Universidade Federal de São Carlos
Centro de Ciências Exatas e de tecnologia
Departamento de Estatística

Estudo do teste SNHT (Standard Normal Homogeneity Test) para detecção de pontos de mudança em séries temporais

Giovanna Garutti Chianezzi

Trabalho de Conclusão de Curso

Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Estudo do teste SNHT (Standard Normal Homogeneity Test) para detecção de pontos de mudança em séries temporais

Giovanna Garutti Chianezzi
Orientador: Maria Sílvia de Assis Moura

Trabalho de Conclusão de Curso a ser apresentado como parte dos requisitos para obtenção do título de Bacharel em Estatística.

São Carlos
30 de Junho de 2021

Giovanna Garutti Chianezzi

Estudo do teste SNHT (Standard Normal Homogeneity Test) para detecção de pontos de mudança em séries temporais

Banca Examinadora

- Maria Sílvia de Assis Moura
- Maria Aparecida de Paiva Franco
- Luis Aparecido Milan

Agradecimentos

Aos meus pais Valdir e Silvana que sempre estiveram ao meu lado me apoiando ao longo de toda a minha trajetória e ofereceram apoio e incentivo aos estudos que serviram de base para as minhas realizações.

Aos amigos de São Carlos, que sempre estiveram ao meu lado, pela amizade incondicional e pelo apoio demonstrado ao longo de todo o período de faculdade.

Deixo um agradecimento especial a minha orientadora, Maria Sílvia de Assis Moura pelo incentivo e por sempre estar presente para indicar a direção que o trabalho deveria tomar.

A Alexandersson H., pelo fornecimento de dados e materiais que foram fundamentais para o desenvolvimento do estudo que possibilitou a realização deste trabalho.

Por último, quero agradecer também à Universidade Federal de São Carlos e todo o seu corpo docente.

Resumo

O estudo de séries temporais é de extrema importância para uma melhor compreensão de como se desenvolvem determinados eventos ao longo do tempo. Para isso, ser capaz de indicar o momento exato em que um determinado fenômeno muda o seu padrão de comportamento é um recurso muito interessante.

Este trabalho se aprofunda na temática da detecção de pontos de mudança em séries temporais por meio da utilização do SNHT (Standard Normal Homogeneity Test), que consiste em um teste estatístico proposto especificamente para este propósito.

Alguns conceitos estatísticos básicos e o teste em sí são explicados detalhadamente. A exemplificação do teste é realizada através da sua aplicação em dados a respeito do número de motoristas mortos ou seriamente feridos em acidentes de trânsito na Grã-Bretanha entre os anos de 1969 - 1984.

Um experimento, a partir de séries de médias móveis simuladas, será realizado para que seja possível ter uma boa noção a respeito do poder do teste e seu desempenho. Esta analise será feita por meio da análise das taxas de *Erro tipo II* obtidas pelo SNHT.

E para finalizar o estudo por completo, o SNHT será aplicado em dados referentes ao número de casos novos de Covid-19 na cidade de São Paulo, na tentativa de se entender melhor o comportamento da pandemia da Covid-19 durante o seu primeiro ano.

Palavras-chave: *Dados Covid-19, Estatística, Ponto de Mudança, Quebra Estrutural, Séries Temporais, SNHT, Teste de Hipóteses.*

Sumário

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução e Objetivos

1.1 Introdução

A necessidade de estudar e compreender melhor ocorrências ao longo da história como, por exemplo, a volumetria de um rio ou a temperatura de uma cidade no decorrer dos anos, deu início ao estudo de séries temporais, que hoje consiste em um importante campo da análise estatística.

A análise de séries temporais se tornou de extrema importância para a compreensão dos mais variados tipos de acontecimentos, desde fenômenos naturais a até mesmo o preço de ativos negociados na bolsa de valores.

Tal relevância pode ser explicada devido ao grande interesse em se entender padrões de comportamento e, principalmente, em ser capaz de gerar previsões de alta confiabilidade acerca de acontecimentos com potencial de causar grande impacto social.

Dessa forma, é clara a necessidade de monitorar a evolução de fatos como o comportamento das ondas do mar ao longo do tempo, para ser possível compreender o seus padrões de movimento durante diferentes estações do ano e ter a capacidade de identificar momentos em que mudanças abruptas ocorrem.

Estudos como esse podem ser feitos a partir da percepção de uma mudança de comportamento dos dados temporais. O que facilmente ocorreria através da detecção de pontos de quebra na série, ou seja, na detecção de momentos em que o evento observado apresenta uma mudança brusca de comportamento.

O entendimento deste evento será possibilitado pela análise minuciosa dos dados temporais e da detecção da presença de pontos de quebra, que causam grandes mudanças na evolução de determinado acontecimento.

A detecção de pontos de mudança em séries temporais pode ser realizada por meio do SNHT (Standard Normal Homogeneity Test), teste que foi originalmente utilizado por Alexandersson (1986) para o estudo de dados climáticos. Devido as características das séries em estudo, o teste foi desenvolvido para ser aplicado em uma série de diferenças gerada a partir da série em estudo em relação a uma série de referência, formada por dados muito semelhantes ao dados em teste, porém com um comportamento mais homogêneo.

Mais recentemente, Haimberger (2007) apresentou o SNHT modificado, chamado de SNHT para amostras de mesmo tamanho, modificação que será comentada e estudada de forma detalhada através de dados simulados mais a frente neste trabalho.

O tipo tipo de análise em questão pode ser muito interessante para o melhor entendimento dos padrões de evolução da pandemia de Covid-19, ocasionada pelo novo corona vírus, iniciada no ano de 2020 e que já impactou a vida de milhões de pessoas ao redor do mundo, mas principalmente a vida dos brasileiros.

No próximo capítulo será comentado quais são os objetivos da realização deste estudo. No capítulo três serão introduzidos alguns conceitos estatísticos importantes para o melhor entendimento do tema abordado. Posteriormente, a metodologia em análise será aplicada em um conjunto de dados, para exemplificar como deve ser utilizada na prática. Então, um experimento com dados simulado será realizado par o melhor compreendimento do poder do teste. Para finalizar, no capítulo seis, o SNHT será aplicado em dados referentes ao número de casos da Covid-19 na cidade de São Paulo e algumas conclusões a respeito da evolução da pandemia poderão ser tiradas.

1.2 Objetivos

Este trabalho tem como objetivo revisar conceitos básicos a respeito do estudo de séries temporais, estudar com mais profundidade meios de detectar se estas possuem pontos de mudança e localizá-los, caso existam.

Para isso, o teste SNHT (Standard Normal Homogeneity Test) será discutido a fundo, com o interesse de entender o seu funcionamento em detalhes e encontrar a melhor forma de utilizá-lo, de acordo com o objetivo do estudo e com as particularidades da série temporal em observação.

O desempenho do teste será medido por meio da sua aplicação em diversas séries de médias móveis com pontos de quebra localizados em posições já previamente conhecidas

geradas por simulação no software R e análise das suas taxas de ocorrência do *Erro tipo II*.

O teste também será aplicado, posteriormente, em dados reais referentes ao número de novos casos registrados de Covid-19 na cidade de São Paulo durante o primeiro ano de pandemia, com o objetivo de se compreender melhor a como se deu a sua evolução ao longo das semanas e analisar seus padrões de comportamento.

Capítulo 2

Metodologia

2.1 Séries temporais

Segundo Morettin e Toloi (2004), uma série temporal consiste em um conjunto de dados coletados ao longo de determinado tempo, afim de monitorar o comportamento do evento em estudo durante este período. De forma prática, trata-se de uma sequência de observações ordenadas, geralmente coletadas em intervalos uniformes, que possibilitam a análise e o entendimento de tal evento em diferentes pontos no tempo.

Uma determinada série de tamanho n pode ser denotada da seguinte maneira: $Y = (Y_1, Y_2, ..., Y_n)$. De forma análoga, Y_i representa um determinado ponto da série Y tal que $i = 1, 2, ..., n$.

Morettin e Toloi (2004) comentam que séries temporais tratam-se de processos estocásticos que, por sua vez, consistem em processos guiados por leis da probabilidade. Tais processos podem ser definidos como sendo uma família de variáveis aleatórias supostamente definidas num mesmo espaço de probabilidades.

As séries temporais ditas estacionárias ou convergentes flutuam em torno da mesma média e possuem variância constante ao longo de todo o período, assim, não apresentam drásticas mudanças de comportamento. Caso contrário, elas são denominadas não estacionárias ou divergentes.

Também de acordo com Morettin e Toloi (2004), uma série temporal pode possuir tendência, o que consiste em apresentar uma mudança crescente ou decrescente em sua média a longo prazo. Entretanto, é difícil de se definir longo prazo. Vale ressaltar que a tendência pode ser uma mudança linear ou não linear.

De acordo com Morettin e Toloi (2004), diferenciar uma série é uma boa forma de

remover a sua componente de tendência e, assim, transformá-la em uma série estacionária, ou seja, em uma série com média constante em todo o período.

O efeito sazonal é definido por Morettin e Toloi (2004) como um comportamento se repete ao longo do tempo em uma série e que pode ser eliminado por meio de análises de médias sazonais.

2.2 Teste de hipóteses

Para Bolfarine e Sandoval (2001) o teste de hipóteses é uma boa alternativa quando existe a necessidade de se tomar uma decisão de rejeitar ou não rejeitar determinada hipótese, com base em um conjunto de evidências.

A hipótese a ser testada é chamada de H_0 (Hipótese nula) e a hipótese que se contrapõe a ela é chamada de H_1 (Hipótese alternativa). A partir de evidências e com um certo nível de confiança estatístico, que consiste em uma certa porcentagem de intervalos necessários para incluir o parâmetro populacional se fossem repetidamente reunidas amostras da mesma população, pode-se então não rejeitar H_0, determinando se esta hipótese está correta ou pode-se rejeitar H_0 em favor da hipótese alternativa.

Apesar da alta confiabilidade estatística, erros podem acontecer de duas maneiras e Bolfarine e Sandoval (2001) os definem da seguinte forma:

- Erro tipo I: Rejeitar H_0, quando H_0 é verdadeiro;
- Erro tipo II: Não rejeitar H_0, quando H_0 não é verdadeiro.

2.3 SNHT (Standard Normal Homogeneity Test)

O teste SNHT foi desenvolvido por Alexandersson (1986) e trata-se de um teste paramétrico, que é muito usado para detecção de pontos de mudança, identificação de tendência não homogênea e alteração de variância em séries temporais.

Conforme Alexandersson (1986), o SNHT encontra a série de diferenças entre a série em estudo e as séries de referência vizinhas. Então, para um determinado ponto t tal que $t \in (1, ..., n)$ na série de diferenças de tamanho n, calcula-se a média de todas as observações anteriores ao ponto t e também a média de todas as observações posteriores para que elas possam ser comparadas.

Alexandersson e Moberg (1997) definem Y como sendo a série candidata em estudo e Y_i como um valor específico da série em questão, enquanto X_j é definido como a j-ésima série de um conjunto de k séries de referência utilizadas no estudo e, analogamente, X_{ji} refere-se à um ponto específico de tal série. Para detectar não homogeneidades relativas, formam-se razões dadas por:

$$\mathrm{Q}_i = Y_i - \frac{\sum_{j=1}^{k} \rho_j^2 X_{ji} \overline{Y} / \overline{X_j}}{\sum_{j=1}^{k} \rho_j^2},$$

em que:

- ρ_j é o coeficiente de correlação entre Y e X_j (deve sempre ser positivo);
- $\overline{Y}$ é o valor médio da série Y;
- $\overline{X_j}$ é o valor médio da série X_j,
- Q_i é a i-ésima observação da série de diferenças Q.

A normalização dos dados é necessária, uma vez que, permite a comparação de séries provindas de períodos distintos e também em unidades de medida diferentes. Além disso, faz com que os valores de Q se distribuam em torno de zero.

É muito importante que os valores de $\overline{Y}$ e $\overline{X_j}$ sejam calculados para o mesmo período de tempo, garantindo que o tamanho das não homogeneidades seja estimado corretamente. Por outro lado, os coeficientes ρ_j não precisam obrigatoriamente serem calculados para períodos em comum, apesar de que também é recomendado que sejam.

