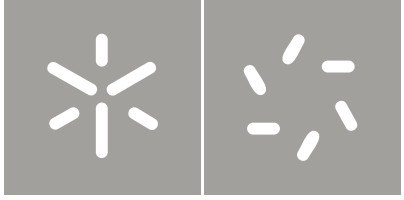
Universidade do Minho
Escola de Ciências

Lara Maria Lopes Teixeira

Análise de *Change-points* em Séries Temporais

Análise de *Change-points* em Séries Temporais
Lara Maria Lopes Teixeira
UMinho | 2012

Outubro de 2012



Universidade do Minho
Escola de Ciências

Lara Maria Lopes Teixeira

Análise de *Change-points* em Séries Temporais

Tese de Mestrado
Estatística

Trabalho efectuado sob a orientação da
Professora Doutora Arminda Manuela Andrade Pereira Gonçalves

e co-orientação do
Professor Doutor Marco André da Silva Costa

Outubro de 2012

Agradecimentos

Gostaria de agradecer a um conjunto de pessoas que de diferentes formas me ajudaram:

À Professora Doutora Arminda Manuela Gonçalves e ao Professor Doutor Marco Costa, pela orientação, pela disponibilidade, humanidade e especialmente pela partilha dos seus conhecimentos;

A todos os professores e colegas que me acompanharam durante o meu percurso académico;

À minha mãe pelo incentivo e pela força que sempre me transmitiu, ao meu pai, à Vanda, minha irmã, e ao Sérgio pela compreensão e apoio que me deram.

Resumo

A análise de *change-points* é um processo importante na análise de séries temporais, permitindo a identificação e o estudo de pontos de mudança na sucessão de observações. O problema relativo à análise de *change-points* tem sido um tópico de interesse de análise estatística, verificando-se um rápido desenvolvimento das técnicas de análise, principalmente nas últimas décadas, devido à melhoria acentuada das ferramentas computacionais e ao facto deste tipo de problemas surgirem em áreas tão importantes como a Medicina, a Economia e as Finanças, a Psicologia e o Ambiente.

Neste trabalho apresentam-se os aspectos principais subjacentes à análise de *change-points*, nomeadamente, são abordados vários tipos de *change-points* que se podem observar e os métodos de análise que têm surgindo. A abordagem informacional é uma técnica geral de selecção de modelos e consiste em utilizar um critério de informação para identificar a posição desconhecida de um *change-point* num modelo, discriminando de entre os vários modelos, o que é mais verosímil para ajustar os dados. Um dos critérios de informação desenvolvidos é o *Schwarz Information Criterion* (SIC). Esta é a metodologia utilizada no estudo das séries temporais relativas à variável de qualidade da água Oxigénio Dissolvido, medida mensalmente desde Janeiro de 1999 a Dezembro de 2011, em oito estações de monitorização da bacia hidrográfica do Rio Ave. As variações temporais de dados ambientais são complexas e pode ser difícil identificar os denominados *change-points* com modelos tradicionais aplicados a este tipo de problemas. Neste estudo, como as séries de observações apresentam um comportamento sazonal, propõe-se uma abordagem alternativa na aplicação da análise de *change-points* tendo em conta esta estrutura dos dados.

A aplicação da análise de *change-points* permitiu detectar *change-points* na média e na variância, simultaneamente, nas oito séries de observações estudadas. Como os pressupostos de normalidade e independência da metodologia aplicada não se verificam em algumas séries temporais estudadas foi realizado um estudo de simulação de modo a avaliar o desempenho da metodologia, quando aplicada a séries de dados não normais e/ou com correlação temporal. A principal conclusão do estudo de simulação é que na presença de correlação a metodologia tende a detectar falsos *change-points*. Contudo, atendendo aos resultados obtidos na aplicação prática, a correlação identificada não coloca em causa a validade da análise efectuada uma vez que os *change-points* continuam significativos mesmo considerando-se significâncias inferiores (mesmo a um nível de 1%).

Abstract

Change-points analysis is an important process in time series analysis, allowing identifying and studying change-points in the observation series. The problem relative to change-points analysis has been a relevant topic in statistical analysis and there has been a swift development of the analysis techniques, mainly during the last decades, due to the sharp improvement of computing tools and to the fact that these types of problems arise in areas of such importance as Medicine, Economy, Finances, Psychology, and Environment.

In this paper are presented the main features underlying the change-points analysis, namely several types of change-points that can be observed and the analysis methods that have arisen throughout time. The informational approach is a general methodology of model selection and consists of using an informational criterion to identify the unknown position of a change-point in a given model, by discriminating among the various models the one more likely to fit the data. One of the informational criteria is the Schwarz Information Criterion (SIC). This is the methodology used in the study of time series relatively to Dissolved Oxygen as a water quality variable measured monthly since January 1999 to December 2011 in eight monitoring stations of the River Ave's hydrographic basin. Time variations in environmental data are complex and it can be difficult to identify the so-called change-points with traditional models applied to this type of problems. In this study, as the series of observations present a seasonal behavior, we propose an alternative approach in the application of the change-points analysis by taking into account this data structure.

The application of change-points analysis allowed detecting change-points in the average and variance simultaneously in the eight observation series under study. As the assumptions of normality and independence of the applied methodology are not present in some time series, we have carried out a simulation study in order to evaluate the methodology's performance when applied to non-normal data series and/or with time correlation. The main conclusion of the simulation study is that in the presence of correlation the methodology tends to detect false change-points. However, by taking into account the results obtained in the practical application, the identified correlation does not jeopardize the analysis validity, since the change-points are still significant even considering lower levels of significance (even at a level of 1%).

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

"Todo o mundo é composto de mudança."[1] A percepção e compreensão de determinadas mudanças podem ajudar a compreender e a dar respostas a diversos problemas com especial relevância na actualidade. Em muitas situações práticas são necessárias metodologias estatísticas que permitam a identificação e o estudo de mudanças numa sucessão de observações ordenadas no tempo, traduzindo-se este problema como a análise de *change-points*.

Um "change-point", em português "ponto de mudança"[2], é um ponto no tempo em que os parâmetros da distribuição subjacente da série temporal ou os parâmetros do modelo utilizado para descrever a série repentinamente se alteram (Beaulieu *et al.*, 2012). A análise de *change-points*, usualmente, divide-se em dois aspectos: o primeiro é detectar se ocorreu alguma mudança na série da variável aleatória observada e o segundo é estimar o número de mudanças e as suas localizações (Chen & Gupta, 2012).

O primeiro estudo de *change-points* foi desenvolvido na década de 1950, com o estudo de Page (1954), que desenvolveu o método *cumulative sum* (CUSUM), em português, método de somas cumulativas, concentrando-se nas mudanças na média, isto é, no comportamento médio das observações. Desde então tem-se observado um rápido desenvolvimento das técnicas de análise de *change-points*, principalmente nas últimas décadas, devido à melhoria acentuada das ferramentas computacionais. O aumento dos estudos deve-se também ao facto deste tipo de problemas surgirem em áreas tão importantes como a Medicina, a Economia e as Finanças, a Psicologia, o Ambiente, entre muitas outras.

O problema da detecção e análise de *change-points* está associado a diferentes mudanças de comportamento da série temporal que podem ocorrer, como por exemplo, mudanças na média, na variância, em ambas simultaneamente e ainda em mudanças associadas a modelos de regressão linear. Uma descrição dos vários tipos de *change-points* pode ser encontrada em Chen & Gupta (2012) e Beleaulieu *et al.* (2012).

[1] Verso de Luís de Camões, 1595.

[2] Tradução retirada do Glossário Inglês-Português de Estatística da Sociedade Portuguesa de Estatística e da Associação Brasileira de Estatística.

Vários métodos, com diferentes abordagens, têm sido desenvolvidos de modo a dar resposta ao problema da análise de *change-points*. Numa abordagem não-paramétrica Hájek (1962) utilizou testes de *ranks* para alterações num modelo de regressão e Milton (1965) para mudanças de nível. Chernoff & Zacks (1964) estudaram mudanças na média baseados numa abordagem bayesiana. O teste da razão de verosimilhanças foi desenvolvido por Hawkins (1977) e mais tarde por Worsley (1979) para mudanças na média (com variância conhecida e desconhecida). O método de somas cumulativas foi desenvolvido por Page (1955), e Schwarz (1978) desenvolveu uma abordagem informacional. Neste contexto utiliza-se a palavra "informacional" para referir-se a critérios baseados na informação da amostra. A maioria dos métodos baseiam-se nos pressupostos de normalidade e independência, não tendo em conta estruturas que podem ser apresentadas em conjuntos de observações no tempo, como por exemplo a sazonalidade e a correlação.

1.1 Dados e motivação

A actividade humana exercida sobre a natureza tem aumentado desde a segunda revolução industrial, reflectindo-se no mundo actual e levando a uma importância crescente das questões de sustentabilidade do Ambiente. Sendo assim, o uso de metodologias diferenciadas para a avaliação do impacto e das mudanças, que vêm ocorrendo, é pertinente e essencial para a gestão dos diversos problemas resultantes destas questões de sustentabilidade.

Neste estudo serão analisados dados relativos a variáveis de qualidade da água, um dos recursos naturais de importância vital. Os dados dizem respeito à bacia hidrográfica do Rio Ave situada no Noroeste de Portugal, onde a monitorização se tem tornado uma prioridade no planeamento e gestão da qualidade da água desta bacia hidrográfica. A base económica do Vale do Ave está ligada fortemente à indústria, sendo a água um factor determinante na localização industrial, mas esta industrialização tem conduzido a uma má qualidade da mesma desde meados da década de 1970. Será utilizada a variável Oxigénio Dissolvido (OD), uma das mais importantes variáveis na avaliação da qualidade das águas superficiais de uma bacia (Costa & Gonçalves, 2011 e Gonçalves e Costa, 2012), medida mensalmente desde Janeiro de 1999 a Dezembro de 2011, em oito estações de monitorização.

Neste trabalho, o estudo do comportamento da série temporal da variável de qualidade da água, Oxigénio Dissolvido, será abordado na linha de investigação de Gonçalves & Costa(2011) e Gonçalves & Alpuim (2011) que estudaram alterações na tendência das séries temporais deste tipo de variáveis de qualidade da água. A utilização de metodologias de análise de *change-points*, procura determinar o tipo de mudanças e o instante em que estas ocorrem.

1.2 Objectivos e estrutura do trabalho

O objectivo principal do trabalho é analisar metodologias de detecção de *change-points*, de modo a possibilitar o estudo e detecção de pontos de mudança no comportamento em séries de dados de qualidade da água, a aplicação é efectuada à variável de qualidade Oxigénio Dissolvido, identificando a natureza do ponto de mudança e em que instantes do tempo ocorrem.

Com esse intuito, no Capítulo 2 será abordada a problemática inerente à análise de *change-points*, com os aspectos principais e essenciais para a compreensão do tema, como a formalização do problema em estudo, os principais tipos de *change-points* que se observam e as metodologias mais usuais. Serão ainda apresentados alguns problemas que surgem na análise de *change-points* e respectivas abordagens, que têm vindo a ser desenvolvidos nos últimos tempos.

No Capítulo 3 será apresentada a metodologia *Schwarz Information Criterion* (SIC), denominada em português por "Critério de Informação de Schwarz", baseada numa abordagem informacional, útil para discriminar os vários modelos de *change-points*. Esta metodologia será a aplicada aos dados em estudo, pois apresenta a vantagem de poder ser adaptada a um conjunto vasto de situações, bem como ser utilizada para detectar diferentes tipos de *change-points*.

A aplicação da metodologia, baseada no Critério de Informação de Schwarz, aos dados de qualidade da água será apresentada no Capítulo 4. Como as variações temporais de dados hidrológicos são complexas, pode ser difícil identificar os denominados *change-points* com os modelos tradicionais aplicados a este tipo de problemas, pois a maioria das séries de dados ambientais apresentam estruturas inerentes como a sazonalidade. Esta variação sazonal surge principalmente no caso de observações mensais e requer o desenvolvimento de outras metodologias. Neste estudo, como os dados referem-se a variáveis observadas mensalmente, apresentando sazonalidade, propõe-se uma abordagem alternativa na aplicação da análise de *change-points* tendo em conta esta estrutura dos dados.

Como os pressupostos de normalidade e independência, da metodologia aplicada no Capítulo 4, não se verificam em todas as séries temporais estudadas, no Capítulo 5 será realizado um estudo de simulação de modo a avaliar o comportamento da metodologia, quando aplicada a séries de dados não normais e com correlação temporal.

As conclusões do trabalho desenvolvido e dos resultados obtidos serão descritas no Capítulo 6, assim como linhas de investigação para o trabalho futuro.

Capítulo 2

Análise de *Change-points*

O problema relativo à análise de *change-points* tem sido um tópico de interesse de análise estatística nas últimas décadas, podendo vários problemas práticos serem encontrados em diversas áreas de conhecimento.

Na área do Ambiente, as técnicas de análise de *change-points* têm sido muito usadas, nomeadamente no contexto das problemáticas associadas à exaustiva exploração da natureza e às suas consequências. No que respeita a estudos de alterações climáticas, por exemplo, Lund & Reeves (2002) estudaram a temperatura média anual em Chula Vista, Califórnia, e Jarušková (2010) estudou as temperaturas médias mensais em Estocolmo. Relativamente à poluição do ar, Barratt *et al.* (2007) estudaram a concentração de monóxido de carbono antes e depois da introdução de uma linha de *bus* na Rua Marylabone, no centro de Londres, e Jarušková (1996) analisou séries temporais relativas à pressão do ar. Chu *et al.* (2012) estudaram mudanças na precipitação máxima anual no sul de Taiwan.

Na área da Economia e Finanças, também podem ser encontrados vários estudos sobre *change-points*, como por exemplo, numa publicação de Inclán & Tiao (1994), onde se analisaram séries de dados relativos ao mercado de acções da *International Business Machines* (IBM) e Hsu (1977) estudou o impacto do caso *Watergate* nas acções dos Estados Unidos da América.

Neste capítulo são apresentados os aspectos fundamentais que envolvem a problemática da análise de *change-points*. Começa-se pela formulação do problema que se pretende estudar, seguindo-se a explicitação dos tipos de *change-points* que se podem encontrar e das metodologias mais usuais para a sua análise. Por fim, será feita uma abordagem do problema de múltiplos *change-points* e de algumas propriedades inerentes a algumas séries de observações.

2.1 Formulação do problema

A inferência estatística sobre *change-points* abrange dois aspectos, detectar se ocorreu alguma mudança na série da variável aleatória observada e estimar o número de mudanças e as suas localizações no tempo.

Sejam $X_1, X_2, \ldots, X_n$ uma sequência de variáveis aleatórias independentes com função de distribuição $F_1, F_2, \ldots, F_n$, respectivamente. Pretende-se, geralmente, testar a seguinte hipótese nula,

$$H_0 : F_1 = F_2 = \ldots = F_n \tag{2.1}$$

versus a hipótese alternativa

$$H_1 : F_1 = \ldots = F_{k_1} \neq F_{k_1+1} = \ldots = F_{k_2} \neq F_{k_2+1} = \ldots F_{k_q} \neq F_{k_q+1} \ldots = F_n, \tag{2.2}$$

onde $1 < k_1 < k_2 < \ldots < k_q < n$, q é o número de *change-points* e $k_1, k_2, \ldots, k_q$ são as respectivas posições desconhecidas que têm de ser estimadas.

Se as distribuições $F_1, F_2, \ldots, F_n$ pertencem à mesma família paramétrica $F(\theta)$, então terá de ser testada a hipótese nula sobre os parâmetros populacionais θ_i, $i = 1, \ldots, n$

$$H_0 : \theta_1 = \theta_2 = \ldots = \theta_n = \theta \quad \text{(desconhecido)} \tag{2.3}$$

versus a hipótese alternativa

$$H_1 : \theta_1 = \ldots = \theta_{k_1} \neq \theta_{k_1+1} = \ldots = \theta_{k_2} \neq \ldots \neq \theta_{k_{q-1}+1} = \ldots = \theta_{k_q} \neq \theta_{k_q+1} = \ldots = \theta_n, \tag{2.4}$$

onde q e $k_1, k_2, \ldots, k_q$ têm de ser estimados. Com estas hipóteses abrangem-se os dois aspectos da inferência estatística sobre *change-points* referidos inicialmente.

De referir ainda que as hipóteses adaptam-se à situação de existir apenas uma mudança na sequência de observações ou existirem múltiplos *change-points*.

2.2 Tipos de *change-points*

As mudanças que podem surgir numa série temporal são várias, sendo os *change-points* na média, na variância, em ambas simultaneamente e, ainda, os *change-points* associados a modelos de regressão linear, os tipos de pontos de mudança mais estudados e que mais se observam em situações práticas.

O caso mais comum de *change-point* é o associado a modelos com erros que seguem uma distribuição normal. Assim, para exemplificar os vários tipos de *change-points* serão apresentadas nesta secção sequências simuladas de valores normalmente distribuídos, com

parâmetros de acordo com o tipo de *change-point* que se pretende detectar.

2.2.1 *Change-point* na média

O problema de pontos de mudança na média foi inicialmente estudado por Page (1954, 1955, 1957) com o desenvolvimento do método de somas cumulativas. Gardner (1969) estudou o mesmo problema, mas sob o ponto de vista bayesiano e Bhattacharya & Johnson (1968) utilizaram uma abordagem não-paramétrica. Mais recentemente podem ser encontrados estudos sobre *change-points* na média em Chen & Gupta (2001), com a utilização do procedimento da razão de verosimilhanças, e em Beaulieu *et al.* (2012) com a aplicação da abordagem informacional.

Assumindo a igualdade de variâncias, $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_n^2$, pretende-se testar a hipótese nula de igualdade de médias,

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_n = \mu \quad \text{(desconhecida)} \tag{2.5}$$

versus a hipótese alternativa

$$H_1 : \mu_I = \mu_1 = \ldots = \mu_k \neq \mu_{k+1} = \ldots = \mu_n = \mu_{II}, \tag{2.6}$$

onde k corresponde à posição onde ocorreu o *change-point*.

O modelo que descreve uma sequência de variáveis com *change-point* na média pode ser descrito como

$$X_t = \begin{cases} \mu_I + \epsilon_t, & \epsilon_t \sim N(0, \sigma^2), \quad t = 1, \ldots, k \\ \mu_{II} + \epsilon_t, & \epsilon_t \sim N(0, \sigma^2), \quad t = k+1, \ldots, n, \end{cases} \tag{2.7}$$

onde μ_I e μ_{II} representam a média antes e depois do *change-point*, respectivamente, e ϵ_t um ruído branco normal de média nula.

Para exemplificar uma mudança na média (Figura 2.1) foram geradas 100 observações de acordo com (2.7), com os parâmetros $\mu_I = 0$, $\mu_{II} = 3$ e $\sigma^2 = 1$ e estabelecendo-se que o *change-point* ocorreu na posição $k = 50$.

2.2.2 *Change-point* na variância

A detecção de *change-points* na variância foi estudada por Hsu (1977) através de dois métodos com a construção de testes estatísticos, um baseado no *Locally Most Powerful Test* e outro baseado no método de somas cumulativas (CUSUM). Este último também foi usado por Inclán & Tiao (1994). Inclán (1993) utilizou procedimentos bayesianos para avaliar a existência de várias mudanças na variância e Chen & Gupta (1997) utilizaram a

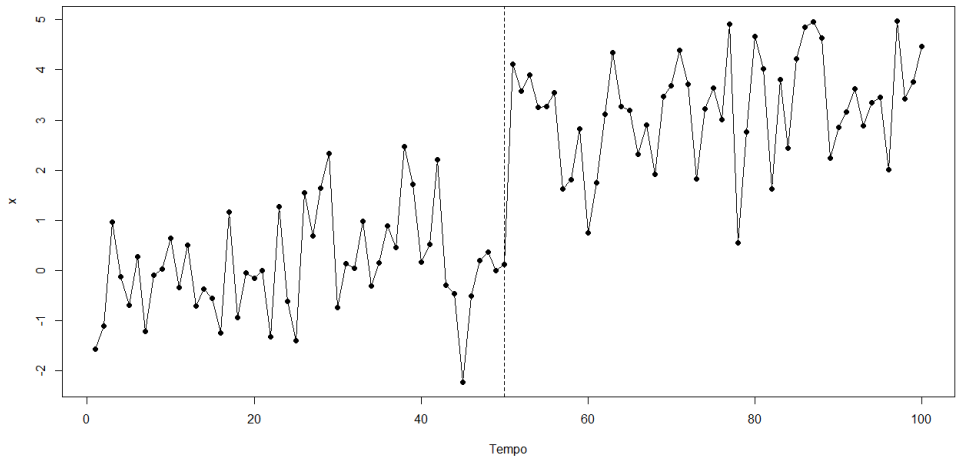


Figura 2.1: Mudança na média numa sequência de observações normais e independentes.

abordagem informacional. Recentemente, Zhao *et al.* (2010) utilizaram o teste da razão de verosimilhanças para mudanças na variância de processos estocásticos lineares.

Para se determinar se existe um *change-point* na variância terá de ser testada a hipótese nula, considerando, a igualdade de médias, $\mu_1 = \mu_2 = \ldots = \mu_n$,

$$H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_n^2 = \sigma^2 \quad \text{(desconhecida)} \tag{2.8}$$

versus a hipótese alternativa

$$H_1 : \sigma_I^2 \sigma_1^2 = \ldots = \sigma_k^2 \neq \sigma_{k+1}^2 = \ldots = \sigma_n^2 = \sigma_{II}^2. \tag{2.9}$$

Uma série de observações com *change-point* na variância pode ser descrita pelo modelo

$$X_t = \begin{cases} \mu + \epsilon_t^I, & \epsilon_t^I \sim N(0, \sigma_I^2), \quad t = 1, \ldots, k \\ \mu + \epsilon_t^{II}, & \epsilon_t^{II} \sim N(0, \sigma_{II}^2), \quad t = k+1, \ldots, n, \end{cases} \tag{2.10}$$

onde σ_I^2 e σ_{II}^2 representam a variância antes e depois do *change-point*, respectivamente.

Na Figura 2.2 encontra-se um exemplo de uma série temporal com *change-point* na variância. Foram geradas 100 observações de acordo com (2.10), onde o *change-point* foi definido em $k = 50$ e os valores usados para os parâmetros foram $\mu = 0$, $\sigma_I^2 = 1$ e $\sigma_{II}^2 = 4$.

2.2.3 *Change-point* na média e na variância

Em algumas situações pode existir um *change-point* na média e na variância, simultaneamente. Este problema não tem sido muito abordado e só em estudos mais recentes

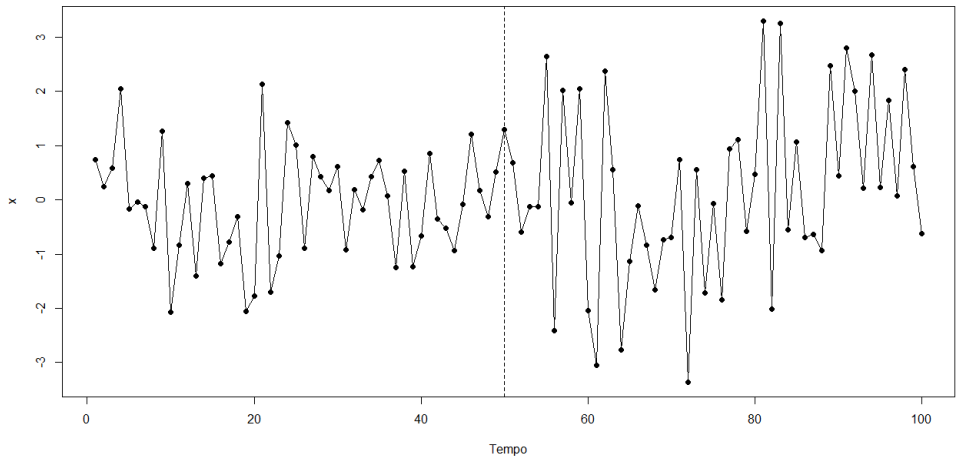


Figura 2.2: Mudança na variância numa sequência de observações normais e independentes.

se podem encontrar exemplos, como Chen & Gupta (1999) com a utilização da abordagem informacional e Hawkins & Zamba (2005) com a utilização do teste da razão de verosimilhanças.

No caso em que se pretende determinar se ocorreu um *change-point* na média e na variância, simultaneamente, terá de ser testada a hipótese nula

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_n = \mu \quad \wedge \quad \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_n^2 = \sigma^2 \tag{2.11}$$

versus a hipótese alternativa

$$H_1 : \mu_1 = \ldots = \mu_I = \mu_k \neq \mu_{k+1} = \ldots = \mu_n = \mu_{II}$$

$$\wedge \tag{2.12}$$

$$\sigma_I^2 = \sigma_1^2 = \ldots = \sigma_k^2 \neq \sigma_{k+1}^2 = \ldots = \sigma_n^2 = \sigma_{II}^2.$$

O modelo que traduz uma situação de mudança na média e variância, simultaneamente, pode ser descrito como

$$X_t = \begin{cases} \mu_I + \epsilon_t^I, & \epsilon_t^I \sim N(0, \sigma_I^2), \quad t = 1, \ldots, k \\ \mu_{II} + \epsilon_t^{II}, & \epsilon_t^{II} \sim N(0, \sigma_{II}^2), \quad t = k+1, \ldots, n, \end{cases} \tag{2.13}$$

onde μ_I e σ_I^2 representam a média e a variância antes do *change-point* e μ_{II} e σ_{II}^2 a média e a variância depois do *change-point*.

Um exemplo de uma série com *change-point* na média e na variância, em simultâneo, pode ser observado na Figura 2.3. Os valores definidos para os parâmetros foram $\mu_I = 0$,

$\sigma_I^2 = 1$, $\mu_{II} = 3$ e $\sigma_{II}^2 = 4$ e a mudança ocorreu em $k = 50$.

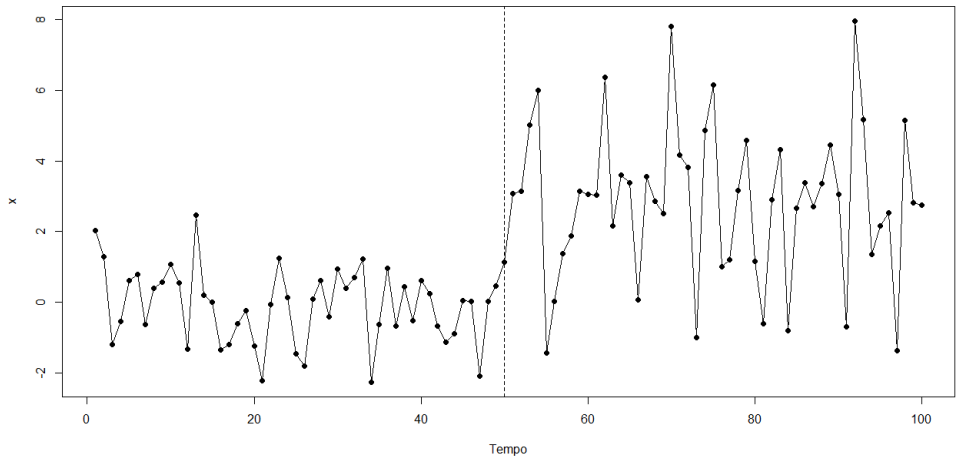


Figura 2.3: Mudança simultânea na média e na variância numa sequência de observações normais e independentes.

2.2.4 *Change-point* relativo a um modelo de regressão linear

Um modelo de regressão linear com mudança na intersepção e/ou no declive é outro tipo de *change-point* que se pode encontrar em várias áreas de estudo. Antes da introdução da hipótese da existência de *change-points* no estudo de modelos de regressão, surgiam problemas de incapacidade de se estabelecer um modelo para alguns conjuntos de dados observados, pois se o comportamento do conjunto de dados muda a partir de um determinado ponto, um só modelo de regressão não consegue explicar devidamente os dados. Quandt (1958, 1960) derivou o teste da razão de verosimilhanças e Ferreira (1975) e Kim (1991) estudaram as mudanças num modelo de regressão através da abordagem bayesiana. Muito recentemente, Beaulieu *et al.* (2012) e Chen & Gupta (2012) utilizaram a abordagem informacional para estudar o mesmo tipo mudanças.

A mudança nos coeficientes do modelo de uma regressão linear pode ocorrer apenas no coeficiente correspondente à intersepção ou no coeficiente de intersepção e no declive.

No caso em que se estuda a mudança apenas no coeficiente relativo à intersepção a hipótese nula que se pretende testar, assumindo-se a igualdade do declive $\beta_{1,1} = \beta_{1,2} = \ldots = \beta_{1,n} = \beta_1$, é a igualdade dos coeficientes relativos à intersepção,

$$H_0 : \beta_{0,1} = \beta_{0,2} = \ldots = \beta_{0,n} = \beta_0 \tag{2.14}$$

versus a hipótese alternativa

$$H_1 : \beta_0^I = \beta_{0,1} = \ldots = \beta_{0,k} \neq \beta_{0,k+1} = \ldots = \beta_{0,n} = \beta_0^{II}. \tag{2.15}$$

O modelo com mudança na intersepção pode ser expresso por

$$X_t = \begin{cases} \beta_0^I + \beta_1 t + \epsilon_t, & \epsilon_t \sim N(0, \sigma^2), \quad t = 1, \ldots, k \\ \beta_0^{II} + \beta_1 t + \epsilon_t, & \epsilon_t \sim N(0, \sigma^2), \quad t = k+1, \ldots, n, \end{cases} \tag{2.16}$$

onde β_0^I e β_0^{II} representam os coeficientes de intersepção antes e depois do *change-point*, respectivamente.

Um exemplo de *change-point* apenas na intersepção pode ser observado na Figura 2.4. Para este caso, consideraram-se os valores para os parâmetros $\beta_0^I = 0$, $\beta_0^{II} = 5$, $\beta_1 = 0,1$ e $\sigma^2 = 4$. Foram geradas 100 observações em que o *change-point* ocorre em $k = 50$.

No entanto, caso se pretenda estudar mudanças no coeficientes de intersepção e no declive, simultaneamente, a hipótese nula a testar será a igualdade dos coeficientes de intersepção e declive,

$$H_0 : \beta_{0,1} = \beta_{0,2} = \ldots = \beta_{0,n} = \beta_0 \quad \wedge \quad \beta_{1,1} = \beta_{1,2} = \ldots = \beta_{1,n} = \beta_1 \tag{2.17}$$

versus a hipótese alternativa

$$H_1 : \beta_0^I = \beta_{0,1} = ... = \beta_{0,k} \neq \beta_{0,k+1} = \ldots = \beta_{0,n} = \beta_0^{II}$$

$$\wedge \tag{2.18}$$

$$\beta_0^I = \beta_{1,1} = ... = \beta_{1,k} \neq \beta_{1,k+1} = \ldots = \beta_{1,n} = \beta_0^{II}.$$

Por sua vez, o modelo com alteração na intersepção e no declive é dado por

$$X_t = \begin{cases} \beta_0^I + \beta_1^I t + \epsilon_t, & \epsilon_t \sim N(0, \sigma^2), \quad t = 1, \ldots, k \\ \beta_0^{II} + \beta_1^{II} t + \epsilon_t, & \epsilon_t \sim N(0, \sigma^2), \quad t = k+1, \ldots, n, \end{cases} \tag{2.19}$$

em que β_0^I e β_1^I são os coeficientes do modelo de regressão antes do *change-point* e β_0^{II} e β_1^{II} do modelo depois do *change-point*.

A Figura 2.5 ilustra um exemplo de mudança nos coeficientes relativos à intersepção e ao declive, simultaneamente. Considerou-se $\beta_0^I = 6$, $\beta_0^{II} = 0$, $\beta_1^I = 0,1$ e $\beta_1^{II} = 0,3$. Foram geradas 100 observações de acordo com (2.19), com uma média nula e $sigma^2 = 4$, e o *change-point* foi estabelecido em $k = 50$.

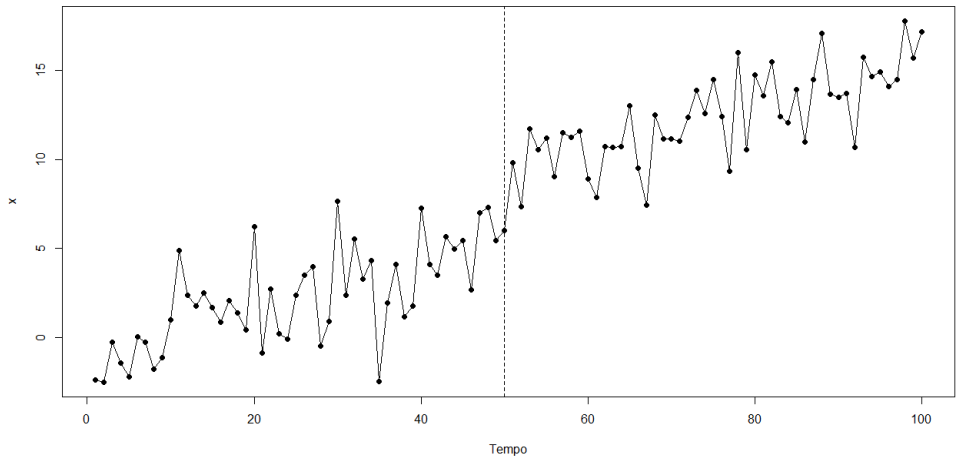


Figura 2.4: Mudança na intersepção de um modelo de regressão linear numa sequência de observações normais e independentes.

