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

Detecção de alteração no regime pluviométrico da Serra da Cantareira via Modelo de Espaço de Estados

Robert Luis Alves de Souza

Trabalho de Conclusão de Curso

Robert Luis Alves de Souza

Modelo de Espaço de Estados para detecção de alteração no regime pluviométrico da Serra da Cantareira

Este exemplar corresponde à redação final do trabalho de conclusão de curso devidamente corrigido e defendido por Robert Luis Alves de Souza e aprovado pela banca examinadora.

São Carlos, 07 de Janeiro de 2021.

Banca Examinadora

- Professor Márcio Luis Lanfredi Viola
- Professor Michel Helcias Montoril

Dedicatória

Dedico este trabalho a todos da minha família e amigos que estiveram sempre presentes na minha vida. Vocês são e sempre serão essenciais.

Agradecimentos

Primeiramente gostaria de agradecer a toda minha família, por todo apoio, por todo conhecimento, histórias de superação e motivação que sempre me deram. Sem vocês eu não seria nada.

Agradeço também a minha noiva que me auxiliou, amparou, cuidou e amou em todos os momentos.

Um agradecimento em especial para os membros da família que não estão mais entre nós, mas que sempre serão lembrados. Infelizmente dona Nina, você não pode me ver entrando na universidade e tão pouco me formando, mas tenho certeza que está fazendo uma tremenda festa aí de cima. E meu amado avô Anedino, o senhor sempre foi e sempre será um exemplo para mim, obrigado por tudo. Para nós que ficamos apenas nos resta a saudade, carinho e admiração. A todos vocês, um dia nós nos encontramos novamente.

Agradeço pelos meus antepassados que tanto lutaram para que hoje eu fosse um dos primeiros membros da família a cursar uma universidade pública de qualidade.

Agradeço também a minha orientadora Maria Sílvia que dedicou tempo, carinho e atenção para que juntos concluíssemos esse trabalho.

Agradeço a todos os docentes e funcionários da UFSCar que viabilizaram todo o caminho do meu aprendizado até o presente momento.

Por fim agradeço a todos os meus amigos por sempre estarem do meu lado e me apoiaram em todos os momentos.

Resumo

No período entre 2014 e 2016 a região sudeste sofreu com um processo de forte estiagem. De acordo com o encarte especial de 2014, elaborado pela Agência Nacional de Águas (2014) do Ministério do Meio Ambiente, o regime de chuvas na região apresentou uma redução intensa e gradual dos níveis pluviométricos, o que causou a crise hídrica, afetando o abastecimento de água na região. Sendo assim, a principal tese deste relatório é de que o regime de chuvas se alterou, baseado no fato de que o comportamento pluviométrico estava muito abaixo da média em diferentes regiões do país.

Com isso, este trabalho teve o intuito de verificar se essa alteração realmente ocorreu, ou foi apenas um evento extremo. Para realizar esse estudo, foram utilizados os dados pluviométricos disponibilizados no site da SABESP (`http://mananciais.sabesp.com.br/Home`) e adotado o modelo espaço de estados, por possuir uma grande flexibilidade e um erro de previsão pequeno.

A partir disso, percebemos que, apesar da variabilidade e presença de estrutura de longa duração, o modelo foi capaz de identificar uma possível alteração no regime pluviométrico no Sistema Cantareira logo no início da previsão, porém vemos que para os passos seguintes o modelo não foi capaz de identificar pontos fora de controle.

Palavras-chave: *Filtro de Kalman, Modelo Espaço de Estados, Previsão.*

Glossário

1. $t = 1, 2, ..., T$;
2. y_t é o vetor de dados observados para cada instante t de ordem $q \times 1$;
3. F_t é a matriz de observação $q \times 1$;
4. x_t é o vetor de dados não observados (conhecido também como vetor de estados) para cada instante t de ordem $p \times 1$;
5. $\boldsymbol{v_t}$ e $\boldsymbol{w_t}$ são ruídos não correlacionados;
6. $\boldsymbol{G_t}$ é chamada matriz de transição de ordem $p \times p$;
7. $\boldsymbol{R}$ é uma matriz de seleção (no nosso caso é a identidade);
8. $\boldsymbol{V}$ e $\boldsymbol{W}$ são matrizes de covariâncias;
9. $\boldsymbol{\mu}$ vetor de médias (no nosso caso é 0)
10. $\widetilde{y}_s$, vetor de informações disponíveis até instante s
11. Θ parâmetro conhecido
12. $x_{t|s}$ estimador do vetor de estados no instante t dado s
13. $P_{t|s}$ é a variância do estimador estimador de estados no instante s
14. $x_{0|0}$ estado inicial
15. $P_{0|0}$ matriz de covariância inicial
16. K_t ganho de Kalman
17. $\boldsymbol{\Psi}$ vetor de hiperparâmetros

Sumário

Lista de Figuras

Capítulo 1

Introdução

De acordo com Freitas (2019), o Brasil é um país de proporções continentais e, devido ao seu tamanho e localização, apresenta diversos climas, dentre eles os principais são: equatorial, semiárido, tropical, tropical de latitude, tropical atlântico e subtropical. Compreender essa diversidade climática do Brasil nos permite traçar uma relação entre o regime de chuvas e as estações do ano.

A região sudeste sofreu com um forte processo de estiagem entre os anos de 2014 e 2017. De acordo com o encarte especial de 2014, elaborado pela Agência Nacional de Águas (2014) (Agência Nacional de Águas que tem como missão implementar e coordenar a gestão compartilhada e integrada dos recursos hídricos e regular o acesso a água, promovendo o seu uso sustentável em benefício da atual e das futuras gerações), do Ministério do Meio Ambiente, o regime de chuvas na região apresentou uma redução intensa e gradual dos níveis pluviométricos, o que causou a crise hídrica, afetando o abastecimento de água na região. Sendo assim, a principal tese deste relatório é de que o regime de chuvas se alterou, baseado no fato de que o comportamento pluviométrico estava muito abaixo da média em diferentes regiões do país.

Devido a este fenômeno atípico, o volume de água armazenado no Sistema Cantareira reduziu drasticamente a ponto de, em 2014, ser utilizado o volume morto, que representa a "reserva de água"para usos emergenciais. De acordo com Puga (2018),

> A correta identificação das prováveis causas da seca é importante para a escolha adequada dos tipos de ferramentas, políticas e medidas de adaptação ou mitigação que devem ser levadas em consideração para o enfrentamento do problema. Para tanto, a redefinição conceitual da seca pode ajudar no enfrentamento e em medidas adaptativas mais adequadas.

Portanto, entender as causas desse fenômeno nos ajuda a tomar ações para que em um futuro próximo a região não sofra com problemas decorrentes do abastecimento de água.

Como possuímos os dados diários das medições dos índices pluviométricos, na região do Sistema Cantareira, e como a região sofreu com uma mudança de comportamento das chuvas, é adequado utilizar uma metodologia que incorpore tais mudanças. Sendo assim, o objetivo deste trabalho é detectar se houve ou não alteração no padrão do regime pluviométrico no Sistema Cantareira através da abordagem clássica do modelo espaço de estados.

Capítulo 2

Material e Métodos

Para realização deste trabalho será utilizado o banco de dados disponíveis no site `http://mananciais.sabesp.com.br/Home` da Sabesp (Companhia de Saneamento Básico do Estado de São Paulo), do qual serão analisados os dados pluviométricos captados diariamente do Sistema Cantareira.

Será utilizado o modelo espaço de estados para verificar se a tese do relatório da Agência Nacional de Águas (2014) está correta, ou seja, verificar se de fato o regime de chuvas se alterou no Sistema Cantareira a partir de 2012.

2.1 Modelo Espaço de Estados

O modelo espaço de estados é um modelo dinâmico que consegue incorporar mudanças de comportamento ao longo do tempo envolvendo variáveis não observáveis, como tendência, sazonalidade, ciclos, etc.

> *A idéia dessa decomposição da série temporal surgiu nos trabalhos de Holt (1957) e Winters (1960), que desenvolveram as técnicas de alisamento exponencial. Aproveitando essa idéia, na década de 60 surgiram alguns trabalhos formalizando a metodologia de modelos estruturais, dentre os quais pode-se citar os de Muth (1960), Theil e Wage (1964) e Nerlove* et al. *(1964). Já na década de 70 surgiram os primeiros modelos de previsão Bayesianos utilizando a modelagem dinâmica, nos trabalhos de Harrison e Stevens (1971, 1976). Entretanto, mediante a dificuldade computacional da época e o aparecimento dos modelos ARIMA de Box e Jenkins (1976), procedimentos utilizando a idéia de decomposição em componentes não-observáveis só voltaram a ser desenvolvidos no final da década de 80. (Franco* et al. *(2009), p.1)*

Com o a criação do programa STAMP (*Structural Time Series Analyser Modeller and Predictor*), no final da década de 80, junto com as pesquisas de Harvey e Fernandes (1989); Harvey (1993), os modelos dinâmicos se popularizaram.

Atualmente existem pacotes disponíveis em diversas linguagens como, por exemplo, *Ox* (Doornik (2009)) e *R* (R Core Team (2019)) que realizam a estimação, suavização e previsão para modelos dinâmicos, como os pacotes *SfPackevers* versão 2.2 (Koopman *et al.* (1999)) e *dlm* (Petris (2008)).

Com o modelo expresso na forma de espaço de estados, é possível a aplicação do filtro de Kalman (Kalman (1960)) para a estimação dos hiperparâmetros desconhecidos do modelo e realização de previsões.

De acordo com Commandeur e Koopman (2007), o modelo espaço de estados é representado na forma multivariada e possui apenas duas equações sendo elas:

$$\boldsymbol{Y_t} = \boldsymbol{F_t X_t} + \boldsymbol{v_t},\ \boldsymbol{v_t} \overset{i.i.d}{\sim} N(\boldsymbol{0}, \boldsymbol{V_t}), \tag{2.1}$$

$$\boldsymbol{x_t} = \boldsymbol{\mu} + \boldsymbol{G_t x_{t-1}} + \boldsymbol{R w_t},\ \boldsymbol{w_t} \overset{i.i.d}{\sim} N(\boldsymbol{0}, \boldsymbol{W}), \tag{2.2}$$

em que,

1. $t = 1, 2, ..., T$;
2. $\boldsymbol{Y_t}$ é o vetor de dados observados para cada instante t de ordem $q \times 1$;
3. $\boldsymbol{F_t}$ é a matriz de observação $q \times 1$;
4. $\boldsymbol{X_t}$ é o vetor de dados não observados para cada instante t de ordem $p \times 1$;
5. $\boldsymbol{v_t}$ e $\boldsymbol{w_t}$ são ruídos não correlacionados;
6. $\boldsymbol{G_t}$ é chamada matriz de transição de ordem $p \times p$;
7. $\boldsymbol{R}$ é uma matriz de seleção (no nosso caso usaremos a identidade);
8. $\boldsymbol{V}$ e $\boldsymbol{W}$ são matrizes de covariâncias;
9. $\boldsymbol{\mu}$ é o vetor de médias (no nosso caso usaremos $\boldsymbol{0}$)

A equação (2.1) é conhecida como equação de observação que relaciona $\boldsymbol{Y_t}$ com um vetor não observado $\boldsymbol{X_t}$. Já a equação (2.2) é conhecida como equação de estado, dependente basicamente de $\boldsymbol{x_{t-1}}$. Através dessas duas equações é possível descrever várias estruturas de séries temporais. Tomando como exemplo, pode-se adicionar variáveis indicadoras de

quebras estruturais, adicionar tendência ou sazonalidade, erros normais ou não normais, adicionar variáveis explicativas, etc.

De acordo com Durbin e Koopman (2001), como o modelo de natureza Markoviana é escrito de maneira recursiva, os cálculos computacionais apresentam uma grande flexibilidade, uma vez que é possível alterar a estrutura com certa facilidade.

Além disso, Harvey (1990) aponta que uma outra vantagem desse modelo é a possibilidade de trabalhar com dados *missing* e agregação temporal, com possibilidade de também reformular o modelo para um espaço de tempo contínuo, tratando assim observações irregularmente espaçadas no tempo.

Segundo Alencar (2006), para realizar a modelagem, devemos seguir os seguintes passos:

1. Definir/Identificar o modelo;

2. Identificar os parâmetros necessários;

3. Obter o vetor de previsões para o vetor de estados;

4. Estimar os parâmetros identificados;

5. Verificar a adequabilidade do modelo.

2.2 Filtro de Kalman

Introdução

O filtro de Kalman surgiu na década de 60 como uma alternativa para solucionar problemas lineares, utilizando uma filtragem a dados discretos. Mais tarde, essa metodologia passou a ser incorporada em áreas como Engenharia e Estatística. Atualmente, existem vários trabalhos publicados que adotam esse tipo de abordagem. Para a construção deste capítulo, tomou-se como base as obras de Aiube (2005), Alencar (2006), Durbin e Koopman (2001), Harvey (1989) e, por fim, Jazwinski (1970).

A título de informação, em meados dos anos 1960, após Kalman publicar seu artigo sobre um procedimento recursivo para solução de problemas lineares por filtragem de dados discretos, Kalman e Bucy (1961) desenvolveram uma versão em tempo contínuo do filtro Kalman, conhecida por filtro de Kalman-Bucy.

Algoritmo do filtro de Kalman

Basicamente, o filtro consiste em um conjunto de operações matemáticas recursivas muito eficientes para estimação de qualquer parâmetro desconhecido de um modelo na forma de espaço de estados, devido a minimização do erro quadrático médio. Por meio da variável de observação, pode-se prever de forma eficiente uma variável não observável, denotada por "variável de estado".

Esse algoritmo tem como objetivo obter o estimador ótimo do vetor de estados no instante t, baseado na informação conhecida até o instante s, onde $s < t$. Ele permite a estimação dos parâmetros desconhecidos via maximização da função de verossimilhança, decompondo o erro de previsão.

Para maior elucidação do funcionamento do algoritmo, a seguir temos um figura que representa de forma esquemática os processos do algoritmo do filtro de Kalman. A composição das equações será discutida nos tópicos a seguir.

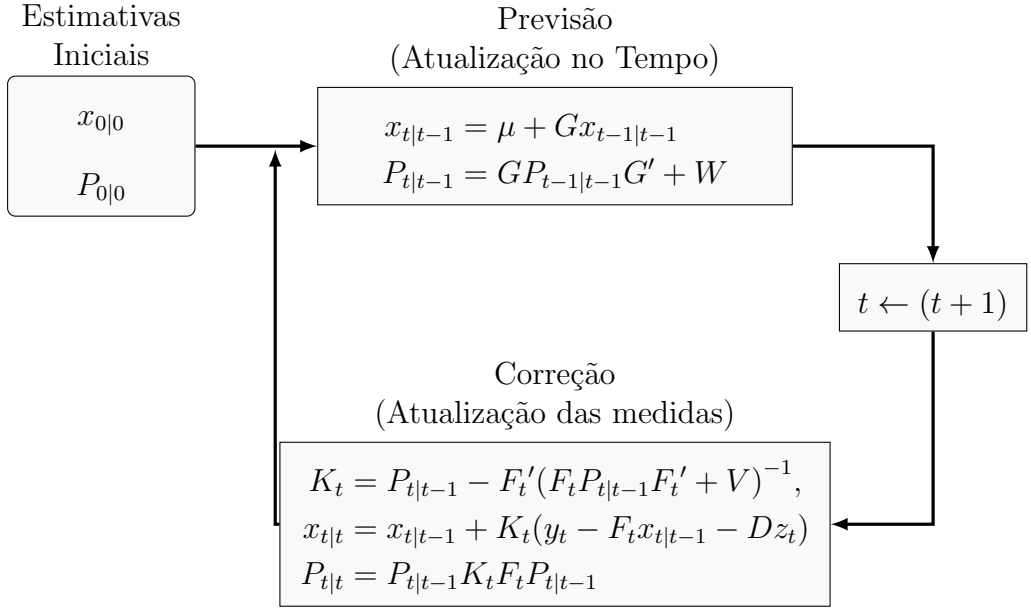

Figura 2.1: Desenho esquemático do Filtro de Kalman.

