temperatura mínima no período de 2008 a 2015 nas 20 cidades em estudo. Nota-se que a temperatura mínima aumenta à medida que se desloca no sentido do Sul para o Norte. Esse fato é devido à proximidade com a linha do equador, região em que a terra recebe maior incidência dos raios solares. As menores temperaturas mínimas são registadas para as cidades de Alto Paraíso (1), Morrinhos (14) e Jataí (11), as quais variam entre 14.3 °C e 15.1 °C.

Os valores médios semanais da temperatura máxima estão apresentados na Figura 2.5. As cidades de Aragarças (2), Goiás (7) e Monte Alegre (13) registaram temperaturas

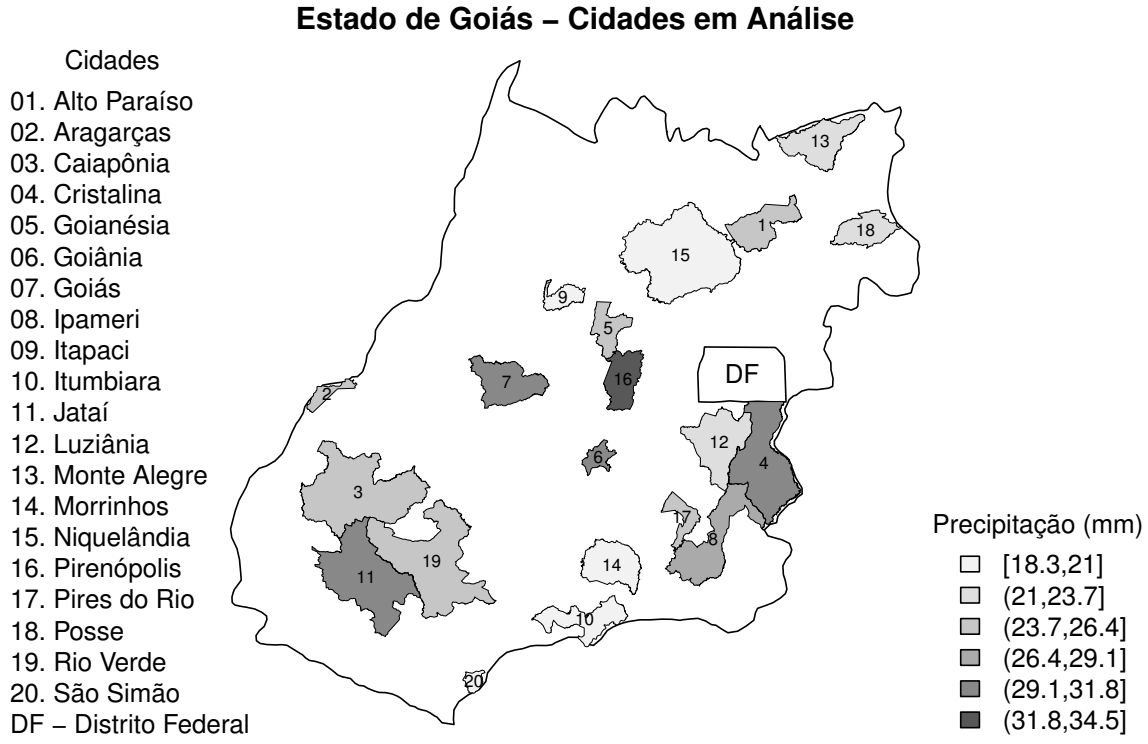


Figura 2.3: Média semanal da precipitação acumulada no período de janeiro de 2008 a março de 2015, dada em milímetros (mm).

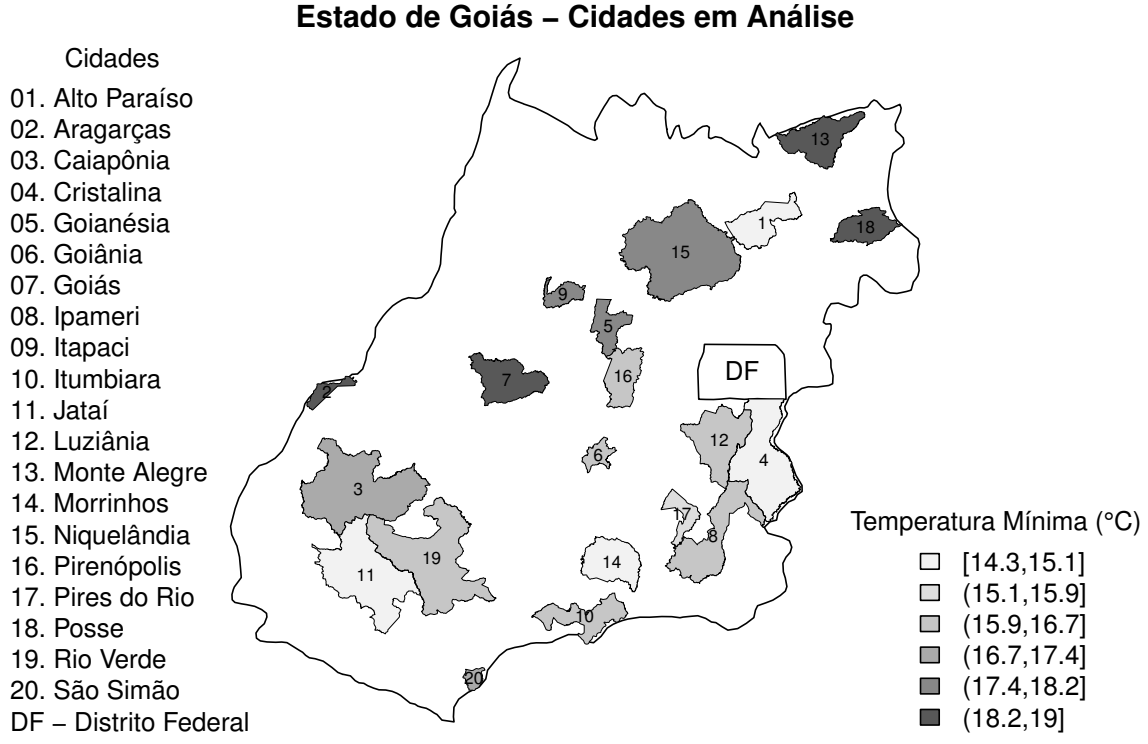


Figura 2.4: Média semanal da temperatura mínima no período de janeiro de 2008 a março de 2015, dada em graus Celsius (°C).

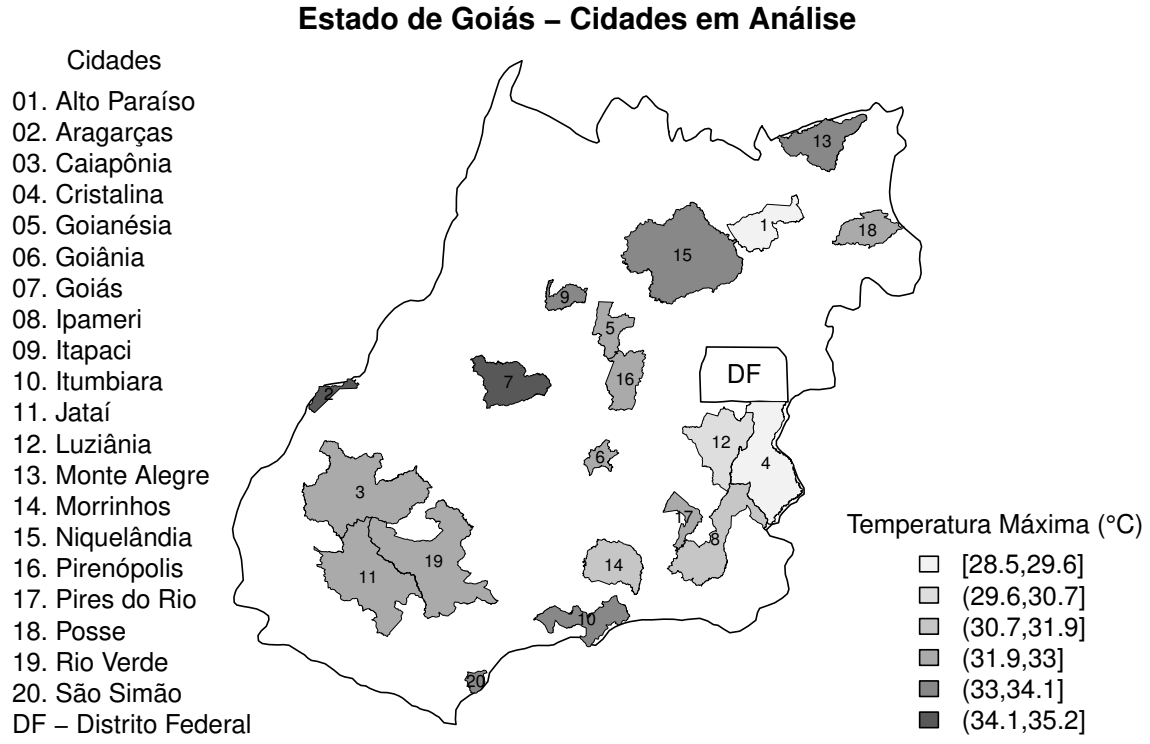


Figura 2.5: Média semanal da temperatura máxima no período de janeiro de 2008 a março de 2015, dada em graus Celsius (°C).

máximas superiores a 33 °C. Observa-se que a média semanal da temperatura máxima varia entre 28.5 °C e 35.2 °C.

Na Figura 2.6 nota-se que a média semanal da humidade relativa do ar diminui à medida que se desloca no sentido do Sul para o Norte, acompanhando o comportamento apresentado pela temperatura mínima, com uma variação média semanal entre 58.8 % na cidade de Posse (Norte do Estado) e 71.9 % na cidade de Jataí (Sul do estado de Goiás).

A Figura 2.7 apresenta os valores médios semanais da velocidade do vento, a qual não parece apresentar uma relação espacial. Os maiores valores médios foram registados nas cidades de Luziânia e Cristalina, com ventos de velocidades superiores a 2.38 m/s.

Na Tabela 2.3 são apresentadas algumas estatísticas descritivas para cada variável. Nota-se alta variabilidade da variável *dengue* (CV = 547.47 %) e da variável *prec* (CV = 144.97 %). As restantes variáveis apresentam um coeficiente de variação menor que 50 %.

Na Tabela 2.4 são apresentados os valores do coeficiente de variação de todas as variáveis em cada cidade, onde se pode confirmar a alta variabilidade da variável *dengue* e da variável *prec* em todas as cidades.

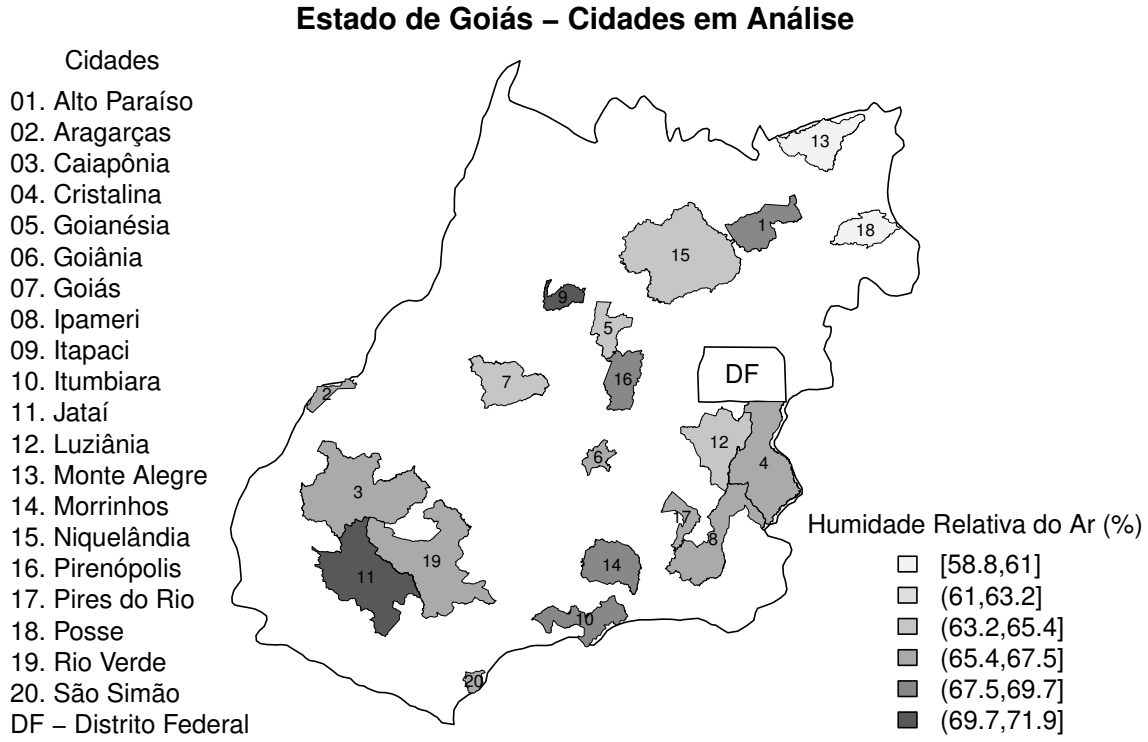


Figura 2.6: Média semanal da humidade relativa do ar no período de janeiro de 2008 a março de 2015, dada em percentagem (%).

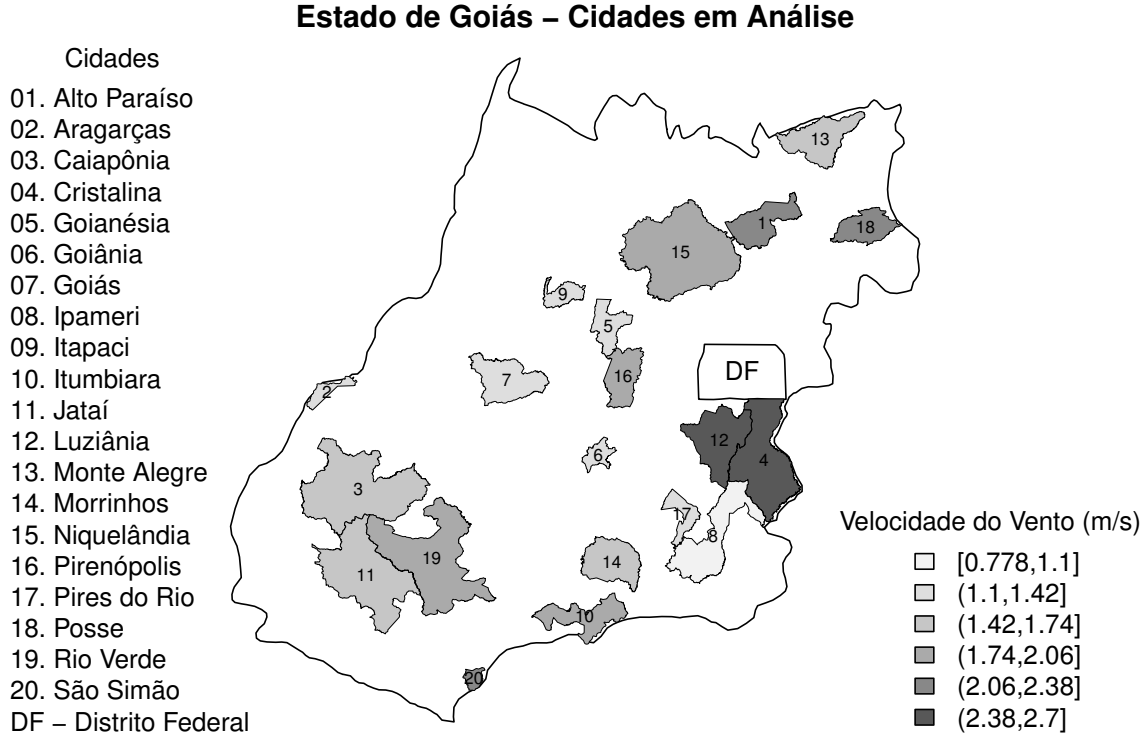


Figura 2.7: Média semanal da velocidade do vento no período de janeiro de 2008 a março de 2015, dada em metros por segundo (m/s).

Tabela 2.3: Estatísticas descritivas.

Variáveis	Mínimo	Média	Máximo	Desvio Padrão	Coef. Variação (%)
dengue (*)	0.00	45.38	5145	248.44	547.47
prec (mm)	0.00	25.06	251.80	36.33	144.97
tmin (°C)	1.60	16.70	26.30	3.25	19.46
tmax (°C)	23.30	32.35	41.70	2.63	8.13
hram (%)	18.12	66.15	91.04	14.82	22.40
vvto (m/s)	0.00	1.70	4.77	0.72	42.35

* Número de casos por semana.

Tabela 2.4: Coeficiente de variação estimado para cada variável nas diferentes cidades, no período de 2008 a 2015.

	COEFICIENTE DE VARIAÇÃO (CV %)					
CIDADE	**dengue**	**prec**	**tmin**	**tmax**	**hram**	**vvto**
01 - Alto Paraíso	329.73	145.86	16.06	6.95	22.22	24.88
02 - Aragarças	172.94	154.35	16.68	6.05	21.04	19.83
03 - Caiapônia	222.46	143.61	16.52	6.11	23.93	18.90
04 - Cristalina	240.96	137.35	15.63	7.09	20.96	21.18
05 - Goianésia	137.20	138.20	11.72	5.98	25.37	33.05
06 - Goiânia	129.30	123.67	20.07	5.93	19.58	27.13
07 - Goiás	202.82	141.85	10.09	6.25	26.63	30.16
08 - Ipameri	237.34	134.28	22.09	6.46	19.53	29.49
09 - Itapaci	172.32	159.37	18.76	5.61	19.96	34.71
10 - Itumbiara	210.25	161.89	23.29	6.48	17.03	30.65
11 - Jataí	177.59	119.40	27.84	6.13	17.33	37.25
12 - Luziânia	234.31	148.54	14.25	6.44	22.64	18.15
13 - Monte Alegre	336.07	161.84	11.06	6.46	29.26	55.48
14 - Morrinhos	198.56	166.05	24.87	6.64	17.62	25.00
15 - Niquelândia	210.27	177.09	9.35	6.02	24.78	24.60
16 - Pirenópolis	337.53	115.45	17.76	5.58	22.16	34.65
17 - Pires do Rio	328.61	139.85	22.18	6.12	18.97	26.50
18 - Posse	245.37	152.53	8.29	6.08	26.89	24.78
19 - Rio Verde	232.62	144.47	20.88	6.41	22.89	39.56
20 - São Simão	293.58	156.90	19.88	6.55	20.38	30.00

Na Figura 2.8, apresentam-se as *boxplots* de cada variável, por cidade, para o período em estudo. Nota-se um alto número de notificações de casos de dengue para a cidade de Goiânia (6), cidade de maior densidade populacional. As restantes variáveis apresentam similaridades em todas as cidades. No Anexo A são apresentadas as *boxplots* das variáveis em cada cidade, tendo em conta as estações do ano. Nota-se um comportamento semelhante das variáveis meteorológicas nas cidades em estudo, considerando o período de janeiro de 2008 a março de 2015.

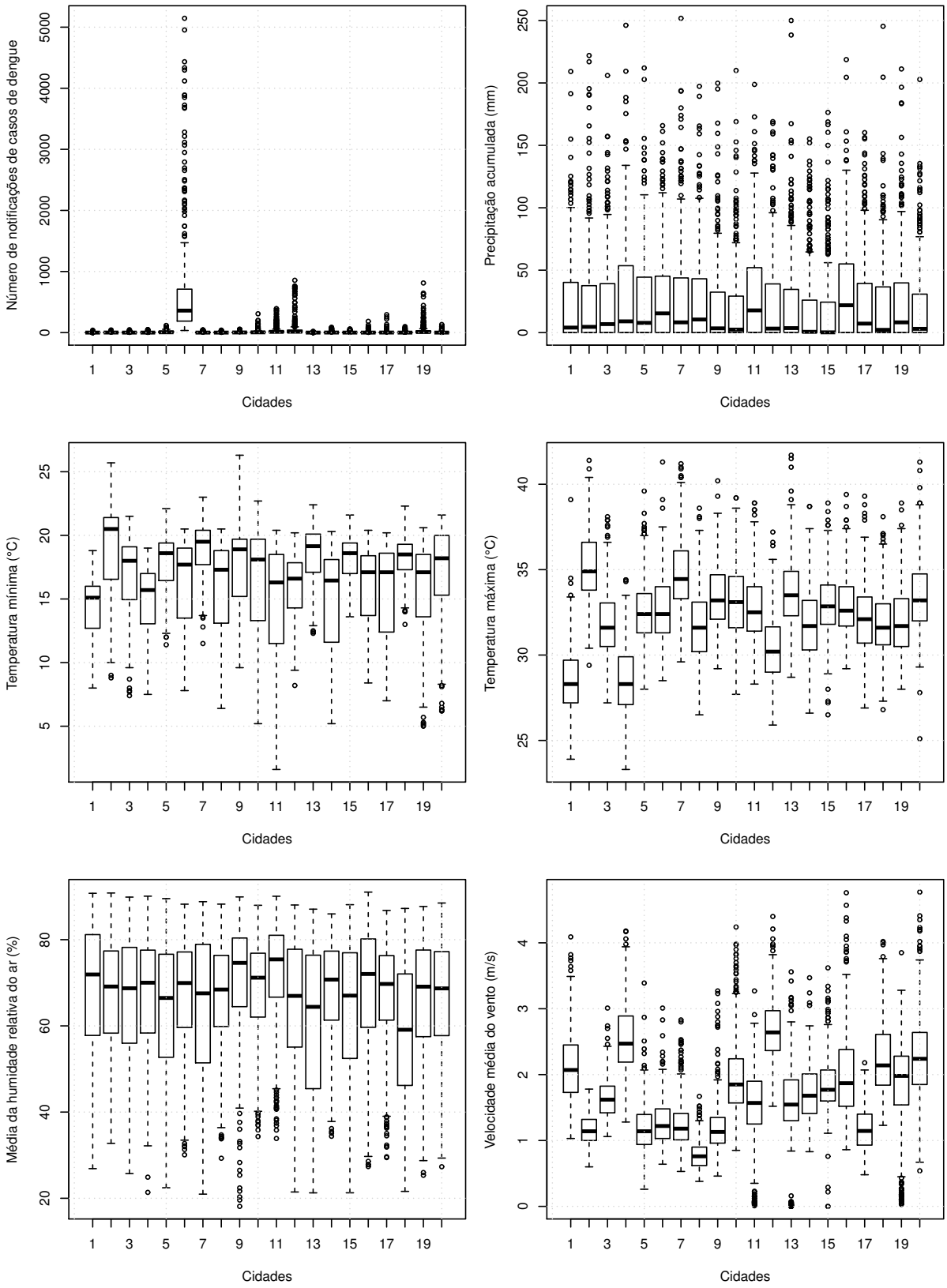


Figura 2.8: *Boxplots* das variáveis para o período de janeiro de 2008 a março de 2015, contabilizadas semanalmente. Os números de 1 a 20 indicam as cidades em estudo conforme apresentadas na Tabela 2.4.

2.4 Análise das correlações

Para caracterizar o período de desenvolvimento do vetor transmissor da dengue até ao momento que a pessoa infectada pelo vírus procura atendimento médico, as variáveis meteorológicas foram desfasadas de zero a dez semanas. A relação entre cada variável meteorológica e o número de notificações de casos de dengue foi analisada utilizando o coeficiente de correlação de *Spearman* (Chen et al., 2010), uma vez que os dados não seguem a distribuição Normal. Os resultados estão apresentados na Tabela 2.5, onde observam-se maiores correlações positivas com a precipitação (*lag* 10), a temperatura mínima (*lag* 10), a temperatura máxima (*lag* 10) e a humidade relativa do ar (*lag* 6). A maior correlação negativa é observada com a velocidade do vento (*lag* 0).

Tabela 2.5: Correlação de *Spearman* entre a variável *dengue* e as variáveis meteorológicas desfasadas no tempo entre zero a dez semanas, contabilizadas para as 20 cidades do estado de Goiás.

lag	*prec (mm)*	*tmin (°C)*	*tmax (°C)*	*hram (%)*	*vvto (m/s)*
0	0.0996***	0.0289***	-0.0808	0.2146***	**-0.1701***
1	0.1181***	0.0613***	-0.0656	0.2285***	-0.1661***
2	0.1412***	0.0872***	-0.0525	0.2434***	-0.1598***
3	0.1629***	0.1141***	-0.0368	0.2559***	-0.1561***
4	0.1846***	0.1363***	-0.0250	0.2696***	-0.1468***
5	0.2038***	0.1595***	-0.0084	0.2778***	-0.1454***
6	0.2264***	0.1774***	0.0002	**0.2885***	-0.1352***
7	0.2319***	0.1950***	0.0168.	0.2830***	-0.1281***
8	0.2474***	0.2006***	0.0191.	0.2859***	-0.1157***
9	0.2570***	0.2072***	0.0345*	0.2790***	-0.1064***
10	**0.2608***	**0.2137***	**0.0485***	0.2713***	-0.0942***

*** $p - value < 0.001$; * $p - value < 0.05$; . $p - value < 0.10$

Para facilitar a visualização, na Figura 2.9 são apresentados os coeficientes de correlação de *Spearman* entre o número de notificações de casos de *dengue* e as variáveis meteorológicas com os desfasamentos. Os pontos assinalados nas linhas indicam que a correlação é significativamente diferente de zero ao nível de significância de 5 %. Em cada linha há um ponto destacado com a letra **X** indicando o desfasamento no qual a variável preditora apresentou maior correlação com a variável resposta.

Na Figura 2.10 são apresentados os valores médios semanais para as variáveis em estudo. Observam-se que as maiores médias de notificações de casos de dengue ocorrem nas últimas semanas e se prolongam nas primeiras semanas até a vigésima semana.

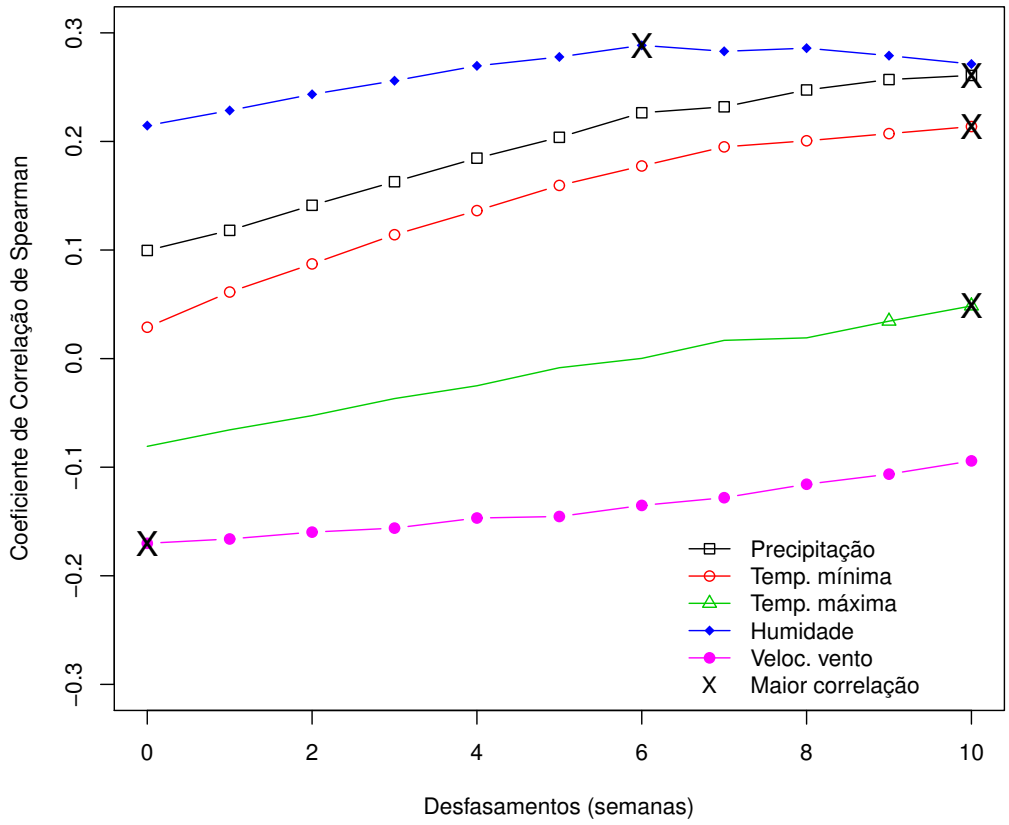


Figura 2.9: Correlações de Spearman entre a variável *dengue* e as variáveis preditoras desfasadas no tempo.

Esse comportamento é acompanhado diretamente nas variações da precipitação, da temperatura mínima e da humidade relativa do ar, e inversamente à temperatura máxima e à velocidade do vento. Nota-se uma maior incidência de notificações de casos de dengue quando a precipitação média está próxima dos 40 mm semanais, temperatura mínima entre 18 °C e 19 °C, temperatura máxima entre 31 °C e 33 °C, a humidade relativa do ar acima de 70 % e com baixa velocidade do vento.

Na Figura 2.11 são apresentadas as médias semanais para o período de janeiro de 2008 a março de 2015, com os respetivos intervalos de confiança para a variável resposta *dengue* e para as variáveis preditoras com os respetivos desfasamentos.

Percebe-se que o número médio de notificações de casos de dengue apresenta maiores valores para os anos de 2010, 2013 e 2015. Esses anos foram considerados epidémicos de acordo com os parâmetros de referência da OMS, por registarem mais de 300 casos de dengue por 100 mil habitantes/ano.

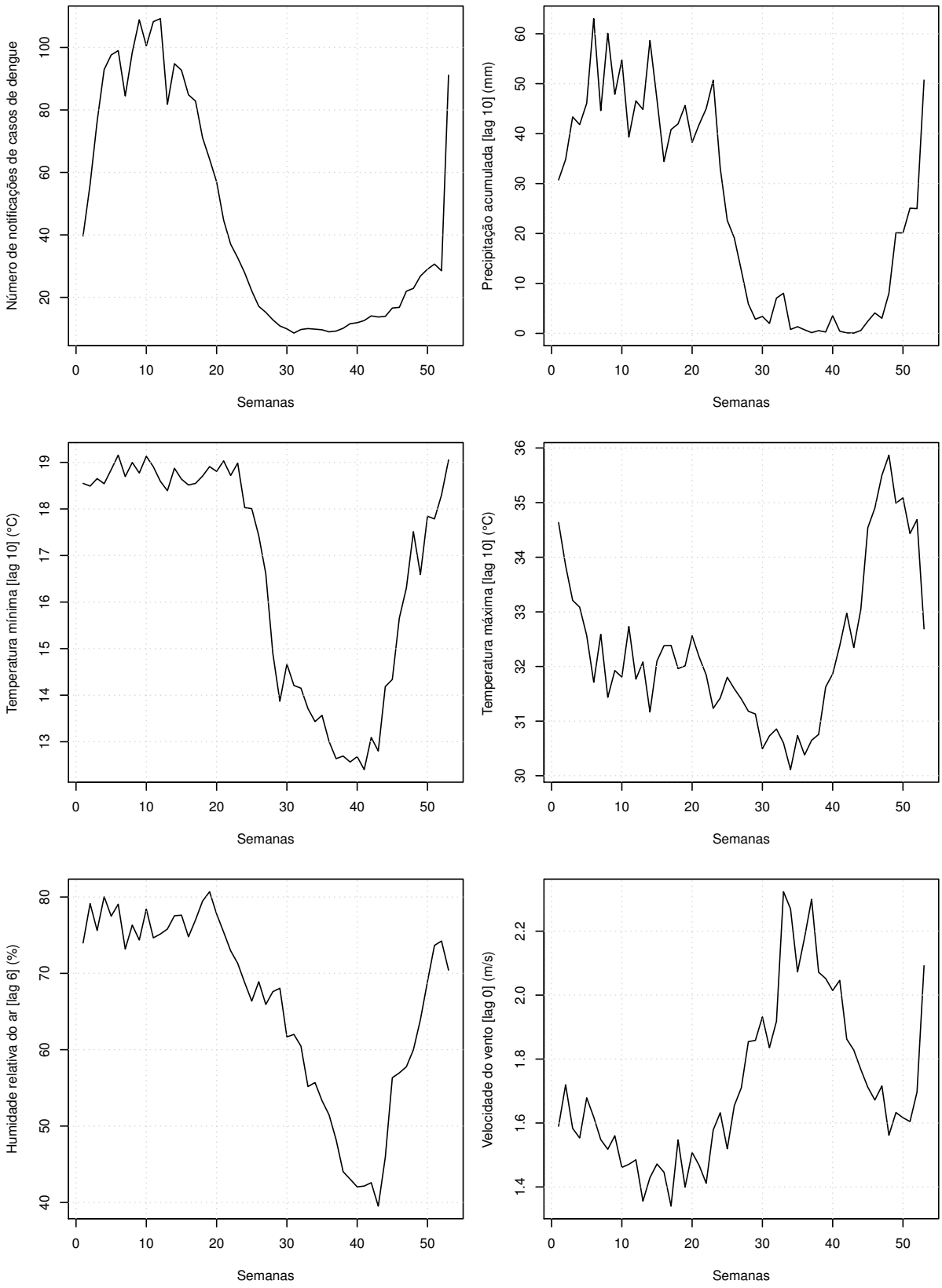


Figura 2.10: Médias semanais da variável dengue e das variáveis preditoras nos respetivos desfasamentos, para o conjunto de dados das 20 cidades e considerando o período de 2008 a 2015.