A série então é normalizada da seguinte forma para que o teste de homogeneidade possa ser aplicado:

$$Z_i = \frac{Q_i - \overline{Q}}{\sigma_Q},$$

em que Z_i representa a série Q com n observações normalizada e σ_Q consiste no desvio padrão da mesma série e é dado por:

$$\sigma_Q = \sqrt{\frac{\sum_{i=1}^{n}(Q_i - \overline{Q})^2}{n-1}}.$$

Pode-se então aplicar o teste SNHT para a detecção de um deslocamento de nível médio, cujas hipóteses consistem em:

$$H_0 : Z_i \sim N(0, \sigma) \quad i \in [1, ..., n] \qquad H_1 : \begin{cases} Z_i \sim N(\mu_1, \sigma) & i \in [1, ..., a] \\ Z_i \sim N(\mu_2, \sigma) & i \in [a+1, ..., n] \end{cases}$$

Assumindo que pode-se utilizar a distribuição normal, denotada por N, Alexandersson e Moberg (1997) consideram que a hipótese nula representa o caso de uma série homogênea, pois apresenta uma média constante e, por sua vez, não possui pontos de mudança. Em contrapartida, a hipótese alternativa representa a possibilidade da série ter sua média alterada em algum momento e, assim, sofrer alguma mudança abrupta em uma determinada posição.

Já o desvio padrão deve, em ambas as hipóteses, permanecer inalterado e, como via de regra, deve ser levemente inferior a 1.

A partir de cálculos realizados por Alexandersson (1986) a respeito da diferença de probabilidades de que H_1 é correto, dada a série observada Z_i, para a probabilidade de que H_0 está correto, foi gerada a seguinte estatística teste:

$$T_{\max} = \max_{1 \leqslant a \leqslant n-1} \{T_a\} = \max_{1 \leqslant a \leqslant n-1} \{a\overline{z}_1^2 + (n-1)\overline{z}_2^2\},$$

em que:

- $\overline{z}_1$ é a média da série antes do ponto de mudança;
- $\overline{z}_2$ é a média da série após o ponto de mudança;
- a é o valor máximo (última posição antes do ponto de mudança).

Se $T^s_{\max}$ encontra-se acima de um dado valor crítico, rejeita-se a H_0 para um determinado nível de significância. As duas diferenças de antes e depois do possível ponto de mudança são dados por:

$$\overline{q}_1 = \sigma_Q \overline{z}_1 + \overline{Q},$$

$$\overline{q}_2 = \sigma_Q \overline{z}_2 + \overline{Q}.$$

A aplicação deste teste não é indicada para detecção de múltiplos pontos de quebra, uma vez que, os testes não seriam realizados de maneira independente. Apesar de existir a possibilidade de generalizar o teste para estes casos de detecção múltipla, é mais recomendado pelo autor que ele seja utilizado em duas ou mais partes consecutivas de uma série, conforme pontos de quebra forem sendo encontrados.

2.4 SNHT para amostras de mesmo tamanho

No teste que foi inicialmente proposto por Alexandersson (1986), os valores de $\overline{z}_1$ e $\overline{z}_2$ deveriam ser calculados a partir de uma quantidade diferente de observações, dependendo da posição em que o ponto t estivesse colocado na série. Porém, para Haimberger (2007) o teste poderia ser melhorado ao se limitar a quantidade de observações que seriam consideradas no cálculo da média antes e depois do ponto t.

Haimberger (2007) explica que, dessa forma, dois problemas existentes na versão original do teste seriam resolvidos. O primeiro é o fato de que o teste possui uma tendência maior em detectar pontos de quebra em posições mais próximas as bordas da série, no inicio e no fim. E o segundo consiste na dificuldade do teste em apontar corretamente o momento exato da quebra quando há presença de sazonalidade.

Dessa maneira, passa a ser possível controlar a quantidade de informação histórica que será considerada para a comparação das médias em qualquer ponto na série. Ou seja,

as médias serão sempre feitas com base no mesmo número p de observações e, assim, os valores serão comparados de maneira mais justa.

Com essa alteração o teste passou a ter mais um parâmetro, que consiste no período que se deseja utilizar para considerar as observações vizinhas ao ponto t que devem ser utilizadas para o cálculo das médias anterior e posterior. Logo, se um período igual a p é escolhido, a estatística teste só poderá ser calculada para pontos Y_i tal que $Y_p \leq Y_i \leq Y_{n-p}$.

Também é valido salientar que, a partir desta mudança, a estatística do teste $T_{\max}$ deixa de ser um valor máximo global e passa a ser o valor máximo local da estatística dentro de cada intervalo que está sendo considerado. Para isso, a estatística $T_{\max}$ passa a ser $T^s_{\max}$ e pode ser descrita da seguinte maneira:

$$T^s_{\max} = \max_{p \leqslant a \leqslant n-p} \quad p((\overline{z}_1 - \overline{z})^2 + (\overline{z}_2 - \overline{z})^2),$$

em que:

- $\overline{z}_1$ é a média da série antes do ponto de mudança;
- $\overline{z}_2$ é a média da série após o ponto de mudança;
- $\overline{z}$ é a média da série inteira;
- p é o número de observações antes e depois do ponto t.

Outra grande mudança na forma com que a estatística do teste é calculada também foi utilizada por Haimberger (2007), alterando a forma com que os dados são inseridos no SNHT. Em vez de se aplicar o teste em uma série de diferenças, gerada através da série em estudo e de outras séries de referência, foi proposto que a série em estudo fosse diretamente testada pelo SNHT, sem se utilizar nenhum outro dado como referência.

Quanto ao cálculo da região crítica do teste, Haimberger (2007) simulou cinco mil séries que não possuíam nenhum ponto de mudança e, posteriormente, aplicou o SNHT em todas elas. A partir deste experimento, o pesquisador pôde observar a taxa de ocorrência do *Erro tipo I* de acordo com os valores que a estatística do teste atingia, o que permitiu que chegasse ao valor limite da região crítica igual a 9.6 com 95% de confiança.

2.5 Modelo de médias móveis MA

Segundo Costa (2019), os modelos de médias móveis são modelos de séries temporais para dados univariados, que possuem estacionariedade fraca, uma vez que, consistem em uma combinação linear de ruídos brancos a_i, em que, a_i é caracterizado por ter média zero, variância constante e ser não-autocorrelacionado. O modelo de médias móveis MA(q), pertencente a família ARMA(p,q), pode ser escrito da seguinte forma:

$$Y_i = \mu + a_i + \phi_1 a_{t-1} + \phi_2 a_{t-2} + ... + \phi_q a_{t-q},$$

com $i = 1, 2, ..., q$, em que:

- Y_i é a i-ésima observação gerada pelo modelo;
- μ é a média da série Y;
- a_i é ruído branco;
- ϕ é o parâmetro do modelo.

Neste trabalho, serão utilizadas séries simuladas através do MA(1), que é dado por:

$$Y_i = \mu + a_i + \phi_1 a_{t-1},$$

apresenta média μ constante e variância finita.

Capítulo 3

Pacotes e definições específicas

3.1 SNHT modificado em estudo

Como foi previamente comentado, este teste é geralmente aplicado em séries de diferenças geradas através da série em estudo e de outras séries utilizadas como referência. Porém, neste trabalho será realizado um estudo específico do SNHT para tentar entender o quão bem ele consegue performar, quando aplicado diretamente na série em estudo, sem se utilizar nenhum outro dado como referência, como foi anteriormente feito por Haimberger (2007) em seu experimento simulado.

Também será estudada uma outra adaptação, desta vez nas hipóteses do teste, que foram originalmente propostas com o objetivo de testar se a série possui pontos de quebra ou não. Entretanto, desta vez, o teste será analisado com um objetivo que vai um pouco mais além do que foi inicialmente proposto, deseja-se verificar com qual nível de assertividade o teste é capaz de detectar o momento exato em que o ponto de mudança ocorre na série.

Este feito será possível através da simulação de diversas séries de médias móveis, em que existem pontos de quebra em posições já conhecidas. Ao se aplicar o SNHT para amostras de mesmo tamanho nestas séries, as seguintes hipóteses serão testadas:

$$H_0 : Z_i \sim N(0,1) \qquad\qquad H_1 : \begin{cases} Z_i \sim N(\mu_1, 1) \\ Z_i \sim N(\mu_2, 1), \end{cases}$$

em que i é um ponto fixo em uma posição já conhecida.

3.2 Pacote SNHT no RStudio

O teste em estudo pode ser aplicado por meio do pacote SNHT presente no RStudio e disponível para consulta em Team (2020). No pacote em questão, a versão do SNHT para amostras de mesmo tamanho pode ser utilizado com muita facilidade e praticidade, portanto, foi o escolhido para ser usado neste trabalho.

Conforme dito em Browning (2017), sob a hipótese nula de que a série não possui pontos de mudança e de que os erros aleatórios são normais, a distribuição da estatística teste pode ser aproximada pela qui-quadrado com um grau de liberdade e, assim, neste pacote, a região crítica do teste é calculada a partir de tal distribuição.

Capítulo 4

Exemplo de aplicação

4.1 Apresentação dos dados

Para exemplificar a aplicação do SNHT neste trabalho, será utilizada uma série temporal, que foi previamente analisada por Harvey e Durbin (1986) e está disponível em Team (2020), contendo o número mensal de motoristas que foram mortos ou que ficaram seriamente feridos em acidentes de trânsito na Grã-Bretanha durante o período de Janeiro de 1969 à Dezembro de 1984.

Sabe-se que em 31 de Janeiro de 1983 foi introduzida a lei do uso obrigatório de cinto de segurança na Grã-Bretanha, o que possivelmente causou efeito direto no número de motoristas mortos ou seriamente feridos em acidentes de trânsito.

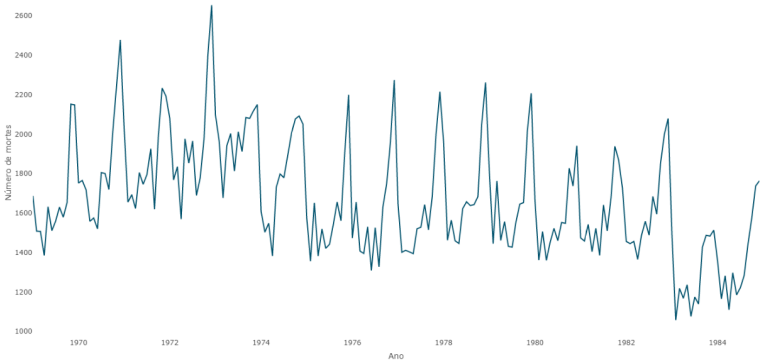

Figura 4.1: Série temporal do número de motoristas mortos ou seriamente feridos na Grã-Bretanha

A partir da figura 4.1, é possível dizer que a série possui uma tendência crescente durante os anos de 1969 à 1974. Após este período, a série apresentou uma forte queda e se manteve então constante até o ano de 1983, onde novamente ocorreu uma queda

significativa no número de motoristas mortos ou seriamente feridos na Grã-Bretanha.

Nota-se também que durante todo o período de coleta de dados, a série apresenta forte indício de que o seu comportamento é sazonal. Uma vez que, ela possui um padrão de obter picos ao longo de todos os anos. .

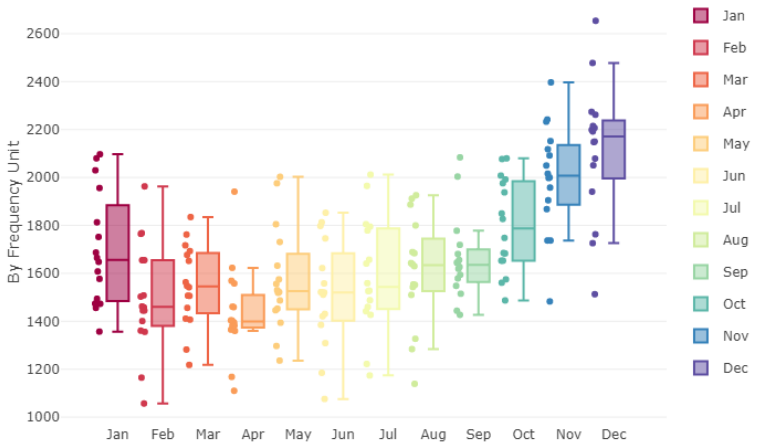

Figura 4.2: Gráfico de boxplots da série temporal de acordo com os meses do ano

A partir da figura 4.2, fica claro que existe tendência ligada aos meses do ano, visto que a distribuição do número de motoristas mortos ou seriamente feridos varia muito de acordo com os meses. De Janeiro a Abril, percebe-se que, em média, ocorre uma queda nos valores observados, de Abril a Julho eles se mantém constantes e de Julho a Dezembro passam a crescer.