2.3 Metodologias

Para cada tipo de *change-point* existem vários métodos para a análise dos mesmos, sendo os mais utilizados e frequentes na literatura o teste da razão de verosimilhanças, o método de somas cumulativas e a abordagem informacional, numa abordagem paramétrica. Ainda se encontram métodos num contexto bayesiano e num contexto não-paramétrico. Nesta secção pretende-se fazer uma breve revisão da literatura sobre as referidas metodologias, sendo apenas explicada com maior detalhe a abordagem informacional no Capítulo 3, pois é a utilizada na aplicação da análise de *change-points* aos dados de qualidade da água.

Chernoff & Zacks (1964) derivaram um estimador bayesiano para a média, utilizando uma distribuição uniforme como *priori* e Gardner (1969) e Sen & Srivastava (1975) derivaram a distribuição assimptótica para problemas de mudança da média de variáveis aleatórias normalmente distribuídas. Por sua vez, Ferreira (1975) estudou mudanças num modelo de regressão, Chin Choy & Broemeling (1980) aplicaram a mesma metodologia fazendo uma generalização do trabalho anterior, e Chalton & Troskie (1999) estudaram o mesmo problema mas para um modelo de regressão múltipla com erros autocorrelacionados. Chen & Gupta (2012) utilizaram também a abordagem bayesiana para mudanças num modelo de regressão linear, num modelo de regressão linear múltipla e ainda para o modelo Gama e para a função risco.

Relativamente ao teste da razão de verosimilhanças, Hawkins (1977) e Worsley (1979) derivaram a distribuição sob a hipótese nula para uma mudança na média, nos casos de variância conhecida e desconhecida. Srivastava & Worsley (1986) aplicaram o teste da razão de verosimilhanças para detectar mudanças nos vectores de médias e aproximaram a

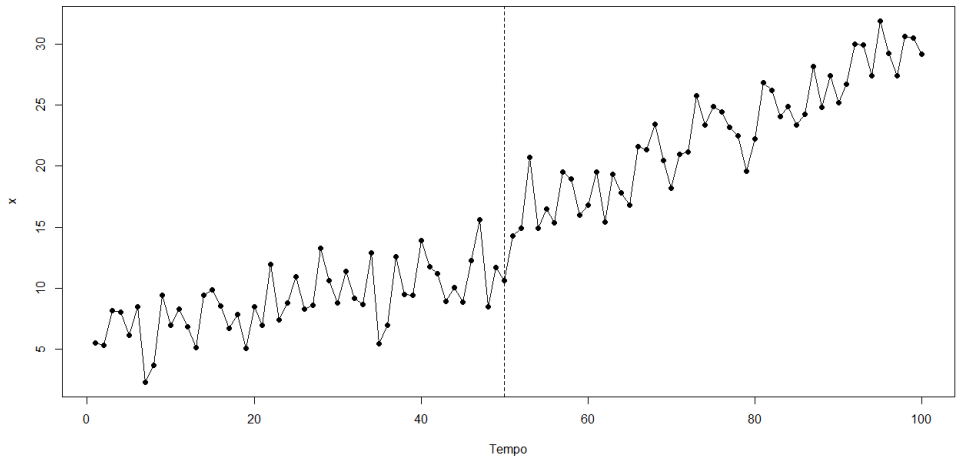


Figura 2.5: Mudança na intersepção e no declive de um modelo de regressão linear numa sequência de observações normais e independentes.

distribuição sob a hipótese nula tomando por base a desigualdade de Bonferroni. James *et al.* (1992) obtiveram aproximações assimptóticas para o teste da razão de verosimilhanças e regiões de confiança para mudanças na média para dados normais multivariados. Também em Chen & Gupta (2012) foi apresentada uma metodologia para mudanças na média e na variância, simultaneamente, e na variância, para modelos multivariados normais e para o modelo exponencial.

O método de somas cumulativas (CUSUM) foi inicialmente desenvolvido por Page (1954, 1955, 1957) para estudar mudanças na média. Hsu (1977) baseado na mesma técnica estudou a detecção de *change-points* na variância, assim como Inclán & Tiao (1994). Por sua vez Pettitt (1980) investigou o problema de *change-points* num modelo binomial e Worsley (1983) estudou a potência dos testes baseados neste modelo.

No que respeita aos métodos não-paramétricos, Hájek (1962) construiu testes de *ranks* para pontos de mudança num modelo de regressão assimptoticamente potentes e Milton (1965) desenvolveu também um método baseado nos *ranks* das probabilidades com aplicações em diversas áreas. Adichie (1967) estudou pontos de mudança num modelo de regressão através do teste de Wilcoxon e de um teste baseado em *scores*. Bhattacharya & Johnson (1968) estudaram duas versões do problema de mudanças de nível.

2.4 Múltiplos *change-points*

As abordagens à problemática da análise de *change-point*, tanto ao nível dos métodos como dos tipos de *change-points* que se pretendem determinar, incidem, na sua maioria, sobre o caso de apenas existir uma única mudança ao longo da sequência de observações,

mas essa hipótese pode ser muitas vezes irrealista.

O problema de múltiplos *change-points* foi abordado por Inclán & Tiao (1994), que utilizam um algoritmo iterativo denominado *iterated cumulative sums of squares* (ICSS) para estudar o problema de múltiplos *change-points* na variância, considerando-se uma sequência de observações independentes, na área das Finanças. Chen & Gupta (1995) derivaram a distribuição assimptótica do procedimento de máxima verosimilhança para testar mudanças simultâneas na média e na covariância sob um modelo multivariado Gaussiano. Srivastava & Worsley (1986) deduziram a estatística de teste e a distribuição aproximada para múltiplas mudanças num vector de médias para uma sequência de vectores aleatórios e gaussianos, utilizando o teste da razão de verosimilhanças. Gerard-Marchant *et al.* (2008) propuseram quatro algoritmos iterativos para detectar múltiplos *change-points* com base em diferentes métodos, que foram implementados em dados de fluxo do rio Flint do Sudoeste da Geórgia.

Os métodos referidos, assim como outros estudos realizados, utilizam procedimentos idênticos, envolvendo processos iterativos, mas específicos para um determinado método, bem como para um determinado tipo de *change-point*, limitando as suas utilizações.

O procedimento de segmentação binária foi proposto por Vostrikova (1981), que provou a sua consistência. Este procedimento de segmentação binária tem sido largamente utilizado para detectar múltiplos *change-points*. Por exemplo, Chen (1998) utilizou a segmentação binária para procurar a existência de vários *change-points* no volume mensal de vendas da Bolsa de Valores de Boston, e tem a vantagem de detectar simultaneamente o número de *change-points* e a sua localização, economizanndo muito tempo computacional e pode ser utilizado para detectar *change-points* de vários tipos, utilizando qualquer metodologia.

O procedimento mencionado pode ser descrito de forma sucinta. Primeiro detecta-se uma única mudança considerando a sequência de observações completa. Se não existir nenhum *change-point* é aceite a hipótese de não existirem mudanças na série em estudo. Se existir um *change-point*, então este divide a sequência original de observações em duas subsequências. Para cada subsequência, inicia-se o procedimento, testando se existe alguma mudança em cada uma e continua-se até não ser detectado nenhum *change-point*, em cada uma das subsequências que vão sendo criadas.

O algoritmo para detectar múltiplos *change-points*, através do procedimento de segmentação binária, pode ser definido através dos seguintes passos:

Passo 1: Testar a hipótese de não existir *change-point*, ou seja, testar a hipótese nula dada por (2.3) contra a hipótese de existir um *change-point*, ou seja, *versus* a seguinte hipótese alternativa

$$H_1 : \theta_1 = \ldots = \theta_k \neq \theta_{k+1} = \ldots = \theta_n, \tag{2.20}$$

onde k é a localização do único *change-point* neste passo. Se H_0 não for rejeitada pára-se o processo, concluindo-se que não existe *change-point*. Se H_0 é rejeitada, existe um *change-point* e prossegue-se para o Passo 2.

Passo 2: Testar se existe um *change-point* nas duas subsequências, antes e depois do *change-point* encontrado no Passo 1, separadamente.

Passo 3: Repetir o processo até não existirem subsequências com *change-points*.

As localizações dos *change-points* encontrados nos passos de 1 a 3 são denotadas por $\{\hat{k}_1, \hat{k}_2, \ldots, \hat{k}_q\}$ e o número total de *change-points* estimados é q. Sendo assim, com o método da segmentação binária apenas é necessário testar a hipótese de existir um único *change-point* e repetir o processo para cada subsequência, até a hipótese de não existirem mudanças não ser rejeitada.

2.5 Características de dados ambientais

A análise de *change-points* inclui conhecer o comportamento da variável em estudo ao longo do tempo. De um modo mais formal, o que se pretende é estudar uma série temporal que pode ser definida como uma sucessão de observações ordenadas no tempo, ou seja, um conjunto de observações $X_{t_1}, X_{t_2}, \ldots, X_{t_n}$ para todos os inteiros n e quaisquer pontos $t_1, t_2, \ldots, t_n$, em regra equidistantes, concretizações de um processo estocástico.

As sequências de dados ambientais observadas ao longo do tempo são muitas vezes complexas, tornando-se o processo de identificação de *change-points* difícil.

Nesta secção pretende-se analisar um conjunto de características presentes em séries temporais, em particular em séries de dados hidrometeorológicos, que têm influência na análise de *change-points*, com referências a alguns métodos e alternativas de análise. Todas as transformações efectuadas nos dados devem ser feitas com muito cuidado pois poderão eliminar comportamentos importantes dos dados, podendo originar alterações que impeçam a detecção de *change-points* existentes ou a aceitação de *change-points* não existentes (os denominados falsos *change-points*).

2.5.1 Não estacionaridade na média e/ou na variância

Para se definir um processo estacionário é necessário primeiro definir-se um processo estocástico.

Num contexto de séries temporais, diz-se que um processo estocástico é qualquer família ou colecção de variáveis aleatórias $X(t), t \in T$ em que T é um conjunto de índices representando o tempo.

O conjunto T é denominado espaço de parâmetros que poderá ser $\mathbb{R}$, $\mathbb{R}^+$, $\mathbb{N}$ ou $\mathbb{Z}$.

O conjunto S, contradomínio das variáveis $X(t)$, é denominado espaço de estados e poderá ser $\mathbb{R}$, $\mathbb{Z}$, $\mathbb{N}$, $\mathbb{R}^2$, etc.

Assim um processo estocástico $\{X(t); t \in \mathbb{R}\}$ diz-se estritamente estacionário se a distribuição conjunta de $(X(t_1), \ldots, X(t_n))$ é igual à distribuição conjunta de $(X(t_1 + \delta), \ldots, X(t_n + \delta))$ qualquer que seja o n-úplo $(t_1, \ldots, t_n)$ e para qualquer δ, ou seja, $F_{(X(t_1),\ldots,X(t_n))}(x_1, \ldots, x_n) = F_{(X(t_1+\delta),\ldots,X(t_n+\delta))}(x_1, \ldots, x_n)$ em todos os pontos $(x_1, \ldots, x_n)$ (Alpuim, 1998).

Um processo estocástico $\{X(t); t \in \mathbb{R}\}$ diz-se estacionário de segunda ordem ou fracamente estacionário se todos os momentos atá à segunda ordem de $(X(t_1), \ldots, X(t_n))$ existem e são iguais aos momentos correspondentes até à segunda ordem de $(X(t_1 + \delta), \ldots, X(t_n + \delta))$. Logo, num processo estacionário de segunda ordem:

- o valor médio não depende de t, i.e., $\mu(t) = \mu$;
- a variância não depende de t, i.e., $\sigma^2(t) = \sigma^2$;
- a covariância entre X_{t_1} e X_{t_2} depende apenas do desfasamento $t_2 - t_1$, i.e., $Cov[X(t_1), X(t_2)] = \gamma(|t_2 - t_1|)$.

A não estacionaridade de séries temporais pode depender da média não constante e/ou da variância não constante.

A tendência de uma série temporal identifica a inclinação, positiva ou negativa, que certas séries apresentam ao longo do tempo. Esta variação do conjunto de dados não necessita de ser constante, mas deverá ser sempre do mesmo sinal. A tendência ou inclinação pode ser consequência do facto dos valores observados dependerem de uma componente determinística, que é função monótona do tempo linear ou não linear.

Muitas das séries que não apresentam, à partida, uma média constante podem ser reduzidas à estacionaridade em relação à média retirando-lhes a tendência, podendo esta ser uma:

- tendência simples, i.e., $X_t = \mu_t + \epsilon_t$;
- tendência linear, i.e., $X_t = \beta_0 + \beta_1 t + \epsilon_t$;
- tendência polinomial, i.e., $X_t = \beta_0 + \beta_1 t + \ldots + \beta_p t^p + \epsilon_t$

onde $E(\epsilon_t) = 0$ e $Var(\epsilon_t) = \sigma^2$.

No que respeita a estabilizar a variância pretende-se determinar o tipo de transformações que o permita fazer.

Na prática, é usual considerar as transformações paramétricas sugeridas por Box & Cox (1964), sendo estas dadas por

$$Y_t = X_t^{(\lambda)} = \begin{cases} (X_t^{\lambda} - 1)/\lambda, & \text{se} \quad \lambda \neq 0 \\ \ln X_t, & \text{se} \quad \lambda = 0 \end{cases} \tag{2.21}$$

onde $\lambda \in [-1, 1]$.

A Tabela 2.1 apresenta os valores mais comuns para λ e as transformações correspondentes.

Tabela 2.1: Transformações de Box & Cox.

Valores de λ	Transformação
-1	$1/X_t$
-0.5	$1/\sqrt{X_t}$
0	$\ln X_t$
0.5	$\sqrt{X_t}$
1	X_t

Algumas destas transformações apenas estão definidas para séries de valores positivos, mas pode-se, no entanto, encontrar uma constante c tal que $X_t + c > 0$ e só depois aplicar as transformações.

Note-se que, quando a não estacionaridade é devida à média e à variância, deve-se estabilizar em primeiro lugar a variância e só depois a média (Alpuim, 1998).

2.5.2 Sazonalidade

Alguns fenómenos apresentam uma variabilidade periódica, a qual se designa por componente sazonal. Tal pode corresponder a um aumento/decréscimo que ocorre regularmente em determinados períodos do ano, originando oscilações que se repetem.

Muitos dados ambientais são recolhidos mensalmente, tendo usualmente a série temporal associada uma forte componente sazonal, podendo esta ser explicada, por exemplo, por causas naturais, tais como as estações do ano.

Uma abordagem simples e sugerida por Jarusková (1997), consiste em subtrair para cada mês a média desse mesmo mês, ou seja, para os dados relativos ao mês de Janeiro subtrair a média global de Janeiro, para os dados relativos ao mês de Fevereiro subtrair a média global de Fevereiro e, assim, sucessivamente. Esta abordagem é mais adequada para séries sem tendência evidente.

Um método alternativo é descrito por Gonçalves & Alpuim (2011). A componente sazonal, s_t, toma doze valores diferentes, $\lambda_i, i = 1, \ldots, 12$, cada um associado a um mês

e expressam o desvio positivo ou negativo dos dados devido ao efeito do mês. Este efeito é usualmente descrito com a ajuda de onze variáveis mudas (*dummy*), e a soma dos coeficientes deve perfazer um total de zero, considerando um modelo linear com termo independente. A componente sazonal é então representada pela combinação linear de onze variáveis explicativas, $s_{t,i}$, definidas por

$$s_{t,i} = \begin{cases} 1, & \text{se os dados no tempo } t \text{ correspondem ao mês } i \\ -1, & \text{se os dados no tempo } t \text{ correspondem ao mês } 12 \\ 0, & \text{caso contrário.} \end{cases} \tag{2.22}$$

A componente sazonal relativa ao mês de Dezembro pode ser calculada a partir dos restantes meses através da fórmula

$$\hat{\lambda}_{12} = -\sum_{i=1}^{11} \hat{\lambda}_i. \tag{2.23}$$

A escolha do mês de Dezembro como combinação linear dos outros meses é arbitrária e qualquer mês pode ser usado para esta finalidade. Aplica-se, por fim, o modelo de regressão múltipla que fornece estimadores óptimos para os parâmetros.

A série dos dados pode ainda apresentar simultaneamente média e sazonalidade, ou seja,

$$X_t = \mu + s_t + \epsilon_t, t \in \mathbb{N}, \tag{2.24}$$

devendo-se estimar a média e os coeficientes de sazonalidade ao mesmo tempo (Alpuim, 1998).

2.5.3 Dependência

Uma característica comum das séries ambientais é a dependência temporal das observações (correlação), principalmente se a escala de tempo é mensal ou menor. A presença de correlação positiva forte cria padrões nas séries temporais que podem ser facilmente confundidos com os *change-points*, sobretudo se a magnitude do *change-point* é pequena (Jarušková, 1997).

Assim, pode-se facilmente interpretar mal as variações destas séries temporais e identificar mudanças aparentes, mesmo que não existam. Este é um problema da detecção de *change-points*, pois a maioria das técnicas foram desenvolvidas para observações independentes. Na presença de correlação, o risco de uma falsa detecção tende a aumentar e o poder de detecção a diminuir (Beaulieu *et al.*, 2012).

O efeito de correlação na detecção de *change-points* em séries temporais tem vindo a ser estudada. Henderson (1986) e Tang & MacNeil (1993) propuseram abordagens que

têm em conta a autocorrelação quando aplicado um teste para mudanças na média. El-Shaarawi & Esterby (1982) abordaram a inferência sobre *change-points* num modelo de regressão, considerando um processo autoregressivo de ordem um para os erros. Antoch *et al.* (1997) mostraram que se as variáveis não são independentes, mas formam uma sequência *autoregressive moving average* (ARMA), então os valores críticos assimptóticos, quando utilizada a abordagem CUSUM, têm de ser multiplicados por $\sqrt{(2\pi f(0)/\gamma)}$, em que γ é a variância e $f()$ denota a densidade específica do processo ARMA correspondente.

Seidel & Lanzante (2004) integraram a autocorrelação na formulação do SIC para *change-points* em modelos de regressão linear. Esta abordagem permite ter em conta na análise um modelo autoregressivo de primeira ordem, AR (1), ou um modelo autoregressivo de segunda ordem, AR (2).

Mais recentemente, Lund *et al.* (2007) desenvolveram um método para a detecção de *change-points* na intersepção de um modelo de regressão linear para séries com características de autocorrelação e periodicidade. Wang (2008) estendeu o teste t e o teste F penalizados para detectar mudanças na média, tendo em conta a autocorrelação de primeira ordem e Robbins *et al.* (2011) propuseram um teste baseado no CUSUM, ajustado para a autocorrelação.

2.5.4 Distribuição não Normal

A maioria dos métodos de detecção de *change-points* assume que as variáveis seguem uma distribuição normal. Contudo, esse pressuposto não se verifica em todas as séries ambientais.

O procedimento utilizado em muitos estudos para resolver a violação deste pressuposto é a transformação das observações, mas muitas vezes, a interpretação dos pontos de mudança fica comprometida. Jarušková (1997) realizou um estudo relativo a uma floresta das montanhas de Erzgebirge, que foi fortemente afectada pelas chuvas ácidas. Aplicou a metodologia à série transformada e os testes detectaram a mudança na média, mas não na variância dos dados transformados, concluindo assim que a forma original permaneceu a mesma mas, a característica escala foi modificada.

Outro procedimento que reduz a assimetria dos dados é o denominado nivelamento, estudar médias anuais em vez de médias mensais, por exemplo, torna o problema da assimetria não tão grave. Contudo, com este procedimento reduz-se a dimensão dos dados e elimina-se o comportamento mensal, que mesmo podendo trazer entraves na aplicação da metodologia, pode ser importante na caracterização da série.

No seguimento da problemática da não normalidade, alguns autores desenvolveram técnicas para a detecção de pontos de mudança nos parâmetros de distribuições diferentes. Chen & Gupta (2012) apresentaram a abordagem informacional e a abordagem bayesiana

para a distribuição Gama e o procedimento da razão de verosimilhanças e a abordagem informacional para a distribuição Exponencial. Jarušková (2007) estudou a mudança nos três parâmetros da distribuição de Weibull, Jarušková & Rencová (2008) estudaram séries de temperatura, utilizando a distribuição *generalized extreme value* (GEV) e Zhao & Chu (2006) desenvolveram uma abordagem para detectar mudanças na contagem de furacões, sendo as contagens modeladas por uma distribuição de Poisson e a intensidade representada por uma distribuição Gama.

Por último, também poderão ser usadas abordagens não-paramétricas de modo a ultrapassar-se o problema da não normalidade das distribuições (Bhattacharya & Johnson, 1968).

Capítulo 3

Critério de Informação de Schwarz

No Capítulo 2 foram abordados, de uma forma sucinta, vários métodos de detecção de *change-points* e os diferentes tipos que podem ser encontrados em séries de dados observadas em diferentes áreas do conhecimento.

Neste capítulo será analisada com mais detalhe a denominada abordagem informacional, que é uma metodologia geral de selecção de modelos e que consiste em utilizar um critério de informação para identificar a posição desconhecida de um *change-point* num modelo, discriminando de entre os vários modelos, o que é mais verosímil para ajustar os dados, isto é, o que melhor descreve a série de dados. Uma das grandes vantagens da abordagem informacional é a de poder ser adaptada a diversas situações e não limitar a sua utilização apenas a um determinado tipo de *change-point*. Além disso, a utilização desta metodologia não exige um desempenho computacional pesado.

O critério de informação que será utilizado será o Critério de Informação de Schwarz.

3.1 Introdução

Akaike (1973) introduziu o *Akaike Information Criterion* (AIC) para selecção de modelos em Estatística. A formulação do AIC para seleccionar um modelo entre M modelos pode ser expressa por

$$AIC_j = -2 \ln L(\hat{\Theta}_j) + 2p_j, \quad j = 1, 2, \ldots, M, \tag{3.1}$$

onde $L(\hat{\Theta}_j)$ é a função de máxima verosimilhança para o modelo j e p_j é o número de parâmetros que têm de ser estimados para o modelo j. O modelo que minimiza o AIC é considerado o modelo mais apropriado.

Este critério tem tido um papel muito importante no desenvolvimento da análise estatística, particularmente em séries temporais, na análise de *outliers* (Kitagawa, 1979) e robustez, análise de regressão e na análise multivariada (Bozdogan *et al.*, 1994). Vários

autores introduziram novos critérios de informação tendo por base o AIC, como Bozdogan (1987) e Rao & Wu (1989).

Uma das modificações do AIC é o Critério de Informação de Schwarz, proposto por Schwarz (1978). O SIC é definido como

$$SIC_j = -2 \ln L(\hat{\Theta}_j) + p_j \ln n, \quad j = 1, 2, \ldots, M, \tag{3.2}$$

onde n é o número de observações. Este critério baseia-se na função de máxima verosimilhança de um determinado modelo penalizado pelo número de parâmetros que são estimados. Também o modelo que minimiza o SIC é considerado o modelo mais apropriado, representando o melhor compromisso entre a parcimónia (poucos parâmetros) e o bom ajustamento (resíduos pequenos).

Aparentemente, a diferença entre o AIC e o SIC é o termo de penalização, em vez de $2p$ é $p \ln n$. Contudo, o SIC dá uma estimativa assimptoticamente consistente da ordem do verdadeiro modelo (Chen & Gupta, 2012).

Em suma, a abordagem informacional com a utilização, neste caso, do Critério de Informação de Schwarz consiste em identificar o tempo mais provável para um *change-point* através da identificação do modelo que minimiza o SIC, que é considerado o modelo mais apropriado, sendo este comparado com o modelo sem nenhum ponto de mudança. Assim, existirão dois modelos, um correspondente à hipótese nula (2.3) e o outro à hipótese alternativa (2.20).

3.2 Formulação dos modelos

A formulação dos diferentes modelos do SIC relativos a cada hipótese, nula e alternativa, e a cada tipo de *change-point* será feita com base nos pressupostos de normalidade e independência das observações.

Apenas para o caso de mudança na média e na variância, em simutâneo, será feita a dedução dos modelos mais detalhadamente, pois será o caso utilizado na aplicação prática do Capítulo 4, sendo a dedução para os restantes modelos similar.

3.2.1 *Change-point* na média e na variância

Pretende-se determinar os modelos para a existência de *change-point* na média e na variância, simultaneamente. Sob a hipótese nula (2.11), os estimadores de máxima verosimilhança para μ e σ^2 são

$$\hat{\mu} = \overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{3.3}$$

e

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2, \tag{3.4}$$

respectivamente. Em seguida, denotando o SIC sob a hipótese nula (2.11) por $SIC(n)$, tem-se

$$SIC(n) = -2\ln\, L_0(\hat{\mu}, \hat{\sigma}^2) + 2\ln n, \tag{3.5}$$

e a função de máxima verosimilhança

$$L_0(\hat{\mu}, \hat{\sigma}^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\left(\frac{-(X_i - \hat{\mu})^2}{2\hat{\sigma}^2}\right). \tag{3.6}$$

O factor 2 da segunda parcela da equação (3.5) representa o número de parâmetros que são necessários estimar: a média e a variância. Atendendo às equações (3.5) e (3.6) obtém-se

$$SIC(n) = -2\sum_{i=1}^{n}\left\{\ln\left[\frac{1}{\sqrt{2\pi\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right]}} \exp\left(\frac{-(X_i - \overline{X})^2}{2\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right]}\right)\right]\right\} + 2\ln n \tag{3.7}$$

e fazendo-se algumas simplificações tem-se

$$SIC(n) = n\ln 2\pi + n\ln\sum_{i=1}^{n}(X_i - \overline{X})^2 + n + (2-n)\ln n. \tag{3.8}$$

Sob a hipótese alternativa de haver mudança na média e na variância, (2.12), têm de ser estimados quatro parâmetros: duas médias e duas variâncias, antes e depois do *change-point*. O SIC sob a hipótese alternativa é denotado por $SIC(k)$ e pode ser obtido através de

$$SIC(k) = -2\ln\, L_1(\hat{\mu}_I, \hat{\mu}_{II}, \hat{\sigma}_I^2, \hat{\sigma}_{II}^2) + 4\ln n. \tag{3.9}$$

A função de máxima verosimilhança é dada por

$$L_1(\hat{\mu}_I, \hat{\mu}_{II}, \hat{\sigma}_I^2, \hat{\sigma}_{II}^2) = \prod_{i=1}^{k}\left\{\frac{1}{\sqrt{2\pi\hat{\sigma}_I^2}} \exp\left(\frac{-(X_i - \hat{\mu}_I)^2}{2\hat{\sigma}_I^2}\right)\right\} \prod_{i=k+1}^{n}\left\{\frac{1}{\sqrt{2\pi\hat{\sigma}_{II}^2}} \exp\left(\frac{-(X_i - \hat{\mu}_{II})^2}{2\hat{\sigma}_{II}^2}\right)\right\}. \tag{3.10}$$

Considerando as equações (3.9) e (3.10) e simplificando-as obtém-se

$$SIC(k) = n\ln 2\pi + k\ln\hat{\sigma}_I^2 + (n-k)\ln\hat{\sigma}_{II}^2 + n + 4\ln n, \tag{3.11}$$

onde

$$\hat{\sigma}_I^2 = \frac{1}{k}\sum_{i=1}^{k}(X_i - \overline{X}_I)^2, \tag{3.12}$$

$$\hat{\sigma}_{II}^2 = \frac{1}{(n-k)}\sum_{i=k+1}^{n}(X_i - \overline{X}_{II})^2, \tag{3.13}$$

$$\overline{X}_I = \frac{1}{k}\sum_{i=1}^{k} X_i \tag{3.14}$$

e

$$\overline{X}_{II} = \frac{1}{n-k}\sum_{i=k+1}^{n} X_i. \tag{3.15}$$

Aplicações do Critério de Informação de Schwarz para mudanças na média e na variância, simultaneamente, podem ser encontradas em Chen & Gupta (1999), onde se estuda a resistência à tracção e o tráfego em Illinois.

3.2.2 *Change-point* na média

Para a formulação do modelo para pontos de mudança na média sob a hipótese nula (2.5) e considerando a variância desconhecida, o $SIC(n)$ é definido por (3.8) pois, apesar da hipótese nula apenas considerar a igualdade de médias, é assumida a igualdade de variâncias.

Sob a hipótese alternativa (2.6), o $SIC(k)$ é definido como

$$\begin{aligned} SIC(k) &= -2\ln L_1(\hat{\mu}_I, \hat{\mu}_{II}, \hat{\sigma}^2) + 3\ln n = \\ &= n\ln 2\pi + n\ln\left[\sum_{i=1}^{k}(X_i - \overline{X}_I)^2 + \sum_{i=k+1}^{n}(X_i - \overline{X}_{II})^2\right] + n + (3-n)\ln n \end{aligned} \tag{3.16}$$

onde $L_1(\hat{\mu}_I, \hat{\mu}_{II}, \hat{\sigma}^2)$ é a função de máxima verosimilhança sob H_1.

Um exemplo da utilização do Critério de Informação de Schwarz para estudar alterações na média pode ser encontrado em Beaulieu *et al.* (2012), onde é utilizado como aplicação no estudo de mudanças na média da captação de dióxido de carbono pela terra em Mauna Loa, Havai.

3.2.3 *Change-point* na variância

Para o caso de mudança na variância, considerando-se a hipótese nula (2.8) e a média desconhecida, o $SIC(n)$ é dado novamente por (3.8) e o $SIC(k)$, considerando a hipótese

alternativa (2.9), é definido por

$$SIC(k) = -2\ln\, L_1(\hat{\mu}, \hat{\sigma}_I^2, \hat{\sigma}_{II}^2) + 3\ln n = \\ = n\ln 2\pi + k\ln\hat{\sigma}_I^2 + (n-k)\ln\hat{\sigma}_{II}^2 + n + 3\ln n. \tag{3.17}$$

Um exemplo da utilização do SIC para estudos de mudança na variância pode ser encontrado em Chen & Gupta (1997), onde se estudam os preços das acções nos Estados Unidos da América.

3.2.4 *Change-point* relativo a um modelo de regressão linear

Nesta secção apresenta-se o caso de alterações nos coeficientes de um modelo de regressão linear. Estes coeficientes são estimados segundo o método de máxima verosimilhança. Considerando-se apenas mudança no coeficiente de intersepção, pretende-se testar a hipótese nula (2.14) em que o $SIC(n)$ é definido por

$$SIC(n) = -2\ln\, L_0(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2) + 3\ln n = \\ = n\ln 2\pi + n\ln\Big[\sum_{i=1}^{n}(X_i - \hat{\beta}_0 - \hat{\beta}_1 i)^2\Big] + n + (3-n)\ln n, \tag{3.18}$$

e o $SIC(k)$ sob a hipótese alternativa (2.15) definido por

$$SIC(k) = -2\ln\, L_1(\hat{\beta}_0^I, \hat{\beta}_0^{II}, \hat{\beta}_1, \hat{\sigma}^2) + 4\ln n =$$

$$= n\ln 2\pi + n\ln\Big[\sum_{i=1}^{k}(X_i - \hat{\beta}_0^I - \hat{\beta}_1 i)^2 + \sum_{i=k+1}^{n}(X_i - \hat{\beta}_0^{II} - \hat{\beta}_1 i)^2\Big] + n + (4-n)\ln n. \tag{3.19}$$

Caso se pretenda estudar se existe um *change-point* no coeficiente de intersepção e ainda no declive, simultaneamente, terá de ser testada a hipótese nula (2.17) e o $SIC(n)$ é dado por (3.18). Por sua vez, o $SIC(k)$ correspondente à hipótese alternativa (2.18), é dado por

$$SIC(k) = -2\ln\, L_1(\hat{\beta}_0^I, \hat{\beta}_0^{II}, \hat{\beta}_1^I, \hat{\beta}_1^{II}, \hat{\sigma}^2) + 5\ln n =$$

$$= n\ln 2\pi + n\ln\Big\{\sum_{i=1}^{k}(X_i - \hat{\beta}_0^I - \hat{\beta}_1^I i)^2 + \sum_{i=k+1}^{n}(X_i - \hat{\beta}_0^{II} - \hat{\beta}_1^{II} i)^2\Big\} + n + (5-n)\ln n, \tag{3.20}$$

em que $\hat{\beta}_0^I$ e $\hat{\beta}_1^I$ são as estimativas dos coeficientes antes do *change-point* e $\hat{\beta}_0^{II}$ e $\hat{\beta}_1^{II}$ depois do mesmo.

Um exemplo da aplicação destes casos pode ser encontrado em Chen & Gupta (2012)

com a aplicação dos mesmos a dados relativos ao mercado de acções.

3.3 Selecção do modelo

A posição mais provável para um ponto de mudança é aquela que minimiza o valor de $SIC(k)$. Note-se que para ser possível obter os estimadores de máxima verosimilhança, apenas se podem detectar mudanças localizadas entre a segunda e a $(n-2)$-ésima posição. Então, a posição do *change-point* é estimada por $\hat{k}$ tal que

$$SIC(\hat{k}) = \min_{2\leq k\leq n-2} SIC(k). \tag{3.21}$$

Chen & Gupta (1997) apresentaram um teorema e a sua prova, que afirma que $\hat{k}$ estimado de acordo com (3.21) é consistente para o verdadeiro *change-point* k_0. Algumas propriedades de $SIC(k)$ foram apresentadas por Chen & Gupta (1999), nomeadamente a função característica, a média e a variância da estatística de teste $S = SIC(k)$.

O modelo com um ponto de mudança, $SIC(k)$, é seleccionado se

$$SIC(k) < SIC(n). \tag{3.22}$$

Caso contrário, o modelo sem nenhum ponto de mudança, $SIC(n)$, é mais provável. Os critérios de informação, como o SIC, apresentam a vantagem de não ser necessário recorrer a uma distribuição da estatística de teste, nem determinar níveis de significância quando apenas se pretende identificar potenciais *change-points* numa análise exploratória inicial. Contudo, caso os valores $SIC(k)$ e $SIC(n)$ estejam muito próximos é questionável se existe realmente um *change-point* ou essa diferença deve-se a flutuações inerentes aos dados. De modo a tirar-se uma conclusão com significância estatística, Chen & Gupta (1997) acrescentaram à regra de decisão um valor crítico.