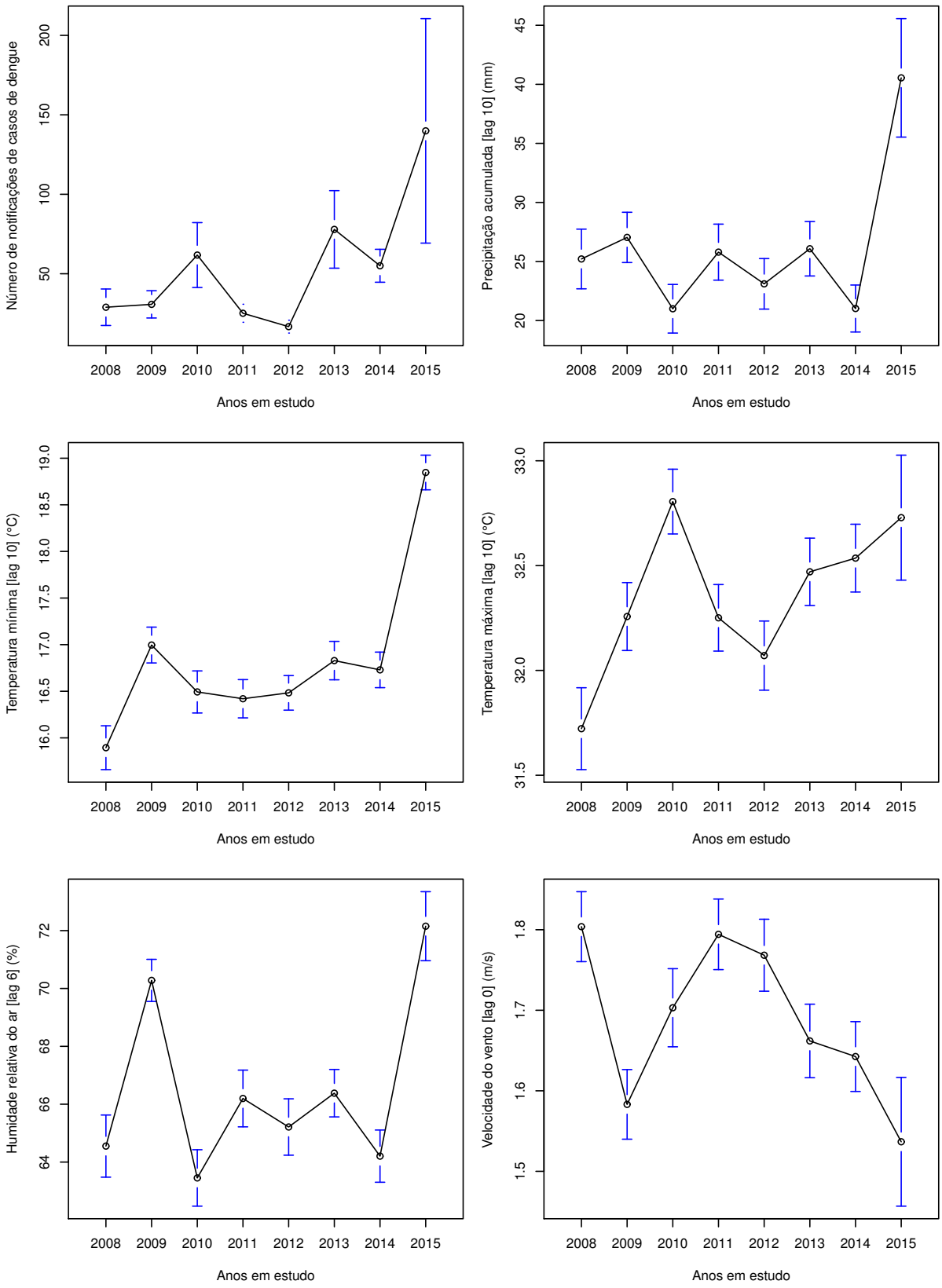


Figura 2.11: Médias semanais nos anos em estudo, com os respetivos intervalos de confiança, da variável dengue e das variáveis preditoras nos respetivos desfasamentos.

2.5 Considerações finais do capítulo

Neste capítulo apresentam-se os resultados da análise exploratória da base de dados, a qual é formada pelos registos semanais das notificações de casos de dengue em 20 cidades do estado de Goias, pelas informações meteorológicas e pelo total de habitantes anual para cada cidade.

A cidade de Goiânia, capital do estado de Goiás, sendo a cidade de maior densidade populacional, apresenta também o maior número de notificações de casos de dengue registado no período em estudo. Conclui-se que as notificações de casos de dengue se relacionam diretamente com o período chuvoso, intensificando-se nas primeiras semanas de cada ano.

Nota-se, ainda, que o período com maior incidência de casos de dengue é caracterizado por apresentar temperaturas entre 19 °C e 31 °C, precipitação média semanal de 40 mm e média da humidade relativa do ar abaixo de 70 %. Essas informações de referências serão utilizadas na modelação das notificações de casos de dengue que será realizada nos capítulos seguintes.

Capítulo 3

Modelação Temporal - Equações de Estimação Generalizadas

Este capítulo restringe-se à análise dos dados referentes à cidade de Goiânia, uma das vinte cidades apresentadas no Capítulo 2. Trata-se da cidade com o número mais elevado de notificações de casos de dengue, para a qual iremos estudar a influência das variáveis meteorológicas no número de notificações de casos de dengue. Esta cidade localiza-se na região Centro-Oeste do Brasil (Latitude: -15.91°, Longitude: -50.13°, Altitude: 512.22 m), tem uma área de 729.02 km^2 e uma população estimada em 1.5 milhão de habitantes para o ano de 2015. A temperatura média varia entre 18 ℃ e 26 ℃ e o verão húmido entre os meses de dezembro a março favorecem a formação de criadouros naturais e a proliferação do mosquito *Aedes aegypti*, sendo que a primeira epidemia de dengue foi registada em 1994, com repetições nos anos de 2008, 2010, 2013 e 2015.

Os dados para análise foram extraídos da base de dados descrita no Capítulo 2. As 377 observações representam as informações do número de notificações de casos de dengue (*dengue*), precipitação acumulada (*prec*), temperatura mínima (*tmin*), temperatura máxima (*tmax*), média da humidade relativa do ar (*hram*) e número total de habitantes (*thab*), contabilizadas semanalmente ao longo dos anos de 2008 a 2015.

Faz-se notar que estes valores semanais poderão ser considerados temporalmente correlacionados. Com o pressuposto de independência das observações violado, os modelos GLM poderão apresentar distorções nas estimativas dos parâmetros.

3.1 Introdução

Para estudar como as variáveis meteorológicas influenciam o número de notificações de casos de dengue, aplica-se a metodologia das equações de estimação generalizadas (*Generalized Estimating Equations* - GEE). Esta metodologia tem-se mostrado extremamente útil na análise de dados agrupados, caso de dados longitudinais, especialmente quando a resposta é de natureza discreta, como dados de contagens. Neste contexto, ajustam-se modelos de regressão com as distribuições de Poisson e Binomial Negativa para prever o número de notificações de casos de dengue em função das variáveis meteorológicas.

O método a ser utilizado deve estimar os parâmetros do modelo de forma que possa ser explicada a incidência semanal de notificações de casos de dengue ao longo dos anos em função das variáveis meteorológicas, dando o devido tratamento à correlação temporal.

A abordagem das GEE (Liang and Zeger, 1986) foi proposta em 1986 como uma extensão aos modelos lineares generalizados (*Generalized Linear Models* - GLM), para analisar dados longitudinais com dependência temporal e independentes entre sujeitos distintos, utilizando a abordagem da quasi-verosimilhança para estimar os parâmetros do modelo de regressão. Em 1988 foram definidas abordagens específicas (Zeger et al., 1988) para modelar a heterogeneidade e as variações da média populacional entre os indivíduos, incorporando possíveis efeitos fixos e aleatórios.

O termo *generalized estimating equations* indica que a equação de estimação não é obtida a partir da função de verosimilhança, sendo baseada em métodos de quasi-verosimilhança que evitam pressupostos paramétricos (Hardin and Hilbe, 2003). A modificação proposta para obter a equação de estimação generalizada tem em consideração componentes de variância de segunda ordem, sendo as dependências entre as observações definidas pela especificação da estrutura de correlação.

Neste sentido, ao assumir a matriz identidade como a matriz de correlação entre as observações, assume-se que as observações semanais repetidas para cada ano são independentes e a abordagem das GEE estima os parâmetros como o GLM. Assim, são adotadas estruturas distintas para a matriz de correlação, de modo a considerar a correlação temporal existentes nos dados. A utilização de uma matriz inadequada poderá afetar a estimação dos coeficientes e acarretar em resultados com menor precisão. A estrutura de correlação especifica a associação entre duas diferentes amostras no mesmo grupo (Liang and Zeger, 1986), ou seja, no nosso caso em estudo, especifica a associação entre duas

diferentes semanas no mesmo ano.

3.2 Metodologia

Para descrever as equações de estimação generalizadas (GEE), apresentaremos, brevemente, os conceitos atinentes aos modelos lineares (LM) e aos modelos lineares generalizados (GLM), que são precedentes às GEE. Os LM aplicam-se quando a variável resposta é procedente de uma distribuição Gaussiana, enquanto que os GLM são caracterizados pela variável resposta seguir uma distribuição pertencente à família exponencial, que inclui a Gaussiana, a Binomial, a Poisson, a Gama, a Gaussiana Inversa, a Geométrica e a Binomial Negativa. As GEE são extensões aos modelos lineares generalizados, que podem ser adotados para a análise de dados discretos com dependência temporal.

3.2.1 Modelos lineares - LM

Considerando a variável resposta Y_i e um vetor de variáveis preditoras $\boldsymbol{x_i} = (x_{i1}, x_{i2}, \ldots, x_{ip})$, o modelo linear (LM) é definido por

$$Y_i = \beta_0 + \beta_1 \times x_{i1} + \beta_2 \times x_{i2} + \cdots + \beta_p \times x_{ip} + \varepsilon_i, \qquad i = 1, \ldots, N$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

onde $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$ é o vetor de parâmetros de regressão e ε_i é o erro aleatório associado à *i-ésima* observação, cuja distribuição de probabilidade se supõe ser Normal com média zero e variância constante σ^2. Tem-se que $E[Y_i|x_i] = \mu = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$, ou seja, o valor esperado da variável resposta é uma função linear das covariáveis. Os pressupostos de normalidade, homogeneidade e independência dos erros devem ser verificados. A violação desses pressupostos pode originar estimativas erradas dos parâmetros e o modelo linear ajustado deve ser rejeitado (Zuur et al., 2009).

Um dos métodos de estimação dos parâmetros de um modelo de regressão linear é o método de mínimos quadrados (*Ordinary Least Squares* - OLS), que consiste em minimizar a soma dos quadrados dos resíduos, ou seja, que minimiza

$$SS(\beta) = \sum_{i=1}^{n} \varepsilon_i^2.$$

Outro método de estimação é o método da máxima verosimilhança, que consiste em maximizar a função de verosimilhança

$$L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}) = (2\pi\sigma^2)^{-n/2} exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \boldsymbol{x_i}\boldsymbol{\beta})^2\right]$$

onde $\boldsymbol{y}$ é uma amostra aleatória, de dimensão n.

3.2.2 Modelos lineares generalizados - GLM

Os modelos lineares generalizados (GLM) são uma extensão dos modelos de regressão linear clássicos (Chatterjee and Hadi, 2015). A extensão é feita em duas direções:

- a distribuição da variável resposta pode ser qualquer distribuição da família exponencial;
- a função que relaciona o valor esperado e o vetor de covariáveis pode ser qualquer função diferenciável.

Os GLM são caracterizados pela seguinte estrutura:

1. Componente aleatória

   Dado o vetor de covariáveis $\boldsymbol{X_i}$, as variáveis Y_i são independentes com distribuição pertencente à família exponencial, com $E[Y_i|\boldsymbol{X_i}] = \mu_i$ para $i = 1, \ldots, n$ e um parâmetro de dispersão ϕ não dependente de i.

2. Componente sistemática

   O valor esperado μ_i está relacionado com o preditor linear $\eta_i = \boldsymbol{Z}_i^T\boldsymbol{\beta}$ através da relação

   $$\mu_i = h(\eta_i) = h(\boldsymbol{Z}_i^T\boldsymbol{\beta}), \qquad \eta_i = g(\mu_i),$$

   onde

   - h é uma função monótona e diferenciável;
   - $g = h^{-1}$ é a função de ligação;
   - $\boldsymbol{\beta}$ é um vetor de parâmetros de dimensão p;

- $\boldsymbol{Z_i}$ é um vetor de especificação de dimensão p, função do vetor de covariáveis $\boldsymbol{x_i}$. Em geral, $\boldsymbol{Z}_i = (1, x_{i1}, x_{i2}, \ldots, x_{ip})^T$.

Na estimação dos parâmetros $\boldsymbol{\beta}$ do modelo pode-se aplicar o método da máxima verosimilhança.

3.2.3 Equações de estimação generalizadas - GEE

Um modelo de regressão GLM utilizando a abordagem GEE é especificado pela definição da componente sistemática, da função de ligação, da variância e da estrutura de correlação existente entre as observações dentro de um mesmo grupo.

Componente sistemática

Considerando a variável resposta Y_{it} e um vetor de variáveis explicativas $\boldsymbol{x}_{it}$, define-se o número de notificações de casos de dengue Y_{it}, no ano i, na semana t, com a parte sistemática dada por

$$\eta_{it} = \boldsymbol{\beta} \times \boldsymbol{x}_{it},$$

onde $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$.

Para a distribuição de Poisson, e para a distribuição Binomial Negativa, o logaritmo natural (ln) é uma função de ligação que define o relacionamento entre a componente sistemática e a média condicional, especificada por

$$E[Y_{it}|X_{it}] = \mu_{it}$$

$$g(\mu_{it}) = ln(\mu_{it}) = \boldsymbol{\beta} \times \boldsymbol{x_{it}}$$

$$E[Y_{it}|x_{it}] = e^{\boldsymbol{\beta} \times \boldsymbol{x_{it}}} = e^{\eta_{it}}.$$

Variância

Para os dados de notificações de casos de dengue, caracterizados como dados de contagens, tem-se que a variância da distribuição de Poisson é dada por:

$$var[Y_{it}|X_{it}] = \mu_{it},$$

e, da distribuição Binomial Negativa, tem-se:

$$var[Y_{it}|X_{it}] = \mu_{it} + \alpha\mu_{it}^2,$$

onde α representa o parâmetro escalar que indica a sobredispersão existentes nos dados.

Estruturas de correlação

A relação entre duas diferentes observações Y_{it} e $Y_{it'}$, tomadas no mesmo grupo i e nos tempos t e t', é especificada pela estrutura de correlação. Seja o ano i e as semanas t e t', destacam-se as seguintes estruturas de correlação:

- **Estrutura independente:** utiliza a matriz identidade (I) como matriz de correlação de trabalho, assumindo que as observações não apresentam correlações entre si. Essa estrutura é dada por:

$$Corr(y_{it}, y_{it'}) = R(\alpha) = \begin{cases} 1 & se \quad t = t' \\ 0 & \forall \quad t \neq t', \end{cases}$$

$$R(\alpha) = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix},$$

  onde y_{it} e $y_{it'}$ representam observações do grupo (ano) i nos tempos (semanas) t e t', respetivamente.

- **Estrutura estacionária dependente de ordem m, M(m):** assume que a correlação entre quaisquer duas observações espaçadas $|t-t'|$ é a mesma e que os dados estão correlacionados até um desfasamento máximo igual a m, tornando-se necessário especificar o valor de m, a partir de quando se pode assumir a independência. A estrutura da matriz de correlação, considerando $m = 2$, é definida como:

$$Corr(y_{it}, y_{it'}) = R(\alpha) = \begin{cases} 1 & se \quad t = t' \\ \alpha_{|t-t'|} & se \quad |t-t'| = 1, ..., m \\ 0 & se \quad |t-t'| > m \end{cases}$$

$$R(\alpha) = \begin{bmatrix} 1 & \alpha_1 & \alpha_2 & \dots & 0 \\ \alpha_1 & 1 & \alpha_1 & \dots & 0 \\ \alpha_2 & \alpha_1 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}$$

- **Estrutura permutável:** utilizada para agrupamento sem ordem lógica, que não se adequa ao nosso caso de estudo, onde a correlação entre as observações dos indivíduos num mesmo grupo é a mesma.

$$Corr(y_{it}, y_{it'}) = \begin{cases} 1 & se \quad t = t' \\ \alpha & se \quad t \neq t' \end{cases}$$

$$R(\alpha) = \begin{bmatrix} 1 & \alpha & \alpha & \dots & \alpha \\ \alpha & 1 & \alpha & \dots & \alpha \\ \alpha & \alpha & 1 & \dots & \alpha \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha & \alpha & \alpha & \dots & 1 \end{bmatrix}$$

- **Estrutura autorregressiva de primeira ordem AR(1):** adequada para dados agrupados ao longo do tempo. Assume que a correlação entre duas observações diminui exponencialmente à medida que aumenta o intervalo de tempo entre elas. Esta estrutura exige que as medições sejam obtidas em intervalos igualmente espaçados (ou aproximadamente iguais) no tempo.

$$Corr(y_{it}, y_{it'}) = \begin{cases} 1 & se \quad t = t' \\ \alpha^{|t-t'|} & \forall \quad t \neq t' \end{cases}$$

$$R(\alpha) = \begin{bmatrix} 1 & \alpha & \alpha^2 & \dots & \alpha^n \\ \alpha & 1 & \alpha & \dots & \alpha^{n-1} \\ \alpha^2 & \alpha & 1 & \dots & \alpha^{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha^n & \alpha^{n-1} & \alpha^{n-2} & \dots & 1 \end{bmatrix}$$

- **Não estruturada:** considera que a correlação entre duas observações de um grupo é sempre diferente, sendo recomendada quando o número de observações é pequeno, os dados são balanceados, ou seja, os grupos possuem o mesmo número de observações. A principal vantagem na utilização desta estrutura é a ausência de pressupostos para os parâmetros de variância e covariância.

$$Corr(y_{it}, y_{it'}) = \begin{cases} 1 & se \quad t = t' \\ \alpha_{t'} & se \quad t \neq t' \end{cases}$$

$$R(\alpha) = \begin{bmatrix} 1 & \alpha_1 & \alpha_2 & \dots & \alpha_n \\ \alpha_1 & 1 & \alpha_1 & \dots & \alpha_{n-1} \\ \alpha_2 & \alpha_1 & 1 & \dots & \alpha_{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha_n & \alpha_{n-1} & \alpha_{n-2} & \dots & 1 \end{bmatrix}$$

3.3 Caso de estudo

O número de notificações de casos de dengue será explicado considerando o ciclo de desenvolvimento do mosquito transmissor, até a manifestação da doença na pessoa infectada, dada pelo registo no serviço de saúde. Para tal, foram criadas variáveis desfasadas em $0, 1, 2, \dots, 10$ semanas, a partir das variáveis precipitação ($prec$), temperatura mínima

($tmin$), temperatura máxima ($tmax$) e média da humidade relativa do ar ($hram$).

Na Tabela 3.1 apresentam-se os coeficientes de correlação de *Spearman* entre o número de notificações de casos de dengue e as variáveis meteorológicas desfasadas (*lag*) em 0 a 10 semanas, para os dados da cidade de Goiânia. Observam-se associações positivas entre o número de notificações de casos de dengue e a precipitação, a temperatura mínima e a média humidade relativa do ar, com as maiores correlações para os desfasamentos de seis, sete e cinco semanas, respetivamente. A temperatura máxima apresenta baixa associação negativa e estatisticamente não significativa, invertendo-se a partir da sexta semana de desfasamento, não sendo incluída no modelo ajustado.

Tabela 3.1: Correlações das variáveis meteorológicas com o número de notificações de casos de dengue contabilizados semanalmente na cidade de Goiânia.

Lag	prec (mm)	tmin (°C)	tmax (°C)	hram (%)
0	0.4040**	0.4162***	**-0.2895***	0.5716***
1	0.4651**	0.4820***	-0.2206	0.6026***
2	0.5064**	0.5392***	-0.1609	0.6156***
3	0.5459***	0.5915***	-0.1233	0.6401***
4	0.5861***	0.6403***	-0.0725	0.6476***
5	0.6044***	0.6652***	-0.0305	**0.6569*****
6	**0.6333*****	0.6821***	0.0152	0.6513***
7	0.6308***	**0.6988*****	0.0726	0.6238***
8	0.6254***	0.6860***	0.1138	0.5960***
9	0.6152***	0.6749***	0.1519	0.5610***
10	0.5849***	0.6556***	0.1978	0.5179***

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

Na Figura 3.1 tem-se o número de notificações de casos de dengue para o período de 2008 a 2015 contabilizados semanalmente. Observam-se valores discrepantes para o período em estudo, caracterizados pelos anos em que a cidade de Goiânia foi considerada em estado de epidemia (anos de 2008, 2010, 2013 e 2015), de acordo com os índices definidos pela OMS.

Na Figura 3.2 tem-se o número de notificações de casos de dengue para o período de 2008 a 2015 contabilizados semanalmente. Nota-se que o número de notificações de casos de dengue até a vigésima semana é superior à média do período em estudo. Essa característica é devida ao período de chuva com um verão quente e húmido, propício à formação de criadouros naturais para o desenvolvimento do mosquito *Aedes aegypti*.

A Figura 3.2 sugere que a modelação da incidência de dengue deve considerar uma componente sazonal. Pelo que, num contexto de análise exploratória dos dados, foi

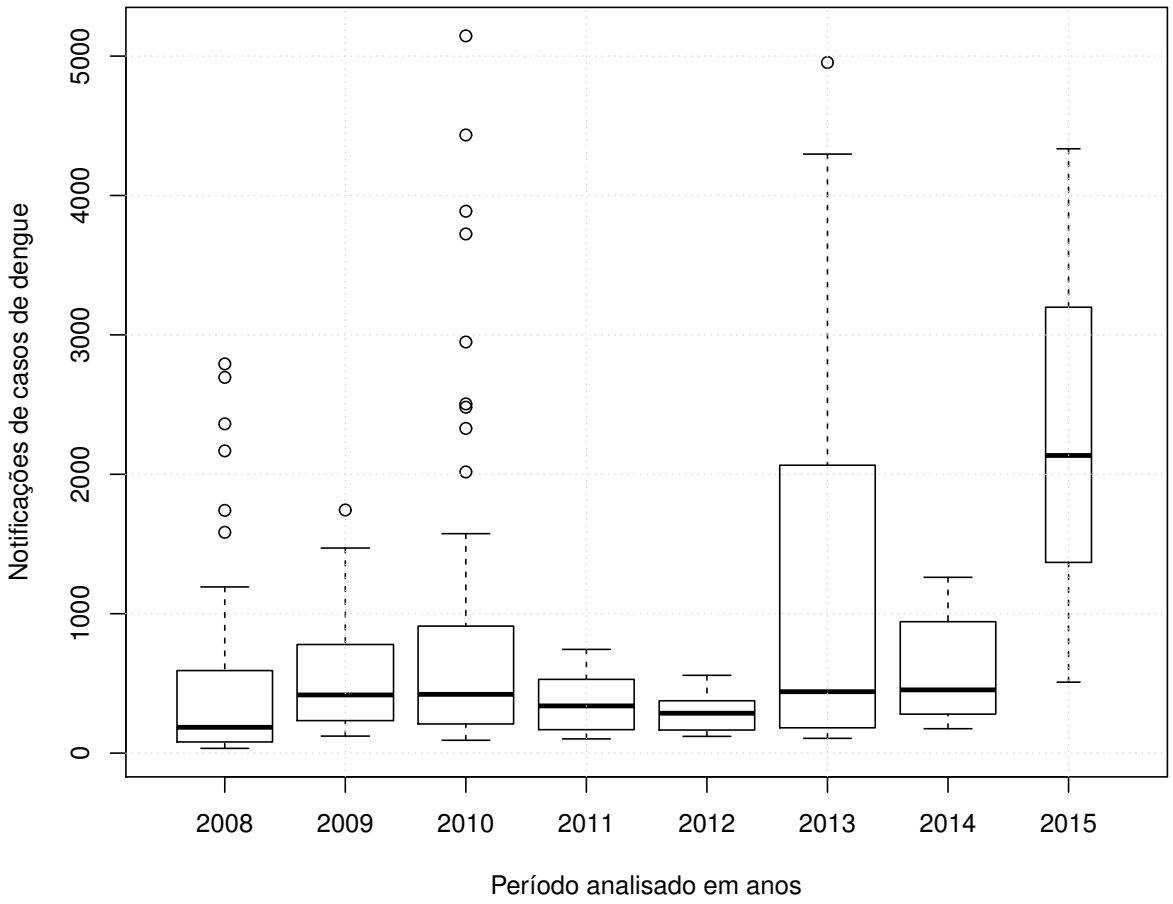


Figura 3.1: Número de notificações de casos de dengue na cidade de Goiânia, para o período de 2008 a 2015, contabilizados semanalmente.

estimada a respetiva função periodograma, confirmando-se a periodicidade anual (período estimado em aproximadamente 52 semanas). Dessa forma, iremos incorporar no modelo uma regressão harmónica, envolvendo funções trigonométricas *seno* e *cosseno* (Bailey et al., 2013). Especificamente, o modelo proposto irá considerar as funções definidas por:

$$cos\left(\frac{2\pi t}{52}\right), \qquad sin\left(\frac{2\pi t}{52}\right),$$

$$t = 1, \ldots, 53,$$

onde t identifica a semana de um determinado ano.

A Figura 3.3 apresenta as variáveis meteorológicas e o número de notificações de casos de dengue para o período de 2008 a 2015 na cidade de Goiânia, contabilizadas semanalmente. Em 2008, 2010, 2013 e 2015, as notificações são superiores a 300 casos por 100 mil habitantes, períodos considerados epidémicos pela Organização Mundial de Saúde (OMS).

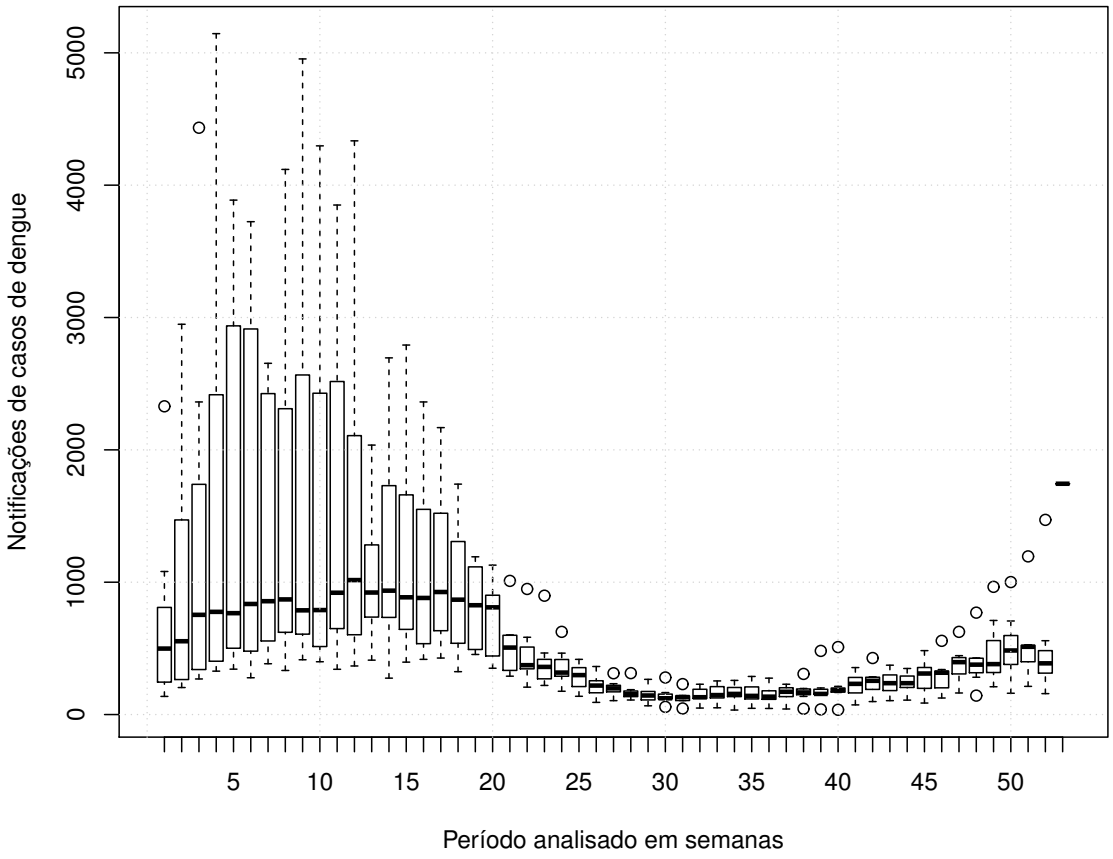


Figura 3.2: Número de notificações de casos de dengue na cidade de Goiânia, para o período de 2008 a 2015, contabilizados semanalmente.

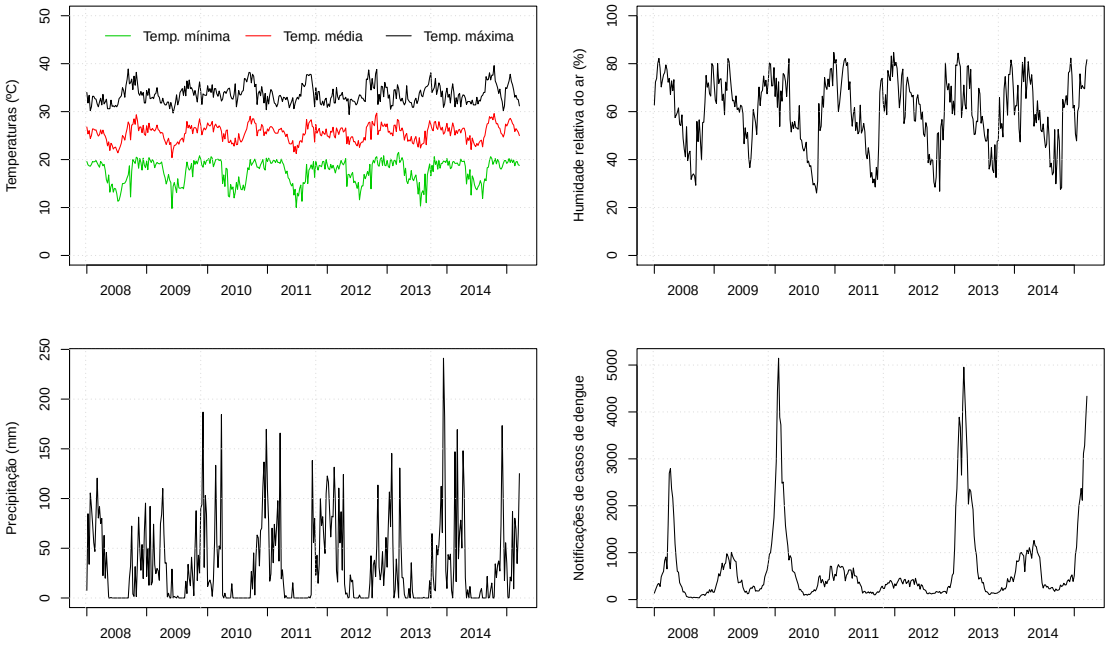


Figura 3.3: Variáveis meteorológicas e número de notificações de casos de dengue na cidade de Goiânia, contabilizadas semanalmente.

Para o período em análise, a cidade de Goiânia apresenta temperatura mínima variando entre 9.8 °C e 21.5 °C e a temperatura máxima entre 29.4 °C e 39.6 °C, com a temperatura entre 21.5 °C e 29.4 °C, correspondendo ao período que se regista o maior número de notificações de casos dengue. A média da humidade relativa do ar varia entre 26.04 % e 84.68 %, registando-se o maior número de notificações de casos de dengue para o período em que se mantém acima da média observada, isto é, 60.31 %. O período de chuva com maior registo de casos de dengue corresponde aos meses de outubro a abril, com a precipitação acumulada semanal acima da média do período, isto é, 34.04 mm.

Dada a existência de correlação temporal subjacente aos nossos dados, recorre-se à metodologia de estimação das GEE, conforme explicado na Secção (3.2.3). Por se tratar de dados de contagem, inicialmente ajustam-se Modelos Lineares Generalizados (*Generalized Linear Models* - GLM) com a distribuição de Poisson.

Assumindo que a variável resposta é proveniente da distribuição de Poisson, e recorrendo à função de ligação logarítmica, então o modelo poderá ser apresentado do seguinte modo:

$$
\begin{aligned}
Y_{it} &\sim Poi(\mu_{it}) \\
ln(\mu_{it}) = \beta_0 + \beta_1 prec_{i(t-lag_1)} + \beta_2 tmin_{i(t-lag_2)} + \beta_3 hram_{i(t-lag_3)} + \\
\beta_4 cos\left(\frac{2\pi t}{52}\right) + \beta_5 sen\left(\frac{2\pi t}{52}\right) + ln(thab_i),
\end{aligned}
\tag{3.1}
$$

onde Y_{it} é o número de notificações de casos de dengue na semana $t = 1, \ldots, 53$ para o ano $i = 1, \ldots, 8$, μ_{it} é a respetiva esperança condicionada a um conjunto de covariáveis, $\beta_0, \beta_1, \ldots, \beta_5$ são os coeficientes do modelo, $thab_i$ é o total de habitantes no ano $i = 1, \ldots, 8$ e lag_1, lag_2, lag_3 representam eventuais desfasamentos para as variáveis meteorológicas.

O número de notificações de casos de dengue apresenta uma variância muito superior à média, fato que sugere a presença de sobredispersão dos dados. Ao identificar-se sobredispersão, optou-se pela distribuição Binomial Negativa (Hilbe, 2011), onde $E[Y] = \mu$ e $Var[Y] = \mu + \alpha\mu^2$, para ajustar um modelo que explicasse a incidência dos casos de dengue semanalmente a partir das variáveis meteorológicas desfasadas no tempo (Gomes et al., 2012).

Ao assumir que a variável resposta é proveniente da distribuição Binomial Negativa, e que se utiliza a função de ligação logarítmica, tem-se:

$$Y_{it} \sim BinNeg(\mu_{it}, \alpha)$$

sendo $ln(\mu_{it})$ dado pela equação (3.1).

Foram ajustados diversos modelos considerando as distribuições de Poisson e Binomial Negativa para a variável resposta, e as estruturas de correlações independente, autorregressiva de primeira ordem e dependente de segunda ordem. Os dados foram agrupados em anos, classicamente referidos como sujeitos, e o efeito intrassujeitos representado pelas semanas.

O Critério de Informação de *Akaike* (AIC), tradicionalmente aplicado a GLM, baseia-se na estimativa de máxima verosimilhança, não podendo ser aplicado à abordagem das GEE. Dessa forma, para selecionar a estrutura de correlação para o modelo que melhor se ajusta aos dados, utiliza-se o critério QIC (*Quasi-Likelihood Under Independence Model Criterion*) (Pan, 2001), desenvolvido para essa finalidade.

3.4 Resultados

Numa primeira fase, ajustaram-se modelos utilizando as covariáveis com os desfasamentos temporais que apresentaram maiores correlações com a variável resposta conforme identificados na Tabela 3.1. No entanto, os modelos ajustados não atenderam aos critérios de seleção, pelo que se procurou identificar um modelo mais adequado. Obteve-se um modelo com precipitação ($prec$), temperatura mínima ($tmin$) e humidade relativa do ar média ($hram$) desfasadas em oito, sete e dez semanas, respetivamente. A expressão geral para o modelo escolhido é representada por:

$$\begin{aligned} ln(\mu_{it}) = \beta_0 + \beta_1 prec_{i(t-8)} + \beta_2 tmin_{i(t-7)} + \beta_3 hram_{i(t-10)} + \\ \beta_4 cos\left(\frac{2\pi t}{52}\right) + \beta_5 sin\left(\frac{2\pi t}{52}\right) + ln(thab_i), \\ i = 1, \ldots, 8, \qquad t = 1, \ldots, 53 \end{aligned}$$

A Tabela 3.2 apresenta um estudo comparativo dos vários modelos, envolvendo distintas estruturas de correlação, apresentadas na Secção 3.2.3. A qualidade de ajustamento foi avaliada utilizando a estatística QIC e a significância estatística das covariáveis.