Também vale ressaltar que alguns meses apresentam uma variabilidade bem superior em relação a outros. Por exemplo, valores observados nos meses de Abril e Setembro não variam muito ao longo dos anos, enquanto os valores em meses como Fevereiro apresentam alta variabilidade.

4.2 Aplicação do teste

Um fato importante a se levar em consideração na escolha do tamanho do período em análise, é de que a série possui sazonalidade de acordo com os meses do ano e, dessa forma, a suposição de normalidade dos erros aleatórios pode ser prejudicada. Com isso, seria mais vantajoso realizar o teste com período igual a doze, dado que é o intervalo de tempo que a série leva para repetir seu padrão sazonal.

O gráfico a seguir demonstra como ficou a estatística do teste em relação à sua região

crítica com período igual a doze e 95% de confiança.

Vale ressaltar que o pacote calcula a região crítica com base na distribuição qui-quadrado com um grau de liberdade e 95% de confiança, portanto, a linha traceja no gráfico, que corresponde ao limite da região crítica, está colocada no valor $3,85$. Porém, para as análises deste trabalho o valor encontrado por Haimberger (2007) de $9,6$ para 95% de confiança é que será considerado.

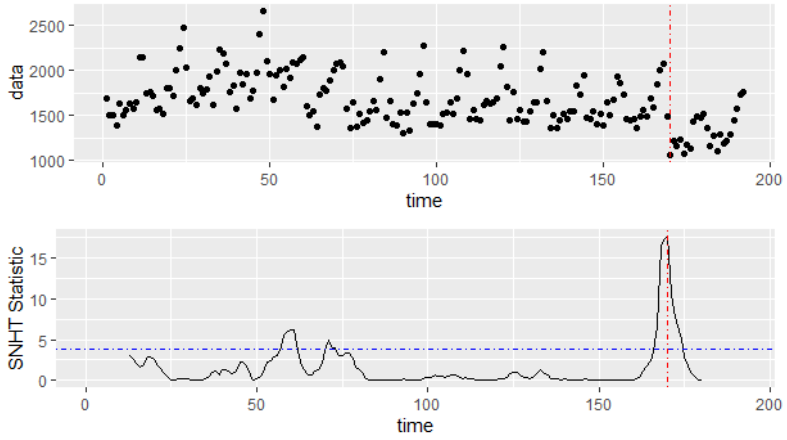

Figura 4.3: Gráfico da série temporal com a demarcação da região crítica do teste SNHT com período igual a doze

Tabela 4.1: Tabela de resultados do SNHT com período igual a doze

Posição de mudança	$\boldsymbol{T^s_{170}}$	**Média à esquerda**	**Média à direita**
170	17.57	1624	1286

De acordo com o que é mostrado na figura 4.1, o ponto de mudança identificado ocorreu no ano de 1983, como se era esperado. Assim, este resultado faz muito sentido, segundo o conhecimento prévio a respeito da série em estudo.

Segundo a tabela 4.1, é possível notar que a mudança foi detectada no ponto 170 da série com uma estatística máxima de 17.57. A média das observações á direita do ponto detectado foi consideravelmente menor do que a média das observações à esquerda. Portanto, assim como o estudo de Harvey e Durbin (1986) concluiu, é possível dizer que há indícios de que a lei do uso de cinto de segurança realmente causou um impacto positivo ao reduzir o número médio de mortes no trânsito.

Conforme os resultados do teste apresentados nesta sessão, ficou claro que o tamanho do período utilizado tem grande impacto nos resultados no teste. Dessa forma, este parâmetro deve ser estudado mais a fundo e ser escolhido cuidadosamente.

Capítulo 5

Aplicação do SNHT em dados simulados

Para possibilitar uma melhor compreensão a respeito do nível de assertividade do teste, séries temporais de médias móveis foram simuladas de tamanho $n = 200$ e com diferentes valores de $\phi \in (0,5; 0,7; 0,9)$, pois, dessa maneira, o desempenho do teste também poderá ser medido de acordo com as características específicas da série em que está sendo aplicado.

Ainda com o objetivo de se entender como é a performance do teste mais profundamente, em todas as séries simuladas, foi colocado um ponto de mudança na posição 120, entretanto, o tamanho desta quebra também sofreu variação de acordo com cada teste Quebra $\in (0,5; 1; 2; 3; 5)$, ou seja, dado que as séries em estudo possuem $\sigma = 1$, a performance do teste será medida frente a quebras de tamanho menor do que σ, de mesmo tamanho e também de algumas vezes maiores do que σ.

Outro fator que irá variar de valor ao longo dos testes será o período p utilizado para o cálculo das médias da série, uma vez que, como já foi comentado anteriormente, o valor de p pode influenciar muito nos resultado dos SNHT. Foi definido que $p \in (7; 14; 28; 35)$, pois são valores que sempre levam em consideração uma periodicidade igual a 7, o que será conveniente para o estudos dos dados referentes a Covid-19 que serão realizados no capitulo a seguir.

Considerando todas as combinações possíveis entre fatores que serão variados, conforme explicado acima, para cada possibilidade foram simuladas 1000 séries e o SNHT foi aplicado. A tabela 5.1 traz os resultados obtidos para a taxa de *Erro tipo II* que ocorreu em cada uma das combinações, levando em conta as seguintes hipóteses adaptadas do

teste:

$$H_0 : Z_i \sim N(0,1) \quad i \in [119;120] \qquad H_1 : \begin{cases} Z_i \sim N(\mu_1, 1), & i \in [119;120] \\ Z_i \sim N(\mu_2, 1) & i \in [119;120], \end{cases}$$

assim, dado que a quebra foi inserida na posição 120, só será contabilizado como um acerto do teste, quando este apontar que houve um ponto de mudança entre as posições 119 e 120.

Tabela 5.1: Taxas de *Erro tipo II* obtidas pelo SNHT aplicado em séries simuladas considerando um intervalo de acerto de duas observações

		ϕ		
Quebra	p	0,5	0,7	0,9
0,5	7	0,969	0,988	0,986
0,5	14	0,957	0,962	0,974
0,5	28	0,954	0,953	0,958
0,5	35	0,938	0,950	0,945
1,0	7	0,924	0,938	0,958
1,0	14	0,871	0,917	0,905
1,0	28	0,828	0,864	0,864
1,0	35	0,813	0,827	0,855
2,0	7	0,706	0,802	0,834
2,0	14	0,501	0,595	0,687
2,0	28	0,429	0,547	0,595
2,0	35	0,453	0,524	0,562
3,0	7	0,369	0,501	0,571
3,0	14	0,193	0,269	0,342
3,0	28	0,177	0,228	0,338
3,0	35	0,168	0,264	0,308
5,0	7	0,043	0,091	0,148
5,0	14	0,011	0,036	0,061
5,0	28	0,009	0,032	0,050
5,0	35	0,015	0,039	0,048

A partir da tabela 5.1, é possível perceber que dentre os valores de ϕ utilizados, os melhores resultados foram obtidos com ϕ = 0,5, dado que foi a coluna que obteve as menores taxas de *Erro tipo II*, ou seja, há indícios de que quanto menor é o valor de ϕ, melhor o SNHT irá performar.

Na tabela 5.1 também é possível notar que, quando a $Quebra = 0,5$, o teste quase nunca consegue acertar, o que faz sentido, visto que a quebra colocada é de valor inferior ao σ da série. Quando a $Quebra = 1$, o teste consegue acertar um pouco mais, mas ainda assim não obtém muito sucesso, continua errando em mais de 80% das vezes. Com $Quebra = 2$ o teste já é capaz de acertar em mais da metade dos testes para alguns valores de p, entretanto, não é um resultado muito positivo, dado que a quebra colocada possui o dobro do valor de σ e esperava-se que o teste conseguiria detecta-la mais vezes. Já com $Quebra = 3$, o SNHT pôde atingir menos de 20% de erro e com $Quebra = 5$, ele conseguiu acertar em quase todas as vezes.

Pelos resultados apresentados também ficou claro que a assertividade do teste é melhorada com os valores de p igual a 28 ou 35, fato que já era esperado, pelo fato de que, com maiores valores de p, mais informação é considerada no cálculo das médias, fazendo com que as comparações sejam melhores.

É valido ressaltar que taxas de erro muito baixas foram obtidas apenas quando a $Quebra = 5$, sendo uma quebra realmente muito grande, cinco vezes o valor do σ da série. Na tentativa de melhorar os resultados obtidos, o experimento será realizado novamente, dessa vez, considerando um intervalo maior em volta do ponto 120 como acerto, pois é plausível que o teste não consiga identificar o ponto de mudança em sua posição exata, mas consiga identificar os seus arredores. Para isso as seguintes hipóteses adaptadas do teste serão consideradas:

$$H_0 : Z_i \sim N(0,1) \quad i \in [115; ...; 125], \qquad H_1 : \begin{cases} Z_i \sim N(\mu_1, 1) & i \in [115; ...; 125] \\ Z_i \sim N(\mu_2, 1) & i \in [115; ...; 125], \end{cases}$$

Assim, dado que a quebra foi inserida na posição 120, será contabilizado como um acerto do teste, quando este apontar que houve um ponto de mudança entre as posições 115 a 125, sendo assim, cinco posições anteriores à quebra e cinco posições posteriores. Os resultados obtidos neste novo experimento estão sendo apresentados na tabela 5.2.

Tabela 5.2: Taxas de *Erro tipo II* obtidas pelo SNHT aplicado em séries simuladas considerando um intervalo de acerto de dez observações

		ϕ		
Quebra	p	0,5	0,7	0,9
0,5	7	0,915	0,921	0,921
0,5	14	0,875	0,857	0,881
0,5	28	0,784	0,815	0,817
0,5	35	0,764	0,789	0,801
1,0	7	0,832	0,856	0,887
1,0	14	0,650	0,704	0,762
1,0	28	0,491	0,547	0,635
1,0	35	0,485	0,532	0,555
2,0	7	0,567	0,658	0,690
2,0	14	0,193	0,316	0,410
2,0	28	0,103	0,170	0,214
2,0	35	0,106	0,157	0,192
3,0	7	0,247	0,374	0,463
3,0	14	0,032	0,070	0,108
3,0	28	0,018	0,030	0,065
3,0	35	0,019	0,023	0,045
5,0	7	0,018	0,052	0,123
5,0	14	0,000	0,000	0,006
5,0	28	0,000	0,000	0,001
5,0	35	0,000	0,000	0,002

Com base na tabela 5.2, nota-se que, em geral, houve uma melhora nos resultados obtidos com relação ao primeiro experimento, uma vez que, como um todo, os valores das taxas de erro encontradas foram consideravelmente menores. Entretanto, é possível se observar alguns padrões que foram mantidos.

Novamente, fica claro que o teste tem sua assertividade melhorada para séries com $\phi = 0,5$ e com p igual a 28 ou 35, também repete-se o fato de que, quanto maior a quebra colocada, menor a taxa de erro obtida. Porém, desta vez, observa-se que a partir de $Quebra = 2$, o SNHT já é capaz de performar muito melhor, apresentando uma taxa de

erro de apenas 10%, o que já é algo relativamente bom.

Desse modo, a partir dos testes realizados com simulação de dados, é possível dizer que o teste consegue detectar bem os arredores de pontos de quebra que são maiores do que o σ da série. Entretanto, ele não é tão preciso em apontar a sua localização exata.

5.1 Exemplos específicos do experimento simulado

A seguir serão detalhados alguns exemplos de séries simuladas através modelo de médias móveis, MA(1), que fizeram parte do conjunto de séries utilizado no experimento anteriormente analizado neste capítulo. Primeiramente, as características de um tipo de série em o teste performou muito mal, em seguida um tipo de série em o teste conseguiu acertar em cerca de metade das tentativas e, por último, um tipo de série em o teste alcançou 100% de aproveitamento.