Então, rejeita-se a hipótese nula de não existirem *change-points* quando

$$\min_{2\leq k\leq n-2} SIC(k) + c_\alpha < SIC(n) \tag{3.23}$$

onde c_α e α têm a seguinte relação

$$1 - \alpha = P\Big[SIC(n) < \min_{2\leq k\leq n-2} SIC(k) + c_\alpha | H_0\Big]. \tag{3.24}$$

Para ser possível obter os valores críticos é necessário o conhecimento da distribuição sob a hipótese nula do $\min_{2\leq k\leq n-2} SIC(k)$, contudo, esta distribuição não é geralmente conhecida. Chen & Gupta (1999) apresentaram a distribuição assimptótica para a hipótese

nula de igualdade de médias e de variâncias, e obtiveram a fórmula aproximada para c_α:

$$c_\alpha \approx -2\ln n + \left\{ -\frac{1}{a(\ln n)} \ln\ln \left[1-\alpha+\exp\left(-2\exp\left[b(\ln n)\right]\right)\right]^{-1/2} + \frac{b(\ln n)}{a(\ln n)} \right\}^2, \quad (3.25)$$

onde $a(\ln n) = (2\ln\ln n)^{1/2}$ e $b(\ln n) = 2\ln\ln n + \ln\ln\ln n$.

Para diferentes níveis de significância α e diferentes tamanhos da amostra n, também determinaram valores de c_α, podendo esses valores ser observados na Tabela 3.1.

Tabela 3.1: Valores aproximados de c_α.

Tamanho	α			
n	0,010	0,025	0,050	0,100
7	35,699	19,631	12,909	7,758
8	25,976	17,232	11,925	7,405
9	23,948	16,423	11,540	7,262
10	23,071	15,994	11,313	7,168
11	22,524	15,691	11,139	7,087
12	22,108	15,445	10,989	7,010
13	21,763	15,233	10,854	6,936
14	21,463	15,044	10,731	6,863
15	21,198	14,873	10,617	6,793
16	20,960	14,717	10,511	6,725
17	20,744	14,574	10,411	6,660
18	20,546	14,441	10,317	6,597
19	20,364	14,317	10,228	6,536
20	20,195	14,201	10,144	6,477
21	20,038	14,092	10,064	6,420
22	19,891	13,989	9,988	6,364
23	19,753	13,892	9,916	6,311
24	19,623	13,799	9,846	6,259
25	19,501	13,711	9,779	6,209
26	19,384	13,627	9,715	6,160
27	19,274	13,547	9,653	6,113
28	19,169	13,470	9,593	6,067
29	19,069	13,397	9,536	6,023
30	18,973	13,326	9,480	5,979
35	18,548	13,008	9,227	5,778
40	18,193	12,737	9,008	5,600
45	17,888	12,501	8,814	5,439
50	17,622	12,292	8,640	5,293
55	17,386	12,104	8,482	5,160
60	17,173	11,937	8,338	5,036
70	16,804	11,635	8,082	4,815
80	16,490	11,377	7,859	4,620
90	16,218	11,151	7,662	4,446
100	15,977	10,950	7,486	4,289
120	15,567	10,604	7,179	4,015
140	15,225	10,313	6,919	3,780
160	14,933	10,061	6,693	3,574
180	14,678	9,840	6,493	3,391
200	14,451	9,643	6,313	3,227

Capítulo 4

Aplicação a Dados de Qualidade da Água

O meio ambiente oferece a todos os seres vivos as condições essenciais para a sua sobrevivência e desenvolvimento. Contudo, a relação entre o Homem e a Natureza não tem sido pacífica.

A pressão exercida sobre os ecossistemas tem aumentado desde a segunda revolução industrial, reflectindo-se no mundo actual e originando uma importância crescente das questões de sustentabilidade ambiental. Estas questões visam actuar sobre várias formas de agressão ao meio ambiente, como por exemplo, melhorar a qualidade da água e do solo, diminuir a poluição atmosférica e desflorestação.

Neste capítulo será apresentada uma aplicação da análise de *change-points* com o objectivo de detectar mudanças no comportamento de variáveis de qualidade da água. Os dados foram obtidos a partir do Sistema Nacional de Informação de Recursos Hídricos (SNIRH) que foi criado pelo Instituto da Água (INAG) e são relativos à bacia hidrográfica do Rio Ave.

Na realização da análise estatística foi utilizado o *software* estatístico livre R (R Development Core Team, 2011), em que foram utilizadas funções já incorporadas e ainda criados novos códigos[1]. O *software* R possui o *package* "changepoint" publicado recentemente, em Fevereiro de 2012. Contudo, na análise realizada não se utilizou este *package* pois pretendia-se utilizar especificamente o Critério de Informação de Schwarz com os valores críticos obtidos por Chen & Gupta (1999), que este *package* não contem.

[1]Todos os códigos estão disponíveis mediante solicitação.

4.1 Caracterização geral

A bacia hidrográfica do Rio Ave situa-se no Noroeste de Portugal e é confrontada a Norte pela bacia hidrográfica do Rio Cávado, a Oriente pela bacia hidrográfica do rio Douro e a Sul pela Bacia Hidrográfica do Rio Leça (Figura 4.1). A bacia hidrográfica ocupa uma área de 1391 km^2, dos quais cerca de 247 km^2 e 340 km^2 correspondem, respectivamente, às áreas das bacias dos seus dois afluentes mais importantes, o Rio Este e o Rio Vizela.

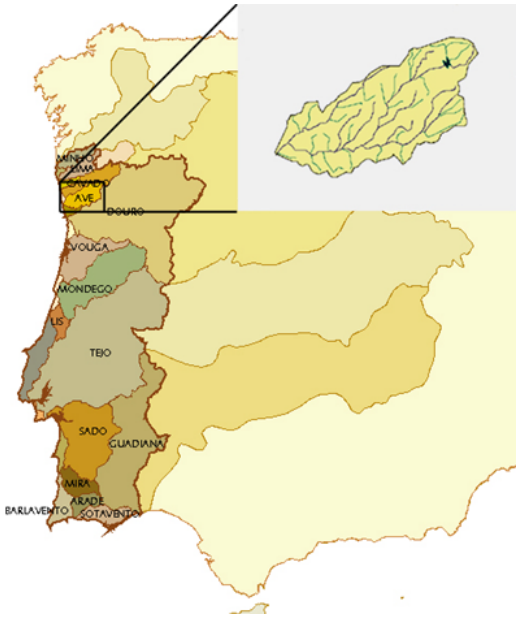


Figura 4.1: Enquadramento geográfico da bacia hidrográfica do Rio Ave.

A área abrangida inclui integral ou parcialmente os concelhos de Guimarães, Vila Nova de Famalicão, Barcelos, Braga, Cabeceiras de Basto, Fafe, Felgueiras, Lousada, Maia, Mondim de Basto, Paços de Ferreira, Póvoa de Lanhoso, Póvoa de Varzim, Santo Tirso, Vieira do Minho e Vila do Conde.

O rio Ave desenvolve-se na direcção geral Este-Oeste e percorre cerca de 100 km desde a sua nascente a 1260 m de altitude, na Serra da Cabreira, até à sua foz, em Vila do Conde, gerando uma bacia hidrográfica vasta e complexa.

Na bacia hidrográfica do Rio Ave, os cursos de água apresentam, de um modo geral, graves perturbações tanto a nível físico-químico como biológico, com excepção dos sectores próximos das nascentes, traduzindo-se pela fraca qualidade da água o que, por sua vez, tem reflexos evidentes nas comunidades aquáticas. Esta situação deve-se fundamentalmente à forte pressão exercida pelos agregados urbanos que se encontram disseminadas ao longo desta bacia. A região da bacia hidrográfica do Rio Ave tem uma economia altamente dependente da indústria, e a água tem desempenhado um papel determinante na localização da mesma neste vale (predominantemente a indústria têxtil e de vestuá-

rio). Uma das principais razões para a extrema poluição destas águas é o facto de que a construção de infra-estrutura para controlar e evitar a poluição não ter acompanhado o desenvolvimento industrial.

A monitorização da qualidade das águas de superfície tem-se tornado, assim, uma prioridade e realiza-se periodicamente devido ao agravamento da situação ambiental que tem levado a que as autoridades se preocupem com o aumento da poluição da água nesta bacia hidrográfica. Desde 1988, como parte de um plano nacional, diversas instituições nacionais e locais oficiais têm trabalhado em conjunto para o controlo rigoroso e regular da qualidade das águas superficiais, nomeadamente a monitorização ficou a cargo do Laboratório de Poluição da Direcção Regional do Ambiente e Recursos Naturais da Região. Como consequência, a bacia hidrográfica chegou a ser monitorizada por vinte estações de amostragem distribuídas pelo Rio Ave e pelos seus principais afluentes. Nestas estações de monitorização de qualidade da água realizam-se medições e análises mensais para obter uma avaliação geral da qualidade da água de superfície da bacia.

No presente estudo tomou-se por base as estações de amostragem de qualidade da Rede Nacional de Qualidade da Água a do Programa de Monitorização em Captações actualmente em funcionamento, perfazendo um total de oito estações de amostragem de qualidade (Tabela 4.1). A sua representação espacial encontra-se na Figura 4.2.

Tabela 4.1: Estações de amostragem de qualidade.

Curso de Água	Estação de Amostragem	Designação utilizada
Rio Ave	Taipas	TAI
	Riba d'Ave	RAV
	Santo Tirso	STI
	Ponte Trofa	PTR
Ribeira de Cantelães	Cantelães	CAN
Rio Ferro	Ferro	FER
Rio Vizela	Golães	GOL
	Vizela (Santo Adrião)	VSA

A variável analisada é o Oxigénio Dissolvido (OD), medido em mg/l, que constitui uma das variáveis indicadoras mais importantes na determinação do grau de poluição existente num curso de água. A oxidação de matéria orgânica, fotossíntese e respiração são processos de transformação que afectam de forma significativa esta variável. Quanto maior for o valor do Oxigénio Dissolvido, melhor será a qualidade da água.

O conjunto de dados utilizado é relativo ao período de Janeiro de 1999 a Dezembro de 2011.

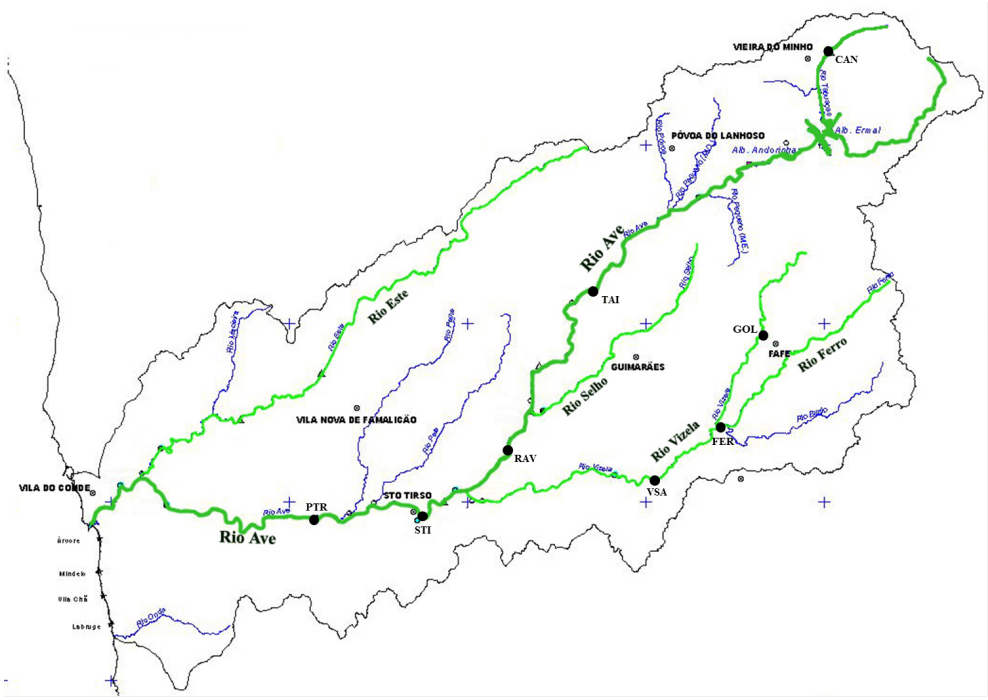


Figura 4.2: Distribuição espacial das estações de amostragem de qualidade na bacia hidrográfica do Rio Ave.

4.2 Análise exploratória dos dados

Nesta secção pretende-se fazer uma análise inicial dos dados, em que se calculam as estatísticas descritivas básicas da variável Oxigénio Dissolvido em cada uma das oito estações de amostragem de qualidade. O objectivo principal é avaliar o comportamento da variável OD nas diferentes estações de amostragem, por forma a permitir uma aplicação mais adequada das metodologiaspara a análise de detecção de *change-points*.

Na Tabela 4.2 encontram-se as principais medidas descritivas, assim como o número de valores em falta. Relativamente aos valores em falta, é a estação de Cantelães que apresenta o maior número e a estação de Riba d'Ave o menor número. Todas as estações de amostragem possuem valores em falta. No que respeita à medida de localização calculada, as estações de Riba d'Ave, Santo Tirso e Ponte Trofa apresentam os valores da média ligeiramente mais baixos quando comparados com as restantes cinco estações, traduzindo uma qualidade da água inferior. Também, relativamente à medida de dispersão, desvio padrão, as mesmas estações apresentam valores semelhantes, sendo os das três estações, Riba d'Ave, Santo Tirso e Ponte Trofa, os mais elevados.

A maior amplitude corresponde à estação de Santo Tirso e a menor à estação de Ferro. O menor e o maior valor de Oxigénio Dissolvido observados correspondem, respectivamente, às estações de Santo Tirso e de Cantelães. Para uma melhor compreensão destes valores pode-se observar a Figura 4.3, onde estão representados para cada estação

de amostragem o diagrama em caixa de bigodes e o histograma. Pode-se observar que as estações de Cantelães, Riba d'Ave, Santo Tirso e Ponte Trofa possuem *outliers*, tendo a estação de Santo Tirso o maior número. As distribuições dos dados relativos a estas estações de amostragem, à excepção de Cantelães, são as que apresentam uma maior assimetria.

Tabela 4.2: Estatísticas descritivas e número de valores em falta da variável Oxigénio Dissolvido para as 8 estações de amostragem.

Estação de Amostragem	Amplitude	Média	Desvio padrão	Assimetria	Número de valores em falta
CAN	$7,40 - 12,80$	$9,76$	$1,03$	$0,20$	6
TAI	$6,60 - 11,72$	$9,34$	$1,11$	$-0,04$	5
RAV	$1,80 - 11,70$	$8,50$	$1,70$	$-0,73$	1
STI	$1,67 - 12,00$	$8,28$	$2,04$	$-0,87$	2
PTR	$2,40 - 11,70$	$8,06$	$1,85$	$-0,73$	2
FER	$7,30 - 11,70$	$9,54$	$1,06$	$0,01$	4
GOL	$7,00 - 11,70$	$9,46$	$1,06$	$0,02$	5
VSA	$7,20 - 12,40$	$9,57$	$1,11$	$0,22$	5

Na Figura 4.4 estão representados os valores observados de Oxigénio Dissolvido ao longo do tempo, em cada estação de amostragem, sendo cada série constituída, no máximo, por 156 observações. Nestas representações podem-se observar os valores mais discrepantes, bem como a indicação de alterações da média e/ou variância das séries (em particular, entre 2004 e 2006).

No que respeita à média, esta aparentemente, aumenta ou diminui conforme a estação de amostragem, mas a variabilidade das observações diminui em todas as estações, sendo mais evidente em algumas. Outra característica importante é a indicação de uma componente sazonal. Esta sazonalidade deve-se à relação entre a concentração do Oxigénio Dissolvido com as condições meteorológicas ao longo do ano, nomeadamente, variações de temperatura e intensidade de precipitação.

4.3 Aplicação da análise de *change-points*

Nesta secção será efectuada a análise de *change-points* a cada uma das oito séries de observações, correspondentes a cada uma das estações de amostragem, de modo a perceber-se se as alterações sugeridas pela análise exploratória efectuada na secção 4.2, relativamente a mudanças silmutâneas na média e variância, são estatisticamente significativas ou apenas se devem à variação inerente dos dados (associada a fenómenos hidrológicos aleatórios). Como os dados em estudo são observações mensais, nos quais foi identificada uma com-

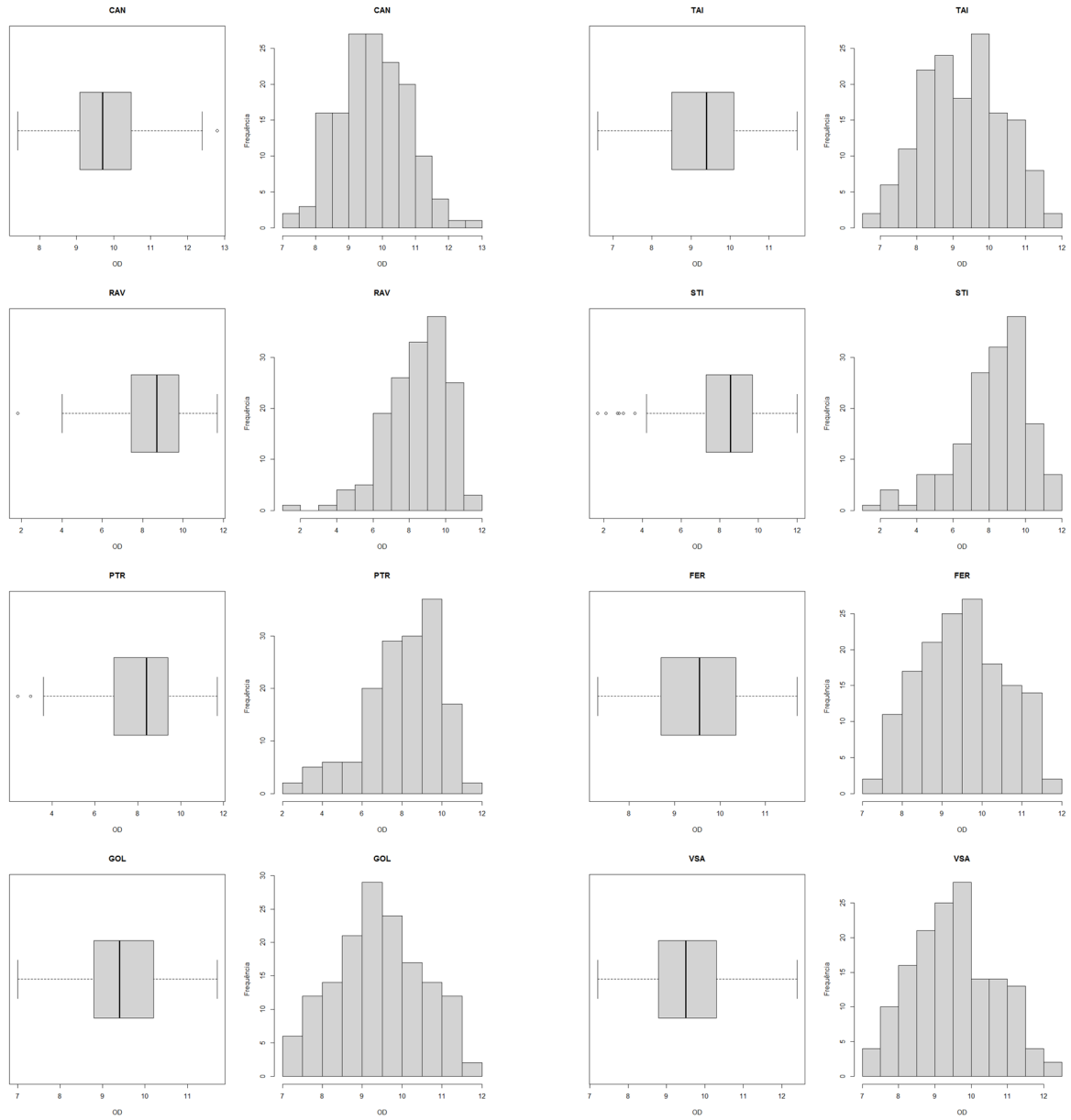


Figura 4.3: Diagrama em caixa de bigodes e histograma da variável Oxigénio Dissolvido para as 8 estações de amostragem.

ponente sazonal, o impacto desta deve ser minimizado e só depois aplicada a metodologia para se detectar a existência de *change-points*.

O método que será utilizado para estimar os coeficientes sazonais será o descrito por Gonçalves e Alpuim (2011) e a sua explicitação encontra-se na secção 2.5.2. Assim, será ajustado o modelo

$$X_t^{(M1)} = \mu + s_t + \epsilon_t, \quad t = 1, \ldots, n, \tag{4.1}$$

onde μ é a média global da série, s_t é a componente sazonal e ϵ_t o erro. Para a análise de detecção de *change-points* considerar-se-á a série dos resíduos $\hat{\epsilon}_t = X_t^{(M1)} - \hat{\mu} - \hat{s}_t, t =$

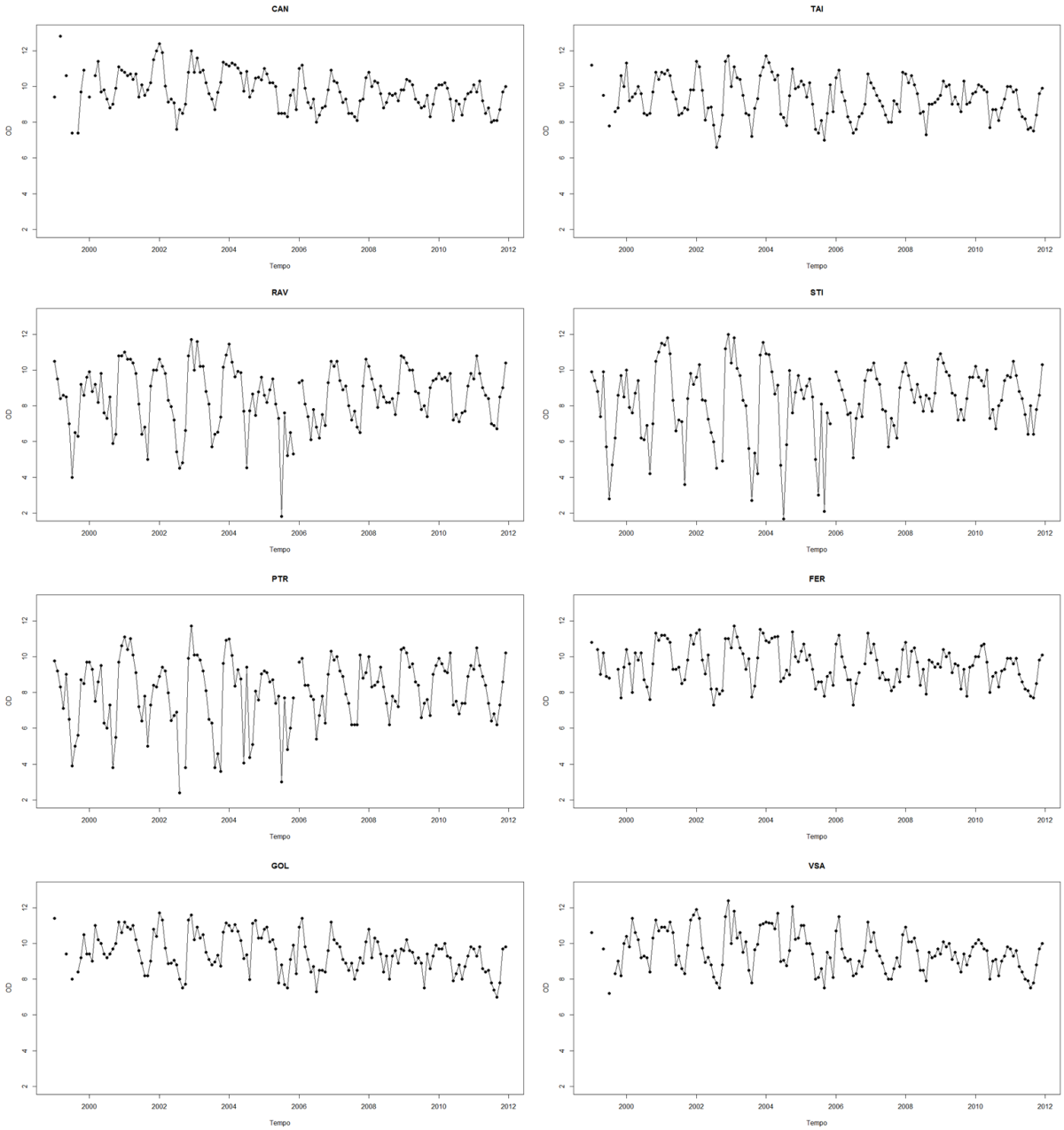


Figura 4.4: Série temporal da variável Oxigénio Dissolvido para as 8 estações de amostragem.

$1, \ldots, n$.

Pretende-se então testar se existem *change-points* na média e na variância, simultaneamente, ou seja, pretende-se testar a hipótese nula (2.11) *versus* a hipótese alternativa (2.12), através da aplicação do Critério de Informação de Schwarz (SIC) à nova série $\{\hat{\epsilon_t}\}_{t=1,\ldots,n}$, correspondendo o $SIC(n)$ ao modelo (3.8) e o $SIC(k)$ ao modelo (3.11). Para uma melhor percepção das diferenças entre os valores do critério de informação dos diferentes modelos serão representados os valores de $SIC(k)$ e o valor de $SIC(n) - c_\alpha$ para

dois níveis de significância, $\alpha = 0,05$ e $\alpha = 0,01$, e nos gráficos estes são representados através de linhas horizontais de referência.

Se, estatisticamente, se detecta um *change-point*, um segundo modelo será ajustado aos dados originais,

$$X_t^{(M2)} = \mu_t + s_t + \epsilon_t, \quad t = 1, \ldots, n, \tag{4.2}$$

onde $\mu_t = \begin{cases} \mu_I & \text{se} \quad t \leq k \\ \mu_{II} & \text{se} \quad t > k \end{cases}$, s_t é a componente sazonal para $t = 1, \ldots, n$ e $\epsilon_t \sim \begin{cases} N(0, \sigma_I^2) & \text{se} \quad t \leq k \\ N(0, \sigma_{II}^2) & \text{se} \quad t > k \end{cases}$.

Após o ajustamento do Modelo M2 (4.2), procede-se à detecção de segundos *change-points*, nas duas séries dos resíduos, antes e depois do *change-point*. Contudo, a posição adoptada nesta análise foi conservadora, no sentido em que no estudo de simulação apresentado no Capítulo 5, e em concordância com Beaulieu *et al.* (2012) a presença de correlação nas observações, mesmo que fraca ($\phi = 0,3$), tende a originar a detecção de falsos *change-points*. Assim, quando os valores obtidos $SIC(n)$ e $SIC(k)$ são próximos, mesmo que o *change-point* seja estatisticamente significativo, tomou-se a decisão de não considerar a existência do segundo *change-point*.

As estimativas da variância aqui consideradas são as estimativas de máxima verosimilhança, equação (3.9), uma vez que é este o estimador utilizado pelo Critério de Informação de Schwarz.

A validade da conclusão de existência do *change-point* está dependente da verificação dos pressupostos de normalidade e independência dos erros, para as duas subséries, antes e depois do *change-point*. A construção dos histogramas da série residual permite obter uma ideia da forma da distribuição subjacente. A normalidade testada pelo teste de Shapiro Wilk (Shapiro & Wilk, 1965), cuja hipótese nula é a de que os erros seguem uma distribuição normal. Quanto à investigação da existência de correlação, em ambas as duas subséries, são estimadas as funções de autocorrelação (FAC) e as funções de autocorrelação parcial ($FACP$).

Por fim, é apresentada a série original bem como as médias estimadas, antes e depois do *change-point*, e ainda os intervalos de confiança empíricos, $\overline{x}_I \pm 1,96 sd_I$ e $\overline{x}_{II} \pm 1,96 sd_{II}$, onde $\overline{x}_I$ e sd_I representam a média e o desvio padrão amostrais antes do ponto de mudança, e $\overline{x}_{II}$ e sd_{II} depois do mesmo.

O nível de significância considerado em todas as decisões nesta secção será de 5%.

4.3.1 Estação de amostragem de Cantelães

O Modelo (4.1) foi ajustado à série de dados de OD relativa à estação de amostragem de Cantelães, apresentando-se na Tabela 4.3 as estimativas dos coeficientes do modelo.

Tabela 4.3: Estimativas dos coeficientes do Modelo (4.1) para a estação de Cantelães.

Parâmetro	Estimativa
μ	$9,77$
s_{JAN}	$0,77$
s_{FEV}	$0,93$
s_{MAR}	$0,77$
s_{ABR}	$0,29$
s_{MAI}	$-0,07$
s_{JUN}	$-0,71$
s_{JUL}	$-0,95$
s_{AGO}	$-0,94$
s_{SET}	$-0,94$
s_{OUT}	$-0,31$
s_{NOV}	$0,43$
s_{DEZ}	$0,73$

A representação da série dos resíduos do Modelo (4.1) pode ser observada na Figura 4.5.

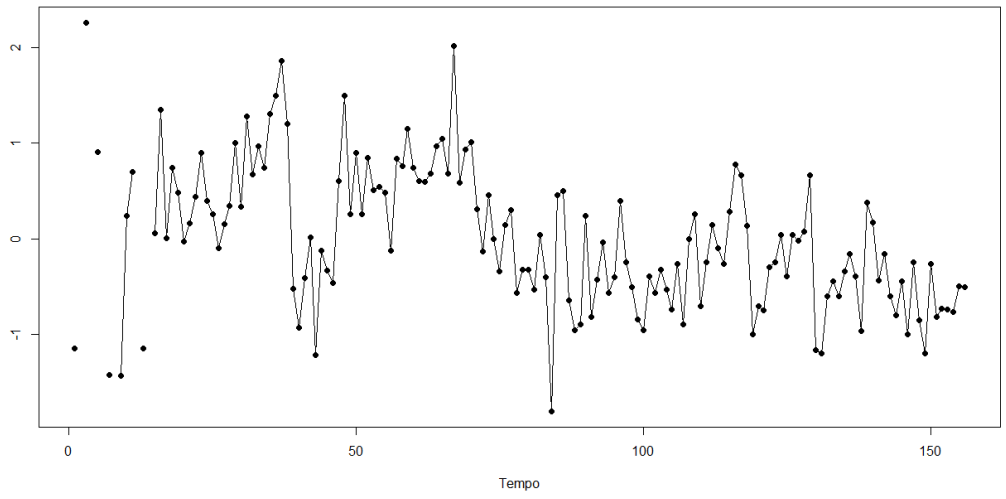


Figura 4.5: Resíduos da série da variável Oxigénio Dissolvido referente à estação de Cantelães depois de ajustado o Modelo (4.1).

O Critério de Informação de Schwarz foi aplicado, obtendo-se $SIC(n) = 345,67$ e $\min_{2\leq k\leq 154} SIC(k) = SIC(73) = 287,25$. O valor crítico para a dimensão de amostra de 150

observações (não se consideraram as 6 observações em falta) e a um nível de significância de 5% é $6,802$, concluindo-se que existe um *change-point* na média e na variância na posição 73, a que corresponde o mês de Janeiro de 2005. Os diferentes valores de $SIC(k)$ podem ser observados na Figura 4.6, assim como os valores de $SIC(n) - c_\alpha$ para $\alpha = 0,05$ e $\alpha = 0,01$.

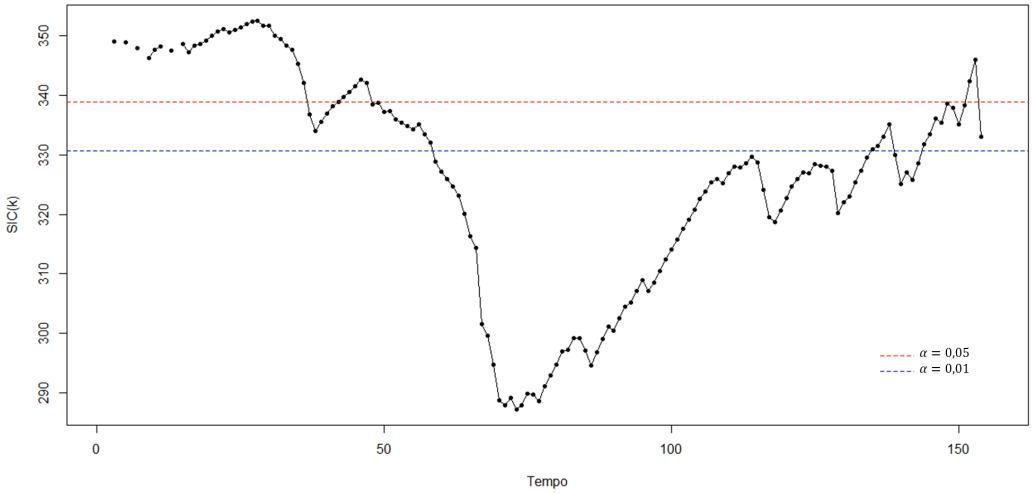


Figura 4.6: Valores de $SIC(k)$ associados à estação de Cantelães e as linhas de referência.