De acordo com a Figura 2.1, primeiramente obtemos as estimativas iniciais e com elas realizamos a previsão para o próximo instante com a informação que possuímos até então. Após isso, temos uma nova observação, sendo necessária a atualização da estimativa, então há o incremento de uma unidade no tempo e corrigimos as estimativas. Essa correção ocorre através do ganho de Kalman que é um tipo de ponderação entre a predição e o valor calculado previamente, ou seja, a estimativa inicial. Com os valores corrigidos, é possível

realizar uma nova predição e, através dessa, damos início a um procedimento recursivo. O critério de parada é definido de acordo com o nível de tolerância de erro de estimativa.

Para a realização desse procedimento, vamos adotar a notação apresentada em Shumway e Stoffer (2000) e Alencar (2006). Seja $\widetilde{\boldsymbol{y}}_{\boldsymbol{s}} = (y_1, ..., y_s)$, o vetor de informações disponíveis até instante s. Vamos definir:

$$\begin{aligned}
\boldsymbol{x}_{\boldsymbol{t|s}} &= \mathbf{E}[\boldsymbol{X}_{\boldsymbol{t}}|\widetilde{\boldsymbol{y}}_{\boldsymbol{s}}], \\
\boldsymbol{P}_{\boldsymbol{t|s}} &= \mathbf{E}[(\boldsymbol{X}_{\boldsymbol{t}} - \boldsymbol{x}_{\boldsymbol{t|s}})(\boldsymbol{X}_{\boldsymbol{t}} - \boldsymbol{x}_{\boldsymbol{t|s}})'|\widetilde{\boldsymbol{y}}_{\boldsymbol{s}}], \\
\boldsymbol{P}_{\boldsymbol{t_1,t_2|s}} &= \mathbf{E}[(\boldsymbol{x}_{\boldsymbol{t_1}} - \boldsymbol{x}_{\boldsymbol{t_1|s}})(\boldsymbol{x}_{\boldsymbol{t_2}} - \boldsymbol{x}_{\boldsymbol{t_2|s}})'|\widetilde{\boldsymbol{y}}_{\boldsymbol{s}}],
\end{aligned}$$

em que $\boldsymbol{x}'$ corresponde ao vetor $\boldsymbol{x}$ transposto.

Sobre a hipótese de normalidade dos dados temos, segundo Aiube (2005), que "A derivação do filtro de Kalman, surge do fato probabilístico de que, tanto os ruídos das equações de medição e transição, como o vetor inicial de estado, possuem distribuição Normal. Sendo assim, basta apenas os dois primeiros momentos para descrever todos os instantes."

Dito isso, sobre suposição de normalidade, obteremos a estimativa $x_{t|s}$ que é a esperança de X_t dado o vetor observado e a variância desse estimador, denotada por $P_{t|s}$. Ressalta-se que o estimador obtido é o que minimiza o erro quadrático médio dos estimadores lineares, mesmo sem a hipótese de normalidade (Shumway e Stoffer,2000).

De acordo com Alencar(2006), através dos valores iniciais $\boldsymbol{x_{0|0}} = \boldsymbol{\mu}$ e $\boldsymbol{P_{0|0}} = \boldsymbol{\Sigma}$ o Filtro de Kalman apresenta equações de predição para $s = t - 1$ e de atualização para $s = t$, com $t = 1, ..., T$.

Equações de predição:

$$\begin{aligned}
\boldsymbol{x_{t|t-1}} &= \boldsymbol{\mu} + \boldsymbol{G}\boldsymbol{x_{t-1|t-1}}, \\
\boldsymbol{P_{t|t-1}} &= \boldsymbol{G}\boldsymbol{P_{t-1|t-1}}\boldsymbol{G'} + \boldsymbol{W}.
\end{aligned}$$

Como conhecemos $\mathbf{Y_t}$ é possível atualizarmos nossa predição através das

Equações de atualização:

$$\begin{aligned}
\boldsymbol{x_{t|t}} &= \boldsymbol{x_{t|t-1}} + \boldsymbol{K_t}(\boldsymbol{y_t} - \boldsymbol{F_t}\boldsymbol{x_{t|t-1}} - \boldsymbol{D}\boldsymbol{z_t}), \\
\boldsymbol{P_{t|t}} &= \boldsymbol{P_{t|t-1}} - \boldsymbol{K_t}\boldsymbol{F_t}\boldsymbol{P_{t|t-1}},
\end{aligned}$$

em que $\boldsymbol{K_t}$ e $\boldsymbol{F_t}$ são, respectivamente, o ganho de Kalman e a matriz variâncias. O ganho de Kalman é dado pela expressão:

$$\boldsymbol{K_t} = \boldsymbol{P_{t|t-1}F_t'(F_tP_{t|t-1}F_t' + V)}^{-1}.$$

Durante o processo de filtragem, obtêm-se o valor esperado de $\boldsymbol{X_t}$, dadas as informações até o instante t. Em termos matemáticos, $\boldsymbol{\hat{X}_t = E(X_t|\widetilde{y}_s)}$.Vale ressaltar que os estimadores suavizados são correspondentes aos estimadores via minimização do erro quadrático médio. Com as equações de suavização é possível obter esse valor baseado em informações posteriores ao instante t da seguinte forma:

Equações de Suavização:

$$\boldsymbol{x_{t-1|T} = x_{t-1|t-1} + J_{t-1}(x_{t|T} - Gx_{t-1|t-1} - \mu)},$$
$$\boldsymbol{P_{t-1|T} = P_{t-1|t-1} + J_{t-1}(P_{t|T} - P_{t|t-1})J_{t-1}'},$$

em que,

$$\boldsymbol{J_{t-1} = P_{t-1|t-1} - G'(P_{t|t-1})}^{-1}.$$

A demonstração das equações do Filtro de Kalman são apresentadas nos trabalhos de Harvey (1989) e Hamilton (1994).

Estimação dos hiperparâmetros por máxima verosimilhança

Seja $\boldsymbol{\Psi} = (\Psi_1, \Psi_2, ..., \Psi_p)'$ o vetor de hiperparâmetros que pertence ao espaço paramétrico $\mathbb{R}_+^p$. Por exemplo, no modelo de nível local, $\boldsymbol{\Psi} = (\sigma_v^2, \sigma_w^2) \in \mathrm{R}_+^2$.

O modelo espaço de estados depende da estimação dessas quantidades, pois componentes como média e variância são funções desses hiperparâmetros. Sendo assim, vamos utilizar a estimação por máxima verossimilhança, por possuir propriedades ótimas sob determinadas condições de regularidade abordadas, por exemplo, em Migon e Gamerman (1999) e Casella e Berger (2002).

A função de verossimilhança pode ser obtida através dos valores resultantes da aplicação do Filtro de Kalman, pois este retorna os valores da variável de observação $\mathbf{Y_t}$. Logo, assumindo a normalidade, podemos dizer que $\boldsymbol{Y_t|y_1, y_2, ..., y_{s-1}} \sim N(\boldsymbol{\widetilde{y}_{t|t-1}, F_t})$. Segundo Franco *et al.* (2009), a função de densidade preditiva é dada por

$$p(\mathbf{Y_t}|\widetilde{y}_{s-1}, \Psi) = (2\pi)^{-1/2}\,|\boldsymbol{F_t}|\,-1/2exp(-\frac{1}{2}\boldsymbol{(Y_t - \widetilde{y}_{t|t-1})^t F_t}^{-1}\boldsymbol{(Y_t - \widetilde{y}_{t|t-1})}).$$

Como as observações são independentes e identicamente distribuídas, fazendo $v_t = Y_t - \widetilde{y}_{t|t-1}$ temos que o logaritmo da função de verossimilhança é

$$\ell(\boldsymbol{Y_n}|\boldsymbol{\Psi}) = -\frac{n}{2}ln(2\pi) - \frac{1}{2}\sum_{t=1}^{n} ln\,|\boldsymbol{F_t}| - \frac{1}{2}\sum_{t=1}^{n} \boldsymbol{v_t}'\,\boldsymbol{F_t}^{-1}\boldsymbol{v_t}, \tag{2.3}$$

em que $\boldsymbol{Y_n}$ é $\boldsymbol{Y_n} = (y_1, y_2, ..., y_n)'$ e Ψ é o vetor de hiperparâmetros dado.

O vetor estimado de hiperparâmetros, $\hat{\Psi}$, é obtido maximizando (2.3). Os estimadores de máxima verossimilhança não podem ser obtidos de forma analítica, logo utilizam-se métodos computacionais numéricos para realizar a maximização, onde os métodos mais

comuns são Newton-Raphson e BFGS.

Capítulo 3

Aplicação da Metodologia

Para comprovar a eficácia do método apresentado, iremos aplicá-lo em um exemplo amplamente discutido na literatura. No capítulo 8 de Commandeur e Koopman (2007), eles abordam uma série de procedimentos para o tratamento de modelos de espaço de estados univariados.

A seguir será realizada a análise dos mesmos conjuntos de dados, modelos e a sequência de análise que os autores utilizaram, para verificar a eficácia da previsão através do filtro de Kalman com o modelo expresso na forma de espaço de estados.

Podemos expressar os modelos como no formato geral mostrado em (2.1) e (2.2), mas lembrando que estaremos utilizando a forma univariada, ou seja, os termos Y_t e ϵ_t são vetores, $\boldsymbol{Z_t}$ é uma matriz de ordem $m \times 1$ denotado por vetor de observação, $\boldsymbol{T_t}$ de ordem $m \times m$, é a matriz de transição, $\boldsymbol{\alpha_t}$ é o vetor de estados de ordem $m \times 1$, em que m se refere ao número de elementos. Para fins de simplificação, os autores adotaram que a matriz de seleção $\boldsymbol{R_t}$ é a matriz identidade de ordem $m \times m$. Serão apresentados a seguir, alguns modelos que serão utilizados nos exemplos.

O modelo mais simples que conseguimos utilizar é o chamado *modelo de nível local* que é caracterizado por conter apenas as componentes de nível (μ_t) e de erro (ϵ_t). Definimos então:

Modelo de nível local

$$\boldsymbol{X_t} = \boldsymbol{\mu_t}, \boldsymbol{v_t} = \boldsymbol{\xi_t}, \boldsymbol{F} = \boldsymbol{G} = \boldsymbol{R} = \mathbb{I}, \boldsymbol{w_t} = \boldsymbol{\sigma_\xi^2},$$

Com isso, as equações de espaço de estado são escritas como:

$$\begin{aligned} \boldsymbol{Y_t} &= \boldsymbol{\mu_t} + \boldsymbol{v_t},\ \boldsymbol{v_t} \sim\ N(0, \sigma_v^2), \\ \boldsymbol{\mu_{t+1}} &= \boldsymbol{\mu_t} + \boldsymbol{\xi_t},\ \boldsymbol{\xi_t} \sim\ N(0, \sigma_\xi^2). \end{aligned} \tag{3.1}$$

Modelo Estrutural Básico

Em séries temporais é frequente a existência de uma componente sazonal, ou seja, um comportamento que ocorre com uma dada periodicidade. Tal sazonalidade pode ser modelada através de componentes trigonométricas ou por fatores. Denotamos a componente sazonal como $\boldsymbol{\gamma_t}$ e veremos a seguir como modelamos essa componente.

Admitindo que existe uma periodicidade s e se o padrão sazonal é constante no tempo, então os valores sazonais de 1 até s podem ser modelados por escalares $\gamma_1^*, ..., \gamma_s^*$ com a

seguinte restrição

$$\sum_{j=0}^{s-1} \gamma_j^*.$$

A título simplificação para entendimento da expressão que virá a seguir, vamos assumir agora que existem s meses por ano, temos que $\boldsymbol{\gamma_t} = \boldsymbol{\gamma_j^*}$ onde $t = s(i-1)+j$ com $i = 1, 2, ...$ e $j = 1, 2, ..., s$. Diante desse cenário, devido a restrição imposta $\sum_{j=0}^{s-1} \boldsymbol{\gamma_{t+1-j}} = 0$ conseguimos isolar o termo γ_{t+1} resultando em

$$\boldsymbol{\gamma_{t+1}} = -\sum_{j=1}^{s-1} \boldsymbol{\gamma_{t+1-j}},$$

para t iniciando em $s-1$ com incrementos unitários.

Na prática, como o padrão sazonal não é determinística existe uma componente de erro associada, logo a modelagem do fator sazonal resulta em

$$\boldsymbol{\gamma_{t+1}} = -\sum_{j=1}^{s-1} \boldsymbol{\gamma_{t+1-j}} + \boldsymbol{\varepsilon_t} \tag{3.2}$$

em que $\boldsymbol{\varepsilon_t} \sim N(0, \sigma_{\boldsymbol{\varepsilon}}^2)$, para $t = 1, ..., s-1$.

Dito isso, podemos agora construir o modelo estrutural básico que consiste em uma componente de nível local ((μ_t)), componente de tendência linear (u_t) e uma componente sazonal. Seguindo o exemplo onde temos $s = 12$ meses, modelo é definido como:

$$\begin{aligned}
\boldsymbol{Y_t} &= \boldsymbol{\mu_t} + \boldsymbol{\gamma_t} + \boldsymbol{v_t},\ \boldsymbol{v_t} \sim\ N(0, \sigma_v^2),\\
\boldsymbol{\mu_{t+1}} &= \boldsymbol{\mu_t} + \boldsymbol{u_t} + \boldsymbol{\xi_t},\ \boldsymbol{\xi_t} \sim\ N(0, \sigma_\xi^2),\\
\boldsymbol{u_{t+1}} &= \boldsymbol{u_t} + \boldsymbol{\delta_t},\ \boldsymbol{\delta_t} \sim\ N(0, \sigma_\delta^2),\\
\boldsymbol{\gamma_{1,t+1}} &= -(\boldsymbol{\gamma_{1,t}} + \boldsymbol{\gamma_{2,t}} + ... + \boldsymbol{\gamma_{11,t}}) + \boldsymbol{\zeta_t},\ \boldsymbol{\zeta_t} \sim\ N(0, \sigma_\zeta^2),\\
\boldsymbol{\gamma_{2,t+1}} &= \boldsymbol{\gamma_{1,t}},\\
&\vdots\\
\boldsymbol{\gamma_{11,t+1}} &= \boldsymbol{\gamma_{10,t}}.
\end{aligned}$$

Vale ressaltar que como visto acima, a componente de sazonalidade requer a introdução de $s-1$ equações. Sendo assim para nosso exemplo com $s = 12$ utilizamos 11 equações. Ademais, no capítulo 4 utilizaremos esse modelo para análise dos dados.

3.1 Análise dos conjuntos utilizados pelos autores

Será utilizado o software R para realizar as análises com base no código elaborado por Radhakrishna (2013), que utiliza o banco de dados disponibilizado no livro dos autores Commandeur e Koopman (2007). Nas análises que virão a seguir serão utilizados os mesmos conjuntos de dados e análises que os autores realizaram no capítulo 8.

Figura 3.1: Estimação do erro da variância para o modelo determinístico sazonal e o nível estocástico aplicado no log do banco UK drivers KSI.

Através da Figura 3.1, observa-se que a variância se inicia com um valor alto, diminuindo drasticamente a ponto de ser praticamente 0 e, ao final, observa-se valores altos. Esse comportamento era esperado uma vez que a incerteza, tanto no começo como no final da série, é maior.

É possível elaborar um intervalo de confiança baseado na distribuição normal, ao nível de 90%, da forma:

$$\mu_t \pm 1.64\sqrt{Var(\mu_t)},$$

onde μ_t é dado pela Equação (3.1).

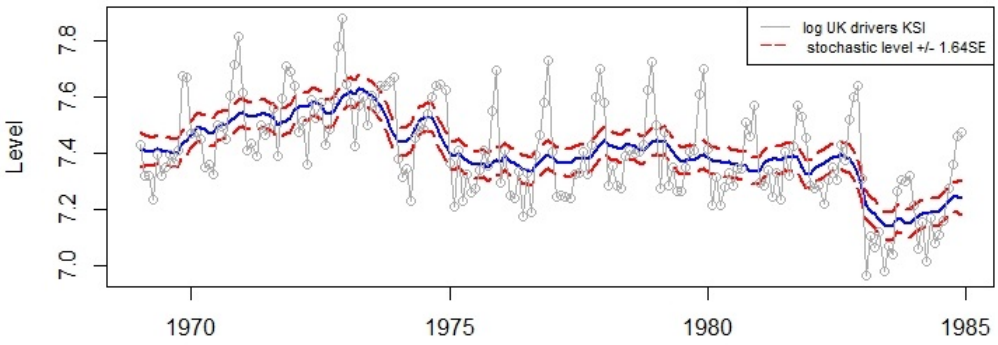

Figura 3.2: Série log UK drivers KSI com intervalo de confiança de 90% baseado na distribuição Normal.