Tabela 3.2: Resultados obtidos utilizando o método das GEE com as distribuições de Poisson e Binomial Negativa, para as estruturas de correlações independente, autorregressiva de primeira ordem (AR1) e dependente de segunda ordem (M2). SC e SS identificam as funções $cos(2\pi t/52)$ e $sin(2\pi t/52)$, respetivamente, sendo importantes para modelar a sazonalidade inerente aos dados.

	GEE - ESTRUTURAS DE CORRELAÇÃO					
	Independente		**AR(1)**		**M(2)**	
	Estimativa	**Erro Padrão**	**Estimativa**	**Erro Padrão**	**Estimativa**	**Erro Padrão**
MODELOS DE POISSON						
Constante	-9.579*	1.0689	-8.031*	0.2845	-8.643*	0.1661
prec (lag 8)	-0.001	0.0018	0.000	0.0003	0.000	0.0005
tmin (lag 7)	0.077	0.0397	0.009	0.0103	0.029*	0.0089
hram (lag 10)	0.004	0.0087	-0.001	0.0016	0.002	0.0029
SC	0.465*	0.2027	0.370	0.2357	0.473	0.1746
SS	0.939*	0.3416	1.198*	0.2528	1.035*	0.2047
ϕ	398.937	-	398.937	-	398.937	-
QIC	126129	-	130965	-	127261	-
MSE	484189	-	495290	-	489589	-
MODELOS BINOMIAL NEGATIVA						
Constante	-9.294*	0.8740	-8.123*	0.1900	-9.066*	0.1562
prec (lag 8)	0.000	0.0016	0.000	0.0003	0.001*	0.0004
tmin (lag 7)	0.051	0.0310	0.011*	0.0054	0.041*	0.0141
hram (lag 10)	0.007	0.0058	-0.001	0.0012	0.006*	0.0026
SC	0.501*	0.1576	0.416*	0.1520	0.472*	0.1250
SS	0.840*	0.2902	1.083*	0.2009	0.867*	0.1699
α	0.353	-	0.353	-	0.353	-
ϕ	1.221	-	1.221	-	1.221	-
QIC	465.432	-	477.988	-	453.949	-
MSE	493052	-	495644	-	496870	-

$*p < 0.05$

3.5 Discussão

Os modelos ajustados aplicando a distribuição de Poisson não apresentaram significância estatística ao nível de 5 % para as covariáveis, com elevado valor de QIC e o parâmetro de dispersão ($\hat{\phi} = 398.93$) indicando a presença de sobredispersão nos dados. Entre os três modelos de Poisson apresentados na Tabela 3.2, o que apresentou melhor ajuste foi obtido assumindo a estrutura de correlação independente.

Conforme referido anteriormente, para mitigar o problema da sobredispersão, recorreu-se à distribuição Binomial Negativa, cujos resultados são apresentados na Tabela 3.2. O modelo ajustado com a estrutura de correlação dependente de segunda ordem foi o que apresentou o menor valor de QIC, realçando-se o facto de todas as covariáveis apresentarem significância estatística ao nível de 5 % e o baixo valor do parâmetro de dispersão ($\hat{\phi} = 1.221$). A heterogeneidade dos dados foi considerada pela estimação do parâmetro

extra de sobredispersão ($\hat{\alpha} = 0.353$), conforme indicado na Tabela 3.2. Baseado nos resultados apresentados na Tabela 3.2, conclui-se que o melhor modelo para expressar o valor médio do número de notificações de casos de dengue, usando as variáveis meteorológicas e os respetivos desfasamentos, é dado pela equação (3.2).

$$ln(\widehat{\mu_{it}}) = -9.066 + 0.001 \times prec_{i(t-8)} + 0.041 \times tmin_{i(t-7)} + 0.006 \times hram_{i(t-10)} + 0.472 \times cos\left(\frac{2\pi t}{52}\right) + 0.867 \times sin\left(\frac{2\pi t}{52}\right) + ln(thab_i), \quad (3.2)$$
$$i = 1, \ldots, 8, \qquad t = 1, \ldots 53$$

A Figura 3.4 apresenta o número de notificações de casos de dengue observado e os valores ajustados pela abordagem das equações de estimação generalizadas (GEE), recorrendo-se às distribuições de Poisson (estrutura de correlação independente) e Binomial Negativa (estrutura de correlação estacionária dependente de ordem 2), ao longo de 377 semanas dos 8 anos de nosso estudo.

Nota-se um comportamento similar entre a curva ajustada pela distribuição de Poisson com a estrutura de correlação independente e a curva ajustada pela distribuição Binomial Negativa com a estrutura dependente de segunda ordem. Para os anos epidémicos, realça-se a maior dificuldade em estimar adequadamente os picos devido à falta de informações, tais como, a população de mosquito transmissor e os programas de controles do vetor, como medidas preventivas adotadas pelo governo.

A Figura 3.5 apresenta as médias semanais para o número de notificações de casos de dengue no período de 2008 a 2015, registado na cidade de Goiânia. Nota-se que a curva ajustada pela distribuição Binomial Negativa, utilizando a estrutura estacionária dependente de ordem 2, acompanha os dados observados com maior precisão. Esse resultado indica que a heterogeneidade dos dados foi corrigida pelo parâmetro de sobredispersão ($\hat{\alpha} = 0.353$). Por outro lado, repare-se que os dados de cada ano são correlacionados até a segunda semana após serem observados, quando se consideram as funções $cos(2\pi t/52)$ e $sin(2\pi t/52)$ para modelar a sazonalidade inerente aos dados.

A cidade de Goiânia é responsável por mais de 80% dos casos de dengue no estado de Goiás (Souza et al., 2010), apresentando correlação positiva entre o número de casos de dengue e a pluviosidade, com maior aumento dos casos de dengue nos primeiros meses de cada ano, indo ao encontro dos resultados aqui apresentados.

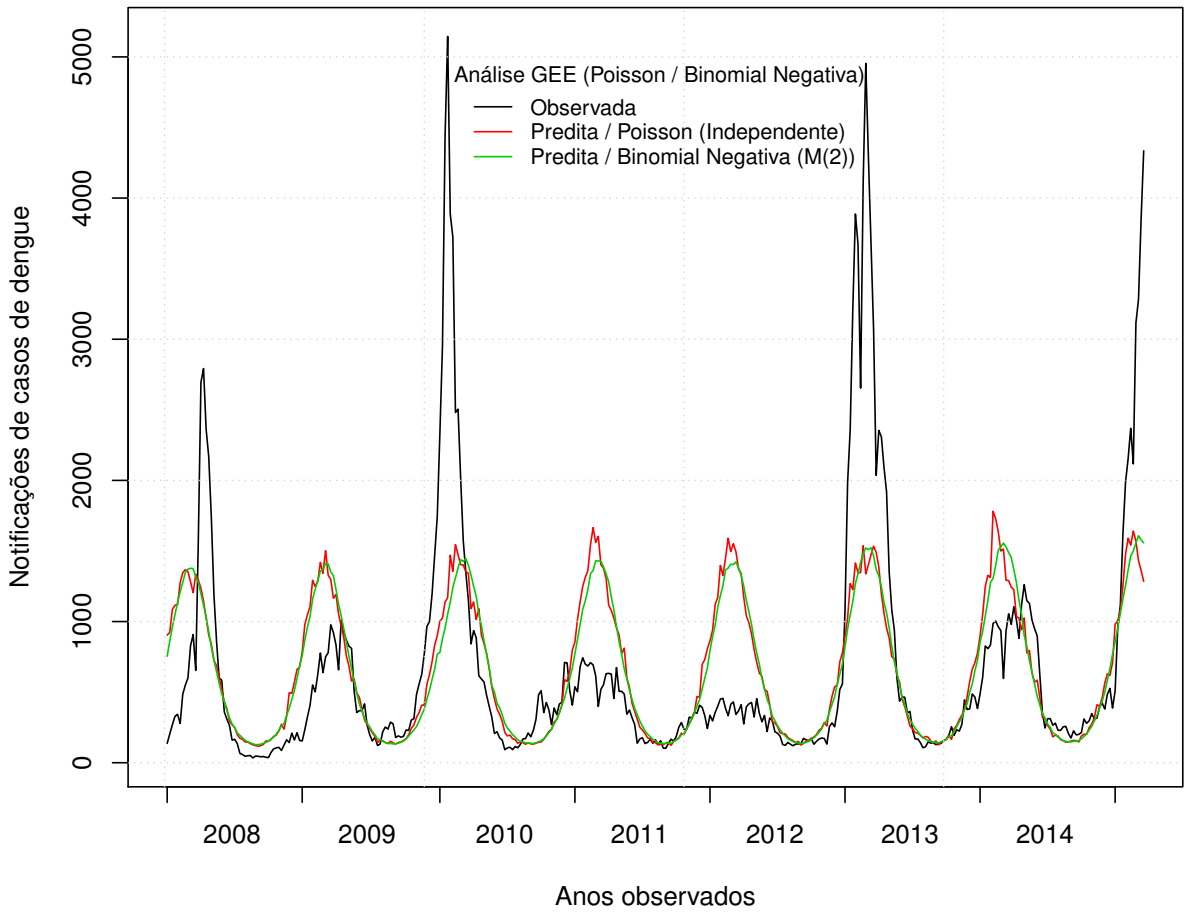


Figura 3.4: Número de notificações de casos de dengue observados e valores ajustados pela abordagem das equações de estimação generalizadas (GEE) utilizando as distribuições de Poisson e Binomial Negativa.

3.6 Considerações finais

O mosquito *Aedes aegypti*, vetor transmissor da dengue no Brasil, desenvolve-se em ambientes com água parada e temperatura média propícia entre 21 °C e 29 °C. Entre os meses de outubro a abril, a cidade de Goiânia apresenta um período quente e húmido com a precipitação acumulada mensal acima de 100 mm, contribuindo para a acumulação de água parada e a formação de criadouros do vetor transmissor da dengue.

Neste capítulo, a taxa de notificação de casos de dengue semanal na cidade de Goiânia foi modelada utilizando a abordagem das GEE com a estrutura de correlação dependente de segunda ordem, por se tratar de dados longitudinais com correlação temporal. Assumiu-se a distribuição Binomial Negativa para a variável resposta considerar a presença de sobredispersão nos dados, e o modelo ajustado foi selecionado tendo em conta o menor valor da estatística QIC.

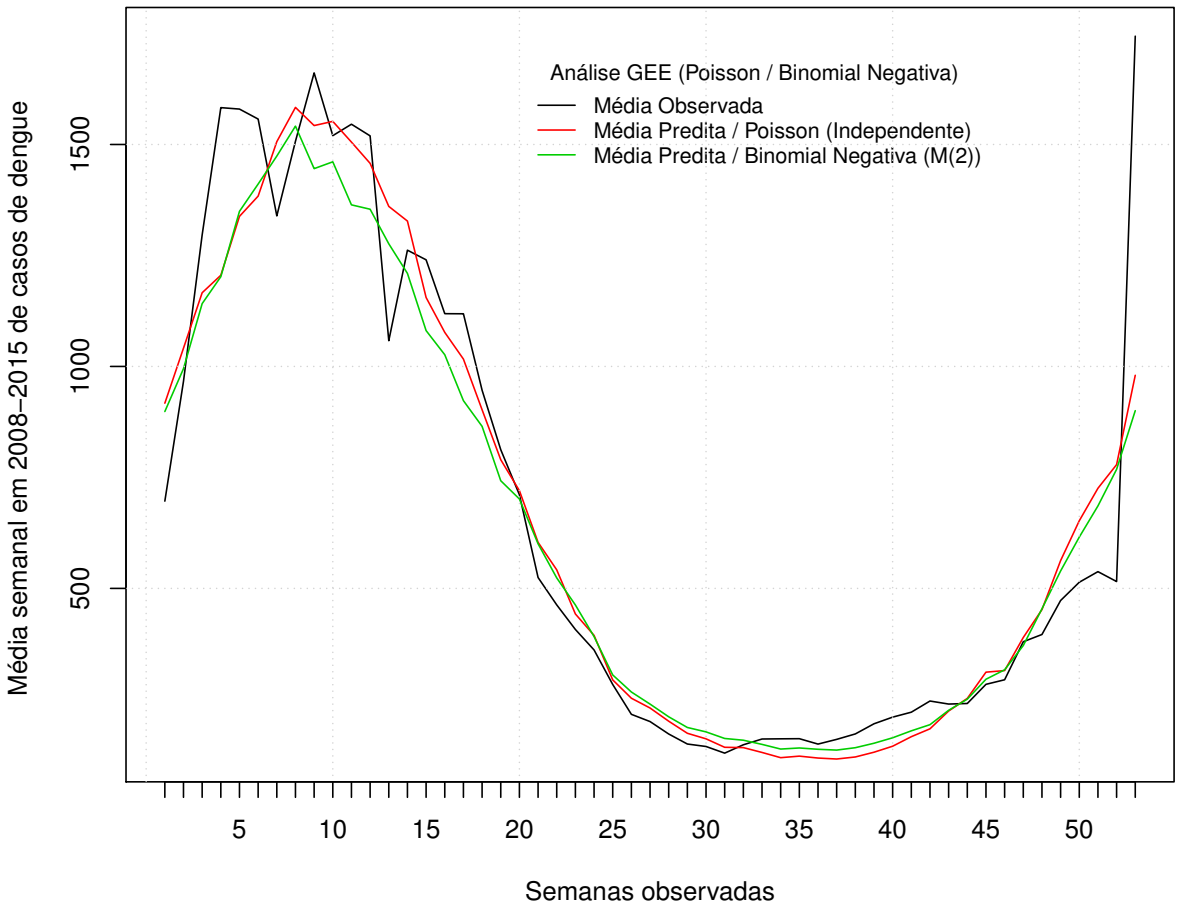


Figura 3.5: Número de notificações de casos de dengue observados e valores ajustados pela abordagem das equações de estimação generalizadas (GEE) utilizando as distribuições de Poisson e Binomial Negativa.

A partir dos valores estimados para os parâmetros e apresentados na Tabela 3.2, para uma alteração no valor de uma variável preditora, mantendo constantes as demais variáveis, temos:

- O aumento de 1 °C na temperatura mínima provoca um aumento esperado de 4.1 % na taxa de notificação de casos de dengue passadas sete semanas;
- A cada incremento de 10 % na humidade relativa do ar, temos um aumento esperado aproximado de 6 % na taxa de notificação de casos de dengue passadas dez semanas;
- O aumento da precipitação acumulada de 100 mm, causa um aumento esperado de 10 % na taxa de notificação de casos de dengue, após oito semanas.

Com o modelo estimado tem-se a indicação que as variações climáticas contribuem positivamente para o aumento dos casos de dengue, com uma maior manifestação nos centros urbanos.

Sob as condições climáticas favoráveis que foram indicadas, tornam-se necessárias ações públicas para a eliminação dos criadouros e dos mosquitos *Aedes aegypti*, para evitar a ocorrência de epidemias e mortes por dengue.

Capítulo 4

Modelação Espaço-Temporal

No Capítulo 2, foi apresentada a base de dados em estudo, a qual é formada pelas informações semanais das notificações de casos de dengue e pelas informações meteorológicas, contemplando 20 cidades do estado de Goiás. O Capítulo 3 foi dedicado ao estudo das notificações de casos de dengue na cidade de Goiânia, capital do Estado, tendo-se recorrido à abordagem das equações de estimação generalizadas (GEE), que permite considerar a correlação temporal inerente aos dados da dengue. Neste capítulo, iremos considerar a base de dados na sua totalidade, ou seja, as 20 cidades do estado de Goiás, no período entre janeiro de 2008 e março de 2015.

Como os dados dizem respeito a informações repetidas no tempo para diversos locais, torna-se necessário considerar a eventual existência de correlação espacial e/ou temporal. Tratando-se de dados agrupados, o número de notificações de casos de dengue será estudado recorrendo-se a modelos lineares generalizados mistos. Os modelos mistos permitem incorporar os efeitos fixos associados às variáveis meteorológicas e os efeitos aleatórios para explicar as influências espaço-temporais, caracterizadas pelos fatores *cidade*, *ano* e *semana*.

4.1 Modelo linear generalizado misto

Os modelos lineares generalizados mistos (GLMM) são uma extensão dos modelos lineares mistos, desenvolvidos para considerar distribuições pertencentes à família exponencial e as variações causadas por efeitos aleatórios associados aos fatores relacionados com o espaço ou tempo. Uma vez que os modelos lineares e os modelos lineares generalizados

já foram apresentados no Capítulo 3, inicia-se com a apresentação dos modelos lineares mistos e depois descreve-se os GLMM.

4.1.1 O modelo linear misto

Os *modelos lineares mistos* podem ser expressos de formas diferentes, mas equivalentes, sendo apropriados para representar dados agrupados, por exemplo, quando os dados são recolhidos ao longo do tempo para vários indivíduos, ou quando as medições são feitas em *clusters* de unidades estatísticas relacionadas (neste estudo, poderão ser cidades ou anos ou semanas).

Estes modelos têm em conta fontes naturais de heterogeneidade na população. Assim, as observações de um determinado grupo da população têm sua própria trajetória média específica sobre o tempo e um subconjunto dos coeficientes da regressão é considerado aleatório. Os modelos lineares mistos devem o seu nome ao facto de incorporarem quer efeitos fixos, isto é, parâmetros associados a toda a população, quer efeitos aleatórios, associados aos indivíduos selecionados aleatoriamente da população. O termo misto é usado para denominar o modelo que contém tanto efeitos fixos como aleatórios (McCulloch and Neuhaus, 2001).

Estes modelos são considerados adequados para dados agrupados (dados com medições repetidas, dados longitudinais, dados multinível), quantificando a relação entre a variável contínua dependente e as variáveis preditoras. Os efeitos fixos estão associados a covariáveis contínuas ou categóricas e os efeitos aleatórios associados com os fatores aleatórios (West et al., 2014). A especificação para uma observação individual Y_{ij}, representando o valor contínuo da variável resposta Y, tomado na *j-ésima* observação do *i-ésimo* grupo, é dada por:

$$\begin{aligned} Y_{ij} = \beta_0 + \beta_1 \times X_{1ij} + \beta_2 \times X_{2ij} + \cdots + \beta_p \times X_{pij} + \\ a_{i1} \times Z_{1ij} + \cdots + a_{iq} \times Z_{qij} + \varepsilon_{ij} \end{aligned} \tag{4.1}$$

$$\boldsymbol{a_i} \sim N(0, D)$$

$$\boldsymbol{\varepsilon_i} \sim N(0, R_i)$$

onde D e R_i são as matrizes de variâncias e covariâncias dos parâmetros dos efeitos aleatórios e dos erros aleatórios, respetivamente. As p covariáveis $X_1, \ldots, X_p$ são associadas

aos efeitos fixos e as q covariáveis $Z_1, \ldots, Z_q$ são associadas aos efeitos aleatórios.

4.1.2 O modelo linear generalizado misto

O modelo linear generalizado misto (GLMM) pode ser apresentado como uma extensão do modelo misto, onde se pode assumir uma distribuição para a variável resposta diferente da Gaussiana. O GLMM pode, ainda, permitir a análise de correlação entre as observações, incorporando os efeitos aleatórios. A partir das definições dos modelos lineares generalizados, apresentadas na Secção 3.2.2, podemos definir os GLMM em três etapas:

1. Distribuição da variável resposta: a distribuição da variável resposta é uma distribuição pertencente à família exponencial (Poisson, Binomial, Binomial Negativa, Gama, Normal). Neste estudo, podemos considerar que a variável dependente condicionada ao efeito aleatório associado, por exemplo, ao fator *cidade* a, segue a distribuição de Poisson:

   $$Y_{ij}|a_i \sim P(\mu_{ij})$$

   onde Y_{ij} representa o valor da variável resposta Y, medido na *j-ésima* semana, da *i-ésima* cidade. O valor médio e a variância condicional são dados, respetivamente, por:

   $$E[Y_{ij}|a_i] = \mu_{ij}$$
   $$Var[Y_{ij}|a_i] = \mu_{ij}$$

2. Componente sistemática: a especificação da componente sistemática em termos das covariáveis incorpora os efeitos fixos e aleatórios, permitindo a análise de dados com observações correlacionadas. Por exemplo, neste estudo para se considerar um efeito específico para cada cidade, tem-se:

   $$\eta_{ij} = \beta_0 + \beta_1 \times X_{1ij} + \beta_2 \times X_{2ij} + \cdots + \beta_p \times X_{pij} + a_i,$$
   $$i = 1, \ldots, 20, \qquad j = 1, \ldots, 53$$

   onde $a_i \sim N(0, \sigma_a^2)$ e representa o efeito aleatório.

3. Função de ligação: a variável resposta relaciona-se linearmente com as variáveis explicativas e os efeitos aleatórios, via preditor linear, utilizando uma função de ligação específica de acordo com a distribuição utilizada. Para a distribuição de Poisson, temos:

$$ln(\mu_{ij}) = \eta_{ij}$$
$$\mu_{ij} = e^{\eta_{ij}}$$

Entre os métodos de estimação dos parâmetros utilizados em GLMM, destaca-se o método da máxima verosimilhança utilizando a aproximação de *Laplace*.

4.2 Caso de estudo com a abordagem GLMM

Tal como referido anteriormente, a base de dados em estudo, apresentada no Capítulo 2, é formada por informações de notificações de casos de dengue e por informações das variações meteorológicas registadas em 20 cidades do estado de Goiás, contabilizadas semanalmente ao longo do período de 8 anos.

A análise da dependência espaço-temporal subjacente aos dados de contagem de dengue foi feita em duas etapas. Primeiramente, analisou-se a dependência espacial à custa do índice de Moran, uma medida de associação espacial. A estimação desta medida depende da definição de uma matriz de pesos, delineada tendo em consideração que as observações mais próximas são mais prováveis de serem semelhantes que as observações distantes. No nosso estudo, o índice de Moran foi estimado utilizando a matriz de distâncias inversas calculadas a partir da distância Euclidiana entre as cidades (identificadas pelas coordenadas Longitude e Latitude dos seus "centroides"). O índice de Moran permite-nos prosseguir com um teste de inferência, sendo o seu valor esperado debaixo da hipótese nula, segundo a qual não há correlação espacial, comparado com o valor do índice calculado à custa dos dados. Os resultados do teste de Moran permitiram-nos concluir que, no caso da variável *dengue* padronizada tendo em conta a população de cada cidade, a auto-correlação espacial não é estatisticamente significativa. Tal poderá ser justificado pelo grande afastamento entre a maioria das 20 cidades (ver Figura 2.1).

Seguidamente, prosseguiu-se com a análise de correlações temporais subjacentes aos nossos dados de contagem de dengue, considerando a existência de 20 séries temporais,

uma por cada cidade da região de estudo. Foram realizados testes de significância estatística de correlações entre semanas, recorrendo-se ao coeficiente de *Spearman*, por não exigir o pressuposto de normalidade dos dados. O nosso estudo envolveu a estimação de auto-correlações dentro de cada série temporal, assim como a estimação de correlações cruzadas entre duas cidades distintas. Os principais resultados são apresentados no Anexo B, concluindo-se que a maioria das auto-correlações, e correlações cruzadas, são significativamente diferentes de zero, com repetições cíclicas confirmando a presença de sazonalidade nas contagens de dengue.

Tendo em conta os resultados anteriores sobre as influências espaço-temporais presentes nos nossos dados, nas secções seguintes iremos analisar modelos mistos alternativos que consideram efeitos aleatórios essencialmente não correlacionados para captar as especificidades de cada cidade e/ou ano e/ou semana. Note-se que as variáveis meteorológicas, consideradas como efeitos fixos, deverão explicar grande parte da correlação semanal identificada para a variável dengue no estudo apresentado no Anexo B.

4.2.1 Abordagem GLMM com a distribuição de Poisson

Considerando o fator *cidade* para identificar uma possível heterogeneidade espacial e os fatores *ano* e *semana* para representarem influências temporais, temos que o número de notificações de casos de dengue, Y_{ijs}, da semana s, no ano j da cidade i, condicionado aos efeitos aleatórios cidade a_i , ano b_j e semana c_s dada por:

$$Y_{ijs}|a_i, b_j, c_s \sim Poisson(\mu_{ijs})$$

$$E[Y_{ijs}|a_i, b_j, c_s] = \mu_{ijs}$$

$$Var[Y_{ijs}|a_i, b_j, c_s] = \mu_{ijs}$$

$$\eta_{ijs} = \beta_0 + \beta_1 \times prec_{ij(s-lag_1)} + \beta_2 \times tmin_{ij(s-lag_2)} + \beta_3 \times tmax_{ij(s-lag_3)} + \beta_4 \times hram_{ij(s-lag_4)} +$$

$$\beta_5 \times vvto_{ij(s-lag_5)} + ln(thab_{ij}) + a_i + b_j + c_s$$

$$ln(\mu_{ijs}) = \eta_{ijs}$$

$$a_i \sim N(0, \sigma_a^2), \qquad i = 1, \ldots, 20$$

$$b_j \sim N(0, \sigma_b^2), \qquad j = 1, \ldots, 8$$

$$c_s \sim N(0, \sigma_c^2), \qquad s = 1, \ldots, 53$$

onde:

a) $Y_{ijs}|a_i, b_j, c_s$ indica o número de notificações de casos de dengue na semana s, do ano j, na cidade i, condicionado aos efeitos aleatórios cidade a_i, ano b_j e semana c_s.

b) a distribuição de Y_{ijs} condicional aos efeitos aleatórios a_i, b_j e c_s é a distribuição de Poisson com média μ_{ijs} e variância igual a μ_{ijs}.

c) η_{ijs} é o preditor linear com o logaritmo natural do total de habitantes ($ln(thab)$) como variável *offset*, para considerar o efeito do número de habitantes de cada cidade/ano em estudo.

d) a_i, b_j, c_s são os efeitos aleatórios associados a cidade i, ano j e semana s, permitindo uma ordenada na origem diferente para cada nível dos efeitos aleatórios.

e) Assume-se que a_i, b_j, c_s são normalmente distribuídos com média zero e variância σ_a^2, σ_b^2 e σ_c^2 respetivamente.

f) Os possíveis desfasamentos para as variáveis meteorológicas são representados por $lag_1, \ldots, lag_5$.

4.2.2 Abordagem GLMM com a distribuição Binomial Negativa

Tal como referido anteriormente, identificada a sobredispersão nos dados pela análise realizada assumindo a distribuição de Poisson, iremos também aqui considerar a distribuição Binomial Negativa. Neste caso, teremos então

$$Y_{ijs}|a_i, b_j, c_s \sim NegBin(\mu_{ijs}, k)$$

$$E[Y_{ijs}|a_i, b_j, c_s] = \mu_{ijs}$$

$$Var[Y_{ijs}|a_i, b_j, c_s] = \mu_{ijs} + \frac{\mu_{ijs}^2}{k}$$

$$\eta_{ijs} = \beta_0 + \beta_1 \times prec_{ij(s-lag_1)} + \beta_2 \times tmin_{ij(s-lag_2)} + \beta_3 \times tmax_{ij(s-lag_3)} +$$
$$\beta_4 \times hram_{ij(s-lag_4)} + \beta_5 \times vvto_{ij(s-lag_5)} + ln(thab_{ij}) + a_i + b_j + c_s$$

$$ln(\mu_{ijs}) = \eta_{ijs}$$

$$a_i \sim N(0, \sigma_a^2)$$

$$b_j \sim N(0, \sigma_b^2)$$

$$c_s \sim N(0, \sigma_c^2)$$

onde:

a) $Y_{ijs}|a_i, b_j, c_s$ indica o número de notificações de casos de dengue na semana s, do ano j, na cidade i, condicionado aos efeitos aleatórios cidade a_i, ano b_j e semana c_s.

b) a distribuição de Y_{ijs} condicional aos efeitos aleatórios a_i, b_j e c_s é a distribuição Binomial Negativa com média μ_{ijs} e variância $\mu_{ijs} + \frac{\mu_{ijs}^2}{k}$.

c) η_{ijs} é a função de predição linear, tal como definido no caso da distribuição de Poisson.

d) a_i, b_j, c_s indicam a adição dos efeitos aleatórios específicos para a cidade i, ano j e semana s, tal como definido no caso da distribuição de Poisson.

e) Assume-se que a_i, b_j, c_s são normalmente distribuídos com média zero e variância σ_a^2, σ_b^2, σ_c^2, tal como definido no caso da distribuição de Poisson.

f) A distribuição Binomial Negativa exige que a variância seja definida à custa de parâmetro extra de sobredispersão k, nomenclatura adotada neste trabalho em conformidade com o *software* R. Na literatura são encontradas duas formas de indicar a sobredispersão, θ e k, sendo $\theta = k^{-1}$. *Softwares*, como o SPSS, utilizam o θ para representar o parâmetro de sobredispersão.

No *software* R há vários pacotes que tratam dos modelos lineares generalizados mistos. No entanto, a maioria não considera a distribuição Binomial Negativa. Dessa forma, o nosso estudo considera os pacotes *lme4* e *glmmADMB*.

4.2.3 Seleção do modelo que melhor se ajusta aos dados

Para selecionar o modelo que melhor se adequa ao nosso estudo de caso, foram consideradas a significância estatística dos coeficientes estimados para os efeitos fixos e a estatística AIC, tendo em conta distintas combinações dos efeitos aleatórios. Os modelos analisados tiveram em consideração diferentes aspectos, nomeadamente:

- Distribuições da variável dependente: Poisson e Binomial Negativa, com função de ligação *log*.
- Efeitos fixos: precipitação *(prec)*, temperatura mínima *(tmin)*, temperatura máxima *(tmax)*, humidade relativa do ar *(hram)*, velocidade do vento *(vvto)* e o logaritmo natural do total de habitantes como variável *offset (ln(thab))*.
- Efeitos aleatórios: *cidade, ano* e *semana.*

Foram tidos em conta desfasamentos para as variáveis meteorológicas com a finalidade de representar o 'ciclo da dengue', desde a geração do mosquito *Aedes aegypti* pela eclosão dos ovos, incluindo as fases de desenvolvimento do mosquito, de incubação da doença no mosquito, de transmissão e a manifestação dos sintomas nos seres humanos, até o momento que é registada a notificação de caso de dengue pelo serviço de saúde.

Inicialmente, foram estimados os coeficientes para as variáveis meteorológicas com os desfasamentos que apresentaram as maiores correlações de *Spearman* (Tabela 2.5). No entanto, as estimativas não apresentaram significância estatística ao nível de 5 % e, tal como esperado, a escolha da distribuição de Poisson indicou a presença de sobredispersão dos dados.

Foram testadas combinações para vários desfasamentos das variáveis meteorológicas (nomeadamente 0, 2, 4, 6, 8 e 10 semanas), assumindo-se um modelo linear generalizado com distribuição Binomial Negativa. Nesta fase, foi identificado o modelo GLM como mais adequado com as seguintes covariáveis: precipitação desfasada em seis semanas ($prec_{(s-6)}$), temperatura mínima desfasada em quatro semanas ($tmin_{(s-4)}$), temperatura máxima na semana corrente ($tmax_s$), humidade relativa do ar desfasada em dez semanas ($hram_{(s-10)}$) e velocidade do vento desfasada em duas semanas ($vvto_{(s-2)}$), sendo os desfasamentos definidos em relação à variável dengue.

Seguidamente, foi analisado o modelo linear generalizado misto assumindo que a variável resposta segue a distribuição Binomial Negativa, considerando distintas combinações

dos efeitos aleatórios associados aos fatores *cidade*, *ano* e *semana*. Tal como anteriormente, a seleção do modelo foi realizada com base no menor valor da estatística AIC e na análise da significância estatística das estimativas dos coeficientes de regressão para os efeitos fixos, tendo-se também em conta o comportamento das componentes de variância para os efeitos aleatórios.

4.3 Análise dos principais modelos ajustados

Na Tabela 4.1 são apresentados os resultados dos principais modelos ajustados, considerando a significância estatística ao nível de 5 % para os coeficientes estimados. Os efeitos aleatórios foram incluídos nos modelos, assumindo uma estrutura de correlação independente, para considerar a influência de cada fator (*cidade, ano* e *semana*) no número de notificações de casos de dengue.

Tabela 4.1: Modelos ajustados com GLM (Poisson e Binomial Negativa) e GLMM (Binomial Negativa). Foram considerados os desfasamentos para as variáveis meteorológicas que apresentaram menores AIC. As estimativas assinaladas com (#) indicam que os valores não foram estatisticamente diferentes de zero ao nível de 5 %. Os modelos M2 a M6 incorporam efeitos fixos e aleatórios.

PARÂMETROS	**M0**	**M1**	**M2**	**M3**	**M4**	**M5**	**M6**
Constante	-13.7679	-12.6564	-12.6231	-12.8740	-11.2647	-11.5195	-10.2061
prec (lag 6)	0.0029	0.0048	0.0037	0.0056	0.0010#	0.0045	6.49e-5#
tmin (lag 4)	0.1790	0.1422	0.174	0.1277	0.0139#	0.1507	0.0531
tmax	0.0337	-0.0261	-0.0378	-0.0129#	0.0410	-0.0693	0.0073#
hram (lag 10)	0.0340	0.0418	0.0406	0.0395	0.0168	0.0385	0.0067
vvto (lag 2)	-0.5174	-0.3183	-0.4725	-0.2940	-0.2073	-0.3309	-0.2778
AIC	303103	38482	37763	37411	38041	36336	37206
k	-	0.4137	0.4837	0.5174	0.4697	0.6550	0.5674
ϕ	42.1643	1.022	1.0152	1.0178	1.0139	1.0044	1.0070
σ^2_{cid}	-	-	0.3358	-	-	0.4111	0.3410
σ^2_{ano}	-	-	-	0.4881	-	0.6598	-
σ^2_{sem}	-	-	-	-	0.6891	-	0.8175

O parâmetro de sobredispersão no nosso modelo é k, em que $k^{-1} = \theta$ (k aumenta, diminui a variância).