Uma vez que se detecta o *change-point*, ajusta-se o Modelo (4.2) e as estimativas dos coeficientes do modelo relativas à estação de Cantelães encontram-se na Tabela 4.4. Antes de se proceder à análise dos resíduos, foi testado se em cada uma das subséries, antes e depois do *change-point*, existe mais algum ponto de mudança estatisticamente significativo. Foi então detectado um ponto de mudança na primeira subsérie, obtendo-se $SIC(n) = 161,36$ e $\min\limits_{2\leq k\leq 71} SIC(k) = SIC(46) = 151,01$. O *change-point* detectado é estatisticamente significativo, pois o valor crítico para um nível de significância de 5% e considerando 67 observações (novamente não foram consideradas as 6 observações em falta) é $8,155$. Contudo, atendendo à posição tomada na secção 4.3, não se considera a existência do segundo *change-point* uma vez que $SIC(k)$ e $SIC(n)$ tomam valores próximos. Assim, considera-se apenas a existência de um *change-point* em Janeiro de 2005.

Na Figura 4.7 pode ser observada a série original dos valores do Oxigénio Dissolvido para a estação de Cantelães e os valores estimado ssegundo o Modelo 4.2. Como se pode verificar os valores estimados estão próximos dos valores originais, notando-se um afastamento superior na primeira parte da série como seria de se esperar visto a variância das observações ser superior.

A série residual encontra-se representada na Figura 4.8, assim como o ponto de mu-

Tabela 4.4: Estimativas dos coeficientes do Modelo (4.2) para a estação de Cantelães.

Parâmetro	Estimativa
μ_I	$10,22$
μ_{II}	$9,41$
σ_I^2	$0,58$
σ_{II}^2	$0,24$
s_{JAN}	$0,70$
s_{FEV}	$1,00$
s_{MAR}	$0,76$
s_{ABR}	$0,31$
s_{MAI}	$-0,09$
s_{JUN}	$-0,69$
s_{JUL}	$-0,96$
s_{AGO}	$-0,92$
s_{SET}	$-0,95$
s_{OUT}	$-0,32$
s_{NOV}	$0,41$
s_{DEZ}	$0,75$

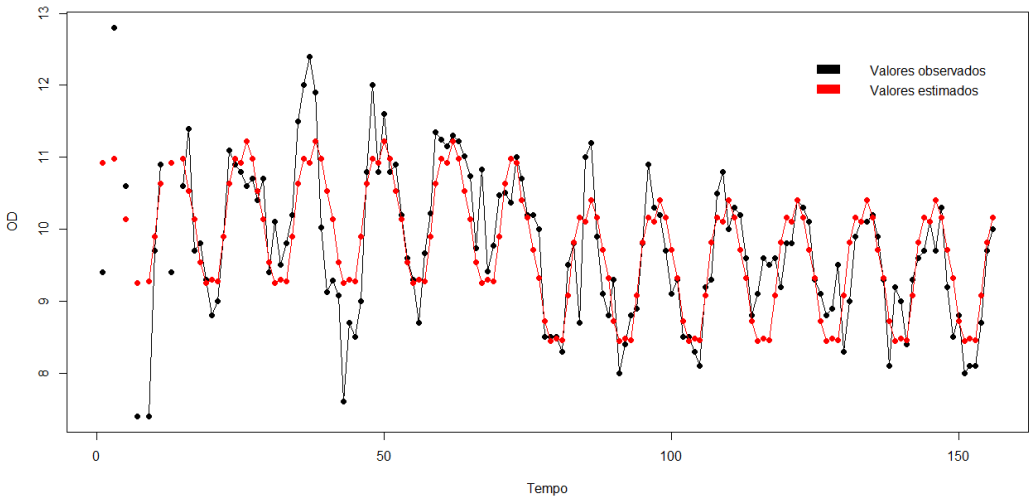


Figura 4.7: Valores observados e estimados do OD na estação de Cantelães.

dança. Como já foi referido, a validade da conclusão de existência do *change-point* está dependente da verificação dos pressupostos de normalidade e independência dos erros.

A observação dos histogramas da Figura 4.9 indica que as distribuições dos erros são simétricas, não sendo rejeitada a normalidade das distribuições pelo teste de Shapiro Wilk, onde se obtiveram os valores de prova $0,073$ e $0,628$ para a primeira e segunda subséries, respectivamente.

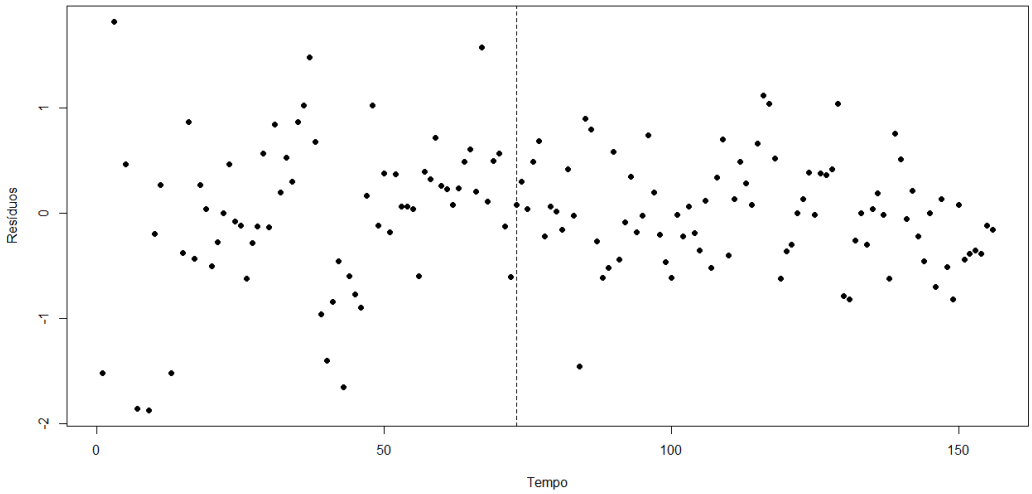


Figura 4.8: Série de resíduos associados à estação de Cantelães e o *change-point* identificado.

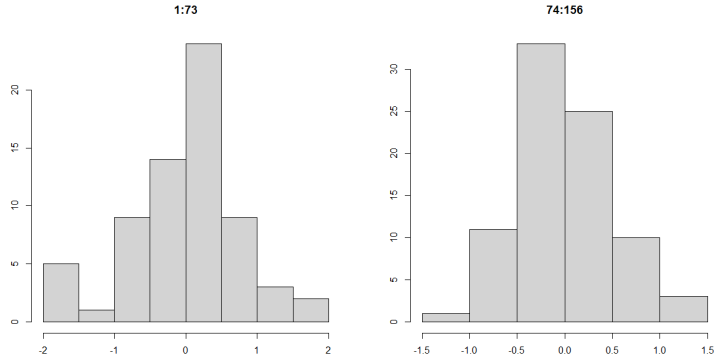


Figura 4.9: Histogramas dos resíduos associados à estação de Cantelães.

Pretende-se agora investigar a existência de correlação nas duas subséries dos erros. Para tal foram estimadas as funções de autocorrelação (FAC) e autocorrelação parcial ($FACP$) que se encontram representadas na Figura 4.10. Pela análise destes gráficos pode-se verificar a existência de uma correlação fraca na primeira subsérie, identificando-se uma estrutura autoregressiva AR(1) com parâmetro autoregressivo estimado $\hat{\phi} = 0,295$.

A representação da série original com as médias estimadas, antes e depois do *change-point* e os intervalos de confiança empíricos, encontra-se na Figura 4.11. Os valores do Oxigénio Dissolvido na estação de Cantelães diminuem, em média, a partir de Janeiro de 2005, o que corresponde a uma degradação da qualidade da água se se considerar apenas esta variável de qualidade da água, diminuindo também a respectiva variabilidade.

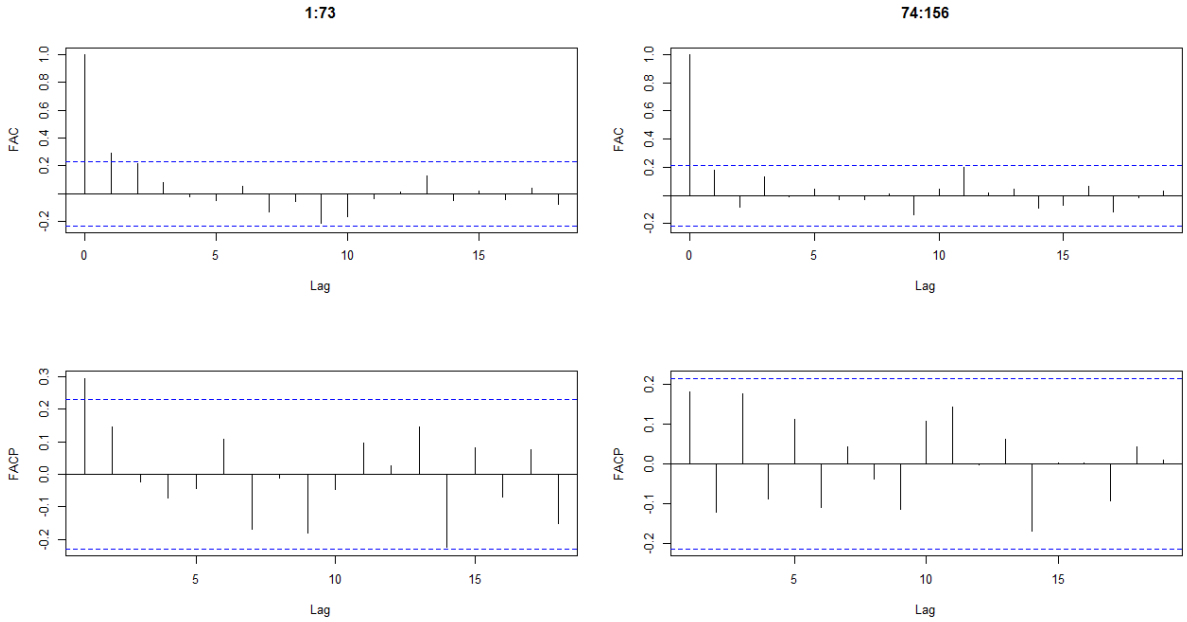


Figura 4.10: FAC e $FACP$ estimadas dos resíduos obtidos para a estação de Cantelães.

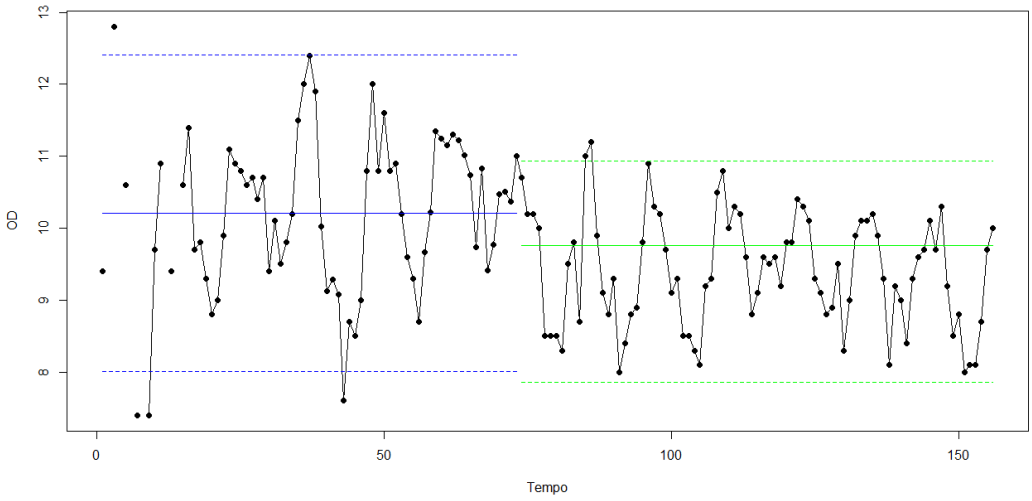


Figura 4.11: Série de observações da estação de Cantelães com as médias estimadas e os intervalos de confiança empíricos, antes e depois do *change-point*.

4.3.2 Estação de amostragem de Taipas

O Modelo (4.1) foi ajustado à série de observações de OD associada à estação de Taipas, e as estimativas obtidas estão apresentadas na Tabela 4.5.

A série relativa aos erros obtidos depois de ajustado o Modelo (4.1) está representada na Figura 4.12. Os valores obtidos considerando o critério de informação são $SIC(n) = 321,79$ e $\min_{2\leq k\leq 154} SIC(k) = SIC(70) = 307,96$. Como o valor crítico para um tamanho de

Tabela 4.5: Estimativas dos coeficientes do Modelo (4.1) para a estação de Taipas.

Parâmetro	Estimativa
μ	$9,34$
s_{JAN}	$1,23$
s_{FEV}	$1,05$
s_{MAR}	$0,69$
s_{ABR}	$0,36$
s_{MAI}	$-0,03$
s_{JUN}	$-0,80$
s_{JUL}	$-1,19$
s_{AGO}	$-1,46$
s_{SET}	$-0,82$
s_{OUT}	$-0,33$
s_{NOV}	$0,50$
s_{DEZ}	$0,80$

amostra de 151 observações (não foram contabilizadas as 5 observações em falta) é $6,791$ conclui-se que existe um *change-point* na posição 70, que corresponde a Outubro de 2004. Na Figura 4.13 estão representados os diferentes valores de $SIC(k)$.

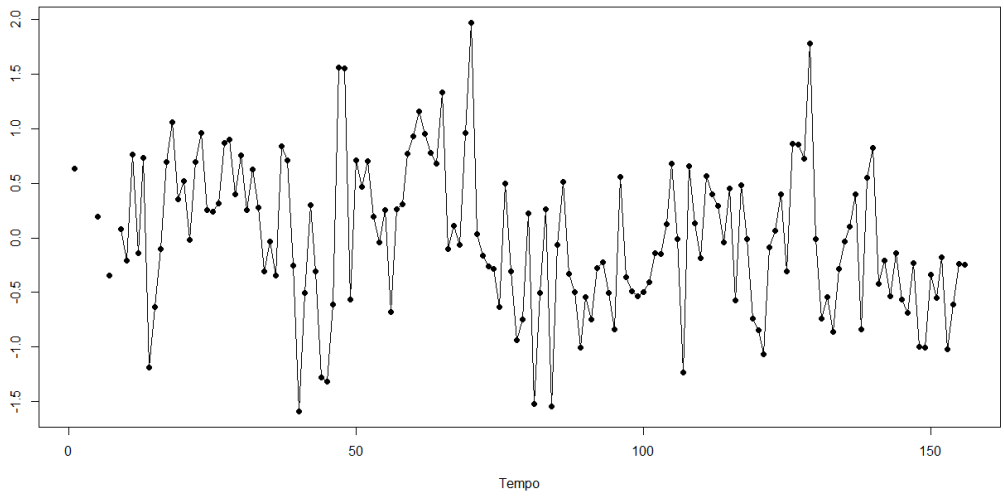


Figura 4.12: Resíduos da série da variável Oxigénio Dissolvido referente à estação de Cantelães depois de ajustado o Modelo (4.1).

O Modelo (4.2) foi ajustado, considerando agora a existência de um ponto de mudança na média e na variância, e obtiveram-se as estimativas dos parâmetros do modelo que se encontram apresentadas na Tabela 4.6.

A possibilidade de existência de mais do que um *change-point* foi estudada e obteve-se

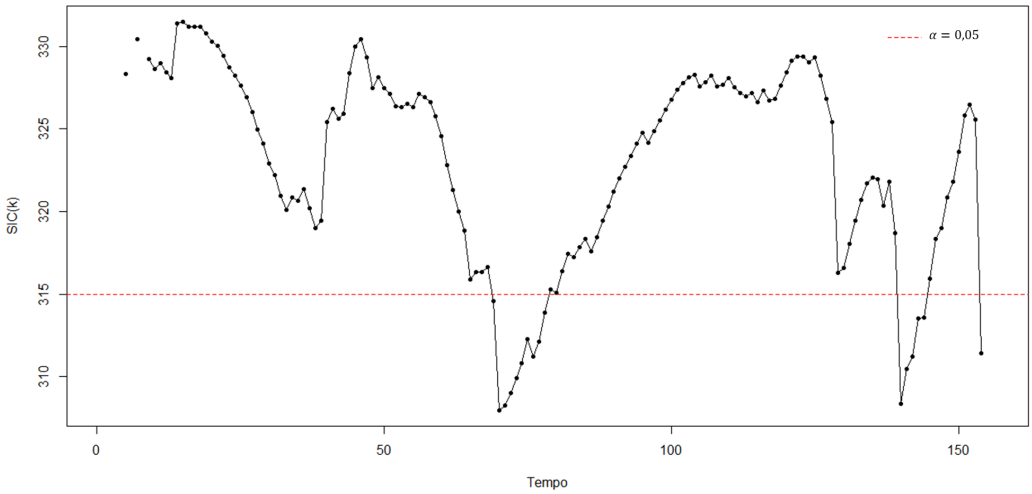


Figura 4.13: Valores de $SIC(k)$ associados à estação de Taipas e as linhas de referência

Tabela 4.6: Estimativas dos coeficientes do Modelo (4.2) para a estação de Taipas.

Componente	Coeficiente
μ_I	$9,62$
μ_{II}	$9,12$
σ_I^2	$0,49$
σ_{II}^2	$0,34$
s_{JAN}	$1,21$
s_{FEV}	$1,06$
s_{MAR}	$0,70$
s_{ABR}	$0,37$
s_{MAI}	$-0,05$
s_{JUN}	$-0,79$
s_{JUL}	$-1,21$
s_{AGO}	$-1,46$
s_{SET}	$-0,83$
s_{OUT}	$-0,35$
s_{NOV}	$0,52$
s_{DEZ}	$0,83$

para a segunda subsérie $SIC(n) = 159,72$, $\min_{2\leq k\leq 84} SIC(k) = SIC(84) = 149,09$, correspondendo à posição 154 da série total, e o valor crítico para uma amostra de 86 observações é $7,738$. Apesar do resultado ser significativo não será considerada a existência do segundo *change-point* devido à decisão tomada na secção 4.2.

Na Figura 4.14 estão representadas as séries dos valores observados e dos valores estimados, considerando e existência de apenas um *change-point* em Outubro de 2004,

constatando-se que na primeira parte da série existe um pior ajustamento devido à maior variabilidade dos dados.

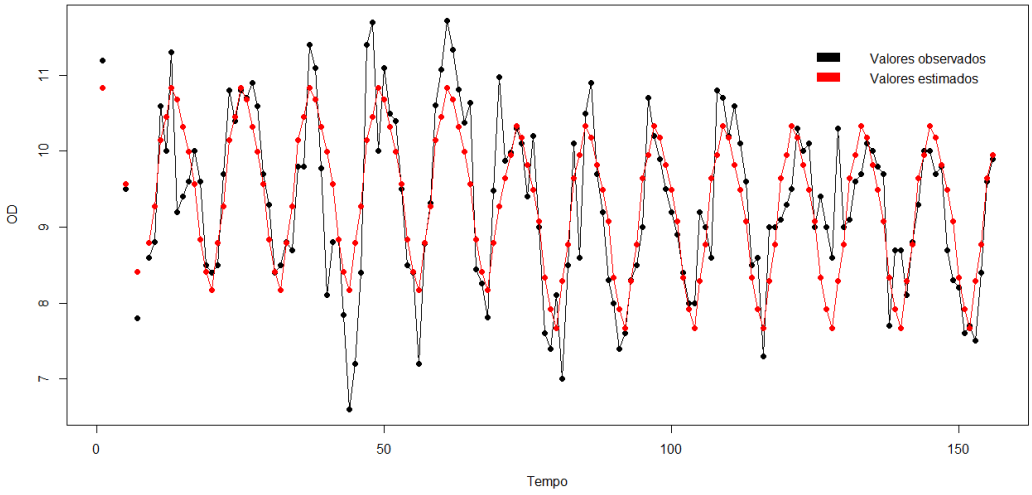


Figura 4.14: Valores observados e estimados do OD na estação de Taipas.

Ajustado o modelo terão então de ser analisados os resíduos, que se encontram representados na Figura 4.15, com o *change-point* identificado. A normalidade das duas subséries verifica-se, como sugerem os histogramas (Figura 4.16) e comprova o teste de Shapiro Wilk, em que se obtiveram os valores de prova $0,252$ e $0,236$, respectivamente. Contudo, no que respeita à independência, esta já não se verifica, como se pode observar pelo comportamento da FAC e da FACP estimadas e representadas na Figura 4.17, tanto para a primeira como para a segunda subsérie, sendo $\hat{\phi} = 0,222 = 0,362$ e $\hat{\phi} = 0,222$, respectivamente.

A representação da série original, assim como o comportamento da média e da variância antes e depois do *change-point*, encontra-se na Figura 4.18. Assim, na estação de Taipas verificou-se uma diminuição, em média, dos valores do Oxigénio Dissolvido em Outubro de 2004, e ainda uma diminuição da variância.

4.3.3 Estação de amostragem de Riba d'Ave

Será agora considerada a estação de amostragem de Riba d'Ave em que a série dos valores de OD também apresenta um comportamento sazonal, tendo-se que ajustar o Modelo (4.1). As estimativas dos parâmetros do modelo encontram-se na Tabela 4.7.

Na Figura 4.19 está representada a série dos resíduos do Modelo 4.1. Utilizando o Critério de Informação de Schwarz obteve-se $SIC(n) = 456,53$ e $\min_{2 \leq k \leq 154} SIC(k) = SIC(89) = 436,03$, podendo ser observados todos os valores de $SIC(k)$ na Figura 4.20.

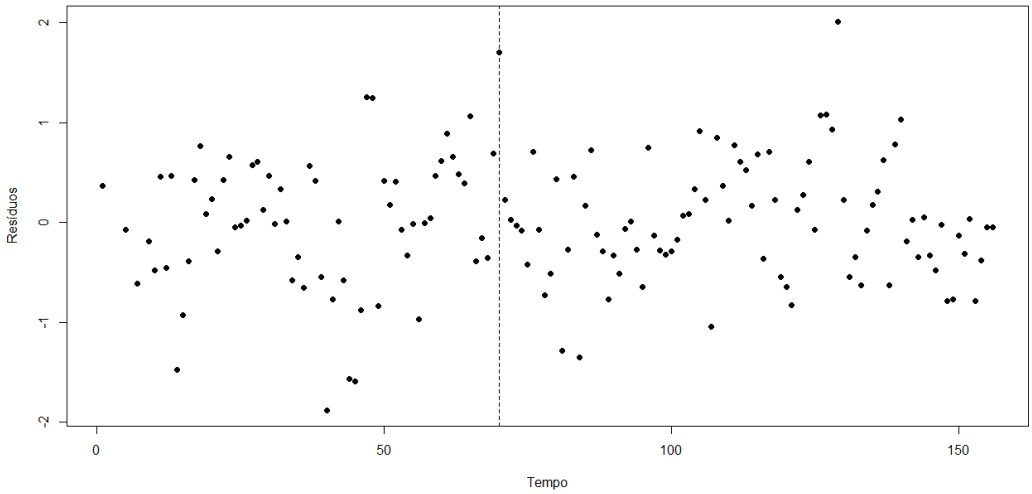


Figura 4.15: Série de resíduos associados à estação de Taipas e o *change-point* identificado.

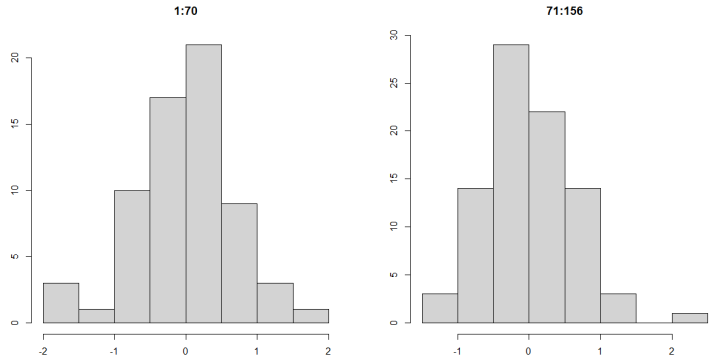


Figura 4.16: Histogramas dos resíduos associados à estação de Taipas.

A estação de Riba d'Ave apenas tem um valor em falta, pelo que para as 155 observações o valor crítico é $6,746$, concluindo-se que existe um *change-point* na média e na variância, na posição 89, que corresponde a Maio de 2006.

Detectado o *change-point*, ajustou-se o Modelo 4.2 e as estimativas obtidas dos parâmetros do modelo constam na Tabela 4.8.

A detecção de um segundo *change-point* ocorreu, neste caso na primeira subsérie. Os valores do critério de informação obtidos foram $SIC(n) = 288,33$ e $\min_{2\leq k\leq 87} SIC(k) = SIC(78) = 273,32$, sendo o valor crítico para 88 observações (não foi contabilizada a observação em falta) $6,700$. No seguimento das opções tomadas anteriormente, manteve-se a mesma posição quanto à existência de um segundo *change-point*, isto é, considera-se apenas o *change-point* em Maio de 2006. Na Figura 4.21 pode ser visualizada a série dos valores observados e a série dos valores ajustados. Constata-se um maior afastamento

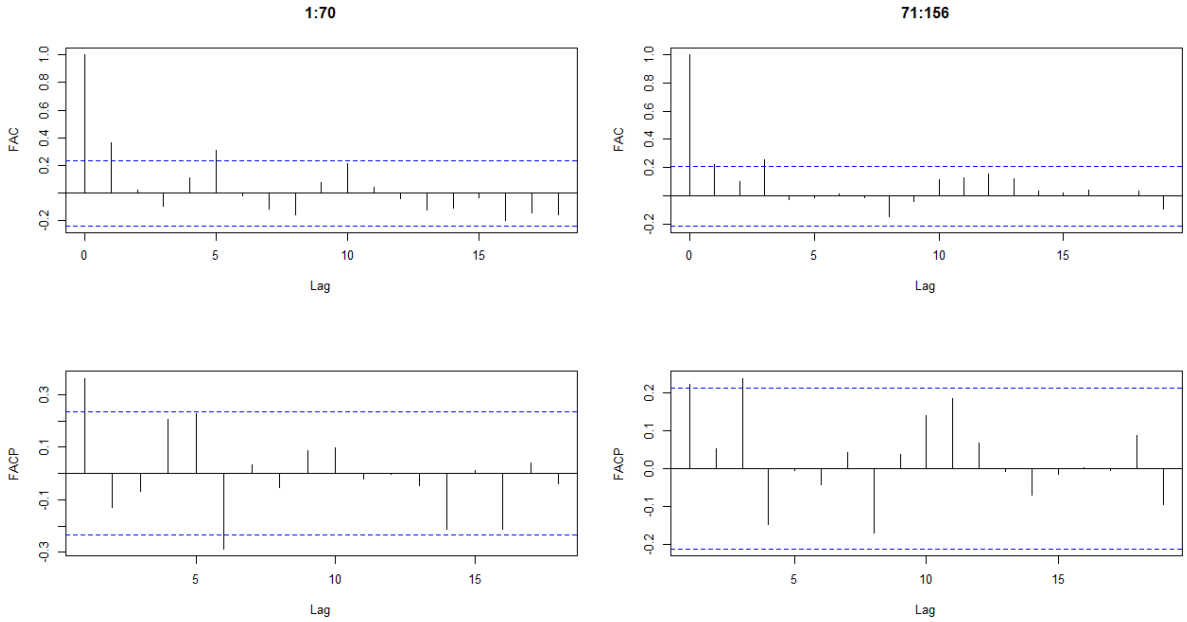


Figura 4.17: FAC e $FACP$ estimadas dos resíduos obtidos para a estação de Taipas.

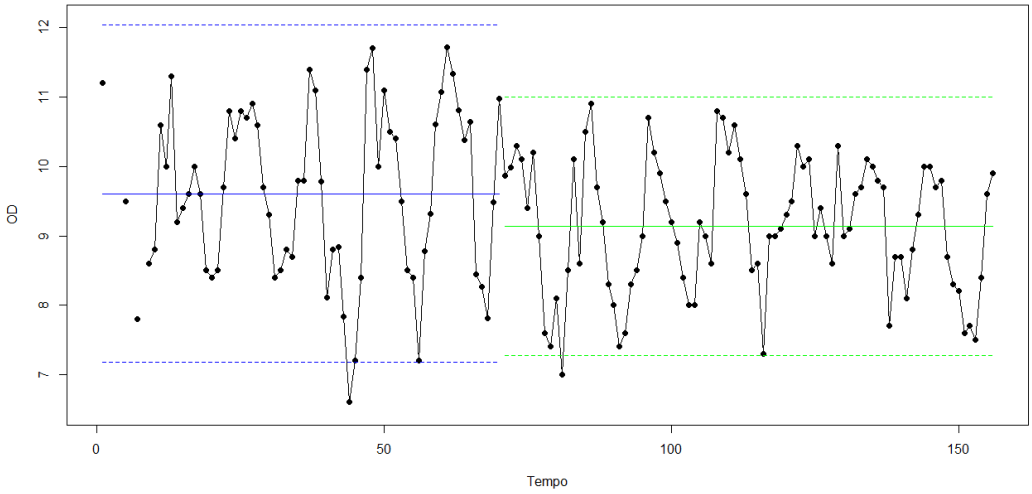


Figura 4.18: Série de observações da estação de Taipas com as médias estimadas e os intervalos de confiança empíricos, antes e depois do *change-point*.

do ajustamento do modelo às observações na primeira parte da série. A observação 79, por apresentar um comportamento discordante relativamente às restantes observações foi retirada para se verificar se esta influenciava a detecção do *change-point*, contudo tal não aconteceu. Assim manteve-se o ponto de mudança em Maio de 2006.

Procedeu-se no final à análise dos resíduos. Os resíduos da série e o respectivo *change-point* estão representados na Figura 4.22. O gráfico indica uma maior variabilidade da

Tabela 4.7: Estimativas dos coeficientes do Modelo (4.1) para a estação de Riba d'Ave.

Parâmetro	Estimativa
μ	$8,51$
s_{JAN}	$1,62$
s_{FEV}	$1,44$
s_{MAR}	$0,91$
s_{ABR}	$0,54$
s_{MAI}	$0,28$
s_{JUN}	$-0,70$
s_{JUL}	$-2,39$
s_{AGO}	$-1,43$
s_{SET}	$-1,84$
s_{OUT}	$-0,92$
s_{NOV}	$0,64$
s_{DEZ}	$1,85$

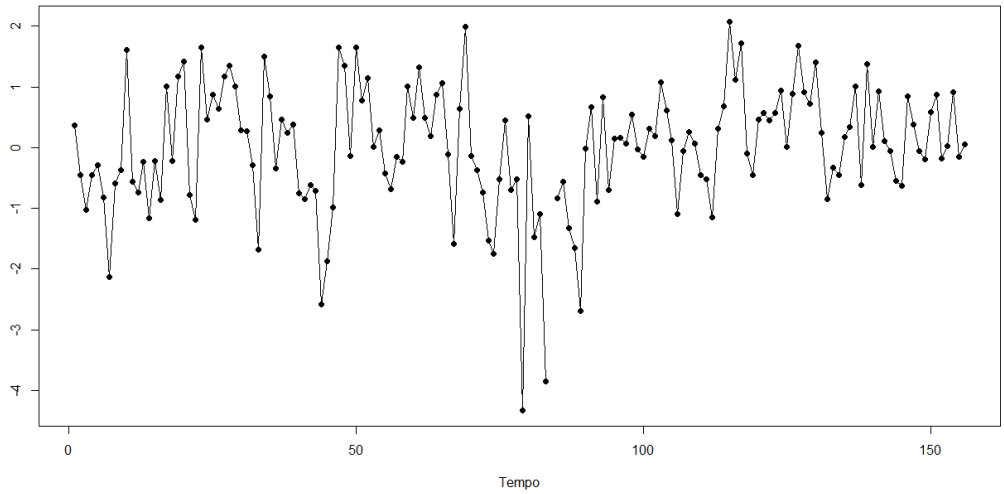


Figura 4.19: Resíduos da série da variável Oxigénio Dissolvido referente à estação de Cantelães depois de ajustado o Modelo (4.1).

primeira parte da série.

Na análise de resíduos, quanto à análise da normalidade, verificou-se que a primeira subsérie apresenta uma assimetria negativa (Figura 4.23) e no teste de Shapiro Wilk obtiveram-se os valores de prova $0,010$ e $0,779$ para a primeira e segunda subsérie, respectivamente. Quanto à independência, também não é verificada na primeira subsérie dos resíduos (Figura 4.24) e $\hat{\phi} = 0,311$.

Na Figura 4.25 está representada a série dos valores observados de Oxigénio Dissolvido associados à estação de amostragem de Riba d'Ave e o *change-point* na média e na variân-

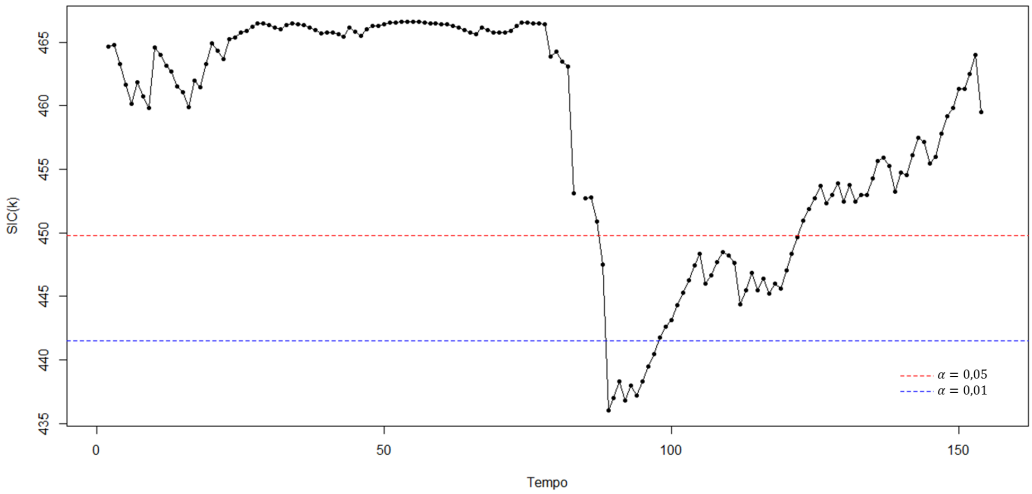


Figura 4.20: Valores de $SIC(k)$ associados à estação de Riba d'Ave e as linhas de referência.