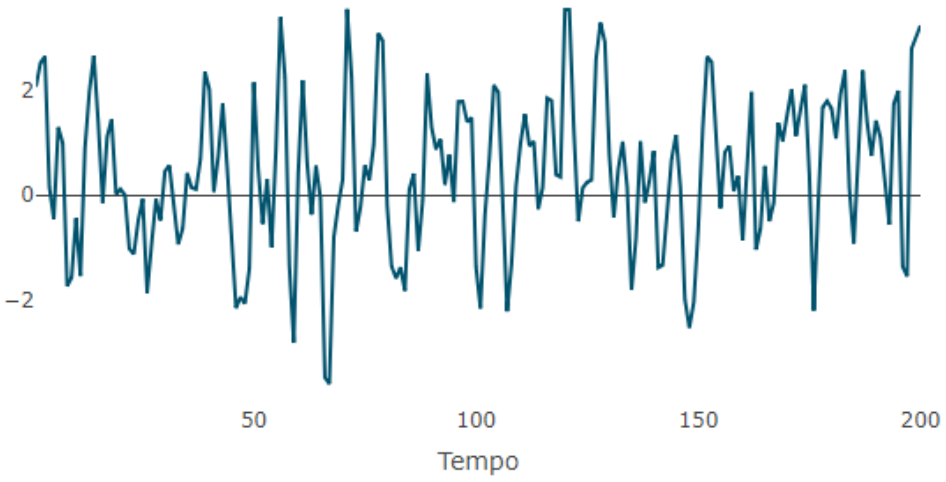

Figura 5.1: Série de médias móveis simulada com $\phi = 0{,}9$ e Quebra 0,5

Na figura 5.1, um exemplo dentre as séries simuladas com $\phi = 0{,}9$ e quebra 0,5 na posição 120 é representada graficamente e, a partir dela, percebe-se que, de fato, a Quebra inserida foi relativamente pequena e nem é possível notar com clareza a mudança do ponto 120 em diante no gráfico.

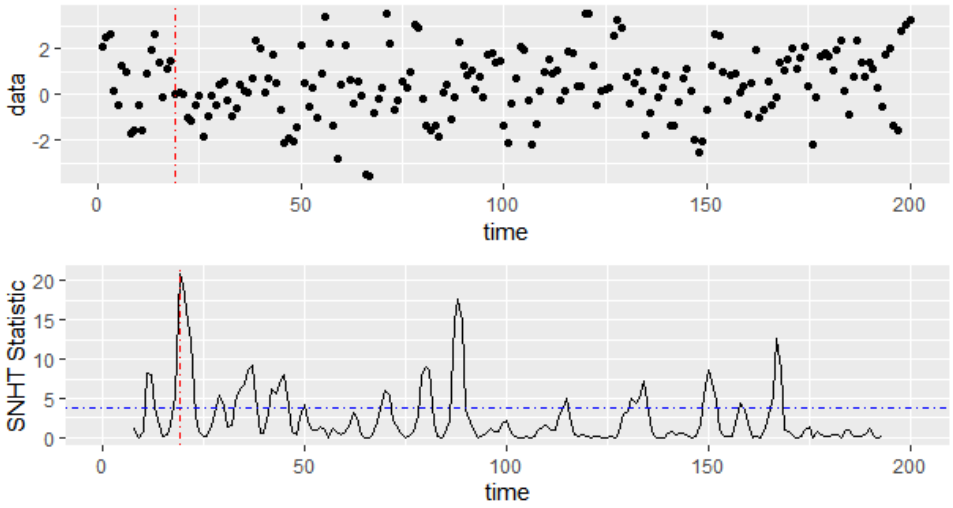

Figura 5.2: Estatística do teste com período 7 na série de $\phi = 0{,}9$ e Quebra 0,5

A figura 5.2 mostra a estatística do teste obtida ao se aplicar o SNHT com $p = 7$ neste exemplo de série simulada com $\phi = 0{,}9$ e Quebra 0,5 na posição 120. Nota-se que a estatística do teste ultrapassa a linha tracejada, que indica o limite da região crítica do teste, em vários momentos, inclusive, na posição 120 a estatística apresentou um pico superior à linha tracejada. Porém, o maior pico, ponto onde a quebra é identificada pelo teste, foi encontrado mais no início da série, próximo à posição 20.

Vale ressaltar que o σ dos dados é superior à quebra colocada. Portanto, não houve uma mudança tão expressiva na série, dado que a sua variabilidade ao longo de todo o período era bem superior ao valor da quebra, assim, era esperado que o teste poderia identificar a quebra em outros momentos, pois na série há realmente outros pontos em que a mudança ocorre de forma mais significativa.

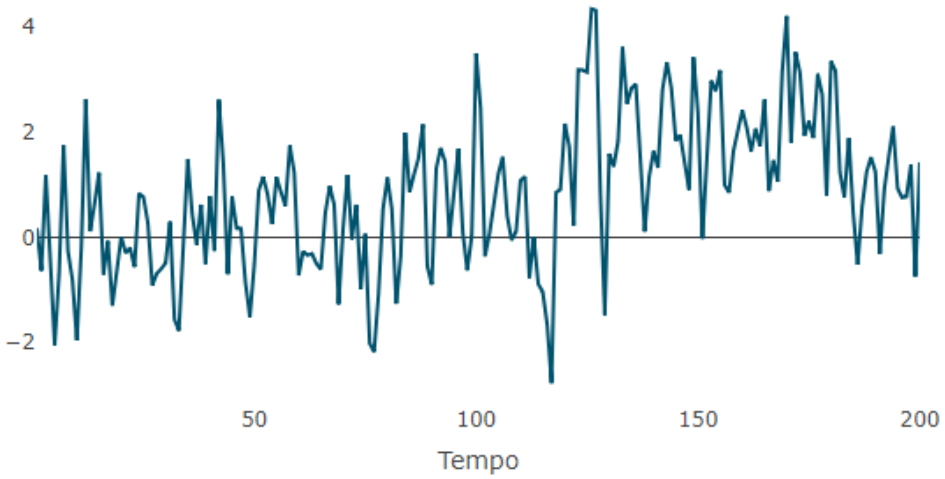

Figura 5.3: Série de médias móveis simulada com $\phi = 0{,}5$ e quebra 2

Na figura 5.3, um exemplo dentre as séries simuladas com $\phi = 0{,}5$ e Quebra 2 na posição 120 é representada graficamente e, com base nela, é visível que a partir da posição 120 houve uma mudança que fez com que os valores da série, daquele momento em diante, fossem consideravelmente maiores do que os anteriores.

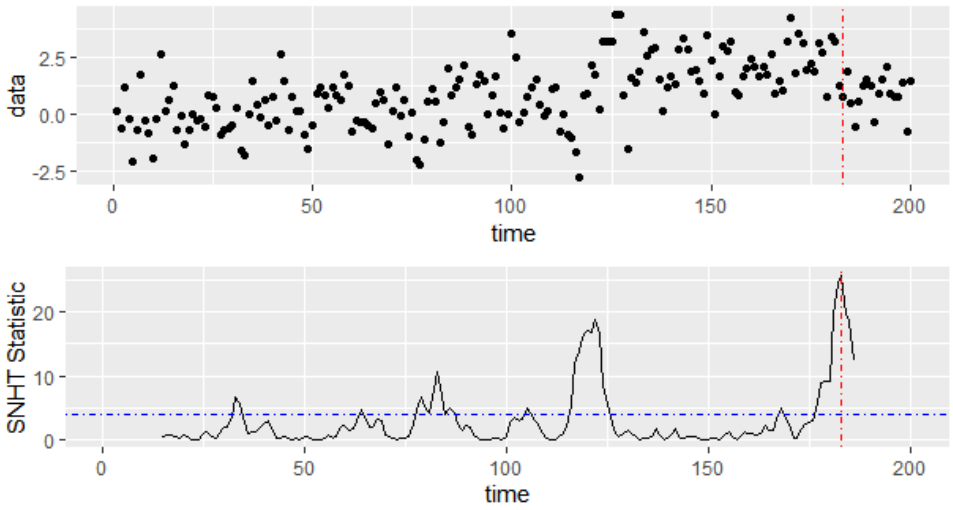

Figura 5.4: Estatística do teste com período 14 na série de $\phi = 0{,}5$ e quebra 2

A figura 5.4 mostra a estatística do teste obtida ao se aplicar o SNHT com $p = 14$ neste exemplo de série simulada com $\phi = 0{,}5$ e quebra 2 na posição 120. Nota-se que, a estatística do teste ultrapassa a linha tracejada em alguns momentos, inclusive, na posição 120 a estatística apresentou um pico muito superior à linha tracejada. Porém, o maior pico, ponto onde a quebra é identificada pelo teste, foi encontrado mais no fim da série, próximo à posição 180.

O grupo de séries simuladas e testadas com estas mesmas características apresentaram uma taxa de erro bem próxima à 50%, o que sugere que o teste não é capaz de identificar com precisão estes pontos de quebra que são facilmente notados a olho nu.

Entretanto, é necessário comentar que na posição em que o SNHT identificou o ponto de mudança, também é possível ver com facilidade que a série parece, de fato, apresentar uma mudança de comportamento e passa a ter valores um pouco menores. Apesar dessa mudança não ter sido propositalmente inserida, o teste a detectou de forma mais incisiva do que a que era esperada.

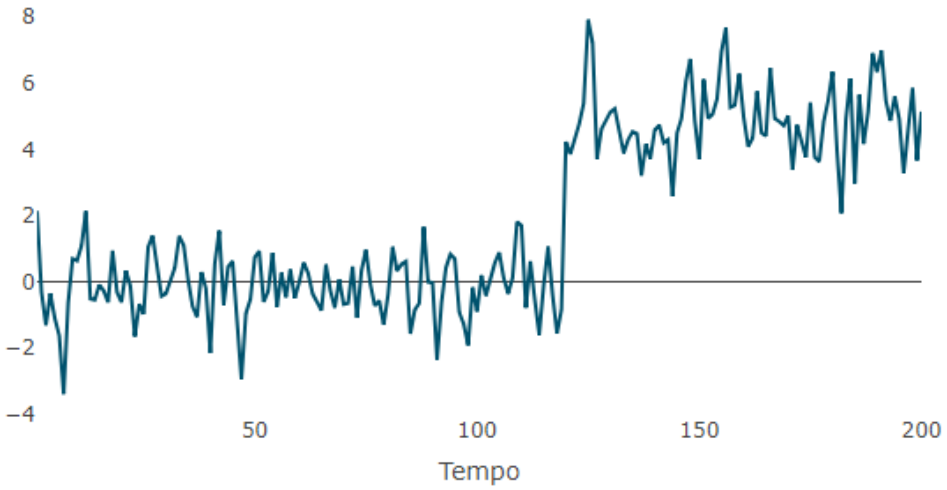

Figura 5.5: Série de médias móveis simulada com $\phi = 0{,}5$ e Quebra 5

Na figura 5.5, um exemplo dentre as séries simuladas com $\phi = 0{,}5$ e Quebra 5 na posição 120 é representada graficamente e, com base nela, é visível que a partir da posição 120 houve uma mudança bastante abrupta, que fez com que os valores da série, daquele momento em diante, fossem muito maiores do que os anteriores.

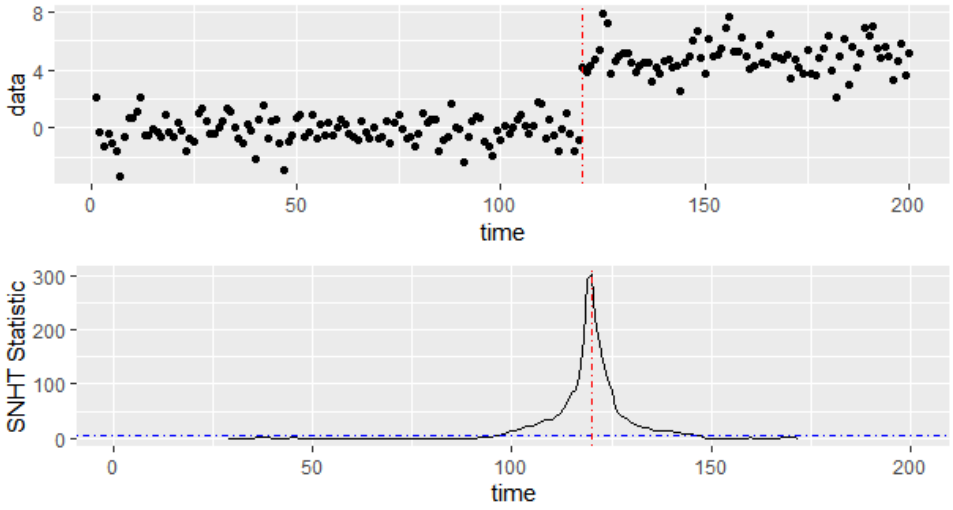

Figura 5.6: Estatística do teste com período 28 na série de $\phi = 0{,}5$ e Quebra 5

A figura 5.6 mostra a estatística do teste obtida ao se aplicar o SNHT com $p = 28$ neste exemplo de série simulada com $\phi = 0{,}5$ e Quebra 5 na posição 120. Nota-se que, a estatística do teste ultrapassa a linha tracejada somente no momento em que a quebra foi inserida, na posição 120, apresentando um pico muito grande. Dessa maneira, fica claro que o teste identificou com bastante facilidade o ponto exato em que sabe-se que a quebra realmente ocorreu.

