Instituto de MATEMÁTICA E ESTATÍSTICA UFRGS

DEPARTAMENTO DE ESTATÍSTICA
E-mail: dest@mat.ufrgs.br

Trabalho de Conclusão de Curso

Granger Causalidade e a Dinâmica Migratória do Vírus da Gripe

Aline Foerster Grande

25 de novembro de 2020

Aline Foerster Grande

Granger Causalidade e a Dinâmica Migratória do Vírus da Gripe

Trabalho de Conclusão apresentado à comissão de Graduação do Departamento de Estatística da Universidade Federal do Rio Grande do Sul, como parte dos requisitos para obtenção do título de Bacharel em Estatística.

Orientador: Prof. Dr. Guilherme Pumi

Porto Alegre
Novembro de 2020

Aline Foerster Grande

Granger Causalidade e a Dinâmica Migratória do Vírus da Gripe

Este Trabalho foi julgado adequado para obtenção dos créditos da disciplina Trabalho de Conclusão de Curso em Estatística e aprovado em sua forma final pela Orientador e pela Banca Examinadora.

Orientador: _____
Prof. Dr. Guilherme Pumi, UFRGS

Banca Examinadora:

Prof. Dr. Rodrigo Citton Padilha dos Reis, UFRGS

Profa. Dra. Gabriela Bettella Cybis, UFRGS

Porto Alegre
Novembro de 2020

Agradecimentos

Gostaria de agradecer aos meus pais, Angela e Pedro, por todo o apoio e suporte que me deram durante o curso. À minha irmã, Amanda, pela torcida e cumplicidade. Aos meus avôs, avós, primos, primas, tios e tias por sempre estarem torcendo por mim. Aos meus amigos e amigas da faculdade e do colégio, pelo aprendizado e pelo companheirismo. Em especial a Daniela, Gabriela, Mariana e Vitor, que sempre estiveram ao meu lado, pela amizade e pelo apoio demonstrado. À todos os professores que me ajudaram na minha formação. À professora Gabriela, pelos ensinamentos e pelo incentivo na elaboração deste trabalho. Ao meu orientador, professor Guilherme, pela orientação, dedicação e por sempre estar disposto a me auxiliar. Ao Humberto, pelo apoio, carinho e pela nossa sintonia.

Resumo

O objetivo desse trabalho é modelar a incidência da gripe no Brasil no mês t a partir dos dados de incidência e de diversidade genética coletados dos meses anteriores no hemisfério norte. Para tal irá se utilizar os métodos de Granger causalidade e regressão com defasagens. Os modelos propostos podem ser utilizados na previsão da incidência da gripe no Brasil, conhecimento que pode ser estratégico para a implementação de políticas de vacinação pelo governo bem como desenvolvimento de estratégias ótimas de controle de epidemias de gripe.

Palavras-Chave: Gripe, Séries Temporais, Diversidade Genética, Granger causalidade.

Abstract

The objective of this work is to model the incidence of influenza in Brazil in the month t from the data of incidence and genetic diversity collected from previous months in the northern hemisphere. For this purpose, the Granger-causality and regression methods with lags will be used. The proposed models can be used to predict the incidence of influenza in Brazil, knowledge that can be strategic for the implementation of vaccination policies by the government as well as the development of optimal strategies for controlling influenza epidemics.

Keywords: Flu, Time Series, Genetic Diversity, Granger causality.

Sumário

1 Introdução

Nesse estudo deseja-se investigar a dinâmica migratória dos vírus da gripe mais recorrentes, a saber, H1N1 e H3N2 (Influenza A). Existem várias análises que podem ser feitas para modelar esta dinâmica, e neste trabalho serão abordadas as análises de Granger causalidade e de regressão com defasagens, que visam determinar a influência de determinadas variáveis na previsão de outras, bem como permite uma modelagem fina de causalidade no contexto de séries temporais. Neste trabalho serão utilizados dados relacionados à incidência (na verdade uma proxy para a incidência da gripe), que diz respeito ao número de casos de gripe reportados e à diversidade genética da gripe, cuja finalidade é servir de indicativo relativo à diversidade genética das cepas de vírus da gripe em circulação em determinado momento. Espera-se que quanto maior a diversidade genética dos vírus da gripe, maior será a incidência deste.

A literatura não é clara com relação ao epicentro da gripe A no mundo, sendo que a China seria o local mais provável, seguida das regiões tropicais (Rambaut et al., 2008). No entanto, já é conhecido que diversas variantes da gripe que afetam o Brasil migram da Europa, América do Norte e Ásia para o Brasil no movimento migratório dos solstícios. Este movimento, se matematicamente bem descrito, tem o potencial de permitir a obtenção de previsões acuradas para a incidência da gripe no Brasil. Diante disso decidiu-se modelar a incidência da gripe no Brasil através dos valores históricos observados da incidência e também da diversidade genética nas demais regiões.

O primeiro passo do presente trabalho é o estudo da dinâmica de migração intercontinental do vírus da gripe sob a ótica da Granger causalidade. Este método visa determinar o sentido causal entre duas variáveis, estipulando que Y_t Granger causa X_t se os valores passados de Y_t ajudam a prever o valor presente de X_t. Nesse trabalho, utiliza-se o histórico da diversidade genética e da incidência da gripe de diversas regiões, com o intuito de verificar a existência de Granger causalidade nas direções migratórias apropriadas da gripe.

A segunda etapa deste estudo compreende a obtenção de modelos preditivos baseado em regressão com séries defasadas para a incidência da gripe no Brasil a partir dos dados de incidência das demais regiões (como Europa, América do Norte, Ásia e América do Sul), bem como dos dados da diversidade genética dos vírus H1N1 e H3N2. Nesta etapa procura-se determinar o melhor modelo preditivo para a incidência da gripe no Brasil, dentre uma série de modelos candidatos.

O trabalho está organizado em três seções. Na primeira seção começa-se abordando os conceitos de gripe, vírus Influenza e sobre a dinâmica migratória da gripe. Em seguida explica-se com mais detalhes as metodologias de Granger causalidade

e de regressão com defasagens. A segunda seção esclarece como foi feita a coleta, a limpeza e a organização dos dados. A terceira seção apresenta os resultados das análises de Granger causalidade e de regressão com defasagens, junto com uma análise descritiva dos dados.

2 Gripe

Uma das doenças mais frequentes no Brasil e no mundo é a gripe comum. A gripe se caracteriza por uma infecção aguda do sistema respiratório e é provocada por diversos vírus RNA. Vírus de RNA são aqueles que têm RNA como material genético e que são mais propensos a sofrer mutações genéticas. Sabe-se que todo ano há diversos casos registrados e até óbitos por influenza, o tipo mais comum de vírus da gripe (Rambaut et al., 2008). Segundo Barr et al. (2010), a vacinação contra a gripe é a medida mais eficaz para a sua prevenção.

2.1 O Vírus Influenza

Todos os anos, cerca de 10% da população mundial contrai o vírus influenza, sendo esse responsável por entre 250.000 a 500.000 mortes anualmente, e seus sintomas mais comuns são tosse, febre, dores de cabeça, de garganta e dores musculares (Eccles, 2005; Rambaut et al., 2008). Existem três tipos comuns de vírus, Influenza A, B e C, sendo que os dois primeiros são responsáveis por epidemias sazonais. Dos três tipos de vírus, o tipo A é o que tem mais capacidade de se multiplicar num organismo e grande parte dos seus casos ocorre no inverno e em países de clima temperado. Segundo Rambaut et al. (2008), a dinâmica evolucionária do vírus influenza A é formada por uma rápida mutação, seleção natural e rearranjo frequente.

De acordo com Forleo-Neto et al. (2003), a Influenza A se divide em subtipos de hemaglutininas (H1, H2 e H3) e de neuraminidases (N1, N2) que são proteínas presentes na superfície do vírus. Nesse estudo deseja-se investigar melhor o comportamento dos vírus da gripe H1N1 (hemaglutinina H1 e neuraminidase N1) e H3N2 (hemaglutinina H3 e neuraminidase N2) separadamente, que são os mais recorrentes.

2.1.1 H1N1

O subtipo H1N1 surgiu em 1918 causando uma pandemia conhecida como Gripe Espanhola, uma das pandemias mais mortais da história, tendo afetado cerca de 1/4 da população mundial e sendo responsável por dezenas de milhões mortes (Garten et al., 2009). O vírus da gripe H1N1 reapareceu em 1977 e desde então as suas epidemias apontam menores taxas de mortalidade quando comparadas com as epidemias da H3N2 (Rambaut et al., 2008). Apesar disso, em 2009 ocorreu a pandemia do subtipo H1N1, onde os primeiros surtos manifestaram-se no México, espalhando-se pelo mundo (Rambaut e Holmes, 2009). Segundo Silva (2015), após o ano de 2009, o vírus H1N1 continua circulando no Brasil, caracterizando-se por gerar epidemias

anuais sazonais com altas taxas de mortalidade. Os novos grupos filogenéticos (de origem) do vírus H1N1 dessas epidemias anuais aparentam iniciar em países do hemisfério norte, indo para o Brasil apenas na epidemia do ano seguinte.

2.1.2 H3N2

O subtipo H3N2 surgiu em 1968 como sendo a terceira pandemia do século XX chamada de Gripe de Hong Kong e tem dominado as epidemias sazonais do vírus influenza A nos últimos anos (Ibiapina et al., 2005). Segundo estudo Kaji et al. (2003), há diferenças nos sintomas das gripes H1N1 e H3N2 entre as quais, podemos citar, o fato da febre ser maior no vírus do subtipo H3N2 quando comparado ao grupo H1N1. Atualmente as linhagens do subtipo H3N2 chegam no Brasil desde países vizinhos da América do Sul e entram no país principalmente pela região Sudeste (Born, 2013).

2.1.3 Dinâmica Migratória da Gripe

De acordo com Rambaut et al. (2008), a China tem sido o lugar proposto como o epicentro do vírus da gripe A (H1N1 e H3N2). O mesmo também fala na possibilidade das regiões tropicais serem as populações de origem de novas mutações sazonais, ou seja, a diversidade genética é gerada nessa população de origem, os trópicos, e depois avança para os hemisférios norte e sul.

Conforme Born (2013), quando se fala do vírus da gripe H3N2 compreende-se que as epidemias de gripe no hemisfério norte ocorrem entre os meses de novembro e março e no hemisfério sul de maio a setembro. Ademais foi possível compreender que as cepas são primeiramente espalhadas para a América do Norte e Europa e só depois para a América do Sul (Russell et al., 2008). Segundo estudos, Hong Kong é o lugar com maior probabilidade de ascendência das novas variantes disseminadas e o Brasil o local com menor probabilidade, onde a sua principal porta de entrada do vírus da gripe H3N2 seria a região Sudeste seguida pelas regiões Sul e Nordeste (Born, 2013). Portanto, seria interessante desenvolver e atualizar vacinas na América do Sul com cepas da Europa e dos Estados Unidos de estações sazonais anteriores.

2.2 Granger Causalidade

Nesse estudo será utilizado o método de Granger causalidade para estudar a dinâmica migratória da gripe (Granger, 1969). Este método tem como objetivo determinar o sentido causal entre duas variáveis, estipulando que X_t Granger causa Y_t se os valores passados de X_t ajudam a prever o valor presente de Y_t, podendo fornecer resultados melhores que considerando apenas o passado de Y_t. Mais especificamente, é uma maneira de verificar se uma série temporal ajuda na previsão de outra série através da modelagem VAR. Para utilizar este método as séries precisam ser equiparadas.

Na estatística, matemática e econometria, uma série temporal é um segmento de uma das possíveis trajetórias de um processo estocástico $\{X_t\}_{t\in\mathbb{Z}}$. Informalmente, uma série temporal de tamanho n, digamos $\{x_t\}_{t=1}^{n}$, é um conjunto de valores observados ao longo do tempo.

2.2.1 Técnica Utilizada

Nesta seção serão abordadas as técnicas estatísticas que estão por trás da metodologia de Granger causalidade.

Modelo de Vetores Autorregressivos (VAR)

O modelo de Vetores Autorregressivos (VAR) é uma extensão dos modelos Autorregressivos (AR) e tem como um dos seus objetivos prever o comportamento futuro de séries temporais inter-relacionadas (Sims, 1980). A equação definida pelo VAR determina uma variável em variáveis defasadas de si própria e de outras variáveis que estão no modelo.

Um processo estocástico k-dimensional X_t é dito ser um processo VAR(p) se puder ser escrito como

$$\mathbf{Y}_t = c + A_1\mathbf{Y}_{t-1} + A_2\mathbf{Y}_{t-2} + \cdots + A_p\mathbf{Y}_{t-p} + \boldsymbol{\varepsilon}_t,$$

onde c é um vetor k de constantes (interceptos), $A_1, \cdots, A_p$ são matrizes $k \times k$ e $\boldsymbol{\varepsilon_t}$ um vetor k-dimensional de erros.

O modelo de previsão de séries temporais inter-relacionadas Y_t e X_t é dado por

$$\mathbf{Y}_t = c + \sum_{i=1}^{s} B_i\mathbf{Y}_{t-i} + \sum_{j=1}^{m} A_j\mathbf{X}_{t-j} + \boldsymbol{\varepsilon}_t,$$

onde $B_1, \cdots, B_s$ e $A_1, \cdots, A_m$ são matrizes constantes que relacionam valores passados das variáveis dependentes e das variáveis independentes e $\boldsymbol{\varepsilon_t}$ é um ruído branco k-dimensional.

Granger causalidade

A ideia por trás de Granger causalidade (para séries temporais univariadas) é considerar o modelo

$$Y_t = \beta_0 + \sum_{i=1}^{k} \beta_i Y_{t-i} + \sum_{j=1}^{m} \alpha_j X_{t-j} + \boldsymbol{\varepsilon}_t$$

onde $\boldsymbol{\varepsilon}_t$ denota o ruído branco. Dizemos que X_t Granger causa Y_t se os valores passados de X_t ajudam a prever o valor presente de Y_t. Para testar se X_t Granger-causa Y_t considere o seguinte teste

$$H_0 : \alpha_1 = \cdots = \alpha_m = 0 \quad \text{vs.} \quad H_1 : \alpha_s \neq 0, \text{ para pelo menos um } s \in \{1, \cdots, m\}.$$

No teste acima, com a rejeição da hipótese nula pode-se concluir que X_t Granger-causa Y_t. Semelhantemente, considerando o modelo dado por

$$X_t = \beta_0 + \sum_{i=1}^{k} \beta_i Y_{t-i} + \sum_{j=1}^{m} \alpha_j X_{t-j} + \boldsymbol{\varepsilon}_t$$

Para testar se Y_t Granger-causa X_t considere o seguinte teste

$$H_0 : \beta_1 = \cdots = \beta_k = 0 \quad \text{vs.} \quad H_1 : \beta_s \neq 0, \text{ para pelo menos um } s \in \{1, \cdots, k\}.$$

No teste acima, com a rejeição da hipótese nula pode-se concluir que Y_t Granger-causa X_t.

2.2.2 Granger Causalidade e Estacionariedade

Antes de aplicar o método de Granger causalidade, precisa-se verificar se as séries são ou não-estacionárias, e para isso são utilizados testes como o de Phillips-Perron (Phillips e Perron, 1988). Nesse teste a hipótese nula é de que a série apresenta pelo menos uma raíz unitária (série não estacionária) e a hipótese alternativa é a ausência de raiz unitária. Desse modo, a série será considerada estacionária se rejeitar-se a hipótese nula. Uma análise gráfica preliminar pode auxiliar na decisão da aplicação ou não de testes de estacionariedade. A ausência de tendências determinísticas visíveis e/ou sazonalidades aparentes são indícios de estacionariedades. Entretanto, não são suficientes para tomada de decisão, que, preferencialmente, deve ser feita através de um teste apropriado como o Phillips-Perron ou ADF (Dickey e Fuller, 1981).

Granger causalidade para séries estacionárias

No método de Granger causalidade, caso ambas séries forem estacionárias, deve-se seguir as seguintes etapas:

1. Verificar se as séries temporais cointegram utilizando o teste de cointegração de Phillips-Ouliaris (Phillips e Ouliaris, 1990).

2. Ajustar um modelo VAR(p), onde p é o número de defasagens. Esse número de defasagens pode ser escolhido utilizando métodos usuais.

3. Aplicar o teste de Granger causalidade no modelo VAR definido anteriormente. Nele precisa-se declarar qual variável acredita-se que Granger causa a outra. A rejeição da hipótese nula indica a existência de Granger causalidade.

Granger causalidade para séries não-estacionárias

Uma forma de aplicar o método de Granger causalidade caso as séries não forem estacionárias é utilizando o procedimento de Toda e Yamamoto, introduzido por Toda e Yamamoto (1995), e que compreende os seguintes passos:

1. Verificar se as séries cointegram. Duas séries cointegram se possuem a mesma ordem de integração, digamos m, e se o resíduo da regressão de uma série pela outra for estacionário, o que pode ser determinado utilizando-se testes como o de Phillips-Perron.

2. Ajustar um modelo VAR(p).

3. Para aplicar o Teste de Wald, precisa-se ajustar um modelo VAR($p+m$) aos dados. Este modelo certamente apresentará várias variáveis não significativas, dados os passos anteriores, mas isto não é problema visto que este modelo não será utilizado diretamente – é apenas um dispositivo para garantir a teoria assintótica. Através da rejeição da hipótese nula pode-se concluir a existência de Granger causalidade na direção testada.

2.2.3 Aplicações de Granger Causalidade

O método de Granger causalidade apresenta diversas áreas de aplicabilidade e é bastante utilizado principalmente no campo da economia. Diversos artigos podem ser encontrados empregando essa teoria, com o objetivo de analisar a relação entre as variáveis. Alguns exemplos de aplicações são dados abaixo.

Na área da economia estudou-se a causalidade entre as principais bolsas de valores do mundo, mostrando como os mercados se comportam entre si e analisando se um mercado apresenta forte influência sobre os demais (Farias e Sáfadi, 2010). Os principais resultados indicaram que o mercado brasileiro apresenta forte influência sobre os mercados chinês e russo, mas o contrário não ocorre.

Na área da agronomia é de interesse estudar se determinadas variáveis agropecuárias e socio-econômicas (como rebanho bovino e densidade demográfica) Granger causam o desmatamento na Amazônia (Diniz et al., 2009). As análises apresentaram resultados de que há uma relação causal entre o desmatamento e as variáveis mencionadas anteriormente.

De acordo com Dantas e Weydmann (2009), na área econômica, uma questão importante é o estudo da existência de uma relação entre os preços externos e internos da carne de frango. A pesquisa apontou que há essa relação e com isso percebe-se que o comportamento do preço externo do frango pode auxiliar no planejamento da produção.

Na área biológica investigou-se a relação causal entre os casos de gripe em humanos e a poluição do ar no Taiwan (Chen et al., 2018). No estudo buscou-se a relação entre os casos de influenza semanais e o PM2.5 (material particulado) acumulado. Os resultados indicaram que a poluição Granger causa os casos de gripe no grupo de idosos (acima de 64 anos).

2.3 Modelos de Regressão

A segunda metodologia estatística escolhida para este estudo é a regressão com defasagens. Para a seleção dos modelos utilizaremos três técnicas distintas, resumidas a seguir.

2.3.1 Seleção de Variáveis via Stepwise

Regressão Stepwise é um método utilizado na construção de modelos que tem como objetivo escolher variáveis preditas através de um algoritmo que verifica a importância delas (Hastie et al., 2009). Os tipos de seleção de variáveis mais conhecidos são a seleção Forward-Stepwise e a seleção Backward-Stepwise. Na seleção do tipo Forward-Stepwise, iniciamos com o modelo mais simples possível, contando apenas com o intercepto, e a cada passo adicionamos um preditor por vez, até que um certo critério de parada seja alcançado. Neste trabalho utilizaremos os critérios do p-valor e do AIC (Akaike, 1987), detalhados adiante. Um ponto positivo deste método é que ele pode ser aplicado mesmo quando o número de preditores for muito maior que o tamanho amostral. A seleção Backward inicia com todos os preditores do modelo, e em seguida remove iterativamente o preditor que tem o menor impacto no ajuste, de acordo com algum critério, um de cada vez. Este procedimento é iterado até que algum critério de parada previamente estipulado seja atingido.

Esta seleção só pode ser utilizada quando o número de preditores for menor que o tamanho amostral. Neste trabalho utilizou-se a seleção Backward com dois critérios de parada: o critério baseado no AIC e o p-valor.

O critério de informação de Akaike, conhecido como AIC se baseia na função de log-verossimilhança e é definida pela equação

$$AIC = -2\ell(\boldsymbol{\theta}) + 2p$$

onde $\ell(\boldsymbol{\theta})$ é a função de log-verossimilhança e p é o número de parâmetros do modelo (Akaike, 1987). O critério AIC tem como objetivo encontrar um balanço entre parcimônia e complexidade do modelo. Ou seja, é um método que procura um balanço entre quão bem um determinado modelo explica o comportamento da variável resposta e o número de parâmetros que este modelo contém. Dentre vários modelos competidores, selecionamos o modelo com o menor valor do AIC. Computacionalmente, esta técnica está implementada através da função `step` do pacote `stats` da versão 4.0.0 do software R (R Core Team, 2020).

Outro método abordado foi Stepwise com o critério de p-valor, que seleciona as variáveis de acordo com as suas estatísticas de Wald, adicionando os termos significativos e eliminando os termos não significativos. Este procedimento ocorre até que uma regra de parada seja alcançada, como por exemplo, quando todas as variáveis tiverem um p-valor abaixo de algum limite escolhido.

2.3.2 LASSO

A Regressão LASSO é um método de penalização que tem como objetivo fornecer modelos menores e mais parcimoniosos (Hastie et al., 2009). A penalização é aplicada nos coeficientes para diminuir o número de parâmetros e, consequentemente, reduzir a incerteza do modelo.