Tabela 4.8: Estimativas dos coeficientes do Modelo (4.2) para a estação de Riba d'Ave.

Parâmetro	Estimativa
μ_I	$8,30$
μ_{II}	$8,78$
σ_I^2	$1,42$
σ_{II}^2	$0,46$
s_{JAN}	$1,64$
s_{FEV}	$1,47$
s_{MAR}	$0,93$
s_{ABR}	$0,56$
s_{MAI}	$0,30$
s_{JUN}	$-0,71$
s_{JUL}	$-2,40$
s_{AGO}	$-1,44$
s_{SET}	$-1,85$
s_{OUT}	$-0,93$
s_{NOV}	$0,63$
s_{DEZ}	$1,80$

cia bem como os intervalos de confiança empíricos. Como se pode verificar, nesta estação, houve ao longo do tempo um aumento do Oxigénio Dissolvido, em média, correspondendo a uma melhoria na qualidade da água e uma diminuição da variabilidade.

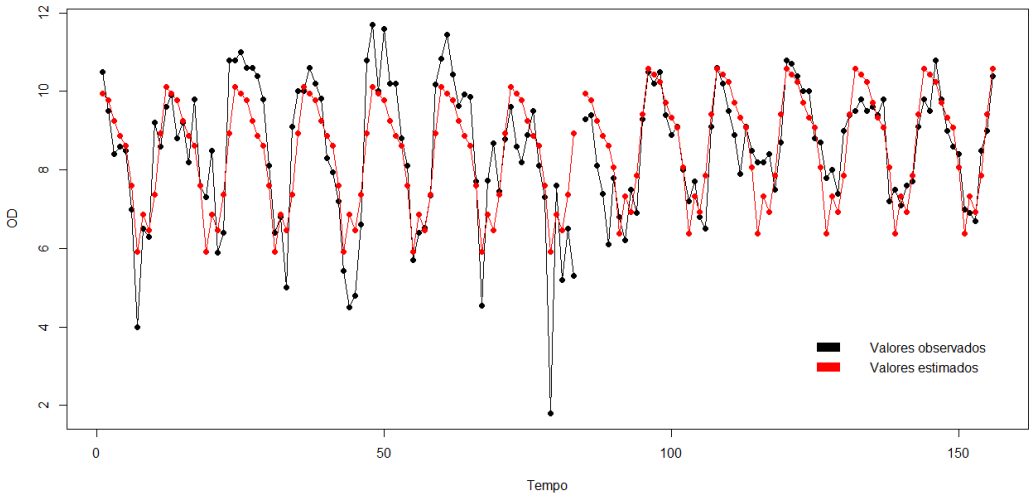


Figura 4.21: Valores observados e estimados do OD na estação de Riba d'Ave.

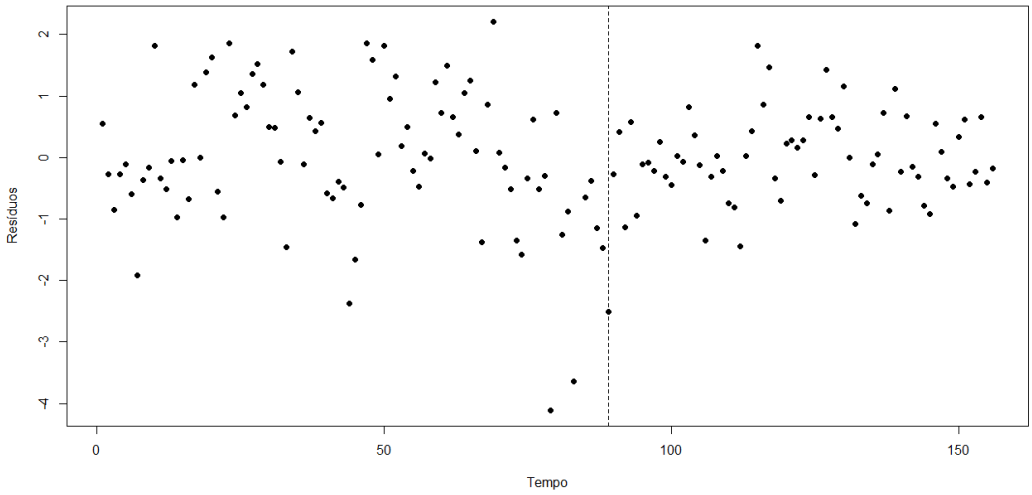


Figura 4.22: Série de resíduos associados à estação de Riba d'Ave e o *change-point* identificado.

4.3.4 Estação de amostragem de Santo Tirso

Ajustou-se o Modelo (4.1) à série de dados da estação de amostragem de Santo Tirso e as estimativas obtidas dos parâmetros do modelo estão apresentadas na Tabela 4.9.

O Critério de Informação de Schwarz foi aplicado à série de resíduos obtidas pelo Modelo 4.1, esta está representada na Figura 4.26. Obteve-se $SIC(n) = 523,33$ e $\min_{2\leq k\leq 154} SIC(k) = SIC(89) = 493,54$, estando todos os valores de $SIC(k)$ representados na Figura 4.27. Como o valor crítico para uma amostra de 154 observações (a estação de Santo Tirso tem 2 valores em falta) é 6,757, conclui-se que existe um *change-point* na

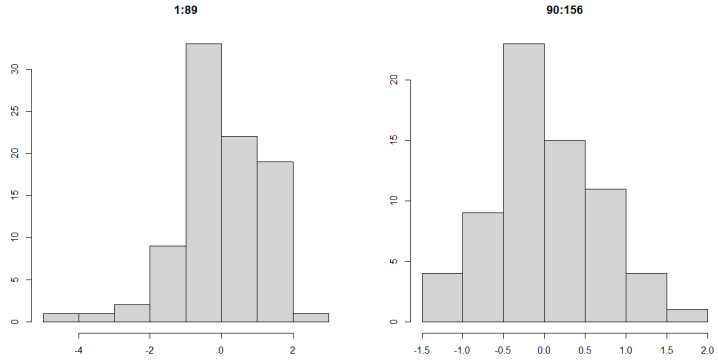


Figura 4.23: Histogramas dos resíduos associados à estação de Riba d'Ave.

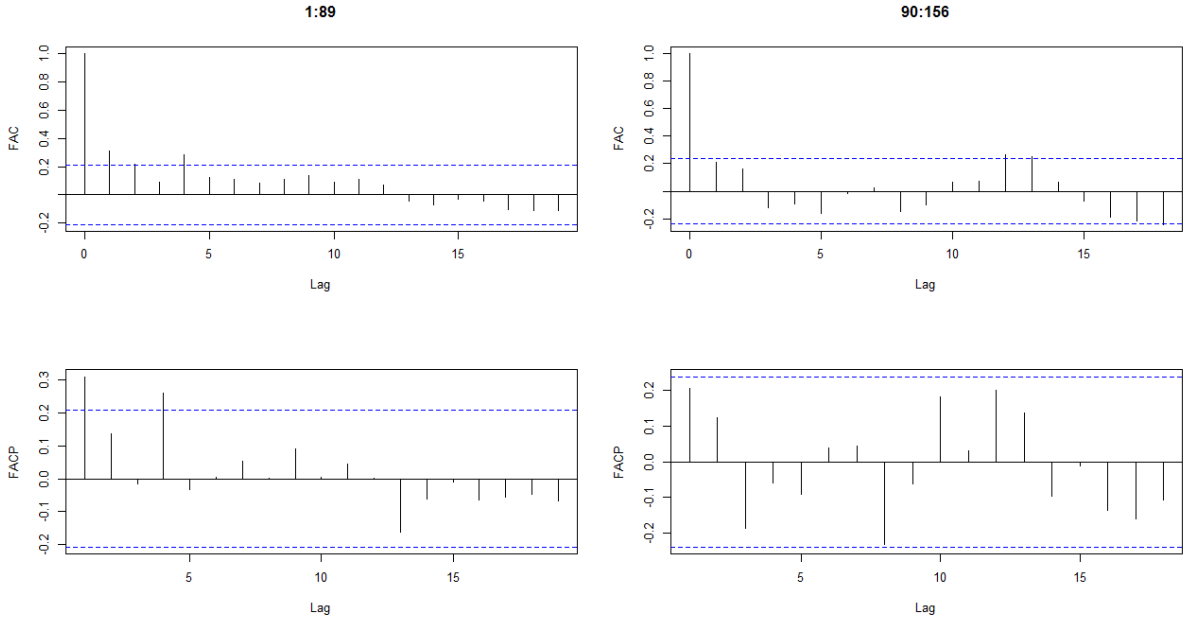


Figura 4.24: FAC e $FACP$ estimadas dos resíduos obtidos para a estação de Riba d'Ave.

posição 89, que corresponde a uma ocorrência em Maio de 2006.

Assim, o Modelo (4.2) foi ajustado aos dados observados e as estimativas estão na Tabela 4.10. Na Figura 4.28 estão representados os valores observados, bem como os valores estimados. Novamente na primeira parte da série, antes do *change-point*, a diferença entre os valores observados e estimados é mais elevada. Na estação de Santo Tirso não foi detectado um segundo *change-point* em nenhuma das duas subséries.

Os resíduos estão representados na Figura 4.29, tal como se esperava, a variabilidade é superior antes do *change-point* devido ao pior ajustamento do modelo na primeira subsérie. No que respeita à normalidade, observa-se uma assimetria positiva da sua distribuição na segunda subsérie (Figura 4.30). No Teste de Shapiro Wilk obtiveram-se os valores de prova $0,429$ e $0,043$, para a primeira e segunda subsérie, respectivamente, rejeitando-se

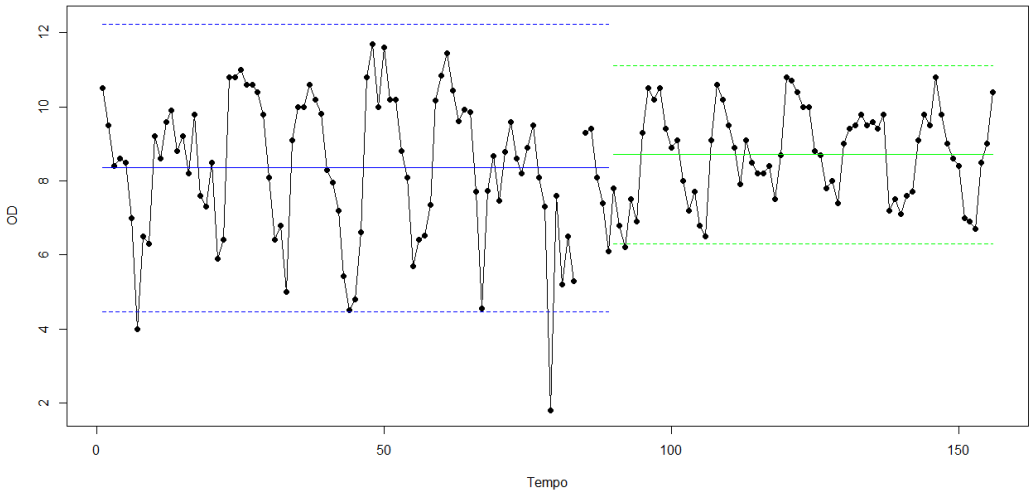


Figura 4.25: Série de observações da estação de Riba d'Ave com as médias estimadas e os intervalos de confiança empíricos, antes e depois do *change-point*.

Tabela 4.9: Estimativas dos coeficientes do Modelo (4.1) para a estação de Santo Tirso.

Parâmetro	Estimativa
μ	$8,28$
s_{JAN}	$1,89$
s_{FEV}	$1,72$
s_{MAR}	$1,09$
s_{ABR}	$0,67$
s_{MAI}	$0,36$
s_{JUN}	$-1,37$
s_{JUL}	$-2,73$
s_{AGO}	$-1,71$
s_{SET}	$-1,91$
s_{OUT}	$-1,05$
s_{NOV}	$1,14$
s_{DEZ}	$1,90$

a hipótese de normalidade para a segunda. A independência dos resíduos também não é verificada na segunda subsérie como se pode observar na Figura 4.31 e para esta subsérie $\hat{\phi} = 0,388$.

Uma representação gráfica dos valores observados do Oxigénio Dissolvido na estação de amostragem de Santo Tirso encontra-se na Figura 4.32, bem como a média e o intervalo empírico, antes e depois do *change-point*. Nesta estação os valores do Oxigénio Dissolvido aumentaram, em média, a partir de Maio de 2006 e é, realmente notória, a diminuição da variância.

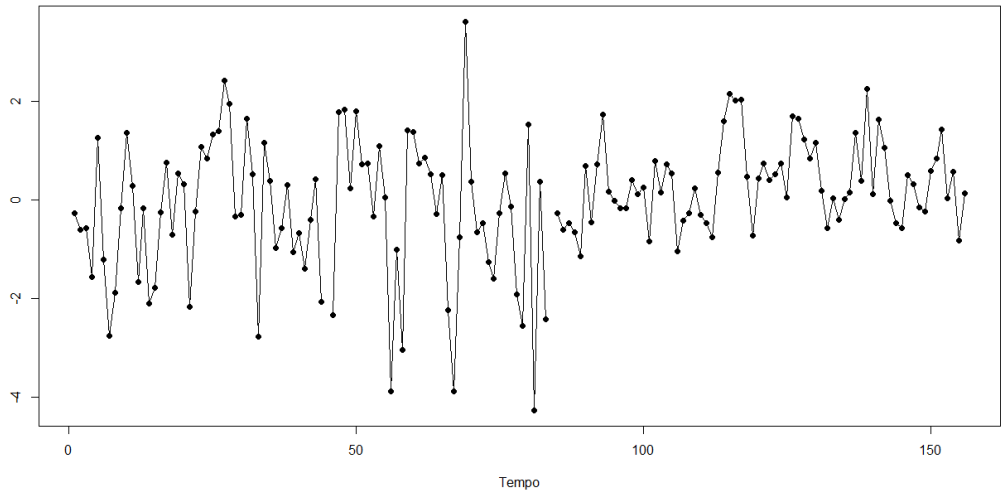


Figura 4.26: Resíduos da série da variável Oxigénio Dissolvido referente à estação de Santo Tirso depois de ajustado o Modelo (4.1).

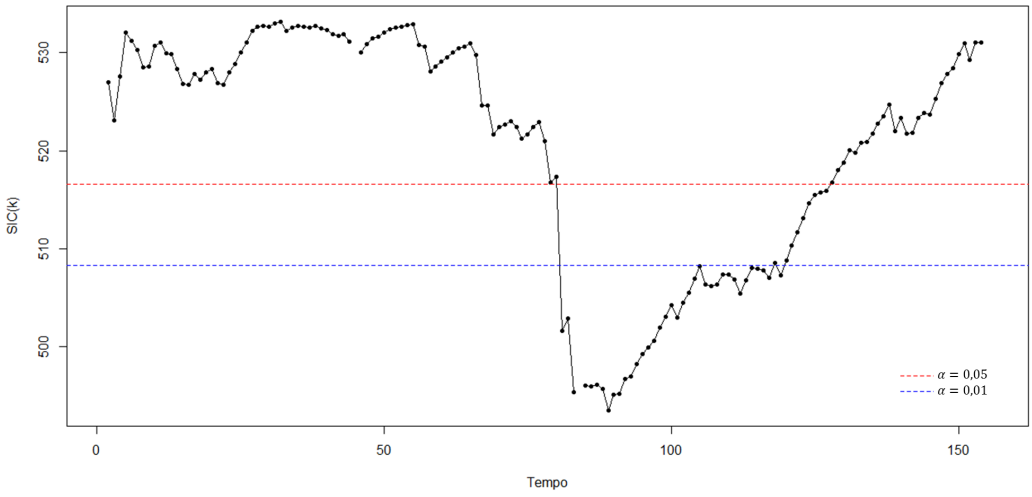


Figura 4.27: Valores de $SIC(k)$ associados à estação de Santo Tirso e as linhas de referência.

4.3.5 Estação de amostragem de Ponte Trofa

Foi ajustado o Modelo (4.1) à série de valores de OD observados na estação de amostragem de Ponte Trofa, constando as estimativas relativas aos coeficientes da média e sazonalidade na Tabela 4.11.

A representação dos resíduos do Modelo 4.1 encontra-se na Figura 4.33. Aplicou-se o SIC e os resultados obtidos foram $SIC(n) = 482,48$ e $\min_{2 \leq k \leq 154} SIC(k) = SIC(119) = 459,24$. O valor crítico para uma amostra de 154 observações (não foram contabilizados

Tabela 4.10: Estimativas dos coeficientes do Modelo (4.2) para a estação de Santo Tirso.

Parâmetro	Estimativa
μ_I	$7,97$
μ_{II}	$8,69$
σ_I^2	$2,21$
σ_{II}^2	$0,64$
s_{JAN}	$1,92$
s_{FEV}	$1,76$
s_{MAR}	$1,13$
s_{ABR}	$0,71$
s_{MAI}	$0,40$
s_{JUN}	$-1,39$
s_{JUL}	$-2,74$
s_{AGO}	$-1,72$
s_{SET}	$-1,96$
s_{OUT}	$-1,06$
s_{NOV}	$1,12$
s_{DEZ}	$1,83$

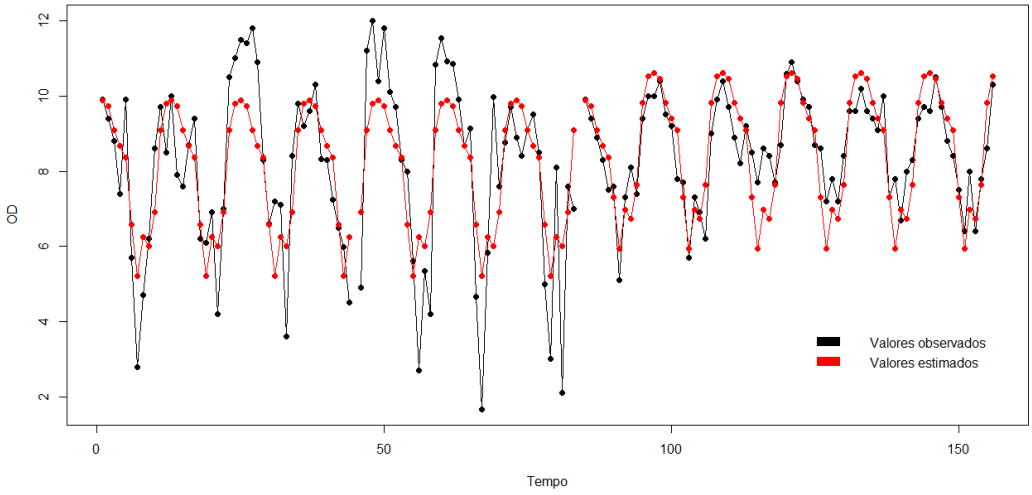


Figura 4.28: Valores observados e estimados do OD na estação de Santo Tirso.

os 2 valores em falta) é $6,757$, logo pode-se concluir é estatisticamente significativa a presença de um *change-point* na posição 119, no mês de Novembro de 2008. Contudo, pela experiência das restantes séries e pela observação da Figura 4.33 o valor correspondente à posição 106 (Outubro de 2007) parece discordante considerando as observações da segunda metade da série, aproximadamente, e por isso esta observação foi retirada. Foi novamente aplicado o Critério de Informação de Schwarz. Os novos resultados obtidos

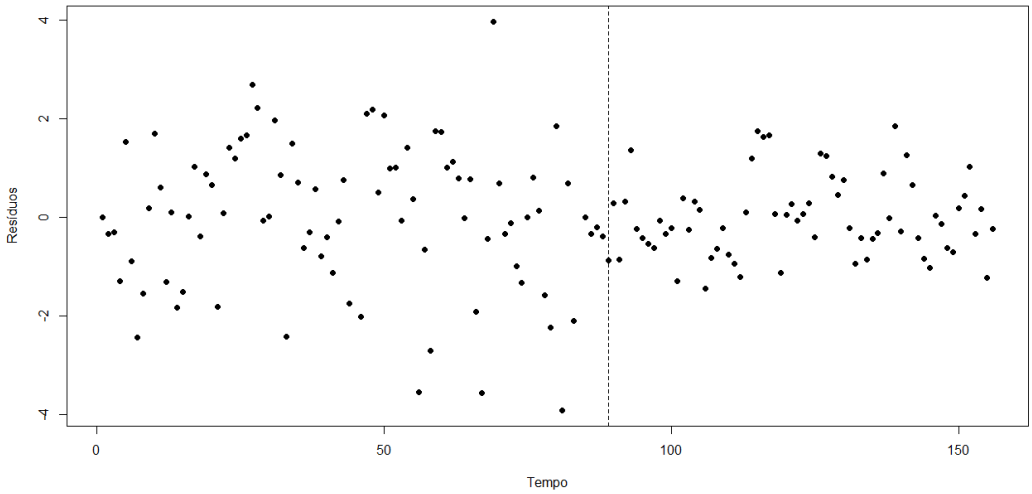


Figura 4.29: Série de resíduos associados à estação de Santo Tirso e o *change-point* identificado.

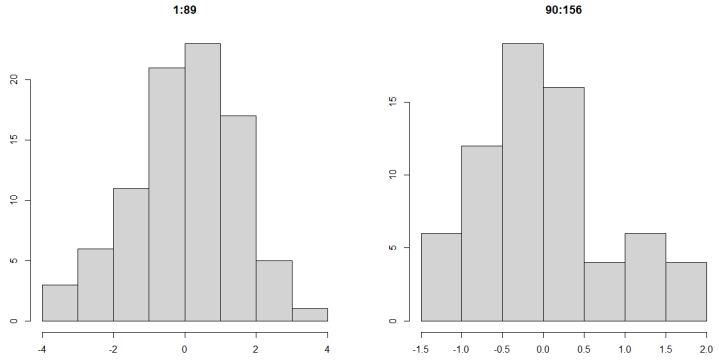


Figura 4.30: Histogramas dos resíduos associados à estação de Santo Tirso.

foram $SIC(n) = 471,44$ e $\min_{2 \leq k \leq 154} SIC(k) = SIC(83) = 443,22$, podendo todos os valores de $SIC(k)$ e os níveis de referência ser observados na Figura 4.34. Como se retirou uma observação o valor crítico passou a ser $6,769$, concluindo-se que a eliminação de uma única observação alterou a posição do *change-point*, passando este a ser na posição 83, que corresponde a Novembro de 2005 e será o considerado para esta estação de amostragem de Ponte Trofa.

Considerado o *change-point* na posição 83, foi ajustado o Modelo (4.2) onde se obtiveram as estimativas apresentadas na Tabela 4.11. Na Figura 4.35 estão representados os valores observados do Oxigénio Dissolvido e os valores ajustados, observando-se uma maior disparidade na primeira parte da série, o que tem vindo a acontecer em todas as estações de amostragem estudadas. Nesta estação, Ponte Trofa, não foi detectado um segundo *change-point*.

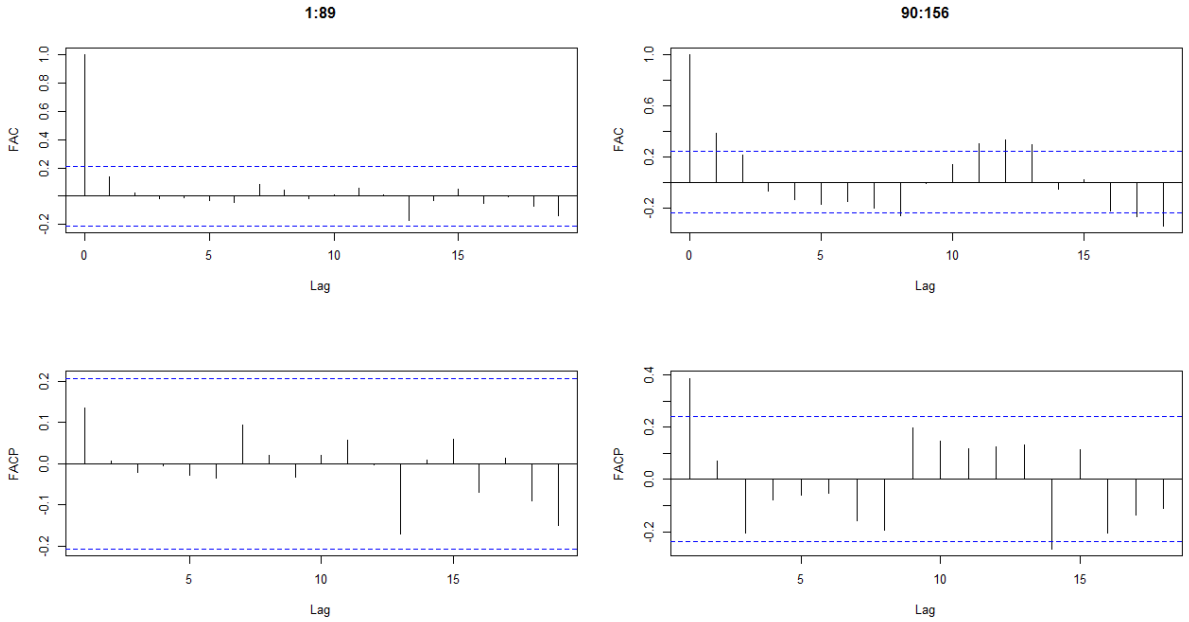


Figura 4.31: FAC e $FACP$ estimadas dos resíduos obtidos para a estação de Santo Tirso.

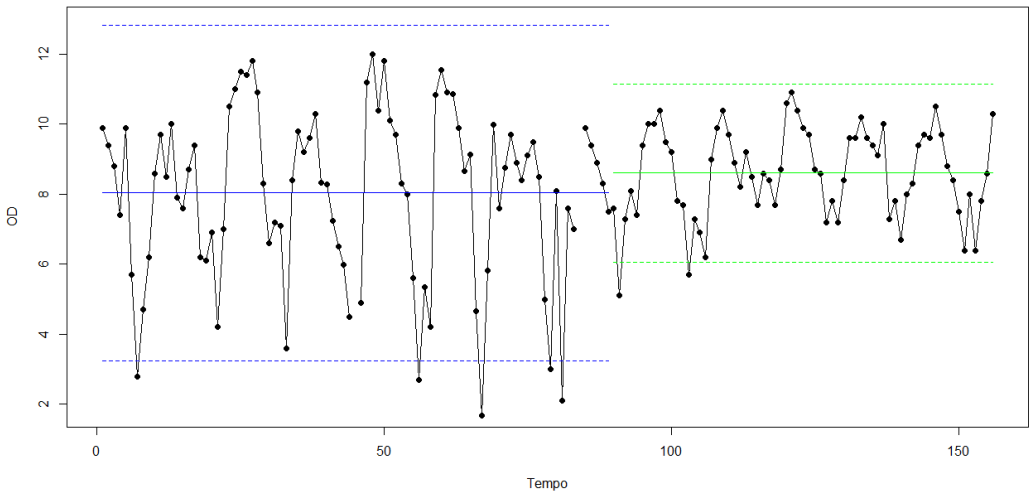


Figura 4.32: Série de observações da estação de Santo Tirso com as médias estimadas e os intervalos de confiança empíricos, antes e depois do *change-point*.

O pior ajustamento na primeira parte da série reflecte-se na maior variância dos erros como se pode observar na Figura 4.36. No que respeita à normalidade, esta não se verifica na segunda subsérie como indica a Figura 4.37 e comprova o teste de Shapiro Wilk, onde se obtiveram os valores de prova $0,431$ e $0,006$ para a primeira e a segunda subsérie, respectivamente. Quanto à independência, esta verifica-se nas duas subséries, como se pode observar na Figura 4.38.

Tabela 4.11: Estimativas dos coeficientes do Modelo (4.1) para a estação de Ponte Trofa.

Parâmetro	Estimativa
μ	$8,05$
s_{JAN}	$1,86$
s_{FEV}	$1,64$
s_{MAR}	$0,93$
s_{ABR}	$0,75$
s_{MAI}	$0,45$
s_{JUN}	$-1,02$
s_{JUL}	$-1,79$
s_{AGO}	$-2,02$
s_{SET}	$-2,06$
s_{OUT}	$-1,26$
s_{NOV}	$0,63$
s_{DEZ}	$1,89$

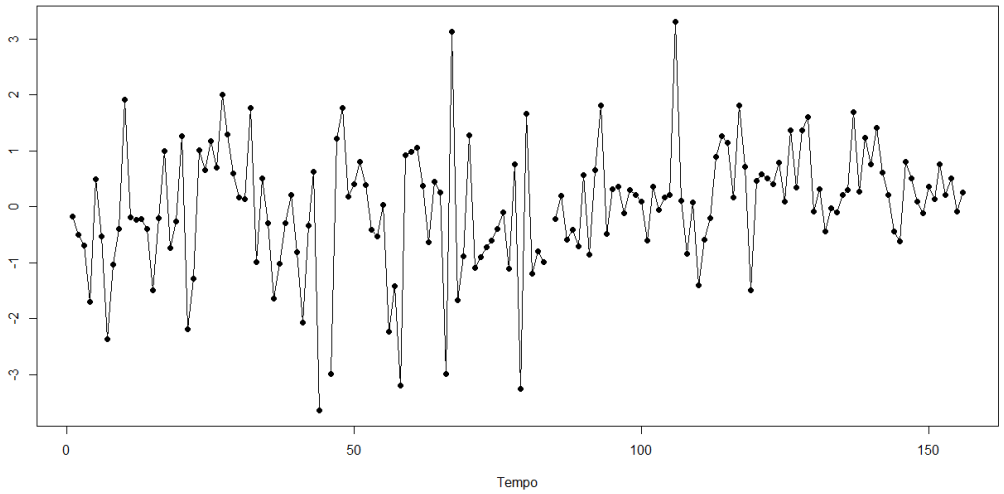


Figura 4.33: Resíduos da série da variável Oxigénio Dissolvido referente à estação de Ponte Trofa depois de ajustado o Modelo (4.1).

A estação de amostragem de Ponte Trofa apresenta um aumento do Oxigénio Dissolvido, em média, que corresponde a uma melhoria da qualidade da água considerando apenas esta variável de qualidade da água, a partir de Novembro de 2005 e uma diminuição da variância (Figura 4.39).

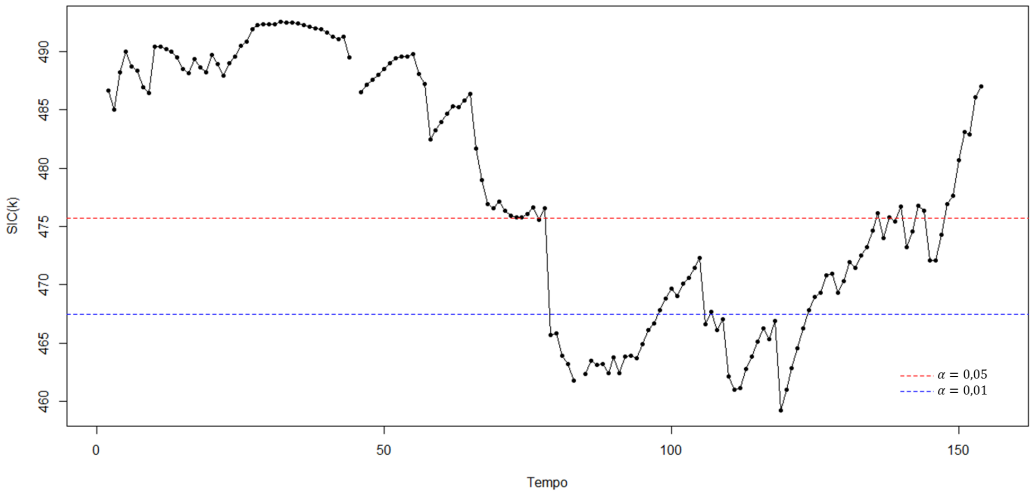


Figura 4.34: Valores de $SIC(k)$ associados à estação de Ponte Trofa e as linhas de referência.

Tabela 4.12: Estimativas dos coeficientes do Modelo (4.2) para a estação de Ponte Trofa.

Parâmetro	Estimativa
μ_I	$7,78$
μ_{II}	$8,37$
σ_I^2	$1,70$
σ_{II}^2	$0,60$
s_{JAN}	$1,87$
s_{FEV}	$1,65$
s_{MAR}	$0,94$
s_{ABR}	$0,75$
s_{MAI}	$0,46$
s_{JUN}	$-1,02$
s_{JUL}	$-1,79$
s_{AGO}	$-2,01$
s_{SET}	$-2,08$
s_{OUT}	$-1,26$
s_{NOV}	$0,63$
s_{DEZ}	$1,86$

4.3.6 Estação de amostragem de Ferro

A série de dados observada na estação de amostragem de Ferro possui um comportamento sazonal pelo que foi aplicado o Modelo (4.1). As estimativas dos coeficientes do modelo encontram-se na Tabela 4.13. A representação da série dos erros associada ao Modelo 4.1 pode ser observada na Figura 4.40.