Capítulo 6

Análise de dados referentes à Covid-19

6.1 Apresentação dos dados

Neste capítulo, serão analisados dados coletados do SUS (2021) [1], referentes a média móvel de sete dias do número de novos casos de Covid-19 registrados na cidade de São Paulo, ou seja, cada valor apresentado consiste na média de casos registrados de seis dias atrás, até aquele dia. Os dados se estendem do período de março de 2020, início da pandemia no Brasil, a abril de 2021, momento em que sabe-se que o país estava saindo de uma segunda grande onda de casos da doença.

[1] https://covid.saude.gov.br/

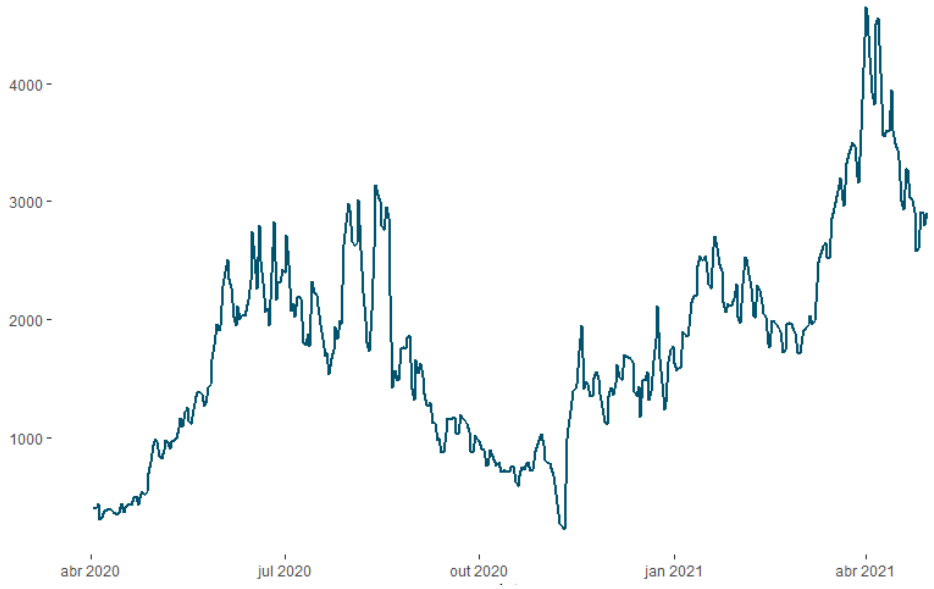

Figura 6.1: Número de novos casos de Covid-19 na cidade de São Paulo

Na figura C.15 os dados são apresentados graficamente durante o seu período em estudo. Logo a primeira vista, a ocorrência de duas grandes ondas de novos casos já fica muito nítida, sendo que há dois grandes picos de casos que estão bastante destacados no gráfico, o primeiro em meados de agosto de 2020 e o segundo, mais recentemente, em abril de 2021, mês em que a pandemia atingiu os seus mais altos níveis em relação a número de casos e número de mortes no Brasil.

Esse padrão de comportamento, em formato de ondas, foi sentido a nível nacional no Brasil e, pelo gráfico, é indiscutível o fato de que a cidade de São Paulo também acompanhou o ritmo brasileiro, sofrendo bastante com a doença em todo o período em análise.

Como já foi muito difundido pela mídia, os casos começaram a se tornar mais frequentes em junho de 2020 até chegarem ao seu primeiro pico de em torno de, 3 mil casos diários em São Paulo, em agosto de 2020 e depois entraram em um processo de queda, fazendo com que os números registrados em novembro de 2020 fossem relativamente baixos. Porém, os casos logo tornaram a crescer novamente, até o início de abril de 2021, período mais crítico da pandemia no Brasil e com mais de 4 mil casos da doença sendo registrados por dia em São Paulo. Felizmente, no final do mês em questão, os números voltaram a cair.

Dado esta alta quantidade de mudanças de comportamento apresentadas pelos dados, soa interessante analisar os momentos exatos em que as quebras no padrão da série

ocorreram. Para isso, o teste SNHT será aplicado nos dados, na tentativa de se identificar corretamente quando essas mudanças aconteceram e descobrir se estas podem estar relacionadas a fatos que foram noticiados na época.

6.2 Aplicação do SNHT

Uma vez que, o objetivo deste estudo é de detectar, se existirem, vários pontos de mudança na série e que não se é indicado apontar mais de um ponto de mudança a cada vez que o teste é aplicado, foi decidido então, que o teste seria aplicado varias vezes. Assim, a cada vez que o teste é aplicado, o ponto detectado é sinalizado e, a partir deste momento em diante, os dados serão novamente submetidos ao teste para a possível identificação de uma nova quebra. Dessa maneira, o processo será repetido sucessivamente até que o teste não consiga mais identificar nenhum ponto.

Conforme descrito acima, a seguir os dados serão testados pelo SNHT e isto ocorrerá de duas formas, primeiramente será utilizado um período de $p = 14$ na testagem e, posteriormente, um período de $p = 28$ será levado em conta. Estes valores foram escolhidos devido ao fato de terem apresentado uma melhor performance durante o experimente simulado e, também, por serem múltiplos de sete, quantidade de dias usados no cálculo das médias móveis, o que certamente irá gerar um padrão sazonal e, portanto, deve ser considerado.

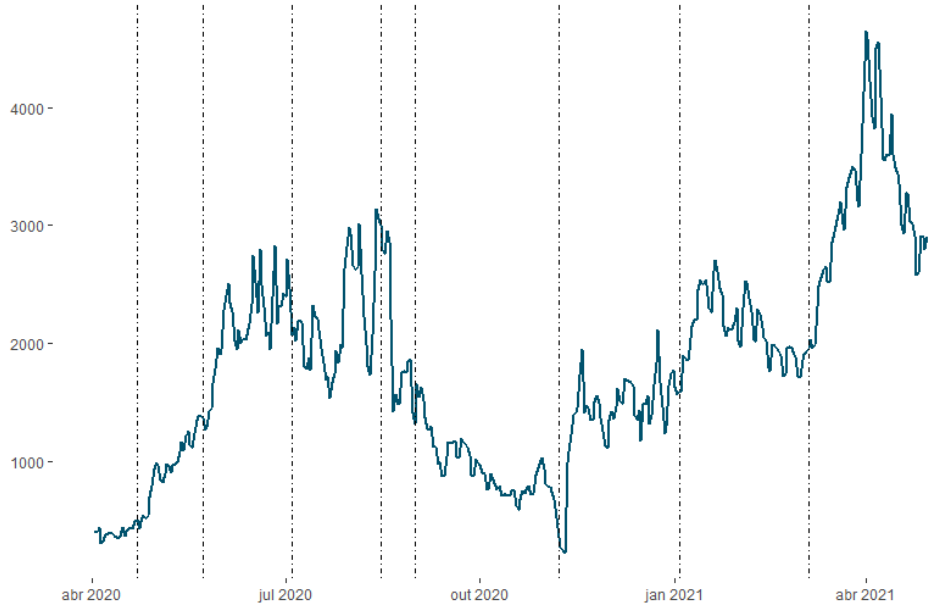

Figura 6.2: Número de novos casos de Covid-19 na cidade de São Paulo sinalizando pontos de quebra segundo o SNHT com $p = 14$

Com base na figura 6.2, é possível observar que o SNHT aplicado com $p = 14$ foi capaz de identificar oito pontos de mudança na série em estudo. É interessante notar que, o primeiro ponto identificado está localizado bem no início da série e se um período maior tivesse sido utilizado, este ponto não poderia ter sido identificado, já que as p primeiras observações são descontadas para o cálculo da estatística do teste.

Todos os momentos em que o teste identificou a ocorrência de pontos de quebra realmente parecem se tratar de dias em que a média móvel dos casos estava sofrendo uma mudança de comportamento, pois eles estão sempre colocados em instantes de aumento ou queda súbita dos valores registrados.

Entretanto, também é importante comentar que no gráfico existem diversos outros momentos em que os dados parecem mudar de comportamento, mas que não foram apontados pelo teste. Isto pode ter acontecido pelo fato de que 14 pontos ao redor da série precisaram ser desconsiderados a cada vez que o teste era aplicado, o que pode ter atrapalhado os cálculos da estatística do teste ou pode ser simplesmente porque realmente não são bons candidatos a ponto de mudança da série.

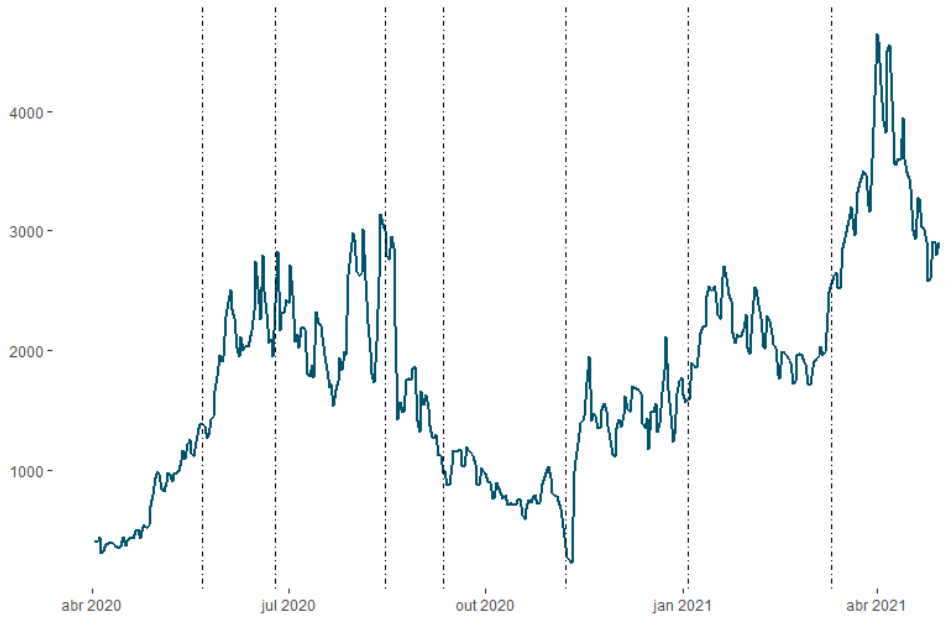

Figura 6.3: Número de novos casos de Covid-19 na cidade de São Paulo sinalizando pontos de quebra segundo o SNHT com $p = 28$

Na figura 6.3 estão sinalizados os pontos de quebra identificados pelo SNHT com $p = 28$ e, assim como na sua aplicação com $p = 14$, nota-se que o teste apontou momentos que parecem ser muito pertinentes mas, ainda assim, existem muitos outros pontos que parecem ser potencialmente pontos de mudança e que não foram detectados pelo teste.

Desta vez o teste identificou sete pontos de mudança, um a menos do que na aplicação anterior e vale ressaltar que nenhum ponto muito próximo aos extremos da série foi demarcado, o que faz sentido, uma vez que ao se considerar um período maior no teste, mais pontos são desconsiderados no seu início e fim. Infelizmente, esta limitação do teste o impossibilita de detectar grandes mudanças que ocorrem próximas aos limites da série, como aquele grande aumento nos valores que aconteceu no início da segunda onda da pandemia.