Todos os modelos apresentados na Tabela 4.1 consideraram as mesmas covariáveis para os efeitos fixos, e os efeitos aleatórios correspondentes à *cidade*, *ano* e *semana* foram incluídos em modelos distintos. A componente fixa do preditor linear é

$$\beta_0 + \beta_1 prec_{ij(s-6)} + \beta_2 tmin_{ij(s-4)} + \beta_3 tmax_{ijs} + \beta_4 hram_{ij(s-10)} + \beta_5 vvto_{ij(s-2)} + ln(thab_{ij})$$

Na Tabela 4.1 temos:

a) Modelo $M0$ correspondente a um GLM assumindo que a variável resposta segue a distribuição de Poisson:

$$Y_{ijs} \sim Poisson(\mu_{ijs})$$

- Note-se que o valor elevado da estimativa do parâmetro de dispersão ϕ (42.164) pode ser justificado pela sobredispersão inerente aos dados. Adicionalmente, faz-se notar o elevado valor da estatística *AIC* (303103). Estes resultados confirmam que a distribuição de Poisson não é adequada para a modelação dos dados, sugerindo que se deve recorrer à distribuição Binomial Negativa.

b) Modelo $M1$ é um modelo linear generalizado, tendo em conta que a variável resposta segue a distribuição Binomial Negativa:

$$Y_{ijs} \sim NegBin(\mu_{ijs}, k)$$

- Ao utilizar a distribuição Binomial Negativa, o novo valor da estimativa do parâmetro de dispersão ϕ (1.022) indica que a sobredispersão dos dados, detectada no modelo de Poisson ($M0$), foi adequadamente tratada pela distribuição Binomial Negativa, com a estimação do parâmetro extra de sobredispersão $k = 0.4137$. Facto que se confirma pelo valor da estatística *AIC* (38482).

c) Os modelos $M2$, $M3$ e $M4$ são modelos GLMM com efeitos aleatórios associados aos fatores *cidade* (*a*), *ano* (*b*) e *semana* (*c*), assumindo que a variável resposta segue a distribuição Binomial Negativa, tem-se:

$$M2: \quad Y_{ijs}|a_i \sim NegBin(\mu_{ijs}, k)$$

$$M3: \quad Y_{ijs}|b_j \sim NegBin(\mu_{ijs}, k)$$

$$M4: \quad Y_{ijs}|c_s \sim NegBin(\mu_{ijs}, k)$$

- Nota-se que incluir efeitos aleatórios no modelo levou à redução da estatística *AIC*, quando comparados com os modelos ajustados sem efeito aleatório.

- No modelo $M2$, incluindo o efeito aleatório da *cidade*, todas as estimativas dos coeficientes de regressão são estatisticamente diferentes de zero e o *AIC* igual a 37763, privilegiando a inclusão do fator *cidade* no modelo final.
- Os modelos $M3$ e $M4$ apresentaram estimativas estatisticamente não significativas ao nível de 5 % para as variáveis *tmax*, *prec6* e *tmin4*, conforme indicado na Tabela 4.1 pelo símbolo #.

d) O modelo $M5$ é um modelo GLMM com os efeitos aleatórios da *cidade* e do *ano*, assumindo que a variável resposta segue a distribuição Binomial Negativa:

$$Y_{ijs}|a_i, b_j \sim NegBin(\mu_{ijs}, k)$$

- Todas as variáveis do modelo são estatisticamente significativas. O valor da estatística *AIC* (36366) sugere que o modelo $M5$ é preferível aos modelos mais simples $M2$ e $M3$, sendo apresentado como:

e) O modelo $M6$ é um modelo GLMM com os efeitos aleatórios da *cidade* e da *semana*, assumindo que a variável resposta segue a distribuição Binomial Negativa:

$$Y_{ijs}|a_i, c_s \sim NegBin(\mu_{ijs}, k)$$

- As variáveis do modelo não apresentam significância estatística para as estimativas associadas à *prec6* e *tmax*, além da estatística *AIC* (37206) não ser menor que a estatística *AIC* do modelo $M5$. Esses resultados sugerem que o fator *semana* não deve ser incluído no modelo final.

f) Dessa forma, $M5$ foi selecionado como o modelo que melhor se ajusta aos dados, por apresentar todas as estimativas dos coeficientes de regressão associados aos efeitos fixos com significância estatística ao nível de 5 % e menor valor da estatística *AIC*. Conforme explicado na Secção 4.2.2, o modelo $M5$ é dado por:

$$Y_{ijs}|a_i, b_j \sim NegBin(\mu_{ijs}, k)$$

$$E[Y_{ijs}|a_i, b_j] = \mu_{ijs}$$

$$Var[Y_{ijs}|a_i, b_j] = \mu_{ijs} + \frac{\mu_{ijs}^2}{k}$$

$$\eta_{ijs} = \beta_0 + \beta_1 \times prec_{ij(s-6)} + \beta_2 \times tmin_{ij(s-4)} + \beta_3 \times tmax_{ijs} + \beta_4 \times hram_{ij(s-10)} +$$

$$\beta_5 \times vvto_{ij(s-2)} + ln(thab_{ij}) + a_i + b_j$$

$$ln(\mu_{ijs}) = \eta_{ijs}$$

$$a_i \sim N(0, \sigma_a^2), \qquad i = 1, \ldots, 20,$$

$$b_j \sim N(0, \sigma_b^2), \qquad j = 1, \ldots, 8,$$

$$s = 1, \ldots, 53$$

Na Figura 4.1 temos o histograma dos resíduos de Pearson para o modelo $M5$, o qual apresenta uma curva assimétrica, com cauda mais pesada à direita, característica das distribuições não normais, como é o caso da distribuição Binomial Negativa.

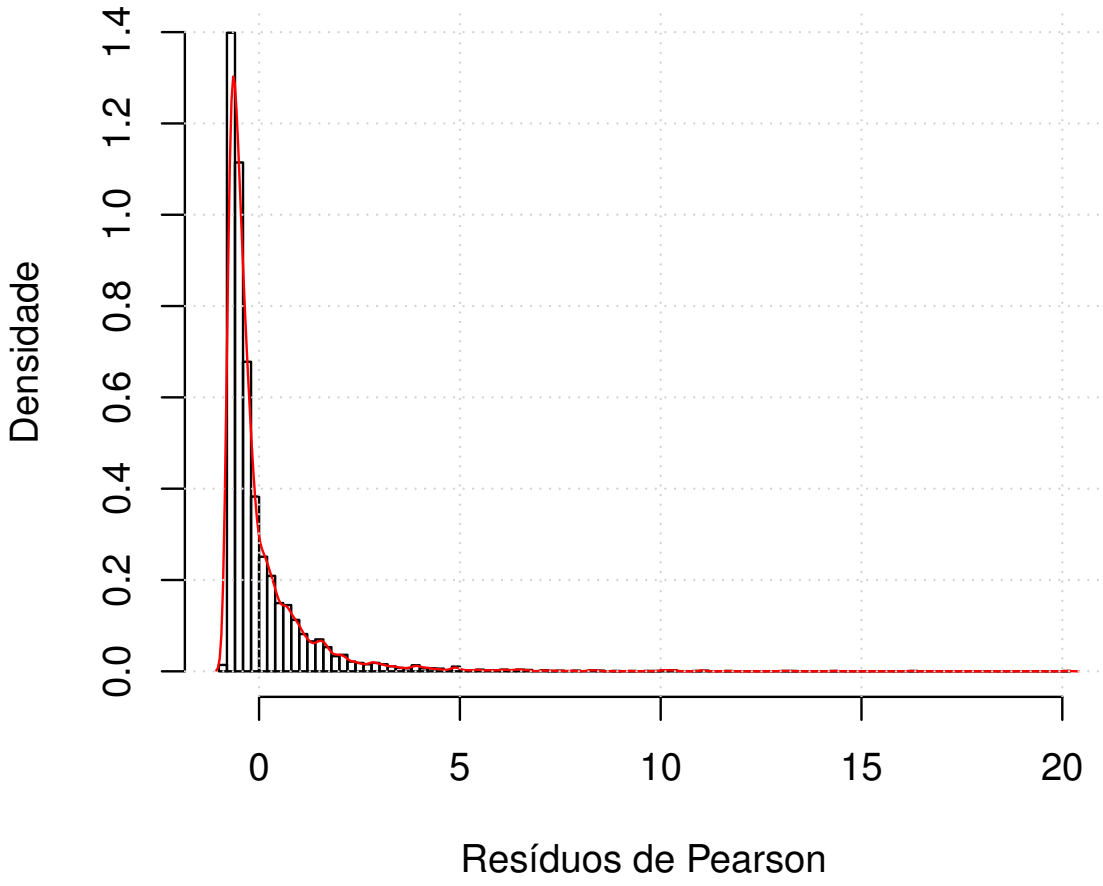


Figura 4.1: Histograma dos resíduos de Pearson com curva assimétrica característica da distribuição Binomial Negativa.

Para a verificação dos pressupostos do modelo ajustado, McCullagh e Nelder (1989) recomendam o uso dos resíduos da desviância, pois estes podem ter distribuições bem próximas da distribuição Gaussiana. Assim, na Figura 4.2 tem-se o histograma dos resíduos da desviância (*deviance residuals*) que mostra uma curva simétrica com aparência de normalidade, apesar de estar centrada à esquerda.

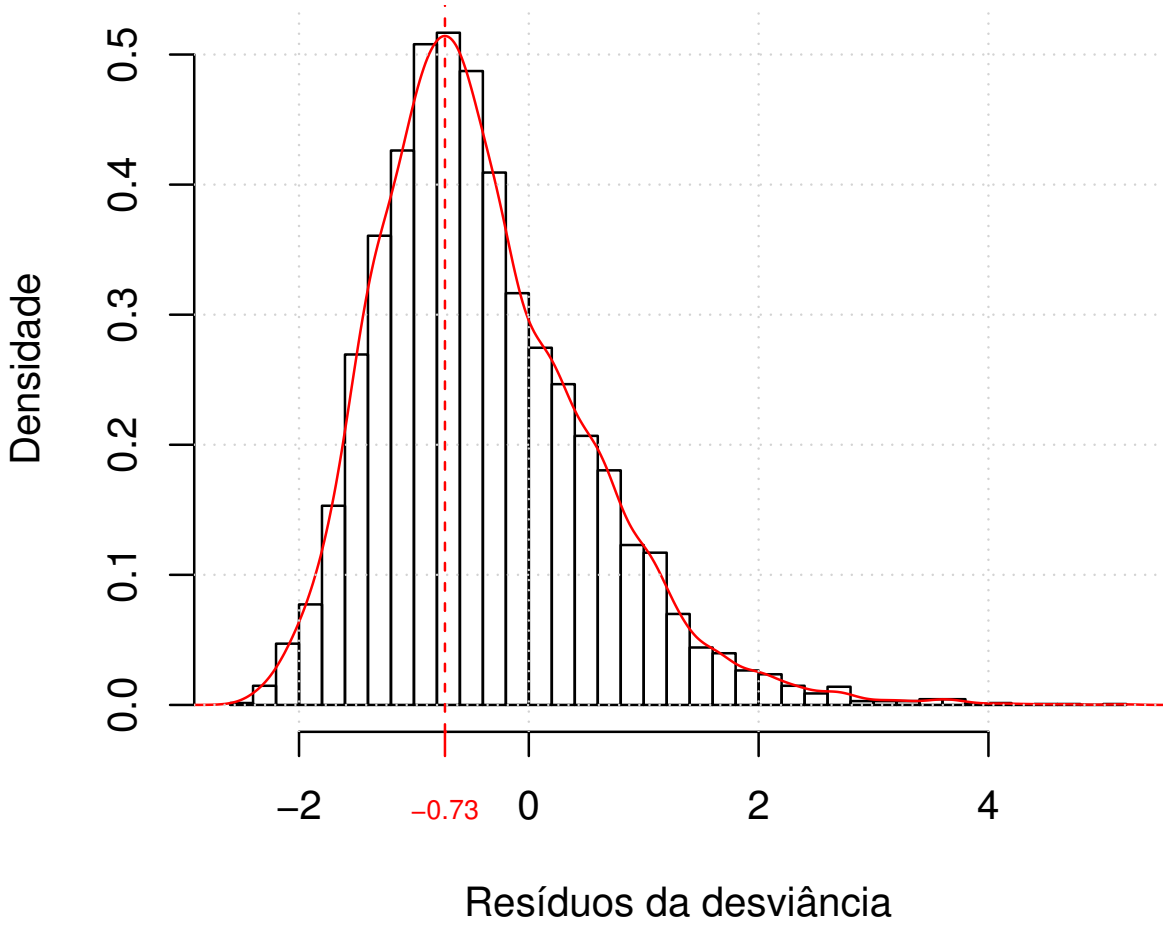


Figura 4.2: Histograma dos resíduos da desviância com densidade estimada pelo método de *kernel*.

Na Figura 4.3, o gráfico de dispersão dos resíduos da desviância padronizados mostra aproximadamente 95 % dos resíduos entre o intervalo [-2, 2], sem uma identificação de ocorrência de padrões. Os pontos fora desse intervalo indicam valores atípicos (*outliers*) correspondendo aos períodos considerados epidémicos, de acordo com os critérios definidos pela OMS.

Na Figura 4.4 (a) temos a representação dos resíduos de Pearson *versus* os valores ajustados e na Figura 4.4 (b) temos a representação dos resíduos da desviância *versus* os valores ajustados. Notam-se grande concentração de pontos no valor zero correspondendo

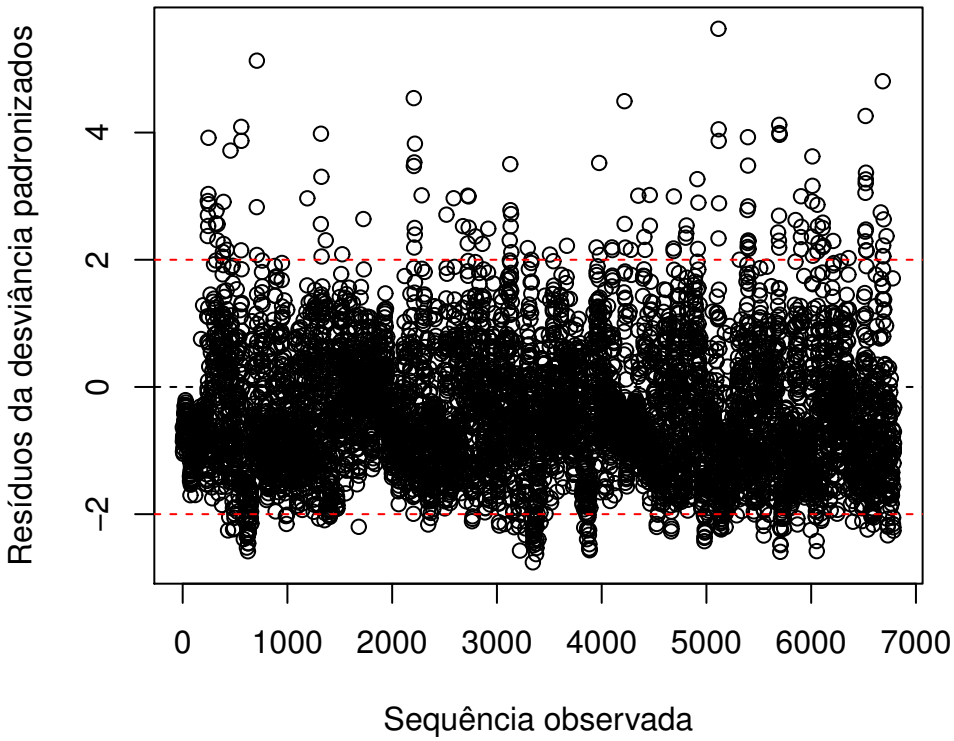


Figura 4.3: Resíduos da desviância padronizados *vs.* sequência de observação.

aos períodos de inverno (períodos de seca) em que não há registos de notificações de casos de dengue, principalmente nas pequenas cidades. Observe-se, também, que a linha de suavização se afasta da linha horizontal em zero indicando os valores que foram mal estimados para os períodos epidémicos ocorridos, principalmente na cidade de Goiânia, nos anos de 2010, 2013, 2014 e 2015.

Os efeitos aleatórios estimados pelo modelo selecionado *M5*, referentes ao fator *cidade*, são apresentados na Figura 4.5, a qual permite identificar as cidades de Goiânia, Jataí e Luziânia como tendo valores esperados de dengue mais elevados. De forma inversa, as cidades de Monte Alegre, Cristalina e Ipameri apresentam valores esperados de dengue mais reduzidos.

A Figura 4.6 mostra os efeitos temporais estimados no modelo *M5* referentes aos anos de 2008 a 2015, permitindo identificar os anos 2010, 2013, 2014 e 2015 como tendo valores esperados de dengue mais elevados. Esses valores são impulsionados pelos períodos epidémicos, principalmente na cidade de Goiânia. Nota-se, também, que os anos 2008 e 2012 foram os que apresentaram menores índices de notificações de casos de dengue.

A interpretação dos coeficientes de regressão estimados, para os efeitos fixos apresen-

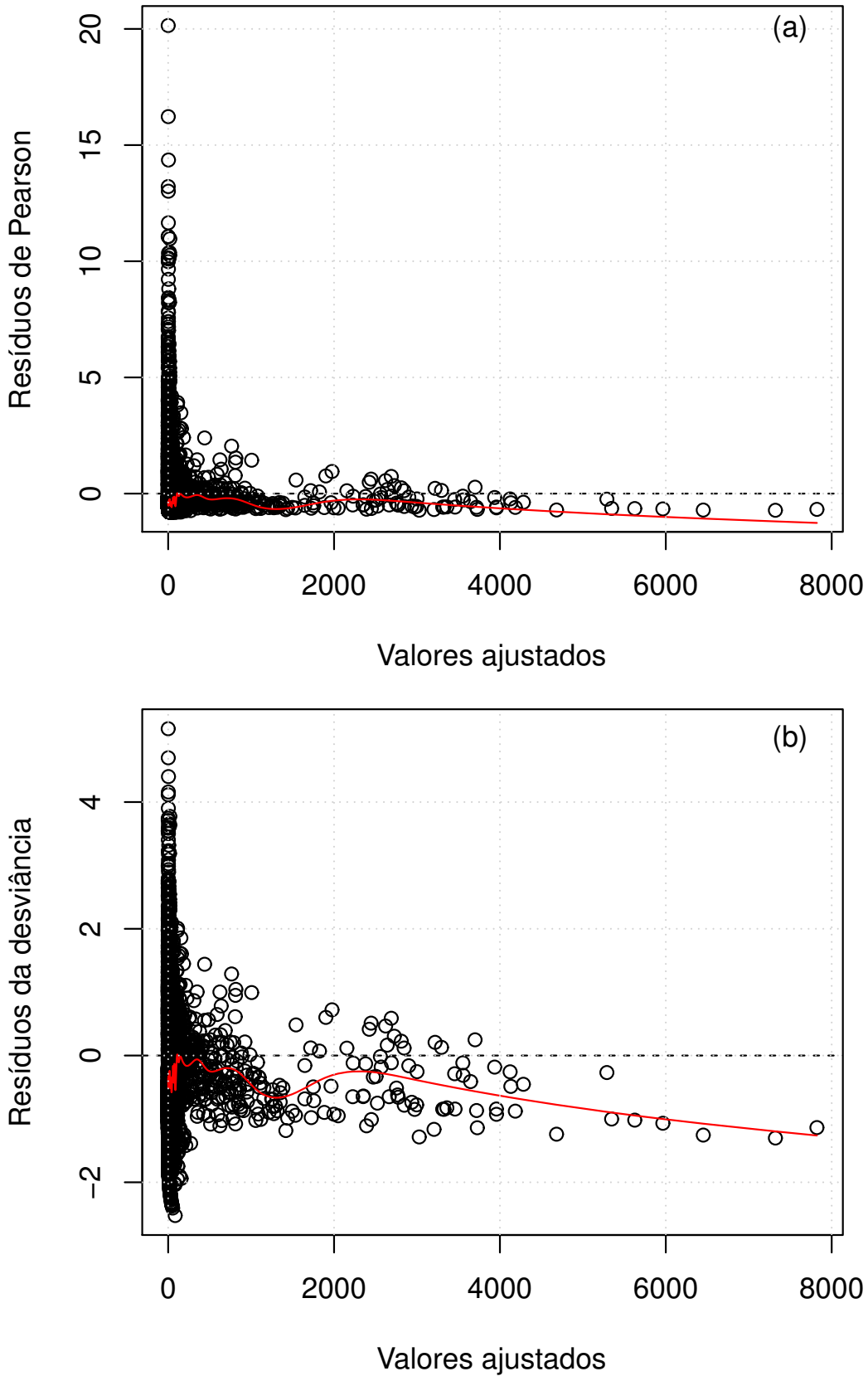


Figura 4.4: Apresentação dos resíduos para o modelo M5. a) Resíduos de Pearson *vs.* Valores ajustados; b) Resíduos da desviância padronizados *vs.* Valores ajustados.

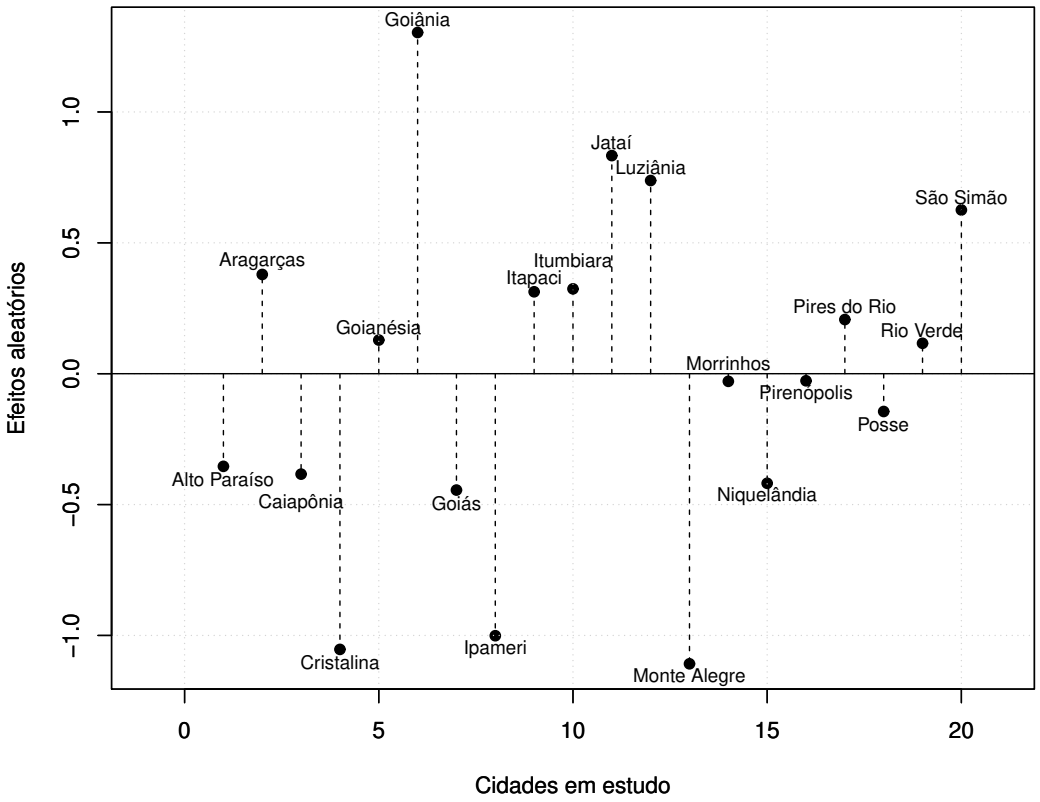


Figura 4.5: Efeitos aleatórios estimados para as 20 cidades do estado de Goiás.

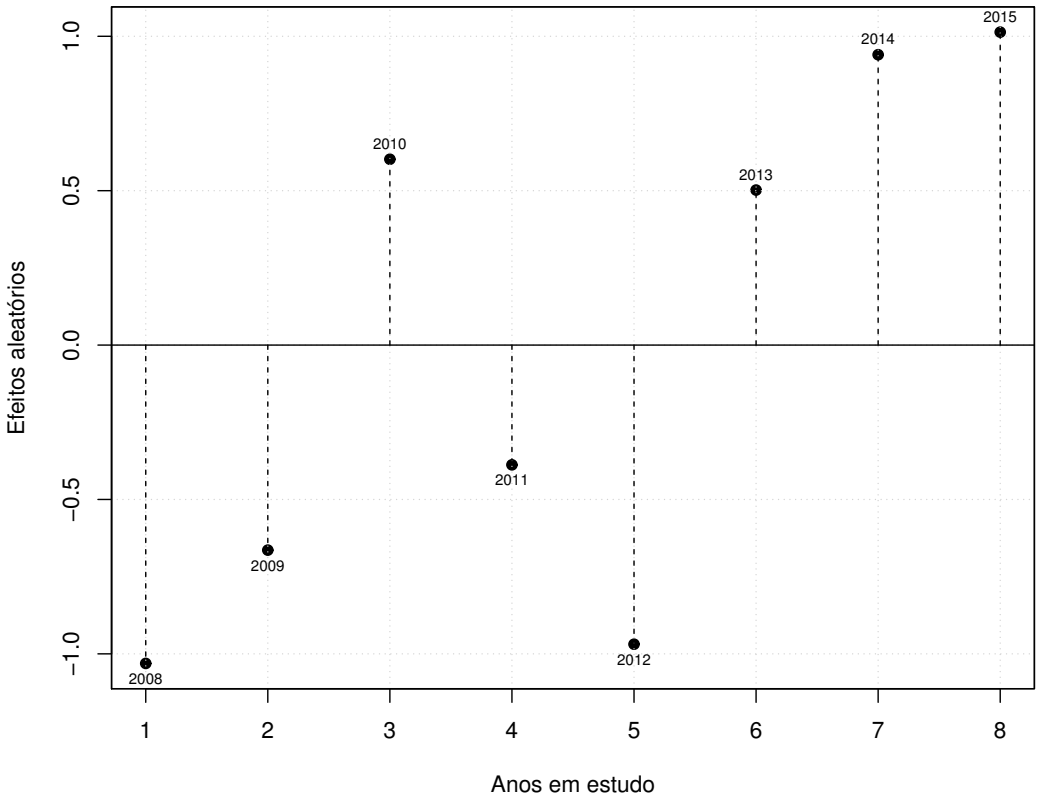


Figura 4.6: Efeitos aleatórios estimados para os anos de 2008 a 2015.

tados na Tabela 4.1, permite concluir que a precipitação, a temperatura mínima e a humidade relativa do ar têm influência positiva no número de notificações de casos de dengue, enquanto a velocidade do vento e a temperatura máxima influenciam negativamente.

Na Figura 4.7, temos a representação do número de notificações de casos de dengue para os anos em estudo. São mostrados os valores médios semanais ajustados com o intervalo de confiança de 95 % no modelo *M5*, juntamente com os valores médios semanais observados. Nota-se uma média semanal sobrestimada para o ano de 2014, influenciada pelo grande número de notificações de casos de dengue ocorrido na cidade de Goiânia no ano de 2013 (média de 1187 casos semanais), ano que foi considerado epidémico em conformidade com os parâmetros estabelecidos pelas OMS.

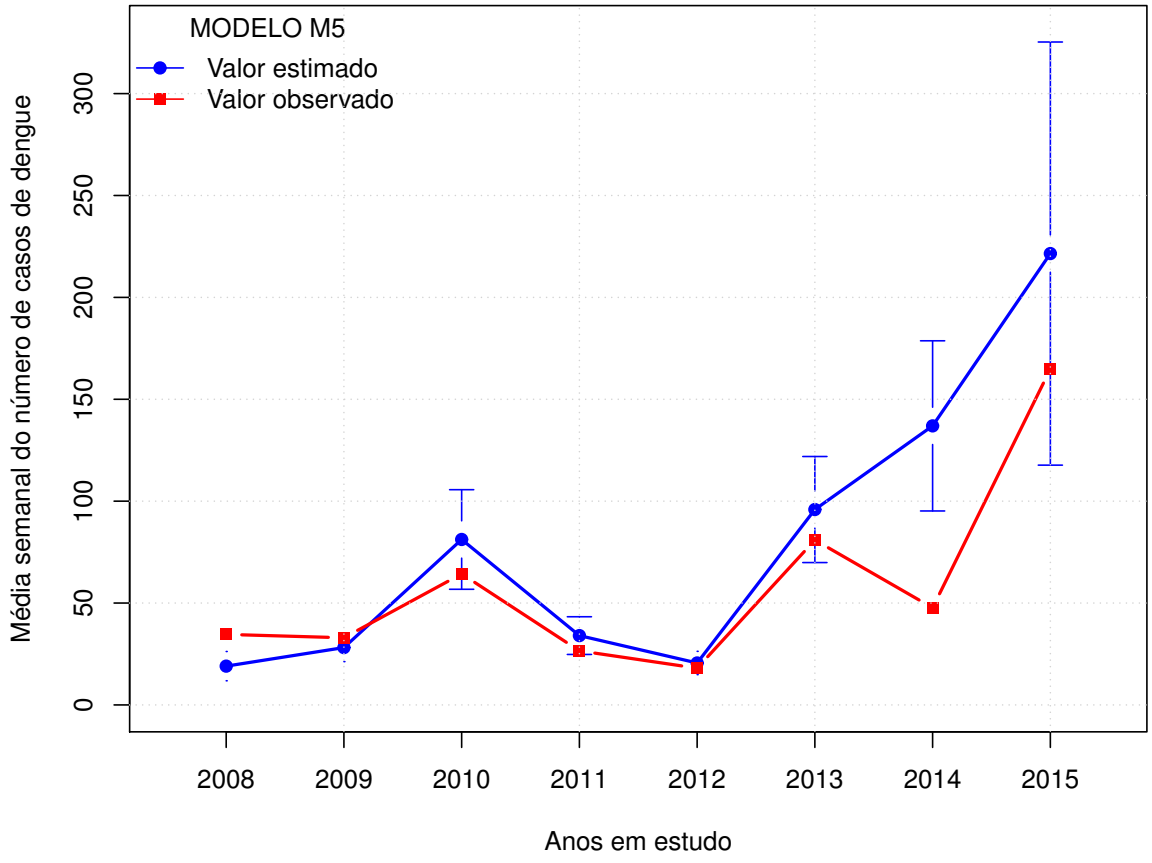


Figura 4.7: Dengue média semanal para os anos de 2008 a 2015, considerando as 20 cidades do estado de Goiás em estudo.

Na Figura 4.8, as médias semanais estão representadas para as cidades em análise. Para uma melhor visualização, a cidade de Goiânia foi omitida no gráfico por apresentar uma média semanal muito superior às demais cidades (690.5). Nota-se que o número de

notificações de casos de dengue observada pertence ao intervalo de confiança de 95 % do número de notificações de casos de dengue ajustado com GLMM. Os resultados apresentados na Figura 4.8 sugerem que o modelo selecionado se ajusta bem aos dados, quando considerados os fatores aleatórios para *cidade* e *ano*, representado as dependências espacial e temporal, respetivamente.

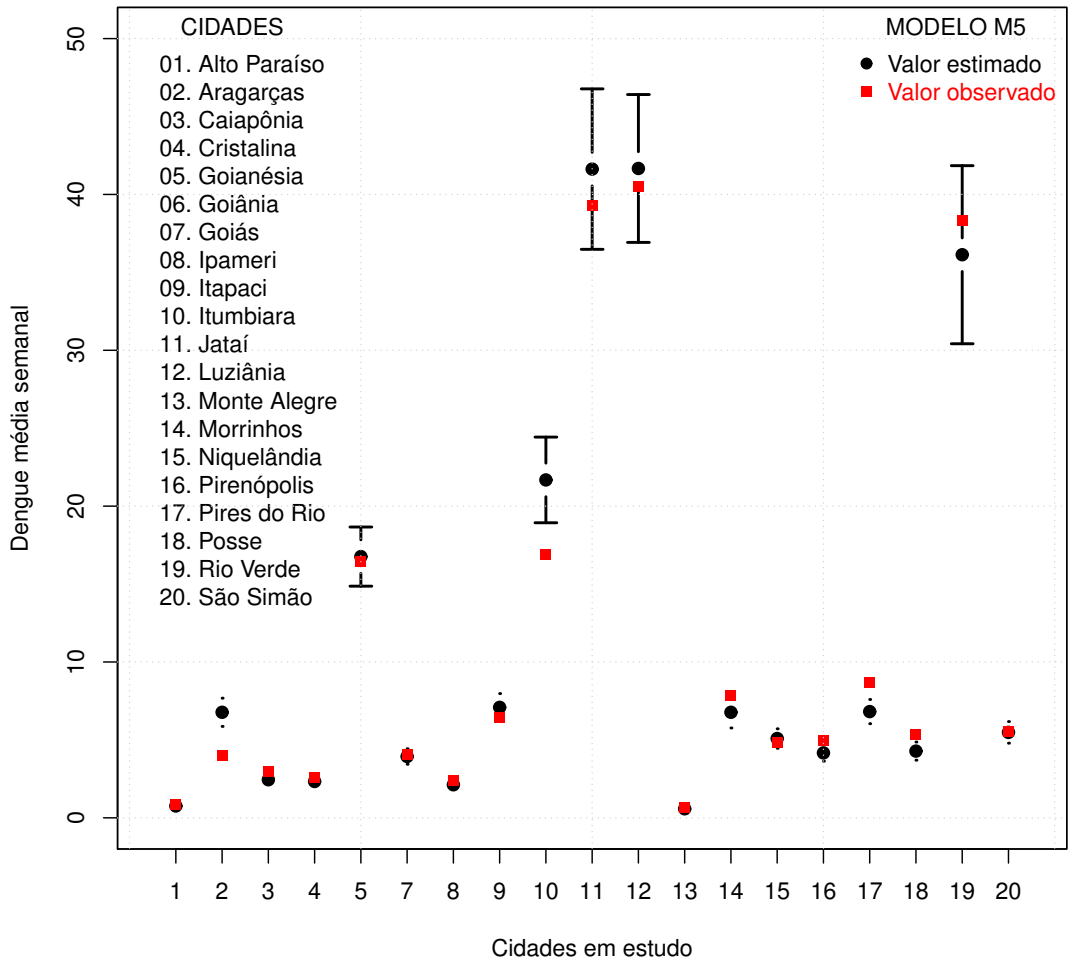


Figura 4.8: Dengue média semanal para 19 cidades do estado de Goiás. A média para a cidade de Goiânia foi omitida no gráfico para permitir uma melhor visualização.

Os testes efetuados sobre as diferentes combinações de um ou dois fatores, associados aos efeitos aleatórios *cidade*, *ano* e *semana*, deveriam assumir diferentes estruturas de correlação. No entanto, os pacotes disponíveis no ambiente R privilegiam a escolha da estrutura de correlação independente, o que no nosso caso nos pareceu aceitável.

O modelo M5, apesar de apresentar bons resultados na estimação do número de notificações de casos de dengue, ainda não permite captar as variações correspondentes aos períodos de epidemia em que temos grande número de notificações de dengue.