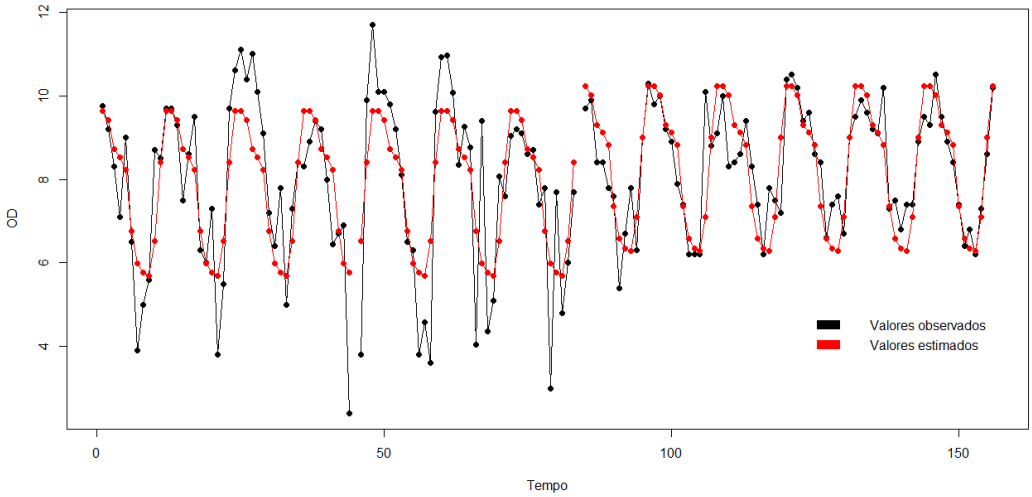


Figura 4.35: Valores observados e estimados do OD na estação de Ponte Trofa.

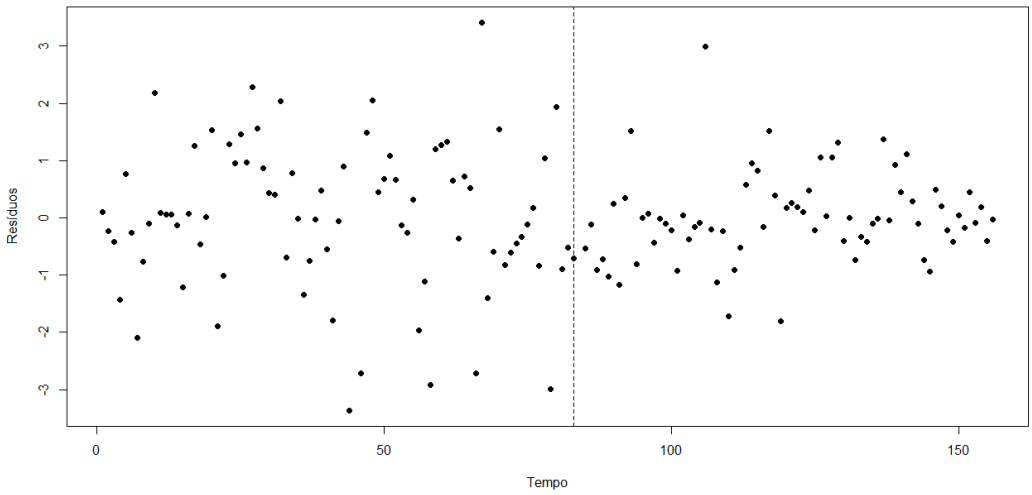


Figura 4.36: Série de resíduos associados à estação de Ponte Trofa e o *change-point* identificado.

Através da aplicação do SIC obteve-se $SIC(n) = 356,58$ e $\min_{2 \leq k \leq 154} SIC(k) = SIC(70) = 341,12$, e para uma amostra de 152 observações (a estação de Ferro tem 4 valores em falta) o valor crítico é $6,780$, concluindo-se assim que existe uma mudança significativa na posição 70, ou seja, em Outubro de 2004. Todos os valores de $SIC(k)$ estão representados na Figura 4.41. Na estação de Ferro não foi detectado um segundo ponto de mudança significativo.

Detectado o *change-point* foi ajustado o Modelo (4.2), considerando a alteração que existe na série de dados. As estimativas dos coeficientes do novo modelo encontram-se

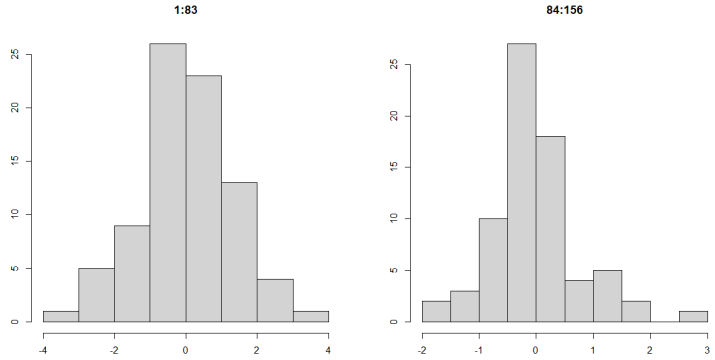


Figura 4.37: Histogramas dos resíduos associados à estação de Ponte Trofa.

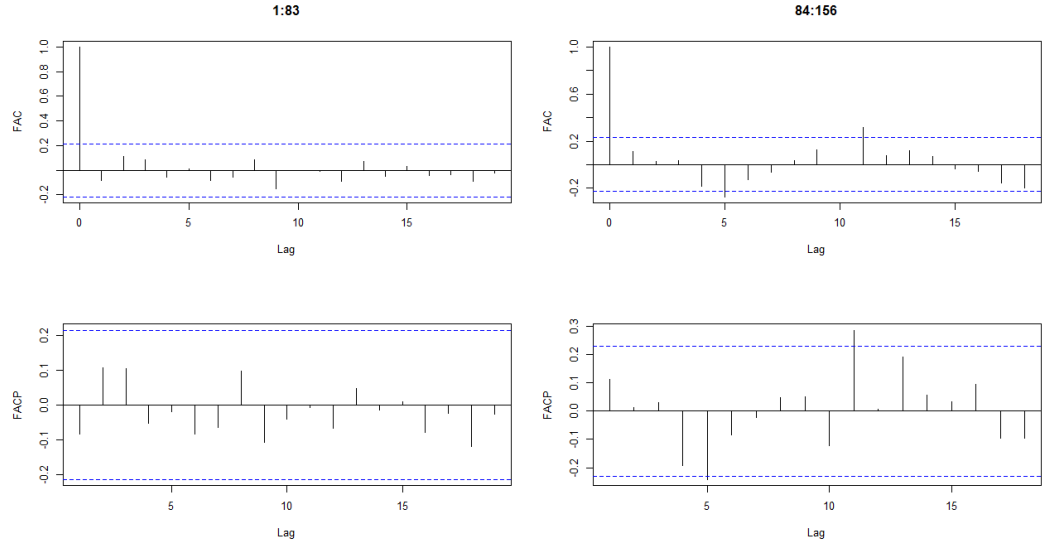


Figura 4.38: FAC e $FACP$ estimadas dos resíduos obtidos para a estação de Ponte Trofa.

na Tabela 4.14. Na Figura 4.42 representam-se os valores observados e os estimados, notando-se uma variabilidade superior nos valores da primeira parte da série (antes do *change-point*).

Os resíduos obtidos (Figura 4.43) reflectem o ajustamento, com uma diminuição da variância depois do *change-point*, não sendo nesta estação de amostragem esta diminuição mais notória. A distribuição das duas subséries está representada através dos histogramas na Figura 4.44. Obtiveram-se os valores de prova $0,055$ e $0,878$ para o teste de Shapiro Wilk, logo pode-se concluir que os resíduos das duas subséries seguem uma distribuição normal. No que respeita à independência, esta não se verifica na primeira subsérie (Figura 4.45), sendo $\hat{\phi} = 0,335$.

A série dos valores observados de Oxigénio Dissolvido com o *change-point*, em Outubro de 2004, na média e na variância foi representada na Figura 4.46, constatando-se

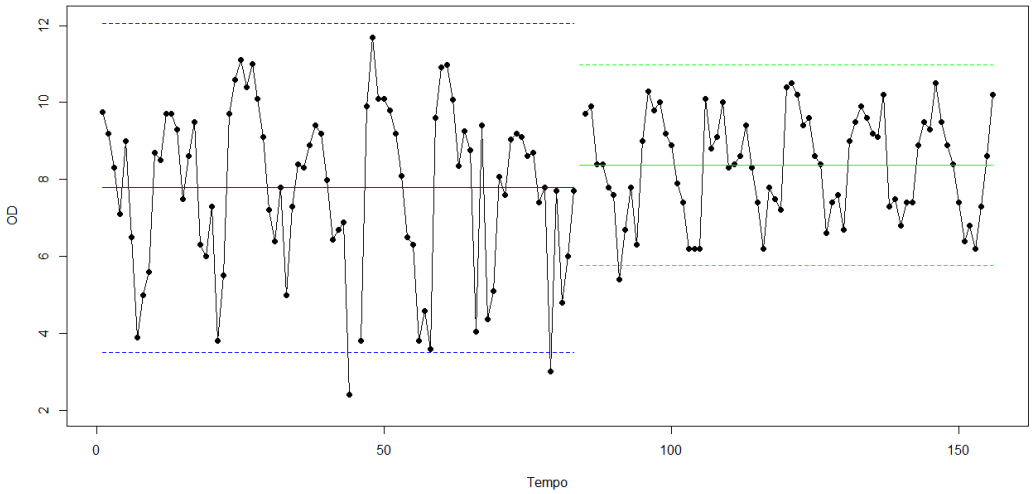


Figura 4.39: Série de observações da estação de Ponte Trofa com as médias estimadas e os intervalos de confiança empíricos, antes e depois do *change-point*.

Tabela 4.13: Estimativas dos coeficientes do Modelo (4.1) para a estação de Ferro.

Parâmetro	Estimativa
μ	$9,53$
s_{JAN}	$0,96$
s_{FEV}	$1,00$
s_{MAR}	$0,60$
s_{ABR}	$0,42$
s_{MAI}	$0,08$
s_{JUN}	$-0,74$
s_{JUL}	$-0,81$
s_{AGO}	$-1,18$
s_{SET}	$-1,04$
s_{OUT}	$-0,24$
s_{NOV}	$0,31$
s_{DEZ}	$0,64$

uma diminuição, em média, dos valores de Oxigénio Dissolvido e uma diminuição da sua variabilidade.

4.3.7 Estação de amostragem de Golães

Na estação de Golães foi ajustado o Modelo (4.1), em que as estimativas dos coeficientes do modelo estão descritas na Tabela 4.15. A série relativa aos erros do Modelo 4.1 pode ser observada na Figura 4.47.

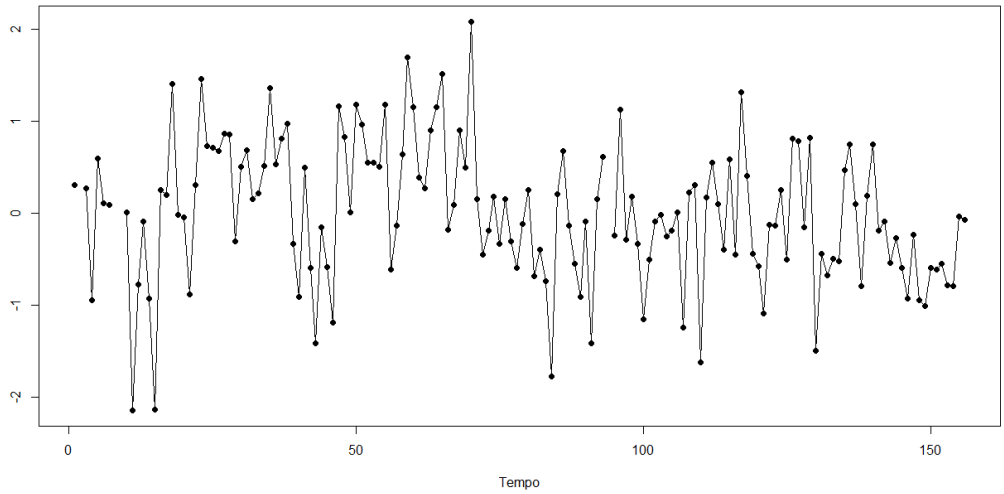


Figura 4.40: Resíduos da série da variável Oxigénio Dissolvido referente à estação de Ferro depois de ajustado o Modelo (4.1).

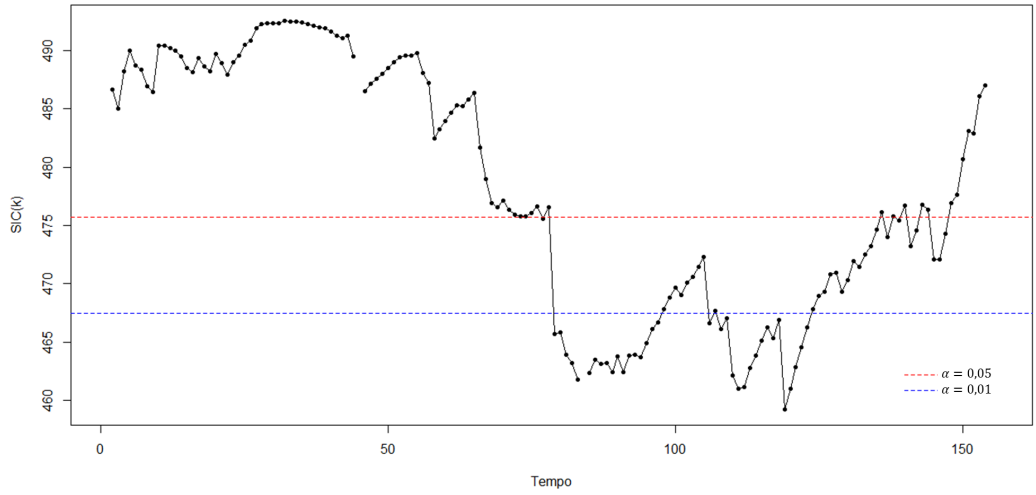


Figura 4.41: Valores de $SIC(k)$ associados à estação de Ferro e as linhas de referência.

Aplicou-se o Critério de Informação de Schwarz e concluiu-se que existe um ponto de mudança estatisticamente significativo na posição 77, que corresponde a Maio de 2005, pois obteve-se $SIC(n) = 348,35$ e $\min_{2 \leq k \leq 154} SIC(k) = SIC(77) = 312,58$ (Figura 4.48), sendo o valor crítico para uma amostra de 151 observações (não foram contabilizadas as 5 observações em falta) $6,791$.

Como foi detectado um *change-point* ajustou-se o Modelo (4.2) à série de dados observados e as estimativas dos coeficientes obtidas encontram-se na Tabela 4.16. Não foi detectado um segundo *change-point* nesta estação de amostragem. Na Figura 4.49

Tabela 4.14: Estimativas dos coeficientes do Modelo (4.2) para a estação de Ferro.

Parâmetro	Estimativa
μ_I	$9,81$
μ_{II}	$9,31$
σ_I^2	$0,70$
σ_{II}^2	$0,37$
s_{JAN}	$0,95$
s_{FEV}	$1,01$
s_{MAR}	$0,59$
s_{ABR}	$0,41$
s_{MAI}	$0,07$
s_{JUN}	$-0,75$
s_{JUL}	$-0,83$
s_{AGO}	$-1,17$
s_{SET}	$-1,03$
s_{OUT}	$-0,27$
s_{NOV}	$0,34$
s_{DEZ}	$0,68$

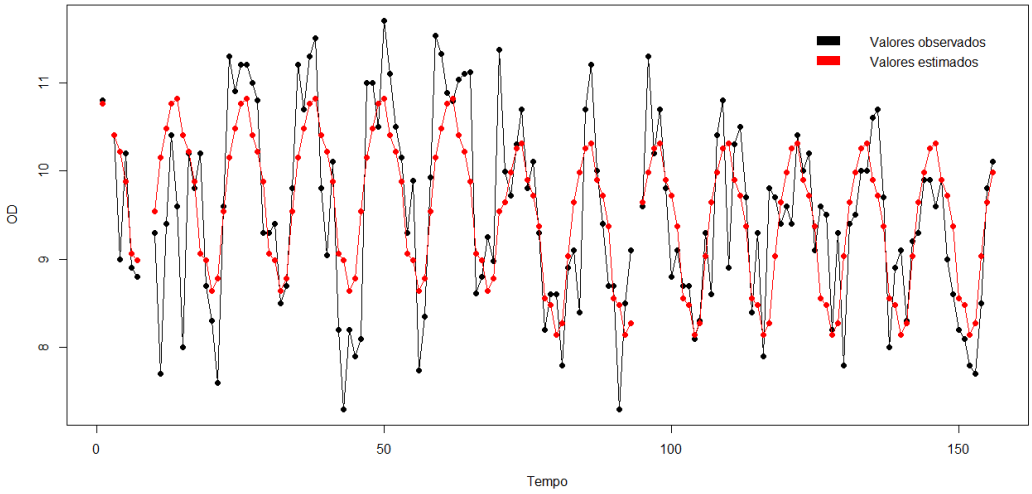


Figura 4.42: Valores observados e estimados do OD na estação de Ferro.

representa-se a série dos valores observados de Oxigénio Dissolvido e dos valores estimados, existindo alguma diferença entre os valores, principalmente, na primeira parte da série.

Os resíduos obtidos estão representados na Figura 4.50 e nesta estação de amostragem a diferença de variância, antes e depois do *change-point*, é a menos notória. A normalidade verifica-se nas duas subséries (Figura 4.51), tendo-se obtido no teste de Shapiro Wilk os

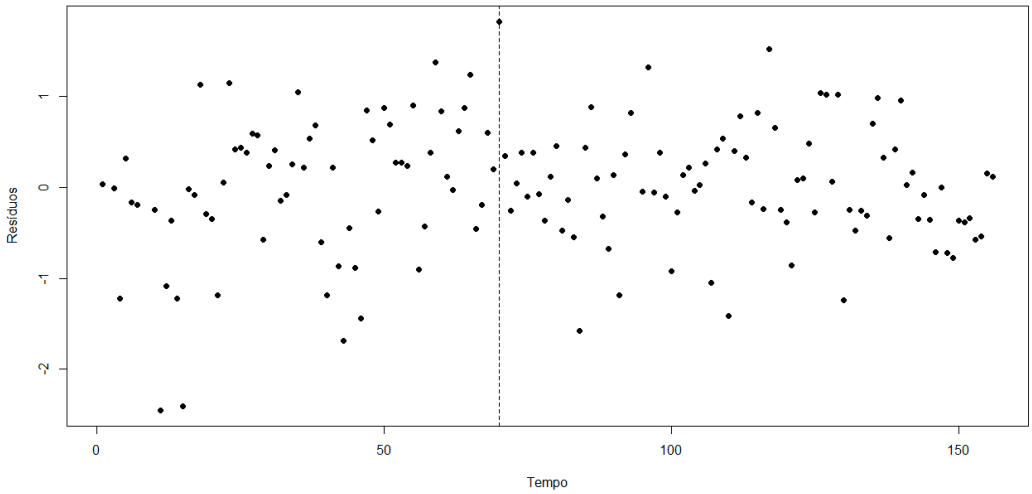


Figura 4.43: Série de resíduos associados à estação de Ferro e o *change-point* identificado.

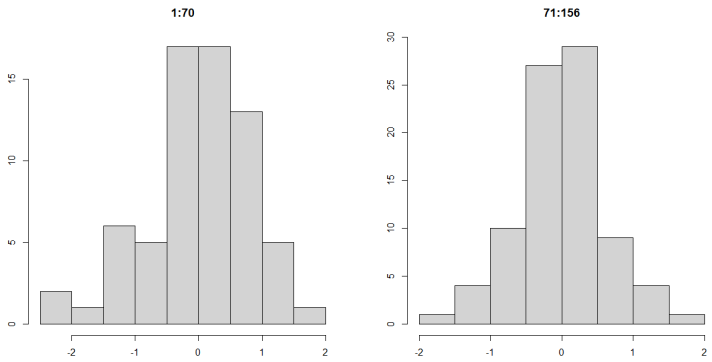


Figura 4.44: Histogramas dos resíduos associados à estação de Ferro.

valores de prova $0,053$ e $0,804$. Relativamente à independência, como pode ser observado na Figura 4.52, esta não se verifica na primeira subsérie, sendo $\hat{\phi} = 0,380$.

A representação dos valores observados de Oxigénio Dissolvida na estação de Golães com a mudança na média e na variância, em simultâneo, pode ser observada na Figura 4.53. Constata-se que houve um decréscimo tanto do valor médio do Oxigénio Dissolvido como o da variabilidade.

4.3.8 Estação de amostragem de Vizela (Santo Adrião)

A última estação a ser estudada é a estação de Vizela (Santo Adrião). Também foi ajustado o Modelo (4.1) à série de dados observados (Tabela 4.17). A série correspondente aos resíduos do Modelo 4.1 está representada na Figura 4.54.

Com a utilização do Critério de Informação de Schwarz obteve-se $SIC(n) = 358,44$

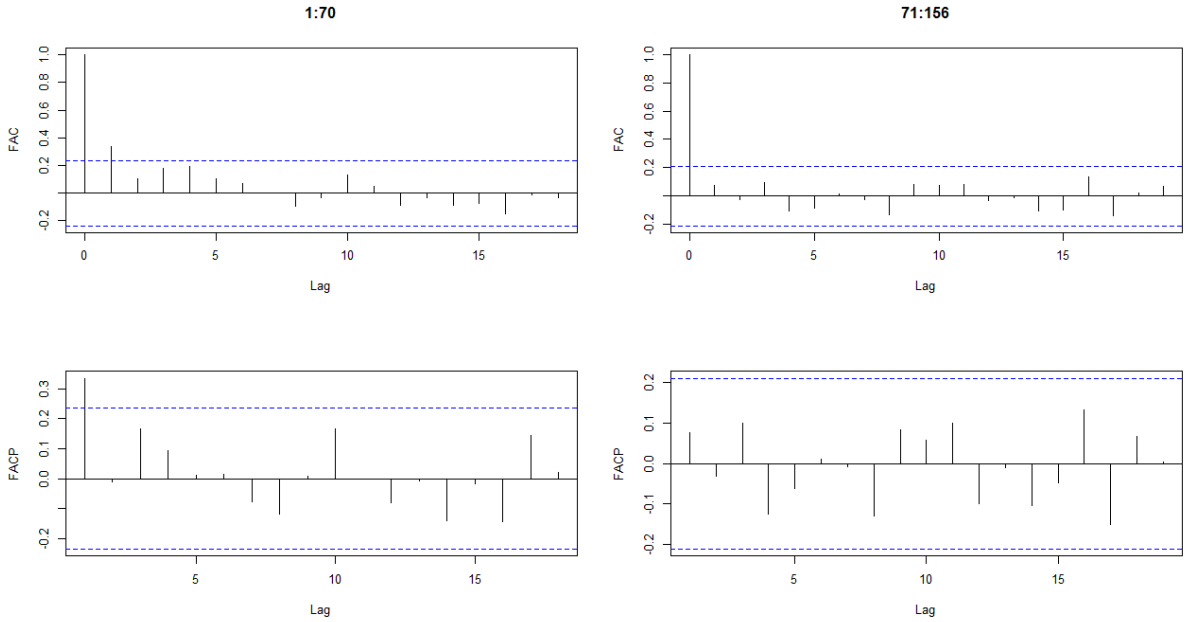


Figura 4.45: FAC e $FACP$ estimadas dos resíduos obtidos para a estação de Ferro

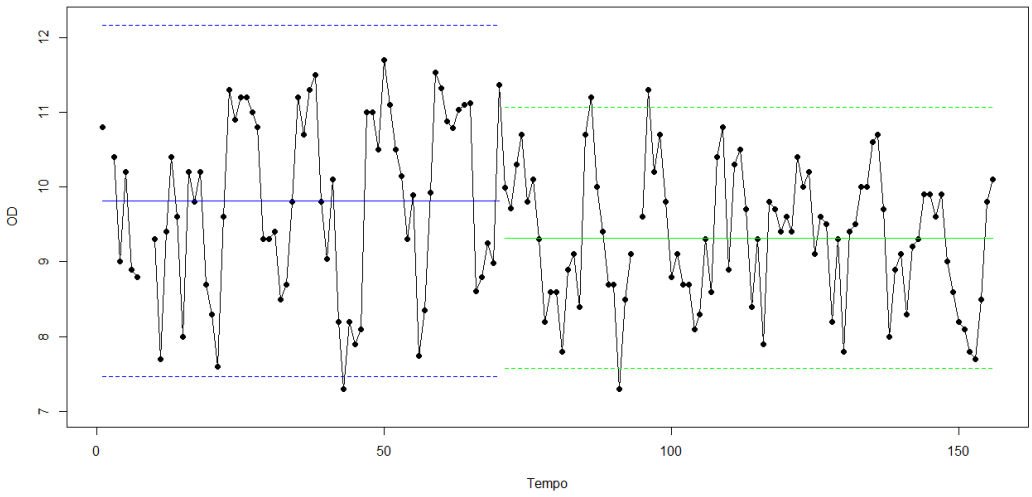


Figura 4.46: Série de observações da estação de Ferro com as médias estimadas e os intervalos de confiança empíricos, antes e depois do *change-point*.

e $\min_{2\leq k\leq 154} SIC(k) = SIC(74) = 321,06$ (Figura 4.55). Como o valor crítico para uma amostra de 151 observações (não foram contabilizadas as 5 observações em falta) é $6,791$ conclui-se que existe um *change-point* na posição 74, que corresponde a Fevereiro de 2005. Ajustou-se então o Modelo (4.2), que tem em conta este ponto de mudança (Tabela 4.18).

Nas duas subséries obtidas foi testado se em cada uma delas existia um segundo *change-point* e obteve-se para a segunda subsérie o $SIC(n) = 144,00$ e o $\min_{2\leq k\leq 80} SIC(k) =$

Tabela 4.15: Estimativas dos coeficientes do Modelo (4.1) para a estação de Golães.

Parâmetro	Estimativa
μ	$9,46$
s_{JAN}	$1,04$
s_{FEV}	$0,83$
s_{MAR}	$0,73$
s_{ABR}	$0,30$
s_{MAI}	$-0,15$
s_{JUN}	$-0,69$
s_{JUL}	$-0,82$
s_{AGO}	$-1,26$
s_{SET}	$-0,81$
s_{OUT}	$-0,44$
s_{NOV}	$0,56$
s_{DEZ}	$0,71$

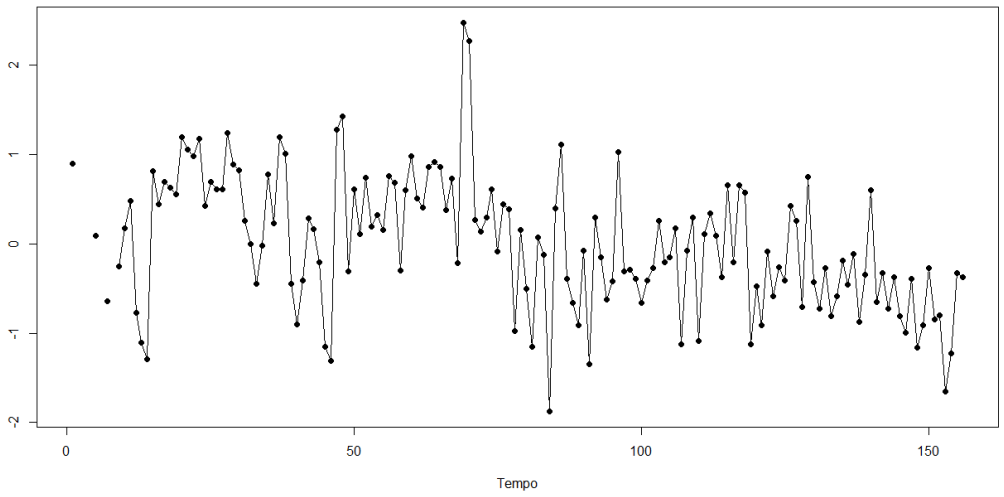


Figura 4.47: Resíduos da série da variável Oxigénio Dissolvido referente à estação de Golães depois de ajustado o Modelo (4.1).

$SIC(66) = 135,57$, e o valor crítico para 82 observações é $7,818$. Apesar da mudança ser significativa, seguindo a decisão da secção 4.2, não será considerado o segundo *change-point*. Os valores observados e estimados pelo Modelo (4.2) estão representados graficamente na Figura 4.56.

Os resíduos obtidos estão representados na Figura 4.57 e pode-se verificar uma diminuição da variância, depois da ocorrência do *change-point*. A normalidade dos resíduos não se verifica na segunda subsérie, como os histogramas da Figura 4.58 indicam, assim como os valores de prova obtidos para o teste de Shapiro Wilk foram $0,995$ e $0,017$. No

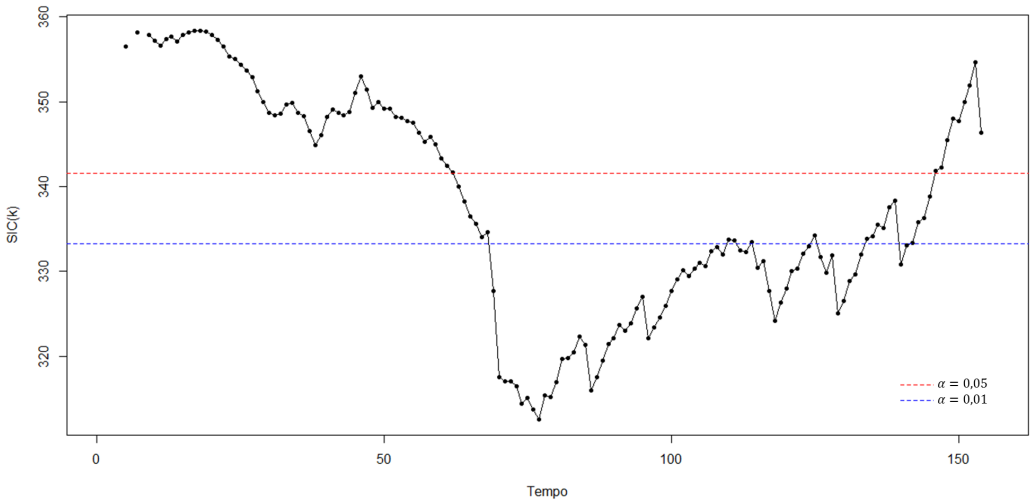


Figura 4.48: Valores de $SIC(k)$ associados à estação de Golães e as linhas de referência.

Tabela 4.16: Estimativas dos coeficientes do Modelo (4.2) para a estação de Golães.

Parâmetro	Estimativa
μ_I	$9,85$
μ_{II}	$9,11$
σ_I^2	$0,51$
σ_{II}^2	$0,34$
s_{JAN}	$1,00$
s_{FEV}	$0,81$
s_{MAR}	$0,71$
s_{ABR}	$0,28$
s_{MAI}	$-0,20$
s_{JUN}	$-0,64$
s_{JUL}	$-0,81$
s_{AGO}	$-1,22$
s_{SET}	$-0,80$
s_{OUT}	$-0,43$
s_{NOV}	$0,57$
s_{DEZ}	$0,73$

que respeita à independência dos resíduoos é agora a primeira subsérie que não satisfaz este pressuposto (Figura 4.59), sendo $\hat{\phi} = 0,290$.

Na estação de amostragem de Vizela (Santo Adrião) verificou-se uma diminuição, em média, dos valores de Oxigénio Dissolvido, assim como uma diminuição da variância, a partir de Fevereiro de 2005 (Figura 4.60).

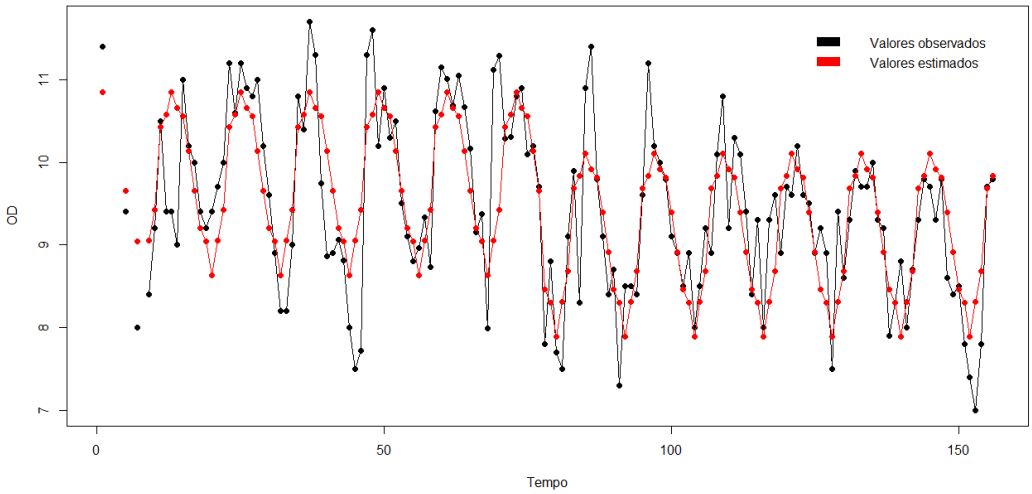


Figura 4.49: Valores observados e estimados de OD na estação de Golães.

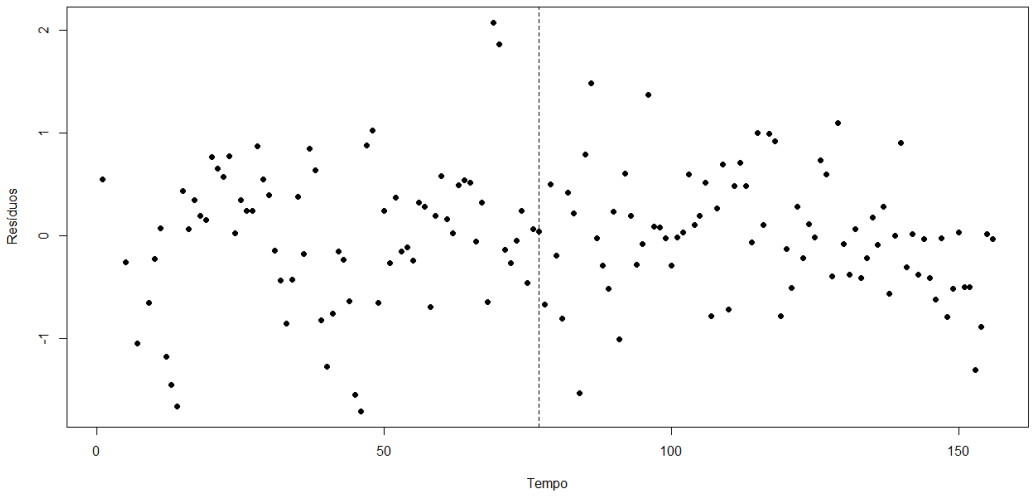


Figura 4.50: Série de resíduos associados à estação de Golães e o *change-point* identificado.