Também é notável que alguns momentos foram destacados em ambas as aplicações mas, por outro lado, alguns dos pontos identificados em cada uma das testagens foram bem distintos. Esta comparação ficará mais clara com o gráfico a apresentado a seguir.

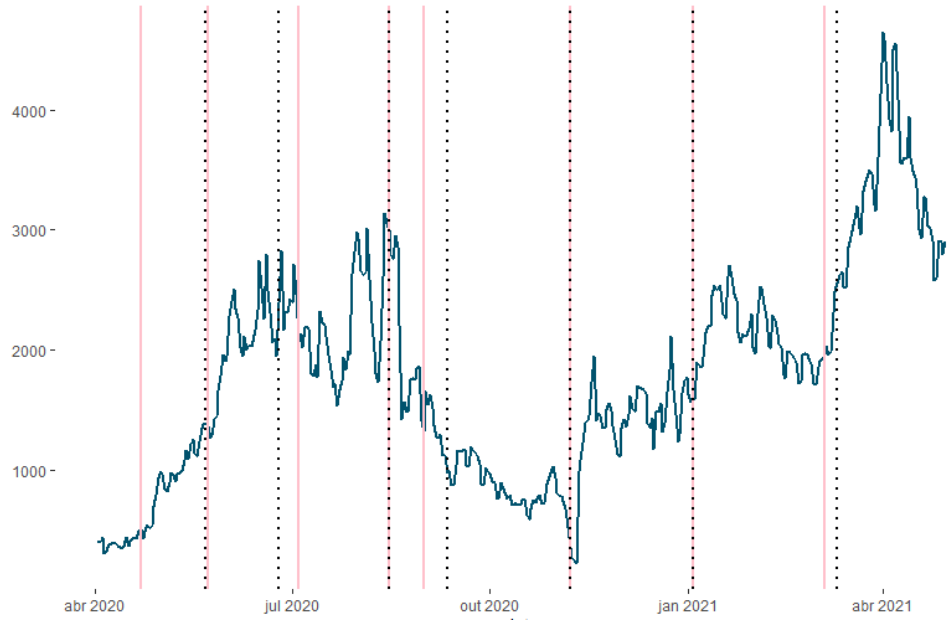

Figura 6.4: Número de novos casos de Covid-19 na cidade de São Paulo sinalizando pontos de quebra segundo o SNHT com $p = 14$ e $p = 28$

A partir da figura 6.4, pode-se observar os pontos identificados pelo SNHT com $p = 14$, sinalizados pela cor rosa, contrapostos aos pontos detectados pelo teste com $p = 28$, demarcados pela cor preta no gráfico.

Como já foi comentado, o primeiro ponto detectado pelo teste com $p = 14$ está localizado muito no início da série e, assim, o teste com $p = 28$ não é capaz de identificá-lo, entretanto, a maioria dos demais pontos encontrados ocorreu em posições muito semelhantes em ambas as aplicações.

Através das duas formas que foram utilizadas para a testagem da série, foram encontrados os mesmos pontos no início de setembro e durante novembro de 2020 e também em janeiro de 2021. Alguns pontos foram identificados em posições muito próximas em ambas as testagens, como ocorreu em maio de 2020 e março de 2021. Já durante julho de 2020 e no final de setembro de 2020, as quebras foram detectadas pelos testes aconteceram em momentos mais distintos.

O primeiro ponto sinalizado em rosa está localizado no dia 22 de abril de 2020 e, de fato, parece ter sido um momento de mudança, dado que a partir dali os valores foram se tornando cada vez maiores. Em torno de 15 dias antes desta data, período que acredita-se ser o necessário para que os efeitos de um aumento na taxa de contagio da doença demorem a ser perceptíveis pelos dados, uma notícia do Portal Governo (2020) foi publicada com a atualização de que o Governador do estado de SP havia decidido prorrogar a quarentena,

devido a previsões de que os casos de Covid-19 poderiam aumentar consideravelmente nos próximos dias. Pelos resultados aqui apresentados, fica claro que esta previsão estava correta e mesmo com a imposição da quarentena, a doença continuou a se proliferar cada vez mais.

O segundo e terceiro pontos marcados no gráfico, em preto e em rosa, respectivamente, ocorreram nos dias 22 e 23 de maio de 2020 e realmente parece ter acontecido uma mudança, visto que, a partir deste momento os valores passaram a crescer mais rapidamente. Em torno de 15 dias antes desta data, uma reportagem de Toledo (2020) foi veiculada pela revista Veja, alertando sobre a necessidade de se decretar o lockdown na cidade de São Paulo, pois um modelo desenvolvido na Universidade Estadual de Campinas (Unicamp) projetava que com os níveis de isolamento registrados na época, a doença tendia a aumentar muito nas próximas semanas. Pelos resultados do SNHT, fica possível perceber que, infelizmente, o lockdown não foi implementado e que as projeções feitas pelo modelo da Unicamp se confirmaram.

No dia 15 de maio de 2020 o gráfico está demarcado em preto e também em rosa, ou seja, mesmo com aplicações diferentes do teste, ele continuou detectando o mesmo lugar na série. Cerca de 15 dias antes desta data, Forato (2020) comentou que a disseminação do corona vírus na cidade de São Paulo estava diminuindo de velocidade e que o casos deveriam apresentar queda, porém, deixou o alerta de que mesmo em ritmo mais lento, a doença ainda causaria muitas vítimas. Com base nos dados analisados, pode-se perceber que o comportamento nas próximos semana apresentou exatamente a mesma tendência comentada.

Em 7 de novembro de 2020, também foi identificado pelo teste duas vezes, mesmo com parâmetros distintos. Neste momento fica nítido que a série apresentou uma queda muito significativa, chegando a registrar seus menores valores desde o início da pandemia, em março daquele ano.

Novamente, o teste aplicado das duas formas detectou o mesmo dia de mudança, desta vez em 3 de janeiro de 2021, momento em que os casos voltaram a atingir patamares muitos elevados. Segundo Gomes (2020), este padrão já era esperado, devido a proximidade aos feriados de fim de ano e a queda nos níveis de isolamento social.

Em cada uma das testagens, outros pontos também foram detectados em locais que a série aparenta ter sofrido mudança de comportamento, entretanto, ficaram localizados em posições mais distantes entre si.

Capítulo 7

Conclusão

Como foi visto, o SNHT para amostras de mesmo tamanho pode apresentar resultados diferentes de acordo com o período utilizado em sua aplicação, fazendo com que o parâmetro p assuma um papel de extrema importância para o teste. Tal importância é ainda mais acentuada na presença de sazonalidade na série em estudo, sendo importante considerar o intervalo do padrão de repetição sazonal dos dados durante a escolha do período do teste.

O experimento com séries de médias móveis MA(1) simuladas com pontos de quebra mostrou que ao se considerar como acerto apenas a detecção do momento exato em que a mudança ocorre, o desempenho do teste não foi tão satisfatório, dado que, a partir e uma quebra de 2σ colocada na série, ele foi capaz que apontá-la em seu local correto em por volta de metade das vezes. E com a quebra de 3σ as taxas do *Erro tipo II* ainda ficaram relativamente altas, registrando mais de 15%. Uma assertividade superior à 95% só foi alcançada com quebras de 5σ, que consistem em mudanças de comportamento realmente muito drásticas.

Porém, ao se considerar um intervalo maior ao redor do ponto de mudança como acerto, cinco observações anteriores e cinco posteriores, o teste passa a ter uma performance consideravelmente melhor. Assim, o SNHT já foi capaz de atingir mais de 95% de assertividade com quebras de 3σ e, desta vez, com a quebra de 5σ ele conseguiu acertar em todas as tentativas realizadas.

Portanto, é possível concluir que, para dados com as mesmas características dos que foram simulados neste trabalho, o teste só consegue atingir altos níveis de assertividade com a precisão exata do instante de mudança, quando a quebra colocada é maior ou igual a 5σ. Entretanto, se pontos ao redor da quebra também forem julgados aceitáveis, o teste

já tem a possibilidade de fazer a sua detecção corretamente em quase todas as tentativas a partir de uma quebra de 3σ.

Infelizmente, essa necessidade de que a quebra seja, no mínimo, de tamanho superior à 3σ se torna um empecilho para a utilização do SNHT em diversos casos. Uma vez que, em problemas reais, as quebras que se deseja detectar são, geralmente, de tamanho bastante inferior à 3σ e muito provavelmente não poderão ser encontradas através do SNHT. Ao mesmo tempo, o teste frequentemente apontou pontos de mudança em locais em que eles não estavam presentes.

Ao se aplicar o teste na série de médias móveis do número de novos casos de Covid-19 na cidade de São Paulo múltiplas vezes para a detecção de diversos pontos de mudança durante o primeiro ano da pandemia e também utilizando dois valores de p distintos, ficou claro que o SNHT apontou momentos em que, pela sua representação gráfica, já era possível suspeitar que a mudança havia ocorrido.

Também foi possível perceber que mesmo se utilizando períodos diferentes na testagem, $p = 14$ e $p = 28$, ambas as aplicações detectaram momentos muito semelhantes e que condiziam com notícias publicadas na época. Estas já faziam menções a respeito do que se era esperado para o comportamento da pandemia nos próximos dias e tais previsões puderam ser provadas com os resultados obtidos neste trabalho.

De forma geral, neste trabalho, o teste SNHT pôde ser exemplificado por meio de dados referentes à acidentes de trânsito no Reino Unido e, a partir disso, tornou-se claro o grande impacto que a escolha correta do parâmetro p pode causar na assertividade do teste. Através do experimento simulado, o seu funcionamento também foi entendido com mais detalhes e foi possível se obter mais conhecimento a respeito do poder do teste. Por fim, a utilização do SNHT em dados da Covid-19 trouxe mais informações interessantes em relação a evolução da pandemia na cidade de São Paulo.

Apêndice A

Símbolos matemáticos

- Y é a série temporal em estudo;
- Y_i é a i-ésima observação da série temporal Y;
- μ é a média da série Y;
- a_i é ruído branco;
- ϕ é o parâmetro do modelo.
- $\overline{Y}$ é o valor médio da série Y;
- X_j é uma série temporal pertencente ao conjunto de séries de referência;
- $X_j i$ é um ponto específico em uma série de referência;
- t é o ponto de mudança presente na série;
- k é o número de séries de referência;
- $\overline{X_j}$ é o valor médio da série X_j;
- ρ_j é o coeficiente de correlação entre Y e X_j (deve sempre ser positivo);
- n é o número de observações na amostra;
- N detona a distribuição normal;
- Q_i é a i-ésima observação da série de diferenças Q;
- $\overline{Q}$ é a média da série de diferenças Q;

- Z é a série de diferenças normalizada;
- σ_Q é o desvio padrão da série de diferenças;
- $T^s_{\max}$ é o ponto máximo da estatística do teste;
- $\overline{z}_1$ é a média da série padronizada antes do ponto de mudança;
- $\overline{z}_2$ é a média da série padronizada após o ponto de mudança;
- a é o valor máximo (última posição antes do ponto de quebra);