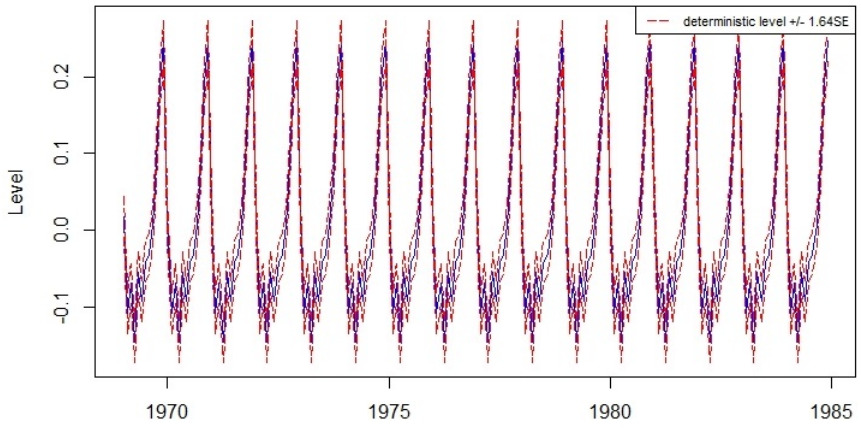

Figura 3.3: Sazonalidade com intervalo de confiança de 90% baseado na distribuição Normal.

A partir das Figuras 3.2 e 3.3 é possível notar que os intervalos acompanham adequadamente a média e a sazonalidade das observações, respectivamente.

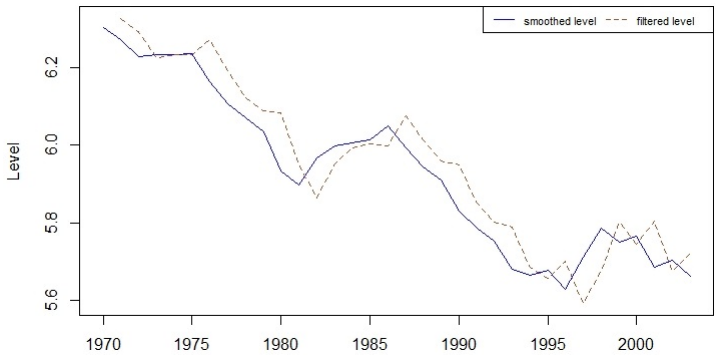

Figura 3.4: Estado suavizado e filtrado do modelo de nível local aplicado aos dados Norwegian road traffic fatalities.

Na Figura 3.4 é possível notar que o estado filtrado apresenta um certo *lag* com relação a série original, no caso, de um ano. Isso ocorre devido ao processo de filtragem utilizar um ano para prosseguir com as iterações.

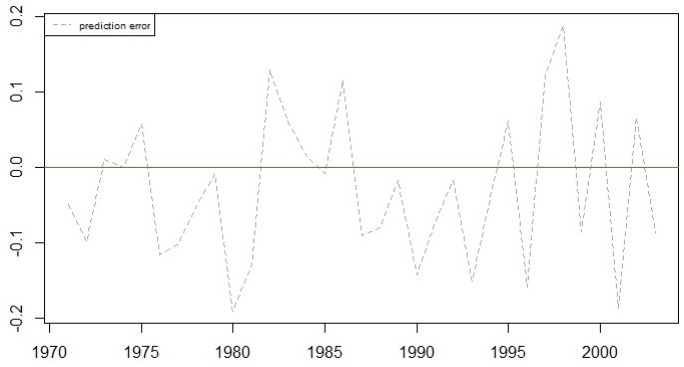

Figura 3.5: Erro da previsão um passo à frente aplicado aos dados Norwegian road traffic fatalities para o modelo de tendência local

Figura 3.6: Variância do erro da previsão um passo à frente aplicado aos dados Norwegian road traffic fatalities para o modelo de tendência local.

Podemos notar pelos gráficos das Figuras 3.5 e 3.6 que a a variância do erro de previsão vai reduzindo conforme o decorrer do tempo. Além disso, podemos dizer que isso também

ocorre para o estado filtrado convergindo para um valor constante, facilitando assim a estimação do filtro de Kalman.

3.2 Análise de diagnóstico

Para a verificação das suposições do modelo, ou seja, resíduos normalmente distribuídos, independentes e com variância constante, deve-se realizar uma análise de diagnóstico. A título de exemplo, os autores optaram por realizá-la usando o conjunto de dados UK drivers KSI.

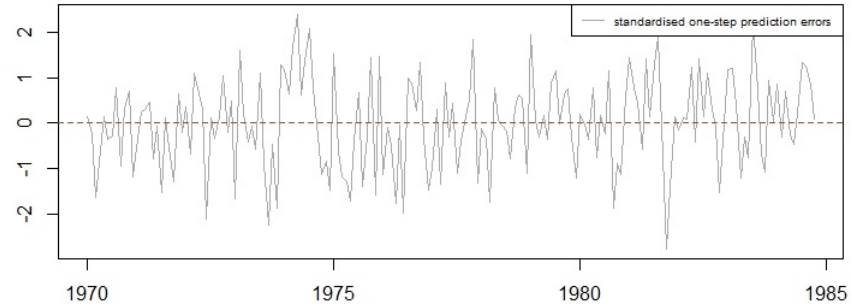

Figura 3.7: Erros de predição um passo à frente padronizados.

Pode-se observar que, majoritariamente, na Figura 3.7, os resíduos se comportam em uma faixa que compreende o eixo das ordenadas de -2 até 2, apresentando um perfil estacionário, o que inclui homocedasticidade.

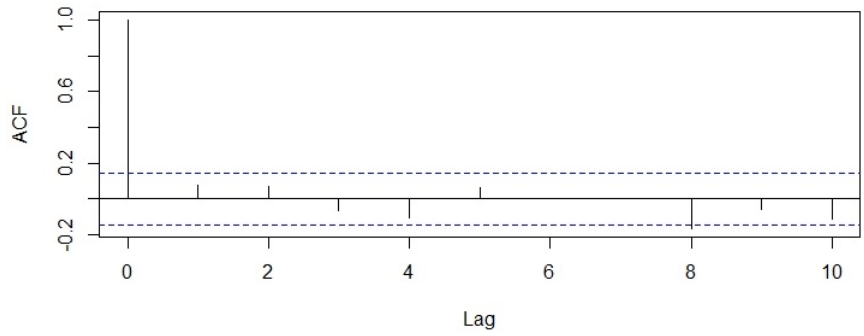

Figura 3.8: Correlograma dos erros de predição um passo à frente padronizados.

Pela Figura 3.8, nota-se que os resíduos se comportam como ruído branco, pois apresentam valores de autocorrelação baixos que não extrapolam os limites estabelecidos para *lags* maiores que 0. Agora devemos verificar a independência dos resíduos. Para isso, calcula-se a estatística do teste de Ljung-Box, que nesse caso resultou em 13.719 com p-valor de 0.1862, ou seja, não há evidências de que os resíduos não se comportem de maneira independente.

Por fim, vamos verificar se os resíduos são normalmente distribuídos.

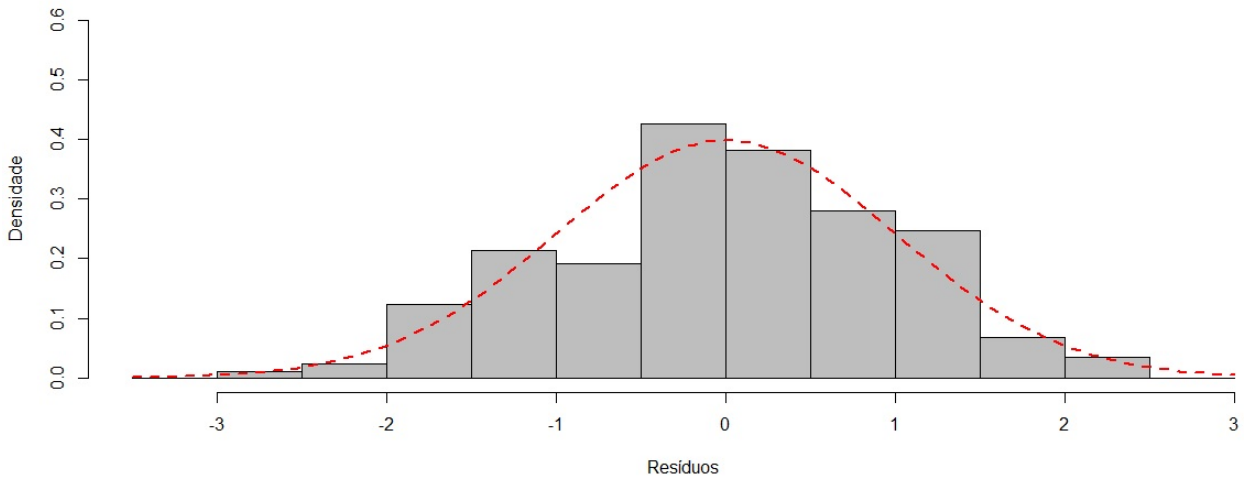

Figura 3.9: Histograma dos erros de predição um passo à frente padronizados.

Através da Figura 3.9 não fica tão evidente que os resíduos são normalmente distribuídos, logo vamos aplicar o teste de Shapiro-Wilk que verifica a aderência à distribuição Normal. Para esse exemplo, o valor da estatística resultou em $W = 0.99427$ com p-valor de 0.7231, ou seja, não há evidências de que os resíduos se comportem de forma diferente da distribuição normal.

3.3 Previsão

Para finalizar os exemplos, vamos agora nos ater a exemplos de previsão para o modelo espaço de estados. Os autores utilizam o banco de dados *Road traffic fatalities in Norway and Finland* .

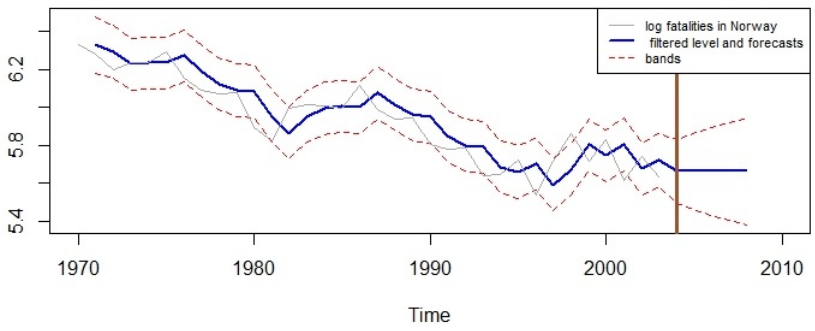

Figura 3.10: Série Log *fatalities in Norway* com intervalo de confiança, filtragem aplicada à tendência e previsão para 5 anos.

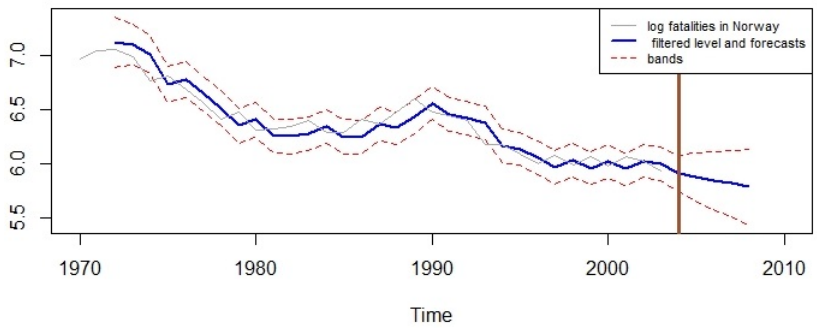

Figura 3.11: Série *fatalities in Norway* com intervalo de confiança, filtragem aplicada à tendência e previsão para 5 anos.

Pode-se observar pelas Figuras 3.10 e 3.11 que o intervalo de previsão vai aumentando para horizontes maiores de tempo, algo que era esperado, visto que a incerteza aumenta para previsões cada vez mais distantes da última observação.

Como conclusão final, salienta-se que os resultados obtidos através do método apresentado foram idênticos aos dos autores, logo podemos dizer que a modelagem empregada em *R* foi idêntica a modelagem dos autores em *Ox*. Sendo assim, prosseguiremos com as análises em *R*.

Capítulo 4

Análise dos dados

4.1 Coleta dos dados

A coleta dos dados foi realizada utilizando a técnica de *web scraping* seguindo as orientações dos autores Munzert *et al.* (2014) no site da SABESP (`http://mananciais.sabesp.com.br/Home`) através do software *R*. De maneira simples, o banco de dados foi constituído visitando o site diversas vezes para coletar os dados diários do dia *01/01/2003* até *01/04/2020*.

4.2 Tratamento inicial dos dados

Com hipótese baseada em senso comum, temos que a precipitação sofre influência das estações do ano, ou seja, é sazonal e para verificar essa hipótese, realizaremos uma análise descritiva.

Para decompormos a série temporal, iremos lidar com os anos bissextos que podem acarretar leves irregularidades, além de ter que utilizar frequência anual de 365.25 ao invés de 365. Sendo assim, ao invés de utilizar a frequência de 365.5, iremos retirar o dia 29 de fevereiro da análise, que ocorreu cinco vezes no nosso período de estudo, 2004, 2008, 2012, 2016 e 2020.

Com a finalidade de não descartar a informação desses dias, consideramos a informação do dia primeiro de março dos anos bissextos como a média entre as observações de 29 de fevereiro e primeiro de março. No caso, estamos executando o procedimento para apenas 5 observações, o que corresponde a cerca de 0,08% de alteração no banco de dados.

4.3 Análise descritiva

A seguir será apresentada a precipitação diária na Figura 4.1. Ressalta-se que, por se tratar de observações diárias, as observações poderiam se sobrepor e, para contornar esse problema, utilizou-se a cor azul com uma tonalidade mais clara e as observações evidenciadas por pontos e linhas.

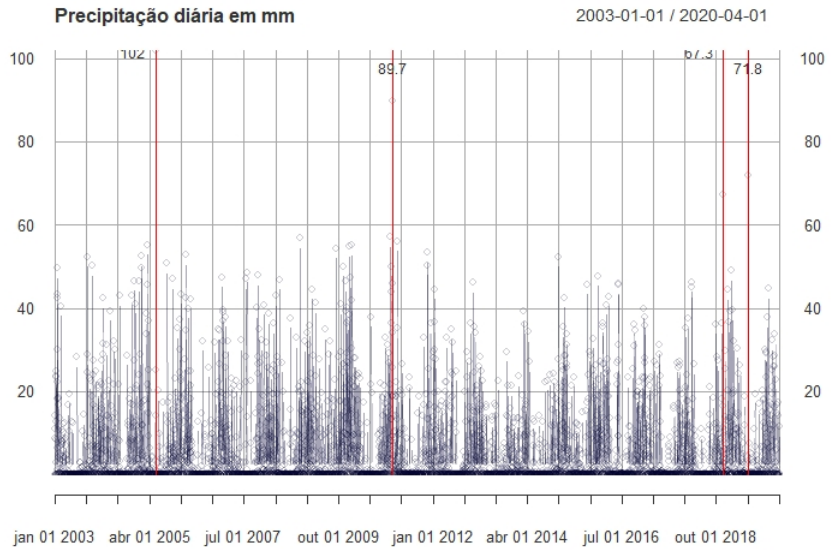

Figura 4.1: Série diária de precipitação em mm.

Nota-se que temos picos acima de 60 mm de precipitação. Em *25/05/2005* temos o maior pico, com 102 mm de precipitação, seguido por *01/11/2011* com 89.7 mm, *05/07/2019* com 71.8 mm e, por fim, *01/12/2018* com 67.3 mm.

Podemos notar também que há uma grande variação e isso é devido à característica dos dados, ou seja, é comum ter dias sem chuva e dias chuvosos. Mais adiante iremos tratar essa variação através de uma transformação.