Para ficar mais fácil de compreender a Regressão LASSO, vamos primeiro recordar a regressão de uma forma resumida. Na regressão quando utiliza-se o método de mínimos quadrados ordinários (MQO), visa-se encontrar o melhor ajuste para um determinado conjunto de dados procurando minimizar a soma dos quadrados das diferenças entre os valores observados e estimados. Todavia, para utilizar o procedimento de MQO é necessário que o erro seja distribuído aleatoriamente e siga uma distribuição normal. Na regressão estamos procurando $\boldsymbol{\beta}$ que satisfaça

$$\hat{\boldsymbol{\beta}}^{MQO} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \sum_{j=1}^{n} \Big[y_j - \Big(\beta_0 + \beta_1 x_{j1} + \cdots + \beta_{p-1} x_{j(p-1)} \Big) \Big]^2 \right\},$$

onde y_j é a variável resposta que queremos predizer, x_{j1}, x_{j2},..., $x_{j(p-1)}$ são as variáveis preditoras, β_0, β_1,..., $\beta_{(p-1)}$ são os coeficientes ou parâmetros do modelo, n é o tamanho da amostra e p é a quantidade de parâmetros.

A ideia dos métodos de penalização é procurar $\hat{\boldsymbol{\beta}}$ tal que

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \sum_{j=1}^{n} \Big[y_j - \Big(\beta_0 + \beta_1 x_{j1} + \cdots + \beta_{p-1} x_{j(p-1)} \Big) \Big]^2 + \lambda g(\boldsymbol{\beta}) \right\},$$

onde $g : \mathbb{R}^p \to [0, \infty)$ representa a função de penalidade (shrinkage penalty) e $\lambda > 0$ controla o impacto dessa função. Este último termo tem como objetivo provocar uma redução nos valores dos β_i's.

Quando a função de penalização g é dada por $g(\boldsymbol{\beta}) = \sum_{j=0}^{p-1} |\beta_j|$ temos o método LASSO. Neste caso a estimativa via LASSO para $\boldsymbol{\beta}$ é obtida através do seguinte problema de optimização:

$$\hat{\boldsymbol{\beta}}^L = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \sum_{j=1}^{n} \Big[y_j - \Big(\beta_0 + \beta_1 x_{j1} + \cdots + \beta_{p-1} x_{j(p-1)} \Big) \Big]^2 + \lambda \sum_{j=0}^{p-1} |\beta_j| \right\}$$

onde o último termo da equação $\lambda \sum_{j=1}^{p-1} |\beta_j|$ visa penalizar os valores dos coeficientes β_j. A ideia dessa penalização é tornar os parâmetros β_j tão pequenos ou até mesmo nulos, para que possamos eliminá-los do modelo, assim eliminando o atributo e reduzindo a dimensionalidade do modelo. Ademais o parâmetro λ controla a "redução" do modelo, e deve ser escolhido de modo a controlar o balanço variância-viés. No software `R` (R Core Team, 2020) este método está implementado no pacote `glmnet`.

3 Dados da Gripe

Este capítulo descreve o banco de dados, apresentando detalhes sobre a coleta dos dados, em quais fontes eles foram obtidos, e quais métodos estatísticos foram utilizados na limpeza, organização e formatação dos dados.

3.1 Número de Casos Positivos da Gripe

Os dados relativos ao número de casos positivos da gripe foram retirados da Flunet, uma ferramenta on-line mantida pela Organização Mundial da Saúde (WHO, 2020). Nela cada país reporta semanalmente o número de casos testados, o número de casos positivos e qual o tipo de vírus. Tipicamente os dados reportados são referentes aos dados coletados em centros de referência em cada país. Desta forma, os dados que utilizamos não são exatamente os dados de incidência da gripe, mas sim os dados reportados pelos centros de referência em cada país que são utilizados como proxy para a incidência da gripe. De qualquer forma, neste trabalho, tais dados serão referenciados como dados de incidência da gripe. Nesse trabalho foram coletados os dados das gripes H1N1 e H3N2 no intervalo de janeiro de 2008 até novembro de 2019, mas devido à problemas de dados faltantes em 2008, os dados utilizados na análise contemplam o período de outubro de 2008 até novembro de 2019.

Na Flunet os dados podem ser agregados por região ou por país. Por simplicidade consideramos os dados agregados por região, com exceção do Brasil. Coletou-se dos seguintes locais: Brasil, Região da América do Norte, Região da América do Sul, Região da América Central, Região Europeia, Região do Sul da Ásia e Região do Pacífico Ocidental. Um detalhe é que no banco de dados, cada região é composta por um determinado número de países, responsáveis por reportar seus dados. Porém, devido à características locais, os dados relativos à diversos países apresentavam uma quantidade alta de dados faltantes, muitas vezes acima dos 50%, resultando em banco de dados locais inúteis para nossos propósitos. Para viabilizar a análise, todos os países apresentando uma quantidade superior a 50% de dados faltantes foram excluídos na construção do banco de dados da respectiva região.

Após a utilização desse critério, o banco de dados relativo à Região Europeia foi constituído utilizando-se os dados dos seguintes países: Bélgica, Dinamarca, Espanha, Estônia, Alemanha, Irlanda, Israel, Itália, Letônia, Holanda, Noruega, Polônia, Federação Russa, Eslovênia, Suécia, Suíça, Turquia e Reino Unido da Grã-Bretanha e Irlanda do Norte. A Região da América do Norte ficou constituída por: Canadá e Estados Unidos. A Região da América do Sul continha: Argentina, Bolívia, Chile,

Colômbia, Equador, Guiana Francesa, Paraguai e Peru. A Região da América Central continha: Costa Rica, Cuba, República Dominicana, El Salvador, Guatemala, Honduras, Jamaica, México, Nicarágua e Panamá. A Região do Sul da Ásia apresentava: Bangladesh, Butão, Índia, Indonésia, Nepal, Sri Lanka e Tailândia. Por fim, a Região do Pacífico Ocidental ficou constituída por: Austrália, Camboja, China, Japão, República Democrática Popular do Laos, Malásia, Nova Caledônia, Filipinas, República da Coreia, Cingapura e Vietnã.

As técnicas de séries temporais que utilizaremos requerem que os dados não contenham dados faltantes. Para resolver isso utilizou-se o método de imputação, que visa preencher os dados faltantes através do seguinte critério: multiplicar a média do número de casos positivos da gripe da respectiva semana pela proporção que aquele país representa de sua região. Em alguns casos, porém, ocorreu que a semana tinha dados faltantes em todos os países, e resolveu-se isso imputando-a através da média entre a semana anterior e posterior.

Após feita a imputação, o banco de dados continha o número de casos positivos da gripe registrado semanalmente por cada país dentro de cada região. Todavia, o fato de cada região possuir vários países reportando as suas incidências da gripe complicava bastante a análise, e em vista disto, os dados foram agregados por região. Esse agrupamento por região pode ser visto na Figura 3.1. Outra modificação feita foi agregar os dados de semanais para mensais.

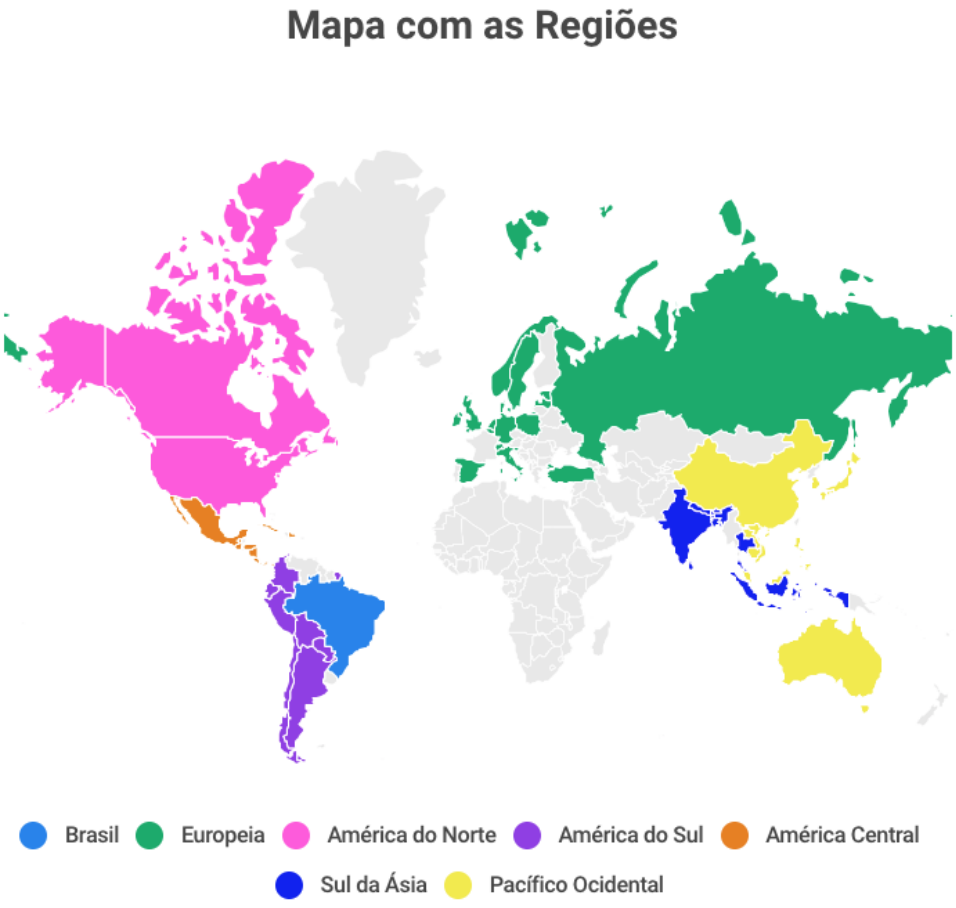


Figura 3.1: Mapa com as diferentes regiões.

3.2 Diversidade Genética

Para a montagem da diversidade genética[1], coletou-se os dados das sequências genéticas, no site da NCBI Influenza Virus Database, uma instituição que fornece acesso à informações biomédicas e genômicas, e que procura sequenciar rapidamente os vírus da gripe de amostras coletadas (NCBI, 2020). O site disponibiliza gratuitamente este material para que pesquisadores utilizem os dados em seus estudos para analisar o comportamento dos vírus, comparando as suas cepas e identificando mais rapidamente novas variantes, auxiliando, por exemplo, no desenvolvimento e atualização de vacinas. Os dados consolidados pelo NCBI são fruto de uma colaboração de diversos pesquisadores, que buscam isolar o vírus para registrar a sua sequência genética.

O banco de dados deste trabalho contém as sequências completas do cromossomo 4 (gene H4), do vírus Influenza A, cujo hospedeiro é o ser humano, de todos os continentes e no intervalo de outubro de 2008 até setembro de 2019. Como os dados estão relacionados às sequências genéticas, então analisou-se os subtipos H1N1 e H3N2 separadamente.

Diversidade genética é uma medida populacional que procura quantificar diferenças entre vírus. Com ela torna-se possível comparar a variedade genética entre subtipos (H1N1 e H3N2) e dentro das populações, focando, por exemplo, em mutações. Esta medida é baseada no número de diferenças de nucleotídeos (responsáveis pela formação dos códigos genéticos) entre as sequências. Esta medida é obtida como a distância entre todos os indivíduos da população. A distância genética é simplesmente uma distância entre dois vírus. A utilização de distâncias genéticas depende da comparação, gene a gene, das sequências de RNA de diferentes vírus. Todavia, estas sequências nem sempre estão "alinhadas", devido à possíveis mutações que ocorrem em consequência de substituições, inserções ou deleções de nucleotídeos. Para que essa comparação entre sequências possa ser feita, existem várias ferramentas que permitem o alinhamento de sequências genéticas. Neste trabalho optou-se por utilizar a ferramenta on-line MAFFT (Yamada et al., 2016).

O próximo passo após o alinhamento das sequências é a criação de uma matriz em que cada entrada contém a distância genética entre duas amostras, calculadas de acordo com uma métrica que veremos adiante. Em seguida combina-se estas medidas (as distâncias) com os períodos em que os vírus foram isolados, obtendo assim um comportamento temporal das distâncias. Em seguida, conforme sugerido em Jesus (2018), a diversidade do vírus é avaliada ao longo do tempo usando o esquema de médias móveis trimestrais.

A matriz de distâncias foi calculada através da distância entre todas as sequências presente no banco de dados, duas a duas, usando a função `dist.dna` do pacote `ape` do software `R`. Este cálculo resultou numa matriz $n \times n$, simétrica, $\mathbf{D}$, com entradas $[\mathbf{D}]_{i,j} = d_{i,j}$ onde $d_{i,j}$ denota a distância genética entre as sequências i e j, onde $i = 1, 2, \ldots, n$ e $j = 1, 2, \ldots, n$, isto é,

[1] Diversidade genética é uma medida que calcula quanto uma sequência genética se diferencia de outra sequência. Esta medida é explicada com mais detalhes nos próximos parágrafos.

$$\mathbf{D} = \begin{pmatrix} 0 & d_{12} & d_{13} & \cdots & d_{1n} \\ d_{21} & 0 & d_{23} & \cdots & d_{2n} \\ d_{31} & d_{32} & 0 & \cdots & d_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & d_{n3} & \cdots & 0 \end{pmatrix}.$$

A matriz **D** é chamada de matriz de distâncias. Observa-se que os elementos da diagonal são zeros, pois, naturalmente, quando $i = j$ temos $d_{ij} = 0$.

Neste trabalho a matriz **D** foi utilizada para calcular a diversidade genética do vírus através da média das distâncias, no esquema de janela deslizante com k = 3, no período de interesse do estudo. Fez-se este cálculo no intervalo de outubro de 2008 até setembro de 2019, separadamente para H1N1 e H3N2, e para seguintes regiões: Ásia, América do Norte e global (todos os continentes).

4 Resultados

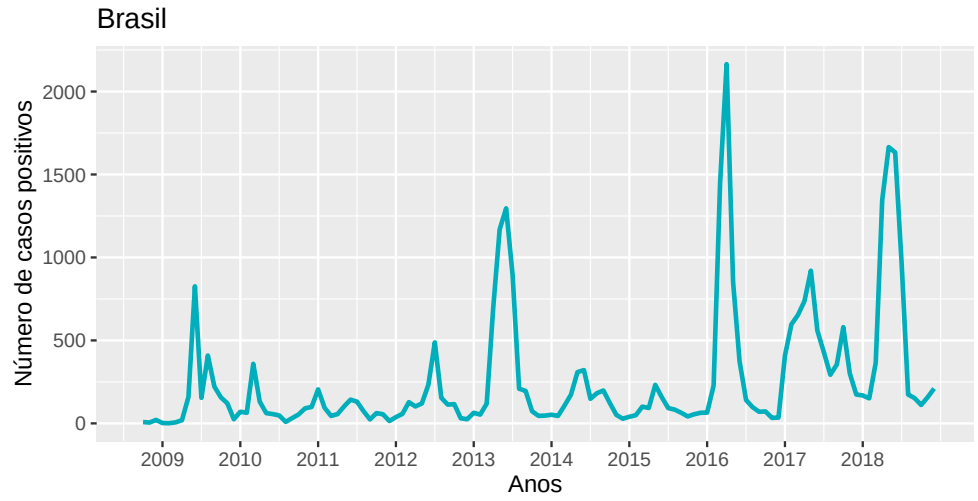
Neste capítulo são apresentados em detalhes os resultados das análises descritas nas Seções 2.2 e 2.3.

4.1 Análise Exploratória

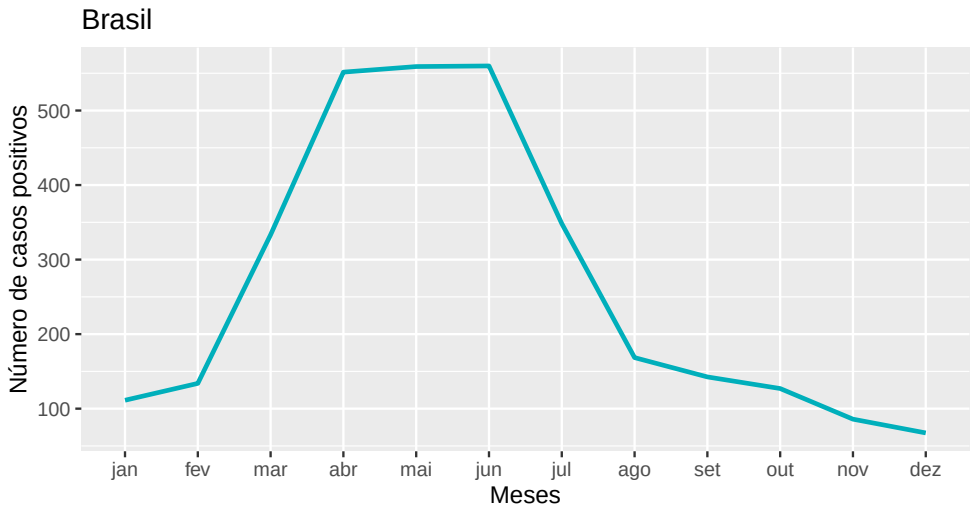
Esta seção apresenta uma análise descritiva dos dados relativos à incidência da gripe e dos dados relacionados à diversidade genética.

4.1.1 Número de Casos Positivos da Gripe

Nesta subseção o objetivo é analisar o comportamento dos dados relacionados à incidência da gripe. A exploração os dados será através da visualização dos respectivos gráficos: série temporal da incidência da gripe, média mensal da incidência da gripe e função de autocorrelação (ACF). Cada um desses três gráficos é apresentado sete vezes, um para cada região. Conclui-se a análise descritiva com uma matriz de correlação das incidências.



(a) Série temporal da incidência da gripe no Brasil.

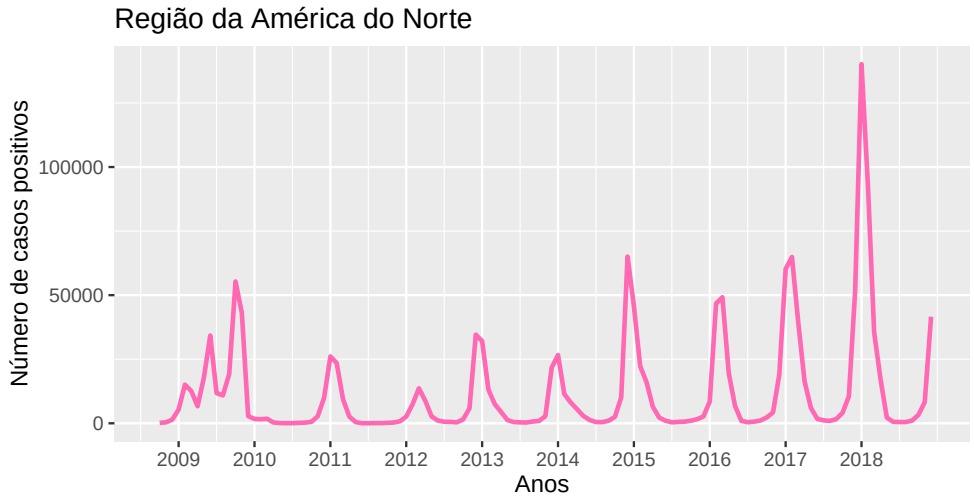


(b) Média mensal da incidência da gripe no Brasil.

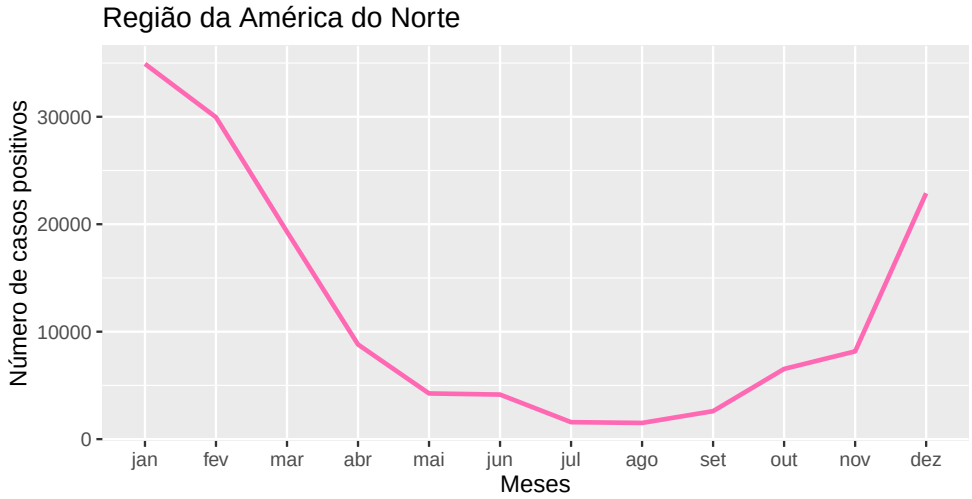
Figura 4.1: Gráficos da incidência da gripe no Brasil (10/2008 − 12/2018).

A Figura 4.1(a) ilustra a série temporal do número de casos positivos da gripe no Brasil. Através de uma análise gráfica percebe-se que a série não aparenta ser estacionária, pois a variável não se comporta de forma aleatória ao longo dos meses. A série não parece apresentar uma tendência, pois não há um padrão de crescimento/decrescimento ao longo do tempo. A sazonalidade vai ser verificada com o gráfico da função de autocorrelação da série.

A Figura 4.1(b) apresenta o número de casos positivos médios a cada mês no Brasil. Observa-se que os meses de abril, maio e junho (próximos ao inverno) são aqueles que apresentam uma maior incidência de gripe. Ademais os meses que apresentam um menor número de casos de gripe são novembro, dezembro e janeiro (época próxima ao verão).



(a) Série temporal da incidência da gripe na Região da América do Norte.