Apêndice B

Códigos utilizados

```
library(snht)
library(cowplot)
library(forecast)
library(TSstudio)
library(dplyr)
library(zoo)
library(xts)
library(plotly)
library(ggplot2)

# load data
UKDriverDeaths

# Plot
ts_plot(UKDriverDeaths,slider = F, line.mode = "lines",width = 2,
title = 'Número de motoristas mortos ou seriamente feridos na
Grã-Bretanha ao longo dos anos',Ytitle = 'Número de mortes',
Xtitle = 'Ano')
# Plot diff
ts_plot(diff(UKDriverDeaths), slider = FALSE, line.mode = "lines",
width = 2,title = 'Número de motoristas mortos ou seriamente feridos
na Grã-Bretanha ao longo dos anos',Ytitle = 'Número diferenciado
de mortes',Xtitle = 'Ano')
```

```
# Plot months
ts_seasonal(UKDriverDeaths, palette_normal='Spectral',type = "normal",
title = 'Número de motoristas mortos ou seriamente feridos na
Grã-Bretanha ao longo dos meses')
# Plot years
ts_seasonal(UKDriverDeaths, palette='Spectral',type = "cycle",
title = 'Número de motoristas mortos ou seriamente feridos na
Grã-Bretanha ao longo dos anos')
# Plot box months
ts_seasonal(UKDriverDeaths, palette='Spectral',type = "box",
title = 'Número de motoristas mortos ou seriamente feridos na
Grã-Bretanha ao longo dos meses')

# Test using period = 12
snhtStatistic12 = snht(data = UKDriverDeaths, period = 12)
# Plot test using period = 12
plotSNHT(data = UKDriverDeaths, stat = snhtStatistic12, alpha = .05)
# Metrics of the test using period = 12
largestStatTime = which.max(snhtStatistic12$score)
snhtStatistic12[largestStatTime, ]


# series simulation and snht testing function
snht_simulation <- function(n ,sample,ma,period,breakPoint,gap){
    result = matrix(ncol=3, nrow=sample)
    # Simulating series
    for(i in 1:sample){
      ts <- arima.sim(list(order = c(0,0,1), ma = ma), n.start=20, n = n)
        # Adding a break point
        for(j in breakPoint:n){
          ts[j] = ts[j] + gap
        }
      # Testing the series
```

```
      snhtStatistic = snht(data = ts, period = period)
      result[i,1] = which.max(snhtStatistic$score)
      result[i,2] = max(snhtStatistic$score,na.rm = TRUE)
      result[i,3] = ifelse(result[i,1] %in% c(breakPoint-1,breakPoint),0,1)
    }
  # Computing test results
  error_rate = sum(result[,3])/sample
  print(result)
  print(error_rate)
}

snht_simulation(n = 200,sample = 1000,ma = 0.9,
                period=35,breakPoint=120,gap=5)

# series simulation and snht testing with 10 points range function
snht_simulation <- function(n ,sample,ar,period,breakPoint,gap){
  result = matrix(ncol=3, nrow=sample)
  # Simulating series
  for(i in 1:sample){
    ts <- arima.sim(list(order = c(0,0,1), ma = ar), n.start=20, n = n)
    # Adding a break point
    for(j in breakPoint:n){
      ts[j] = ts[j] + gap
    }
    # Testing the series
    snhtStatistic = snht(data = ts, period = period)
    result[i,1] = which.max(snhtStatistic$score)
    result[i,2] = max(snhtStatistic$score,na.rm = TRUE)
    result[i,3] = ifelse(result[i,1] %in% c(breakPoint-5,breakPoint-4,
                          breakPoint-3,breakPoint-2,breakPoint-1,breakPoint,
                          breakPoint+1,breakPoint+2,breakPoint+3,breakPoint+4,
                          breakPoint+5),0,1)
  }
```

```
  # Computing test results
  error_rate = sum(result[,3])/sample
  print(result)
  print(error_rate)
}

snht_simulation(n = 200,sample = 1000,ar = 0.9,
                period=35,breakPoint=120,gap=5)

# Read covid data
df = read.csv("D:/Downloads/HIST_PAINEL_COVIDBR_03mai2021.csv",
                header=TRUE, sep=";")
# Filter SP data
sp = df %>% filter(municipio=='SÃ£o Paulo') %>% select(data,casosNovos)

# moving averages function
ma_function <- function(arr, n){
  res = arr
  for(i in n:length(arr)){
    res[i] = mean(arr[(i-n):i])
  }
  res
}

# Adjust data
sp$ma = ma_function(sp$casosNovos, n=7)
sp$data = sp$data %>% as.Date()
ts_ma= ts(sp$ma[7:400], start=c(2020,04,02), frequency = 365)

# Plot covid data
ts_ggplot(sp[7:400,] %>% select(ma,data)) +
  theme(panel.grid = element_blank(),panel.background = element_blank()) +
  geom_line(size=1,color='#00526d')
```

```
############## Test using period = 14 #################
snhtStatistic = snht(data = ts_ma, period = 14)
# Plot test using period = 14
plotSNHT(data = ts_ma, stat = snhtStatistic, alpha = .05)
# Metrics of the test using period = 14
largestStatTime = which.max(snhtStatistic$score)
snhtStatistic[largestStatTime, ]
summary(snhtStatistic)

snhtStatistic = snht(data = ts_ma[27:394], period = 14)
# Plot test using period = 14
plotSNHT(data = ts_ma[27:394], stat = snhtStatistic, alpha = .05)
# Metrics of the test using period = 14
largestStatTime = which.max(snhtStatistic$score)
snhtStatistic[largestStatTime, ]
summary(snhtStatistic)

snhtStatistic = snht(data = ts_ma[58:394], period = 14)
# Plot test using period = 14
plotSNHT(data = ts_ma[58:394], stat = snhtStatistic, alpha = .05)
# Metrics of the test using period = 14
largestStatTime = which.max(snhtStatistic$score)
snhtStatistic[largestStatTime, ]
summary(snhtStatistic)

snhtStatistic = snht(data = ts_ma[58:344], period = 14)
# Plot test using period = 14
plotSNHT(data = ts_ma[58:344], stat = snhtStatistic, alpha = .05)
# Metrics of the test using period = 14
largestStatTime = which.max(snhtStatistic$score)
snhtStatistic[largestStatTime, ]
```

```
summary(snhtStatistic)

snhtStatistic = snht(data = ts_ma[58:283], period = 14)
# Plot test using period = 14
plotSNHT(data = ts_ma[58:283], stat = snhtStatistic, alpha = .05)
# Metrics of the test using period = 14
largestStatTime = which.max(snhtStatistic$score)
snhtStatistic[largestStatTime, ]
summary(snhtStatistic)

snhtStatistic = snht(data = ts_ma[58:226], period = 14)
# Plot test using period = 14
plotSNHT(data = ts_ma[58:226], stat = snhtStatistic, alpha = .05)
# Metrics of the test using period = 14
largestStatTime = which.max(snhtStatistic$score)
snhtStatistic[largestStatTime, ]
summary(snhtStatistic)

snhtStatistic = snht(data = ts_ma[58:158], period = 14)
# Plot test using period = 14
plotSNHT(data = ts_ma[58:158], stat = snhtStatistic, alpha = .05)
# Metrics of the test using period = 14
largestStatTime = which.max(snhtStatistic$score)
snhtStatistic[largestStatTime, ]
summary(snhtStatistic)

snhtStatistic = snht(data = ts_ma[58:142], period = 14)
# Plot test using period = 14
plotSNHT(data = ts_ma[58:142], stat = snhtStatistic, alpha = .05)
# Metrics of the test using period = 14
largestStatTime = which.max(snhtStatistic$score)
snhtStatistic[largestStatTime, ]
summary(snhtStatistic)
```

```
############## Test using period = 28 #################
snhtStatistic = snht(data = ts_ma, period = 28)
# Plot test using period = 28
plotSNHT(data = ts_ma, stat = snhtStatistic, alpha = .05)
# Metrics of the test using period = 28
largestStatTime = which.max(snhtStatistic$score)
snhtStatistic[largestStatTime, ]
summary(snhtStatistic)

# Test using period = 28
snhtStatistic = snht(data = ts_ma[57:394], period = 28)
# Plot test using period = 28
plotSNHT(data = ts_ma[57:394], stat = snhtStatistic, alpha = .05)
# Metrics of the test using period = 28
largestStatTime = which.max(snhtStatistic$score)
snhtStatistic[largestStatTime, ]
summary(snhtStatistic)

# Test using period = 28
snhtStatistic = snht(data = ts_ma[57:169], period = 28)
# Plot test using period = 28
plotSNHT(data = ts_ma[57:169], stat = snhtStatistic, alpha = .05)
# Metrics of the test using period = 28
largestStatTime = which.max(snhtStatistic$score)
snhtStatistic[largestStatTime, ]
summary(snhtStatistic)

# Test using period = 28
snhtStatistic = snht(data = ts_ma[57:142], period = 28)
# Plot test using period = 28
plotSNHT(data = ts_ma[57:142], stat = snhtStatistic, alpha = .05)
```

```
# Metrics of the test using period = 28
largestStatTime = which.max(snhtStatistic$score)
snhtStatistic[largestStatTime, ]
summary(snhtStatistic)

# Test using period = 28
snhtStatistic = snht(data = ts_ma[169:394], period = 28)
# Plot test using period = 28
plotSNHT(data = ts_ma[169:394], stat = snhtStatistic, alpha = .05)
# Metrics of the test using period = 28
largestStatTime = which.max(snhtStatistic$score)
snhtStatistic[largestStatTime, ]
summary(snhtStatistic)

# Test using period = 28
snhtStatistic = snht(data = ts_ma[226:394], period = 28)
# Plot test using period = 28
plotSNHT(data = ts_ma[226:394], stat = snhtStatistic, alpha = .05)
# Metrics of the test using period = 28
largestStatTime = which.max(snhtStatistic$score)
snhtStatistic[largestStatTime, ]
summary(snhtStatistic)

# Test using period = 28
snhtStatistic = snht(data = ts_ma[226:350], period = 28)
# Plot test using period = 28
plotSNHT(data = ts_ma[226:350], stat = snhtStatistic, alpha = .05)
# Metrics of the test using period = 28
largestStatTime = which.max(snhtStatistic$score)
snhtStatistic[largestStatTime, ]
summary(snhtStatistic)

# Test using period = 28
```

```
snhtStatistic = snht(data = ts_ma[283:350], period = 28)
# Plot test using period = 28
plotSNHT(data = ts_ma[283:350], stat = snhtStatistic, alpha = .05)
# Metrics of the test using period = 28
largestStatTime = which.max(snhtStatistic$score)
snhtStatistic[largestStatTime, ]
summary(snhtStatistic)

# Plot with snht p = 14 breakpoints
ts_ggplot(sp[7:400,] %>% select(ma,data)) +
  theme(panel.grid = element_blank(),panel.background = element_blank()) +
  geom_line(size=1,color='#00526d') +
  geom_vline(xintercept = as.Date(c("2020-04-22","2020-05-23","2020-07-04",
                                    "2020-08-15","2020-08-31","2020-11-07",
                                    "2021-01-03","2021-03-05")),
                                     linetype=4, colour="black")

# Plot with snht p = 28 breakpoints
ts_ggplot(sp[7:400,] %>% select(ma,data)) +
  theme(panel.grid = element_blank(),panel.background = element_blank()) +
  geom_line(size=1,color='#00526d') +
  geom_vline(xintercept = as.Date(c("2020-05-22","2020-06-25","2020-08-15",
                                    "2020-09-11","2020-11-07","2021-01-03",
                                    "2021-03-11")),
                                     linetype=4, colour="black")

# Plot with snht p = 14 and p = 28 breakpoints
ts_ggplot(sp[7:400,] %>% select(ma,data)) +
  theme(panel.grid = element_blank(),panel.background = element_blank()) +
  geom_line(size=1,color='#00526d') +
  geom_vline(xintercept = as.Date(c("2020-04-22","2020-05-23","2020-07-04",
                                    "2020-08-15","2020-08-31","2020-11-07",
                                    "2021-01-03","2021-03-05")),
```

```
                                    linetype=1, colour="pink",lwd=.8) +
  geom_vline(xintercept = as.Date(c("2020-05-22","2020-06-25","2020-08-15",
                                    "2020-09-11","2020-11-07","2021-01-03",
                                    "2021-03-11")),
                                    linetype=3, colour="black",lwd=.8)
```