Como não está evidente que haja alguma tendência ou sazonalidade, vamos observar os dados agrupados por semanas e meses. Os dados semanais são apresentados na Figura 4.2.

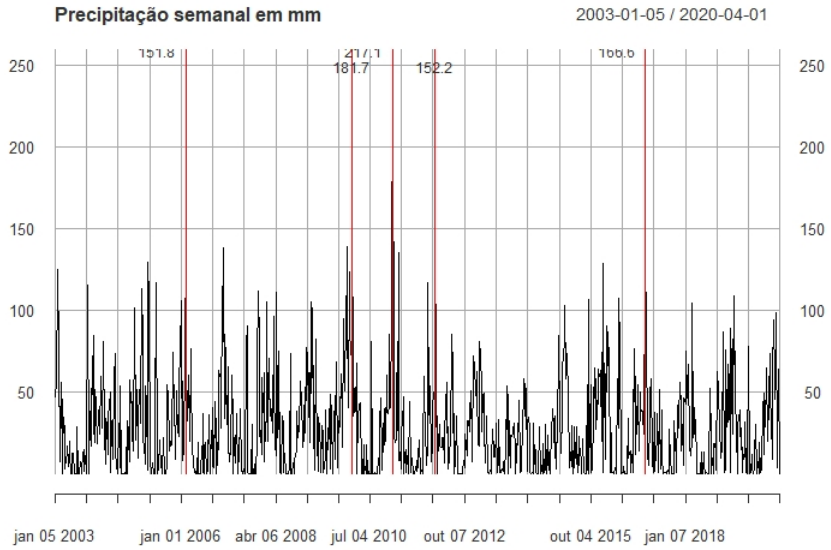

Figura 4.2: Série semanal de precipitação em mm.

Através da Figura 4.2, pode-se notar 5 picos acima de 150mm, *16/01/2011* com 217.1 mm, seguido de *31/01/2010* com 181.7 mm, *22/01/2017* com 166.6 mm, *22/01/2012* com 152.2 mm, *12/02/2006* com 151.8 mm. Podemos salientar ainda que já é possível notar

que a sazonalidade se torna cada vez mais evidente, mas ainda não parece ter algum tipo de tendência.

Vamos acumular os dados diários em meses para tentar evidenciar mais ainda a sazonalidade e tentar verificar a existência de alguma tendência.

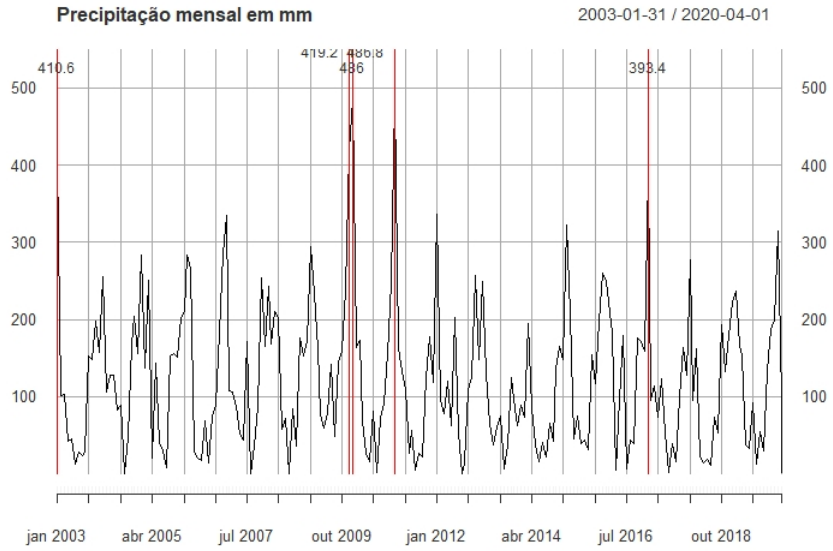

Figura 4.3: Série mensal de precipitação em mm.

Na Figura 4.3, os maiores picos acima de 350 mm na respectiva ordem ocorrem em *Jan/2011* com 486.8 mm, *Jan/2010* com 486.0 mm, *Dez/2009* com 419.2 mm, *Jan/2003* com 410.6mm, *Jan/2017* com 393.4mm. Salienta-se que esse tipo de comportamento foi observado anteriormente nos meses correspondentes ao verão, com uma precipitação maior, e nos meses de inverno, com precipitação menor. Ainda é possível dizer que há indícios de sazonalidade, sustentando a hipótese baseada no senso comum.

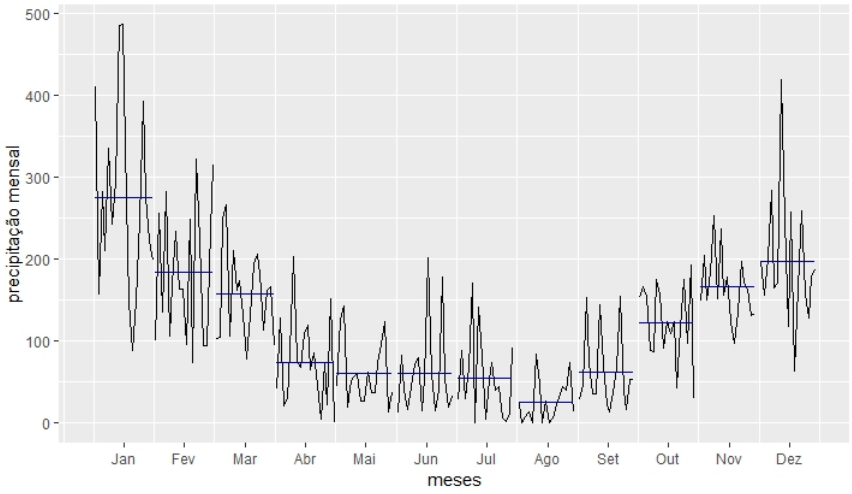

Figura 4.4: Subsérie mensal de precipitação em mm.

Na Figura 4.4, para cada mês, temos uma mini série temporal que relaciona todas as observações referentes a cada mês de todos os anos e, a barra azul representa a média de cada mês. Pode-se observar um formato côncavo o que deixa evidente, mais uma vez, a sazonalidade da série, e que as estações tem influência sobre a chuva.

4.4 Variabilidade dos dados

Vimos na seção anterior que nossos dados possuem grande variabilidade e, caso sejam mantidos assim, vão impactar na previsão. Logo, vamos aplicar a transformação de Box e Cox (1964) para estabilizar a variância que está descrita a seguir

$$Y_i(\lambda) = \begin{cases} \ln(X_i)\ , & se\ \lambda = 0 \\ \frac{X_i^{\lambda}-1}{\lambda}, & se\ \lambda \neq 0 \end{cases} .$$

Escolheu-se $\lambda = 0.12$ após algumas interações seguindo a regra do trapézio. Como confirmação de que o λ é adequado, iremos realizar o gráfico de média por amplitude e vamos verificar se o coeficiente de inclinação da reta é significante a 5%.

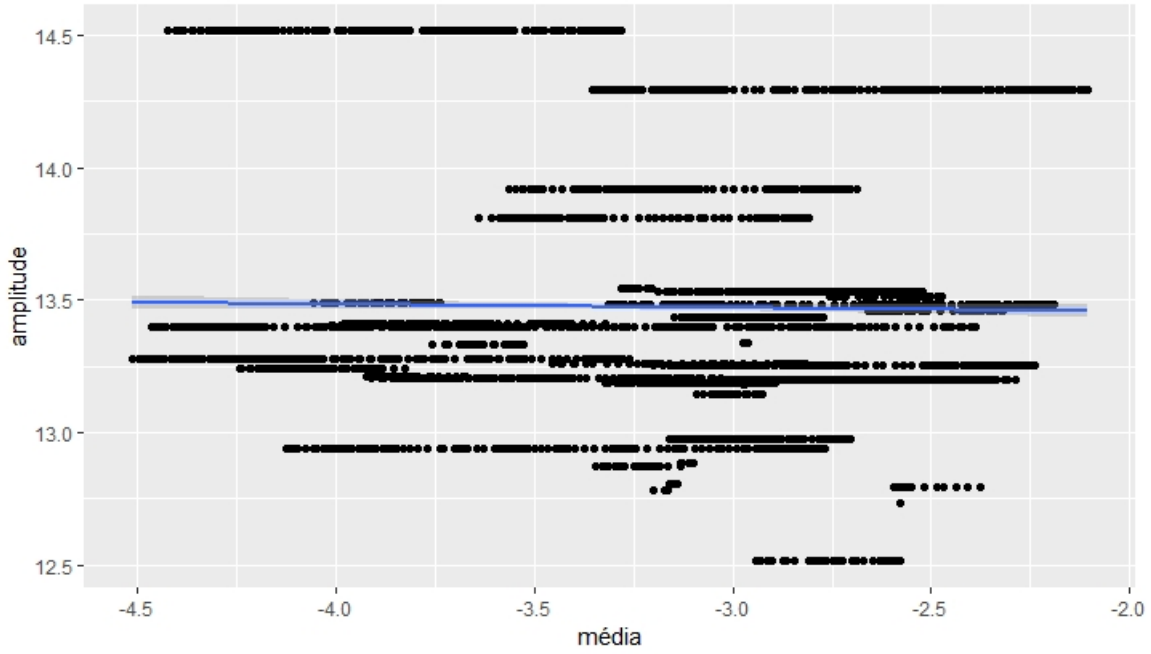

Figura 4.5: Gráfico de média × amplitude.

A partir da Figura 4.5 vemos que a reta está praticamente sem inclinação e, além disso, o p-valor referente ao coeficiente angular da reta é 0.17, sendo não significativo ao nível de significância de 5%, ou seja, não há evidências de que o coeficiente angular da reta seja diferente de 0. Sendo assim, podemos afirmar que a transformação reduziu a variabilidade dos dados.

4.5 Dados transformados

Aplicamos a transformação nos dados e vamos observar novamente o gráfico da série diária, semanal e mensal.

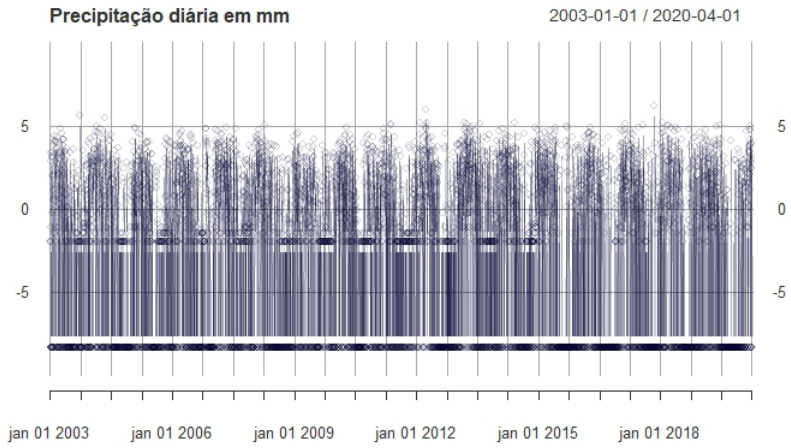

Figura 4.6: Série transformada diária de precipitação em mm.

Como podemos ver, a escala já reduziu drasticamente. Continuamos com a sazonalidade compondo a série e não temos evidências de tendências.

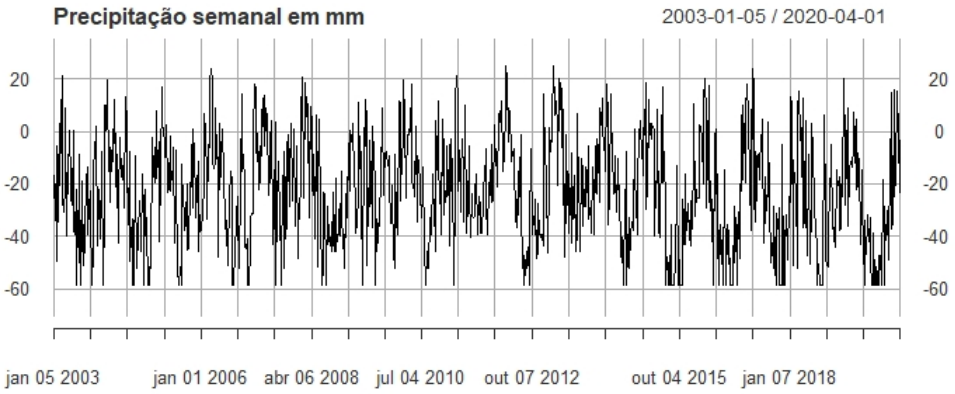

Figura 4.7: Série transformada semanal de precipitação em mm.

Na precipitação semanal, podemos notar novamente que possui sazonalidade, mas aparentemente a tendência não é presente.

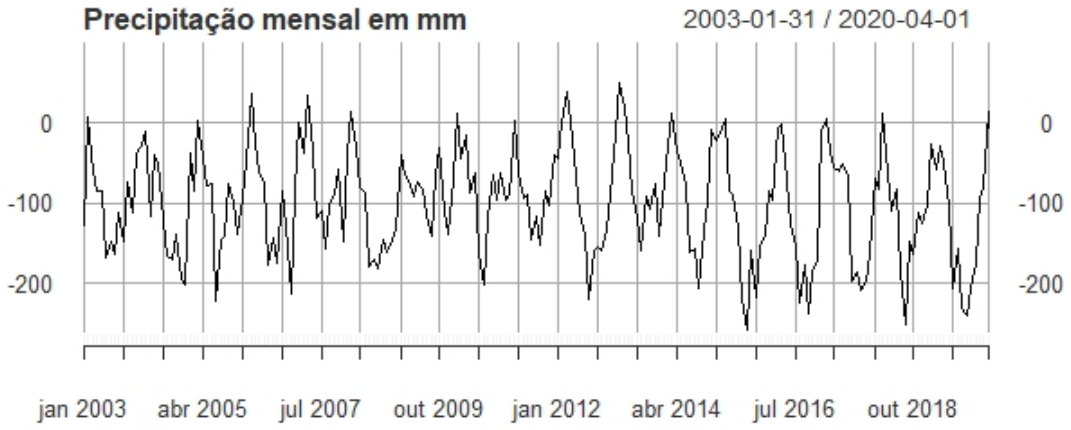

Figura 4.8: Série transformada mensal de precipitação em mm.

Por fim, a série mensal mostra as mesmas informações, com verões mais chuvosos e invernos mais secos.

Para verificar a hipótese de que a série não possui tendência, realizamos o teste Dickey-Fuller e Phillips-Perron com defasagem de 33 e com nível de significância de 1%, devido a termos muitos dados disponíveis.

A estatística de Dickey-Fuller resultou em -7.6979, o que confere um p-valor inferior a 1%. Já a estatística do teste de Phillips-Perron resultou em -7536.1, conferindo também um p-valor inferior a 1%. Com isso, ambos os testes apontam para a rejeição da hipótese nula, ou seja, podemos concluir que a série não tem tendência.

Por fim vamos olhar os gráficos de autocorrelação e autocorrelação parcial.

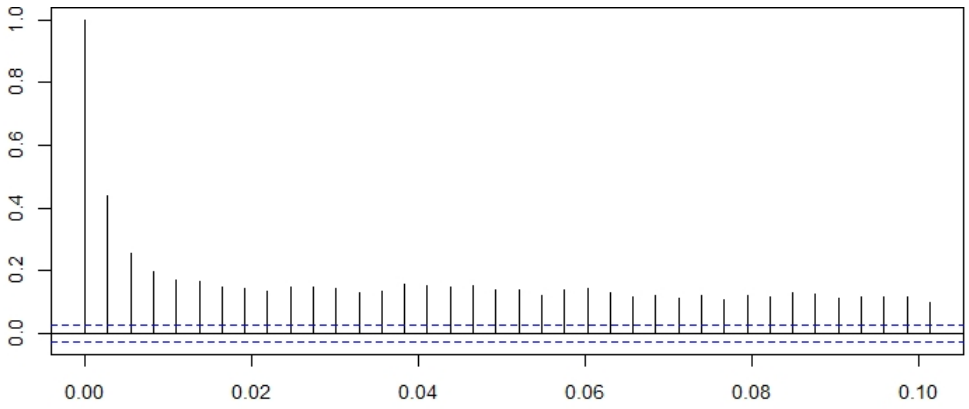

Figura 4.9: Função de autocorrelação com *lag* 37.