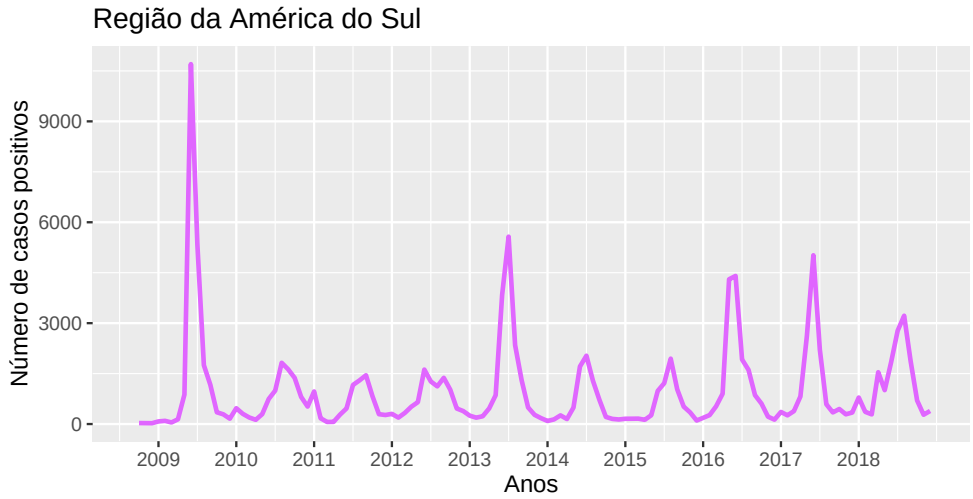


(b) Média mensal da incidência da gripe na Região da América do Norte.

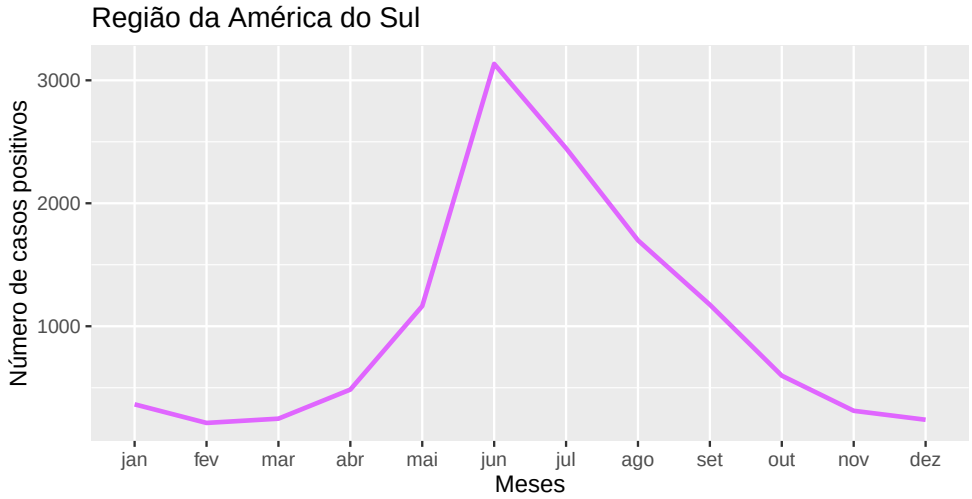
Figura 4.2: Gráficos da incidência da gripe na Região da América do Norte (10/2008 − 12/2018).

A Figura 4.2(a) apresenta a série temporal do número de casos positivos da gripe na Região da América do Norte. A série apresenta sazonalidade, pois há padrões de comportamento que se repetem, e no caso dessa série há diversos picos anuais que ocorrem no início/fim do ano (época do inverno), onde a incidência da gripe aumenta. A série não parece ser estacionária, pois a média e a variância não são constantes ao longo do tempo. O gráfico não parece apresentar uma tendência, pois não há um comportamento de crescimento/decrescimento ao longo dos meses.

A Figura 4.2(b) ilustra o número de casos positivos médios a cada mês na Região da América do Norte. Percebe-se que os meses que representam o inverno no hemisfério norte (dezembro, janeiro, fevereiro e março) são aqueles que apresentam uma maior incidência de gripe, com no mínimo 20.000 casos.



(a) Série temporal da incidência da gripe na Região da América do Sul.

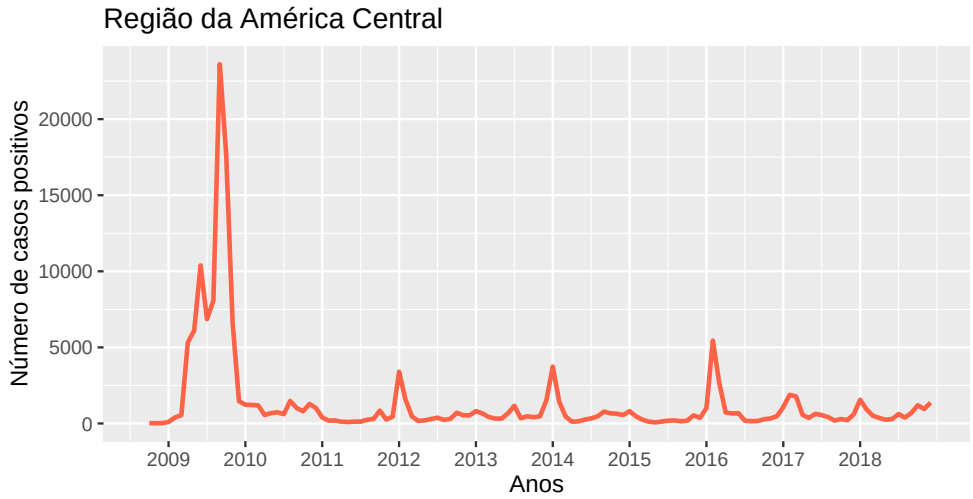


(b) Média mensal da incidência da gripe na Região da América do Sul.

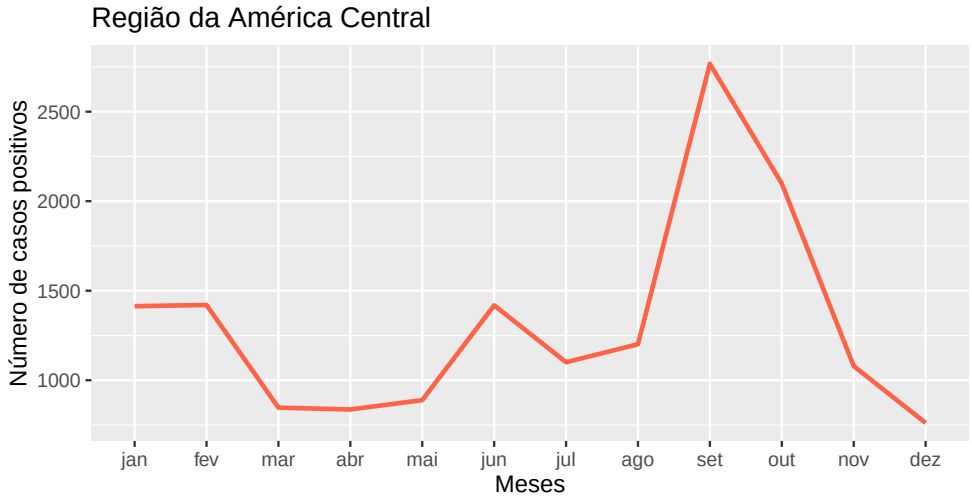
Figura 4.3: Gráficos da incidência da gripe na Região da América do Sul (10/2008 − 12/2018).

A Figura 4.3(a) apresenta a série temporal do número de casos positivos da gripe na Região da América do Sul. Como as demais, a série temporal apresenta uma notável sazonalidade anual, sendo, portanto, não estacionária.

A Figura 4.3(b) apresenta o número de casos positivos médios a cada mês na Região da América do Sul. Observa-se que conforme o inverno vai chegando no hemisfério sul, o número de casos vai aumentando. Os meses que apresentam uma maior incidência de gripe são junho, julho e agosto (época de inverno), onde o pico deles é o mês de junho, apresentando uma média maior que 3.000 casos.



(a) Série temporal da incidência da gripe na Região da América Central.

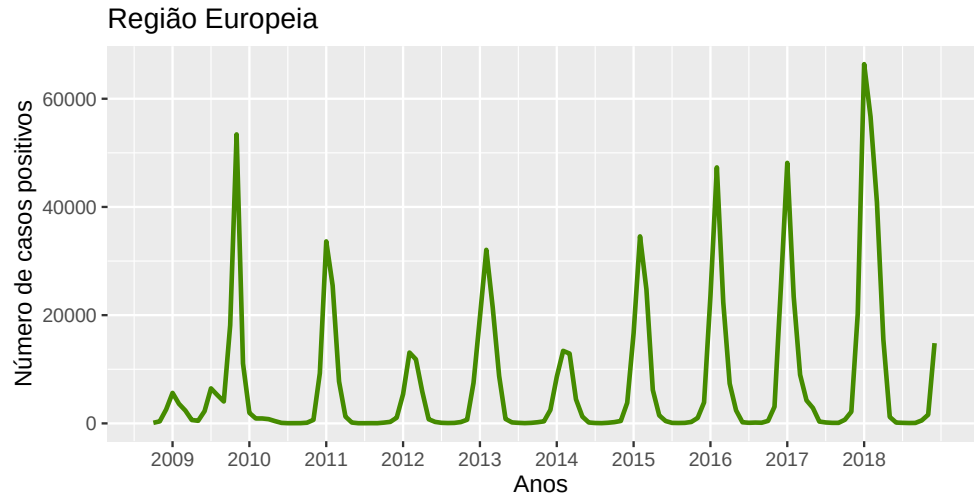


(b) Média mensal da incidência da gripe na Região da América Central.

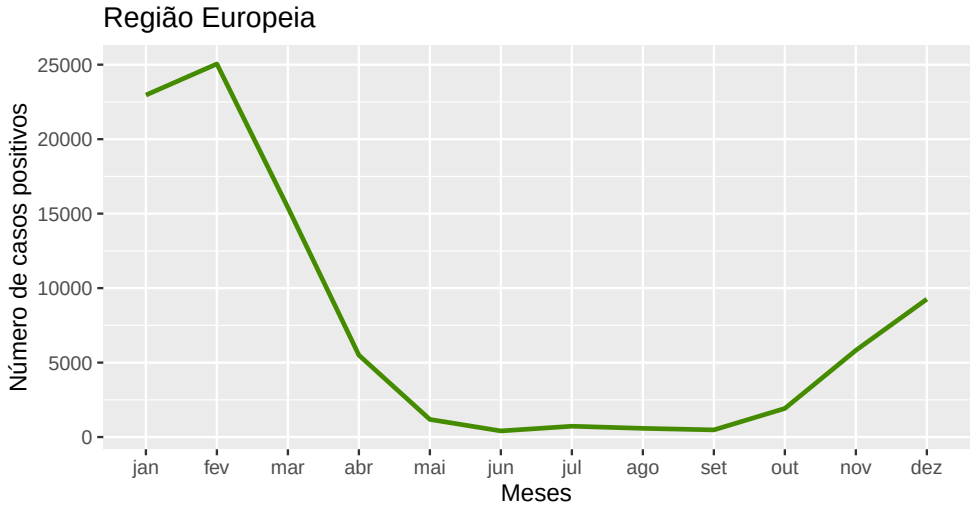
Figura 4.4: Gráficos da incidência da gripe na Região da América Central (10/2008 − 12/2018).

A Figura 4.4(a) apresenta a série temporal do número de casos positivos da gripe na Região da América Central. Observa-se que entre os anos de 2009 e 2010 a incidência da gripe é muito maior que os demais anos. A gripe suína (H1N1) é a principal justificativa desse crescimento, visto que o México é tido como epicentro desta epidemia. Ademais a série apresenta uma sazonalidade anual.

A Figura 4.4(b) apresenta o número de casos positivos médios a cada mês na Região da América Central. O gráfico apresenta um pico da incidência da gripe no mês de setembro, com uma média de 2.750 casos, mas sem apresentar um comportamento sazonal.



(a) Série temporal da incidência da gripe na Região Europeia.

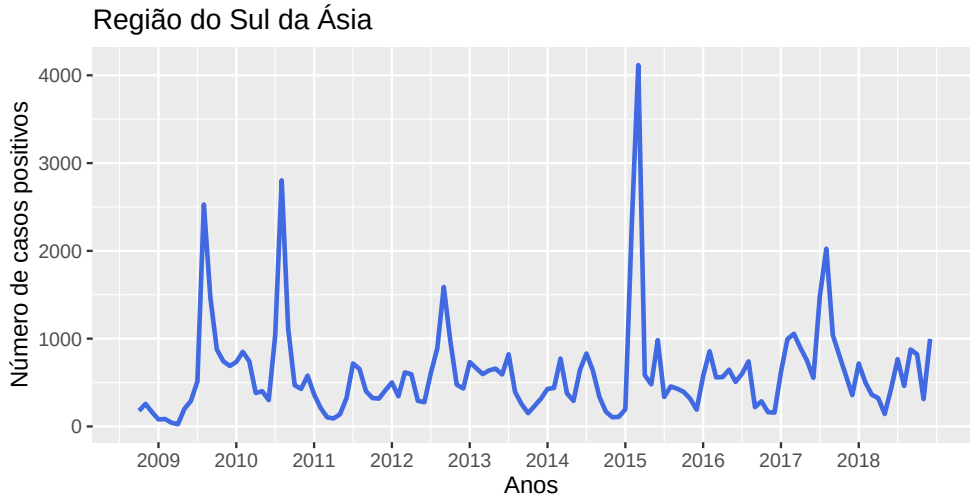


(b) Média mensal da incidência da gripe na Região Europeia.

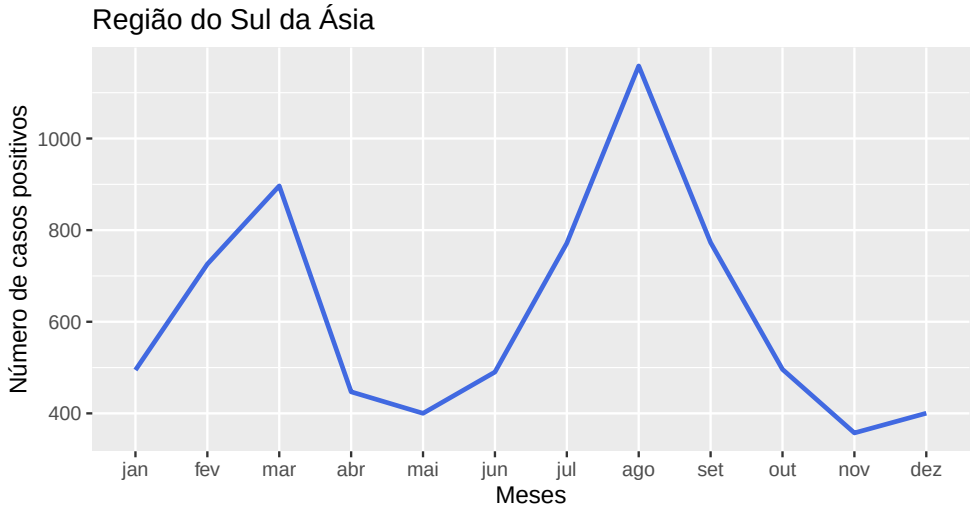
Figura 4.5: Gráficos da incidência da gripe na Região Europeia (10/2008 − 12/2018).

A Figura 4.5(a) apresenta a série temporal do número de casos positivos da gripe na Região Europeia. Analogamente aos casos anteriores, a série apresenta uma clara sazonalidade anual e não apresenta outras tendências determinísticas perceptíveis.

A Figura 4.5(b) ilustra o número de casos positivos médios a cada mês na Região Europeia. Este gráfico apresenta um comportamento semelhante ao da Região da América do Norte, pois os meses que apresentam uma maior incidência de gripe são dezembro, janeiro, fevereiro e março (época de inverno). Observa-se que o pico na Região Europeia encontra-se no mês de fevereiro, com uma média de 25000 casos.



(a) Série temporal da incidência da gripe na Região do Sul da Ásia.

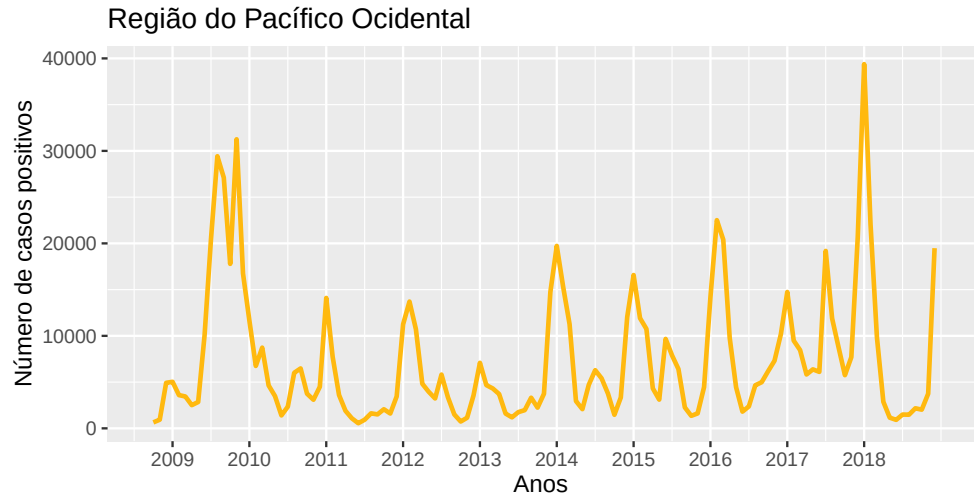


(b) Média mensal da incidência da gripe na Região do Sul da Ásia.

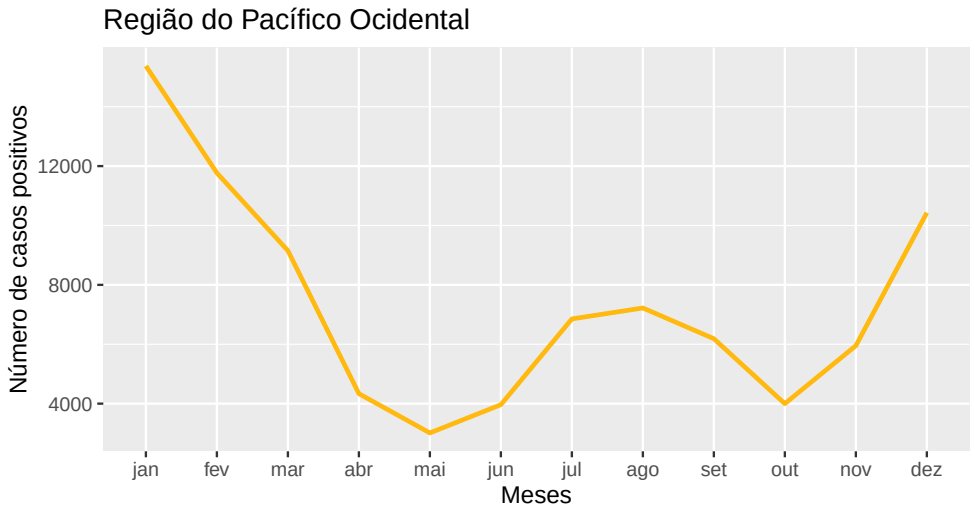
Figura 4.6: Gráficos da incidência da gripe na Região do Sul da Ásia. (10/2008 – 12/2018).

A Figura 4.6(a) apresenta a série temporal do número de casos positivos da gripe na Região do Sul da Ásia. o comportamento da série é parecido com as das demais regiões com forte sazonalidade anual e ausência de outras tendências determinísticas.

A Figura 4.6(b) apresenta o número de casos positivos médios a cada mês na Região do Sul da Ásia. No gráfico há dois picos na incidência da gripe, um no mês de março com média de 900 casos, e outro no mês de agosto com média acima de 1.100 casos.



(a) Série temporal da incidência da gripe na Região do Pacífico Ocidental.



(b) Média mensal da incidência da gripe na Região do Pacífico Ocidental.

Figura 4.7: Gráficos da incidência da gripe na Região do Pacífico Ocidental. (10/2008 − 12/2018).

A Figura 4.7(a) apresenta a série temporal do número de casos positivos da gripe na Região do Pacífico Ocidental. Assim como as demais, o comportamento da série apresenta forte sazonalidade anual e ausência de outras tendências determinísticas.

A Figura 4.7(b) ilustra o número de casos positivos médios a cada mês na Região do Pacífico Ocidental. Apesar dessa região conter países do hemisfério norte e do hemisfério sul (por exemplo, China e Austrália), o gráfico apresenta o comportamento sazonal típico do hemisfério norte, pois da China vem a grande maioria dos dados da região. Devido à isso, o número de casos positivos é maior no meses de dezembro, janeiro, fevereiro e março (época de inverno do hemisfério norte). Na Região do Pacífico Ocidental o pico se encontra no mês de janeiro, apresentando uma média maior que 14.000 casos.