Apêndice C

Detecção de pontos em dados referentes a Covid-19

C.1 SNHT com $p = 14$

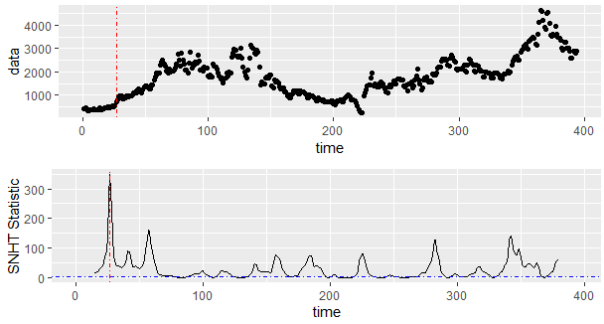

Figura C.1: SNHT com $p = 14$ detectando mudança em 2020-04-22

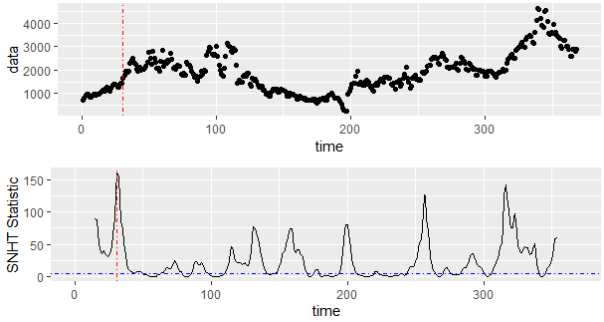

Figura C.2: SNHT com $p = 14$ detectando mudança em 2020-05-23

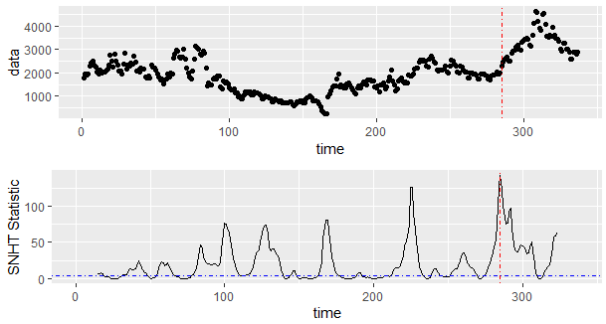

Figura C.3: SNHT com $p = 14$ detectando mudança em 2021-03-11

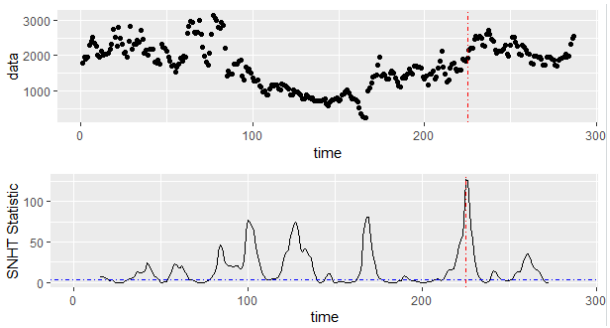

Figura C.4: SNHT com $p = 14$ detectando mudança em 2021-01-03

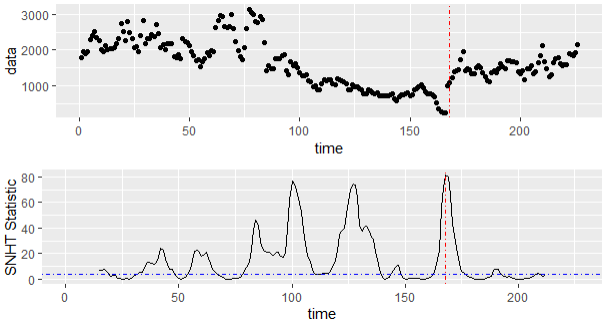

Figura C.5: SNHT com $p = 14$ detectando mudança em 2020-11-07

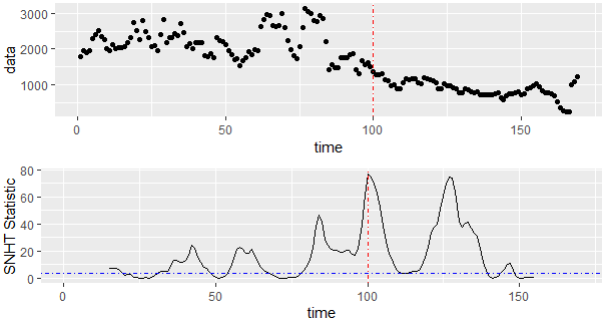

Figura C.6: SNHT com $p = 14$ detectando mudança em 2020-08-31

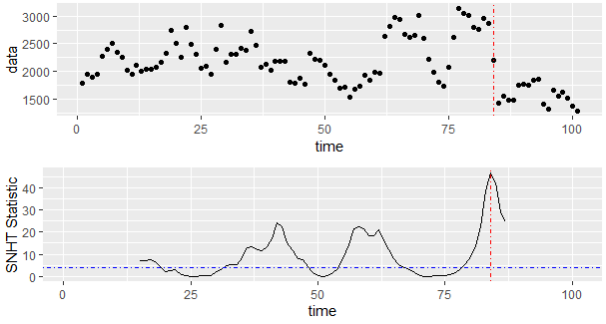

Figura C.7: SNHT com $p = 14$ detectando mudança em 2020-08-15

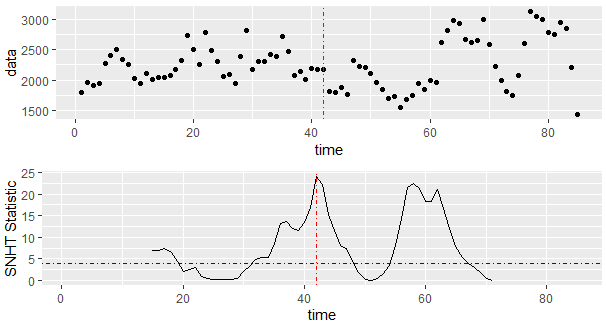

Figura C.8: SNHT com $p = 14$ detectando mudança em 2020-07-04

C.2 SNHT com $p = 28$

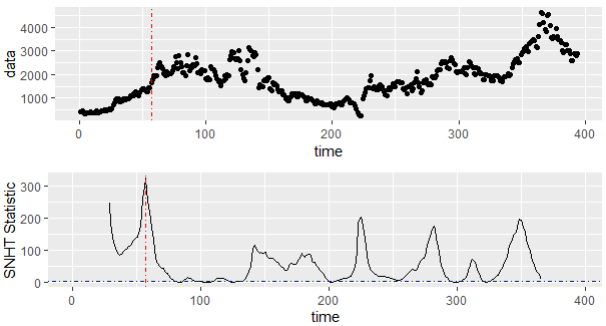

Figura C.9: SNHT com $p = 28$ detectando mudança em 2020-05-22

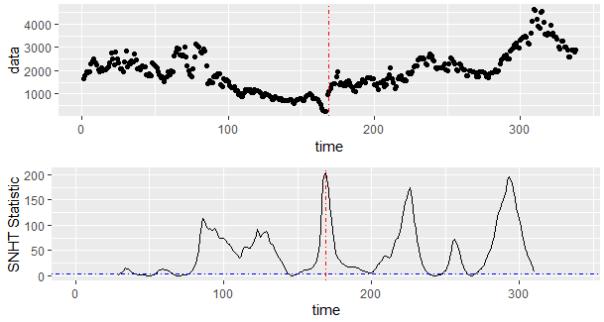

Figura C.10: SNHT com $p = 28$ detectando mudança em 2020-09-11

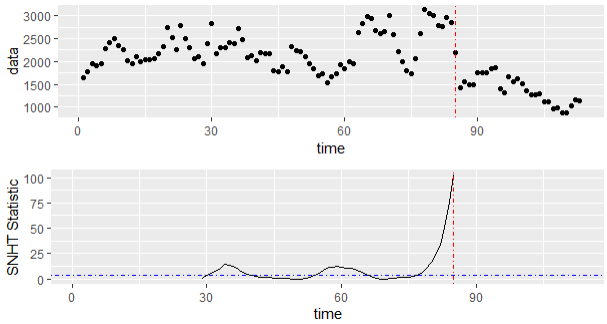

Figura C.11: SNHT com $p = 28$ detectando mudança em 2020-08-15

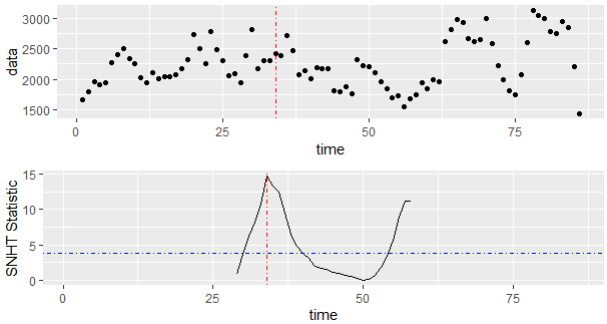

Figura C.12: SNHT com $p = 28$ detectando mudança em 2020-06-25

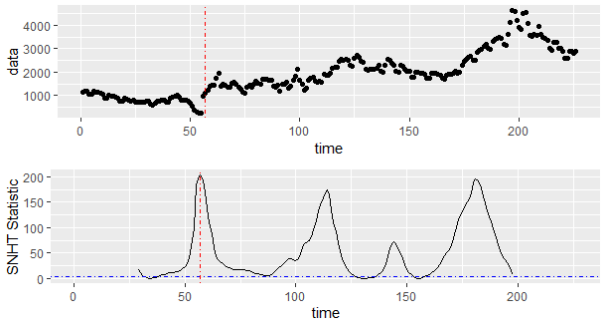

Figura C.13: SNHT com $p = 28$ detectando mudança em 2020-11-07

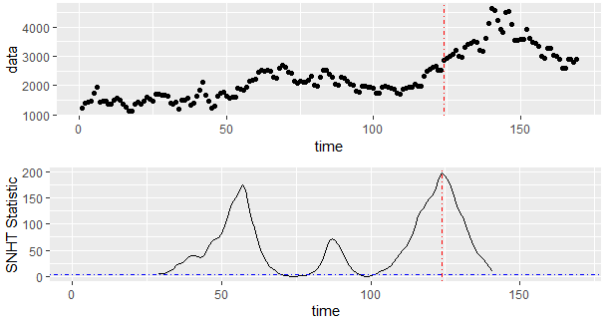

Figura C.14: SNHT com $p = 28$ detectando mudança em 2021-03-11

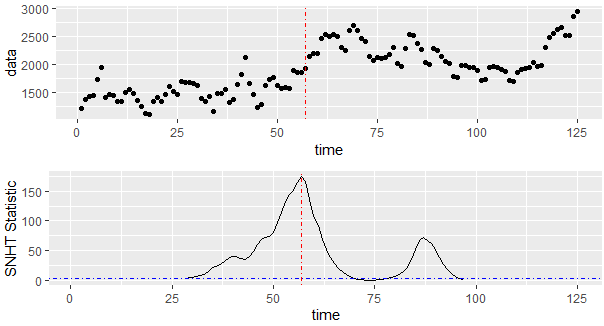

Figura C.15: SNHT com $p = 28$ detectando mudança em 2021-01-03

Referências Bibliográficas

Alexandersson, H. (1986). A homogeneity test applied to precipitation data. *Journal of climatology*, **6**(6), 661–675.

Alexandersson, H. e Moberg, A. (1997). Homogenization of swedish temperature data. part i: Homogeneity test for linear trends. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, **17**(1), 25–34.

Bolfarine, H. e Sandoval, M. C. (2001). *Introdução à inferência estatística*, volume 2. SBM.

Browning, J. (2017). Robust and non-robust snht tests for changepoint detection.

Costa, H. C. (2019). *Modelos de Médias Móveis (MA)*. RPubs.

Forato, F. (2020). São paulo pode chegar a 720 mil casos da covid-19 até 15 de agosto, diz pesquisa. https://canaltech.com.br/saude/sao-paulo-pode-chegar-a-720-mil-casos-da-covid-19-ate-15-de-agosto-diz-pesquisa-169271/. 2021-05-29.

Gomes, R. (2020). São paulo deve ter mil mortes pela covid-19 por semana às vésperas do natal. https://www.redebrasilatual.com.br/saude-e-ciencia/2020/12/natal-covid-19-sao-paulo/. 2021-05-29.

Governo, P. (2020). Governo de sp prorroga quarentena até 22 de abril de 2020. https://www.saopaulo.sp.gov.br/spnoticias/governo-de-sao-paulo-prorroga-quarentena-ate-22-de-abril/. 2021-05-29.

Haimberger, L. (2007). Homogeneização de séries temporais de temperatura de radiossonda usando estatísticas de inovação. *Journal of Climate*.

Harvey, A. C. e Durbin, J. (1986). The effects of seat belt legislation on british road

casualties: A case study in structural time series modelling. *Journal of the Royal Statistical Society: Series A (General)*, **149**(3), 187–210.

Morettin, P. A. e Toloi, C. (2004). *Análise de séries temporais*. Blucher.

SUS, D. (2021). Painel coronavírus. https://covid.saude.gov.br/. 2021-05-02.

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

Toledo, K. (2020). Coronavírus: lockdown será inevitável em são paulo se isolamento não subir. https://saude.abril.com.br/medicina/coronavirus-lockdown-sera-inevitavel-em-sao-paulo-se-isolamento-nao-subir/. 2021-05-29.