Pela Figura 4.9 podemos verificar picos levemente amortecidos com decaimento muito suave a partir do *lag* 3, aparentando ser uma estrutura em forma hiperbólica de longa duração levemente amortecida. Vamos tomar um *lag* maior para verificarmos se possui sazonalidade.

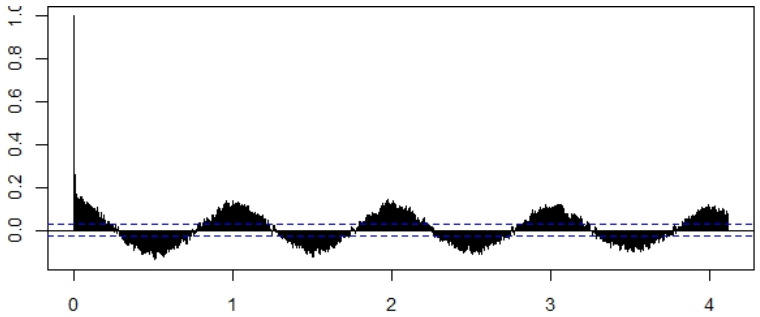

Figura 4.10: Função de autocorrelação com *lag* 1500 com eixo das abscissas em anos.

Pela figura 4.10 é possível notar que existe de fato uma estrutura de longa duração levemente amortecida e podemos observar que existe sazonalidade, por possuir valores altos de correlações em *lags* altos e possuir comportamento cíclico.

Capítulo 5

Análise dos dados

5.1 Ajuste dos modelos

Na literatura é comum obsevarmos a comparação do modelo espaço de estados com o clássico modelo ARIMA. No artigo de Proietti (2000) o autor faz a comparação do modelo estrutural básico com alguns desses modelos clássicos, e tomando por base esse artigo, vamos comparar o ajuste do nosso modelo com esses modelos para verificarmos qual é o mais adequado para fins de previsão considerando essa estrutura de dados.

5.1.1 Modelo Espaço de Estados

Petrone e Petris (2011) apresentam em seu artigo uma melhor elucidação do uso do pacote *dlm* do software *R*. Contudo, Petris (2008) comenta que o *SructTS()* da base do software *R*, é útil para analisarmos séries temporais univariadas. Outro aspecto importante é a agilidade que o *SructTS()* ajusta o modelo, com predições idênticas ao *dlm*, que por sua vez é um pouco mais complicado de se ajustar. Sendo assim, optou-se por ajustar o modelo pelo o *SructTS()*.

Como parâmetros dessa função passamos a nossa série transformada de precipitação mensal, o tipo *"BSM"* que significa *Basic Structural Model* por decompor em tendência linear local e sazonalidade, por fim o vetor de início para utilizar no método L-BFGS-B.

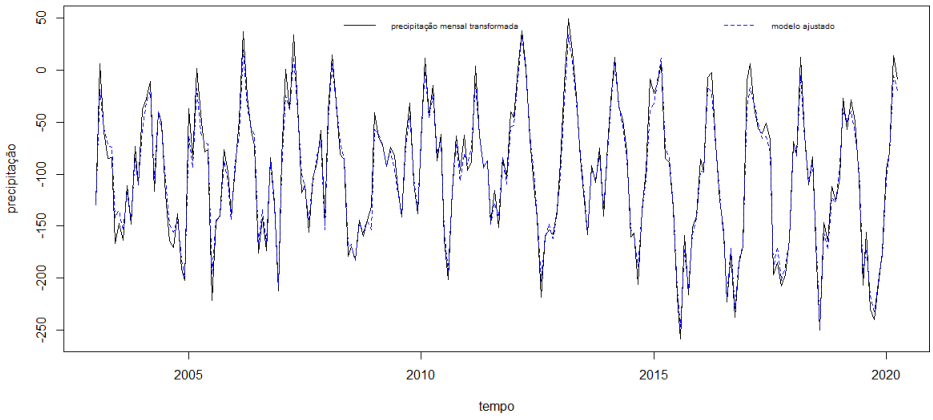

Figura 5.1: modelo ajustado a serie de precipitação mensal transformada

A partir da Figura 5.1 podemos observar que o modelo se ajusta bem aos dados com

pequenas divergências entre o ajustado e a série transformada.

Análise diagnóstico do modelo

Vemos que o modelo se ajustou bem aos dados, mas agora temos que verificar se a estrutura de longa duração permaneceu, se os resíduos são normais e a variabilidade permanece homogênea.

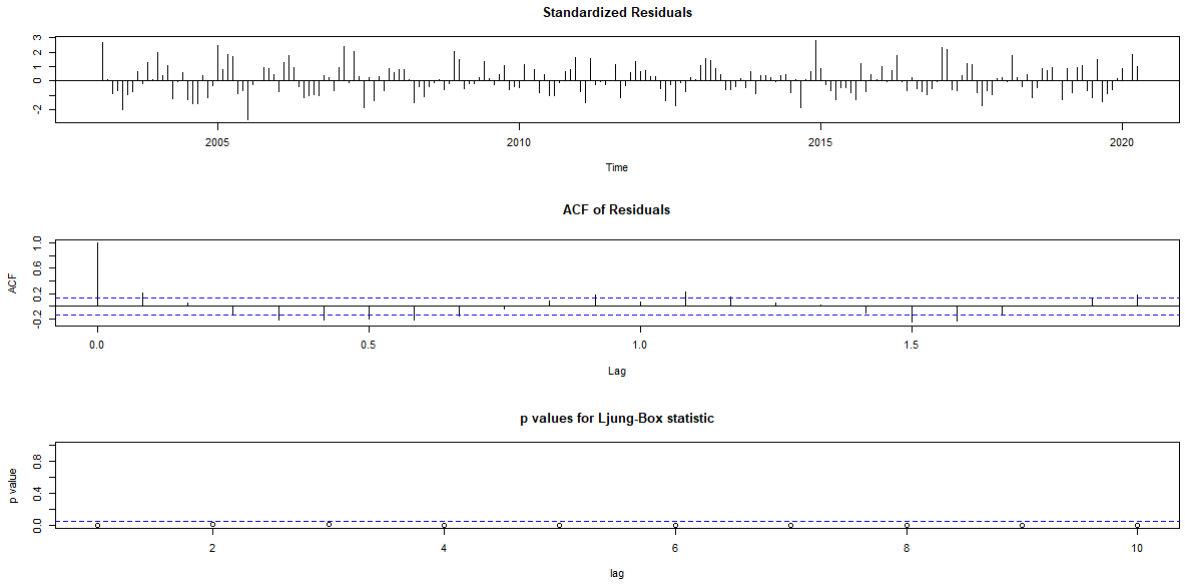

Figura 5.2: Gráficos de análise diagnóstico.

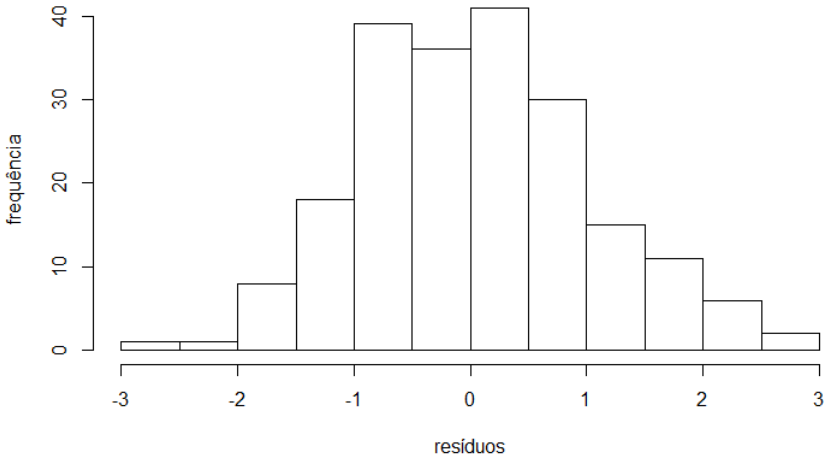

Figura 5.3: Histograma dos resíduos.

Pela Figura 5.2 podemos observar pelos resíduos padronizados e pelo gráfico da função de autocorrelação que a estrutura de longa duração permanece. O teste de Ljung box, rejeita a hipótese de que os resíduos são independentes e identicamente distribuídos para todos os *lag* apresentados.

Continuando a análise dos resíduos, pela Figura 5.3 temos a indicação de os resíduos são normais. Para termos uma confirmação, foi realizado o teste de Shapiro Wilk que

resultou em um p-valor de 0.3858. Considerando um nível de significância de 10%, como p-valor é maior que o nível de significância, não rejeitamos a hipótese de que os resíduos sejam normalmente distribuídos.

5.1.2 Ajuste do modelo ARIMA

Após ajustar diversos modelos, o que melhor se adaptou aos dados foi o ARIMA(2,1,0). Vamos ver então o ajuste e a análise diagnóstico do modelo

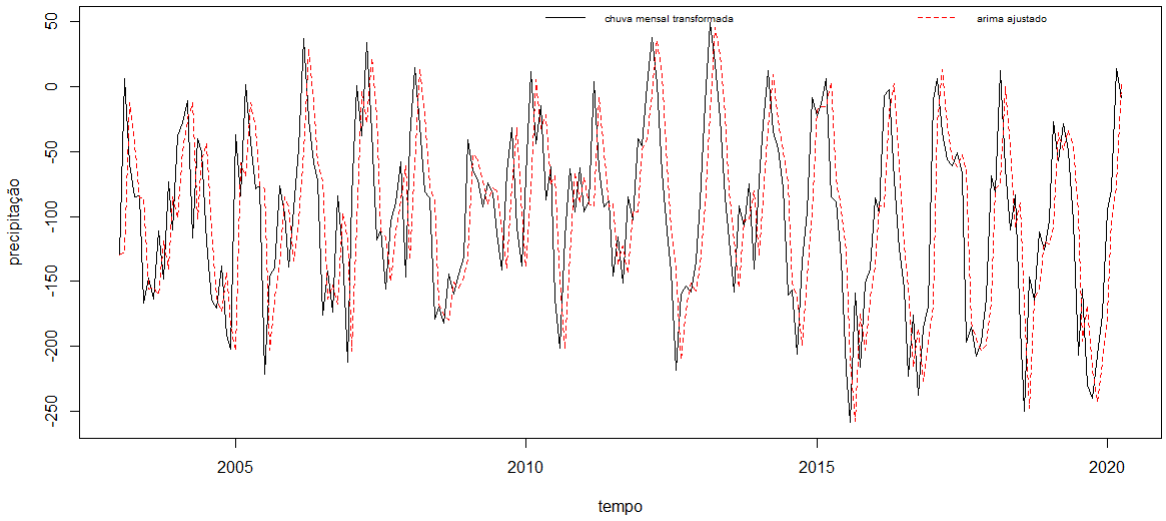

Figura 5.4: ARIMA(2,1,0) ajustado.

Nota-se pela Figura 5.4 que os valores ajustados pelo modelo apresentam um comportamento similar ao da série temporal com uma defasagem pequena.

Análise dignóstico ARIMA(2,1,0)

Uma vez que vimos que o modelo está bem ajustado a serie, vamos executar o diagnóstico do modelo.

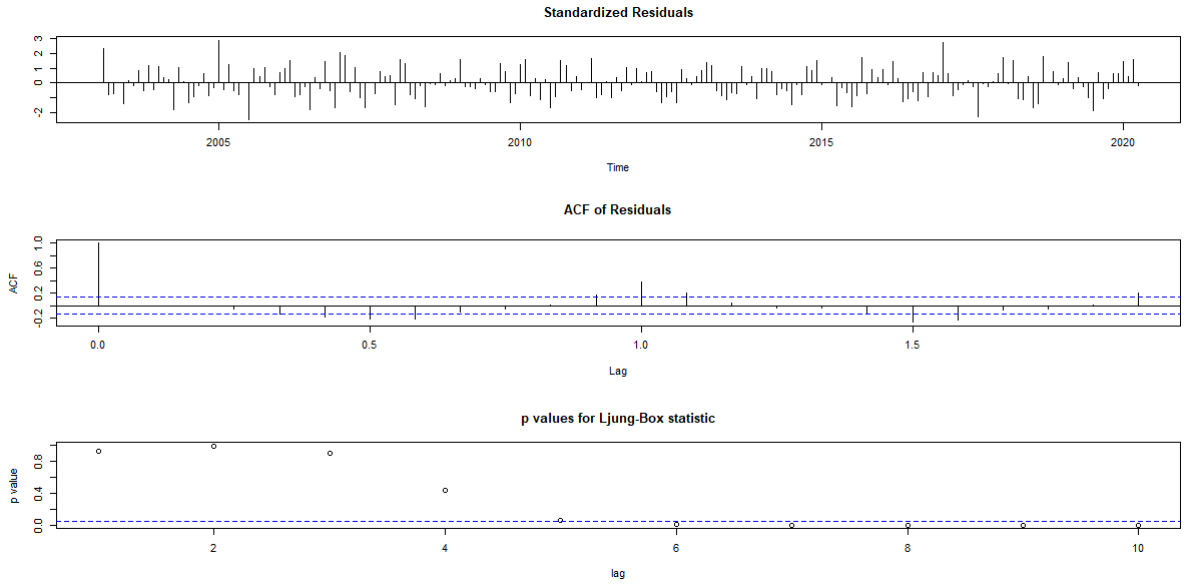

Figura 5.5: Análise diagnóstico para o ARIMA.

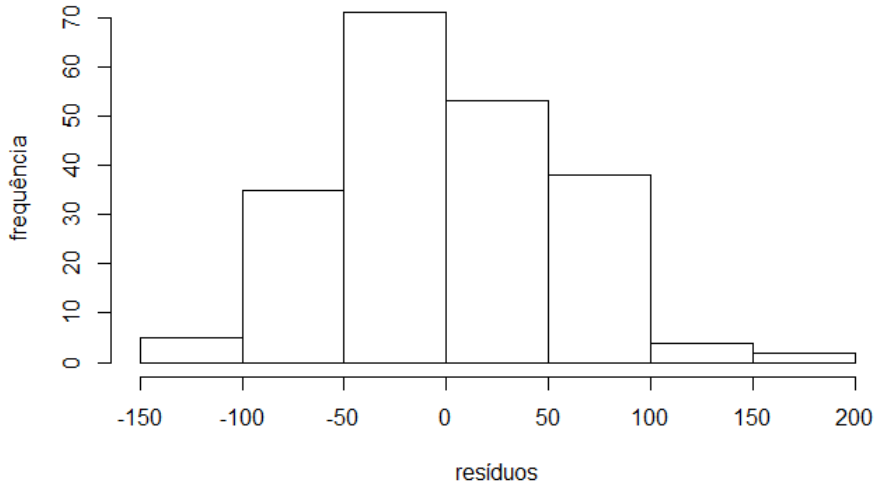

Figura 5.6: Histograma dos resíduos modelo ARIMA.

Nota-se pela Figura 5.5 que da mesma forma do modelo estrutural básico, temos uma estrutura de longa duração. Por outro lado, o teste de Ljung-Box até o quarto *lag*, não rejeita a hipótese de que os resíduos não sejam independentes e identicamente distribuídos, com *lags* maiores que quatro passamos a rejeitar a hipótese.

Vemos ainda que o histograma da Figura 5.6 é aproximadamente normal. Para confirmarmos essa evidência, realizou-se o teste de Shapiro-Wilk que resultou em um p-valor de 0.419. A interpretação é análoga do modelo estrutural básico, ou seja, podemos dizer que os residuos se aderem à distribuição Normal.

5.2 Validação cruzada

Com a intenção de avaliar qual o melhor modelo para fazermos previsão, vamos utilizar a metodologia de validação cruzada. Como temos dados temporais, ao usar os modelos usuais de validação cruzada, teremos o problema da invariância no tempo. Para contornar esse problema, iremos adotar uma variação do *K-fold*, que é conhecido como *nested cross-validation*.