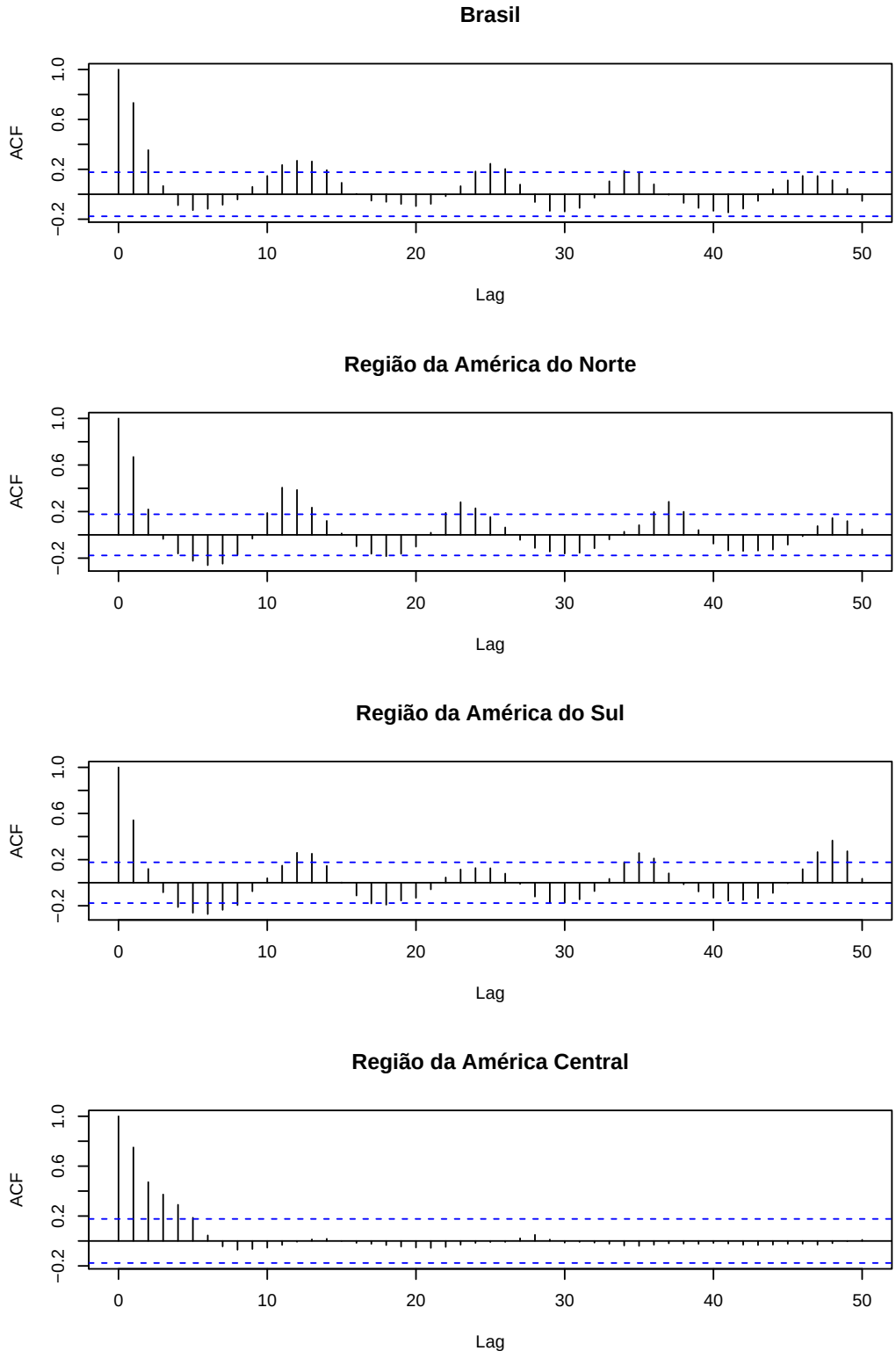


Figura 4.8: Gráfico da função de autocorrelação (ACF) do Brasil, da Região da América do Norte, da Região da América do Sul e da Região da América Central.

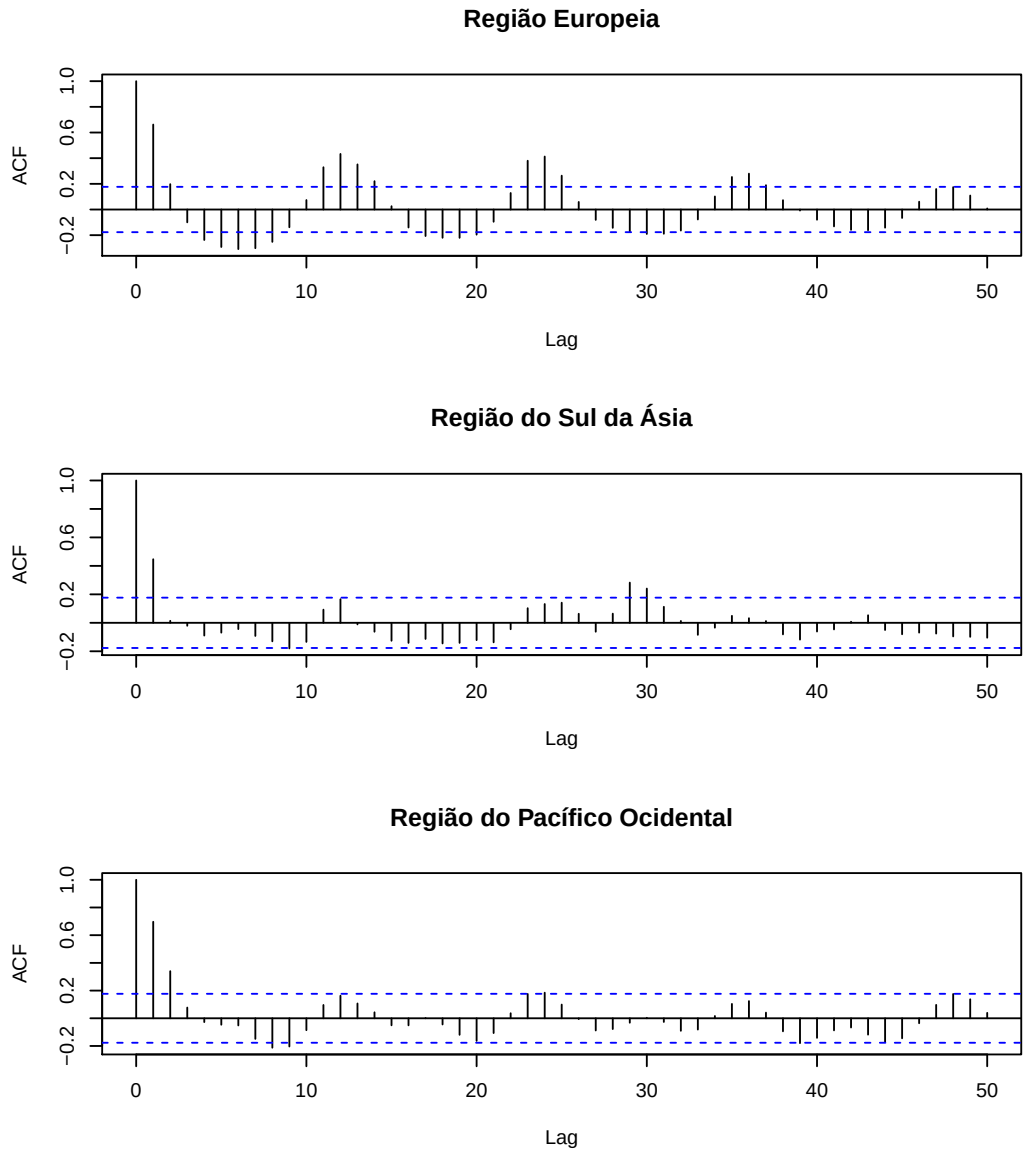


Figura 4.9: Gráfico da função de autocorrelação (ACF) da Região Europeia, da Região do Sul da Ásia e da Região do Pacífico Ocidental.

As Figuras 4.8 e 4.9 apresentam as funções de autocorrelação (ACF) de cada região. Encontrou-se o efeito da sazonalidade nas seguintes regiões: Brasil, América do Norte, América do Sul, Europa e Pacífico Ocidental. Em todos casos a série apresenta padrões alternados de defasagens positivas e negativas, e o período sazonal é de aproximadamente 12 meses ($lag = 12$). No entanto, na Região do Sul da Ásia há baixa evidência de sazonalidade, pois a maior parte das linhas verticais, que visam à indicar o grau de correlação entre as diferentes defasagens, estão dentro do intervalo de confiança de 95%. Ademais, a Região da América Central não apresenta quatro estações bem definidas, o que explica a ausência da sazonalidade no gráfico.

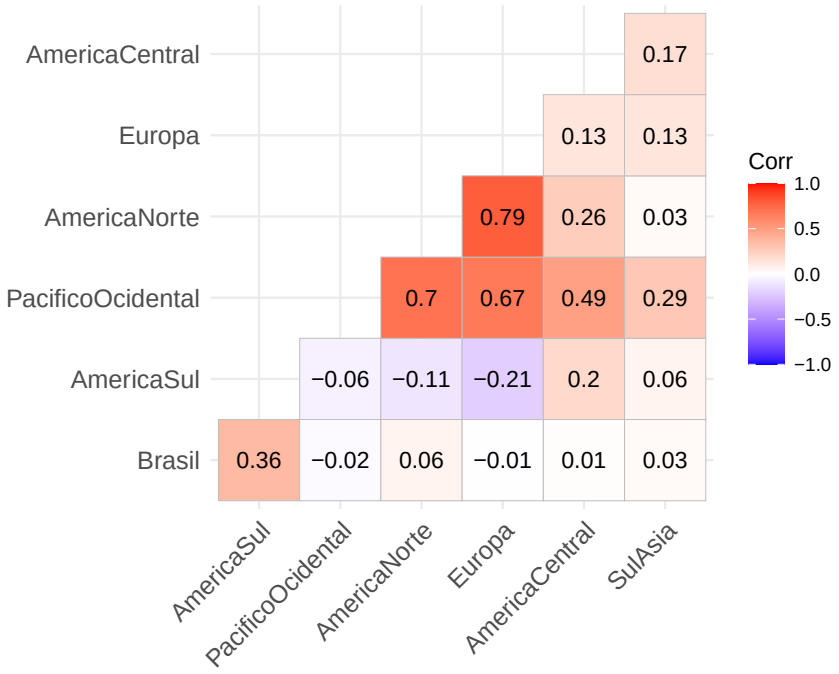


Figura 4.10: Matriz de correlação da incidência da gripe nas diversas regiões consideradas.

A matriz de correlação (Figura 4.10) tem por objetivo descrever a associação entre as diversas regiões consideradas no trabalho. Nesse caso as variáveis são as incidências da gripe das sete regiões. Observa-se que o Brasil não apresenta nenhuma correlação com as demais regiões. Entretanto, há uma correlação positiva entre a Região da América do Norte e a Região Europeia (0.79), entre a Região da América do Norte e a Região do Pacífico Ocidental (0.70) e entre Região Europeia e a Região do Pacífico Ocidental (0.67). As correlações entre as incidências podem ser confirmadas através de uma análise gráfica, que visa comparar as séries temporais das diferentes regiões.

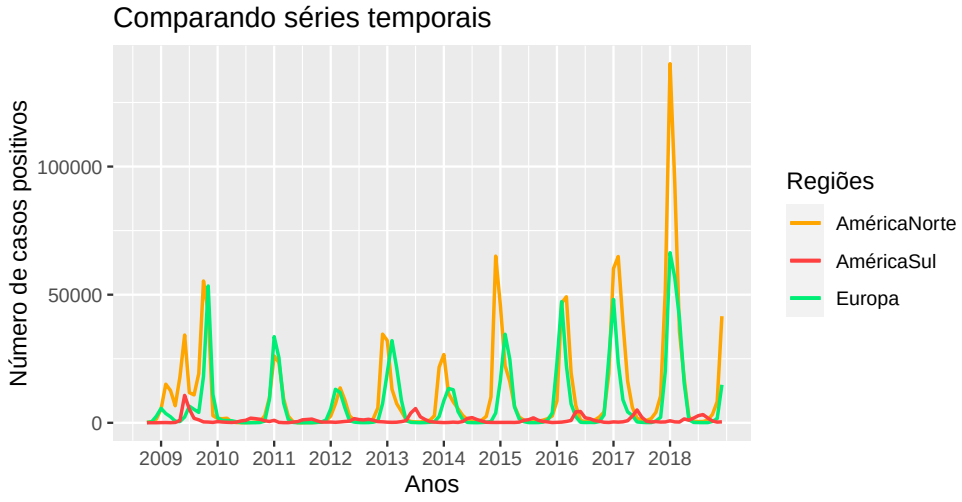


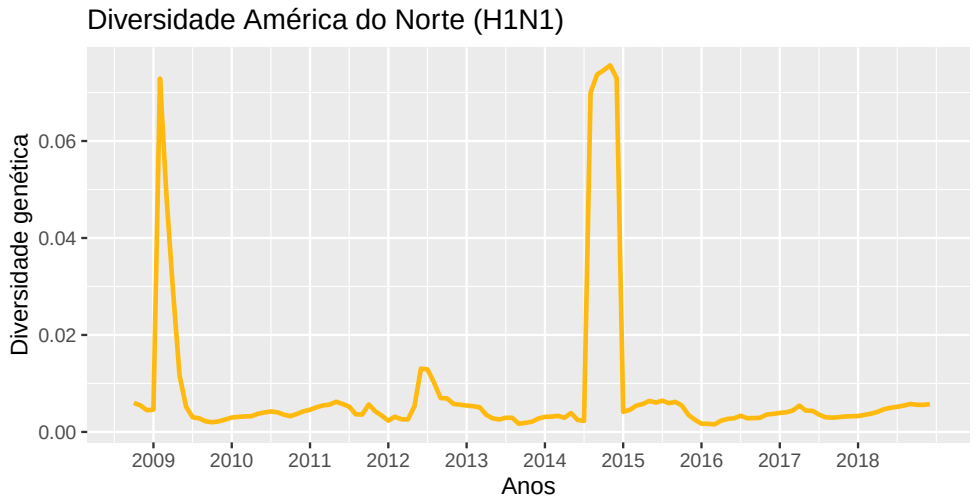
Figura 4.11: Comparando séries temporais.

A Figura 4.11 compara as séries temporais da Região Europeia, Região da Amé-

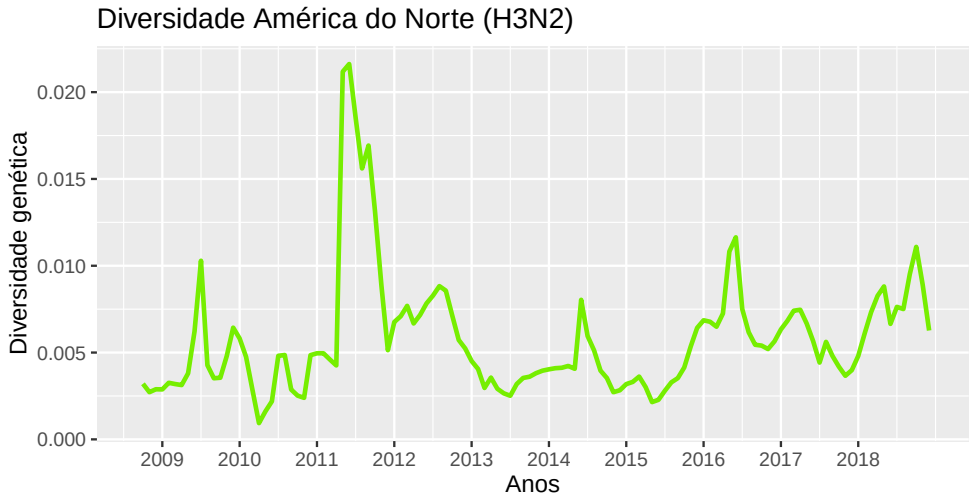
rica do Norte e Região da América do Sul. Observa-se que as séries da Europa e da América do Norte apresentam um comportamento muito parecido e isso justifica a correlação delas ser de 0.79. No entanto, a série da América do Sul é muito diferente das demais séries, onde a magnitude é um dos principais motivos, e por consequência disso que a sua correlação com a Região Europeia é de apenas -0.21, e com a América do Norte é de somente -0.11.

4.1.2 Diversidade Genética

Nesta subseção o objetivo é analisar o comportamento dos dados relativos à diversidade genética da gripe. A análise descritiva será apenas uma visualização e interpretação das séries temporais das diversidades genéticas. Neste caso são apresentadas 6 séries temporais, uma para cada região (global, América do Norte e Ásia) para os vírus das gripes H1N1 e H3N2.



(a) Diversidade genética da América do Norte (subtipo H1N1).

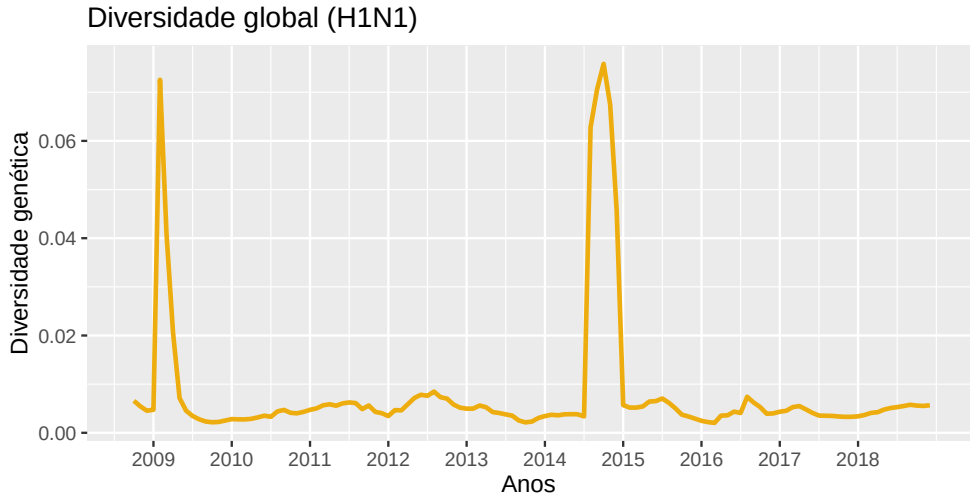


(b) Diversidade genética da América do Norte (subtipo H3N2).

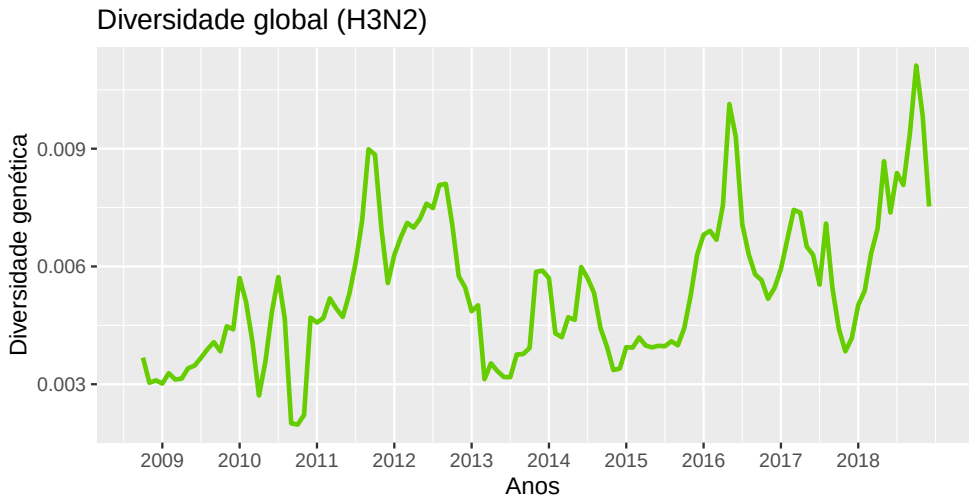
Figura 4.12: Série temporal trimestral da diversidade genética da América do Norte (10/2008−12/2018).

A Figura 4.12 apresenta as diversidades genéticas dos vírus H1N1 e H3N2 na Região da América do Norte. Na Figura 4.12(a) do vírus H1N1 observa-se uma alta diversidade genética no ano de 2009, que é decorrente da pandemia de gripe A. Há um pico no ano de 2014 que, segundo Linderman et al. (2014), foi um período em que as cepas correntes do vírus H1N1 causaram um nível inusitadamente alto da doença em adultos de meia-idade, pois a mutação desse período (2013-2014) foi muito particular, evitando as respostas imunológicas no grupo dos adultos. Ademais, nota-se que a dimensão da diversidade genética do vírus H1N1 é maior que a do vírus H3N2.

Na Figura 4.12(b) da diversidade do vírus H3N2, percebe-se que há um pico no ano de 2011 e uma pequena tendência crescente a partir do ano de 2013. No entanto, não identificou-se nenhuma sazonalidade na diversidade.



(a) Diversidade genética global (subtipo H1N1).



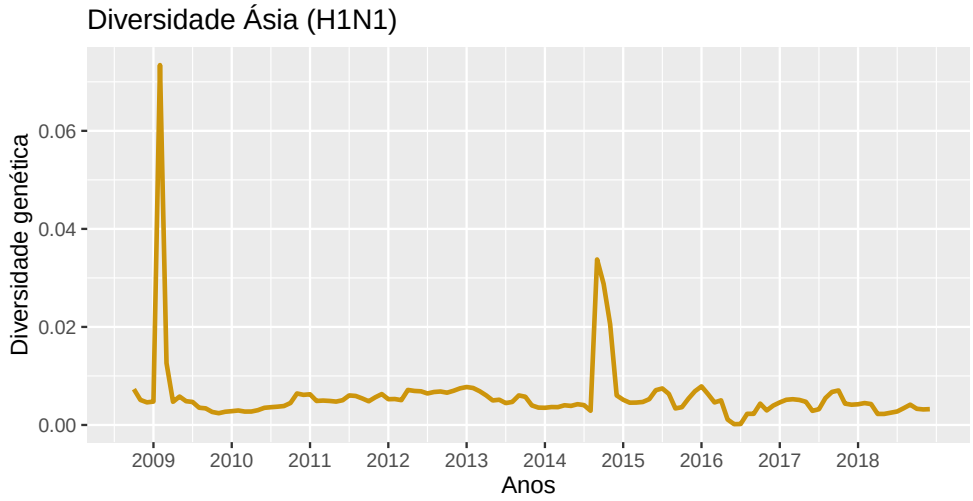
(b) Diversidade genética global (subtipo H3N2).