4.4 Resultados

Nesta secção pretende-se fazer um resumo dos resultados obtidos, pela análise efectuada, nas oito estações de amostragem.

Nas oito séries de Oxigénio Dissolvido associadas às estações de amostragem foram detectados *change-points* na média e na variância, simultaneamente.

Na Tabela 4.19 encontra-se um resumo para cada estação de amostragem, nomeadamente a média e a variância estimadas com base no Modelo (4.2).

Em todas as estações houve uma diminuição da variância e, no que respeita à mé-

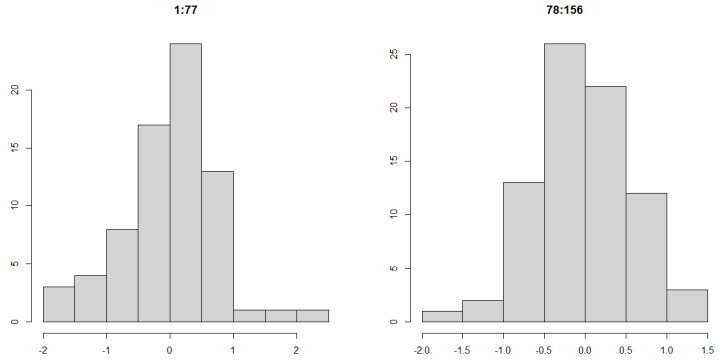


Figura 4.51: Histogramas dos resíduos associados à estação de Golães.

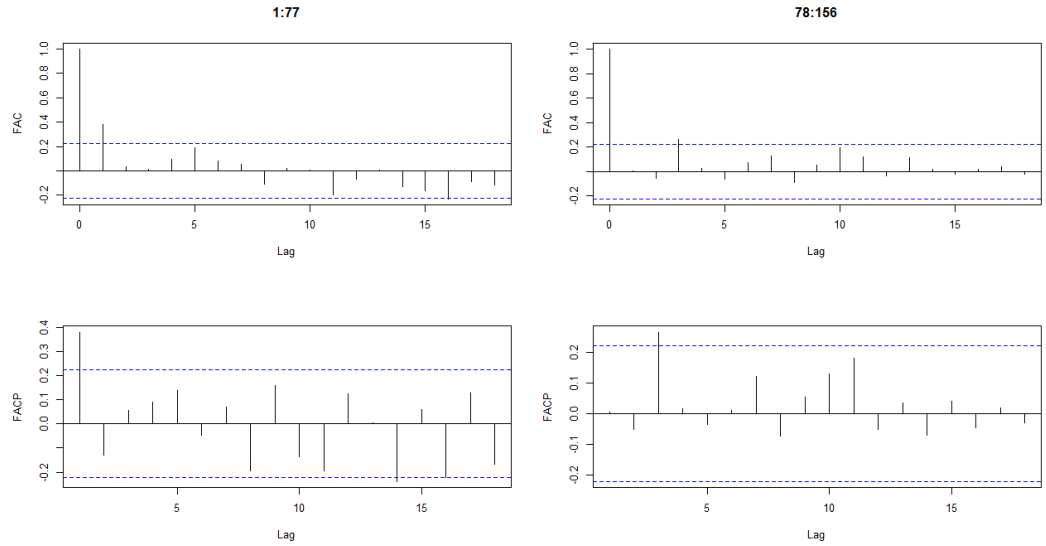


Figura 4.52: FAC e $FACP$ estimadas dos resíduos obtidos para a estação de Golães.

dia, tem-se um primeiro grupo, constituído por Cantelães, Taipas, Ferro, Golães e Vizela (Santo Adrião), que apresenta valores do Oxigénio Dissolvido em média superiores na primeira subsérie quando comparada com as observações da segunda subsérie.O segundo grupo, constituído pelas estações de amostragem Riba d'Ave, Santo Tirso e Ponte Trofa, apresenta valores médios inferiores antes do *change-point* que aumentam, em média, depois deste. Relativamente às posições dos *change-points*, no primeiro grupo estes ocorrem no final de 2004, início de 2005, e no segundo grupo ocorrem um pouco mais tarde, no final de 2005, início de 2006. Esta análise indicia a existência de dois grupos distintos de estações de amostragem, um grupo que ao longo do tempo observado apresenta uma melhoria da qualidade da água em termos de concentração média do Oxigénio Dissolvido, enquanto que o outro grupo apresenta uma degradação da qualidade da água. A identificação destes dois grupos corrobora os resultados obtidos para a mesma bacia hidrográfica

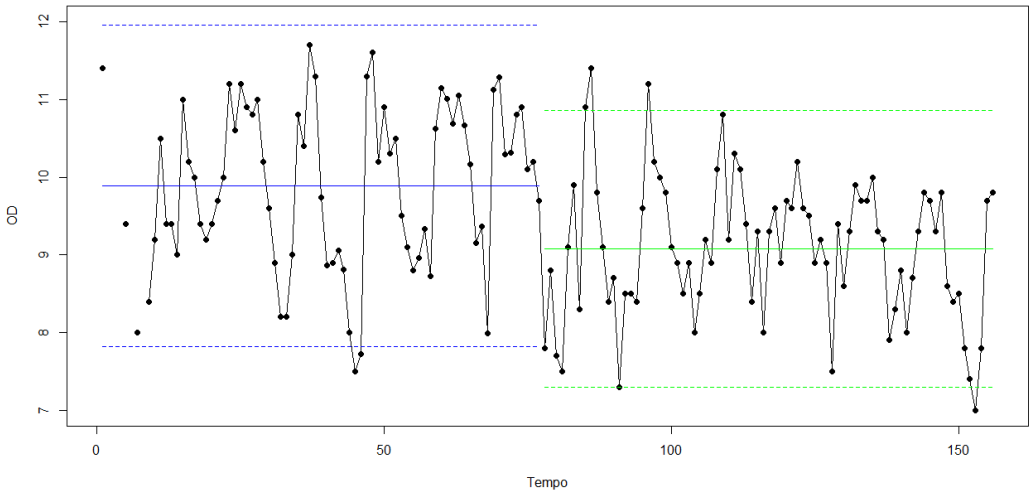


Figura 4.53: Série de observações da estação de Golães com as médias estimadas e os intervalos de confiança empíricos, antes e depois do *change-point*.

Tabela 4.17: Estimativas dos coeficientes do Modelo (4.1) para a estação de Vizela (Santo Adrião).

Parâmetro	Estimativa
μ	$9,57$
s_{JAN}	$0,95$
s_{FEV}	$1,08$
s_{MAR}	$0,60$
s_{ABR}	$0,37$
s_{MAI}	$0,04$
s_{JUN}	$-0,80$
s_{JUL}	$-1,10$
s_{AGO}	$-1,24$
s_{SET}	$-0,98$
s_{OUT}	$-0,09$
s_{NOV}	$0,33$
s_{DEZ}	$0,84$

em Gonçalves & Costa (2011).

A menor diferença de médias observada, antes e depois do *change-point*, é de $0,48$ na estação de amostragem de Riba d'Ave e a maior diferença é de $0,80$ na estação de amostragem de Cantelães. Quanto à variância, a menor diferença corresponde a $0,15$ na estação de Taipas e a maior a $1,57$ na estação de Santo Tirso.

Quanto à verificação dos pressupostos de normalidade e independência, estes nem sempre se verificaram, sendo a maior correlação observada de $0,388$. No Capítulo 5

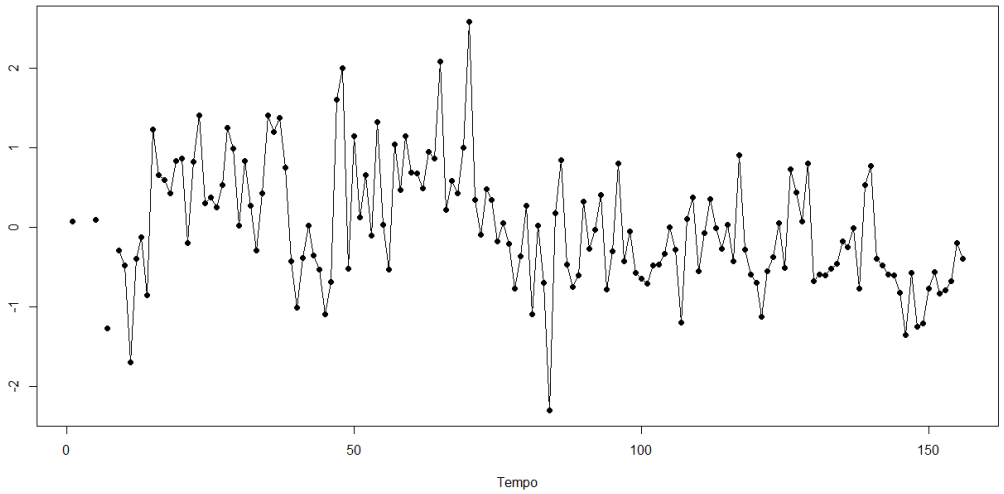


Figura 4.54: Resíduos da série da variável Oxigénio Dissolvido referente à estação de Vizela (Santo Adrião) depois de ajustado o Modelo (4.1).

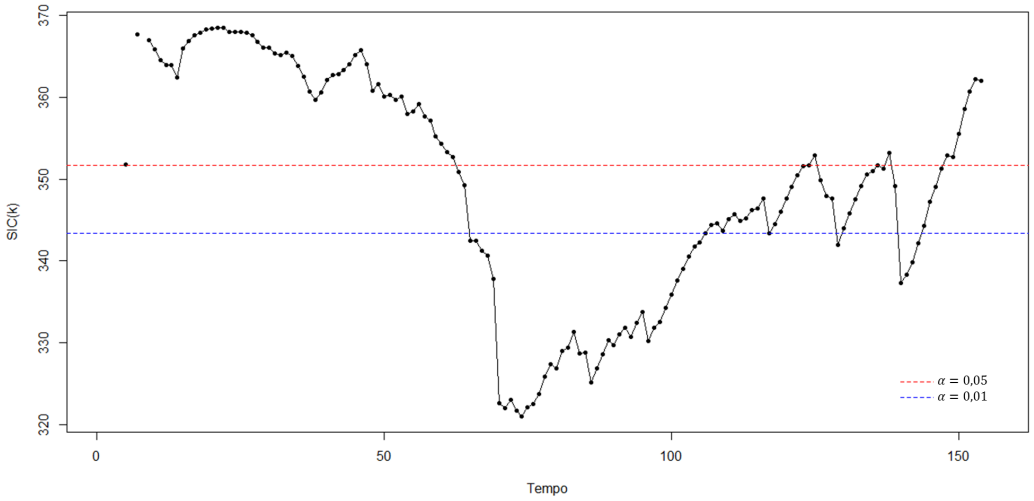


Figura 4.55: Valores de $SIC(k)$ associados à estação de Vizela (Santo Adrião) e as linhas de referência.

será delineado um estudo de simulação de modo a aferir o efeito da correlação e não normalidade das séries temporais na detecção de *change-points*, adoptando a metodologia baseada no Critério de Informação de Schwarz.

Tabela 4.18: Estimativas dos coeficientes do Modelo (4.2) para a estação de Vizela (Santo Adrião).

Parâmetro	Estimativa
μ_I	$9,96$
μ_{II}	$9,24$
σ_I^2	$0,65$
σ_{II}^2	$0,31$
s_{JAN}	$0,89$
s_{FEV}	$1,05$
s_{MAR}	$0,63$
s_{ABR}	$0,40$
s_{MAI}	$0,03$
s_{JUN}	$-0,77$
s_{JUL}	$-1,11$
s_{AGO}	$-1,21$
s_{SET}	$-0,98$
s_{OUT}	$-0,10$
s_{NOV}	$0,32$
s_{DEZ}	$0,85$

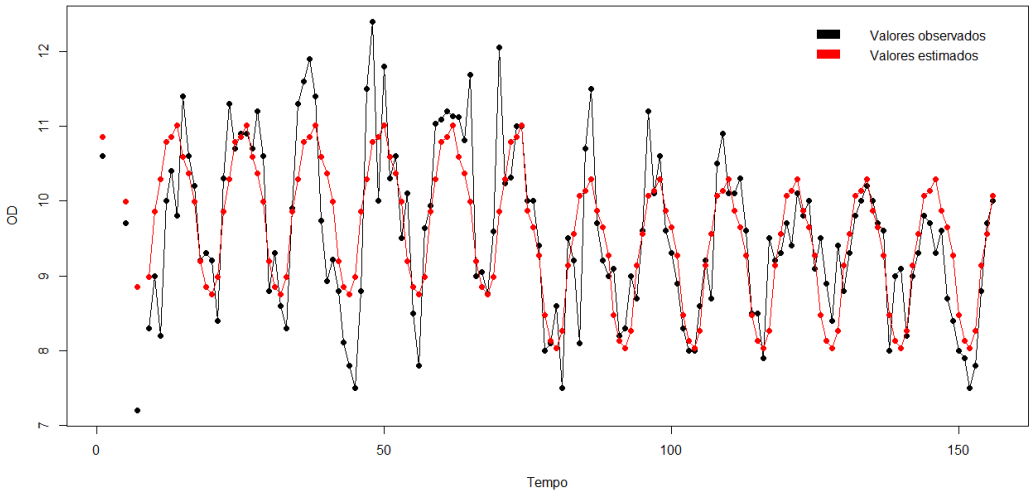


Figura 4.56: Valores observados e estimados do OD na estação de Vizela (Santo Adrião).

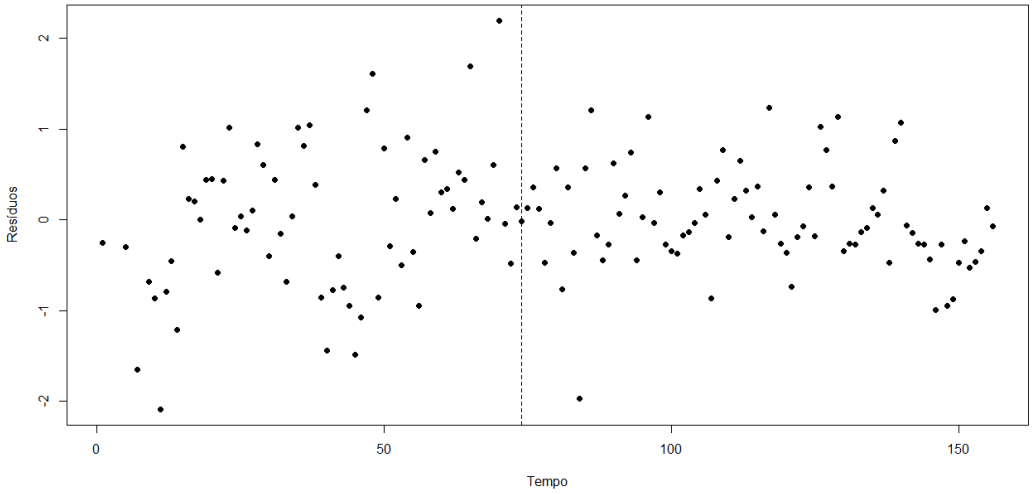


Figura 4.57: Série de resíduos associados à estação de Vizela (Santo Adrião) e o *change-point* identificado.

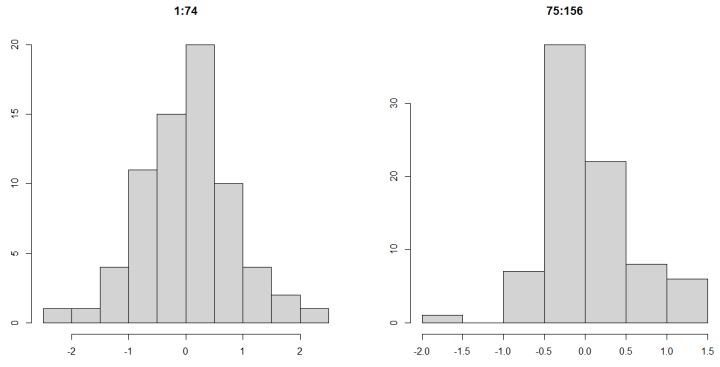


Figura 4.58: Histogramas dos resíduos associados à estação de Vizela (Santo Adrião).

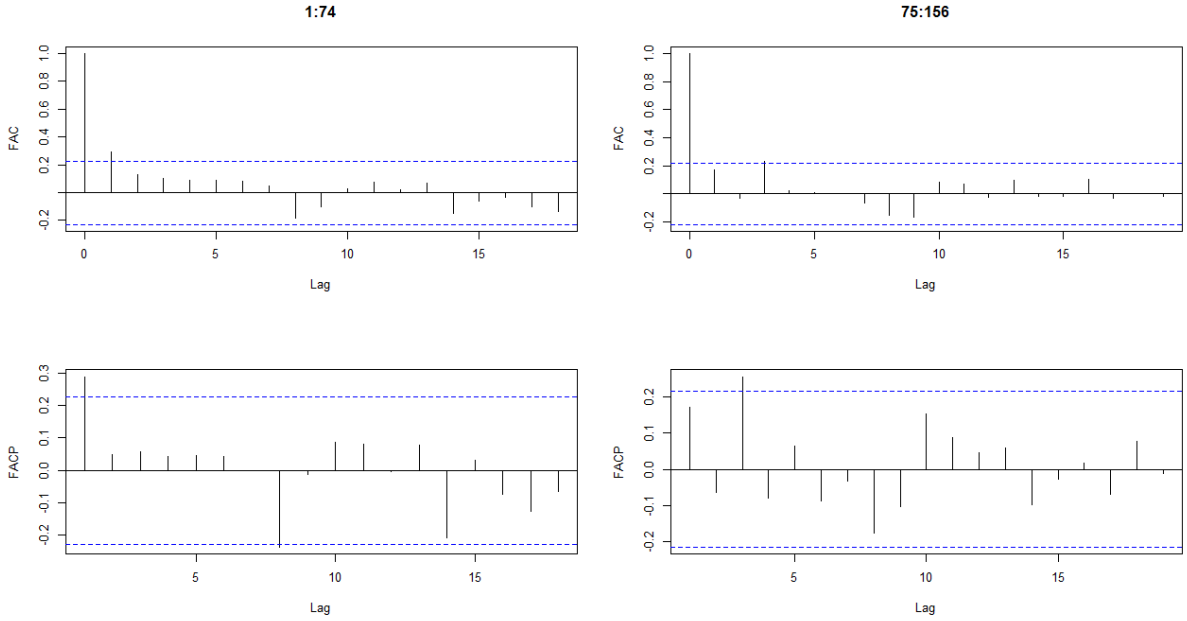


Figura 4.59: FAC e $FACP$ estimadas dos resíduos obtidos para a estação de Vizela (Santo Adrião).

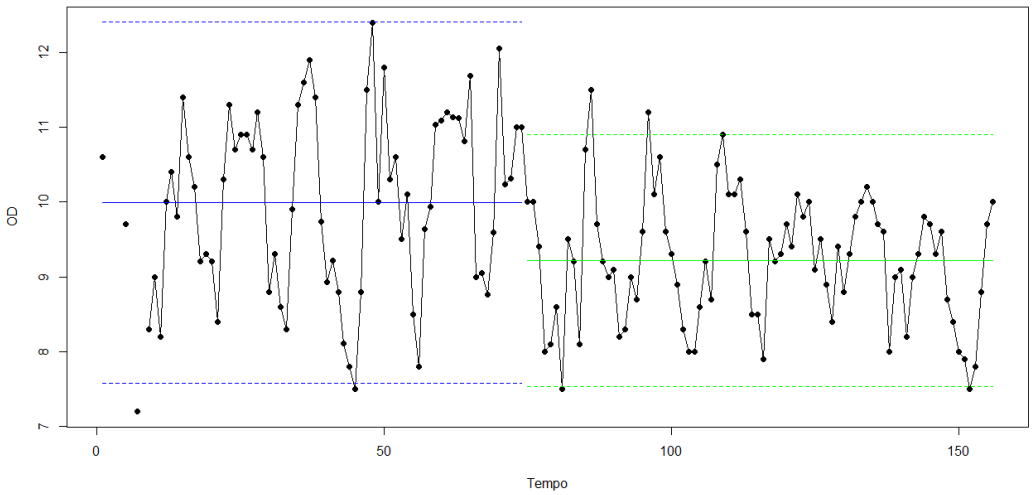


Figura 4.60: Série de observações da estação de Vizela (Santo Adrião) com as médias estimadas e os intervalos de confiança empíricos, antes e depois do *change-point*.

Tabela 4.19: Quadro resumo das características das séries.

Estação de amostragem	Série	*Change-point*	$\hat{\mu}$	$\hat{\sigma^2}$
CAN	1:73	Jan/05	10,22	0,58
	74:156		9,41	0,24
TAI	1:70	Out/04	9,62	0,49
	71:156		9,12	0,34
RAV	1:89	Maio/06	8,30	1,42
	90:156		8,78	0,46
STI	1:89	Maio/06	7,97	2,21
	90:156		8,69	0,64
PTR	1:83	Nov/05	7,78	1,70
	84:156		8,37	0,60
FER	1:70	Out/04	9,81	0,70
	71:156		9,31	0,37
GOL	1:77	Maio/05	9,85	0,51
	78:156		9,11	0,34
VSA	1:74	Fev/05	9,96	0,65
	75:156		9,24	0,31

Capítulo 5

Estudo de Simulação

O estudo de simulação foi realizado com o objectivo de analisar o comportamento da abordagem informacional com a utilização do Critério de Informação de Schwarz, a metodologia aplicada às séries de dados no Capítulo 4, quando os pressupostos de normalidade e independência não se verificam. Do ponto de vista prático é relevante a investigação do desempenho da metodologia adoptada de modo a avaliar em que medida as conclusões do Capítulo 4 são apropriadas, mesmo quando não se verificam as condições de normalidade e independência. As conclusões aqui extraídas apenas são válidas para os cenários analisados uma vez que estes são estabelecidos de modo a englobar os comportamentos das séries estudadas, nomeadamente, as diferentes alterações da média e da variância, a presença de "dependência" e a não normalidade da distribuição dos erros. As funções para a realização do estudo de simulação foram desenvolvidas recorrendo ao *software* estatístico R (R Development Core Team, 2011).

5.1 Delineamento do estudo

Num primeiro cenário base são consideradas séries sem qualquer *change-point* induzido. No segundo cenário base, é imposto um *change-point*.

No cenário sem *change-point* são geradas séries de acordo com o modelo

$$X_t = \mu + \epsilon_t, t = 1, \ldots, n, \tag{5.1}$$

onde μ é a média, ϵ_t o erro e n o tamnho da amostra.

Quando um *change-point* é imposto no segundo cenário, este é induzido no instante $t = \frac{n}{2}$. Esta opção deve-se ao facto de que os *change-points* detectados nos dados reais estudados no Capítulo 4 ocorrem em instantes centrais das séries. Neste caso as séries

são simuladas de acordo com o modelo

$$X_t = \begin{cases} \mu_I + \epsilon_t^I, t = 1, \ldots, k \\ \mu_{II} + \epsilon_t^{II}, t = k+1, \ldots, n, \end{cases} \tag{5.2}$$

onde μ_I e μ_{II} são as médias antes e depois do *change-point* e ϵ_t^I e ϵ_t^{II} são os erros com média nula e variâncias σ_I^2 e σ_{II}^2, respectivamente.

Um estudo comparativo será realizado de modo a estimar o erro de tipo I, sendo este estimado pelo nível de significância empírico calculado através da proporção de rejeições da hipótese nula (2.11), quando a série gerada não tem *change-point*. Também será realizado um estudo comparativo da potência do teste, sendo esta estimada pela proporção de rejeições da hipótese nula quando foi induzido um *change-point* na série gerada. Neste último estudo é importante também avaliar em que medida é que o *change-point* é detectado de uma forma adequada.

Em cada um dos cenários anteriores (séries sem *change-point* e séries com *change-point* induzido) adoptam-se erros com estruturas estocásticas diferenciadas (observações independentes e observações com correlação), bem como com distribuições distintas, nomeadamente, Normal e Exponencial. Esta última é considerada devido à sua forte assimetria.

Nos casos em que se consideram erros com uma estrutura de dependência, esta é assumida como sendo caracterizada por um processo autoregressivo de primeira ordem ($AR(1)$), ou seja, obedecem à estrutura $\epsilon_t = \phi\epsilon_{t-1} + a_t$, com $\mid \phi \mid < 1$, em que a_t é um ruído branco. Neste estudo será considerado $\phi = 0,3$ representando a correlação que se detectou em algumas séries do Capítulo 4. De facto, os resíduos dos modelos lineares de algumas séries apresentam FAC e $FACP$ similares a um processo $AR(1)$ com parâmetros autoregressivos na ordem de grandeza de $0,3$.

Relativamente à distribuição dos erros, a normalidade é considerada uma vez que é um dos pressupostos da metodologia adoptada e serve de referência para comparar com as séries geradas a partir de erros exponenciais. Neste caso, os erros são obtidos fazendo-se $\epsilon_t = Y_t - \frac{1}{\lambda}$, onde $Y_t \sim Exp(\lambda)$ e $E(Y_t) = \frac{1}{\lambda}$.

No estudo são consideradas amostras pequenas, $n = 50$, amostras de tamanho aproximado das séries estudadas no Capítulo 4, $n = 150$, e ainda amostras de dimensão elevada, $n = 500$.

Para cada n, é considerada uma combinação de parâmetros que caracteriza o modelo simulado. No caso de uma série sem *change-point* considera-se o vector de parâmetros $\Theta = \{\mu, \sigma^2, \phi\}$, quando existe um *change-point* induzido considera-se o vector de parâmetros $\Theta = \{\mu_I, \mu_{II}, \sigma_{\epsilon^I}^2, \sigma_{\epsilon^{II}}^2, \phi\}$.

No caso das séries sem *change-point*, a média considerada é $\mu = 0$, sem perda de generalidade. Quando um *change-point* é induzido são considerados três cenários com

diferentes discrepâncias, nomeadamente, considera-se $\mu_I = 0$ e $\mu_{II} = 0,2$, $\mu_{II} = 0,5$ e $\mu_{II} = 0,8$. Note-se que estes valores foram considerados de acordo com os resultados práticos obtidos nas séries estudadas no Capítulo 4.

Relativamente à variabilidade dos erros, consideram-se várias combinações de valores baseados nos resultados empíricos do Capítulo 4. Assim, para séries sem *change-point* admitem-se erros com variâncias $\sigma_\epsilon^2 = 0,5$, $\sigma_\epsilon^2 = 1$ e $\sigma_\epsilon^2 = 1,5$.

Quando um *change-point* é induzido consideram-se as seguintes combinações (0,6, 0,3) e (2, 0,6) para o par (σ_I^2,σ_{II}^2).

Nos casos em que se consideram erros provenientes de um processo $AR(1)$, o ruído branco a_t é simulado com variância $\sigma_a^2 = (1-\phi^2)\sigma_\epsilon^2$.

Além disso, quando os erros têm distribuição Exponencial, estes são obtidos considerando-se $\lambda = \sqrt{\frac{1}{\sigma_\epsilon^2}}$, quando não há correlação e $\lambda = \sqrt{\frac{1}{\sigma_a^2}}$, caso contrário.

O estudo de simulação está delineado de modo a gerarem-se 2000 réplicas para cada cenários, sem *change-point* e com *change-point*, e para cada combinação de parâmetros Θ, considerando as distribuições Normal e Exponencial para os erros.

A cada uma das réplicas obtidas foi aplicada a metodologia baseada no SIC considerando o ponto crítico associado a uma significância de 5% (Tabela 3.1).

5.2 Resultados

As Tabelas 5.1 e 5.2 apresentam as significâncias empíricas obtidas nas séries simuladas sem *change-point*, ou seja, valores estimados da significância do teste. Como era de esperar, nos casos em que as observações são independentes ($\phi = 0$), as significâncias empíricas obtidas são muito próximas da significância considerada de 5% mesmo para amostras de dimensão reduzida ($n = 50$).

Tabela 5.1: Significância empírica para 2000 réplicas considerando os erros com distribuição Normal.

		$n = 50$		$n = 150$		$n = 500$	
μ	σ^2	$\phi = 0$	$\phi = 0.3$	$\phi = 0$	$\phi = 0.3$	$\phi = 0$	$\phi = 0.3$
0	0,5	0,0500	0,1325	0,0515	0,1375	0,0475	0,1605
	1	0,0480	0,1135	0,0465	0,1455	0,0385	0,1630
	1,5	0,0475	0,1230	0,0420	0,1410	0,0435	0,1680

No entanto, quando a metodologia é aplicada a observações com correlação ($\phi = 0,3$) verifica-se que as significâncias empíricas são superiores à da significância adoptada, sendo mesmo o dobro ou o triplo. Assim, este estudo é concordante com os resultados referidos em Beaulieu *et al.* (2012). Salienta-se, no entanto, que o impacto da correlação acentua-se

Tabela 5.2: Significância empírica para 2000 réplicas considerando os erros com distribuição Exponencial.

		$n = 50$		$n = 150$		$n = 500$	
μ	σ^2	$\phi = 0$	$\phi = 0,3$	$\phi = 0$	$\phi = 0,3$	$\phi = 0$	$\phi = 0,3$
0	0,5	0,3065	0,4030	0,4430	0,5440	0,5705	0,6865
	1	0,3160	0,3875	0,4420	0,5545	0,5685	0,6765
	1,5	0,2980	0,3745	0,4305	0,5315	0,5850	0,6850

para amostras de maior dimensão. Da pesquisa bibliográfica efectuada não se encontrou nenhuma referência a este facto.

Os resultados obtidos evidenciam que a grandeza da variabilidade das observações não tem impacto significativo no desempenho da metodologia adoptada, uma vez que as significâncias empíricas obtidas são semelhantes para os diversos valores de σ^2_ϵ.

Os histogramas representados nas Figuras A.1, A.2, A.3, A.4, A.5 e A.6 mostram que os falsos *change-points* identificados correspondem a instantes próximos dos instantes inicial e final das séries, predominantemente quando os erros são gaussianos. De facto, quando os erros são exponenciais a detecção dos falsos *change-points* é mais uniforme no intervalo de tempo das séries. No entanto, para amostras de dimensão superiores os resultados são mais próximos dos obtidos para os erros gaussianos.

As Tabelas 5.3 e 5.4 apresentam a potência empírica do teste adoptado nos casos em que os erros têm distribuição Normal e Exponencial, respectivamente, nos cenários em que H_1 é verdadeira. Nos resultados obtidos persiste a propensão para que a percentagem de *change-points* detectados seja superior, quando existe uma estrutura de dependência nas observações. Como seria de esperar, nos casos em que as diferenças $\mu_{II} - \mu_I$ são maiores, a potência empírica é superior, isto tanto para erros normais como exponenciais.

Tabela 5.3: Potência empírica para 2000 réplicas considerando os erros com distribuição Normal.

				$n = 50$		$n = 150$		$n = 500$	
μ_1	μ_n	σ^2_1	σ^2_n	$\phi = 0$	$\phi = 0,3$	$\phi = 0$	$\phi = 0,3$	$\phi = 0$	$\phi = 0,3$
0	0,2	0,6	0,3	0,1295	0,2795	0,4745	0,7095	0,9950	0,9995
		2	0,6	0,3095	0,4425	0,9385	0,9530	1,0000	1,0000
0	0,5	0,6	0,3	0,3555	0,7160	0,9620	0,9960	1,0000	1,0000
		2	0,6	0,4090	0,6210	0,9850	0,9950	1,0000	1,0000
0	0,8	0,6	0,3	0,7725	0,9725	1,0000	1,0000	1,0000	1,0000
		2	0,6	0,6170	0,8655	0,9980	1,0000	1,0000	1,0000

Note-se que, relativamente à influência da diferença nas variâncias, não está patente um padrão global relativamente ao desempenho da metodologia. Contudo, os resultados

Tabela 5.4: Potência empírica para 2000 réplicas considerando os erros com distribuição Exponencial.

				$n = 50$		$n = 150$		$n = 500$	
μ_I	μ_{II}	σ_I^2	σ_{II}^2	$\phi = 0$	$\phi = 0,3$	$\phi = 0$	$\phi = 0,3$	$\phi = 0$	$\phi = 0,3$
0	0,2	0,6	0,3	0,4015	0,5205	0,7165	0,8885	0,9980	1,0000
		2	0,6	0,5380	0,6080	0,9105	0,9580	1,0000	1,0000
0	0,5	0,6	0,3	0,6005	0,8640	0,992	1,0000	1,0000	1,0000
		2	0,6	0,5950	0,7855	0,9940	0,9985	1,0000	1,0000
0	0,8	0,6	0,3	0,9185	0,9950	1,0000	1,0000	1,0000	1,0000
		2	0,6	0,7810	0,9380	1,0000	1,0000	1,0000	1,0000

obtidos indicam que quando a diferença das médias é menor, a potência empírica é superior quando a diferença das variâncias é superior. Quando a diferença das médias é de $0,8$ e está associada a uma maior alteração na grandeza das variâncias, a potência empírica tende a diminuir.