Assim, iremos prosseguir com o estudo de um modelo adequado para dados organizados numa hierarquia com estrutura aninhada, introduzindo interações entre os fatores de efeitos aleatórios.

4.4 GLMM incorporando efeitos aleatórios aninhados

Para considerar diferentes origens de variabilidades dos dados, mantendo um modelo simples, iremos então assumir que os fatores de efeitos aleatórios obedecem a uma estrutura hierárquica aninhada, introduzindo interações para criar sub-agrupamentos nos diversos níveis dos efeitos aleatórios, tendo como base o modelo *M5* analisado na Secção 4.3.

Inicialmente, para se ter em conta a variabilidade dos dados associada ao fator semana, iremos considerar uma nova variável explicativa "estação do ano" com quatro níveis (inverno, outono, primavera e verão), adicionada ao modelo como um efeito fixo.

O fator *cidade* representará a variabilidade associada às cidades. A variabilidade associada ao fator *ano* será representada por uma interação entre os fatores *cidade* e *ano*, permitindo efeitos diferentes para cada nível de agrupamento gerado pela interação. Esses efeitos serão modelados como variáveis aleatórias não observadas.

O modelo com interação passa a tratar os efeitos aleatórios numa estrutura hierárquica aninhada com o agrupamento *cidade* representado por 20 níveis e, dentro de cada nível *cidade*, tem-se um nível para cada *ano* conforme mostrado na Figura 4.9 e descrito na sua sequência.

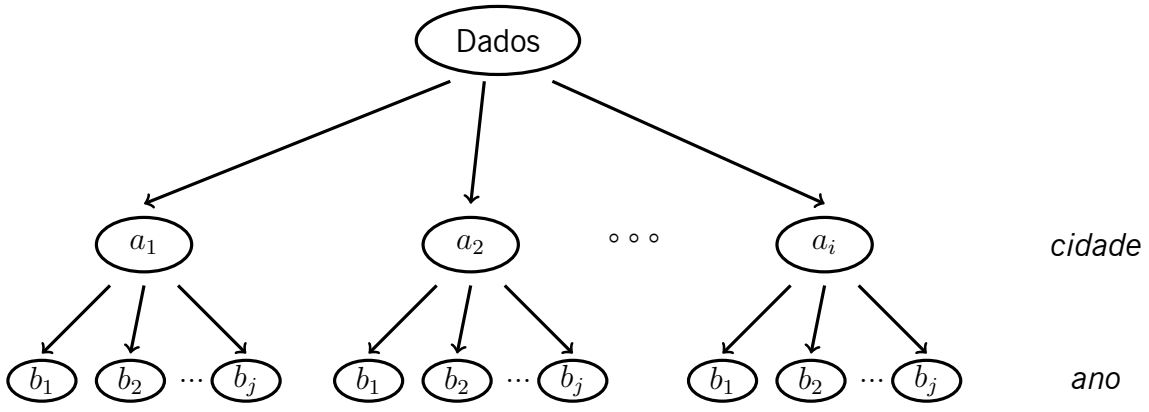


Figura 4.9: Estrutura de dados com agrupamentos hierárquicos aninhados para os grupos *cidade:ano*.

$$Y_{ijs}|a_i, a_i : b_j \sim NegBin(\mu_{ijs}, k)$$

$$E[Y_{ijs}|a_i, a_i : b_j] = \mu_{ijs}$$

$$Var[Y_{ijs}|a_i, a_i : b_j] = \mu_{ijs} + \frac{\mu_{ijs}^2}{k}$$

$$\eta_{ijs} = \beta_0 + \beta_1 prec_{ij(s-6)} + \beta_2 tmin_{ij(s-4)} + \beta_3 tmax_{ijs} + \beta_4 hram_{ij(s-10)} + \beta_5 vvto_{ij(s-2)} + \beta_6 est_{outono} + \beta_7 est_{primavera} + \beta_8 est_{verao} + a_i + a_i : b_j + ln(thab_{ij}/1000)$$

$$ln(\mu_{ijs}) = \eta_{ijs}$$

$$a_i \sim N(0, \sigma_a^2)$$

$$a_i : b_j \sim N(0, \sigma_{a:b}^2)$$

onde:

a) $Y_{ijs}|a_i, a_i : b_j$ indica o número de notificações de casos de dengue na semana s, no ano j, na cidade i, condicionado aos efeitos aleatórios cidade a_i, ano b_j da cidade a_i.

b) a distribuição de Y_{ijs} condicional aos efeitos aleatórios a_i e $a_i : b_j$ é a distribuição Binomial Negativa com média μ_{ijs} e variância $\mu_{ijs} + \frac{\mu_{ijs}^2}{k}$.

c) η_{ijs} é a função de predição linear com o logaritmo natural do total de habitantes ($ln(thab/1000)$) como variável *offset*, para considerar o efeito populacional de cada cidade em estudo.

d) $a_i, a_i : b_j$ indicam a adição dos efeitos aleatórios específicos para a cidade i e para o ano j da cidade i, permitindo a ordenada na origem diferente para cada nível dos efeitos aleatórios.

e) Assume-se que $a_i, a_i : b_j$ são normalmente distribuídos com média zero e variância $\sigma_a^2, \sigma_{a:b}^2$, que determinarão efeitos aleatórios diferentes para cada cidade e para cada ano dentro de uma determinada cidade.

Na Tabela 4.2 tem-se os valores estimados para os parâmetros considerando o modelo GLMM com interações entre os efeitos aleatórios *cidade* e *ano*, numa estrutura hierárquica aninhada para os efeitos aleatórios. Nota-se que a inclusão da interação entre os fatores

aleatórios permitiu um menor valor da estatística *AIC* e manteve a significância estatística ao nível de 5 % para todas as covariáveis, indicando que o modelo se ajusta melhor aos dados, quando comparado com o modelo com os efeitos aleatórios numa estrutura cruzada, analisado na Secção 4.3.

Tabela 4.2: Estimativas dos parâmetros para o modelo GLMM ajustado, assumindo a distribuição Binomial Negativa, incorporando efeitos aleatórios para *cidade* e para a interação *cidade:ano*, e efeitos fixos associados às variáveis meteorológicas e às estações do ano.

PARÂMETROS	**Estimativa**	**Erro Padrão**	**Est. de teste**	**Pr(>|z|)**
Constante	-3.8752	0.4018	-9.6452	< 2e-16
prec (lag 6)	0.0035	0.0005	7.4053	1.31e-13
tmin (lag 4)	0.0853	0.0073	11.6617	< 2e-16
tmax	-0.0580	0.0092	-6.2854	3.27e-10
hram (lag 10)	0.0202	0.0018	11.4572	< 2e-16
vvto (lag 2)	-0.2186	0.0393	-5.5676	2.58e-08
est.outono	1.1020	0.0602	18.2921	< 2e-16
est.primavera	0.1740	0.0651	2.6726	0.0075
est.verão	0.9797	0.0667	14.6778	< 2e-16
AIC	34426.5	-	-	-
σ^2_{cid}	0.4094	-	-	-
$\sigma^2_{cid:ano}$	1.7358	-	-	-
k	1.1238	-	-	-

Na Figura 4.10 apresenta-se o histograma dos resíduos de Pearson com a curva densidade de probabilidade mostrando uma assimetria típica da distribuição Binomial Negativa, com a longa cauda à direita indicando os períodos com grande número de registos de casos de dengue.

Na Figura 4.11, o histograma dos resíduos da desviância e a respetiva curva de densidade de probabilidade indicam que a distribuição dos valores é aproximadamente simétrica e centrada em -0.38, aproximando-se de uma distribuição normal.

Na Figura 4.12 nota-se que os resíduos padronizados do desvio *versus* a ordem de amostragem se distribuem aleatoriamente e não apresentam tendência nem padrão definido no gráfico de dispersão, indicadores da independência dos erros.

Nas Figuras 4.13 e 4.14 temos os resíduos de Pearson *versus* os valores ajustados e os resíduos da desviância *versus* os valores ajustados, respetivamente. Nota-se uma melhor aproximação da linha de suavização com o eixo da linha zero, quando comparados com os resultados do modelo M5 apresentados na Secção 4.3. Observa-se, ainda, uma grande concentração de pontos para os valores ajustados menores que 400, acentuando-se nas

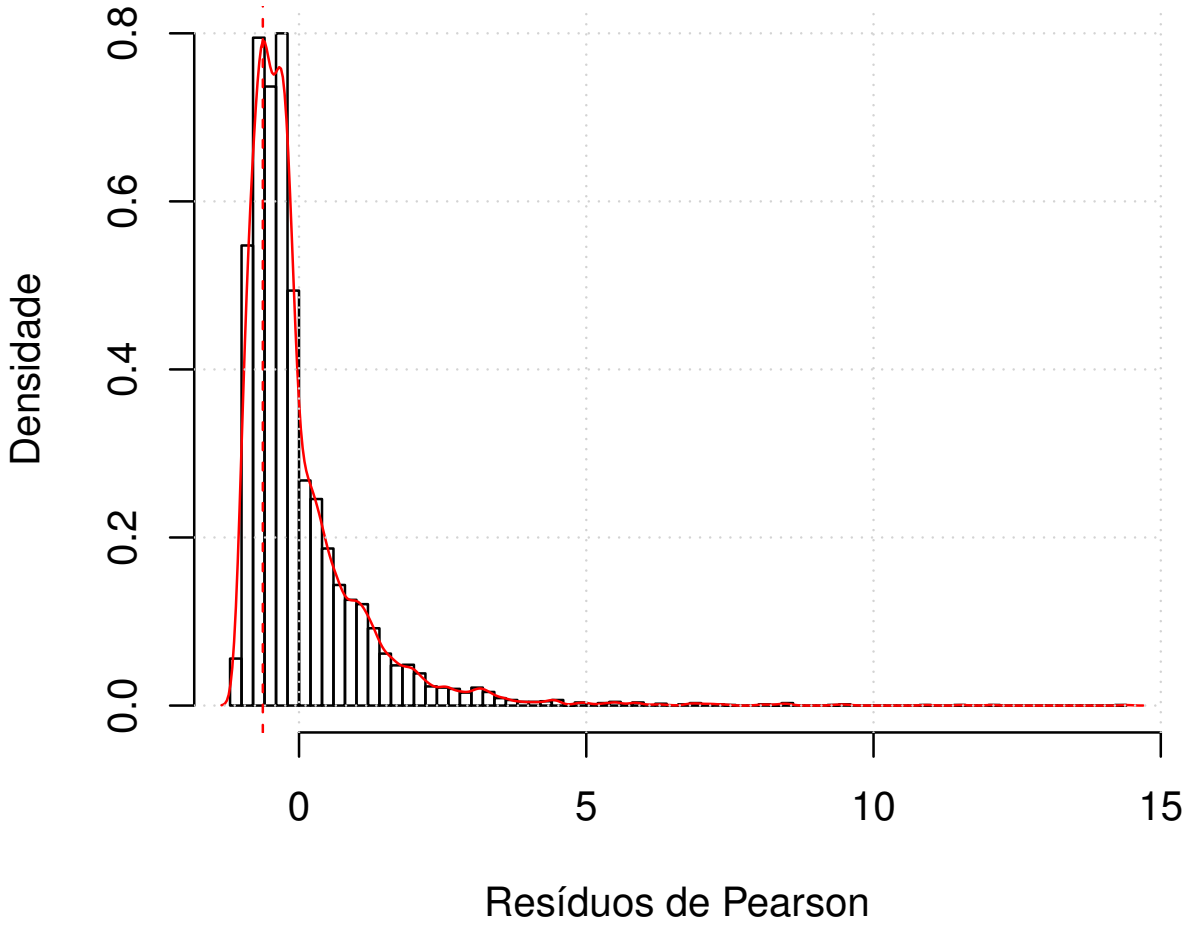


Figura 4.10: Histograma dos resíduos de Pearson com curva assimétrica com cauda pesada à direita caracterizando a distribuição binomial negativa.

proximidades do valor zero, indicando os períodos que não há registos de casos de dengue. Os pontos com valores ajustados maiores que 500 referem-se aos períodos epidémicos.

Na Figura 4.15 observa-se o número médio semanal de notificações de casos de dengue observado e estimado para os anos de 2008 a 2015. Note-se que o valor médio observado pertence ao intervalo de confiança a 95 % calculado para o valor médio estimado, mesmo tendo em conta que os valores estimados são fortemente influenciados pelo número elevado de notificações de casos de dengue registados na cidade de Goiânia. A importância de se considerarem efeitos aleatórios aninhados é evidenciada pelos melhores resultados apresentados na Figura 4.15 *versus* os resultados apresentados na Figura 4.7.

Para validar os pressupostos do modelo GLMM ajustado, iremos recorrer ao pacote *DHARMa* (Hartig, 2017) do *software* R, o qual utiliza uma abordagem baseada em simulações para, a partir de 1000 simulações do modelo ajustado, gerar resíduos padronizados.

Na Figura 4.16 apresenta-se o *QQ-plot* dos resíduos do modelo GLMM indicando um

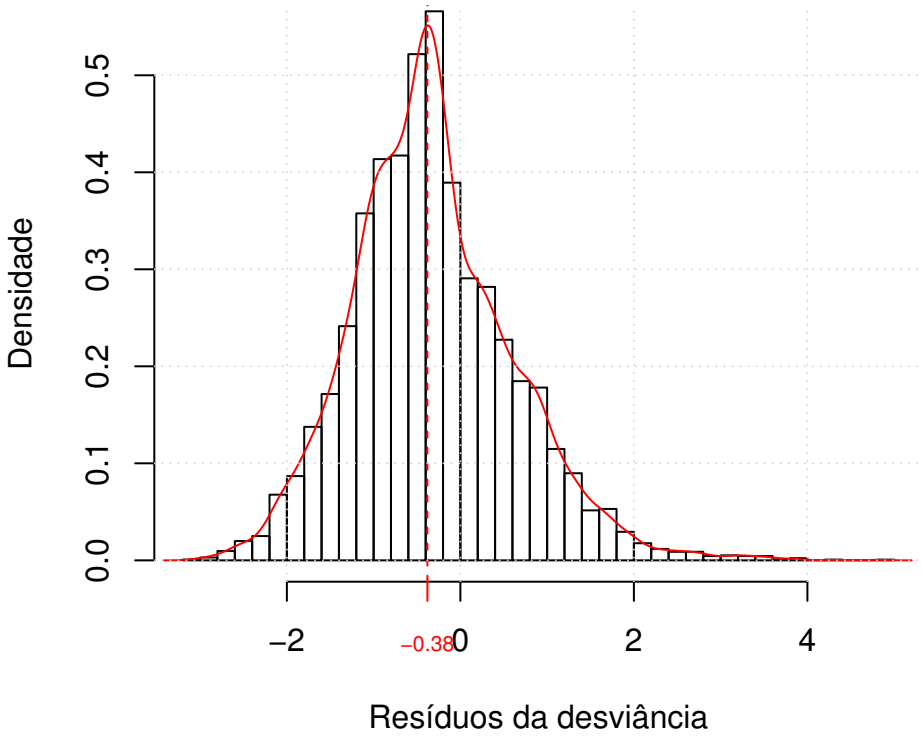


Figura 4.11: Histograma dos resíduos da desviância com curva densidade de probabilidade estimada pelo método de *kernel* para o modelo ajustado com interações entre os efeitos aleatórios.

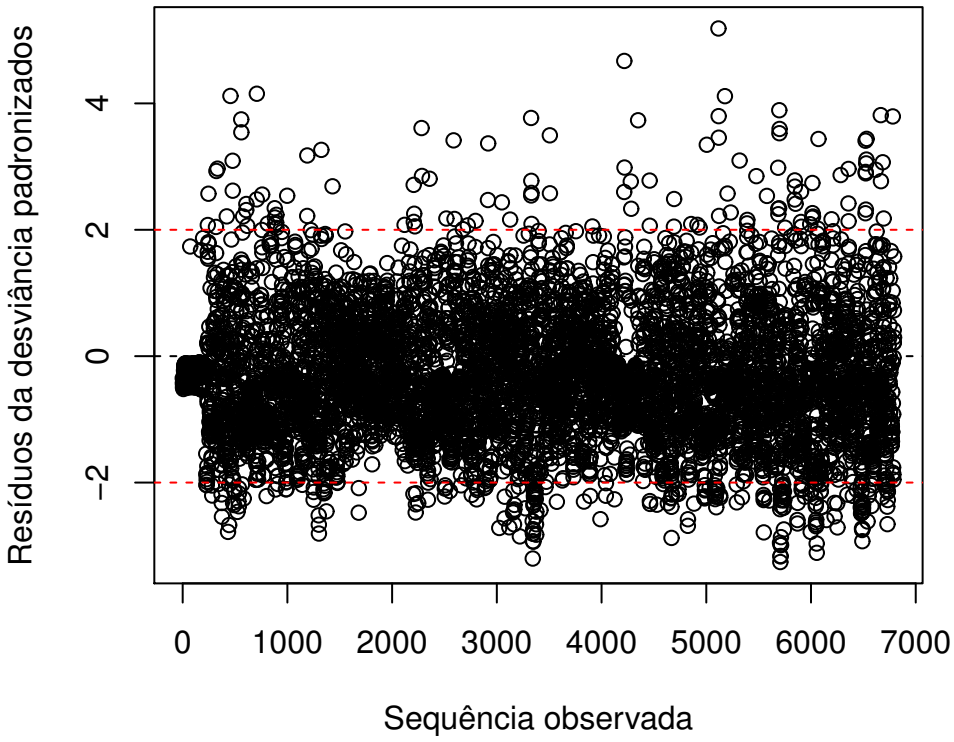


Figura 4.12: Resíduos da desviância padronizados *vs.* sequência observada.

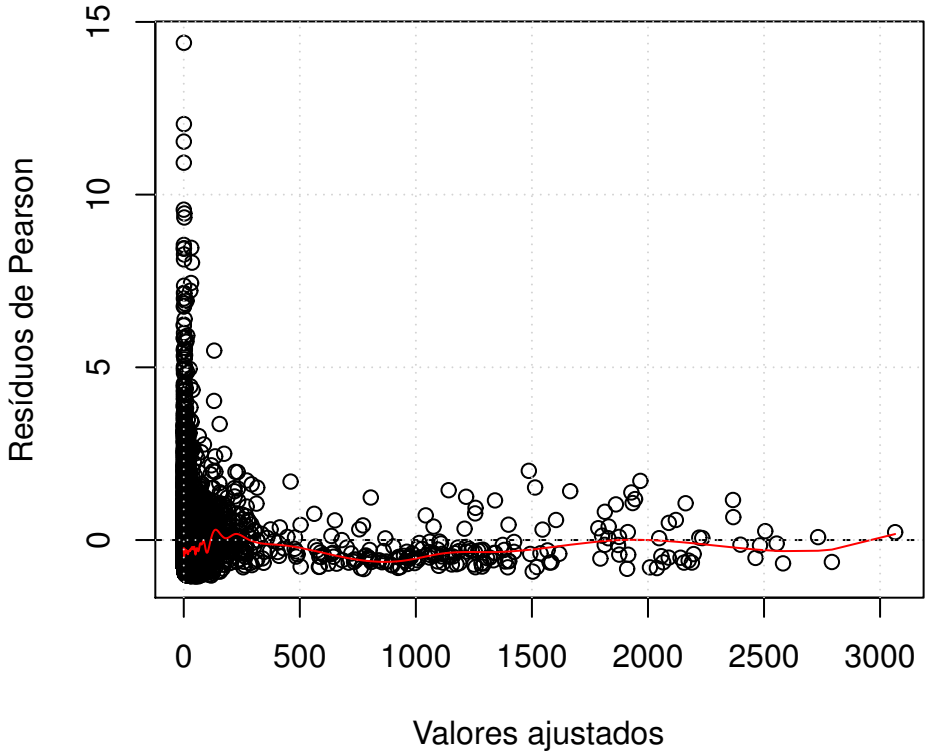


Figura 4.13: Apresentação dos resíduos de Pearson *versus* os valores ajustados para o modelo com interação entre os efeitos aleatórios.

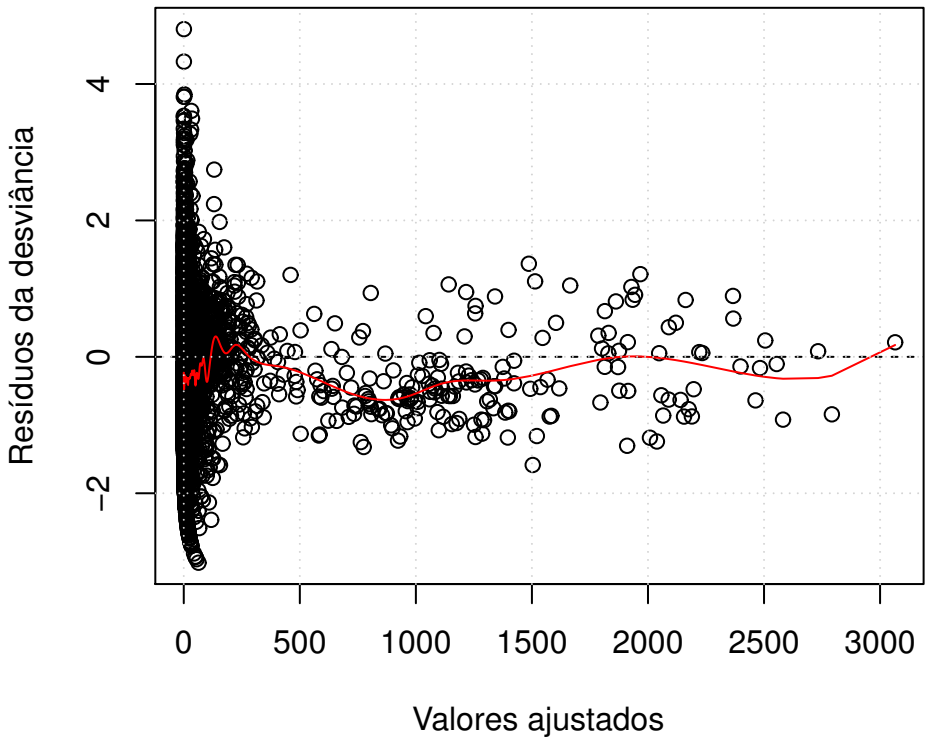


Figura 4.14: Apresentação dos resíduos da desviância *versus* os valores ajustados para o modelo com interação entre os efeitos aleatórios.

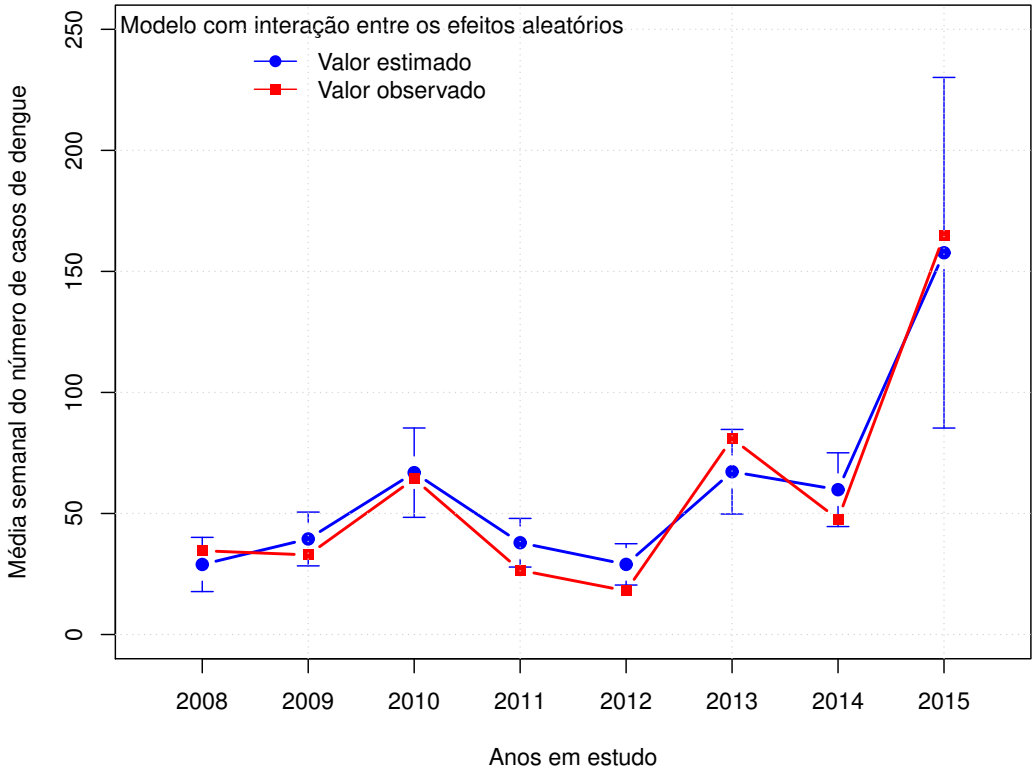


Figura 4.15: Dengue média semanal para os anos de 2008 a 2015, considerando os dados das 20 cidades do estado de Goiás em estudo.

ajuste adequado. O gráfico dos resíduos padronizados *versus* os valores preditos (lado direito) sugere indicações que o modelo foi especificado de forma correta.

Note-se maior concentração dos pontos para os valores preditos menores que 40, equivalentemente à média de registos de casos de notificações de dengue para as cidades em estudo, quando não são considerados os registos para a cidade de Goiânia conforme mostrado na Figura 4.8.

4.4.1 Interpretação dos coeficientes estimados

A partir das estimativas apresentadas na Tabela 4.2, para as variáveis meteorológicas do modelo ajustado que considera a interação entre os efeitos aleatórios *cidade* e *ano* numa estrutura hierárquica aninhada, retiram-se as seguintes principais conclusões sobre a alteração no valor de uma variável preditora, mantendo as restantes variáveis constantes:

- Um aumento de 10 mm na precipitação ocasiona um aumento de 3.5 % na taxa de notificação de casos de dengue, passadas seis semanas.

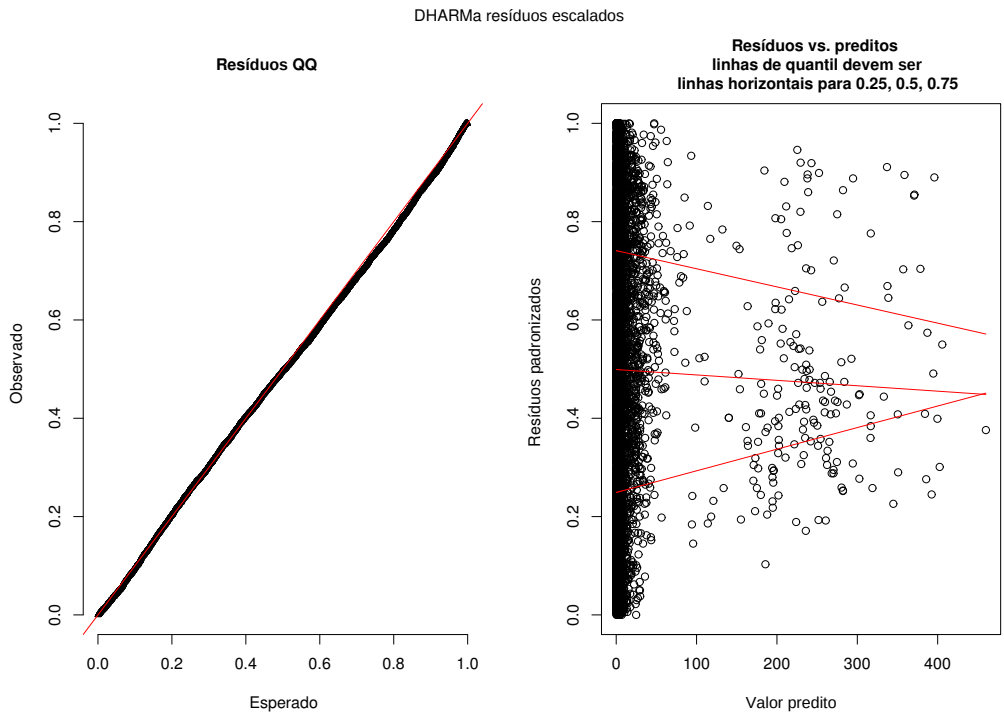


Figura 4.16: Resíduos padronizados obtidos a partir do modelo GLMM ajustado com efeitos aleatórios aninhados.

- Um grau Celsius a mais na temperatura mínima corresponde a um aumento de 8.9 % na taxa de notificação de casos de dengue, passadas quatro semanas.
- Um aumento na temperatura máxima provoca um decréscimo na taxa de notificação de casos de dengue, sendo que a cada grau Celsius corresponde a um decréscimo de 5.64 % na taxa de notificação de casos de dengue registado.
- Um aumento na média da humidade relativa do ar em 1 % provoca um aumento de 2 % na taxa de notificação de casos de dengue, passadas dez semanas.
- Um aumento na velocidade do vento em 1 m/s produz uma diminuição de aproximadamente 20 % na taxa de notificação de casos de dengue, passadas duas semanas.

A taxa de notificação de casos de dengue também é influenciada pelas estações do ano. Tendo o inverno como estação de referência, conclui-se:

- Na primavera há um aumento de 19 % na taxa de notificação de casos de dengue comparado com o inverno.
- No outono, a taxa de notificação de casos de dengue corresponde a cerca de três vezes a taxa de notificação registada no inverno.

- No verão, a taxa de notificação de casos de dengue aumenta cerca de duas vezes em relação ao inverno.

Relativamente às variâncias associadas aos efeitos aleatórios temos moderada variabilidade para o agrupamento cidade ($\widehat{\sigma^2}_{cid} = 0.4094$) e uma alta variabilidade para a interação *cidade:ano* ($\widehat{\sigma^2}_{cid:ano} = 1.7358$), o que reflete a necessidade de se considerar um efeito para cada ano dentro de cada cidade.

O parâmetro de sobredispersão estimado da distribuição Binomial Negativa ($k = 1.1238$) indica que esta distribuição permitiu tratar a alta variabilidade existente nos dados. A alta variabilidade é influenciada pelos dados da cidade de Goiânia e tendo em conta que os anos de 2010, 2013 e 2015 foram considerados epidémicos de acordo com os indicadores de referência definidos pela Organização Mundial de Saúde.

4.5 Considerações finais do capítulo

Neste capítulo foram apresentados conceitos relacionados com a modelação de dados espaço-temporais, recorrendo-se a modelos lineares generalizados mistos, incorporando efeitos fixos e aleatórios, adequados para dados com medições repetidas, como é o caso das notificações de dengue ao longo de oito anos.

Foram analisados modelos para dados agrupados, assumindo uma estrutura hierárquica cruzada ou assumindo uma estrutura hierárquica aninhada, tendo as variáveis meteorológicas e as estações do ano modeladas como efeitos fixos, e os fatores *cidade* e *ano* considerados como efeitos aleatórios.

Conclui-se ainda que as estações do ano apresentam grandes contribuições para o aumento da taxa de notificação de casos de dengue registados no estado de Goiás, com maiores índices no outono e no verão.

As variáveis meteorológicas precipitação, temperatura mínima e média da humidade relativa do ar contribuem positivamente, enquanto a temperatura máxima e a velocidade do vento contribuem negativamente para a taxa de notificação de casos de dengue.

A estrutura hierárquica aninhada, para os fatores associados aos efeitos aleatórios, permitiu a estimação dos coeficientes de regressão com menor erro padrão, possibilitando um valor estimado mais preciso para o número esperado de notificações de casos de dengue em diferentes cidades do estado de Goiás.

Capítulo 5

Análise de significância dos efeitos aleatórios

5.1 Introdução

No Capítulo 4 foram apresentadas metodologias para estimar as componentes de variância associadas aos efeitos aleatórios, recorrendo-se aos modelos lineares generalizados mistos. Estimadas as componentes de variância, há a necessidade de quantificar a precisão de tais estimativas. Os métodos *bootstrap* permitem quantificar incertezas, calculando o erro padrão e os intervalos de confiança, e realizar testes de significância para as componentes de variância.

Neste capítulo analisa-se, por meio de modelação e simulação, a significância das estimativas das componentes de variância associadas aos efeitos aleatórios, aplicando métodos de reamostragem *bootstrap*. Os resultados obtidos nas simulações serão aplicados em cenários com dados reais, referentes ao número de notificações de casos de dengue no estado de Goiás.

Espera-se obter resultados satisfatórios nos cenários com dados reais, fazendo-se a reamostragem de dados em dimensão inferior à dimensão da amostra inicial, permitindo a análise das componentes de variância, pelas estimativas do erro padrão (SE) e dos intervalos de confiança (IC), com menor carga computacional.

5.2 Motivação

Este estudo foi motivado pela necessidade de analisar a relação entre as variáveis meteorológicas e o número de notificações de casos de dengue no estado de Goiás, Brasil. Acredita-se que o número de notificações de casos de dengue está associado aos efeitos aleatórios dos fatores *cidade* e *ano*. Neste sentido, aplicam-se os GLMM e a metodologia *bootstrap* para analisar cenários com dados simulados e cenários com dados reais referentes às notificações de casos de dengue no estado de Goiás.

São utilizados os efeitos aleatórios associados ao fator *cidade* para representar as variações espaciais e os efeitos aleatórios associados ao fator *ano* para representar as variações temporais, tendo em conta que os efeitos fixos estão associados às variáveis meteorológicas. Dessa forma, alterações no número de notificações de casos de dengue não captadas pelo modelo ajustado são associadas a fatores não observados, como as políticas públicas para o combate ao vetor transmissor da dengue, adotadas anualmente em cada cidade.

Estudos similares são encontrados na literatura. Em Johnson *et al.* (2015) é analisada a carga de carrapatos em pintainhos silvestres, aplicados a cenários com fatores aninhados (local, ninhos, pintainhos). O número de localizações necessárias para conseguir uma dada precisão dos parâmetros estimados, foi definido recorrendo à análise de potência em GLMM. Em Thai *et al.* (2013), métodos *bootstrap* aplicados a modelos lineares de efeitos mistos (LMM) são comparados com base no enviesamento relativo dos parâmetros estimados. Num guia prático para estudos em ecologia utilizando GLMM, Bolker *et al.* (2009) concluíram que, ao incorporar os efeitos aleatórios em projetos com respostas de contagens ou proporcionais, os GLMM permitem a generalização das conclusões para tempos, lugares e espécies. Sinha (2009) utiliza a metodologia *bootstrap* para analisar a significância das variâncias em GLMM, e os resultados do estudo de simulações demonstraram que os testes de *bootstrap* têm um nível de significância próximo dos habituais níveis de referências.

5.3 *Bootstrap*

Os modelos lineares de efeitos mistos apresentam grande complexidade na estimação de erros padrão e intervalos de confiança das componentes de variância associadas aos

efeitos aleatórios. Dessa forma, recorre-se à metodologia *bootstrap* para se obter tais estimativas.