Figura 4.13: Série temporal trimestral da diversidade genética global (10/2008−12/2018)

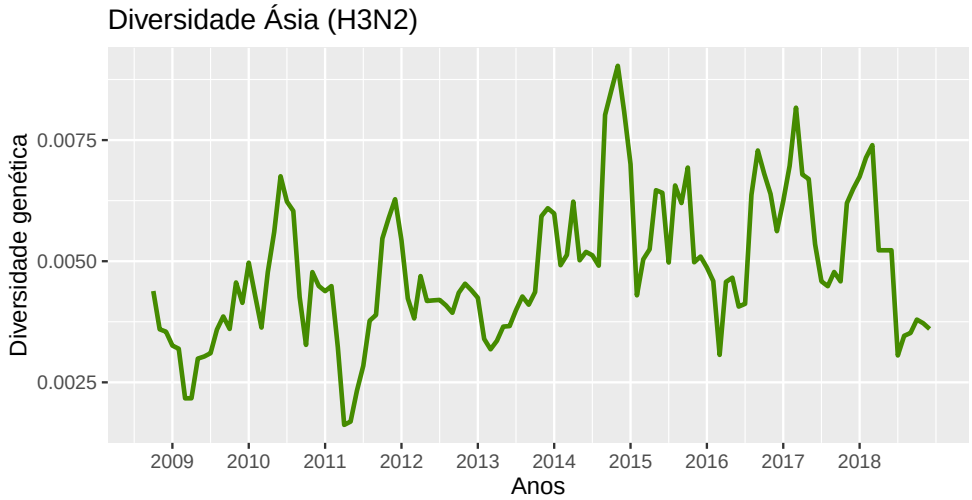
A Figura 4.13 apresenta as diversidades genéticas globais dos vírus H1N1 e H3N2. A Figura 4.13(a), relativa ao subtipo H1N1, apresenta um grande aumento da diver-

sidade no ano de 2009, coincidindo com o período da pandemia desse vírus. Há um pico da diversidade no ano de 2014, justificado pela alta mutação daquele período causando um aumento da incidência da doença em adultos. Ademais, ao comparar-se o gráfico da diversidade global com o da América do Norte (H1N1), observa-se que ambos gráficos são muito semelhantes, consequência do fato que grande parte dos dados globais de diversidade genética provém da América do Norte.

A Figura 4.13(b) apresenta a diversidade global do vírus H3N2. Percebe-se uma tendência crescente ao longo dos anos, indicando uma possível adaptabilidade do vírus. Todavia o impacto é inferior que a diversidade da América do Norte (H3N2), pois a dimensão da diversidade genética global é menor. A série não apresenta média e variância constantes ao longo do tempo, indicando não estacionariedade.



(a) Diversidade genética da Ásia (subtipo H1N1).



(b) Diversidade genética da Ásia (subtipo H3N2).

Figura 4.14: Série temporal trimestral da diversidade genética da Ásia (10/2008−12/2018)

A Figura 4.14 apresenta as diversidades genéticas dos vírus H1N1 e H3N2 na Ásia. Na Figura 4.14(a), do subtipo H1N1, há com comportamento semelhante às demais regiões (América do Norte H1N1 e global H1N1), com exceção do ano de

2014, onde o pico da diversidade genética da Ásia é menor. Ademais a série não apresenta tendência ou sazonalidade.

Na Figura 4.14(b) da diversidade do vírus H3N2 há uma pequena tendência no aumento das diversidades genéticas ao longo do tempo, podendo indicar novamente uma adaptabilidade do vírus. Entretanto, a dimensão da diversidade da Ásia é menor quando comparada a diversidade da América do Norte (H3N2).

4.2 Granger Causalidade

Esta seção apresenta os resultados obtidos da análise de Granger Causalidade utilizando dados relativos à incidência da gripe e dados relacionados à distância genética.

4.2.1 Número de Casos Positivos da Gripe

Nesta análise utilizou-se os dados descritos na Seção 3.1 (número de casos positivos da gripe) para aplicar o método de Granger causalidade. Primeiramente, o interesse é verificar se os dados de incidência do Brasil podem ser explicados através dos dados de incidência das demais regiões. Para isso adotou-se o teste de Granger causalidade (Subseção 2.2.1), e os resultados seguem abaixo.

Tabela 4.1: Resultado do Teste de Wald (séries não-estacionárias).

Hipótese Nula	p-valor	Lag
Região da América do Norte não Granger-causa Brasil	0.49	–
Região Europeia não Granger-causa Brasil	0.02*	3
Região da América Central não Granger-causa Brasil	0.94	–
Região da América do Sul não Granger-causa Brasil	0.01*	2
Região do Sul da Ásia não Granger-causa Brasil	0.34	–
Região do Pacífico Ocidental não Granger-causa Brasil	0.83	–

Considerando $\alpha = 0.05$, nota-se na Tabela 4.1 que as regiões da Europa e da América do Sul Granger-causam o Brasil. Isso significa que o número de casos positivos da gripe destas regiões Granger-causam o número de casos positivos da gripe no Brasil. Este resultado sugere que tanto a incidência da Região Europeia quanto a da Região da América do Sul ajudam a prever o valor presente da incidência no Brasil.

Em seguida escolheu-se fazer uma segunda análise, sem envolver o Brasil. Como explicado na Subseção 2.1.3, na literatura acredita-se que o epicentro da gripe no mundo se encontra no continente asiático, em particular na China, de onde o vírus migra para o hemisfério norte durante o solstício de inverno. Com isso decidiu-se verificar se a Região do Pacífico Ocidental (região que contém grande parte dos dados da Ásia), Granger-causa as regiões do hemisfério norte.

Tabela 4.2: Resultado do Teste de Wald (séries não-estacionárias).

Hipótese Nula	p-valor	Lag
Pacífico Ocidental não Granger-causa a América do Norte	0.03*	3
Pacífico Ocidental não Granger-causa a Europa	0.00*	3

Observa-se na Tabela 4.2 que, tomando $\alpha = 0.05$, a Região do Pacífico Ocidental Granger-causa ambas regiões. Portanto, pode-se dizer que a incidência da Região do Pacífico Ocidental ajuda a prever o valor presente da incidência da Região da América do Norte e da Região Europeia.

Uma consideração que pode-se fazer com esta análise é que, embora não há evidências de que a incidência da Região do Pacífico Ocidental não Granger-causa a incidência do Brasil, existe um efeito indireto da Região do Pacífico com o Brasil, pois o Pacífico Granger-causa a Região Europeia que Granger-causa o Brasil. Este efeito indireto não é detectável pois a análise de Granger causalidade não é associativa e nem comutativa. Uma outra observação é que descobriu-se que a incidência da Região da América do Norte Granger-causa a incidência da Região do Pacífico Ocidental ($p - \text{valor} = 0.029$ e $\text{lag} = 3$), logo percebe-se que há um efeito indireto da Região da América do Norte no Brasil, pois a Região da América do norte Granger-causa a Região do Pacífico Ocidental que Granger-causa a Região Europeia que Granger-causa o Brasil.

Por último, optou-se em analisar se alguma região Granger-causa outra região, mas sem ter necessariamente uma relação direta ou indireta com o Brasil. Os resultados encontrados são apresentados na Tabela 4.3.

Tabela 4.3: Resultado do Teste de Wald (séries não-estacionárias).

Hipótese Nula	p-valor	Lag
América Central não Granger-causa América do Norte	0.06	3
América Central não Granger-causa Europa	0.01*	3
América Central não Granger-causa América do Sul	0.06	5
América Central não Granger-causa Pacífico Ocidental	0.00*	3

A Tabela 4.3 mostra que, tomando $\alpha = 0.05$, a incidência da Região da América Central Granger-causa a incidência da Região Europeia e da Região do Pacífico Ocidental. Todavia, se considerarmos um nível de significância um pouco maior, a Região da América Central Granger-causa também as regiões da América do Norte e da América do Sul. Uma das justificativas mais prováveis da Granger causalidade da América Central nas outras regiões é a ocorrência da gripe suína (H1N1) nos anos de 2009 e 2010. O México foi a origem e o epicentro dessa pandemia, o que justifica o aumento da incidência da gripe primeiro na Região da América Central e seguidamente para as demais regiões.

4.2.2 Diversidade Genética

Nesta análise utilizou-se os dados descritos nas seções 3.1 (número de casos positivos da gripe) e 3.2 (diversidade genética) para aplicar o método de Granger causalidade. O interesse inicial é verificar se os dados de incidência do Brasil podem ser explicados através dos dados de diversidade genética. Para isso empregou-se o teste de Granger causalidade (Subseção 2.2.1) e os resultados seguem na tabela abaixo. Para visualizar melhor os resultados, nomeou-se a diversidade genética da América do Norte como North America, da Ásia como Asia e a global como All.

Tabela 4.4: Resultado do Teste de Wald (séries não-estacionárias).

Hipótese Nula	p-valor
North America (H1N1) não Granger-causa Brasil	0.59
All (H1N1) não Granger-causa Brasil	0.52
Asia (H1N1) não Granger-causa Brasil	0.55
North America (H3N2) não Granger-causa Brasil	0.23
All (H3N2) não Granger-causa Brasil	0.12
Asia (H3N2) não Granger-causa Brasil	0.36

Considerando $\alpha = 0.05$, percebe-se que nenhuma diversidade genética Granger-causa o Brasil, ou seja, a medida de diversidades genéticas não ajudam a prever o valor presente da incidência da gripe no Brasil. Em razão desses resultados, optou-se em fazer uma segunda análise, com o interesse de verificar se os dados de incidência das demais regiões podem ser explicados pelos dados da diversidade genética. Porém como há muitas combinações de regiões e diversidades, decidiu-se em apresentar apenas os resultados relevantes, que seguem na tabela abaixo.

Tabela 4.5: Resultado do Teste de Wald (séries não-estacionárias).

Hipótese Nula	p-valor	Lag
North America (H1N1) não Granger-causa Sul da Ásia	0.01*	3
All (H1N1) não Granger-causa Sul da Ásia	0.07	2
Asia (H3N2) não Granger-causa Sul da Ásia	0.07	2
North America (H3N2) não Granger-causa América Central	0.05*	2
Asia (H1N1) não Granger-causa América Central	0.00*	6

A Tabela 4.5 mostra que, considerando $\alpha = 0.05$, as diversidades genéticas da América do Norte (H3N2) e da Ásia (H1N1) ajudam na previsão da incidência da gripe da Região da América Central. Ademais, que a América do Norte (H1N1) Granger-causa o Sul da Ásia. No entanto, se considerarmos um nível de significância um pouco maior, ocorre que as diversidades da Ásia (H3N2) e global (H1N1) Granger-causam a incidência da Região do Sul da Ásia.

Para finalizar essa seção, optou-se em montar um diagrama que resume e agrupa diversas Granger causalidades encontradas nesse trabalho.

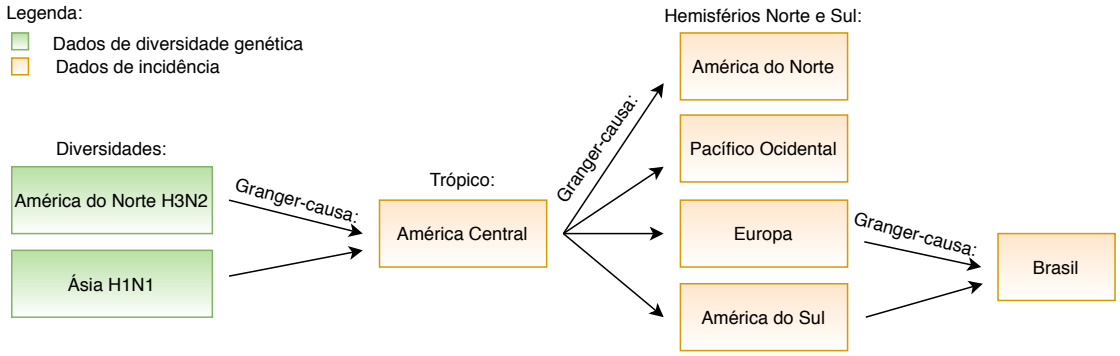


Figura 4.15: Diagrama de Granger causalidade.

A Figura 4.15 mostra que as diversidades genéticas da América do Norte (H3N2) e da Ásia (H1N1) Granger-causam a incidência da gripe da América Central, que Granger-causa as incidências da América do Norte, América do Sul, Europa e Pacífico Ocidental. Ademais que as incidências das regiões da América do Sul e da Europa ajudam na previsão da incidência no Brasil.

A Tabela 4.6 apresenta um resultado adicional encontrado. Nela percebe-se que a diversidade genética da América do Norte (H1N1) Granger-causa a diversidade genética da Ásia (H1N1).

Tabela 4.6: Resultado do Teste de Wald (séries não-estacionárias).

Hipótese Nula	p-valor	Lag
North America (H1N1) não Granger-causa Asia (H1N1)	0.00*	2

4.3 Regressão com Defasagens

Esta seção apresentamos os resultados obtidos da análise de regressão com defasagens utilizando dados relativos à incidência da gripe e dados relacionados à distância genética.

4.3.1 Número de Casos Positivos da Gripe

Nesta análise foram utilizados os dados descritos na Seção 3.1 (número de casos positivos da gripe) para representar a incidência da gripe no Brasil (denotado por B), na Europa (E), na América do Norte (A), na América Central (C), na América do Sul (S), no Sul da Ásia (s) e no Pacífico Ocidental (W). Este estudo tem como objetivo encontrar um modelo linear para a incidência da gripe no Brasil no tempo atual, digamos mês t, sendo indicado por B_t. Procura-se explicar esta incidência através dos dados históricos das regiões consideradas nos últimos 11 meses (defasagens), que no caso do Brasil denota-se por $B_{t-1}, B_{t-2}, B_{t-3}, \cdots, B_{t-11}$, na América do Norte indica-se por $A_{t-1}, A_{t-2}, A_{t-3}, \cdots, A_{t-11}$ e semelhantemente para as demais regiões.

Para a modelagem escolheu-se três modelos que utilizam o critério de seleção de variáveis devido ao grande número de variáveis explicativas, onde dois modelos utilizam o método de penalização LASSO, um com cinco e outro com dez variáveis (denotados por p), e o último trata-se do método Stepwise com o critério do p-valor, apresentando um nível de significância de 4.11%. Recorda-se que para o ajuste dos modelos utilizou-se dados mensais, logo cada defasagem resulta em um mês de atraso.

O primeiro modelo ajustado foi o modelo LASSO com $p = 5$ variáveis, o segundo foi o modelo LASSO com $p = 10$ variáveis e o terceiro foi o modelo Stepwise. A Tabela 4.7 apresenta as covariáveis escolhidas e seus respectivos coeficientes ajustados para cada modelo considerado.

Tabela 4.7: Covariáveis e respectivos coeficientes dos modelos LASSO com 5 variáveis, LASSO com 10 variáveis e Stepwise.

Covariáveis	LASSO 5	LASSO 10	Stepwise
Intercepto	59.486	50.519	-4.535
B_{t-1}	0.53585	0.58964	0.85254
B_{t-2}	–	-0.03595	-0.34971
B_{t-3}	–	-0.01192	–
A_{t-2}	–	–	-0.00410
A_{t-4}	0.00032	0.00072	–
C_{t-3}	-0.00079	-0.00864	-0.02094
S_{t-3}	–	–	0.03652
W_{t-7}	–	-0.00024	–
E_{t-1}	–	0.00084	0.00587
E_{t-2}	0.00578	0.00656	0.01071
E_{t-3}	0.00242	0.00169	–
E_{t-4}	–	0.00088	0.00638

A Tabela 4.7 mostra que as variáveis B_{t-1} (número de casos positivos no Brasil no tempo $t-1$), C_{t-3} (número de casos positivos na América Central no tempo $t-3$) e E_{t-2} (número de casos positivos na Europa no tempo $t-2$) aparecem em todos os três modelos. Uma outra maneira de interpretar os resultados é analisando-se a coluna dos coeficientes. Nela pode-se perceber a direção do impacto que as variáveis explicativas têm sobre a variável resposta B_t. Por exemplo, a variável explicativa B_{t-1}, presente em todos modelos, apresenta um coeficiente positivo, indicando que na medida em que o número dos casos no Brasil no tempo $t-1$ cresce, o número de casos no Brasil no tempo t (variável resposta) tende a crescer também.

Uma segunda análise optou por ser feita considerando cinco modelos (onde três modelos são os mesmos utilizados anteriormente, o quarto é um Stepwise com mais variáveis e o quinto é um StepAIC), mas acrescentando-se uma nova variável ao modelo, chamada μ_t, cujo objetivo é representar a média do número de casos positivos da gripe no Brasil em cada um dos doze meses. O interesse era saber se essa nova variável explicativa seria selecionada ou não dentre os cinco modelos. Os resultados da análise são apresentados nas Tabelas 4.8 a 4.12, uma para cada análise, que mostram as covariáveis escolhidas e seus respectivos coeficientes.

Tabela 4.8: Covariáveis e respectivos coeficientes do modelo ajustado utilizando o modelo LASSO com 5 variáveis.

Covariáveis	Coeficientes
Intercepto	60.821
B_{t-1}	0.52843
A_{t-4}	0.00019
E_{t-2}	0.00548
E_{t-3}	0.00242
μ_t	0.01431

Tabela 4.9: Covariáveis e respectivos coeficientes do modelo ajustado utilizando o modelo LASSO com 10 variáveis.

Covariáveis	Coeficientes
Intercepto	47.833
B_{t-1}	0.56373
B_{t-2}	-0.00001
B_{t-3}	-0.02262
A_{t-4}	0.00073
C_{t-3}	-0.00755
E_{t-1}	0.00066
E_{t-2}	0.00653
E_{t-3}	0.00180
E_{t-4}	0.00049
μ_t	0.01347

Tabela 4.10: Covariáveis e respectivos coeficientes do modelo ajustado utilizando o modelo Stepwise 1 (mais variáveis).

Covariáveis	Coeficientes
Intercepto	-0.843
B_{t-1}	0.8470
B_{t-2}	-0.34248
A_{t-2}	-0.00414
C_{t-1}	0.01515
C_{t-3}	-0.02490
E_{t-1}	0.00535
E_{t-2}	0.01117
E_{t-4}	0.00642
E_{t-8}	0.00255

Tabela 4.11: Covariáveis e respectivos coeficientes do modelo ajustado utilizando o modelo Stepwise 2.

Covariáveis	Coeficientes
Intercepto	30.188
B_{t-1}	0.83433
B_{t-2}	-0.31969
A_{t-2}	-0.00418
C_{t-1}	0.01433
C_{t-3}	-0.02511
E_{t-1}	0.00470
E_{t-2}	0.01113
E_{t-4}	0.00560

Tabela 4.12: Covariáveis e respectivos coeficientes do modelo ajustado utilizando o modelo StepAIC.

Covariáveis	Intercepto	B_{t-1}	B_{t-2}	B_{t-3}	B_{t-6}	B_{t-9}
Coeficientes	245.813	0.99823	-0.53320	0.19909	0.28382	0.19111
Covariáveis	B_{t-11}	A_{t-2}	A_{t-8}	C_{t-2}	C_{t-3}	C_{t-4}
Coeficientes	-0.11843	-0.00909	0.00226	0.02714	-0.03346	0.04356
Covariáveis	C_{t-6}	C_{t-7}	C_{t-8}	S_{t-2}	S_{t-3}	S_{t-4}
Coeficientes	0.03575	-0.07103	0.02461	-0.07482	0.06675	-0.09691
Covariáveis	S_{t-6}	S_{t-8}	S_{t-10}	S_{t-11}	W_{t-1}	W_{t-3}
Coeficientes	-0.06438	-0.09074	0.05997	0.03880	0.00995	0.01160
Covariáveis	W_{t-4}	W_{t-5}	W_{t-7}	W_{t-8}	W_{t-10}	s_{t-1}
Coeficientes	-0.03125	0.01487	-0.02470	0.02208	-0.00744	-0.06890
Covariáveis	s_{t-2}	s_{t-3}	s_{t-5}	s_{t-6}	s_{t-7}	s_{t-9}
Coeficientes	0.07105	-0.10485	0.06589	-0.15519	0.10050	0.05256
Covariáveis	E_{t-1}	E_{t-2}	E_{t-4}	E_{t-6}	E_{t-7}	E_{t-8}
Coeficientes	0.00923	0.01030	0.01163	-0.00531	0.00745	-0.01071
Covariáveis	μ_t					
Coeficientes	-0.55066					

Analisando-se as Tabelas 4.8 a 4.12 percebe-se que a nova variável μ_t aparece nos modelos LASSO com 5 e 10 variáveis e também no StepAIC, mas não nas modelagens Stepwise. Comparando as Tabelas do LASSO com $p = 5$, LASSO com $p = 10$ e Stepwise 2 com as da análise anterior constata-se que variáveis selecionadas de um modo geral são as "mesmas" em todos modelos.

Após as modelagens optou-se em fazer uma análise de predição in-sample (predição dentro da amostra) e out-of-sample (predição fora da amostra) utilizando os últimos cinco modelos. Conforme a Seção 3.1, os dados são de outubro de 2008 até novembro de 2019, e destes, de 2008 até 2018 foi utilizado na modelagem e o ano

de 2019 foi reservado para fazer a previsão out-of-sample. Um exemplo de predição fora amostra seria prever o valor de fevereiro de 2019 utilizando os modelos citados anteriormente (com dados até 2018) e acrescentando os dados de janeiro de 2019 nas covariáveis. As Figuras 4.16 a 4.20 apresentam os valores preditos ($h = 1$ passos à frente) dentro e fora da amostra para cada um dos modelos considerados.

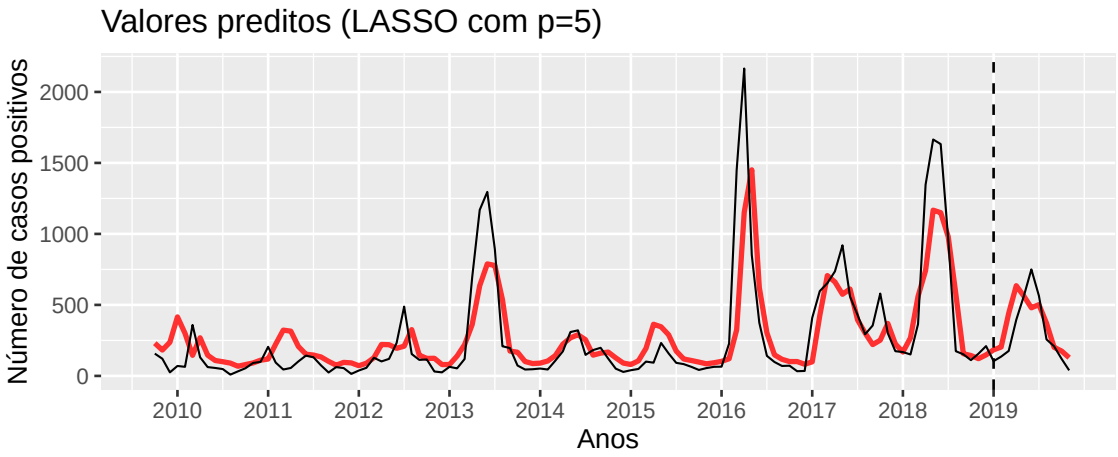


Figura 4.16: Gráfico predição $h = 1$ passo à frente (Modelo LASSO com $p = 5$).