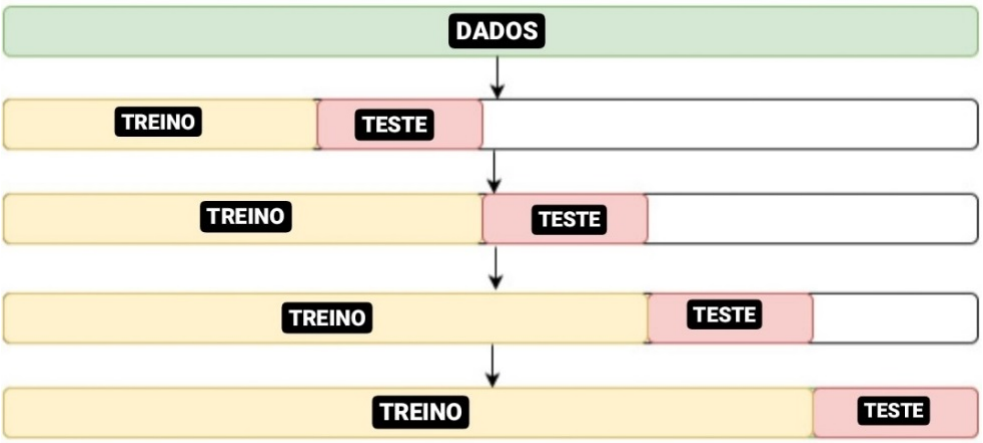

Figura 5.7: Desenho esquemático do *nested cross-validation*.

Na Figura 5.7 podemos ver um desenho esquemático de como funciona a metodologia.

Primeiramente pegamos uma parte dos nossos dados para treinar nosso modelo, prevemos o conjunto teste, calculamos uma medida de erro e, por fim, incorporamos o conjunto teste no conjunto de treino, dando início a mais uma iteração, fazendo isso até que o bloco de teste chegue ao final da série. No nosso conjunto de dados, dividimos em 6 partes, sendo as cinco primeiras com 36 observações e a última com 28 observações, e atribuímos essas partes para o treinamento e teste conforme as iterações do método. A escolha de 36 observações para treino, e realizar a previsão para as próximas 36 observações é para verificar qual modelo seria mais adequado para fazer as previsões de 2012 até 2015, citadas previamente no relatório da Agência Nacional de Águas (2014).

Como medida de erro, optou-se pela raiz quadrada do erro médio (do inglês, *Root-mean-square error (RMSE)*), pois nossos dados possuem zeros e calcular medidas como erro percentual absoluto médio poderia causar o clássico problema da divisão por zero.

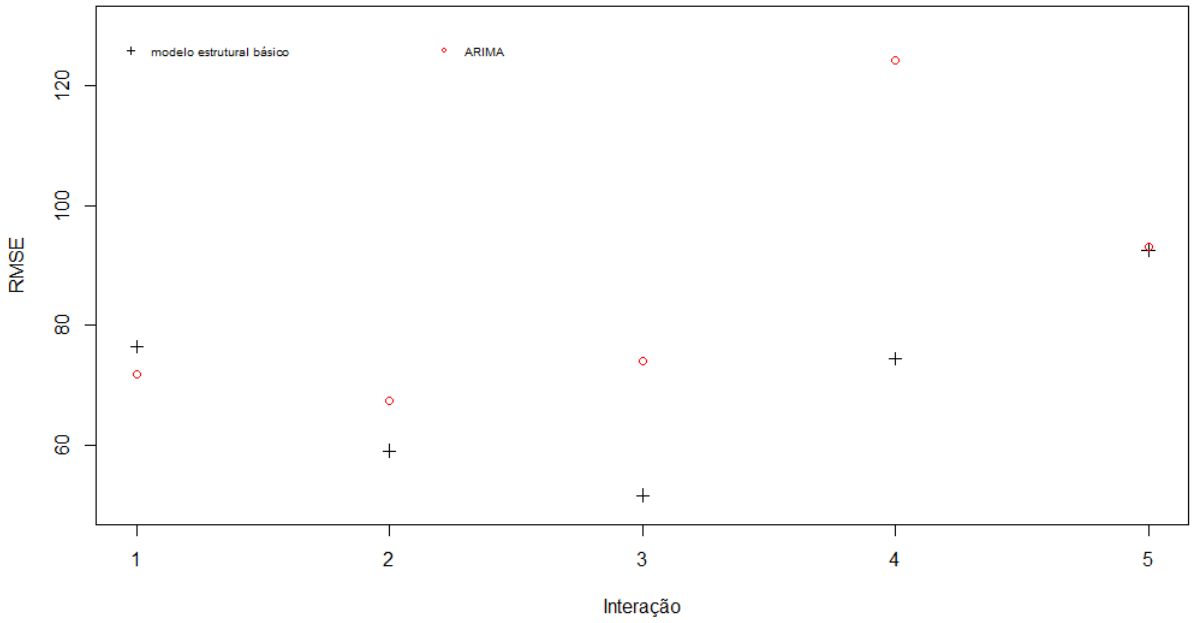

Figura 5.8: Gráfico dispersão RMSE para os modelos.

Após todo o procedimento, fazemos um gráfico de dispersão do *RMSE* para avaliarmos comparativamente como vão se comportar os modelos. Na Figura 5.8 vemos que, no geral, o modelo estrutural básico apresenta uma medida menor, sendo assim, vamos considerar que esse é o modelo mais adequado.

Capítulo 6

Previsão monitoramento da seca

Sabendo agora qual modelo produziu menores medidas de erro, vamos nos ater a realizar a previsão para os anos de 2013 a 2015 a partir dos dados históricos de 2003 até 2012, com a intenção de verificar se para esse período a série extrapola o intervalo de predição.Este período de tempo foi escolhido pois compreende um ano antes da estiagem e um ano após.

Vale ressaltar que estaremos prevendo muitos passos a frente, podendo causar grandes distorções no intervalo de predição. Logo optou-se por adotar um nível de confiança de 80% para evitar o aparecimento de muitos pontos fora de controle. Além disso, consideraremos que um ponto saiu do controle, quando ele extrapolar o intervalo de confiança de 80%, baseado na distribuição normal. Por fim, salienta-se que o modo como o intervalo de confiança foi colocado na Figura 6.1 é apenas para fins visuais.

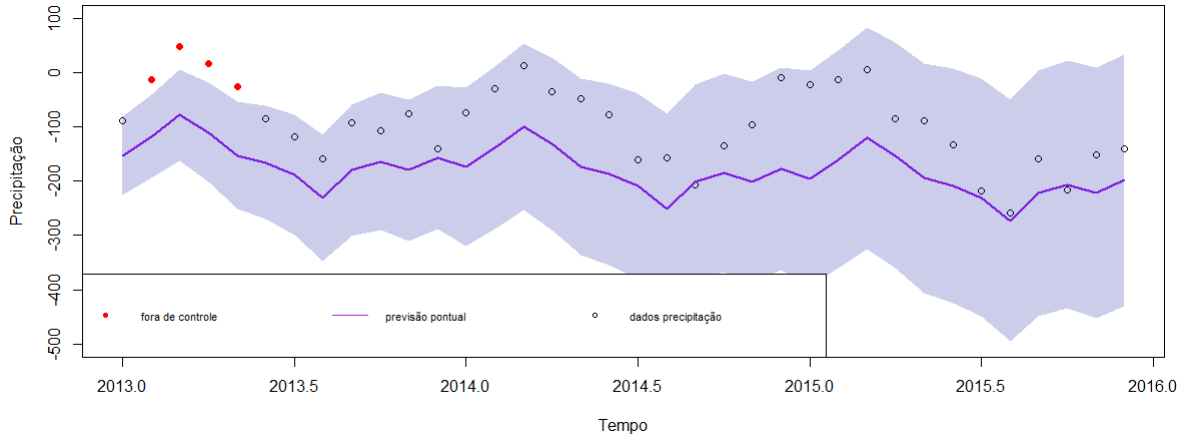

Figura 6.1: Gráfico de controle.

Podemos ver pela Figura 6.1 que temos os meses de fevereiro, março, abril e maio de 2013 fora de controle, indicando assim uma possível alteração no regime de chuvas. Porém, devido a estarmos realizando previsões muitos passos à frente, quanto maior o tempo elas se tornam menos acuradas causando um impacto nos intervalos de predição e, por consequência, nas previsões fora de controle.

Capítulo 7

Conclusão da análise dos dados e ajuste do modelo

No desenvolvimento do trabalho vimos que dados pluviométricos possuem uma certa dificuldade para se modelar devido à natureza dos dados que originam grandes períodos de zeros, ocasionando assim uma variabilidade grande. Além disso, vimos também que possuem a característica de serem cíclicos.

Para lidar com este problema, buscamos por uma transformação que fosse capaz de reduzir e estabilizar essa variabilidade para produzir previsões mais acuradas. Após estabilizada a variância, conseguimos perceber uma estrutura de longa duração com sazonalidade e comportamento cíclico presente nos dados.

Vimos, também, que o modelo estrutural básico se adequou bem aos dados e, no comparativo através da validação cruzada com o modelo ARIMA, o modelo estrutural básico apresentou menores valores de erro no geral. Após a escolha do modelo com menor erro, realizamos a previsão para os anos de 2013, 2014 e 2015, estabelecemos um intervalo de predição e vimos quais pontos da nossa série de precipitação saíram do intervalo estabelecido.

Podemos concluir que, apesar da variabilidade e presença de estrutura de longa duração, o modelo foi capaz de identificar uma possível alteração no regime pluviométrico do Sistema Cantareira logo no início da previsão, porém vemos que para os passos seguintes o modelo não foi capaz de identificar pontos fora de controle.

Capítulo 8

Conclusão do trabalho e trabalhos futuros

Podemos dizer que o trabalho alcançou seu objetivo primário de detectar se houve alteração no regime pluviométrico no Sistema Cantareira através da abordagem clássica do modelo espaço de estados.

Além do objetivo primário, alcançamos o objetivo pedagógico de ter uma visão geral sobre modelo espaço de estados e suas aplicações com enfoque no software *R*.

Como sugestão para trabalhos futuros, recomendamos buscar por modelos que incorporem estruturas de longa duração. Além disso, recomendamos o estudo das mais variadas formas de validação cruzada, assim como o impacto das medidas utilizadas.

Por fim, seria interessante adotar o ponto de vista multivariado, considerando fatores como temperatura, umidade relativa do ar, precipitação etc.

Referências Bibliográficas

Agência Nacional de Águas (2014). Encarte especial sobre a crise hídrica - Brasília: ANA, 2015. Disponível em: http://www.snirh.gov.br/portal/snirh/centrais-de-conteudos/conjuntura-dos-recursos-hidricos/informes2014.pdf. Acesso em: 03 de setembro de 2019.

Aiube, F. A. L. (2005). Modelagem dos preços futuros de commodities: abordagem pelo filtro de partículas. Tese (Doutorado em Engenharia de Produção) - Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, Brasil).

Alencar, A. P. (2006). Modelo espaço de estados com mudança markoviana de regime e probabilidades de transição modeladas com ondaletas. Tese (Doutorado em Estatística) - Universidade de São Paulo, São Paulo.

Box, G. E. e Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, **26**(2), 211–243.

Box, G. E. e Jenkins, G. M. (1976). *Time series analysis: Forecasting and Control*. Holden Day, San Francisco, second edition.

Casella, G. e Berger, R. L. (2002). Statistical inference. Pacific Grove, CA: Duxbury.

Commandeur, J. J. e Koopman, S. J. (2007). *An introduction to state space time series analysis*. Oxford University Press, New York.

Doornik, J. A. (2009). An object-oriented matrix programming language ox 6.

Durbin, J. e Koopman, S. (2001). *Time Series Analysis by State Space Methods*. Oxford University Press, Oxford.

Franco, G. C., Gamerman, D. e Santos, T. (2009). Modelos de espaço de estados: abordagens clássica e bayesiana.Dissertação (Mestrado em Ciências) - Universidade Federal de Minas Gerais, Belo Horizonte.

Freitas, E. (2019). Clima brasileiro. *Disponível em: ⟨https:// brasilescola.uol.com.br/ brasil/ clima-brasileiro.htm⟩*. Acesso em: 25 de novembro de 2019.

Hamilton, J. (1994). Time series analysis. v. 10. Princeton University Press.

Harrison, P. e Stevens, C. (1971). A bayesian approach to short-term forecasting. *Journal of the Operational Research Society*. v.22, n.4 , p. 341-362.

Harrison, P. J. e Stevens, C. F. (1976). Bayesian forecasting. *Journal of the Royal Statistical Society: Series B (Methodological)*. v.38, n.3, p. 205-228.

Harvey, A. C. (1989). Forecasting, structural time series models and the kalman filter. Cambridge University Press, Cambridge.

Harvey, A. C. (1990). Forecasting, structural time series models and the kalman filter.

Harvey, A. C. (1993). Time series models. *MIT press*.

Harvey, A. C. e Fernandes, C. (1989). Time series models for count or qualitative observations. *Journal of Business & Economic Statistics*. v.7, n.4, p.407-417.

Holt, C. (1957). Forecasting trends and seasonal by exponentially weighted moving averages. *ONR Memorandum*. v.52.

Jazwinski, A. H. (1970). Stochastic processes and filtering theory. Academic Press, Cambridge.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of ASME: Series D , Journal of Basic Engineering*. v.82, n.1, p.35-45.

Kalman, R. E. e Bucy, R. S. (1961). New results in linear filtering and prediction theory. v. 83, n. 1, p. 95-108, Journal of Fluids Engineering, American Society of Mechanical Engineers.

Koopman, S. J., Shephard, N. e Doornik, J. A. (1999). Statistical algorithms for models in state space using ssfpack 2.2. *The Econometrics Journal*. v.2, n.1, p.107-160.

Migon, H. S. e Gamerman, D. (1999). Statistical inference: an integrated approach. London: Arnold.

Munzert, S., Rubba, C., Meißner, P. e Nyhuis, D. (2014). *Automated data collection with R: A practical guide to web scraping and text mining*. John Wiley & Sons.

Muth, J. F. (1960). Optimal properties of exponentially weighted forecasts. *Journal of the American Statistical Association*. v.55, n.290, p.299-306.

Nerlove, M., Wage, S. *et al.* (1964). On the optimality of adaptive forecasting. *Management Science*. v.10, n.2, p.207-224.

Petris, G. (2008). dlm: Mle and bayesian analysis of dynamic linear models. *Aplicativo disponível em: http:// cran.r-project.org/ web/ packages/ dlm/ index.html*.

Petrone, S. e Petris, G. (2011). State space models in R. *Journal of Statistical Software. Disponível em: https:// www.jstatsoft.org/ article/ view/ v041i04* , **41**.

Proietti, T. (2000). Comparing seasonal components for structural time series models. International Journal of Forecasting, v.16. n.2. p.247-260. Elsevier.

Puga, B. P. (2018). Governança dos recursos hídricos e eventos climáticos extremos: a crise hídrica de são paulo. Tese (Doutorado em Desenvolvimento Econômico) - Universidade Estadual de Campinas, Campinas.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Radhakrishna (2013). An introduction to state space time series analysis summary. *Publicação disponível em: https:// radhakrishna.typepad.com/ book-summary-and-r-code-1. pdf*.

Shumway, R. e Stoffer, D. (2000). Time series analysis and its application. New York: Springer-Verlag.

Theil, H. e Wage, S. (1964). Some observations on adaptive forecasting. *Management Science*. v.10, n.2, p.198-206.

Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Management science*. v.6, n.3, p. 324-342.