Bootstrap (Efron, 1979) é uma técnica estatística introduzida para estimar o erro padrão de estimadores complexos. A técnica de *bootstrap* baseia-se na análise de amostras obtidas a partir de uma distribuição empírica, que substitui a distribuição da população desconhecida, considerando a amostra observada como representativa de toda a população.

Tradicionalmente, o método *bootstrap* consiste na reamostragem da amostra inicial, com reposição, para obter uma amostra *bootstrap* de igual dimensão à amostra original. Isto é, são realizados *n* sorteios com reposição, sendo *n* o número de observações disponíveis na amostra original. A distribuição *bootstrap*, formada pelas estimativas dos parâmetros das amostras *bootstrap*, converge para a distribuição verdadeira quando o número de amostras *bootstrap* (*B*) tende ao infinito.

Dada a amostra $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ formada de *n* observações, selecionam-se *B* amostras *bootstrap* independentes $\boldsymbol{X^{*1}}, \boldsymbol{X^{*2}}, \ldots, \boldsymbol{X^{*B}}$, cada uma de dimensão *n*, identicamente distribuídas, obtida de forma aleatória e com reposição. Quando se pretende estimar o parâmetro desconhecido θ, com base numa amostra $\boldsymbol{X}$, utilizando o estimador $\boldsymbol{S(X)}$, calcula-se uma estimativa de θ dada por $\hat{\theta} = S(x)$. Em cada amostra *bootstrap* $\boldsymbol{X^*}$, uma réplica *bootstrap* da estimativa de $\hat{\theta}$ resulta da aplicação da estatística $S(.)$ à amostra $\boldsymbol{X^*}$ que se representa por $\hat{\theta^*} = S(x^*)$. A partir das B réplicas bootstrap da estimativa de $\hat{\theta}$, $\hat{\theta^*_b} = S(X^{*b})$, $b = 1, \ldots, B$, forma-se a distribuição *bootstrap*, a qual permite a estimação de parâmetros como erro padrão e intervalo de confiança. Na Figura 5.1 apresenta-se um diagrama esquemático do processo de *bootstrap*.

Bootstrap paramétrico

O método *bootstrap* paramétrico é indicado quando se conhece, *à priori*, alguma informação sobre a distribuição dos dados. Este método consiste em utilizar o parâmetro θ (ou vetor de parâmetros), conhecido ou estimado, da distribuição conhecida para gerar B amostras, $\boldsymbol{X^{*1}}, \boldsymbol{X^{*2}}, \ldots, \boldsymbol{X^{*B}}$, baseadas nessa distribuição de probabilidade e, a partir das amostras geradas, obtêm-se B estimativas do parâmetro (ou vetores de parâmetros) $\widehat{\theta}^*_{Par}$, para formar a distribuição *bootstrap* paramétrica.

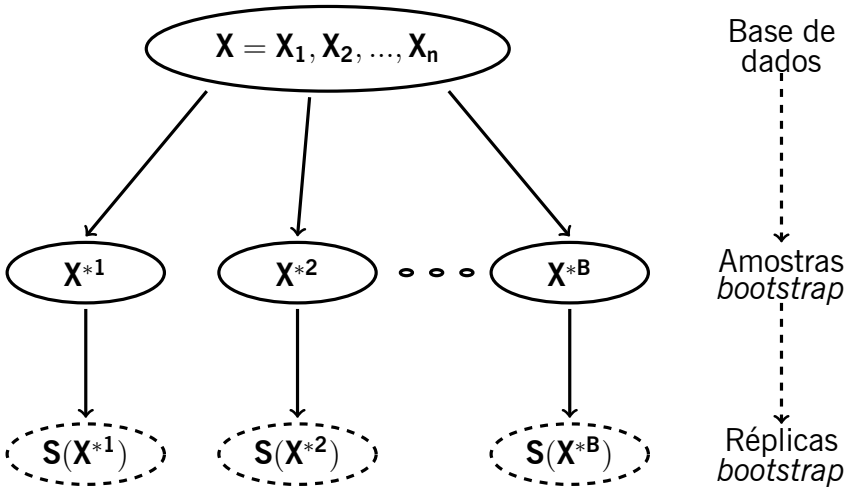


Figura 5.1: Processo esquemático de *bootstrap* para a estimação de erro padrão e intervalo de confiança da estatística S(X), adaptado de *Efron* e *Tibshirani* (1994).

Bootstrap não paramétrico

O método *bootstrap* não paramétrico é utilizado quando não se conhece ou não se fazem suposições sobre a distribuição dos dados (Efron and Tibshirani, 1994). Tradicionalmente, a partir de uma amostra de dimensão n, obtida da população, são selecionadas B amostras *bootstrap* de dimensão n com reposição. Das amostras *bootstrap* $\boldsymbol{X^{*1}}, \boldsymbol{X^{*2}}, \ldots, \boldsymbol{X^{*B}}$ obtêm-se B estimativas do parâmetro (ou vetores de parâmetros) $\widehat{\theta}^*_{NPar}$ para aproximar a distribuição *bootstrap* não paramétrica.

Estimação *bootstrap* de parâmetros

A estimativa *bootstrap* do erro padrão é dada pelo desvio padrão das réplicas *bootstrap*, sendo:

$$\widehat{SE}_B = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^*_b - \tilde{\theta}_B)^2} \tag{5.1}$$

onde B é o número de amostras *bootstrap*, $\hat{\theta}^*_b$ é a estimativa do parâmetro para a *b-ésima* amostra *bootstrap* e $\tilde{\theta}_B$ é a mediana da distribuição *bootstrap*.

O intervalo de confiança (IC) para o parâmetro θ pode ser construído considerando a distribuição dos estimadores assintoticamente normal. Dado o quantil $1-\alpha/2$ da distribuição normal padrão $z_{1-\alpha/2}$, o IC será calculado por:

$$\left[\tilde{\theta}_B - \widehat{SE}_B \times z_{1-\alpha/2}, \tilde{\theta}_B + \widehat{SE}_B \times z_{1-\alpha/2}\right]$$

No entanto, a distribuição dos estimadores aproxima-se da distribuição normal apenas quando a dimensão da amostra é elevada, o que pode levar nos outros casos a um grande viés na estimação do parâmetro θ. Para uma maior precisão na estimação do intervalo de confiança *bootstrap*, utiliza-se o método dos percentis que apresenta resultados com menores viés e é dado por:

$$\left[\widehat{\theta}^*_{(\frac{\alpha}{2}\times B)}, \widehat{\theta}^*_{(1-\frac{\alpha}{2})\times B}\right] \tag{5.2}$$

onde $\widehat{\theta}^*_{(\frac{\alpha}{2}\times B)}$ é o percentil de ordem $100\times\alpha/2$ dos valores de $\widehat{\theta}^*_b$, ou seja, é o $[B \times \alpha/2]$-*ésimo* valor das B réplicas *bootstrap* da estimativa de $\hat{\theta}$ ordenadas por ordem crescente, e $\widehat{\theta}^*_{(1-\frac{\alpha}{2})\times B}$ é o percentil de ordem $100 \times (1-\alpha/2)$% dos valores de $\hat{\theta}^*_b$. O nível de confiança mais utilizado é o que resulta de se considerar $\alpha = 5$ %. No entanto, outros valores podem ser utilizados.

Essa abordagem será aplicada aos dados de notificações de casos de dengue, com a finalidade de avaliar a precisão das estimativas das componentes de variância associadas aos efeitos aleatórios, quando se faz variar o número de elementos reamostrados por grupo. À medida que se aumenta a dimensão da amostra, os parâmetros estimados tornam-se mais precisos, pois a amostra representa melhor a população (Chernick, 2011). Espera-se, com os dados das notificações de casos de dengue, identificar o número mínimo de elementos reamostrados por grupo que permita obter resultados satisfatórios e diminuir a carga computacional associada à estimação dos parâmetros.

5.4 Estudo de simulação

Para analisar distintos cenários aplicando a metodologia *bootstrap* na estimação de erros padrão e intervalos de confiança das componentes de variância associadas aos efeitos aleatórios, será realizado um estudo de simulação utilizando o *software* R (R Core Team, 2016) e os pacotes *GLMMmisc* (Johnson, 2015) e *lme4* (Bates et al., 2015), que são apropriados para este estudo.

Estudos de simulações permitem reproduzir cenários da vida real e, no nosso estudo, serão aplicados para estimar as componentes de variância associadas aos efeitos aleatórios

do fator *cidade*, para representar a influência espacial, e as componentes de variância associadas ao fator *ano*, para representar a influência temporal. Desta forma, cada base de dados simulada será estruturada para representar o número de notificações de casos de dengue de 20 cidades do estado de Goiás, durante um período de 12 anos, totalizando 240 amostras, agrupadas por *cidade* e *ano*.

Será avaliado o comportamento dos efeitos aleatórios sub-amostrando a base de dados para os grupos *cidade* e *ano*. Para gerar as amostras para o *bootstrap* paramétrico será utilizada a função *sim.glmm* desenvolvida por Johnson (Johnson, 2015), a qual gera a variável resposta a partir de um conjunto de parâmetros θ de distribuição conhecida. Caso θ seja desconhecido, então pode ser estimado com base na amostra original.

Neste trabalho utiliza-se como vetor de parâmetros verdadeiros, o vetor dos parâmetros estimados no Capítulo 4, a partir da base de dados apresentada no Capítulo 2, que representa 20 cidades do estado de Goiás. Como cada cidade apresenta um total de habitantes diferente para os anos em estudo, a variável resposta será a taxa de notificações de casos de dengue calculada pela razão do número de notificações de casos de dengue e o respetivo total de habitantes da cidade no referido ano, dado em milhares de habitantes. Assim, o total de habitantes de cada cidade será indicado no preditor linear por uma variável *offset* definida pelo logaritmo natural do total de habitantes $(thab)$ dividido por 1000. O modelo a simular é definido por

$$Y_{ij}|a_i, b_j \sim Poisson(\mu_{ij})$$

$$Var[Y_{ij}|a_i, b_j] = E[Y_{ij}|a_i, b_j] = \mu_{ij}$$

$$ln(\mu_{ij}) = \eta_{ij} = \beta_0 + ln(thab_{ij}/1000) + a_i + b_j$$

$$a_i \sim N(0, \sigma^2_{cid})$$

$$b_j \sim N(0, \sigma^2_{ano})$$

onde

Y_{ij} indica o número de notificações de casos de dengue no ano $j = 1, ..., 12$, na cidade $i = 1, ..., 20$.

$Y_{ij}|a_i, b_j$ segue uma distribuição de Poisson com parâmetro μ_{ij}.

η_{ij} é o preditor linear.

a_i é o efeito aleatório associado à cidade i e b_j é o efeito aleatório associado ao ano j.

Os efeitos aleatórios a_i, $i = 1, ..., 20$, são independentes entre si com uma distribuição normal de média zero e variância σ^2_{cid}. Os efeitos aleatórios b_j, $j = 1, ..., 12$, são independentes entre si com uma distribuição normal de média zero e variância σ^2_{ano}.

Para representar a população, utiliza-se uma base de dados de referência gerada pela função *sim.glmm*, utilizando o vetor de parâmetros θ definido com base nos valores estimados no Capítulo 4, a partir dos dados reais. Outros valores iniciais para o vetor de parâmetros foram testados (resultados não apresentados) para auxiliar a análise da probabilidade de cobertura *bootstrap*.

Os cenários simulados contemplam modelos com dois efeitos aleatórios (*cidade* e *ano*). Estimam-se as probabilidades de cobertura *bootstrap* para as duas componentes de variância (σ^2_{cid} e σ^2_{ano}), fazendo a reamostragem nos dois grupos *cidade* e *ano*, simultaneamente e com reposição, em dimensões que possam ser representativas da população em estudo.

A avaliação da precisão das estimativas das componentes de variância associadas aos efeitos aleatórios será realizada utilizando a raiz quadrada do erro quadrático médio (RMSE) e o erro absoluto médio (MAE), estimados a partir das R medianas obtidas nos cenários simulados. Isso, por se tratar de dados de contagem com sobredispersão, que são melhores representados pela mediana que é uma medida de localização mais robusta. Valores mais baixos de RMSE e MAE indicam melhor precisão debaixo do cenário simulado.

Seja *R* o número de simulações em cada cenário, θ_{TRUE} o valor verdadeiro do parâmetro e $\tilde{\theta_r}$ o valor da mediana na *r-ésima* simulação. O $RMSE$ e o MAE são obtidos por:

$$RMSE = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\theta_{TRUE} - \tilde{\theta}_r)^2} \tag{5.3}$$

$$MAE = \frac{1}{R}\sum_{r=1}^{R}|\theta_{TRUE} - \tilde{\theta}_r| \tag{5.4}$$

Para analisar a metodologia *bootstrap* proposta nos cenários simulados, calcula-se a probabilidade de cobertura *bootstrap* dos parâmetros estimados em cada cenário simulado, a qual indica a percentagem de vezes que o valor verdadeiro do parâmetro pertence ao

intervalo de confiança e é obtida por:

$$pcBoot = \frac{1}{R}\sum_{r=1}^{R} I_{(\theta_{TRUE} \geq IC_r^{inf} \quad \& \quad \theta_{TRUE} \leq IC_r^{sup})} \times 100 \tag{5.5}$$

onde I_A identifica a função indicatriz, assumindo o valor 1 se A é verdadeiro, e o valor 0 caso contrário. IC_r^{inf} e IC_r^{sup} são os limites inferior e superior do intervalo de confiança do parâmetro estimado na simulação r e R o número de simulações de cada cenário.

Tendo em conta que os dados de notificações de casos de dengue apresentam influências espaciais e/ou temporais, torna-se necessário indicar uma proposta adequada para a geração das amostras *bootstrap* $\boldsymbol{X}^{*b}$, $b = 1, \ldots, B$. Esta proposta passa por recorrer às técnicas de subamostragem, onde se extraem amostras de dimensões menores que a amostra original, por forma a garantir a representatividade de cada tipo distinto do grupo.

Na Tabela 5.1 apresentam-se os valores dos parâmetros utilizados nas simulações.

Tabela 5.1: Valores iniciais para estruturar os cenários simulados.

Parâmetro	**Valor**	**Definição**
nCid	20	número de cidades
nAno	12	número de anos
R	500	número de simulações de cada cenário
B	500	número de réplicas *bootstrap*
β_0	1.61	ordenada na origem
σ^2_{cid}	0.4	variância para o grupo cidade
σ^2_{ano}	0.4	variância para o grupo ano
nRepCid	8, 10, 14, 20	número de cidades reamostradas
nRepAno	4, 6, 8, 12	número de anos reamostrados
θ_{TRUE}	$\beta_0, \sigma^2_{cid}, \sigma^2_{ano}$	vetor de parâmetros verdadeiros

O estudo de simulação consiste nos seguintes passos:

Passo 1: Repetir de $r = 1$ até R, os passos 2 até 5.1.

Passo 2: A partir de θ_{TRUE}, que assumirá um dos valores iniciais apresentados na Tabela 5.1, para os parâmetros $\beta_0, \sigma^2_{cid}, \sigma^2_{ano}$, simular a base de dados de referência ($dREF$) do modelo Poisson, incluindo o logaritmo natural do total de habitantes das cidades goianas, dividido por 1000.

Passo 3: Estimar o parâmetro de referência $\hat{\theta}_{REF}$, que assumirá cada um dos valores para $\hat{\beta}_0, \hat{\sigma}^2_{cid}, \hat{\sigma}^2_{ano}$.

Passo 4: *Bootstrap* paramétrico: para o parâmetro $\hat{\theta}_{REF}$, estimado no passo 3., gerar B amostras *bootstrap* $\boldsymbol{X}^{*\mathbf{1}}, \boldsymbol{X}^{*\mathbf{2}}, \ldots, \boldsymbol{X}^{*\boldsymbol{B}}$, conforme explicado na Secção 5.3, e estimar os B valores $\hat{\theta}^{*b}_{Par}$ para $\hat{\beta}^b_0, \hat{\sigma}^{2b}_{cid}, \hat{\sigma}^{2b}_{ano}$, $b = 1, \ldots, B$, para se obter a distribuição *bootstrap* desse parâmetro.

Passo 4.1: Calcular o intervalo de confiança dos parâmetros, utilizando o método dos percentis, para se obter IC^{inf}_r e IC^{sup}_r.

Passo 5: *Bootstrap* não paramétrico: para a base de dados $dREF$, simulada no passo 2., gerar B amostras *bootstrap* $\boldsymbol{X}^{*\mathbf{1}}, \boldsymbol{X}^{*\mathbf{2}}, \ldots, \boldsymbol{X}^{*\boldsymbol{B}}$, conforme explicado na Secção 5.3 e estimar os B valores do parâmetros $\hat{\theta}^{*b}_{NPar}$, respetivamente $\hat{\beta}^b_0$, $\hat{\sigma}^{2b}_{cid}$, $\hat{\sigma}^{2b}_{ano}$, $b = 1, \ldots, B$, necessários para se formar a distribuição *bootstrap*.

Passo 5.1: Calcular o intervalo de confiança dos parâmetros, utilizando o método dos percentis, para se obter IC^{inf}_r e IC^{sup}_r.

Passo 6: Determinar a probabilidade de cobertura *bootstrap*, recorrendo à Equação (5.5), quer para o *bootstrap* paramétrico quer para o *bootstrap* não paramétrico.

Para gerar a amostra $\boldsymbol{X}^{*\boldsymbol{b}}$ do passo 5, no caso do *bootstrap* não paramétrico, recorreu-se ao método de reamostragem com reposição para o grupo *cidade* e para o grupo *ano*, conforme os valores indicados na Tabela 5.1, para *nRepCid* e *nRepAno*, respetivamente.

5.4.1 Resultados do estudo de simulação

Na Tabela 5.2 são apresentados o RMSE e o MAE para a ordenada na origem β_0 e para as componentes de variância σ^2_{cid} e σ^2_{ano}, estimados pelas Equações (5.3) e (5.4), respetivamente, aplicando o método *bootstrap* paramétrico. Observa-se que os valores do RMSE e do MAE na estimação de σ^2_{cid} são menores que os respetivos valores na estimação de σ^2_{ano}. Isso se deve ao facto do grupo *cidade* apresentar um maior número de elementos, proporcionando uma estimativa mais precisa.

Na Tabela 5.3 são apresentados os resultados do RMSE e do MAE para os cenários simulados com os parâmetros apresentados na Tabela 5.1. Os valores foram estimados utilizando as Equações (5.3) e (5.4), respetivamente, e aplicando o método *bootstrap* não paramétrico. Observa-se que:

Tabela 5.2: Erro médio absoluto e raiz quadrada do erro quadrático médio obtidos utilizando o método *bootstrap* paramétrico.

	β_0	σ^2_{cid}	σ^2_{ano}
MAE	0.19363	0.09995	0.12862
RMSE	0.24311	0.12612	0.15548

- Os valores obtidos nas estimativas do RMSE e do MAE utilizando o método *bootstrap* não paramétrico são semelhantes aos valores obtidos pelo método *bootstrap* paramétrico, no caso em que são reamostrados 8 anos e 8 ou 10 cidades (verTabela 5.2).

- No caso do método de reamostragem não paramétrico, os valores do RMSE e do MAE variam de acordo com as dimensões dos grupos amostrados.

- Quando os dois grupos são reamostrados simultaneamente, obtém-se bons resultados fazendo uma reamostragem na ordem dos 50 % em cada grupo em estudo.

Tabela 5.3: Erro médio absoluto e raiz quadrada do erro quadrático médio para os cenários simulados aplicando o método *bootstrap* não paramétrico.

nRepAno	**nRepCid**	**MAE**			**RMSE**		
		β_0	σ^2_{cid}	σ^2_{ano}	β_0	σ^2_{cid}	σ^2_{ano}
4	8	0.19257	0.10342	0.15436	0.24289	0.12687	0.17666
	10	0.19267	0.10291	0.15567	0.24250	0.12753	0.17756
	14	0.19249	0.10130	0.15920	0.24254	0.12991	0.18080
	20	0.19403	0.10460	0.15988	0.24318	0.13543	0.18122
6	8	0.19177	0.10375	0.13599	0.24165	0.12739	0.16010
	10	0.19342	0.10117	0.13608	0.24243	0.12589	0.15989
	14	0.19349	0.10198	0.13764	0.24332	0.12983	0.16105
	20	0.19260	0.10448	0.13865	0.24246	0.13578	0.16193
8	8	0.19287	0.10438	0.12904	0.24355	0.12766	0.15788
	10	0.19220	0.10120	0.12938	0.24177	0.12587	0.15689
	14	0.19062	0.10205	0.13122	0.24018	0.13101	0.15922
	20	0.19094	0.10458	0.13115	0.24057	0.13557	0.15875
12	8	0.19265	0.10418	0.13162	0.24285	0.12686	0.16606
	10	0.19146	0.10138	0.13082	0.24096	0.12542	0.16463
	14	0.19161	0.09972	0.13064	0.24209	0.12792	0.16390
	20	0.19129	0.10350	0.13093	0.24129	0.13450	0.16414

A probabilidade de cobertura foi calculada tendo em conta cada cenário repetido 250 mil vezes (500 repetições vezes 500 amostras *bootstrap*), totalizando quatro milhões de execuções.

Na Figura 5.2 representam-se as probabilidades de cobertura *bootstrap* para a variância associada ao fator *cidade*. Observa-se que ao aplicar o método *bootstrap* não paramétrico com a reamostragem de 50 % ou menos do total de cidades, obtém-se uma probabilidade de cobertura superior a 90 %, independente do número de anos reamostrados. Ao aumentar o número de cidades reamostradas, a probabilidade de cobertura diminui. Aplicando o método *bootstrap* paramétrico, obtém-se uma probabilidade de cobertura de 90 %.

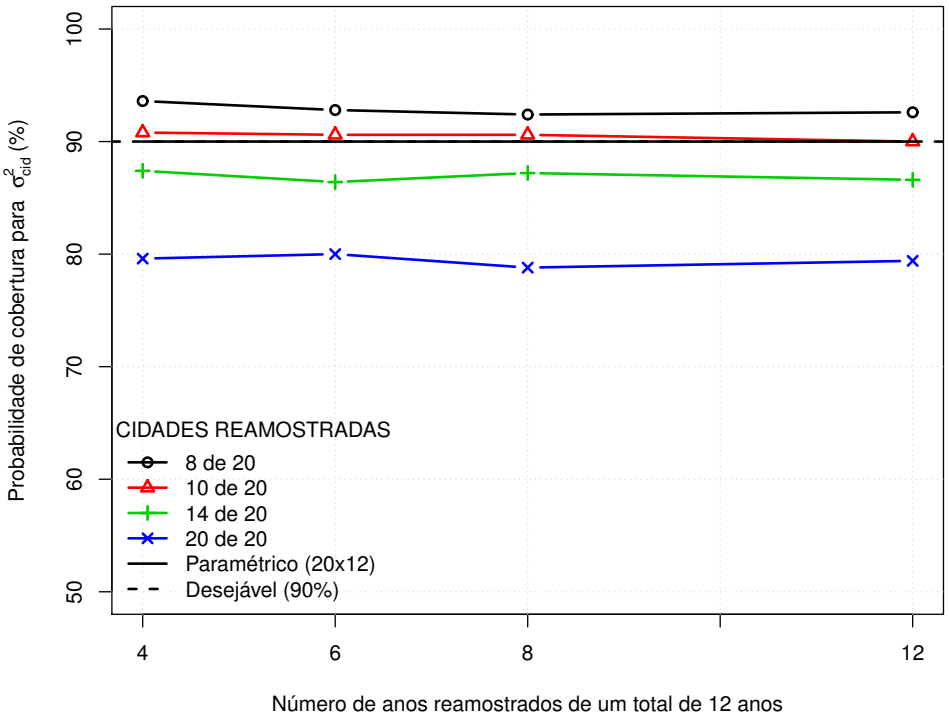


Figura 5.2: Probabilidade de cobertura para a variância associada ao efeito aleatório *cidade* debaixo do *bootstrap* paramétrico e distintos cenários do *bootstrap* não paramétrico.

Na Figura 5.3 apresentam-se as probabilidades de cobertura *bootstrap* da variância associada ao fator *ano*. Nota-se que ao aumentar o número de anos reamostrados, a probabilidade de cobertura diminui, enquanto que ao aumentar o número de cidades reamostradas não influência, significativamente, a probabilidade de cobertura *bootstrap* da variância associada ao fator *ano*. A probabilidade de cobertura, aplicando *bootstrap* paramétrico, apresenta melhor resultado comparado com os resultados obtidos com o método não paramétrico quando a reamostragem do fator *ano* é superior a 50 % do total de anos. Ao aplicar o *bootstrap* não paramétrico e ao reamostrar todos os 12 anos, obteve-se uma probabilidade de cobertura inferior a 80 %, enquanto que para se obter uma probabilidade de cobertura superior a 80 % necessitam-se reamostrar 50 % ou menos do total de anos.

Relativamente à ordenada na origem (β_0), a probabilidade de cobertura *bootstrap* é inferior a 90 % somente na reamostragem dos fatores *ano* e *cidade* iguais ou superiores a 50 % do total de cada fator. Nos restantes casos, a probabilidade de cobertura *bootstrap* permanece acima dos 90 %, sendo que esta probabilidade sob o método *bootstrap* não paramétrico apresenta resultados superiores ao método *bootstrap* paramétrico, conforme apresentado na Figura 5.4.

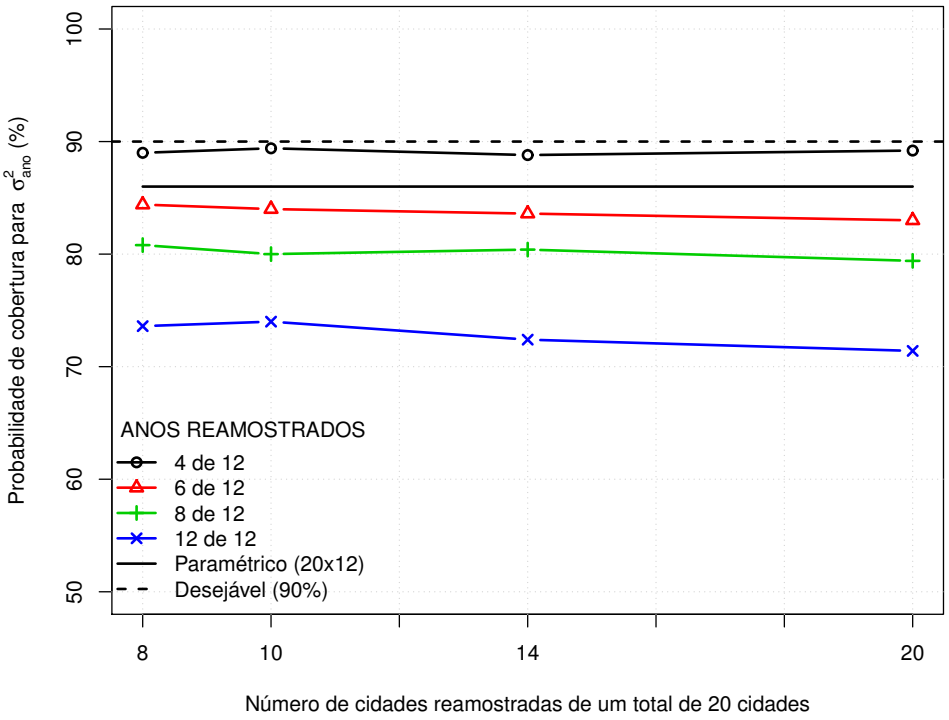


Figura 5.3: Probabilidade de cobertura para a variância associada ao efeito aleatório do *ano* debaixo do *bootstrap* paramétrico e distintos cenários do *bootstrap* não paramétrico.

A partir dos resultados obtidos nas simulações, conclui-se que aplicar o método *bootstrap* não paramétrico, fazendo uma reamostragem na ordem dos 50% do total do fator em estudo, permite obter probabilidades de coberturas superiores a 90% para as componentes de variância associadas aos efeitos aleatórios. Esta percentagem de 50% será aplicado em cenários com dados reais.

5.5 Aplicação dos métodos *bootstrap* em dados reais

Para aplicar os métodos *bootstrap* aos dados reais, utilizam-se os dados que representam o número de notificações dos casos de dengue em 16 cidades do estado de Goiás,

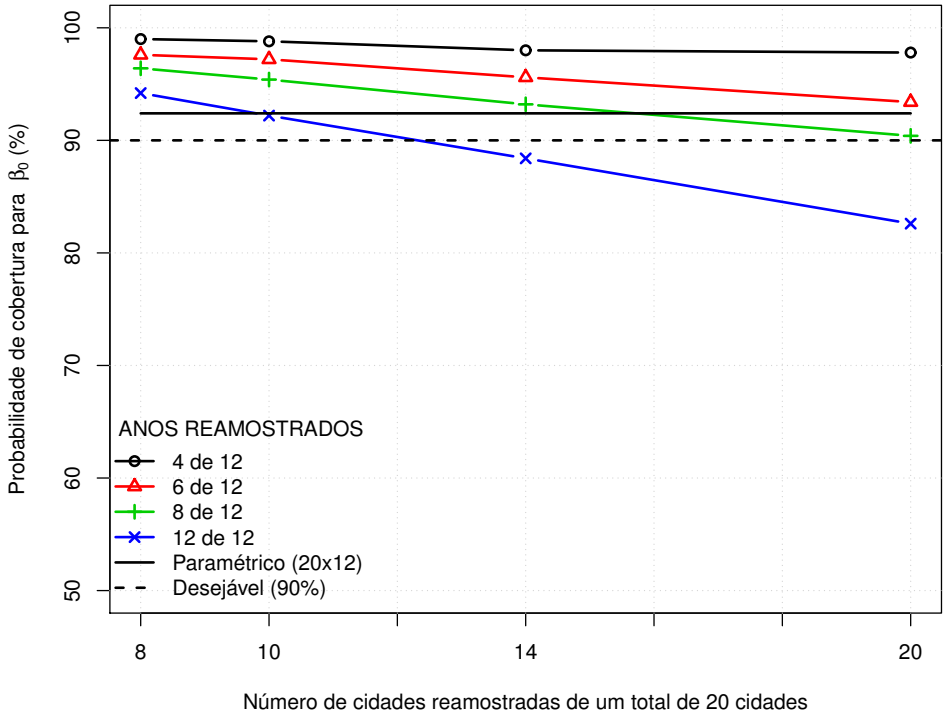


Figura 5.4: Probabilidade de cobertura *bootstrap* para a ordenada na origem debaixo do *bootstrap* paramétrico e distintos cenários do *bootstrap* não paramétrico.

para o período compreendido entre janeiro de 2008 e dezembro de 2014. Os efeitos fixos são representados pelas variáveis meteorológicas e os efeitos aleatórios são representados pelos fatores *cidade* e *ano*.

A base de dados utilizada é um subconjunto da base de dados apresentada no Capítulo 2, da qual foram suprimidas as quatro cidades com maior número de habitantes (Goiânia, Itumbiara, Luziânia e Rio Verde) de modo a nos restringirmos a um grupo de cidades mais homogêneo. Adicionalmente, o ano de 2015 foi também desconsiderado, por não apresentar informações para todas as semanas.

Estimam-se os intervalos de confiança e o erros padrão das componentes de variância associadas aos efeitos aleatórios dos fatores *cidade* e *ano* (σ^2_{cid} e σ^2_{ano}), utilizando os métodos *bootstrap* paramétrico e não paramétrico. As alterações climáticas registadas no estado de Goiás são representadas pelas variáveis meteorológicas, com os seus respetivos desfasamentos que melhor explicam as variações registadas para o número de notificações dos casos de dengue. Conforme analisado no Capítulo 4, o modelo é dado por:

$$Y_{ijs}|a_i, b_j \sim Poisson(\mu_{ijs}|a_i, b_j) \tag{5.6}$$

$$Var[Y_{ijs}|a_i, b_j] = E[Y_{ijs}|a_i, b_j] = \mu_{ijs}$$

$$ln(\mu_{ijs}) = \eta_{ijs} = \beta_0 + \beta_1 prec_{ij(s-6)} + \beta_2 tmin_{ij(s-4)} + \beta_3 tmax_{ijs} + \beta_4 hram_{ij(s-10)} + \beta_5 vvto_{ij(s-2)} + ln(thab_{ij}/1000) + a_i + b_j$$

$$a_i \sim N(0, \sigma_a^2)$$

$$b_j \sim N(0, \sigma_b^2)$$

onde

Y_{ijs} indica o número de notificações de casos de dengue registado na semana $s = 1, \ldots, 53$, do ano $j = 1, \ldots, 7$, na cidade $i = 1, \ldots, 16$.

$Y_{ijs}|a_i, b_j$ segue uma distribuição de Poisson com parâmetro μ_{ijs}.

η_{ijs} é o preditor linear.

β_0 é a ordenada na origem e $\beta_1, \ldots, \beta_5$ são os coeficientes das variáveis meteorológicas.

a_i é o efeito aleatório associado à cidade i e b_j é o efeito aleatório associado ao ano j.

Os efeitos aleatórios a_i, $i = 1, \ldots, 16$, são independentes entre si com uma distribuição normal de média zero e variância σ^2_{cid}. Os efeitos aleatórios b_j, $j = 1, \ldots, 7$, são independentes entre si com uma distribuição normal de média zero e variância σ^2_{ano}. Um resumo das estimativas dos parâmetros do modelo (5.6) está incluído na Tabela 5.4.

Além das estimativas de $\beta_0, \ldots, \beta_5$, σ^2_{cid} e σ^2_{ano}, é preciso definir outras configurações necessárias para gerar novas amostras *bootstrap* a partir dos dados originais, tais como informações para o processo de reamostragem não paramétrico. A Tabela 5.4 resume todas as configurações de *bootstrap* necessárias para aplicar os métodos propostos ao conjunto de dados de notificações da dengue, das 16 cidades no período de 2008 a 2014.

A seguir, apresentamos um procedimento para as abordagens de *bootstrap* não paramétricas e paramétricas, com o objetivo de estimar o erro padrão (SE) para as componentes de variância associadas aos efeitos aleatórios, e para construir o intervalo de confiança *bootstrap* correspondente (IC):

Passo 1: Definir $\hat{\theta}$, que assumirá um dos valores iniciais apresentados na Tabela 5.4, para os parâmetros $\hat{\beta}_0, \ldots, \hat{\beta}_5, \hat{\sigma}^2_{cid}, \hat{\sigma}^2_{ano}$.

Passo 2: *Bootstrap* paramétrico: a partir dos valores de $\hat{\theta}$, definidos no Passo 1., gerar B bases de dados, assumindo uma distribuição de Poisson, e estimar os respetivos valores do parâmetro $\widehat{\theta}^{*b}_{Par}$, para $\hat{\beta}^{b}_{0}, \hat{\sigma}^{2b}_{cid}, \hat{\sigma}^{2b}_{ano}$, $b = 1, \ldots, B$.