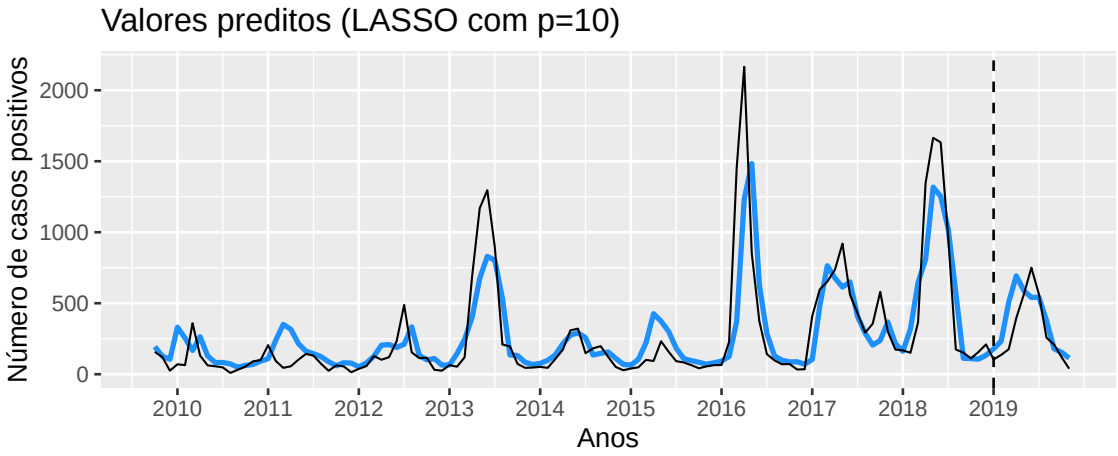


Figura 4.17: Gráfico predição $h = 1$ passo à frente (Modelo LASSO com $p = 10$).

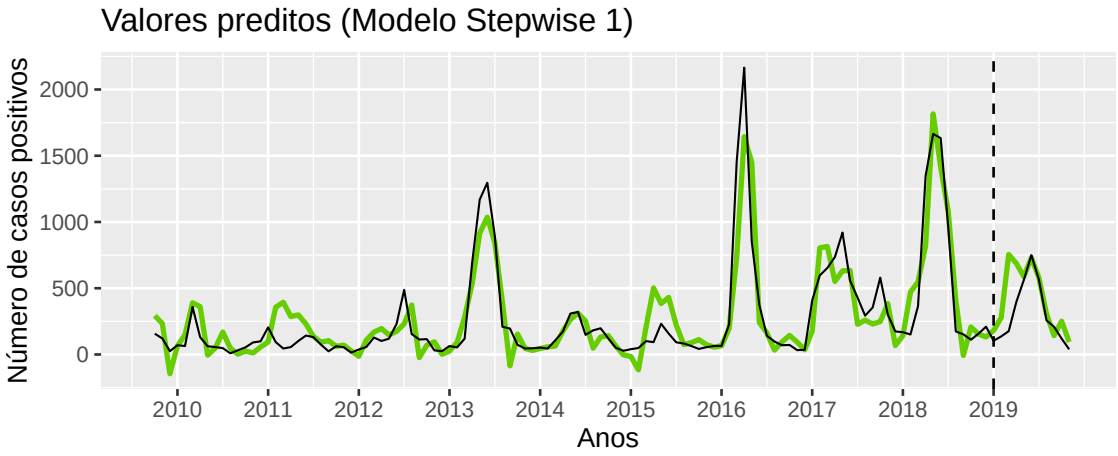


Figura 4.18: Gráfico predição $h = 1$ passo à frente (Modelo Stepwise 1).

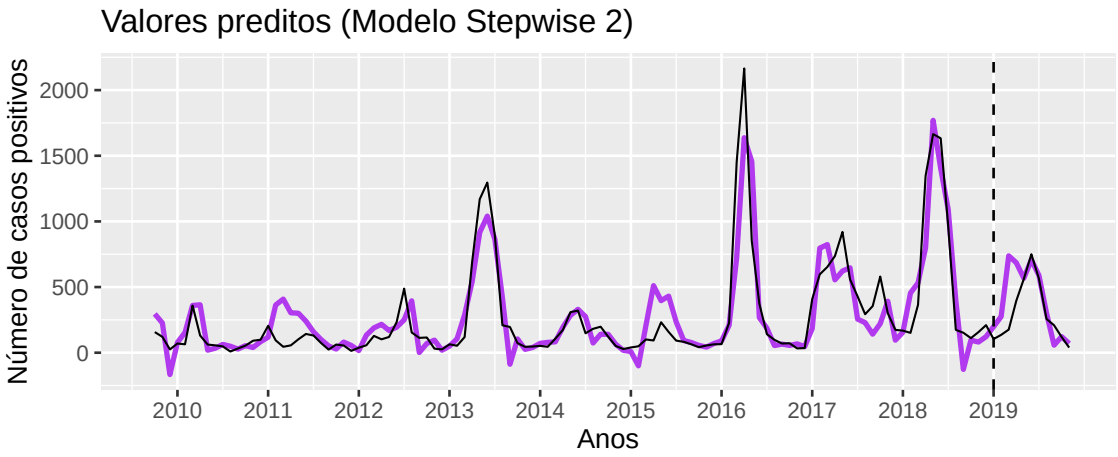


Figura 4.19: Gráfico predição $h = 1$ passo à frente (Modelo Stepwise 2).

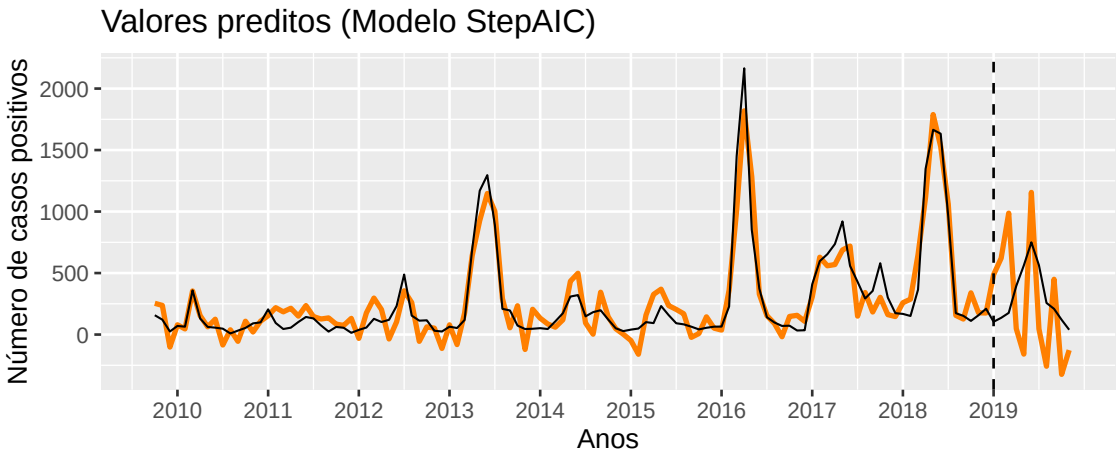


Figura 4.20: Gráfico predição $h = 1$ passo à frente (Modelo StepAIC).

As Figuras 4.16 a 4.20 apresentam os valores preditos da variável resposta B_t, dentro e fora da amostra, e fazem uma comparação entre essa predição com os seus valores verdadeiros, que estão sendo representados pela cor preta. Percebe-se que as predições in-sample parecem estar razoáveis em todos os cinco modelos, com exceção do ano de 2011, onde os modelos erraram o pico. Mas quando se analisa as predições out-of-sample nota-se que todos os modelos possuem um desempenho semelhante, exceto pelo StepAIC que, comparativamente, apresenta uma previsão fora da amostra muito ruim. Esta deficiência na previsão out-of-sample do StepAIC pode ser atribuída ao número grande de variáveis incluídas no modelo ($p = 42$), que pode ocasionar o overfitting (sobreajuste) - quando um modelo se ajusta muito bem aos dados, mas é ineficiente na predição de novos resultados. Ademais, apesar de alguns modelos apresentarem previsões negativas, como o objetivo é predizer os picos de incidência da gripe, então esses valores negativos não serão considerados um problema. A Tabela 4.13 apresenta o erro quadrático médio (EQM) e o erro percentual absoluto médio (MAPE) das previsões de cada modelo, dentro e fora da amostra.

Tabela 4.13: Erro quadrático médio e erro percentual absoluto médio de previsão de cada modelo.

Medidas/Modelos	Stepwise 1	Stepwise 2	LASSO 5	LASSO 10	StepAIC
EQM (in-sample)	31586.7	32515.1	52788.1	46673.9	18124.7
EQM (out-of-sample)	43228.2	41513.0	21805.3	25453.8	241818.1
MAPE (in-sample)	91.2	85.8	105.2	88.0	105.7
MAPE (out-of-sample)	81.2	69.8	64.3	65.7	241.5

Analisando o EQM percebe-se que o modelo StepAIC apresentou, por grande margem, o melhor desempenho em termos de previsão in-sample. Em termos de EQM fora da amostra, os modelos LASSO foram melhores, com pequena vantagem para LASSO com 5 variáveis. Em relação ao MAPE nota-se que dentro da amostra o melhor foi o modelo Stepwise 2 e fora da amostra o modelo LASSO com $p = 5$.

Em um segundo momento, optou-se por analisar a capacidade preditora dos modelos de um até 11 passos à frente. Portanto, como os dados dos modelos são de outubro de 2008 até dezembro de 2018, então as previsões serão desde janeiro de 2019 (1 passo à frente) até novembro de 2019 (11 passos à frente). As modelagens LASSO com $p = 5$ variáveis (Tabela 4.8), LASSO com $p = 10$ variáveis (Tabela 4.9), Stepwise 1 (Tabela 4.10) e Stepwise 2 (Tabela 4.11), vistas anteriormente, vão ser utilizadas nesta análise. Apenas o modelo StepAIC (Tabela 4.12) não vai ser utilizado, pois na análise anterior de predição ele apresentou problemas de overfitting. Abaixo encontram-se dois gráficos que mostram os valores preditos de janeiro de 2019 ($h = 1$ passo à frente) até novembro de 2019 ($h = 11$ passos à frente) dos diferentes modelos, onde o primeiro gráfico apresenta a série temporal inteira e o segundo apenas parte da série.

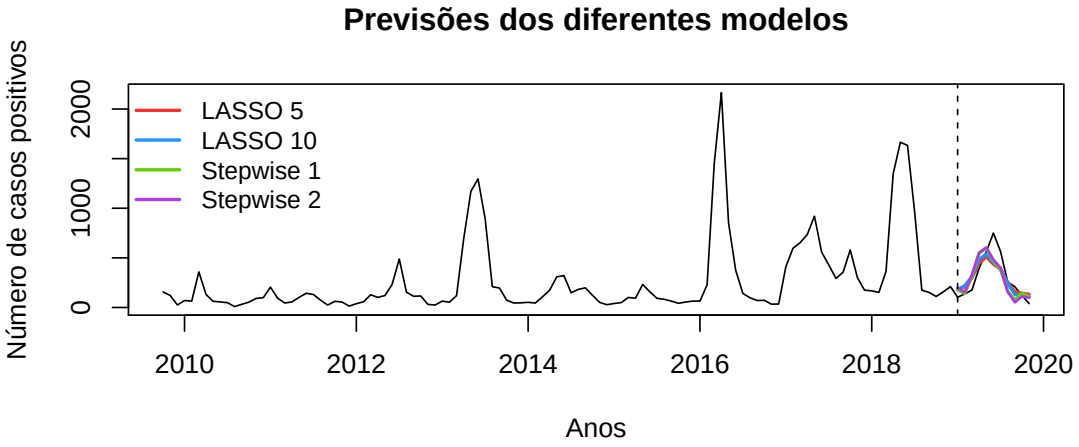


Figura 4.21: Gráfico predição $h = 11$ passos à frente.

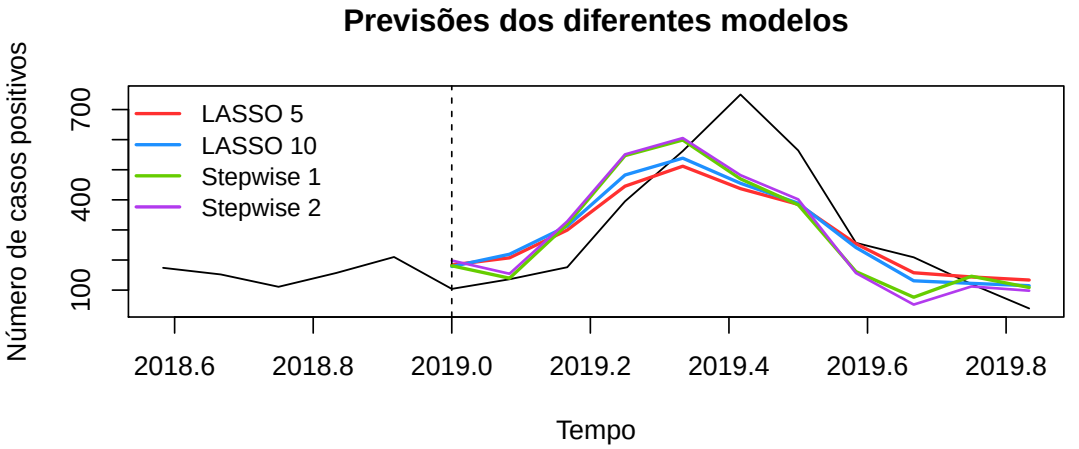


Figura 4.22: Gráfico predição $h = 11$ passos à frente (parte da série temporal).

A Figura 4.21 mostra os valores preditos da variável resposta B_t no ano de 2019 nos diferentes modelos, e os valores verdadeiros de 2009 até 2019 em preto. Percebe-se que as predições para o ano de 2019 parecem estar próximas dos seus valores verdadeiros. No entanto na Figura 4.22 a predição do ano de 2019 não aparenta estar tão boa quanto aparentava estar no primeiro gráfico, apesar de serem os mesmos resultados. Conclui-se com ambos gráficos que há um comportamento parecido entre os dados verdadeiros e os dados preditos para o ano de 2019, porém os dados preditos parecem estar com alguns lags (meses) adiantado, pois tanto a previsão quanto os dados reais indicam uma tendência crescente no início do ano e uma tendência decrescente no final do ano. Para saber qual dos modelos fez o melhor papel na predição, montou-se a Tabela 4.14 que apresenta o erro quadrático médio de cada modelo e em cada um dos 11 passos.

Tabela 4.14: Erro quadrático médio em cada um dos 11 passos.

Passos/Modelos	Stepwise 1	Stepwise 2	LASSO 5	LASSO 10
1 passo à frente	5904.13	8880.90	6346.41	**5511.15**
2 passos à frente	**2958.38**	4608.45	5712.86	6216.09
3 passos à frente	**8736.86**	10816.13	8895.71	10372.02
4 passos à frente	12323.75	14196.65	**7305.05**	9692.25
5 passos à frente	10129.28	11732.39	**6349.04**	7864.72
6 passos à frente	21543.97	21750.63	21599.59	**21059.42**
7 passos à frente	23130.21	**22416.22**	23092.44	22439.56
8 passos à frente	21381.31	20870.76	20206.98	**19664.79**
9 passos à frente	20962.64	21295.01	18254.56	**18157.53**
10 passos à frente	18934.49	19171.77	16485.37	**16342.35**
11 passos à frente	17653.39	17745.75	15798.62	**15377.26**

Por meio da Tabela 4.14, percebe-se que na predição de um passo à frente (previsão para janeiro de 2019) o modelo que apresentou o menor EQM e portanto o melhor resultado foi o LASSO com $p = 10$ variáveis. No entanto na predição de dois passos à frente (previsão para janeiro de 2019 e fevereiro de 2019) nota-se que o Stepwise 1 apresentou o menor EQM. Para facilitar a visualização desta tabela destacou-se em negrito o menor EQM em cada horizonte de previsão. Analisando os menores EQM em cada um dos 11 passos, percebe-se que o modelo que apresentou melhores resultados foi o LASSO com $p = 10$ variáveis, pois obteve o menor EQM em 6 dos 11 cenários.

4.3.2 Modelos utilizando Diversidade Genética

Nesta análise utilizou-se os dados descritos na Seção 3.1 (número de casos positivos da gripe) para caracterizar a incidência da gripe no Brasil (denotado por B), na Europa (E), na América do Norte (A), na América Central (C), na América do Sul (S), no Sul da Ásia (s) e no Pacífico Ocidental (W). Ademais utilizou-se os dados descritos na Seção 3.2 para representar a diversidade genética da gripe na América do Norte (denotado por N para o subtipo H1N1 e n para o subtipo H3N2), na Ásia (P para o subtipo H1N1 e p para o subtipo H3N2) e no mundo (M para o subtipo H1N1 e m para o subtipo H3N2).

Este estudo tem como objetivo encontrar um modelo linear para a incidência da gripe no Brasil no mês t (tempo atual), denotado por B_t. O objetivo é explicar esta incidência através dos dados históricos das regiões consideradas nos últimos 6 meses (defasagens), denotado por $B_{t-1}, B_{t-2}, \cdots, B_{t-6}$ para o Brasil, por $A_{t-1}, A_{t-2}, \cdots, A_{t-6}$ para a América do Norte, por $E_{t-1}, E_{t-2}, \cdots, E_{t-6}$ para a Europa e semelhantemente para as demais regiões. Da mesma forma, procura-se explicar B_t através dos dados históricos das diversidades genéticas consideradas nos últimos 6 meses que, no caso da diversidade genética da América do Norte para a H3N2, denota-se por $n_{t-1}, n_{t-2}, \cdots, n_{t-6}$, na diversidade genética da Ásia H1N1 indica-se por $P_{t-1}, P_{t-2}, \cdots, P_{t-6}$ e semelhantemente para as demais diversidades genéticas.

Para a modelagem escolheu-se cinco modelos que utilizam o critério de seleção de variáveis devido ao grande número variáveis explicativas, onde dois destes abordam

o método LASSO, um com cinco e outro com dez variáveis, outros dois modelos se tratam do método Stepwise com o critério do p-valor, onde o primeiro apresenta um nível de significância de 10% e o segundo de 5%, e o último se refere ao método StepAIC, que é a seleção Backward com o critério AIC. Abaixo encontram-se cinco tabelas, uma para cada modelo, que mostram as covariáveis escolhidas e seus respectivos coeficientes.

Tabela 4.15: Covariáveis e respectivos coeficientes do modelo ajustado utilizando o modelo LASSO com 5 variáveis.

Covariáveis	Coeficientes
Intercepto	54.872
B_{t-1}	0.50499
A_{t-4}	0.00037
E_{t-2}	0.00535
E_{t-3}	0.00219
μ_t	0.07586

Tabela 4.16: Covariáveis e respectivos coeficientes do modelo ajustado utilizando o modelo LASSO com 10 variáveis.

Covariáveis	Coeficientes
Intercepto	42.789
B_{t-1}	0.55347
B_{t-3}	-0.02839
A_{t-4}	0.00114
C_{t-3}	-0.00689
E_{t-1}	0.00077
E_{t-2}	0.00618
E_{t-3}	0.00147
μ_t	0.07361
P_{t-4}	2301.50
P_{t-5}	-2223.09

Tabela 4.17: Covariáveis e respectivos coeficientes do modelo ajustado utilizando o modelo Stepwise 1 (mais variáveis).

Covariáveis	Coeficientes
Intercepto	38.789
B_{t-1}	0.85717
B_{t-2}	-0.34293
A_{t-4}	0.00286
C_{t-2}	0.02204
C_{t-3}	-0.03545
S_{t-2}	0.02604
E_{t-2}	0.00893
P_{t-4}	6482.74
P_{t-5}	-9221.79

Tabela 4.18: Covariáveis e respectivos coeficientes do modelo ajustado utilizando o modelo Stepwise 2.

Covariáveis	Coeficientes
Intercepto	60.916
B_{t-1}	0.82104
B_{t-2}	-0.28483
A_{t-4}	0.00287
C_{t-2}	0.02541
C_{t-3}	-0.03639
E_{t-2}	0.00852
P_{t-4}	5949.73
P_{t-5}	-8932.70

Tabela 4.19: Covariáveis e respectivos coeficientes do modelo ajustado utilizando o modelo StepAIC.