Apêndice A

Códigos em R utilizados para os exemplos

```
# Notação utilizada p/ o modelo espaço estado
# Y_t = F_t (Teta)_t + vt; v~N(0; V_t)
# (Teta)_t = G_t (Teta)_{t1} + wt; wt~N(0;W_t)

library(dlm)

# leitura dos dados
data.1 <- log(read.table("UKdriversKSI.txt",skip=1))
colnames(data.1) <- "logKSI"
data.1 <- ts(data.1, start = c(1969),frequency=12)
intervention <- rep(1,dim(data.1)[1])
intervention[1:169]<- 0
intervention <- ts(intervention, start = c(1969),frequency=12)
prices <- (read.table("logUKpetrolprice.txt",skip=1))
prices <- ts(prices, start = c(1969),frequency=12)
data.2 <- log(read.table("NorwayFinland.txt",skip=1))
data.2 <- data.2[,2,drop=F]
colnames(data.2) <- "logNorFatalities"
data.2 <- ts(data.2 , start = c(1970,1))
data.3 <- log(read.table("NorwayFinland.txt",skip=1))
data.3 <- data.3[,3,drop=F]
colnames(data.3) <- "logFinFatalities"
data.3 <- ts(data.3 , start = c(1970,1))

#intervalo de confiança

fn <- function(params){
  mod <- dlmModPoly(order = 1, dV =exp(params[1]), dW = exp(params[2])) +
    dlmModSeas(frequency = 12,dV =exp(params[1]) , dW= rep(0,11))
  return(mod)
}
fit <- dlmMLE(data.1, rep(0,2),fn)
mod <- fn(fit$par)
filtered <- dlmFilter(data.1,mod)
```

```
smoothed <- dlmSmooth(filtered)
cov <- dlmSvd2var(smoothed$U.S, smoothed$D.S)
lev.var <- sapply(cov, function(x){x[1,1]})
mu <- ((smoothed$s)[,1])[-1]
nu <- ((smoothed$s)[,2])[-1]
res <- residuals(smoothed,sd=F)
lev.ts <- ts(lev.var[-1],start = 1969, frequency =12)
wid <- qnorm(0.05, lower = FALSE) *sqrt(lev.ts)
temp <- cbind(mu, mu + wid %o% c(-1, 1))
temp <- ts(temp,start = 1969,frequency =12)
par(mfrow=c(1,1))
plot(lev.ts,xlab="",ylab = "level estimation variance")
par(mfrow=c(1,1))
plot(temp, plot.type = "s", type = "l",lty = c(1, 5, 5),
ylab = "Level", xlab = "", ylim = range(data.1),col=c("blue","red","red"),lwd=2)
lines(data.1, type = "o", col = "darkgrey")
legend("topright",
         leg = c("log UK drivers KSI"," stochastic level +/- 1.64SE"),
         cex = 0.7, lty = c(1, 5),col = c("darkgrey","red"),
         bty = "y", horiz = F)
nu.var <- sapply(cov, function(x){x[2,2]})
nu.var.ts <- ts(nu.var[-1],start = 1969, frequency =12)
wid <- qnorm(0.05, lower = FALSE) *sqrt(nu.var.ts)
temp <- cbind(nu, nu + wid %o% c(-1, 1))
temp <- ts(temp,start = 1969,frequency =12)
par(mfrow=c(1,1))
plot(temp, plot.type = "s", type = "l",lty = c(1, 5, 5),
       ylab = "Level", xlab = "",col=c("blue","red","red"),lwd=1)
legend("topright",
         leg = "deterministic level +/- 1.64SE",
         cex = 0.7, lty = c(5),col = c("red"),
         bty = "y", horiz = F)
wid1 <- qnorm(0.05, lower = FALSE) *sqrt(lev.ts)
temp1 <- cbind(mu, mu + wid %o% c(-1, 1))
temp1 <- ts(temp1,start = 1969,frequency =12)
wid2 <- qnorm(0.05, lower = FALSE) *sqrt(nu.var.ts)
temp2 <- cbind(nu, nu + wid %o% c(-1, 1))
temp2 <- ts(temp2,start = 1969,frequency =12)
temp3 <- temp1+temp2
par(mfrow=c(1,1))
plot(temp3, plot.type = "s", type = "l",lty = c(1, 5, 5),
       ylab = "Level", xlab = "",col=c("blue","red","red"),lwd=1)
legend("topright",
         leg = "signal +/- 1.64SE",
         cex = 0.7, lty = c(5),col = c("red"),
         bty = "y", horiz = F)

#filtro e predição
```

```
fn <- function(params){
  mod <- dlmModPoly(order = 1, dV =exp(params[1]), dW = exp(params[2]))
  return(mod)
}
fit <- dlmMLE(data.2, rep(0,2),fn)
mod <- fn(fit$par)
filtered <- dlmFilter(data.2,mod)
smoothed <- dlmSmooth(filtered)
mu.s <- dropFirst(smoothed$s)
mu.f <- dropFirst(filtered$a)
mu.m <- dropFirst(filtered$m)
temp <- cbind( mu.s,mu.f)
par(mfrow=c(1,1))
plot(temp, plot.type = "s", type = "l",lty = c(1, 2),
       ylab = "Level", xlab = "",col=c("blue","sienna"),lwd=1)
legend("topright",leg = c("smoothed level","filtered level"),
         cex = 0.7,lty = c(1, 2), col = c("blue","sienna"),
         bty = "y", horiz = T)

temp <- window(cbind( data.2,mu.f),start = 1978,end=1983)
plot(temp, plot.type = "s", type = "l",lty = c(1, 2),
       ylab = "Level", xlab = "",col=c("darkgrey","blue"),lwd=1)
legend("topright",leg = c("log fatalities Norway","filtered level"),
         cex = 0.6,lty = c(1, 2), col = c("darkgrey","blue"),
         bty = "y", horiz = F)
temp <- data.2 - mu.f
par(mfrow=c(1,1))
plot(temp, type = "l", ylab = "", xlab = "",lty=2, col = "darkgrey")
abline(h=0, col = "sienna")
legend("topleft",leg = c("prediction error"),
       cex = 0.6,lty = c(2), col = c("darkgrey"),
       bty = "y", horiz = F)
cov <- (dlmSvd2var(filtered$U.R, filtered$D.R))

var=function(cov){
  var=0
  for(i in 1:length(cov)){
    var[i]=mod$FF%*%cov[[i]]%*%t(mod$FF)
  }
  return(var)
}
var=var(cov)

#var <- (sapply(cov,function(x) mod$FF%*%x%*%t(mod$FF)))+V(mod)

lev.ts <- ts(var[-1],start = 1969, frequency =1)
par(mfrow=c(1,1))
plot(lev.ts,xlab="",ylab = "",lty=2)
legend("topright",leg = c("prediction error"),
```

```
       cex = 0.6,lty = c(2), col = c("black"),
       bty = "y", horiz = F)
#teste diagnóstico

X <- data.frame(intervention = intervention, logpprice = prices)
fn <- function(params){
  level <- dlmModPoly(order = 1,dV=exp(params[1]))
  seas <- dlmModSeas(frequency = 12,dV=exp(params[1]),dW=rep(0,11))
  reg <- dlmModReg(X,addInt=FALSE,dV=exp(params[1]),dW=c(0,0))
  mod <- level+seas+reg
  diag(W(mod))[1:2] <- c(exp(params[2]),0)
  mod
}
fit <- dlmMLE(data.1, rep(0,2),fn)
mod <- fn(fit$par)
filtered <- dlmFilter(data.1,mod)
smoothed <- dlmSmooth(filtered)
sm <- dropFirst(smoothed$s)
lam.1 <- sm[1,13]
beta.1 <- sm[1,14]
mu <- c(sm[,1])
nu <- c(sm[,2])
res <- c(residuals(filtered,sd=F))
res <- ts(res[-c(1:14)], start = 1970.167)
par(mfrow=c(1,1))
plot(ts(res,start = c(1970,1),frequency = 12),col = "darkgrey",ylab="",xlab="")
abline(h=0,col="sienna",lty=2)
legend("topright",leg = c("standardised one-step prediction errors"),
       cex = 0.6,lty = c(1), col = c("darkgrey"),
       bty = "y", horiz = F)
par(mfrow=c(1,1))
acf(res,lag=10,main ="",xlim=c(0,10))
par(mfrow=c(1,1))
hist(res,breaks=seq(-3.5,3,length.out=14),
       prob=T,col="grey",main = "",ylim=c(0,0.6),xlab="Resíduos",ylab="Densidade")
curve(dnorm(x,mean=mean(res),sd=sd(res)),col=2,lty=2,lwd=2,add=TRUE)
shapiro.test(res)
Box.test(res, lag = 10, type = "Ljung")
sapply(1:15,function(l){round(Box.test(res, lag=l, type = "Ljung-Box")$p.value,4)})

#previsão

fn <- function(params){
  dlmModPoly(order= 1, dV= exp(params[1]) , dW = exp(params[2]))
}
fit <- dlmMLE(data.2, rep(0,2),fn)
mod <- fn(fit$par)
filtered <- dlmFilter(data.2,mod)
var <- unlist(dlmSvd2var(filtered$U.R, filtered$D.R))
```

```
wid <- qnorm(0.05, lower = FALSE) *sqrt(c(var))
temp <- cbind(filtered$f, filtered$f + wid,filtered$f-wid)
temp <- ts(temp,start = 1970,frequency =1)
forecast <- dlmForecast(filtered,nAhead=5)
var.2 <- unlist(forecast$Q)
wid.2 <- qnorm(0.05, lower = FALSE) *sqrt(c(var.2))
temp.2 <- cbind(forecast$f, forecast$f +wid.2 , forecast$f- wid.2)
temp.3 <-ts(rbind(temp,temp.2),start = 1970, frequency= 1,end=2008)
par(mfrow=c(1,1))
plot(dropFirst(temp.3),plot.type="s",col=c("blue","red","red"),
       xlim = c(1970,2010),lty=c(1,2,2),lwd=c(2,1,1),ylab="")
lines(data.2,col="darkgrey")
abline(v=2004,col = "sienna",lwd=3)
legend("topright",
         leg = c("log fatalities in Norway",
                 " filtered level and forecasts",
                 "bands"),
         cex = 0.7, lty = c(1, 1,2),
         lwd =c(1,2,1),col = c("darkgrey","blue","red"),
         bty = "y", horiz = F)

fn <- function(params){
  dlmModPoly(order= 2, dV= exp(params[1]) , dW = exp(params[2:3]))
}
fit <- dlmMLE(data.3, rep(0,3),fn)
mod <- fn(fit$par)
filtered <- dlmFilter(data.3,mod)
cov <- (dlmSvd2var(filtered$U.R, filtered$D.R))
var=function(cov){
  var=0
  for(i in 1:length(cov)){
    var[i]=mod$FF%*%cov[[i]]%*%t(mod$FF)
  }
  return(var)
}
var=var(cov)

#var <- (sapply(cov,function(x) mod$FF%*%x%*%t(mod$FF)))+V(mod)
wid <- qnorm(0.05, lower = FALSE) *sqrt(c(var))
filt.trend <- filtered$f[,1]
temp <- cbind(filt.trend, filt.trend + wid,filt.trend-wid)
temp <- ts(temp[-c(1:2),],start = 1972,frequency =1)
forecast <- dlmForecast(filtered,nAhead=5)
var.2 <- unlist(forecast$Q)
wid.2 <- qnorm(0.05, lower = FALSE) *sqrt(c(var.2))
temp.2 <- cbind(forecast$f, forecast$f +wid.2 , forecast$f- wid.2)
temp.3 <-ts(rbind(temp,temp.2),start = 1972, frequency= 1,end=2008)
par(mfrow=c(1,1))
plot(temp.3,plot.type="s",col=c("blue","red","red"),
```

```
       xlim = c(1970,2010),lty=c(1,2,2),lwd=c(2,1,1),ylab="")
lines(data.3,col="darkgrey")
abline(v=2004,col = "sienna",lwd=3)
legend("topright",
         leg = c("log fatalities in Norway",
                 " filtered level and forecasts",
                 "bands"),
         cex = 0.7, lty = c(1, 1,2),
         lwd =c(1,2,1),col = c("darkgrey","blue","red"),
         bty = "y", horiz = F)

start <- time(data.1)[1]
data.1.m <- ts(c(c(data.1)[1:169],rep(NA,23)),frequency =12,start = start)
x <- prices
fn <- function(params){
  mod <- dlmModPoly(order = 1 ) +
    dlmModSeas(frequency =12)+
    dlmModReg(x, addInt=FALSE)
  V(mod) <- exp(params[1])
  diag(W(mod))[1] <- exp(params[2])
  mod
}
fit <- dlmMLE(data.1.m, rep(0,2),fn)
mod <- fn(fit$par)
filtered <- dlmFilter(data.1.m,mod)
forecasted <- ts(c(filtered$f)[171:193],
start = time(data.1)[171],frequency =12)
forecasted.mod.1 <- forecasted
sig.plus.forecast <- filtered$f
par(mfrow=c(1,1))
plot(forecasted.mod.1,col="grey",main = "",xlab = "",ylab="")
X <- data.frame(intervention = intervention, logpprice = prices)
fn <- function(params){
  level <- dlmModPoly(order = 1,dV=exp(params[1]))
  seas <- dlmModSeas(frequency = 12,dV=exp(params[1]),dW=rep(0,11))
  reg <- dlmModReg(X,addInt=FALSE,dV=exp(params[1]),dW=c(0,0))
  mod <- level+seas+reg
  diag(W(mod))[1:2] <- c(exp(params[2]),0)
  mod
}
fit <- dlmMLE(data.1, rep(0,2),fn)
mod <- fn(fit$par)
filtered <- dlmFilter(data.1,mod)
smoothed <- dlmSmooth(filtered)
sm <- dropFirst(smoothed$s)
mu <- c(sm[,1])
nu <- c(sm[,2])
par(mfrow=c(1,1))
temp <- sm[,1]+sm[,2]+ sm[,13]*intervention+sm[,14]*prices
```

```
temp<- ts(cbind(c(data.1),temp),start = 1969,
            frequency = 12)
temp <- window(temp, start = 1982.5)
temp.2 <- window(sig.plus.forecast, start = 1982.5)
plot.ts(temp , plot.type="single" , col =c("darkgrey","blue"),lty=c(1,2),
          xlab="",ylab = "log KSI",ylim=c(6.99,7.63))
legend("topright",
         leg = c("log UK drivers KSI",
                 "signal plus forecasts","signal complete model"),
         cex = 0.7, lty = c(1, 2,1),col = c("darkgrey","blue","sienna"),
         pch=c(3,NA),bty = "n", horiz = T)
lines(sig.plus.forecast, col = "sienna")
```

Apêndice B

Códigos em R utilizados para Webscraping do site da Sabesp

```
library(httr)
library(lubridate)

baixar_sabesp <- function(data) {
  u_sabesp <- paste0(
  "http://mananciais.sabesp.com.br/api/Mananciais/ResumoSistemas/",
  data)
  r_sabesp <- httr::GET(u_sabesp,timeout(60000))
  results <- httr::content(r_sabesp, simplifyDataFrame = TRUE)
  dados=cbind(data=date(data),results$ReturnObj$sistemas[1,c(2,3,9,10,11)])
  return(dados)
}

dados=baixar_sabesp(Sys.Date())

#iniciando a sequencia, se der erro por time out rodar daqui
datas=seq(min(dados$data)-1,as.Date("2003-01-01"),-1)
#2003-01-01 é o início da série
for(i in 1:(length(datas))){
  dados=rbind(dados,baixar_sabesp(datas[i]))
}
```