Passo 2.1: Calcular o SE e o IC a 95% para os valores $\widehat{\theta}^{*b}_{Par}$, $b = 1, \ldots, B$, utilizando as Equações (5.1) e (5.2).

Passo 3: *Bootstrap* não paramétrico: a partir dos dados reais, gerar B bases de dados reamostrando com reposição e estimar os B valores $\widehat{\theta}^{*b}_{NPar}$, para $\hat{\beta}^{b}_{0}$, $\hat{\sigma}^{2b}_{cid}$, $\hat{\sigma}^{2b}_{ano}$, $b = 1, \ldots, B$.

Passo 3.1: Calcular o SE e o IC a 95 % para os valores $\widehat{\theta}^{*b}_{NPar}$, $b = 1, \ldots, B$, utilizando as Equações (5.1) e (5.2).

Tabela 5.4: Configurações *bootstrap*.

Parâmetro	**Valor**	**Definição**
$nCid$	16	número total de cidades
$nAno$	7	número total de anos
B	500	número de réplicas *bootstrap*
β_0	-5.5285	ordenada na origem
β_1	0.0025	coeficiente para precipitação desfasada em 6 semanas
β_2	0.1679	coef. para temperatura mínima desfasada em 4 semanas
β_3	-0.0563	coef. para temperatura máxima
β_4	0.0413	coef. para humidade relativa do ar desfasada em 10 semanas
β_5	-0.4596	coef. para velocidade do vento desfasada em 2 semanas
σ^2_{cid}	0.2336	variância para o grupo cidade
σ^2_{ano}	0.5085	variância para o grupo ano
$nRepCid$	6, 8, 16	número de cidades amostradas
$nRepAno$	4, 7	número de anos amostrados

A Tabela 5.5 apresenta os IC *bootstrap* a 95 % para as componentes de variância, obtidos pela Equação (5.2), aplicando os métodos *bootstrap* paramétrico e não paramétrico. Aplicando o *bootstrap* não paramétrico, foram reamostrados simultaneamente os fatores *cidade* e *ano*. A nomenclatura utilizada indica os cenários analisados e o método *bootstrap* aplicado, conforme o exemplo: *NPar_16x7* corresponde a reamostrar 16 cidades e 7 anos, para o método *bootstrap* não paramétrico.

Da Tabela 5.5, ao aplicar o método *bootstrap* paramétrico obtém-se intervalos de confiança com menores amplitudes para as componentes de variância associadas aos efeitos aleatórios. Esse resultado ocorre devido aos dados manterem a estrutura e a dimensão

Tabela 5.5: Intervalo de confiança para as componentes de variância associadas aos efeitos aleatórios *cidade* e *ano* (σ^2_{cid} e σ^2_{ano}), obtidas a partir dos dodos reais, com amostras de dimensões diversas.

CENÁRIOS	**IC para σ^2_{cid}**			**IC para σ^2_{ano}**		
	2.5%	97.5%	Amplitude	2.5%	97.5%	Amplitude
Paramétrico	0.09333	0.42080	0.32747	0.08376	1.04154	0.95778
NPar_16x7	0.14719	1.42306	1.27587	0.18114	1.09474	0.91360
NPar_16x4	0.15267	1.80407	1.65140	0.01253	1.06855	1.05602
NPar_8x7	0.08688	1.00562	0.91874	0.15546	1.46827	1.31281
NPar_8x4	0.09917	2.50894	2.40977	0.02108	1.52458	1.50350
NPar_6x7	0.05548	1.22582	1.17034	0.15852	1.77614	1.61762
NPar_6x4	0.05263	2.21524	2.16261	0.03241	1.76255	1.73014

originais da base de dados reais. Aplicando o *bootstrap* não paramétrico, diminuindo a dimensão da amostra, aumenta a amplitude do intervalo de confiança.

Tomando os resultados obtidos nas simulações da Secção (5.4), pode-se reamostrar cerca de 50 % do total de cada cidade e de cada ano e utilizar o cenário *NPar_8x4*, para o qual o intervalo de confiança deverá permitir uma probabilidade de cobertura superior a 90 %.

Na Tabela 5.6 apresentam-se os valores para os erros padrão estimados para a mediana e para a média das componentes de variância σ^2_{cid} e σ^2_{ano}, nos cenários não paramétricos considerados para os dados reais. Atendendo as indicações do estudo de simulação apresentado na Secção 5.4, para atingir uma probabilidade de cobertura *bootstrap* acima de 90 %, deve-se ter uma amostra de dimensão igual ou inferior a 50 % da dimensão dos grupos amostrados. Assim, mais uma vez, ter-se-á o "melhor" cenário *bootstrap* não paramétrico identificado por *NPar_8x4*, fazendo a reamostragem de 8 cidades, de um total de 16 cidades, e de 4 anos, de um total de 7 anos.

Os resultados apresentados nas Tabelas 5.5 e 5.6 permitem-nos confirmar a significância estatística dos valores estimados para as variâncias σ^2_{cid} e σ^2_{ano}, fundamentando a importância de incluir as respetivas componentes aleatórias no modelo final.

5.6 Conclusão

Neste capítulo recorreu-se a métodos *bootstrap* paramétrico e não paramétrico para estimar os erros padrão (SE) e os intervalos de confiança (IC) para as componentes de variância associadas aos efeitos aleatórios, em modelos lineares generalizados mistos (GLMM).

Tabela 5.6: Erros padrão para a mediana e para a média estimados para as componentes de variância σ^2_{cid} e σ^2_{ano}, a partir dos dados reais, com diversas dimensões de amostras.

CENÁRIOS	**SE para a mediana**		**SE para a media**	
	$\boldsymbol{\sigma^2_{cid}}$	$\boldsymbol{\sigma^2_{ano}}$	$\boldsymbol{\sigma^2_{cid}}$	$\boldsymbol{\sigma^2_{ano}}$
Paramétrico	0.08189	0.25463	0.08137	0.25258
NPar_16x7	0.29520	0.22191	0.28602	0.22030
NPar_16x4	0.48377	0.25723	0.45161	0.25711
NPar_8x7	0.27569	0.35503	0.27014	0.34919
NPar_8x4	0.60876	0.40149	0.57932	0.39544
NPar_6x7	0.38144	0.46351	0.37308	0.44282
NPar_6x4	0.62983	0.47014	0.60881	0.46302

Foram simulados cenários de diferentes dimensões, utilizando dois fatores de efeitos aleatórios, para definir o número de elementos a ser reamostrado de cada fator. A precisão das estimativas das componentes de variância associadas aos efeitos aleatórios foram avaliados utilizando as estatísticas RMSE, MAE e a probabilidade de cobertura *bootstrap* do intervalo de confiança construídos para as componentes de variância associadas aos efeitos aleatórios.

Os resultados obtidos nas simulações mostram que aumentar o número de elementos reamostrados por grupo produz intervalos de confiança mais estreitos e menores desvios padrão, com a média e a mediana dos parâmetros estimados aproximando-se do valor verdadeiro. Todavia, aumentar o número de elementos reamostrados por grupo diminui a probabilidade de cobertura *bootstrap*, sendo necessário ajustar a dimensão da amostra para atingir a probabilidade de cobertura desejada.

Caso se pretenda evitar qualquer tipo de suposição sobre a distribuição da população, o método *bootstrap* não paramétrico deve ser aplicado fazendo a reamostragem, com reposição, numa proporção de 50 % da dimensão do grupo de interesse. Sendo razoável algum pressuposto sobre a distribuição da população, pode-se utilizar o método *bootstrap* paramétrico para estimar o SE e os IC dos parâmetros. Esses procedimentos permitem obter uma probabilidade de cobertura para o IC superior a 90 %.

Os resultados obtidos nas simulações foram aplicados aos dados reais do número de notificações de casos de dengue registados no estado de Goiás, para os quais foram estimados os intervalos de confiança e os erros padrão do estimador das componentes de variância associadas aos efeitos aleatórios a partir da mediana e da média. O comportamento apresentado nos cenários simulados foram percebidos na análise dos IC nos cenários com

dados reais. Ao aumentar o número de elementos reamostrados, o SE diminui, isto, tendo em conta que os SE são proporcionais à raiz quadrada inversa do tamanho da amostra. Os SE para a média apresentaram resultados levemente inferiores quando comparados com os SE para a mediana, comportamento atribuído às características dos dados de notificações de casos de dengue.

Capítulo 6

Conclusões e trabalhos futuros

A Saúde Pública, em países em desenvolvimento, é caracterizada pela forte incidência de doenças que são agravadas pela falta de infraestrutura básica, ocasionada pelo crescimento desordenado, proporcionando a proliferação de agentes causadores de epidemias, como o mosquito *Aedes aegypti* que é o principal transmissor da dengue no Brasil. Considerado país de clima tropical devido à sua localização geográfica, o Brasil apresenta condições climáticas favoráveis para a proliferação do vetor transmissor da dengue, o qual requer um combate contínuo e sistemático para evitar períodos epidémicos.

Neste trabalho, o número de notificações de casos de dengue no estado de Goiás, situado na região central do Brasil, foi estimado utilizando modelação por regressão incorporando dependência espacial e temporal. Foram consideradas as influências das variações meteorológicas em cada período do ano e em diferentes cidades, tendo em conta os efeitos fixos associados às variáveis meteorológicas e os efeitos aleatórios associados aos anos e às cidades em estudo.

No Capítulo 2 apresenta-se uma análise exploratória da base de dados, a qual aponta indícios que existe uma relação/associação entre o número de notificações de casos de dengue e as variáveis meteorológicas, intensificando-se nas primeiras semanas de cada ano. No estado de Goiás, Goiânia é a cidade que regista o maior número de notificações de casos de dengue por 100 mil habitantes, facto que requer um estudo específico.

No Capítulo 3, tendo em conta a correlação temporal inerente aos dados em estudo, o número de notificações de casos de dengue semanal na cidade de Goiânia foi modelado utilizando a abordagem das GEE com a estrutura de correlação estacionária dependente de segunda ordem. Recorreu-se à distribuição Binomial Negativa devido à presença de

sobredispersão nos dados, e foi selecionado o modelo ajustado mais adequado, o modelo com o menor valor do QIC.

Os resultados obtidos permitem concluir que a temperatura mínima, a humidade relativa do ar e a precipitação contribuem positivamente para o aumento do número de notificações de casos de dengue. Admitindo que a região em estudo apresenta condições climáticas favoráveis à proliferação do vetor transmissor da dengue, torna-se necessário a implementação de ações públicas e continuadas para a erradicação de criadouros e dos mosquitos *Aedes aegypti*, para evitar a ocorrência de epidemias e mortes por dengue.

No Capítulo 4 foram apresentados conceitos relacionados com a modelação de dados espacial e temporal, recorrendo-se aos modelos lineares generalizados mistos, incorporando efeitos fixos e aleatórios, para analisar o número de notificações de casos de dengue em 20 cidades do estado de Goiás. Foram ajustados modelos assumindo uma estrutura hierárquica cruzada ou assumindo uma estrutura hierárquica aninhada. As variáveis meteorológicas e as estações do ano foram modeladas como efeitos fixos, e os fatores *cidade* e *ano* foram modelados como variáveis aleatórias não observadas.

Conclui-se que as estações do ano apresentam grandes contribuições para o aumento do número de notificações de casos de dengue registados no estado de Goiás. O verão e o outono são as estações do ano com maior número de casos de dengue registados. A precipitação, a temperatura mínima e a humidade relativa do ar contribuem positivamente, enquanto a temperatura máxima e a velocidade do vento contribuem negativamente para o aumento do número de notificações de casos de dengue.

A estrutura hierárquica aninhada, para os fatores associados aos efeitos aleatórios, permitiu a estimação dos coeficientes de regressão com menor erro padrão, possibilitando uma estimativa mais precisa do valor esperado do número de notificações de casos de dengue.

No Capítulo 5, propõe-se métodos *bootstrap* paramétrico e não paramétrico para estimar o erro padrão (SE) e construir intervalos de confianças (IC) para as componentes de variância associadas aos efeitos aleatórios, em modelos lineares generalizados mistos, utilizando a distribuição de Poisson. A precisão das estimativas foram avaliadas utilizando as estatísticas RMSE e MAE, e os métodos *bootstrap* propostos foram aplicados aos dados reais, permitindo confirmar a significância estatística dos valores estimados das componentes de variância associadas aos efeitos aleatórios.

Os resultados obtidos mostram que aumentar o número de elementos por grupo no processo de reamostragem produz intervalos de confiança mais estreitos e menores desvios padrão, e a média e a mediana das estimativas dos parâmetros aproximam-se do valor verdadeiro. Todavia, a probabilidade de cobertura *bootstrap* diminui.

Conclui-se que o método *bootstrap* não paramétrico é o mais indicado quando não se pretende fazer qualquer suposição sobre a distribuição da população, realizando o processo de reamostragem no grupo de interesse numa proporção de 50 % do total de níveis do grupo, permitindo obter uma probabilidade de cobertura superior a 90 %. Conhecendo-se a lei de distribuição subjacente à população, as estimativas das componentes de variância podem ser estimadas utilizando o método *bootstrap* paramétrico.

6.1 Trabalhos futuros

Como trabalhos futuros, pretende-se alargar o estudo dos métodos *bootstrap*, tendo em conta a possibilidade da variável de interesse ser proveniente da distribuição Binomial Negativa. É também importante considerar interações entre os efeitos aleatórios, com a finalidade de se obter resultados mais precisos. Estes objetivos terão associados novos desafios em termos de peso computacional, que deverão também ser investigados.

Pretende-se, também, desenvolver o algoritmo para modelos com zeros inflacionados, uma vez que não há registos de notificações de casos de dengue em algumas semanas, principalmente quando a estação do ano é o inverno.

Relativamente à base de dados, poder-se-á tentar incorporar informações de levantamento rápido de índice de infestação por *Aedes aegypti* (LIRAa), que potenciam a identificação de criadouros predominantes e situações de infestação numa determinada cidade.

Bibliografia

Bailey, H., Corkrey, R., Cheney, B., and Thompson, P. M. (2013). Analyzing temporally correlated dolphin sightings data using generalized estimating equations. *Marine Mammal Science*, 29(1):123–141.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48. doi:10.18637/jss.v067.i01.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., and White, J. S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in ecology & evolution*, 24(3):127–135.

Brasil, da Saúde (MS), M., de Vigilância em Saúde, S., and de Vigilância Epidemiológica, D. (2009). *Diretrizes nacionais para a prevenção e controle de epidemias de dengue*. Ministério da Saúde, Brasília(DF).

Brasil and Ministério da Saúde (MS) (2010). Portaria n. 2472, de 31 de agosto de 2010. Define as terminologias adotadas em legislação nacional, conforme disposto no Regulamento Sanitário Internacional 2005 (RSI 2005), a relação de doenças, agravos e eventos em saúde pública de notificação compulsória em todo o território nacional e estabelecer fluxo, critérios, responsabilidades e atribuições aos profissionais e serviços de saúde. Diário Oficial da República Federativa do Brasil, Brasília(DF); 2010.

Chatterjee, S. and Hadi, A. S. (2015). *Regression analysis by example*. John Wiley & Sons.

Chen, S. C., Liao, C. M., Chio, C. P., Chou, H. H., You, S. H., and Cheng, Y. H. (2010). Lagged temperature effect with mosquito transmission potential explains dengue va-

riability in southern taiwan: insights from a statistical analysis. *Science of the total environment*, 408(19):4069–4075.

Chernick, M. R. (2011). *Bootstrap Methods: A Guide for Practitioners and Researchers*, volume 619. John Wiley & Sons.

Depradine, C. and Lovell, E. (2004). Climatological variables and the incidence of dengue fever in barbados. *International journal of environmental health research*, 14(6):429–441.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Gomes, A. F., Nobre, A. A., and Cruz, O. G. (2012). Temporal analysis of the relationship between dengue and meteorological variables in the city of rio de janeiro, brazil, 2001-2009. *Cadernos de Saúde Pública*, 28(11):2189–2197.

Hardin, J. W. and Hilbe, J. M. (2003). *Generalized estimating equations*. Wiley Online Library.

Hartig, F. (2017). *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models*. R package version 0.1.5. `https://CRAN.R-project.org/package=DHARMa`.

Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge University Press.

Johnson, P. (2015). *GLMMmisc: Miscellaneous functions for GLMMs*.

Johnson, P. C., Barry, S. J., Ferguson, H. M., and Müller, P. (2015). Power analysis for generalized linear mixed models in ecology and evolution. *Methods in ecology and evolution*, 6(2):133–142.

Jury, M. R. (2008). Climate influence on dengue epidemics in puerto rico. *International Journal of Environmental Health Research*, 18(5):323–334.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

Lowe, R., Bailey, T. C., Stephenson, D. B., Graham, R. J., Coelho, C. A., Carvalho, M. S., and Barcellos, C. (2011). Spatio-temporal modelling of climate-sensitive disease risk: Towards an early warning system for dengue in brazil. *Computers & Geosciences*, 37(3):371–381.

Lu, L., Lin, H., Tian, L., Yang, W., Sun, J., and Liu, Q. (2009). Time series analysis of dengue fever and weather in guangzhou, china. *BMC Public Health*, 9(1):1.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*, volume 37. CRC press.

McCulloch, C. E. and Neuhaus, J. M. (2001). *Generalized linear mixed models*. Wiley Online Library.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics*, 57(1):120–125.

Pinto, E., Coelho, M., Oliver, L., and Massad, E. (2011). The influence of climate variables on dengue in singapore. *International journal of environmental health research*, 21(6):415–426.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. `https://www.R-project.org/`.

Released, I. C. (2015). IBM SPSS Statistics for Windows, version 23.0.

Sinha, S. K. (2009). Bootstrap tests for variance components in generalized linear mixed models. *Canadian Journal of Statistics*, 37(2):219–234.

Souza, S. S., Silva, H. H. G., and Silva, I. G. (2010). Associação entre incidência de dengue, pluviosidade e densidade larvária de aedes aegypti, no estado de goiás. *Revista da Sociedade Brasileira de Medicina Tropical*, pages 152–155.

Thai, H. T., Mentré, F., Holford, N. H., Veyrat-Follet, C., and Comets, E. (2013). A comparison of bootstrap approaches for estimating uncertainty of parameters in linear mixed-effects models. *Pharmaceutical statistics*, 12(3):129–140.

Viana, D. V. and Ignotti, E. (2013). A ocorrência da dengue e variações meteorológicas no Brasil: revisão sistemática. *Revista Brasileira de Epidemiologia*, 16(2):240–256.

West, B. T., Welch, K. B., and Galecki, A. T. (2014). *Linear mixed models: a practical guide using statistical software*. CRC Press.

Yu, H. L., Yang, S. J., Yen, H. J., and Christakos, G. (2011). A spatio-temporal climate-based model of early dengue fever warning in southern taiwan. *Stochastic Environmental Research and Risk Assessment*, 25(4):485–494.

Zeger, S. L., Liang, K. Y., and Albert, P. S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, pages 1049–1060.

Zuur, A., Ieno, E., Walker, N., Saveliev, A., and Smith, G. (2009). *Mixed effects models and extensions in ecology with R. Gail M, Krickeberg K, Samet JM, Tsiatis A, Wong W, editors*.

Zuur, A., Ieno, E. N., and Smith, G. M. (2007). *Analyzing ecological data*. Springer Science & Business Media.

Anexo A

Boxplots por cidade

Nas Figuras A.1 a A.7 têm-se as *boxplots* das variáveis meteorológicas nas cidades em estudo, no período de janeiro de 2008 a março de 2015, tendo em conta as estações do ano. Nota-se que as variáveis apresentam comportamento semelhante em todas as cidades, sendo:

- Figuras A.1 e A.2: Notificações de casos de dengue em cada cidade por estação do ano, apresenta maior número de casos registados no verão e no outono.
- Figura A.3: a precipitação teve maiores registos nas estações verão e outono, acompanhando o número de notificações de casos de dengue registados.
- Figura A.4: a temperatura mínima teve menores registos no inverno e, maiores registos, no verão.
- Figura A.5: a temperatura máxima apresenta maiores registos no inverno e na primavera, comportamento inverso ao número de notificações de casos de dengue.
- Figura A.6: a humidade relativa do ar acompanha a tendência do número de notificações de casos de dengue, com maiores registos no verão e no outono.
- Figura A.7: a velocidade do vento não apresenta padrão definido, com valores oscilando em 2 m/s.

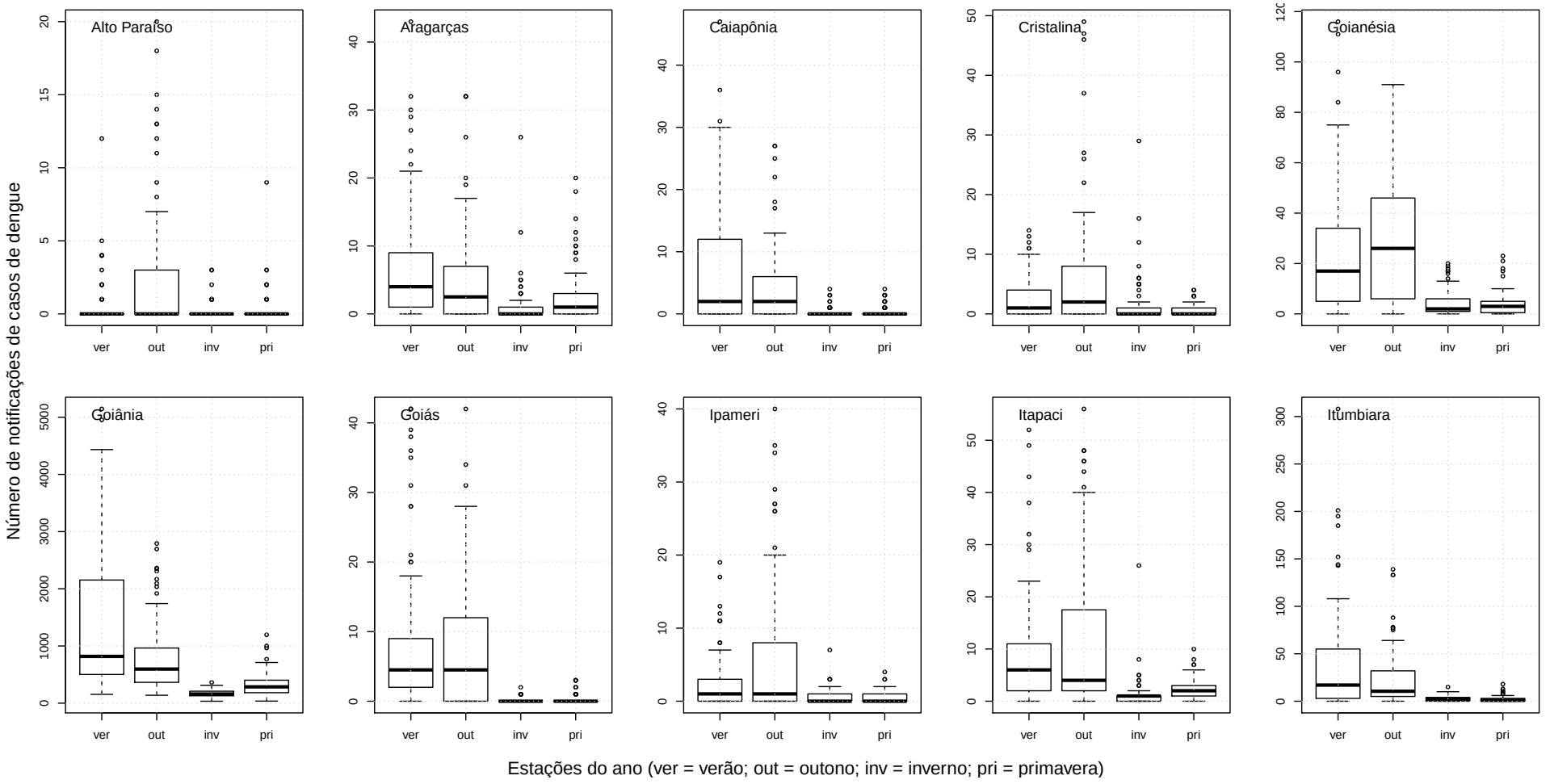


Figura A.1: Notificações de casos de dengue nas cidades em estudo, no período de janeiro de 2008 a março de 2015, tendo em conta as estações do ano.

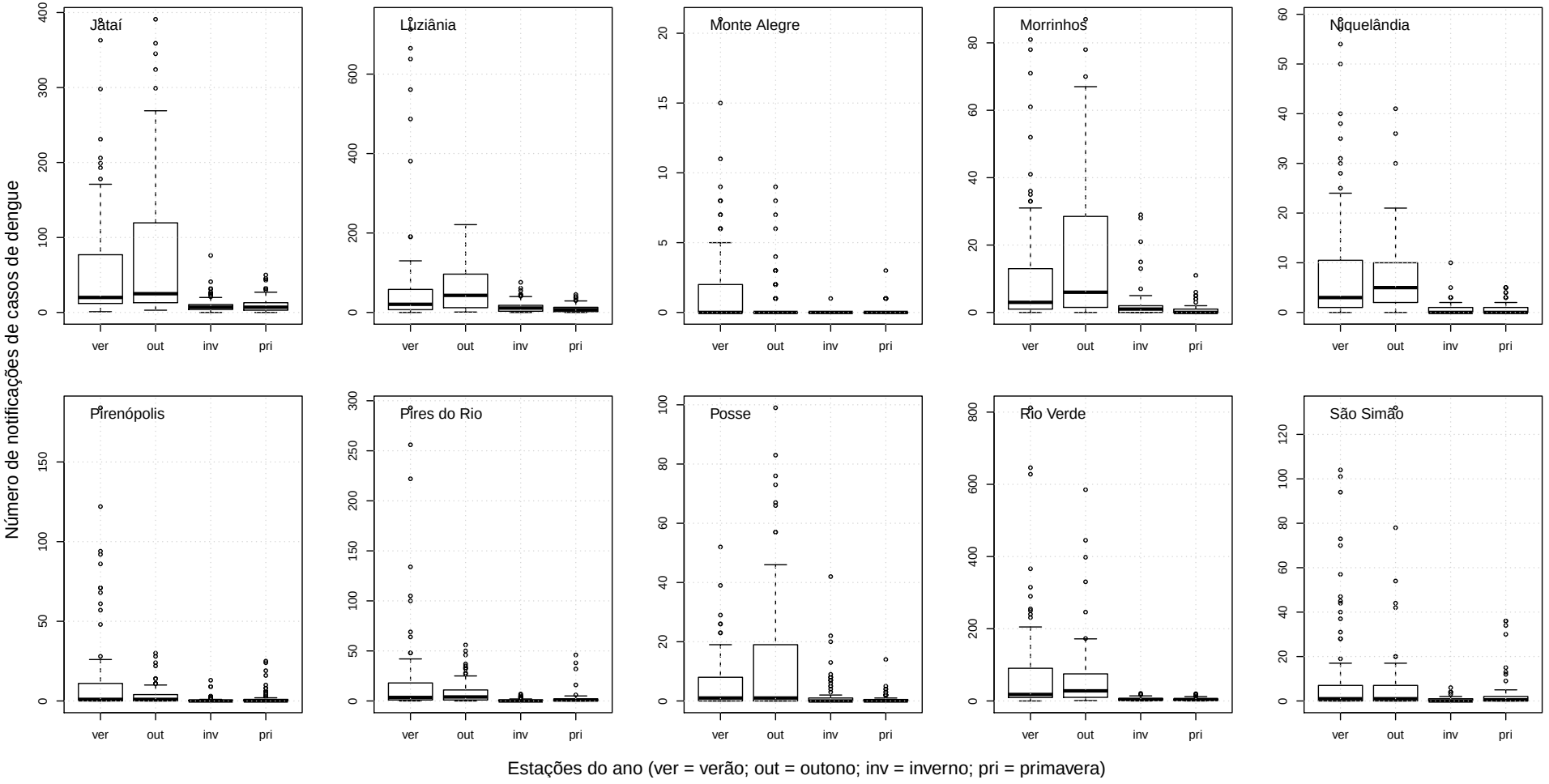


Figura A.2: Notificações de casos de dengue nas cidades em estudo, no período de janeiro de 2008 a março de 2015, tendo em conta as estações do ano.

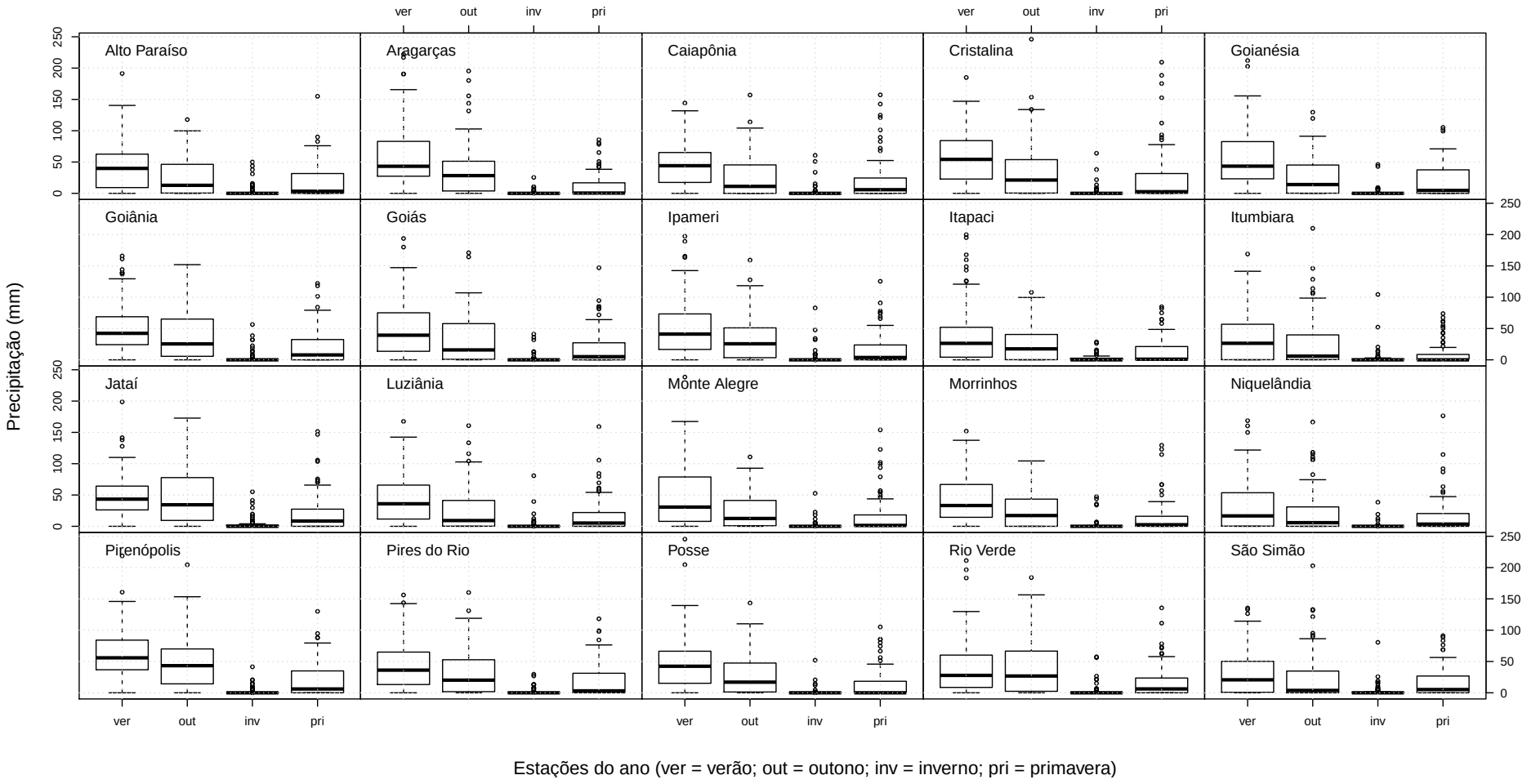


Figura A.3: Precipitação nas cidades em estudo, no período de janeiro de 2008 a março de 2015, tendo em conta as estações do ano.

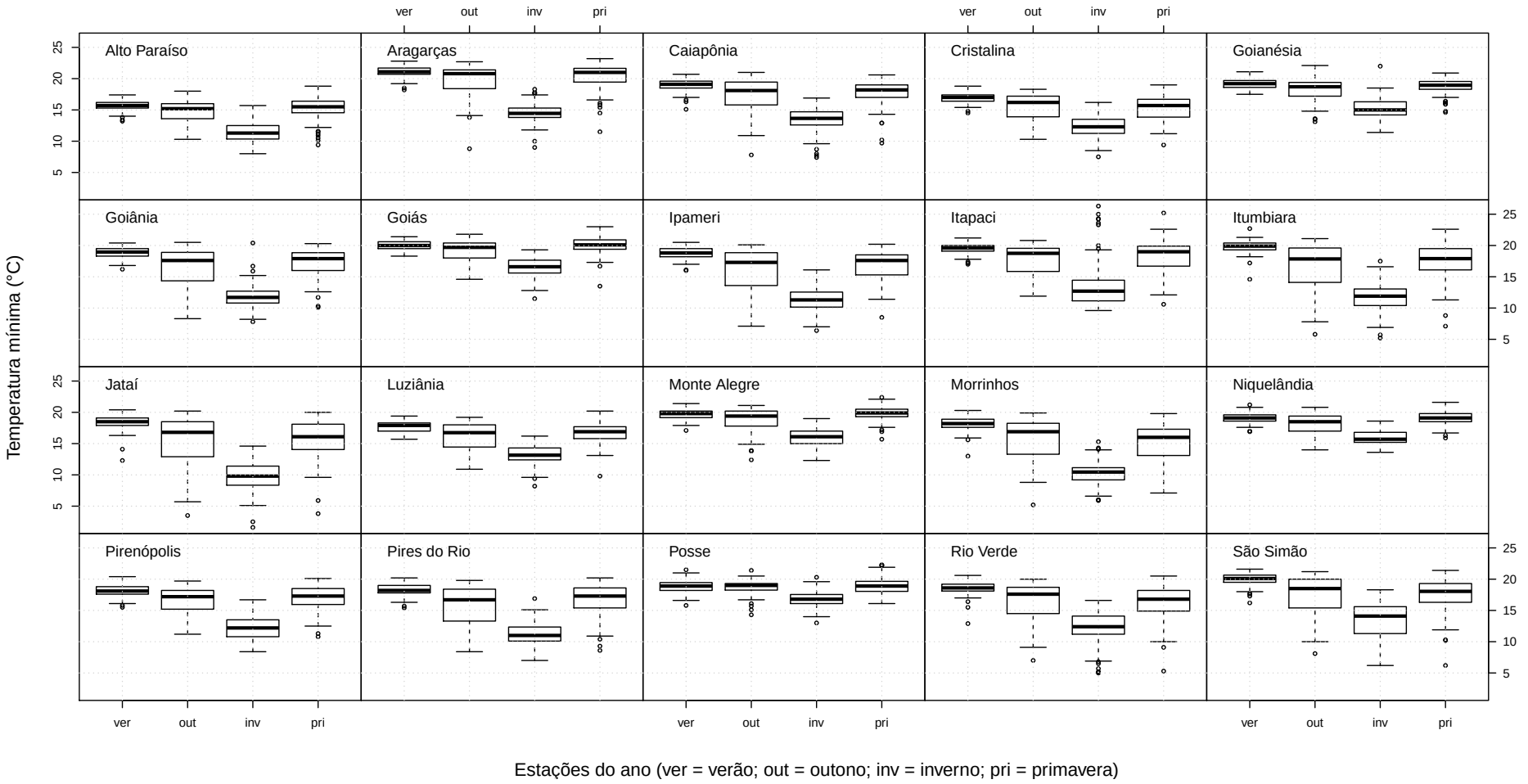


Figura A.4: Temperatura mínima nas cidades em estudo, no período de janeiro de 2008 a março de 2015, tendo em conta as estações do ano.