Covariáveis	Intercepto	B_{t-1}	B_{t-2}	B_{t-6}	A_{t-1}	A_{t-2}
Coeficientes	-451.560	0.76639	-0.46651	0.16839	-0.00339	-0.00738
Covariáveis	A_{t-4}	A_{t-5}	A_{t-6}	C_{t-2}	C_{t-3}	C_{t-4}
Coeficientes	-0.00362	0.00324	-0.00406	0.04852	-0.06649	0.03907
Covariáveis	C_{t-5}	S_{t-2}	S_{t-3}	S_{t-5}	S_{t-6}	W_{t-3}
Coeficientes	-0.02277	0.05985	0.02626	-0.06751	0.05491	0.01587
Covariáveis	W_{t-4}	s_{t-2}	s_{t-4}	s_{t-6}	E_{t-1}	E_{t-2}
Coeficientes	-0.01493	0.06594	-0.08059	-0.10778	0.01320	0.01059
Covariáveis	E_{t-4}	E_{t-6}	M_{t-2}	M_{t-4}	M_{t-6}	m_{t-2}
Coeficientes	0.01843	0.00554	-62816.42	7304.19	-12771.59	-27333.42
Covariáveis	N_{t-2}	n_{t-5}	P_{t-1}	P_{t-2}	P_{t-3}	P_{t-5}
Coeficientes	45461.21	14107.46	13791.42	18133.72	7364.63	-7898.63
Covariáveis	P_{t-6}	p_{t-1}	p_{t-4}	p_{t-6}		
Coeficientes	10365.44	51570.26	-31538.14	64825.66		

Nas Tabelas 4.15 a 4.19 observamos que as variáveis de incidência que aparecem em todos os cinco modelos são: B_{t-1} (número de casos positivos no Brasil com uma defasagem), E_{t-2} (número de casos positivos na Europa com duas defasagens) e A_{t-4} (número de casos positivos na América do Norte com quatro defasagens). Ademais quando analisa-se as variáveis relacionadas à diversidade genética percebe-se que as covariáveis P_{t-4} (diversidade genética da gripe H1N1 na Ásia com quatro defasagens) e P_{t-5} (diversidade genética da gripe H1N1 na Ásia com cinco defasagens) aparecem em quase todos os modelos. Pode-se interpretar os resultados analisando a coluna dos coeficientes. Nela percebe-se o impacto que as variáveis explicativas têm sobre a variável resposta B_t. Por exemplo, a variável de diversidade genética P_{t-4}, presente em vários modelos, apresenta um coeficiente positivo, indicando que na medida em que a diversidade genética da gripe H1N1 na Ásia no tempo $t-4$ cresce, o número de casos no Brasil no tempo t (variável dependente) tende a crescer também.

Após as modelagens optou-se em fazer uma análise de predição in-sample (predição dentro da amostra) e out-of-sample (predição fora da amostra), do mesmo modo que foi feito na Subseção 4.3.1. Nesta análise foram utilizados os dados das Seções 3.1 (número de casos positivos da gripe) e 3.2 (distância genética da gripe), no período de outubro de 2008 até agosto de 2019, onde 2008 até 2018 foi utilizado na modelagem e o ano de 2019 foi reservado para fazer a previsão out-of-sample. As Figuras 4.23 a 4.27 mostram os valores preditos ($h = 1$ passos à frente) dentro e fora da amostra.

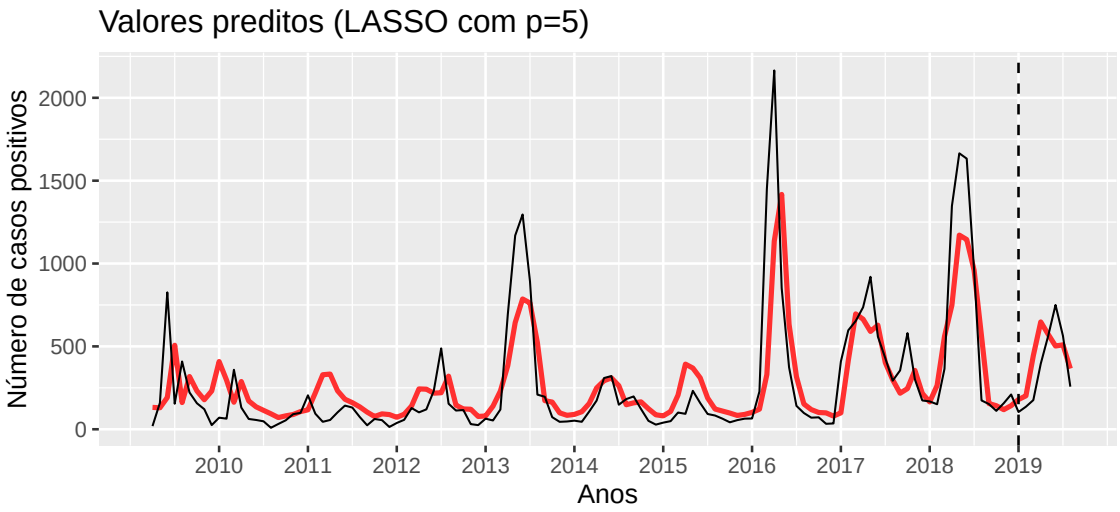


Figura 4.23: Gráfico predição $h = 1$ passo à frente (Modelo LASSO com $p = 5$).

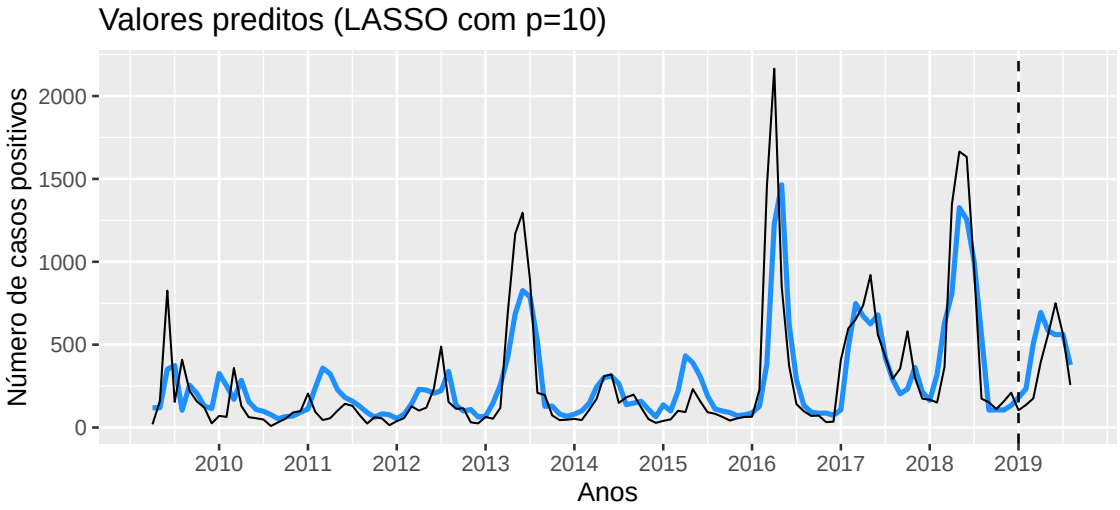


Figura 4.24: Gráfico predição $h = 1$ passo à frente (Modelo LASSO com $p = 10$).

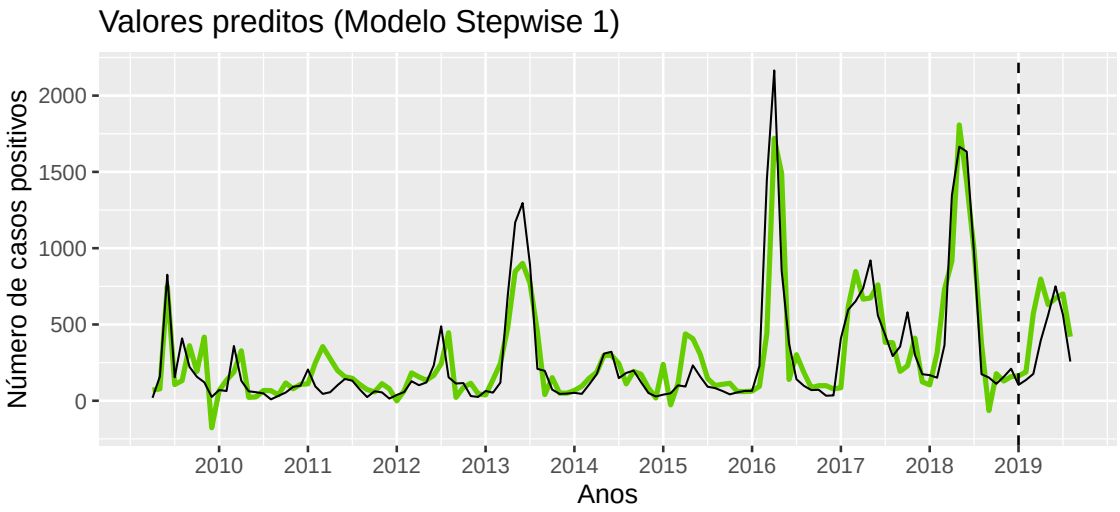


Figura 4.25: Gráfico predição $h = 1$ passo à frente (Modelo Stepwise 1).

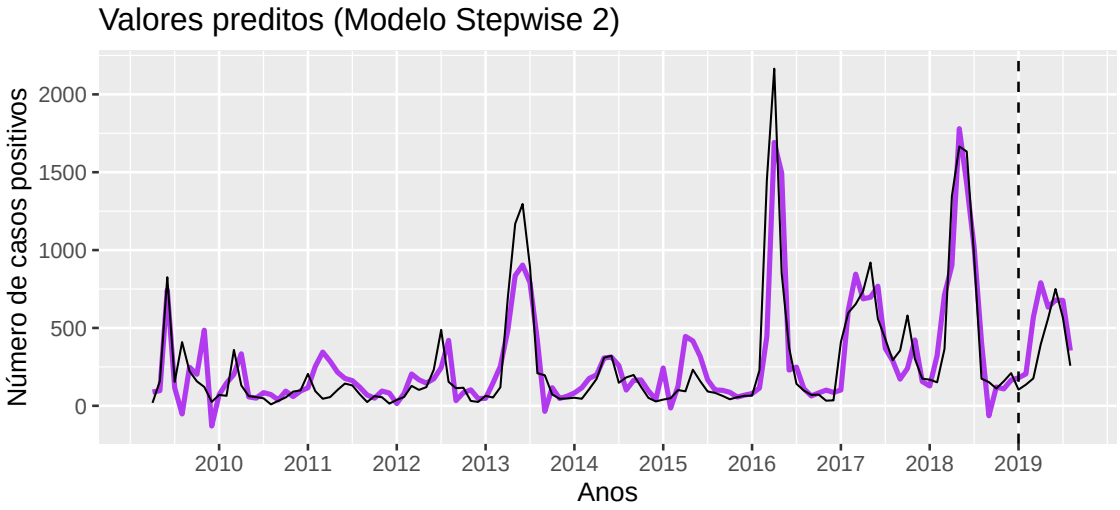


Figura 4.26: Gráfico predição $h = 1$ passo à frente (Modelo Stepwise 2).

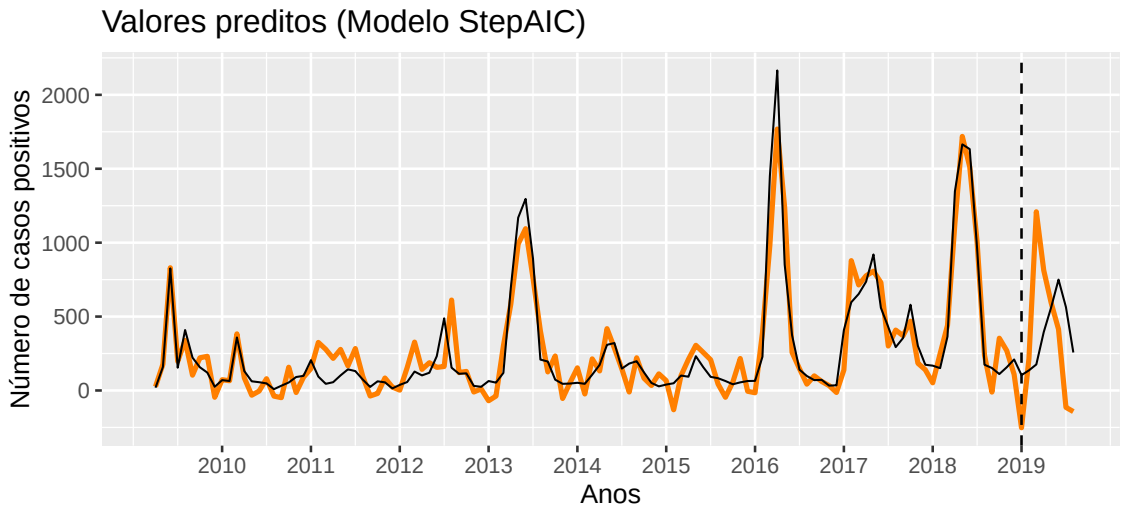


Figura 4.27: Gráfico predição $h = 1$ passo à frente (Modelo StepAIC).

As Figuras 4.23 a 4.27 apresentam as previsões da variável resposta B_t, dentro e fora da amostra, bem como os seus valores verdadeiros, representados pela cor preta. Nota-se que, no geral, as predições in-sample aparentam estar boas em todos os cinco modelos, com exceção ano de 2011, onde os modelos erraram o pico. Porém quando se analisa as predições out-of-sample percebe-se que apenas o StepAIC está muito ruim, pois ocorre overfitting no modelo. No entanto para saber qual modelo fez um papel melhor na previsão, construiu-se a Tabela 4.20 que compara o erro quadrático médio (EQM) e o erro percentual absoluto médio (MAPE) de cada modelo e dentro e fora da amostra.

Tabela 4.20: Erro quadrático médio e erro percentual absoluto médio de previsão de cada modelo.

Medidas/Modelos	Stepwise 1	Stepwise 2	LASSO 5	LASSO 10	StepAIC
EQM (in-sample)	34567.0	35555.7	55306.2	47175.5	18479.5
EQM (out-of-sample)	47623.4	44089.8	27781.2	33250.2	263571.0
MAPE (in-sample)	91.2	89.8	110.9	95.5	87.7
MAPE (out-of-sample)	66.5	65.8	53.1	60.7	177.4

Analisando o EQM nota-se que, dentro da amostra, o modelo StepAIC foi muito melhor que os demais modelos. Em termos de EQM fora da amostra os modelos LASSO foram melhores, com vantagem para LASSO com 5 variáveis. Em relação ao MAPE percebe-se que dentro da amostra o melhor foi o modelo StepAIC e fora da amostra foi o modelo LASSO com $p = 5$.

Em um segundo momento, optou-se em prever não somente um passo à frente, mas desde um até 11 passos à frente, da mesma maneira que foi feita na Subseção 4.3.1. Neste caso, como os dados são de 2008 até 2018, então as previsões serão de janeiro de 2019 (1 passo à frente) até novembro de 2019 (11 passos à frente). Os modelos abordados nessa análise são: LASSO com $p = 5$ variáveis (Tabela 4.15), LASSO com $p = 10$ variáveis (Tabela 4.16), Stepwise 1 (Tabela 4.17) e Stepwise 2 (Tabela 4.18). Entretanto, o modelo StepAIC não vai ser utilizado, pois anteriormente ele apresentou problemas de overfitting. As Figuras 4.28 e 4.29 apresentam

os valores preditos de janeiro de 2019 ($h = 1$ passo à frente) até novembro de 2019 ($h = 11$ passos à frente) dos diferentes modelos, onde o primeiro gráfico apresenta a série temporal inteira e o segundo apenas parte da série.

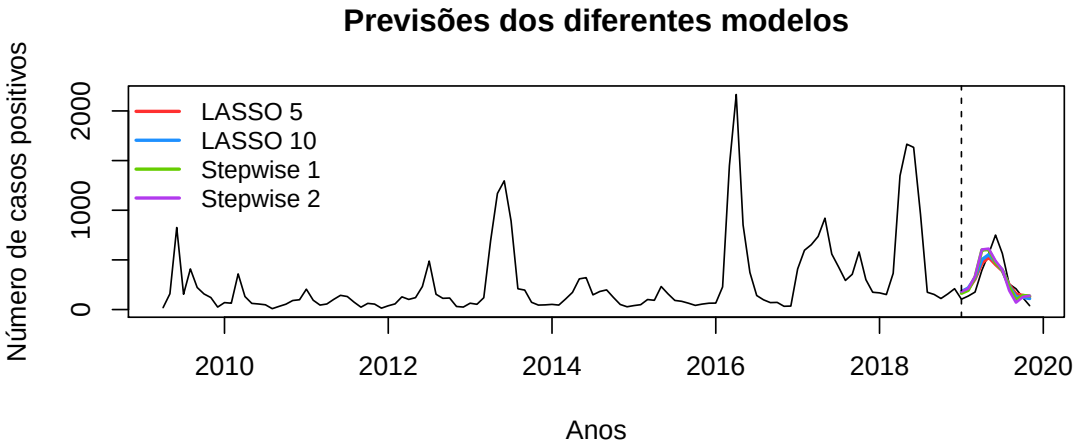


Figura 4.28: Gráfico predição $h = 11$ passos à frente.

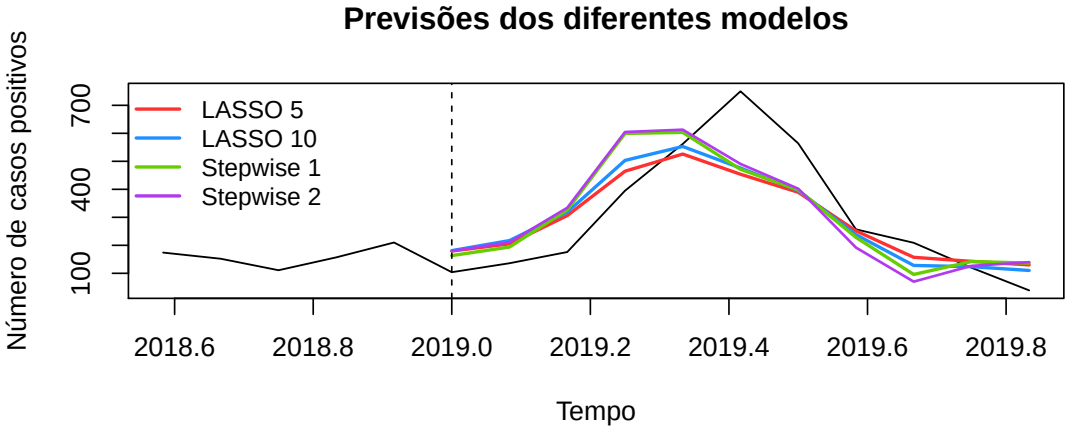


Figura 4.29: Gráfico predição $h = 11$ passos à frente (parte da série temporal).

As Figuras 4.28 e 4.29 mostram os valores verdadeiros da variável resposta B_t em preto, e os valores preditos dos diferentes modelos em vermelho, azul, verde e roxo. Analisando o ano de 2019 na Figura 4.28, nota-se que os valores preditos estão bem semelhantes com os valores verdadeiros. Todavia ao visualizar a Figura 4.29, quando comparamos os dados reais e a previsão, nota-se que a predição está boa, mas com algumas defasagens (meses) adiantados, o mesmo ocorrido na análise anterior de previsão de $h = 11$ passos a frente (Figura 4.22). Quanto aos modelos, uma comparação entre eles é feita na Tabela 4.21, que apresenta o erro quadrático médio de cada um deles e em cada um dos 11 passos.

Tabela 4.21: Erro quadrático médio em cada um dos 11 passos.

Passos/Modelos	Stepwise 1	Stepwise 2	LASSO 5	LASSO 10
1 passo à frente	**3512.23**	5763.15	5701.45	5931.42
2 passos à frente	**3428.68**	5700.51	5255.09	6250.04
3 passos à frente	9823.78	12121.87	**9153.06**	10925.23
4 passos à frente	17825.10	20082.56	**8076.48**	11122.06
5 passos à frente	14600.88	16585.83	**6721.95**	8914.02
6 passos à frente	25108.41	25039.21	20228.37	**19934.59**
7 passos à frente	25466.23	25203.61	21693.13	**21128.41**
8 passos à frente	22385.60	22583.82	18985.19	**18531.94**
9 passos à frente	21318.42	22215.48	**17174.71**	17199.67
10 passos à frente	19235.41	19998.24	15509.01	**15481.04**
11 passos à frente	18353.35	19104.09	14851.20	**14534.22**

Por meio da Tabela 4.21, percebe-se que na predição de um passo à frente (previsão para janeiro de 2019) o modelo Stepwise 1 foi o que apresentou o menor EQM e, consequentemente, o melhor resultado. Entretanto na predição de três passos à frente (previsão para janeiro de 2019, fevereiro de 2019 e março de 2019) o modelo LASSO com $p = 5$ apresentou o menor EQM. Analisando todos os passos, nota-se que o modelo que apresentou melhores resultados foi o LASSO com $p = 10$ variáveis, pois obteve o menor EQM em cinco dentre os 11 horizontes considerados.

4.3.3 Comparação das Previsões

Esta subseção tem como objetivo comparar os resultados obtidos nas Subseções 4.3.1 e 4.3.2 em relação às predições. Estamos interessados em comparar os modelos com apenas dados de incidência com os modelos com dados de incidência e de diversidade genética quanto a seu poder preditivo. Nos gráficos abaixo o termo "incidência" será utilizado para os modelos contendo apenas dados de incidência enquanto o termo "genética" será utilizado para os modelos elaborados com os dados de incidência e de diversidade genética. A Figura 4.30 apresenta os erros quadráticos médios obtidos na previsão fora da amostra (out-of-sample).