Apêndice C

Códigos em R utilizados para análise dos dados

```
#bibliotecas
library(TSstudio);library(lubridate)
library(xts);library(hydroTSM)
library(fpp);library(forecast)
library(randtests);library(ggplot2)
library(dlm);library(zoo);library(rucm)
# devtools::install_github("KevinKotze/tsm")
# install.packages("dlm", repos = "https://cran.rstudio.com/", dependencies = T)
library(tsm)
setwd("C:/Banco de dados")
# alteração em função executada para exibir meses em portugues
# library(magrittr)
# library(dplyr)
# ggAddExtras <- function(xlab=NA, ylab=NA, main=NA) {
#   dots <- eval.parent(quote(list(...)))
#   extras <- list()
#   if ("xlab" %in% names(dots) || is.null(xlab) || any(!is.na(xlab))) {
#     if ("xlab" %in% names(dots)) {
#       extras[[length(extras) + 1]] <- ggplot2::xlab(dots$xlab)
#     }
#     else {
#       extras[[length(extras) + 1]] <-
ggplot2::xlab(paste0(xlab[!is.na(xlab)], collapse = " "))
#     }
#   }
#   if ("ylab" %in% names(dots) || is.null(ylab) || any(!is.na(ylab))) {
#     if ("ylab" %in% names(dots)) {
#       extras[[length(extras) + 1]] <- ggplot2::ylab(dots$ylab)
#     }
#     else {
#       extras[[length(extras) + 1]] <-
ggplot2::ylab(paste0(ylab[!is.na(ylab)], collapse = " "))
#     }
#   }
```

```
#   if ("main" %in% names(dots) || is.null(main) || any(!is.na(main))) {
#     if ("main" %in% names(dots)) {
#       extras[[length(extras) + 1]] <- ggplot2::ggtitle(dots$main)
#     }
#     else {
#       extras[[length(extras) + 1]] <-
ggplot2::ggtitle(paste0(main[!is.na(main)], collapse = " "))
#     }
#   }
#   if ("xlim" %in% names(dots)) {
#     extras[[length(extras) + 1]] <- ggplot2::xlim(dots$xlim)
#   }
#   if ("ylim" %in% names(dots)) {
#     extras[[length(extras) + 1]] <- ggplot2::ylim(dots$ylim)
#   }
#   return(extras)
# }
#
#
# ggsubseriesplot <- function(x, labels = NULL, times = time(x),
phase = cycle(x), ...) {
#   if (!requireNamespace("ggplot2", quietly = TRUE)) {
#     stop("ggplot2 is needed for this function to work.
Install it via install.packages(\"ggplot2\")", call. = FALSE)
#   }
#   else {
#     if (!inherits(x, "ts")) {
#       stop("ggsubseriesplot requires a ts object, use x=object")
#     }
#
#     if (round(frequency(x)) <= 1) {
#       stop("Data are not seasonal")
#     }
#
#     if("1" %in% dimnames(table(table(phase)))[[1]]){
#       stop(paste("Each season requires at least 2 observations.",
#                  ifelse(frequency(x)%%1 == 0,
#                         "Your series length may be too short for this graphic.",
#
# "This may be caused from specifying a time-series with non-integer frequency.")
#       )
#       )
#     }
#
#     data <-
# data.frame(y = as.numeric(x), year = trunc(time(x)), season = as.numeric(phase))
#     seasonwidth <- (max(data$year) - min(data$year)) * 1.05
#     data$time <- data$season + 0.025 + (data$year - min(data$year)) / seasonwidth
#     avgLines <- stats::aggregate(data$y, by = list(data$season), FUN = mean)
```

```
#     colnames(avgLines) <- c("season", "avg")
#     data <- merge(data, avgLines, by = "season")
#
#     # Initialise ggplot object
#     # p <- ggplot2::ggplot(ggplot2::aes_(x=~interaction(year, season), y=~y,
# group=~season), data=data, na.rm=TRUE)
#     p <- ggplot2::ggplot(
#       ggplot2::aes_(x = ~time, y = ~y, group = ~season),
#       data = data, na.rm = TRUE
#     )
#
#     # Remove vertical break lines
#     p <- p + ggplot2::theme(panel.grid.major.x = ggplot2::element_blank())
#
#     # Add data
#     p <- p + ggplot2::geom_line()
#
#     # Add average lines
#     p <- p + ggplot2::geom_line(ggplot2::aes_(y = ~avg), col = "#0000AA")
#
#     # Create x-axis labels
#     xfreq <- frequency(x)
#     if (xfreq == 4) {
#       xbreaks <- c("Q1", "Q2", "Q3", "Q4")
#       xlab <- "Quarter"
#     }
#     else if (xfreq == 7) {
#       xbreaks <- c(
#         "Sunday", "Monday", "Tuesday", "Wednesday", "Thursday",
#         "Friday", "Saturday"
#       )
#       xlab <- "Day"
#     }
#     else if (xfreq == 12) {
#       xbreaks <- c("Jan","Fev","Mar","Abr",
# "Mai","Jun","Jul","Ago","Set","Out","Nov","Dez")
#       xlab <- "Month"
#     }
#     else {
#       xbreaks <- 1:frequency(x)
#       xlab <- "Season"
#     }
#
#     # X-axis
#     p <- p + ggplot2::scale_x_continuous(breaks = 0.5 + (1:xfreq), labels = xbrea
#
#     # Graph labels
#     p <- p + ggAddExtras(ylab = deparse(substitute(x)), xlab = xlab)
#     return(p)
```

```
#   }
# }

##############################################################################
#
# Grafico de amplitude x media
#
# chutar o valor de k,
#
# se a série for, sabidamente sazonal, usar o período
#
##############################################################################
med.var<-function(x,k)
{N<-length(x)
x.m<-rep(0,(N-k))
x.r<-rep(0,(N-k))
for (i in 1:(N-k)) x.m[i]<-mean(x[i:(i+k)])
for (i in 1:(N-k)) x.r[i]<-max(x[i:(i+k)])-min(x[i:(i+k)])
df=data.frame(x=x.m,y=x.r)
aa1=lm(x.r~x.m)
if(summary(aa1)$coefficients[2,4]<=0.05){
  print("Coeficiente beta é significativo a 5%")
}
else{
  print("Coeficiente beta NÃO é significativo a 5%")
}
ggplot(data=df,aes(x=x.m,y=x.r))+geom_point()+
geom_smooth(method = "lm")+xlab("média")+ylab("amplitude")
return(summary(aa1))
}
#############################################################

# lendo dados e tratando
dados=read.csv2("dados_completos_mananciais_sabesp.csv")
dados=dados[,-c(1,3)]
dados$data=dmy(dados$data)

# tratando ano bissexto, assumindo que 1/03 vai ter 48 horas
# ai faço a média da pluviometria para MATAR anos bissextos
# estou fazendo isso 5 vezes (em tese não dará problemas)
# Faço isso pra não precisar ficar usando 365.5

bissextos_pos=which(day(dados$data)==29 & month(dados$data)==2)
dados[bissextos_pos-1,-1]=
round((dados[bissextos_pos-1,-1]+dados[bissextos_pos,-1])/2,1)

dados=dados[format(dados$data, "%m %d") != "02 29",] #removendo bissexto
rm(bissextos_pos)
```

```
# criando objeto xts
chuva=xts(dados$PrecDia, order.by = as.Date(dados[,1],"%y%m%d"))

# criando ts
chuva_ts=ts(dados$PrecDia,start=c(2003,1),frequency = 365)

# plotando a série diária
# plot.xts(chuva,type='b',main="Precipitação diária em mm",lwd=0.1)
# imprimindo com transparencia
plot.xts(chuva,type='b',main="Precipitação diária em mm",
col=rgb(0,0,50,50,maxColorValue=255),lwd=0.1)
addEventLines(chuva[which(chuva>60),],pos=c(2,1),cex=0.8,col="red")

# chuva semanal
chuva_semanal=apply.weekly(chuva,sum)
chuva_semanal_ts=xts_to_ts(chuva_semanal)
plot.xts(chuva_semanal,main="Precipitação semanal em mm",
ylim=c(0,260),lwd=0.05,major.ticks = "weeks")
eventos=xts(as.character(date(chuva_semanal[
which(chuva_semanal>150),])),as.Date(date(chuva_semanal[
which(chuva_semanal>150),])))
addEventLines(chuva_semanal[which(chuva_semanal>150),],
pos=c(2,1),cex=0.8,col=rgb(50,0,0,50,maxColorValue=255))
addEventLines(eventos,pos=c(1,2),cex=0.8,col="red")

# chuva mensal
chuva_mensal=apply.monthly(chuva,sum)
chuva_mensal_ts=xts_to_ts(chuva_mensal)
plot.xts(chuva_mensal,main="Precipitação mensal em mm",
ylim=c(0,550),lwd=0.5,major.ticks = "months",grid.ticks.on = "months")
eventos=xts(as.character(date(chuva_mensal[which(chuva_mensal>350),])),
as.Date(date(chuva_mensal[which(chuva_mensal>350),])))
addEventLines(chuva_mensal[which(chuva_mensal>350),],pos=c(1,2),
cex=0.8,col="red")
addEventLines(eventos,pos=c(1,2),cex=0.8,col="red")


ggsubseriesplot(chuva_ts)
ggsubseriesplot(chuva_semanal_ts)
ggsubseriesplot(chuva_mensal_ts,ylab="precipitação mensal",xlab="meses")
#alterei xbreaks na função do pacote para aparecer em meses em portugues

# ggseasonplot(chuva_ts, year.labels=TRUE, year.labels.left=TRUE)
# ggseasonplot(chuva_semanal_ts, year.labels=TRUE, year.labels.left=TRUE)
# ggseasonplot(chuva_mensal_ts, year.labels=TRUE, year.labels.left=TRUE)

# Como notamos a série possui variabilidade muito grande
# tratando a variabilidade com Transformação de Box Cox
med.var(BoxCox(chuva_ts,lambda = 0.12),k=365)
```

```
# lambda=BoxCox.lambda(chuva_ts) #dá o lambda ótimo (autor:Guerrero)
chuva_ts=BoxCox(chuva_ts,lambda = 0.12)

a=data.frame(date=as.Date(rev(dados[,1])),precip=as.matrix(chuva_ts))
chuva=xts(a$precip,a$date);rm(a)
plot.xts(chuva,type='b',main=
"Precipitação diária em mm",col=rgb(0,0,50,50,maxColorValue=255)
,ylim=c(-10,10),lwd=0.1)

chuva_semanal=apply.weekly(chuva,sum)
chuva_mensal=apply.monthly(chuva,sum)
chuva_semanal_ts=xts_to_ts(chuva_semanal)
chuva_mensal_ts=xts_to_ts(chuva_mensal)

plot.xts(chuva_semanal,main="Precipitação semanal em mm",ylim=c(-70,35),lwd=0.05)
plot.xts(chuva_mensal,main="Precipitação mensal em mm",ylim=c(-260,100),lwd=0.5)

# verificando tendência
# ago,set, out de  forma pronunciada e (nov e dez)  não pronunciada
# nos ultimos anos puxa a média pra baixo
# fim do inverno e primavera mais seca
# de fato tenho sazonalidade anual (oque corrobora pro conhecimento geral nosso)
# invernos mais secos e verão umido
# primaveras mais secas e finais de invernos um poucos mais secos
# (isso pros ultimos anos)


ggsubseriesplot(chuva_ts)
ggsubseriesplot(chuva_semanal_ts)
ggsubseriesplot(chuva_mensal_ts)
# série com dados mensais, somando os valores diarios de cada mes

# Dickey Fuller ou Phillips-Perron  (Testes de raiz unitária)
# 0.01 pois a serie é longa
adf.test(chuva_ts,k=33);pp.test(chuva_ts,lshort=F) #truncando no lag 33

# ACF para complementar sazonalidade
acf(chuva_ts)
acf(chuva_ts,1500)  #picos LEVEMENTE amortecidos

# aponta pra mesma informação de sazonalidade anual
# estrutura de longa duração e amortecimento de forma hiperbólica

pacf(chuva_ts)


modelChuva=StructTS(chuva_mensal_ts,type="BSM",optim.control = c(0,0,0,0))
?KalmanRun()
```

```
# em modelos dinamicos as notações confundem,
#olhe no detalhe e veja as representações
plot(fitted(modelChuva))
plot(chuva_mensal_ts,lwd=1.6,xlab="tempo",
ylab="precipitação");lines(index(modelChuva$data),
apply(modelChuva$fitted,1,sum),col="blue",lty=2);
legend(x=2010,y=100,leg = c("chuva mensal transformada","modelo ajustado"),
cex = 0.6, lty = c(1,2),col = c("black","blue"),bty = "n", horiz = T)
#análise diagnóstico
tsdiag(modelChuva)
hist(residuals(modelChuva),main="",xlab="resíduos",ylab="frequência")
shapiro.test(residuals(modelChuva))
acf(residuals(modelChuva),1000)

#Filtering
filter=KalmanRun(chuva_mensal_ts,modelChuva$model)
plot(chuva_mensal_ts);lines(index(chuva_mensal_ts),
apply(filter$states,1,sum),col="blue",type="l")
#Smoothing
lines(index(modelChuva$data),apply(tsSmooth(modelChuva),1,sum),
col="orange",type="l")
#forecast
forecast=KalmanForecast(n.ahead = 12, modelChuva$model)
#KalmanForescast e forecast apresentaram estimativas pontuais
#identicas com 5 casas decimais
lines(seq(2020.250,2021.163,by=0.083),forecast$pred,col="red")

library(astsa)
a=arima(chuva_mensal_ts, order =c(2,1,0))
plot(chuva_mensal_ts,lwd=1.6,xlab="tempo",ylab="precipitação");
lines(fitted(a),col="red",lty=2);legend(x=2010,y=70,leg = c("chuva mensal
transformada","arima ajustado"),


tsdiag(a)
hist(residuals(a),main="",xlab="resíduos",ylab="frequência")
shapiro.test(residuals(a))
acf(residuals(a),1000)

#nested Time series Cross Validation
#length(chuva_mensal_ts)/36
#train_index=list(c(1),c(1,2),c(1,2,3),c(1,2,3,4),c(1,2,3,4,5))
#test_index=list(2,3,4,5,6)
split_train=list(subset(chuva_mensal_ts,start=1,end=36),
subset(chuva_mensal_ts,start=1,end=72),
subset(chuva_mensal_ts,start=1,end=108),
subset(chuva_mensal_ts,start=1,end=144),
subset(chuva_mensal_ts,start=1,end=180))
```

```
split_test=list(subset(chuva_mensal_ts,start=37,end=72),
subset(chuva_mensal_ts,start=73,end=108),subset(chuva_mensal_ts,start=109,end=144),

rmse_mod1=rep(NA,length(split_test))
for(i in 1:5){
  model_train=StructTS(split_train[[i]],type="BSM",optim.control = c(0,0,0,0))
  forecast=KalmanForecast(n.ahead = 36, model_train$model)
  rmse_forecast=sqrt(mean((forecast$pred-split_test[[i]]
  [1:length(split_test[[i]])])^2))
  rmse_mod1[i]=rmse_forecast
}
rmse_mod2=rep(NA,length(split_test))
for(i in 1:5){
  model_train=arima(split_train[[i]], order =c(2,1,0))
  forecast=KalmanForecast(n.ahead = 36, model_train$model)
  rmse_forecast=sqrt(mean((forecast$pred-split_test[[i]]
  [1:length(split_test[[i]])])^2))
  rmse_mod2[i]=rmse_forecast
}
plot(rmse_mod1,xlab="Interação",ylab="RMSE",pch=3,ylim=c(50,130),
lwd=1.5);points(rmse_mod2,col="red");legend(
x=0.9,y=130,leg = c("modelo estrutural básico","ARIMA"),
cex = 0.6, pch = c(3,1),col = c("black","red"),bty = "n", horiz = T)
#conclusão
#ok, modelo proposto é melhor que o arima do sentido de possuir
rmse mais baixo no geral

#finalmente vamos a previsão de 2012 a 2015 para saber se está fora de controle
#estabelecemos que um ponto sai fora de controle quando sai do intervalo estabelecido
# de 95% ou seja, o próprio ponto, +-1.96/sqrt(n)
#forescast(), coincide com o KalmanForecast, logo vamos utilizar o forecast
#por já me retornar o intervalo de confiança

janela_treino=subset(chuva_mensal_ts,start=1,end=120)
janela_prev=subset(chuva_mensal_ts,start=121,end=156)
model_train=StructTS(janela_treino,type="BSM",optim.control = c(0,0,0,0))
#model_train=arima(janela_treino, order =c(2,1,0))
#forecasts=KalmanForecast(n.ahead = 36, model_train$model)
forecasts=forecast(model_train,h=36);forecasts

#pontual
pontual=forecasts[["mean"]]
#intervalar
lower=ts(forecasts[["lower"]][,1],start=c(2013,1),frequency = 12)
upper=ts(forecasts[["upper"]][,1],start=c(2013,1),frequency = 12)


plot(janela_prev,ylim=c(-500,100),type="p",ylab="Precipitação",xlab="Tempo")
lines(pontual,col="purple",lwd=2)
```

```
polygon(c(time(janela_prev),rev(time(janela_prev))), c(upper,rev(lower)),
        col=rgb(0,0,0.6,0.2), border=FALSE)

fora=(janela_prev < lower | janela_prev > upper)
points(time(janela_prev)[fora], janela_prev[fora], pch=19,col="red")
legend("bottomleft",leg = c("fora de controle",
"previsão pontual","dados precipitação"),
       cex = 0.7, lty = c(0,1,0), pch=c(16,NA,1),
       col = c("red","purple","black"),bty="o",horiz = T)
```