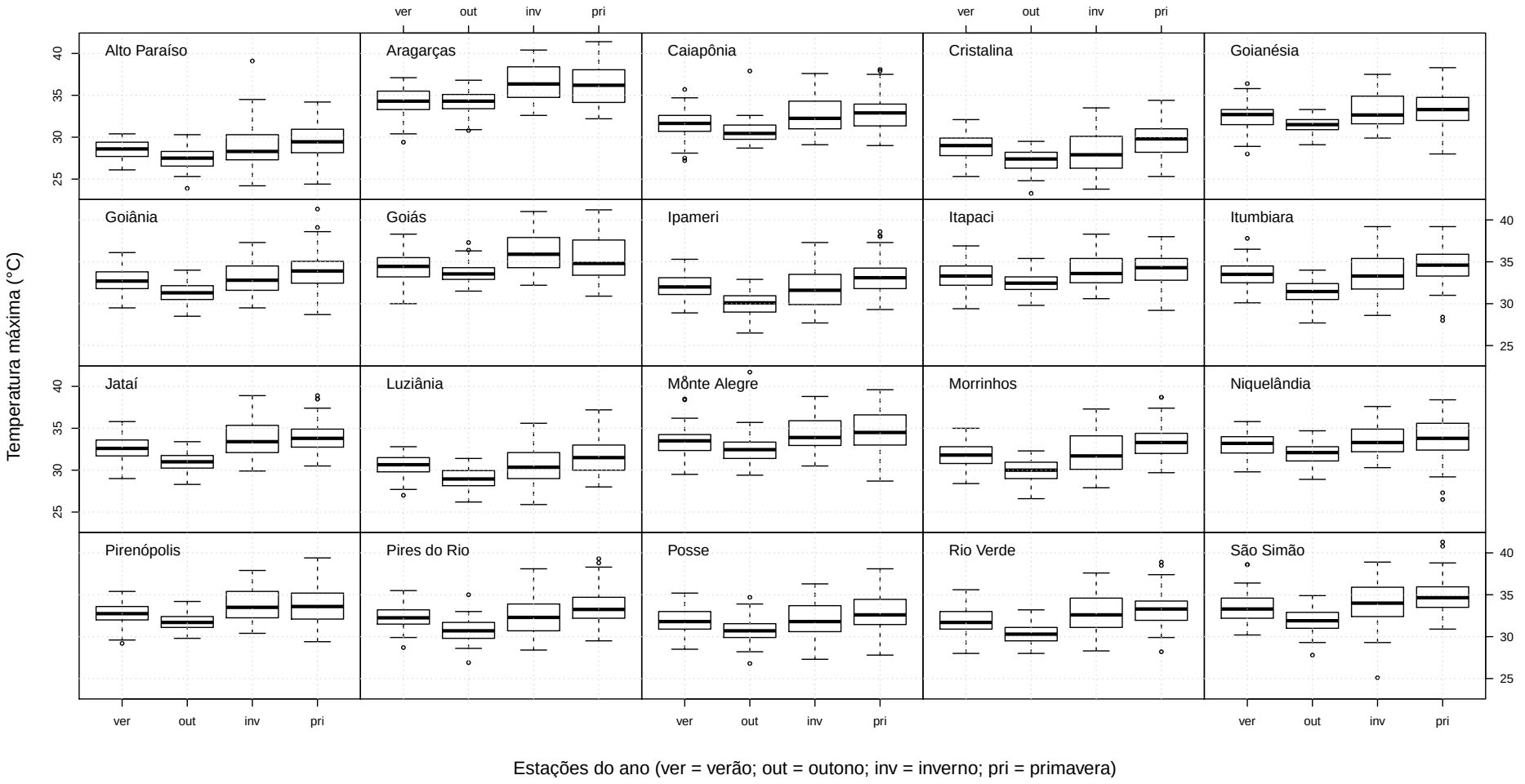


Figura A.5: Temperatura máxima nas cidades em estudo, no período de janeiro de 2008 a março de 2015, tendo em conta as estações do ano.

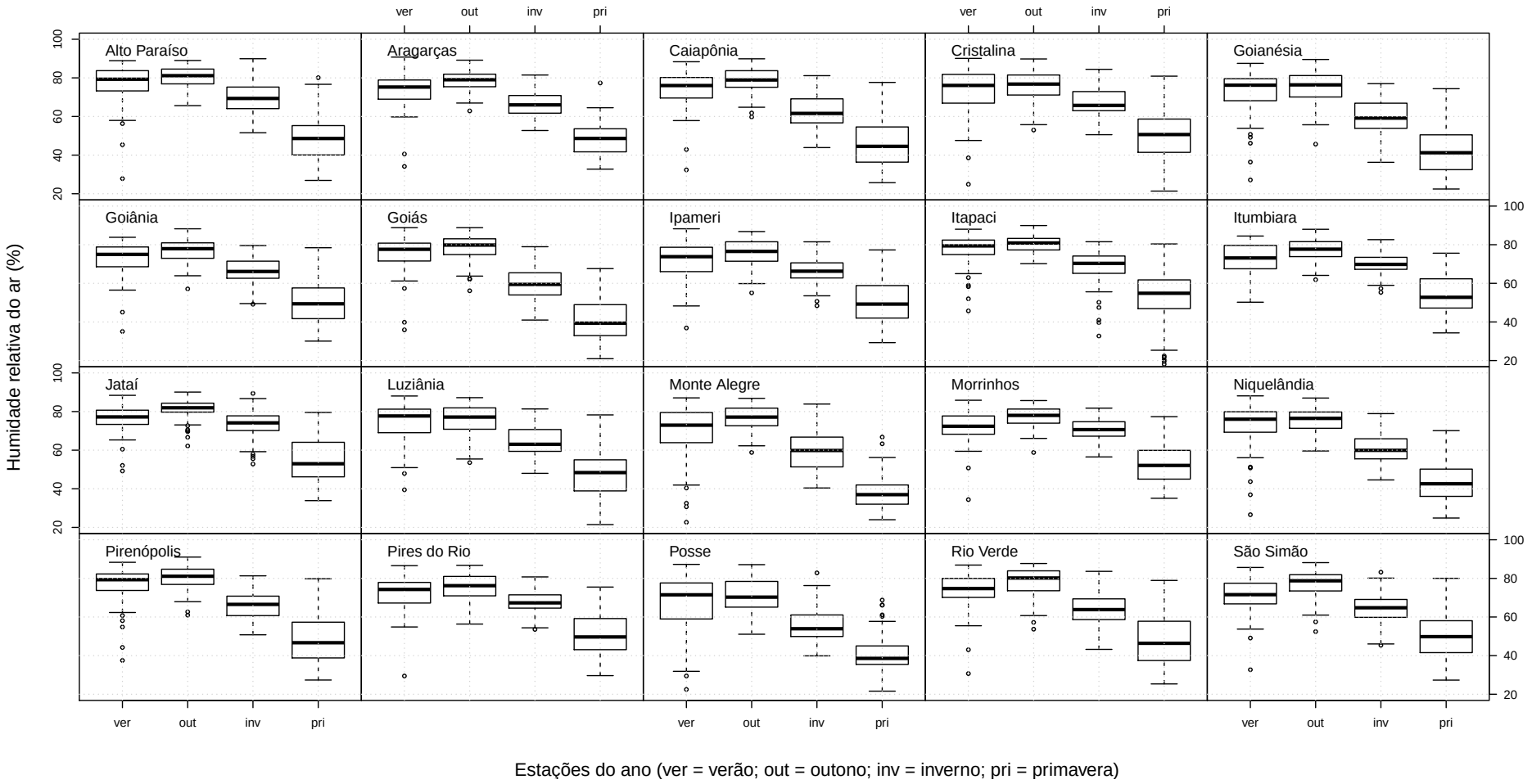


Figura A.6: Humidade relativa do ar nas cidades em estudo, no período de janeiro de 2008 a março de 2015, tendo em conta as estações do ano.

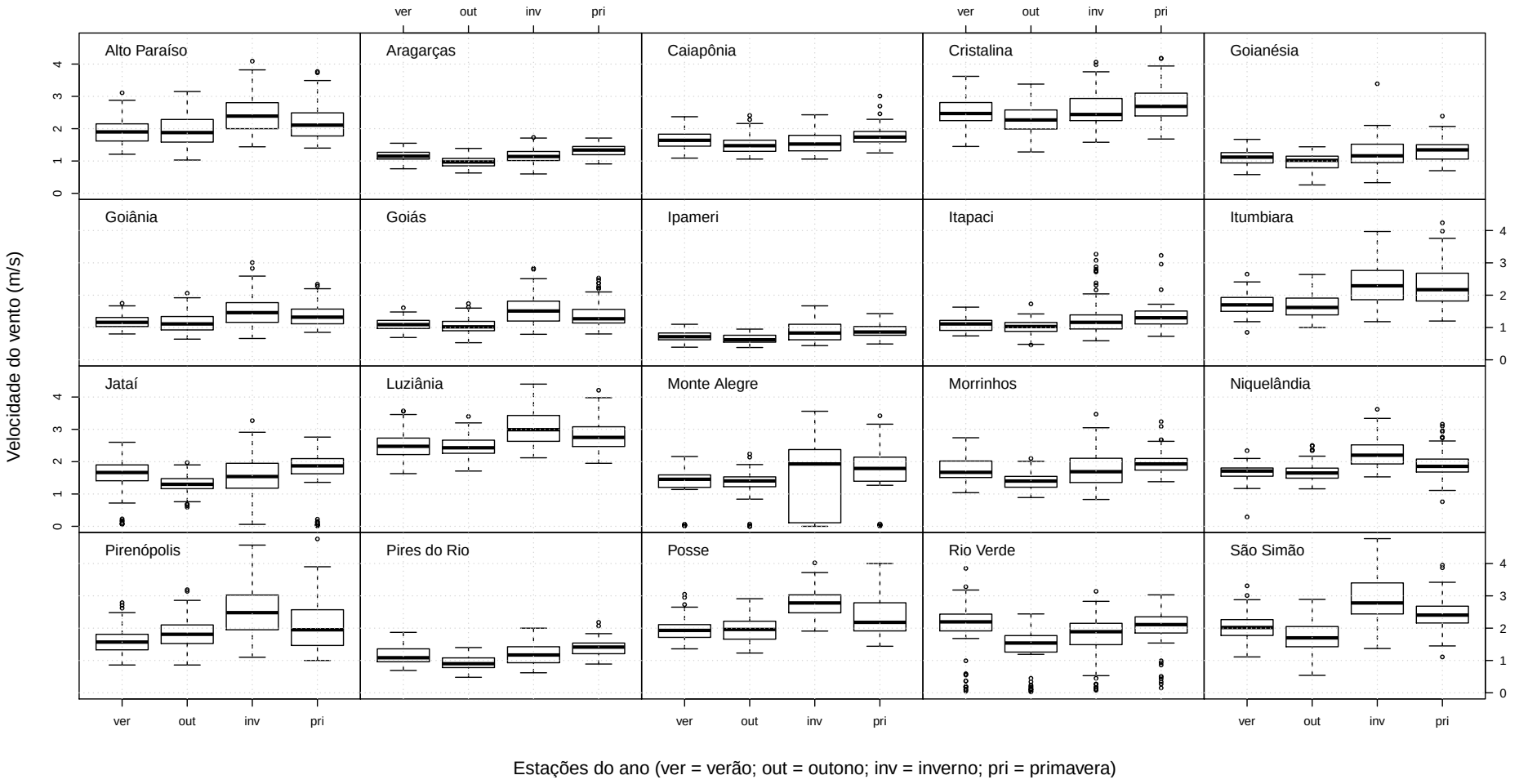


Figura A.7: Velocidade do vento nas cidades em estudo, no período de janeiro de 2008 a março de 2015, tendo em contaas estações do ano.

Anexo B

Análise de correlações para as séries temporais da variável "dengue"

B.1 Autocorrelação para a variável "dengue"

Seja Y_t o número de notificações de casos de dengue numa determinada cidade, na semana t, a função de autocorrelação indicará se existe dependência temporal entre as variáveis Y_t e Y_{t+k}, com $k = 1, 2, 3, ..., N$ para essa cidade. Esta relação é quantificada pelo coeficiente de correlação de *Pearson* e é definido sempre entre -1 e 1, sendo indicada para inferir padrões na série temporal (Zuur et al., 2007).

Nas Figuras B.1 e B.2 são apresentadas as autocorrelações para as 20 séries temporais, uma por cada cidade, do número de notificações de casos de dengue para um desfasamento k máximo dado por $0.4 \times N$, sendo N a dimensão da série. As correlações foram obtidas à custa do coeficiente de correlação de *Pearson*, relaxando-se o pressuposto de normalidade da variável Y, uma vez que tal foi feito num contexto de análise exploratória e não de modelação. Nota-se padrões cíclicos com período anual (52 semanas).

B.2 Correlação cruzada para a variável "dengue"

Seja Y_t e X_{t-k} o número de notificações de casos de dengue na semana t numa determinada cidade e na semana $t - k$ numa cidade distinta, respetivamente. A correlação cruzada quantificará a associação entre estas duas variáveis com um atraso de k semanas. Nos testes de correlações foi adotado o método de *Spearman* por não exigir pressupostos de

normalidade. A Tabela B.1 apresenta as correlações estimadas entre as 20 séries temporais do número de notificações de casos de dengue. Nota-se que a maioria das correlações são significativamente diferentes de zero.

A máxima correlação cruzada obtida entre duas séries temporais está apresentada na Tabela B.2. O painel superior mostra as máximas correlações e o painel inferior indica os desfasamentos em que ocorreram. Verifica-se que a máxima correlação cruzada entre a Cidade 6 (Goiânia) e a Cidade 14 (Morrinhos) é de 0.66 e ocorre com um atraso de 5 semanas ($k = -5$).

B.3 Correlação cruzada para os resíduos de um modelo GLM

Na Tabela B.3 apresentam-se as correlações entre as 20 séries temporais dos resíduos resultantes do seguinte modelo GLM, que assume a distribuição Binomial Negativa para a variável dependente:

$$dengue \sim prec_{t-6} + tmin_{t-4} + tmax_t + hram_{t-10} + vvto_{t-2} + ln(thab_t)$$

Nota-se um número menor de correlações significativamente diferentes de zero associadas aos resíduos.

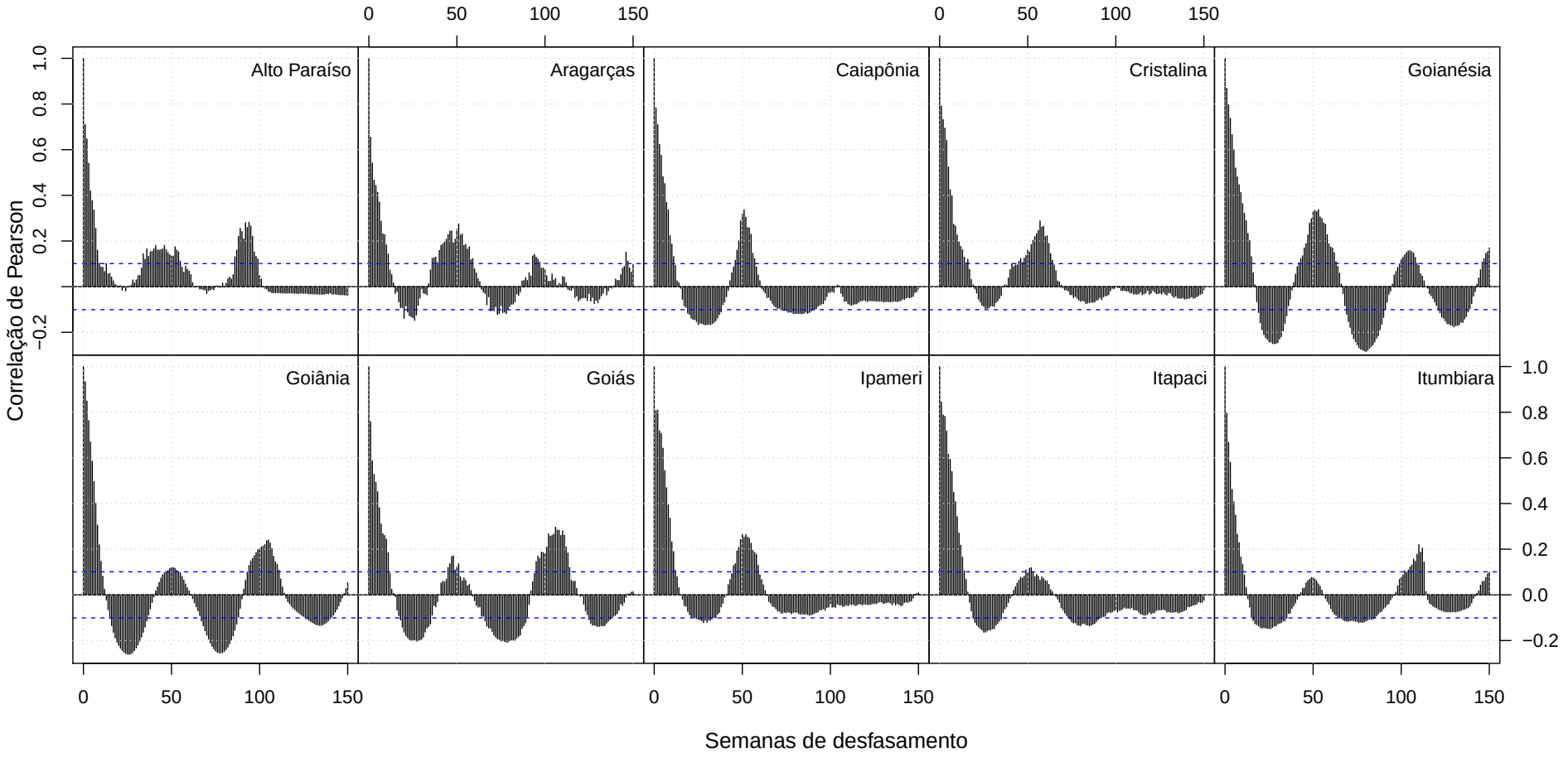


Figura B.1: Autocorrelação das séries temporais do número de notificações de casos de dengue nas cidades do estado de Goiás, considerando-se um desfasamento k máximo de 150 semanas $(lag.max = 150)$.

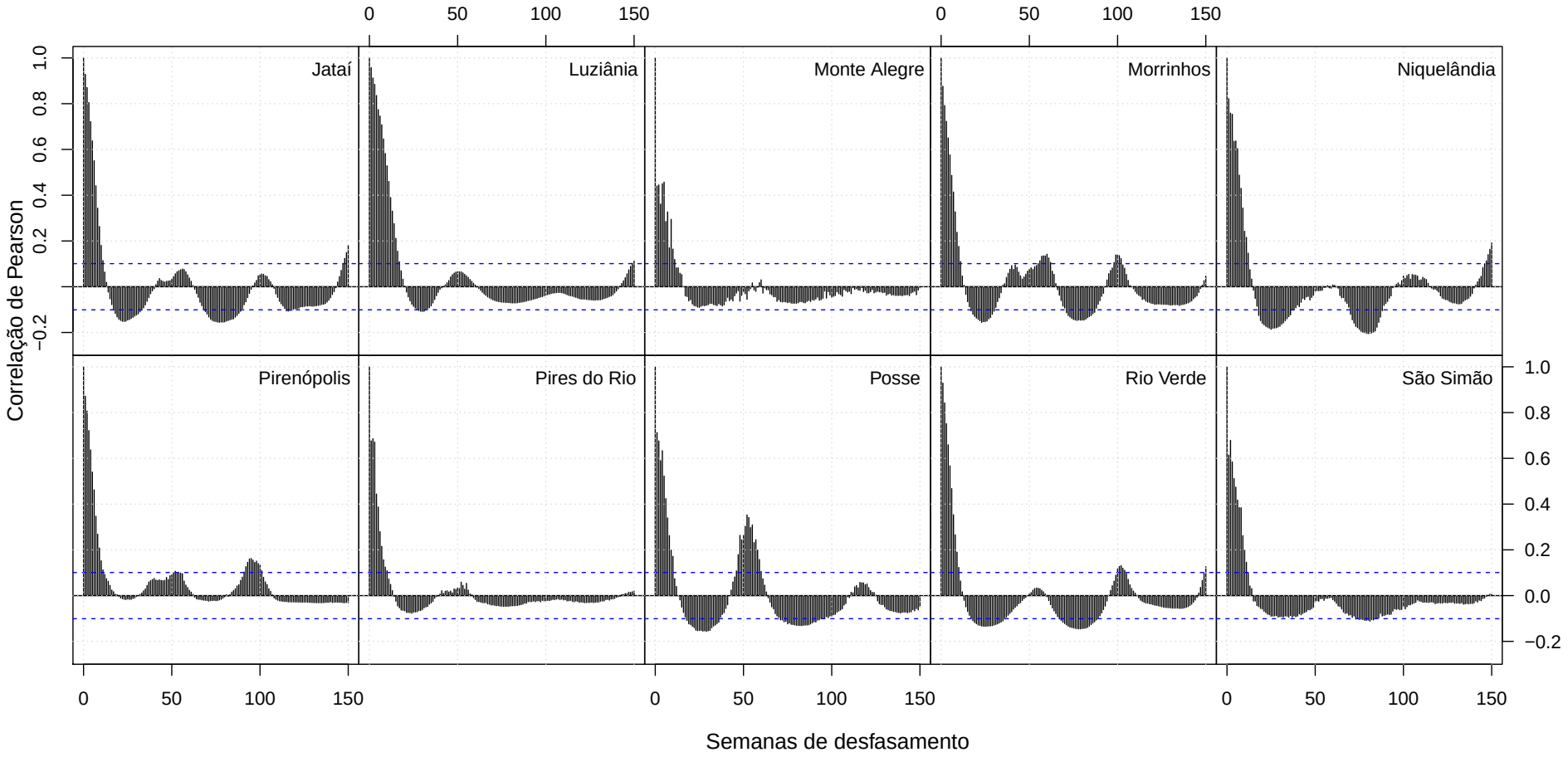


Figura B.2: Autocorrelação das séries temporais do número de notificações de casos de dengue nas cidades do estado de Goiás, considerando-se um desfasamento k máximo de 150 semanas $(lag.max = 150)$.

Tabela B.1: **Correlação cruzada.** Painel superior: correlação entre as 20 séries temporais do número de notificações de casos de dengue nas cidades goianas. Painel inferior: indica se os valores correspondentes no painel superior são significativamente diferentes de 0 ao nível de 5%. 1: $p.value < 0.05$; 0: $p.value >= 0.05$ (para facilitar a visualização, os valores com $p.value >= 0.05$ aparecem a negro). Os números de 1 a 20 correspondem às cidades em análise.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1		-0.25	0.38	0.52	0.44	0.35	0.34	0.46	0.41	**0.05**	0.49	0.45	0.23	0.5	0.26	0.59	0.21	0.4	0.31	0.36
2	1		0.33	**-0.06**	0.14	0.45	0.27	**0.06**	0.27	0.41	0.2	**-0.05**	0.22	0.13	0.19	**-0.09**	0.17	0.15	0.35	0.14
3	1	1		0.46	0.58	0.66	0.54	0.51	0.6	0.42	0.61	0.39	0.45	0.55	0.42	0.36	0.45	0.52	0.58	0.49
4	1	0	1		0.54	0.36	0.48	0.49	0.49	0.2	0.59	0.63	0.3	0.54	0.39	0.41	0.35	0.55	0.42	0.39
5	1	1	1	1		0.63	0.61	0.57	0.51	0.33	0.68	0.71	0.39	0.59	0.55	0.41	0.6	0.56	0.7	0.43
6	1	1	1	1	1		0.67	0.47	0.63	0.6	0.71	0.39	0.43	0.56	0.57	0.47	0.48	0.54	0.76	0.51
7	1	1	1	1	1	1		0.49	0.52	0.5	0.65	0.48	0.32	0.53	0.51	0.4	0.51	0.54	0.61	0.46
8	1	0	1	1	1	1	1		0.51	0.21	0.62	0.62	0.34	0.52	0.38	0.41	0.46	0.48	0.5	0.42
9	1	1	1	1	1	1	1	1		0.41	0.64	0.45	0.35	0.58	0.48	0.33	0.38	0.51	0.55	0.43
10	0	1	1	1	1	1	1	1	1		0.49	0.21	0.16	0.44	0.56	0.21	0.26	0.3	0.54	0.19
11	1	1	1	1	1	1	1	1	1	1		0.62	0.37	0.74	0.58	0.56	0.49	0.62	0.67	0.48
12	1	0	1	1	1	1	1	1	1	1	1		0.27	0.49	0.46	0.32	0.51	0.52	0.5	0.28
13	1	1	1	1	1	1	1	1	1	1	1	1		0.36	0.26	0.17	0.29	0.39	0.38	0.39
14	1	1	1	1	1	1	1	1	1	1	1	1	1		0.56	0.46	0.32	0.56	0.61	0.41
15	1	1	1	1	1	1	1	1	1	1	1	1	1	1		0.19	0.39	0.49	0.62	0.37
16	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1		0.23	0.32	0.35	0.33
17	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		0.32	0.55	0.37
18	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		0.47	0.47
19	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		0.45
20	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	

Cidades: 1. Alto Paraíso; 2. Aragarças; 3. Caiapônia; 4. Cristalina; 5. Goianésia; 6. Goiânia; 7. Goiás; 8. Ipameri; 9. Itapaci; 10. Itumbiara; 11. Jataí; 12. Luziânia; 13. Monte Alegre; 14. Morrinhos; 15. Niquelândia; 16. Pirenópolis; 17. Pires do Rio; 18. Posse; 19. Rio Verde; 20. São Simão.

Tabela B.2: **Correlação cruzada.** Painel superior: máxima correlação cruzada para as séries temporais do número de notificações de casos de dengue nas cidades goianas. Painel inferior: mostra os espaços de tempo para os valores máximos obtidos no painel superior. Os números de 1 a 20 correspondem às cidades em estudo. O desfasamento máximo considerado entre duas séries temporais distintas foi de 100 semanas (aproximadamente 25% da dimensão da série dengue).

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1		-0.17	0.33	0.55	0.37	0.31	0.39	0.53	0.36	0.25	0.39	0.50	0.22	0.48	0.14	0.78	0.64	0.39	0.56	0.36
2	86		0.31	-0.20	-0.31	0.44	0.30	-0.21	0.25	0.39	0.30	0.21	0.30	-0.25	0.42	-0.16	0.25	-0.17	0.33	0.26
3	47	-51		0.58	0.54	0.51	0.44	0.72	0.72	0.44	0.66	0.56	0.55	0.60	0.51	0.50	0.56	0.56	0.51	0.40
4	38	-34	-9		0.61	0.48	0.34	0.76	0.66	0.41	0.58	0.71	0.42	0.67	0.27	0.77	0.71	0.70	0.68	0.54
5	52	-84	0	4		0.60	0.52	0.61	0.54	0.43	0.59	0.61	0.45	0.61	0.44	0.46	0.49	0.53	0.58	0.50
6	0	-1	5	65	1		0.51	0.47	0.44	0.60	0.66	0.61	0.38	**0.66**	0.65	0.50	0.49	0.47	0.76	0.64
7	-2	-57	-50	-41	0	0		0.42	0.44	0.43	0.50	0.46	0.44	0.46	0.49	0.41	0.45	0.42	0.45	0.37
8	44	-87	-6	4	1	-59	49		0.77	0.49	0.62	0.74	0.46	0.60	0.33	0.76	0.70	0.70	0.62	0.53
9	44	-53	-2	4	-2	-58	49	0		0.54	0.68	0.70	0.55	0.64	0.49	0.60	0.59	0.56	0.52	0.45
10	13	0	55	70	55	0	0	61	58		0.36	0.51	0.37	0.32	0.37	0.31	0.63	0.32	0.53	0.38
11	0	-10	-4	2	1	-7	0	0	0	0		0.49	0.54	0.85	0.75	0.35	0.33	0.52	0.65	0.48
12	44	-62	-3	6	0	-56	-55	1	3	-55	-52		0.38	0.43	0.41	0.77	0.71	0.55	0.77	0.62
13	42	-3	2	6	4	2	59	5	7	-51	10	2		0.54	0.55	0.30	0.33	0.46	0.30	0.39
14	0	-88	-5	2	-3	**-5**	0	0	0	0	0	53	-7		0.63	0.46	0.44	0.58	0.64	0.51
15	100	-6	0	3	0	-5	52	2	0	-49	3	55	-7	1		0.23	0.19	0.36	0.56	0.49
16	1	-85	-46	-35	-99	-100	4	-43	-43	6	1	-43	-45	-37	-100		0.81	0.62	0.80	0.54
17	51	-51	1	14	-2	-52	53	9	7	-57	10	8	4	13	-52	50		0.57	0.75	0.55
18	38	25	-8	2	-3	-9	-4	-3	-2	-65	-1	-4	-4	-2	-1	37	-12		0.53	0.50
19	100	-7	50	63	-1	-2	0	56	56	-7	3	55	-4	1	-1	99	49	6		0.65
20	95	-2	5	62	1	0	4	57	57	-5	8	52	3	6	5	100	47	8	3	

Cidades: 1. Alto Paraíso; 2. Aragarças; 3. Caiapônia; 4. Cristalina; 5. Goianésia; 6. Goiânia; 7. Goiás; 8. Ipameri; 9. Itapaci; 10. Itumbiara; 11. Jataí; 12. Luziânia; 13. Monte Alegre; 14. Morrinhos; 15. Niquelândia; 16. Pirenópolis; 17. Pires do Rio; 18. Posse; 19. Rio Verde; 20. São Simão.

Tabela B.3: **Correlação cruzada dos resíduos.** Painel superior: correlação entre as 20 séries temporais dos resíduos do modelo ajustado com GLM e a distribuição Binomial Negativa, dos dados em estudo. Painel inferior: indica se os valores correspondentes no painel superior são significativamente diferentes de 0 ao nível de 5%; 1: $p.value < 0.05$; 0: $p.value >= 0.05$ (para facilitar a visualização, os valores com $p.value >= 0.05$ aparecem a negro). Os números de 1 a 20 correspondem às cidades em estudo.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1		-0.27	0.36	0.44	0.12	0.22	**0.11**	0.39	0.22	**-0.08**	0.35	0.18	0.34	0.42	**-0.03**	0.52	0.16	0.35	**0.05**	0.31
2	1		**0.05**	-0.13	**-0.11**	0.24	**0.06**	**-0.1**	**0.12**	0.31	**0.01**	-0.13	**-0.11**	**0.01**	**0.04**	-0.25	**-0.02**	**-0.01**	0.16	**0.01**
3	1	0		0.34	0.4	0.42	0.23	0.42	0.38	0.24	0.45	0.17	0.28	0.42	0.17	0.22	0.35	0.36	0.45	0.38
4	1	1	1		0.36	0.23	0.35	0.43	0.32	**0.02**	0.55	0.43	0.15	0.45	0.21	0.3	0.27	0.49	0.26	0.3
5	1	0	1	1		0.29	0.38	0.44	0.19	**0**	0.44	0.61	0.22	0.22	0.33	0.15	0.52	0.42	0.46	0.24
6	1	1	1	1	1		0.29	0.36	0.38	0.39	0.56	0.17	0.09	0.49	0.22	0.36	0.24	0.37	0.55	0.46
7	0	0	1	1	1	1		0.37	0.18	0.17	0.46	0.34	0.04	0.3	0.18	0.24	0.32	0.36	0.31	0.24
8	1	0	1	1	1	1	1		0.34	**0.06**	0.6	0.45	0.24	0.47	0.15	0.34	0.4	0.47	0.4	0.37
9	1	0	1	1	1	1	1	1		0.25	0.35	**0.06**	0.14	0.42	0.19	0.12	0.15	0.31	0.33	0.22
10	0	1	1	0	0	1	1	0	1		0.22	-0.14	**-0.07**	0.32	0.23	**0.08**	**0.05**	**0.04**	0.26	**0.08**
11	1	0	1	1	1	1	1	1	1	1		0.47	0.12	0.64	0.33	0.44	0.36	0.52	0.58	0.42
12	1	1	1	1	1	1	1	1	0	1	1		0.12	0.18	0.17	0.19	0.49	0.36	0.32	**0.03**
13	1	0	1	1	1	0	0	1	1	0	1	1		**0.1**	0.16	**0.07**	0.19	0.26	0.2	0.22
14	1	0	1	1	1	1	1	1	1	1	1	1	0		0.18	0.54	**0.04**	0.47	0.34	0.38
15	0	0	1	1	1	1	1	1	1	1	1	1	1	1		**-0.04**	**0.11**	0.36	0.42	0.13
16	1	1	1	1	1	1	1	1	1	0	1	1	0	1	0		0.12	0.2	0.15	0.24
17	1	0	1	1	1	1	1	1	1	0	1	1	1	0	0	1		0.23	0.42	0.25
18	1	0	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1		0.35	0.39
19	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		0.43
20	1	0	1	1	1	1	1	1	1	0	1	0	1	1	1	1	1	1	1	

Cidades: 1. Alto Paraíso; 2. Aragarças; 3. Caiapônia; 4. Cristalina; 5. Goianésia; 6. Goiânia; 7. Goiás; 8. Ipameri; 9. Itapaci; 10. Itumbiara; 11. Jataí; 12. Luziânia; 13. Monte Alegre; 14. Morrinhos; 15. Niquelândia; 16. Pirenópolis; 17. Pires do Rio; 18. Posse; 19. Rio Verde; 20. São Simão.