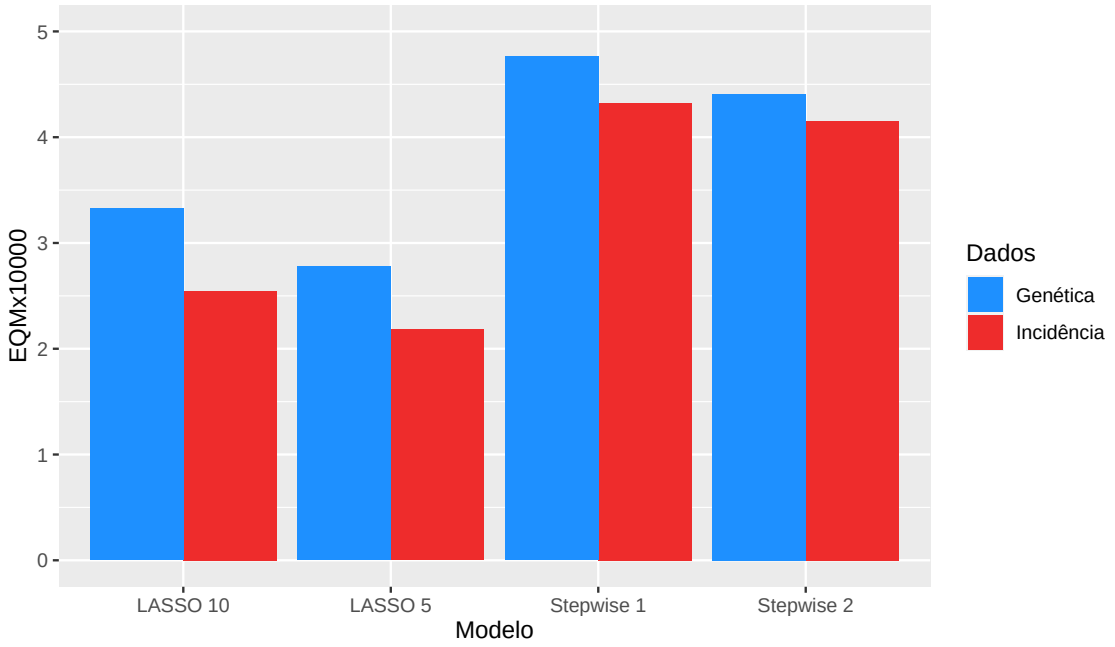


Figura 4.30: Comparação dos erros quadráticos médios (EQM) entre os dados de incidência e os dados de genética na predição $h = 1$ passo à frente.

Na Figura 4.30 percebe-se que em todas as quatro modelagens (LASSO com $p = 5$, LASSO com $p = 10$, Stepwise 1 e Stepwise 2) os dados de "incidência" foram melhores, pois apresentaram um EQM menor. Ademais nota-se que o modelo que mostrou-se preferível foi o LASSO com $p = 5$.

Em seguida comparam-se os modelos na predição de um até 11 passos à frente por meio de um gráfico de linhas. Nesse caso é necessário um gráfico para cada tipo de modelagem e nele é feita uma analogia entre os dados de "incidência" e os de "genética" apresentando os erros quadráticos médios obtidos em cada um dos 11 passos.

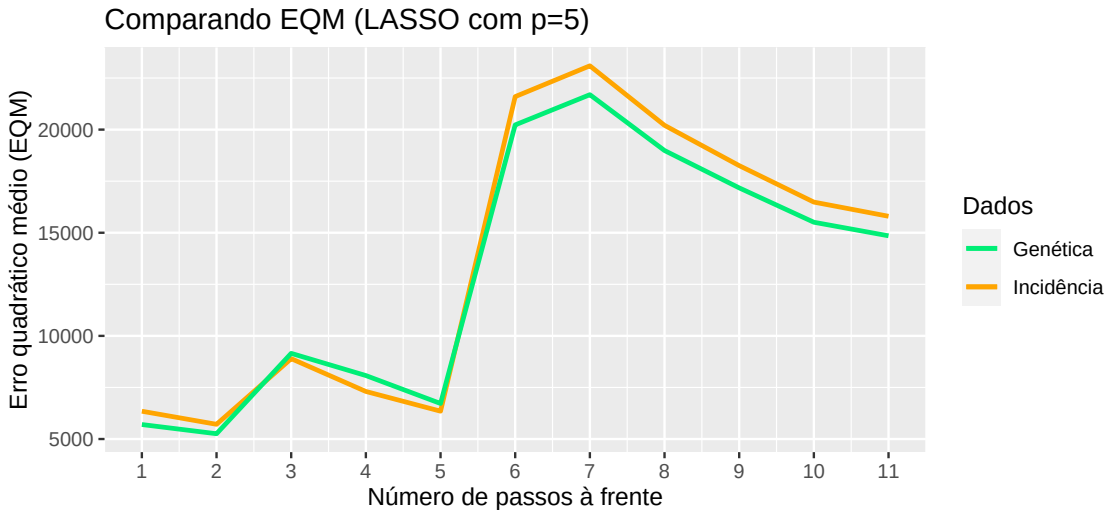


Figura 4.31: Comparação dos erros quadráticos médios (EQM) entre os dados de incidência e os dados de genética na predição $h = 11$ passos à frente (Modelo LASSO com $p = 5$).

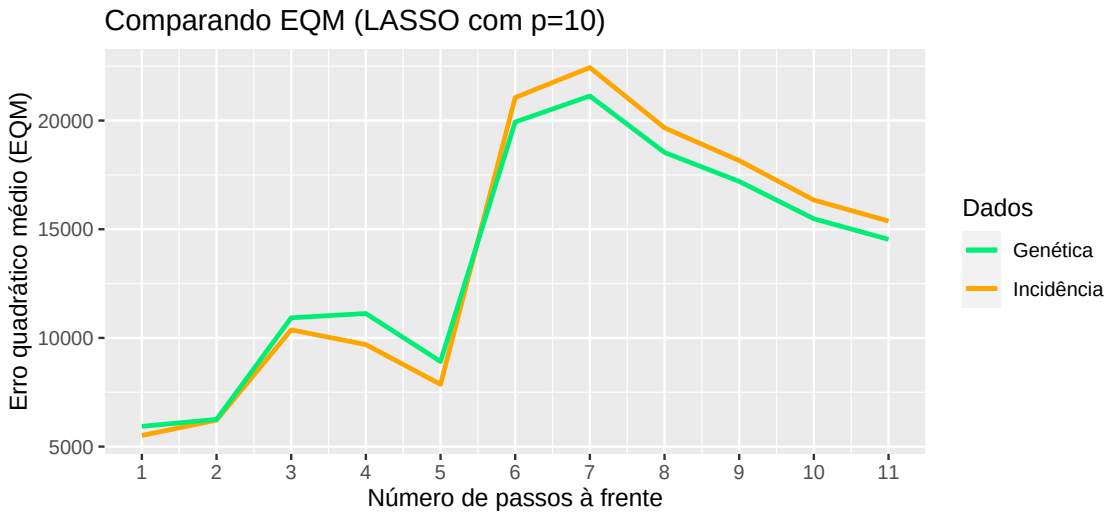


Figura 4.32: Comparação dos erros quadráticos médios (EQM) entre os dados de incidência e os dados de genética na predição $h = 11$ passos à frente (Modelo LASSO com $p = 10$).

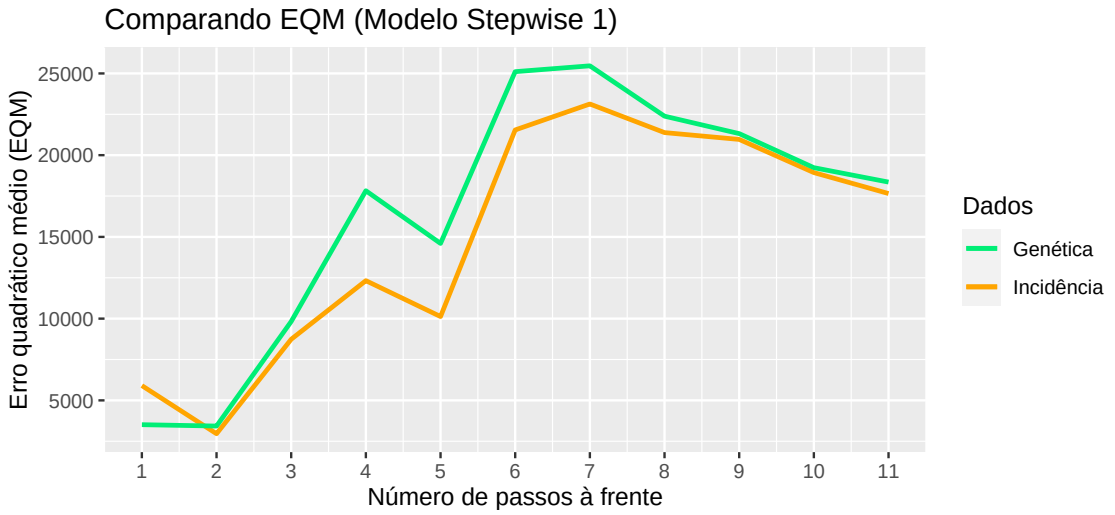


Figura 4.33: Comparação dos erros quadráticos médios (EQM) entre os dados de incidência e os dados de genética na predição $h = 11$ passos à frente (Modelo Stepwise 1).

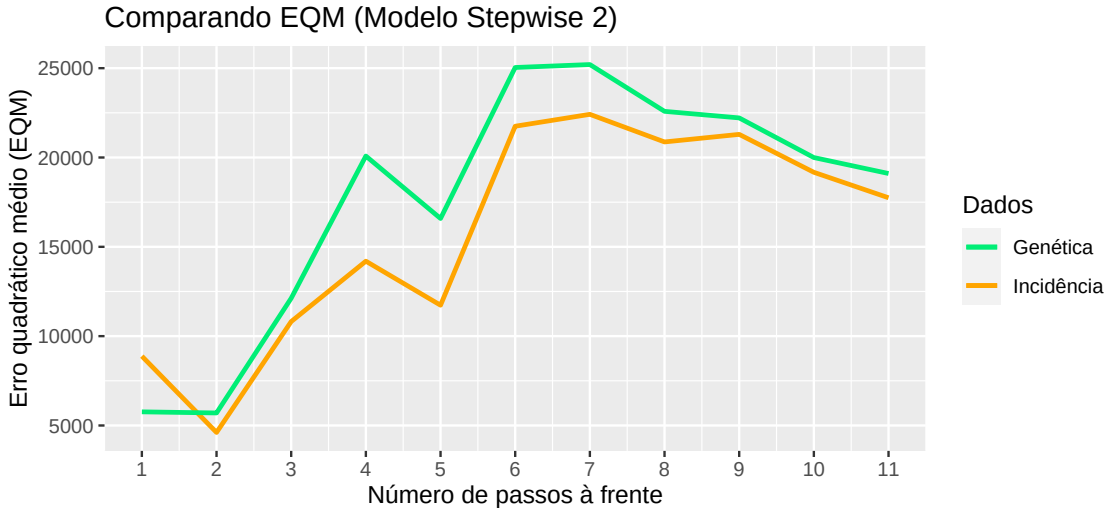


Figura 4.34: Comparação dos erros quadráticos médios (EQM) entre os dados de incidência e os dados de genética na predição $h = 11$ passos à frente (Modelo Stepwise 2).

As Figuras 4.31 a 4.34 indicam que os dados de "incidência" foram vantajosos nas modelagens Stepwise 1 e Stepwise 2, pois obtiveram um EQM menor em quase todos os passos. Todavia, nos modelos LASSO com $p = 5$ e LASSO com $p = 10$, os dados de "incidência" mostraram-se promissores na previsão de um até cinco passos à frente e os dados de "genética" apresentaram resultados mais favoráveis na predição de seis até 11 passos à frente.

5 Discussão

Este trabalho foi desenvolvido com intuito de determinar quais são as variáveis importantes para a previsão do número de casos de gripe no Brasil. O maior interesse é estudar a dinâmica migratória do vírus da gripe da América do Norte e Europa para o Brasil. O objetivo é utilizar a informação temporal (dados históricos da gripe de forma autoregressiva) para modelar o número de casos no Brasil a partir dos dados recentes de número de casos observados e da diversidade genética observados nas demais regiões.

Na Seção 4.2 (análise de Granger causalidade) descobriu-se que os valores passados da incidência da gripe da Região Europeia e da Região da América do Sul ajudam a prever o valor presente da incidência da gripe no Brasil. Percebeu-se ainda que há um efeito indireto da Região do Pacífico Ocidental e da Região da América do Norte no Brasil. Ademais, o diagrama apresentado na Figura 4.15 mostra que as diversidades genéticas da América do Norte (H3N2) e da Ásia (H1N1) Granger-causam a incidência da gripe da América Central, que Granger-causa as incidências da América do Norte, América do Sul, Europa e Pacífico Ocidental (muito provavelmente devido a pandemia da gripe A em 2009). Ademais que as incidências das regiões da América do Sul e da Europa Granger-causam a incidência no Brasil, como já mencionado anteriormente. Estes resultados são intrigantes quando pensa-se em desenvolver e atualizar vacinas no Brasil com dados relacionados às cepas provenientes da Europa e da América do Norte de estações sazonais anteriores, onde a Europa apresenta um efeito direto no Brasil e a Região da América do Norte um efeito indireto.

Quanto a análise de regressão com defasagens (Seção 4.3), descobriu-se que as variáveis que melhor explicam a incidência da gripe no Brasil são: B_{t-1} (número de casos positivos da gripe no Brasil com uma defasagem), A_{t-4} (número de casos positivos na América do Norte com quatro defasagens), E_{t-2} (número de casos positivos na Europa com duas defasagens), E_{t-3} (número de casos positivos na Europa com três defasagens), C_{t-3} (número de casos positivos na América Central com três defasagens), μ_t (média mensal no Brasil), P_{t-4} (diversidade genética da gripe H1N1 na Ásia com quatro defasagens) e P_{t-5} (diversidade genética da gripe H1N1 na Ásia com cinco defasagens).

Quanto à predição $h = 1$ passo à frente, em ambas análises das Subseções 4.3.1 e 4.3.2, o modelo que melhor previu a incidência da gripe no Brasil (variável resposta) foi o LASSO com $p = 5$ variáveis. Quanto à predição $h = 11$ passos à frente, em ambas análises apresentadas, o melhor modelo foi o LASSO com $p = 10$ variáveis. Quanto ao resultado das previsões, percebeu-se que tanto os resultados para os hori-

zontes de previsão $h = 1$ quanto $h = 11$ passos à frente foram interessantes, pois de um modo geral, as previsões foram boas, e elas podem ser úteis no desenvolvimento de políticas públicas de imunização contra a gripe e na atualização de vacinas.

Por fim, na Subseção 4.3.3 estudou-se comparativamente o poder preditivo de modelos com apenas dados de incidência e modelos com dados de incidência e de diversidade genética. No caso da predição $h = 1$ passo à frente observou-se que os modelos da Subseção 4.3.1 se saíram melhor, pois apresentaram menores EQM's em todos os modelos. Quanto à previsão $h = 11$ passos à frente, nas modelagens LASSO com $p = 5$ e com $p = 10$ variáveis os dados da Subseção 4.3.1 se mostraram mais promissores, e nas modelagens Stepwise 1 e Stepwise 2, de um até cinco passos à frente os modelos da Subseção 4.3.1 se mostraram favoráveis, enquanto que de seis até 11 passos à frente as modelagens da Subseção 4.3.2 foram preferíveis. Em futuras contribuições procura-se possíveis novas abordagens que possam melhorar as previsões.

Referências Bibliográficas

Akaike, H. (1987). Factor analysis and AIC. Psychometrika, 52(3):317–332.

Barr, I. G., McCauley, J., Cox, N., Daniels, R., Engelhardt, O. G., Fukuda, K., Grohmann, G., Hay, A., Kelso, A., Klimov, A., Odagiri, T., Smith, D., Russell, C., Tashiro, M., Webby, R., Wood, J., Ye, Z., e Zhang, W. (2010). Epidemiological, antigenic and genetic characteristics of seasonal influenza A(H1N1), A(H3N2) and B influenza viruses: Basis for the WHO recommendation on the composition of influenza vaccines for use in the 2009–2010 Northern Hemisphere season. Vaccine, 28(5):1156–1167.

Born, P. S. (2013). Análises filogenéticas e filogeográficas dos vírus influenza A(H3N2) papel do Brasil no cenário de dispersão global e ajuste temporal entre as cepas vacinais e os vírus circulantes no período de 1999 a 2012. PhD thesis, Instituto Oswaldo Cruz.

Chen, C. W. S., Hsieh, Y.-H., Su, H.-C., e Wu, J. J. (2018). Causality test of ambient fine particles and human influenza in taiwan: Age group-specific disparity and geographic heterogeneity. Environment International, 111:354–361.

Dantas, F. e Weydmann, C. L. (2009). Carne de frango: uma análise da relação entre os preços dos produtores e de exportação. Revista de Economia e Agronegócio, 7(1):31–53.

Dickey, D. A. e Fuller, W. A. (1981). Likelihood ratio statistics for autoregressive time series with a unit root. Econometrica: Journal of the Econometric Society, 49(4):1057–1072.

Diniz, M. B., Junior, J. N. O., Neto, N. T., e Diniz, M. J. T. (2009). Causas do desmatamento da Amazônia: uma aplicação do teste de causalidade de Granger acerca das principais fontes de desmatamento nos municípios da Amazônia Legal brasileira. Nova Economia, 19(1):121–151.

Eccles, R. (2005). Understanding the symptoms of the common cold and influenza. The Lancet Infectious Diseases, 5(11):718–725.

Farias, H. P. e Sáfadi, T. (2010). Causalidade entre as principais bolsas de valores do mundo. RAM. Revista de Administração Mackenzie, 11(2):96–122.

Forleo-Neto, E., Halker, E., Santos, V. J., Paiva, T. M., e Toniolo-Neto, J. (2003). Influenza. Revista da Sociedade Brasileira de Medicina Tropical, 36(2):267–274.

Garten, R. J., Davis, C. T., Russell, C. A., Shu, B., Lindstrom, S., Balish, A., Sessions, W. M., Xu, X., Skepner, E., Deyde, V., Okomo-Adhiambo, M., Gubareva, L., Barnes, J., Smith, C. B., Emery, S. L., Hillman, M. J., Rivailler, P., Smagala, J., de Graaf, M., Burke, D. F., Fouchier, R. A. M., Pappas, C., Alpuche-Aranda, C. M., López-Gatell, H., Olivera, H., López, I., Myers, C. A., Faix, D., Blair, P. J., Yu, C., Keene, K. M., Jr., P. D. D., Boxrud, D., Sambol, A. R., Abid, S. H., George, K. S., Bannerman, T., Moore, A. L., Stringer, D. J., Blevins, P., Demmler-Harrison, G. J., Ginsberg, M., Kriner, P., Waterman, S., Smole, S., Guevara, H. F., Belongia, E. A., Clark, P. A., Beatrice, S. T., Donis, R., Katz, J., Finelli, L., Bridges, C. B., Shaw, M., Jernigan, D. B., Uyeki, T. M., Smith, D. J., Klimov, A. I., e Cox, N. J. (2009). Antigenic and Genetic Characteristics of Swine-Origin 2009 A(H1N1) Influenza Viruses Circulating in Humans. Science, 325(5937):197–201.

Granger, C. W. J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. Econometrica: Journal of the Econometric Society, 37(3):424–438.

Hastie, T., Tibshirani, R., e Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, 2 edition.

Ibiapina, C. C., Costa, G. A., e Faria, A. C. (2005). Influenza A aviária (H5N1) - A gripe do frango. Jornal Brasileiro de Pneumologia, 31(5):436–444.

Jesus, R. G. (2018). Caracterização e visualização da diversidade genética do vírus influenza ao longo do tempo. Monografia (Bacharel em Estatística), UFRGS (Universidade Federal do Rio Grande do Sul), Porto Alegre.

Kaji, M., Watanabe, A., e Aizawa, H. (2003). Differences in clinical features between influenza A H1N1, A H3N2, and B in adult patients. Respirology, 8(2):231–233.

Linderman, S. L., Chambers, B. S., Zost, S. J., Parkhouse, K., Li, Y., Herrmann, C., Ellebedy, A. H., Carter, D. M., Andrews, S. F., Zheng, N.-Y., Huang, M., Huang, Y., Strauss, D., Shaz, B. H., Hodinka, R. L., Reyes-Terán, G., Ross, T. M., Wilson, P. C., Ahmed, R., Bloom, J. D., , e Hensley, S. E. (2014). Potential antigenic explanation for atypical H1N1 infections among middle-aged adults during the 2013–2014 influenza season. Proceedings of the National Academy of Sciences, 111(44):15798–15803.

NCBI (2020). National Center for Biotechnology Information. Last accessed 12 June 2020.

Phillips, P. C. B. e Ouliaris, S. (1990). Asymptotic Properties of Residual Based Tests for Cointegration. Econometrica: Journal of the Econometric Society, 58(1):165–193.

Phillips, P. C. B. e Perron, P. (1988). Testing for a unit root in time series regression. Biometrika, 75(2):335–346.

R Core Team (2020). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rambaut, A. e Holmes, E. (2009). The early molecular epidemiology of the swine-origin A/H1N1 human influenza pandemic. PLoS Currents, 1:RRN1003.

Rambaut, A., Pybus, O. G., Nelson, M. I., Viboud, C., Taubenberger, J. K., e Holmes, E. C. (2008). The genomic and epidemiological dynamics of human influenza A virus. Nature, 453(7195):615–619.

Russell, C. A., Jones, T. C., Barr, I. G., Cox, N. J., Garten, R. J., Gregory, V., Gust, I. D., Hampson, A. W., Hay, A. J., Hurt, A. C., de Jong, J. C., Kelso, A., Klimov, A. I., Kageyama, T., Komadina, N., Lapedes, A. S., Lin, Y. P., Mosterin, A., Obuchi, M., Odagiri, T., Osterhaus, A. D. M. E., Rimmelzwaan, G. F., Shaw, M. W., Skepner, E., Stohr, K., Tashiro, M., Fouchier, R. A. M., e Smith, D. J. (2008). The Global Circulation of Seasonal Influenza A (H3N2) Viruses. Science, 320(5874):340–346.

Silva, P. C. R. (2015). Dinâmica molecular dos vírus Influenza A (H1N1) pandêmico em cinco anos de circulação no Brasil. Master's thesis, Instituto Oswaldo Cruz.

Sims, C. A. (1980). Macroeconomics and Reality. Econometrica, 48(1):1–48.

Toda, H. Y. e Yamamoto, T. (1995). Statistical inference in vector autoregressions with possibly integrated processes. Journal of Econometrics, 66(1-2):225–250.

WHO (2020). FluNet. Last accessed 12 June 2020.

Yamada, K. D., Tomii, K., e Katoh, K. (2016). Application of the MAFFT sequence alignment program to large data—reexamination of the usefulness of chained guide trees. Bioinformatics, 32(21):3246–3251.