Verifica-se que a metodologia adoptada apresenta um desempenho muito bom para amostras grandes ($n = 500$), apresentando potências empíricas próximas de 100% em quase todos os cenários.

Nas Figuras A.7, A.8, A.12, A.10, A.11 e A.12 apresentam-se os histogramas relativos aos *change-points* detectados, quando H_1 é verdadeira. De modo a permitir uma análise mais aprofundada do desempenho da metodologia proposta quanto à precisão da localização dos *change-points* detectados, foram calculadas as percentagens de *change-points* detectados "perto" do verdadeiro *change-point* simulado de entre as 2000 réplicas. Para este efeito estabeleceram-se limites inferiores e superiores entre os quais um *change-point* foi detectado, com uma precisão razoável. Estes limites foram estabelecidos com base na experiência empírica dos dados analisados no Capítulo 4. Assim, consideraram-se os intervalos $19 - 31$, $63 - 87$ e $226 - 274$ para as séries de dimensão 50, 150 e 500, respectivamente.

Considerando os *change-points* detectados, calculou-se a percentagem dos que se situam dentro dos limites estabelecidos, de modo a permitir a comparação do desempenho da metodologia, em particular nos cenários de observações não correlacionadas ou com correlação ($\phi = 0,3$). Nas Tabelas A.1 e A.2 apresentam-se os resultados obtidos.

Os resultados obtidos mantêm a tendência de que nos casos em que as observações são correlacionadas, a metodologia adoptada apresenta melhores desempenhos, mesmo considerando os change-points situados nos limites estabelecidos.

Comparando os resultados para as séries com erros normais e exponenciais verifica-se que, neste último caso, o desempenho é inferior comparativamente ao caso das observações normais.

Globalmente podemos dizer que a diferença dos desempenhos nos cenários, com e sem correlação, é atenuada quando analisamos as percentagens de *change-points* localizados nos limites, de entre os detectados, principalmente para amostras de menor dimensão ($n = 50$).

Capítulo 6

Conclusões

Os métodos de análise de *change-points* são vários e a sua utilização está dependente da série temporal em estudo. Atendendo à natureza dos dados de qualidade da água em estudo e à análise exploratória realizada no Capítulo 4, a abordagem informacional com a utilização do Critério de Informação de Schwarz (SIC) surge como adequada para alcançar os objectivos de detectar os *change-points* nas séries relativas à concentração do Oxigénio Dissolvido.

Como as séries de dados em estudo apresentam um comportamento sazonal foi necessária uma atenção especial, abordando-se esta característica através de modelos lineares. Esta abordagem surge como uma estratégia para dar resposta ao problema da existência da componente sazonal em séries temporais, principalmente na análise de dados ambientais.

Foram detectados *change-points* na média e na variância, simultaneamente, nas séries de dados das oito estações de amostragem da bacia hidrográfica do Rio Ave. A análise do comportamento das séries temporais permitiu verificar que em todas as estações de amostragem houve uma diminuição da variabilidade. Em cinco das estações, nomeadamente Cantelães, Taipas, Ferro, Golães e Vizela (Santo Adrião), verificou-se uma diminuição da média, que se traduz numa degradação da qualidade da água, considerando apenas a concentração do Oxigénio Dissolvido. Nas restantes estações, Riba d'Ave, Santo Tirso e Ponte Trofa, verificou-se uma melhoria da qualidade da água, contudo, estas três estações de amostragem continuam a apresentar as menores concentrações médias de DO, isto é, apresentam a água com menor qualidade. Os resultados obtidos estão de acordo com os resultados apresentados por Costa & Gonçalves (2011).

A análise dos resíduos dos modelos ajustados permitiu concluir que alguns dos pressupostos da metodologia adoptada não se verificaram na totalidade em algumas séries, nomeadamente a independência e a normalidade dos erros. Neste sentido, o estudo de simulação desenvolvido no Capítulo 5 permitiu melhor aferir o impacto da não verificação

destes pressupostos na detecção de *change-points*. A principal conclusão deste estudo é que na presença de correlação, mesmo que fraca, a metodologia tende a detectar falsos *change-points*, ou seja, a significância real é superior à considerada para efeitos da determinação do ponto crítico. Por exemplo, para amostras de dimensão 150 (dimensão próxima das séries estudadas) a significância empírica obtida é aproximadamente 14%, considerando um ponto crítico associado a uma significância de 5%. Contudo, atendendo aos gráficos relativos à representação dos valores de $SIC(k)$, mesmo considerando um ponto crítico associado a uma significância de 1% a metodologia continua a detectar os *change-points* em todas as séries relativas às estações de amostragem, excepto na série relativa à estação de amostragem de Taipas. Este facto leva-nos a concluir que a fraca correlação identificada em resíduos de alguns modelos não colocam em causa a validade da análise feita.

Não foi possível com as diligências efectuadas junto das entidades oficiais identificar factores ou acções concretos que possam estar na origem dos *change-points* identificados. No entanto, um resultado consistente das análises efectuadas foi a diminuição da variabilidade da concentração do Oxigénio Dissolvido, facto que pode estar associado ao melhoramento dos instrumentos de medida.

6.1 Sugestões para trabalho futuro

Apesar do crescente desenvolvimento dos métodos de análise de *change-points*, estes ainda apresentam limitações, porque a maioria das metodologias de análise de *change-points* é baseada nos pressupostos de normalidade e independência das observações das séries temporais. Como foi constatado no Capítulo 4, estes pressupostos nem sempre se verificam pois as séries temporais, nomeadamente as relativas a dados ambientais, apresentam características como a sazonalidade, a correlação, a não normalidade e a não estacionaridade de vários tipos. A incorporação destas características nas metodologias de *change-point* é, assim, muito importante no estudo de séries temporais de dados reais, que usualmente não apresentam comportamentos tão restritos como os de independência e de normalidade impostos pela maioria destas abordagens. O desenvolvimento de novos métodos ou a extensão das metodologias existentes para outros tipos de distribuições, através de abordagens paramétricas e não paramétricas, é desta forma essencial.

Uma outra limitação destas metodologias é a necessidade de determinar as distribuições assimptóticas para a detecção de *change-points* com significância estatística. Assim, para a determinação destas distribuições é necessário implementar novos estudos de simulação e desenvolver novas técnicas computacionais.

Também o estudo da análise de *change-points* num contexto multivariado ainda está

muito pouco desenvolvido. A utilização de metodologias estatísticas da análise multivariada nesta área pode trazer vantagens, na medida em que adiciona uma maior informação ao processo de detecção de *change-points* em problemas envolvendo diversas variáveis, o que acontece na maioria das situações envolvendo dados reais.

Bibliografia

[1] Adichie, J. N. (1967). Asymptotic efficiency of a class of nonparametric tests for regression parameters. *Ann. Math. Statist.* **38**, 884-893.

[2] Alpuim, T. (1998). *Séries Temporais*. Associação dos Estudantes da Faculdade de Ciências de Lisboa, 2ª edição.

[3] Antoch, J., Hušková, M., Prášková, Z. (1997). Effect of dependence on statistics for determination of change. *J. Statist. Plan. Inf.* **60**, 291 – 310.

[4] Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEE Trans. Auto. Control.* **19**, 716 - 723.

[5] Barratt, B., Atkinson, R., Anderson, H. R., Beevers, S., Kelly, F., Mudway, I., Wilkinson, P. (2007). Investigation into the use of the CUSUM technique in identifying changes in mean air pollution levels following introduction of a traffic management scheme. *Atmospheric Environment.* **41**, 1784-1791.

[6] Beaulieu, C., Chen, J., Sarmiento, J.L. (2012). Change-point analysis as a tool to detect abrupt climate variations. *Phil. Trans. R. Soc. A.* **370**, 1228-1249.

[7] Bhattacharya, G.K., Johnson, R.A. (1968). Nonparametric tests for shift at an unknown time point. *Annals of Mathematical Statistics.* **39**, 1731-1743.

[8] Box, G. E. P., Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B.* **26**, 211-252.

[9] Bozdogan, H. (1987). Model selection and Akaike's Information criterion (AIC): The general theory and its analytical extension. *Psychometrika.* **52**, 345-370.

[10] Bozdogan, H., Sclove, S.L., and Gupta, A.K. (1994). AIC-Replacements for some multivariate tests of homogeneity with applications in multisample clustering and variable selection. In Proceedings of the *First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach.* V. 2. Kluwer Academic, Dordrecht, 199-232.

[11] Chalton, D.O., Troskie, C.G. (1999). Parameter changes in the multiple regression model with autocorrelated errors: Bayesian analysis. *Communication in Statistics-Theory and Methods.* **28**, 137-142.

[12] Chen, J. (1998). Testing for a change point in linear regression models. *Communications in Statistics – Theory and Methods.* **27**, 2481-2493.

[13] Chen, J., Gupta, A.K. (1995). Likelihood procedure for testing change points hypothesis for multivariate Gaussian model. *Random Operators and Stochastic Equations.* **3**, 235-244.

[14] Chen, J., Gupta, A. K. (1997). Testing and Locating variance Changepoints with Application to Stock Prices. *Journal of the American Statistical Association.* **92**, No. 438, 739-747.

[15] Chen, J., Gupta, A. K. (1999). Change point analysis of a Gaussian model. *Statistical Papers.* **40**, 323-333.

[16] Chen, J., Gupta, A. K. (2001). On change point detection and estimation. *Communications in Statistics-Simulation and Computation.* **30**, 665-697.

[17] Chen, J., Gupta, A.K. (2012). *Parametric Statistical Change Point analysis.* Second Edition, Birkhauser.

[18] Chernoff, H., Zacks, S. (1964). Estimating the current mean of a normal distribution which is subject to changes in time. *Annals of Mathematical Statistics.* **35**, 999-1018.

[19] Chin, Choy, J.H., Broemeling, L.D. (1980). Some Bayesian inferences for a changing linear model. *Technometrics.* **22**, 71-78.

[20] Chu, H. J., Pan, T. Y., Liou, J. J. (2012). Change-point detection of long-duration extreme precipitation and the effect on hydrologic design: a case study of south Taiwan. *Stoch Environ Risk Assess.* (doi: 10.1007/s00477-012-05066-0)

[21] Costa, M., Gonçalves, A. M. (2011). Clustering and forecasting of dissolved oxygen concentration on a river basin. *Stochastic Environmental Research and Risk Assessment.* **25**, 151-163.

[22] El-Shaarawi, A.H., Esterby, S.R. (1982). Inference About the Point of Change in A Regression Model With A Stationary Error Process. *Time Series Methods in Hydrosciences - Proceedings of an International Conference Held at Canada Centre for Inland Waters.* 55-67.

[23] Ferreira, P.E. (1975). A Bayesian analysis of a switching regression model: Known number of regimes. *Journal of the American Statistical Association.* **70**, 370-374.

[24] Gardner, L.A. (1969). On detection change in the mean of normal variates. *Annls of Mathematical Statistics.* **40**, 116-126.

[25] Gerard-Marchant, P. G. F., Stooksbury, D. E., Seymour, L. (2008). Methods for starting the detection of undocumented multiple changepoints. *J. Clim.* **21**, 4887-4899.

[26] Gonçalves, A. M., Alpuim, T. (2011). Water quality monitoring using cluster analysis and linear models. *Environmetrics.* **22**, 933-945.

[27] Gonçalves, A. M., Costa, M. (2011). Application of Change-Point Detection to a Structural Component of Water Quality Variables. Em proceedings of the International Conference on Numerical Analysis and Applied Mathematics, AIP Conference Proceedings **1389**, 1565-1568.

[28] Gonçalves, A. M., Costa, M. (2012). Predicting seasonal and hydro-meteorological impact in environmental variables modelling via Kalman filtering. *Stochastic Environmental Research and Risk Assessment.* (doi: 10.1007/s00477-012-0640-7)

[29] Hájek, J. (1962). Asymptotically most powerful rank order tests. *Ann. Math. Statist.* **33**, 1124-1147.

[30] Hawkins, D.M. (1977). Testing a sequence of observations for a shift in location. *Journal of the American Statistical.* **72**, 180-186.

[31] Hawkins, D.M., Zamba, K.D. (2005). Statistical process control for shifts in mean or variance using a changepoint formulation. *Technometrics.* **47**, 164-173.

[32] Henderson, R. (1986). Change-point problem with correlated observations, with an application in material accountancy. *Technometrics.* **28**, 381-389.

[33] Hsu, D. A. (1977). Tests for Variance Shift at an Unknown Time Point. *Journal of the Royal Statistical Society.* **26**, No. 3, 279-284.

[34] Inclán, C. (1993). Detection of multiple changes of variance using posterior odds. *Journal of Business and Economics Statistics.* **11**, 189-300.

[35] Inclán, C., Tiao, G. C. (1994). Use of comulative sums of squares for retrospective detection of changes of variance. *Journal of the American Statistical Association.* **89**, 913-923.

[36] James, B., James, K.L., Siegmund, D. (1992). Asymptotic approximations for likelihood ratio tests and confidence regions for a change point in the mean of a multivariate normal distribution. *Statistica Sinica.* **2**, 69-90.

[37] Jarušková, D. (1996). Change-Point Detection in Meteorological Measurement. *Mon Weather Rev.* **124**, 1535-1543.

[38] Jarušková, D. (1997). Some problems with application of change-point detection methods to environmental data. *Environmetrics.* **8**, 469-483.

[39] Jarušková, D. (2007). Maximum log-likelihood ratio test for a change in three parameter Weibull distribution. *J. Stat. Plann. Inf.* **137**, 1805-1815.

[40] Jarušková, D., Rencová, M. (2008). Analysis of annual maximal and minimal temperatures for some European cities by change point methods. *Environmetrics.* **19**, 221-233.

[41] Jarušková, D. (2010). Asymptotic behavior of a test statistic for detection of change in mean of vectors. *Journal of statistical Planning and Inference.* **140**, 616-625.

[42] Kim, D.C. (1991). A Bayesian significance test of the stationarity of regression parameters. *Biometrika.* **78**, 667-675.

[43] Kitagawa, G. (1979). On the use of AIC for the detection of outliers. *Technometrics.* **21**, 193-199.

[44] Lund, R., Reeves, J. (2002). Detection of Undocumented Changepoints: A Revision of the Two-Phase Regression Model. *Journal of Climate.* **15**, 2547-2554.

[45] Lund, R., Wang, X., Lu, Q., Reeves, J. Gallagher, C., Feng, Y. (2007). Changepoint Detection in Periodic and Autocorrelation Time Series. *J. Climate.* **20**, 5178–5190.

[46] Page, E.S. (1954). Continuous inspection schemes. *Biometrika.* **41**, 100-116.

[47] Page, E.S. (1955). A test for a change in a parameter occurring at an unknown point. *Biometrika.* **42**, 523-527.

[48] Page, E.S. (1957). On problem in which a change in a parameter occurs at an unknown points. *Biometrika.* **44**, 248-252.

[49] Pettitt, A.N. (1980). A simple cumulative sum type statistic for the change point problem with zero-one observations. *Biometrika.* **67**, 79-84.

[50] Quandt, R.E. (1958). The estimation of the parameters of a linear regression system obeys two separate regimes. *Journal of the American Statistical Association.* **53**, 873-880.

[51] Quandt, R.E. (1960). Tests of the hypothesis that a linear regression system obeys two separate regimes. *Journal of the American Statistical Association.* **55**, 324-330.

[52] R Development Core Team (2011). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. http//www.R-project.org.

[53] Rao, C.R., Wu, Y. (1989). A strongly consistent procedure for model selection in a regression problem. *Biometrika.* **76**, 369-374.

[54] Robbins, M., Gallagher, C., Lund, R., Aue, A. (2011). Mean shift testing in correlated data. *J. Time Ser. Anal.* **32**, 498-511.

[55] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* **6**, 461-464.

[56] Seidel, D. J., Lanzante, J. R. (2004). An assessment of three alternatives to linear trends for characterizing global atmospheric temperature changes. *J. Geophys. Res. – Atmos.* **109**, D14108.

[57] Sen. A.K., Srivastava, M.S. (1975). On tests for detecting change in mean. *Annals of Statistics.* **3**, 98-108.

[58] Shapiro, S.S., Wilk, M.B. (1965). An analysis of variance test for normality (complete samples). *Biometrika.* **52** (3/4), 591-611.

[59] Srivastava, M. S., Worsley, K. J. (1986). Likelihood ratio test for a change in the multivariate normal mean. *J. Amer. Statist. Assoc.* **81**, 199-204.

[60] Tang, S. M., MacNeill, I. B. (1993). The effect of serial correlation on tests for parameter change at unknown time. *Ann. Stat.* **21**, 552-575.

[61] Vostrikova, L. J. (1981). Detecting "disorder" in multidimensional random processes. *Soviet Mathematics Doklady.* **24**, 55-59.

[62] Wang, Y. Z. (1995). Jump and sharp cusp detection by wavelets. *Biometrika.* **82**, 385-397.

[63] Wang, Y. Z. (2008). Accounting for autocorrelation in detecting mean shifts in cliate data series using the penalized maximal t or F test. *J. Appl. Meteorol. Climatol.* **47**, 2423-2444.

[64] Worsley, K.J. (1979). On the likelihood ratio test for a shift in location of normal populations. *Journal of the American Statistical Association.* **74**, 365-367

[65] Worsley, K.J. (1983). The power of likelihood ratio and cumulative sum tests for a change in a binomial probability. *Biometrika.* **70**, 455-464.

[66] Zhao, X., Chu, P. S. (2006). Bayesian multiple changepoint analysis of hurricane activity in the eastern North Pacific: a Marcov chain Monte Carlo approach. *J. Clim.* **19**, 564-578.

[67] Zhao, W.Z., Tian, Z., Xia, Z.M. (2010). Ratio test for variance change point in linear process with long memory. *Stat. Papers.* **51**, 397-407.

Apêndice A

Apêndice

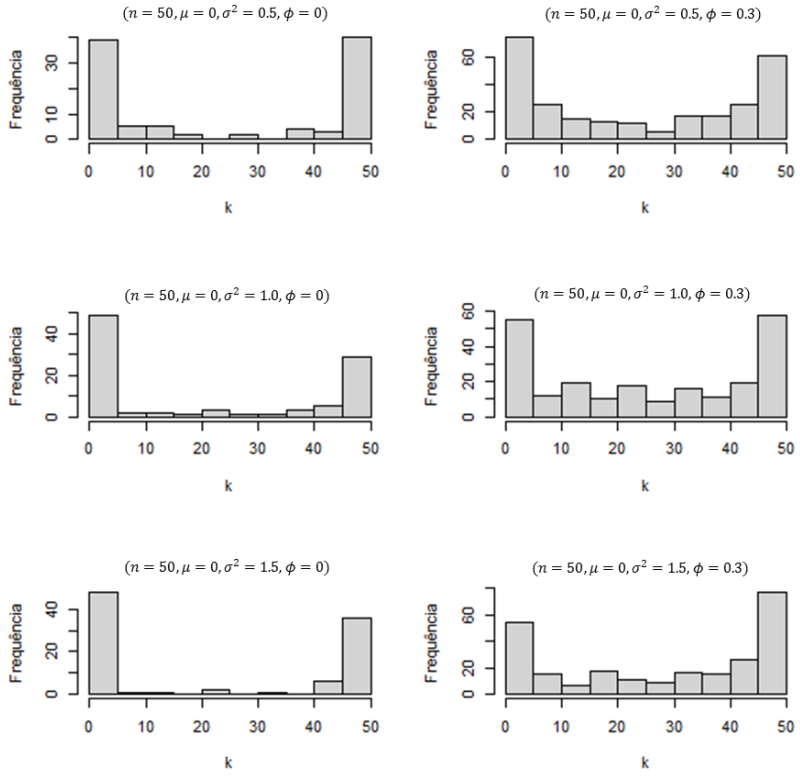


Figura A.1: Histogramas dos falsos *change-points* identificados considerando os erros com distribuição Normal e $n = 50$.

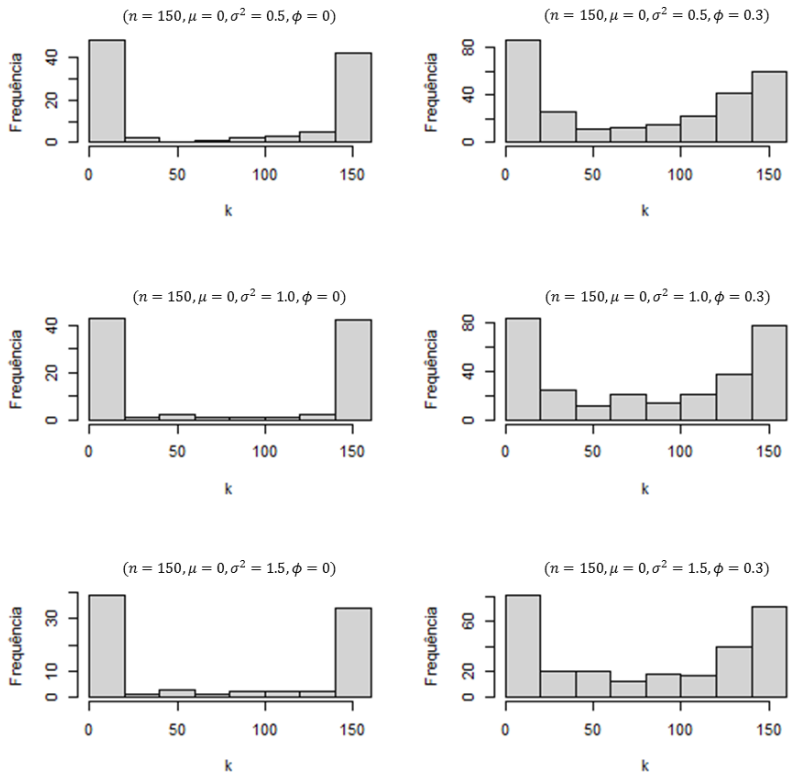


Figura A.2: Histogramas dos falsos *change-points* identificados considerando os erros com distribuição Normal e $n = 150$.

Tabela A.1: Percentagem de *change-points* identificados nos limites estabelecidos considerando os erros com distribuição Normal.

				$n = 50$		$n = 150$		$n = 500$	
μ_I	μ_{II}	σ_I^2	σ_{II}^2	$\phi = 0$	$\phi = 0,3$	$\phi = 0$	$\phi = 0,3$	$\phi = 0$	$\phi = 0,3$
0	0,2	0,6	0,3	0,0545	0,1265	0,3355	0,4260	0,8885	0,8745
				(0,4208)	(0,4508)	(0,7060)	(0,6004)	(0,8925)	(0,8749)
		2	0,6	0,2135	0,2725	0,8250	0,7730	0,9835	0,9750
				(0,6898)	(0,6147)	(0,8785)	(0,8106)	(0,9835)	(0,9750)
0	0,5	0,6	0,3	0,2515	0,5335	0,8390	0,8940	0,9895	0,9860
				(0,7075)	(0,7451)	(0,8716)	(0,8971)	(0,9895)	(0,9860)
		2	0,6	0,3010	0,4500	0,8960	0,8750	0,9935	0,9900
				(0,7359)	(0,7246)	(0,9096)	(0,8789)	(0,9935)	(0,9900)
0	0,8	0,6	0,3	0,6890	0,8855	0,9800	0,9865	1,0000	1,0000
				(0,8913)	(0,9100)	(0,9800)	(0,9865)	(1,0000)	(1,0000)
		2	0,6	0,5115	0,7130	0,9590	0,9385	0,9970	0,9955
				(0,8282)	(0,8238)	(0,9604)	(0,9385)	(0,9970)	(0,9955)

()Percentagem dos *change-points* detectados que se situam dentro dos limites.

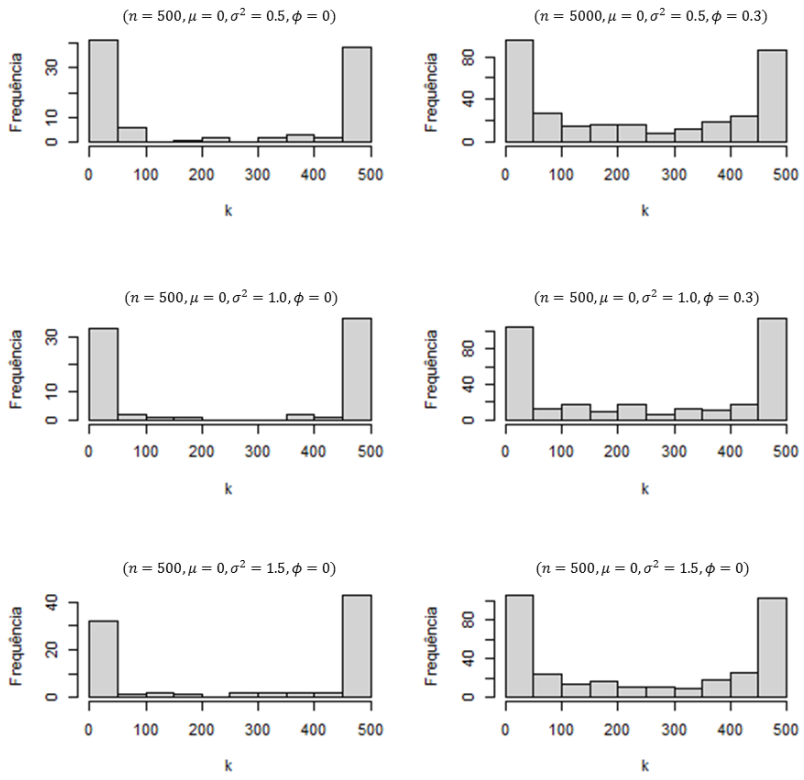


Figura A.3: Histogramas dos falsos *change-points* identificados considerando os erros com distribuição Normal e $n = 500$.

Tabela A.2: Percentagem de *change-points* identificados nos limites estabelecidos considerando os erros com distribuição Exponencial.

				$n = 50$		$n = 150$		$n = 500$	
μ_I	μ_{II}	σ_I^2	σ_{II}^2	$\phi = 0$	$\phi = 0,3$	$\phi = 0$	$\phi = 0,3$	$\phi = 0$	$\phi = 0,3$
0	0,2	0,6	0,3	0,1360	0,2075	0,2950	0,4520	0,7485	0,8430
				(0,3387)	(0,3987)	(0,4117)	(0,5087)	(0,7500)	(0,8430)
		2	0,6	0,2465	0,3125	0,5715	0,6390	0,8720	0,9195
				(0,4582)	(0,5140)	(0,6277)	(0,6670)	(0,8720)	(0,9195)
0	0,5	0,6	0,3	0,3645	0,6195	0,8250	0,9090	0,9830	0,9970
				(0,6070)	(0,7170)	(0,8317)	(0,9090)	(0,9830)	(0,9970)
		2	0,6	0,3480	0,5305	0,8030	0,8670	0,9450	0,9760
				(0,5849)	(0,6754)	(0,9078)	(0,8683)	(0,9450)	(0,9760)
0	0,8	0,6	0,3	0,7770	0,9090	0,9630	0,9825	0,9980	0,9990
				(0,8459)	(0,9136)	(0,9630)	(0,9825)	(0,9980)	(0,9990)
		2	0,6	0,5885	0,7650	0,9195	0,9450	0,9715	0,9955
				(0,7535)	(0,8156)	(0,9195)	(0,9450)	(0,9715)	(0,9955)

()Percentagem dos *change-points* detectados que se situam dentro dos limites.

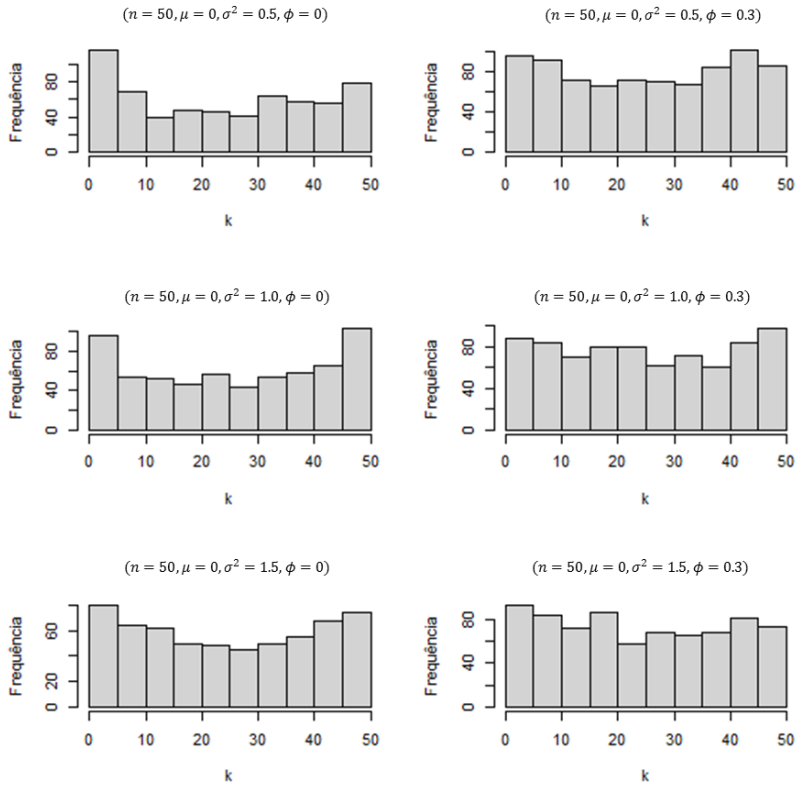


Figura A.4: Histogramas dos falsos *change-points* identificados considerando os erros com distribuição Exponencial e $n = 50$.

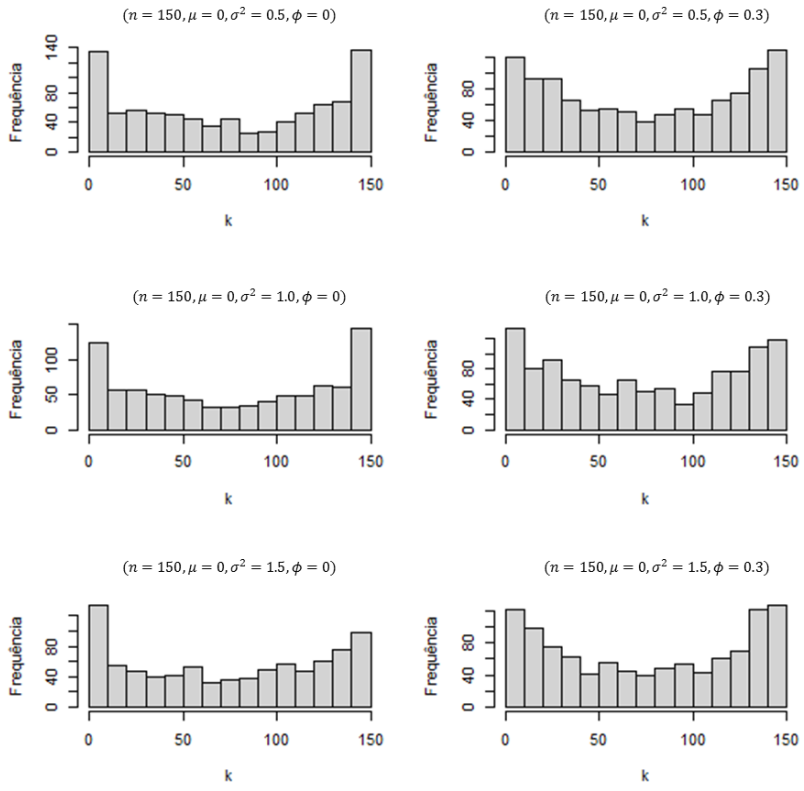


Figura A.5: Histogramas dos falsos *change-points* identificados considerando os erros com distribuição Exponencial e $n = 150$.

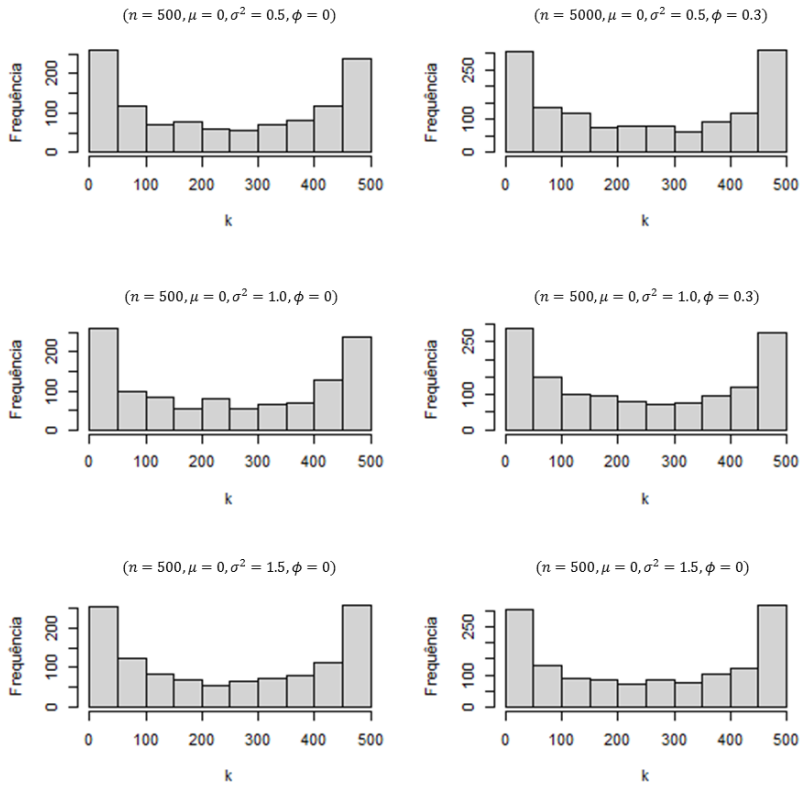


Figura A.6: Histogramas dos falsos *change-points* identificados considerando os erros com distribuição Exponencial e $n = 500$.

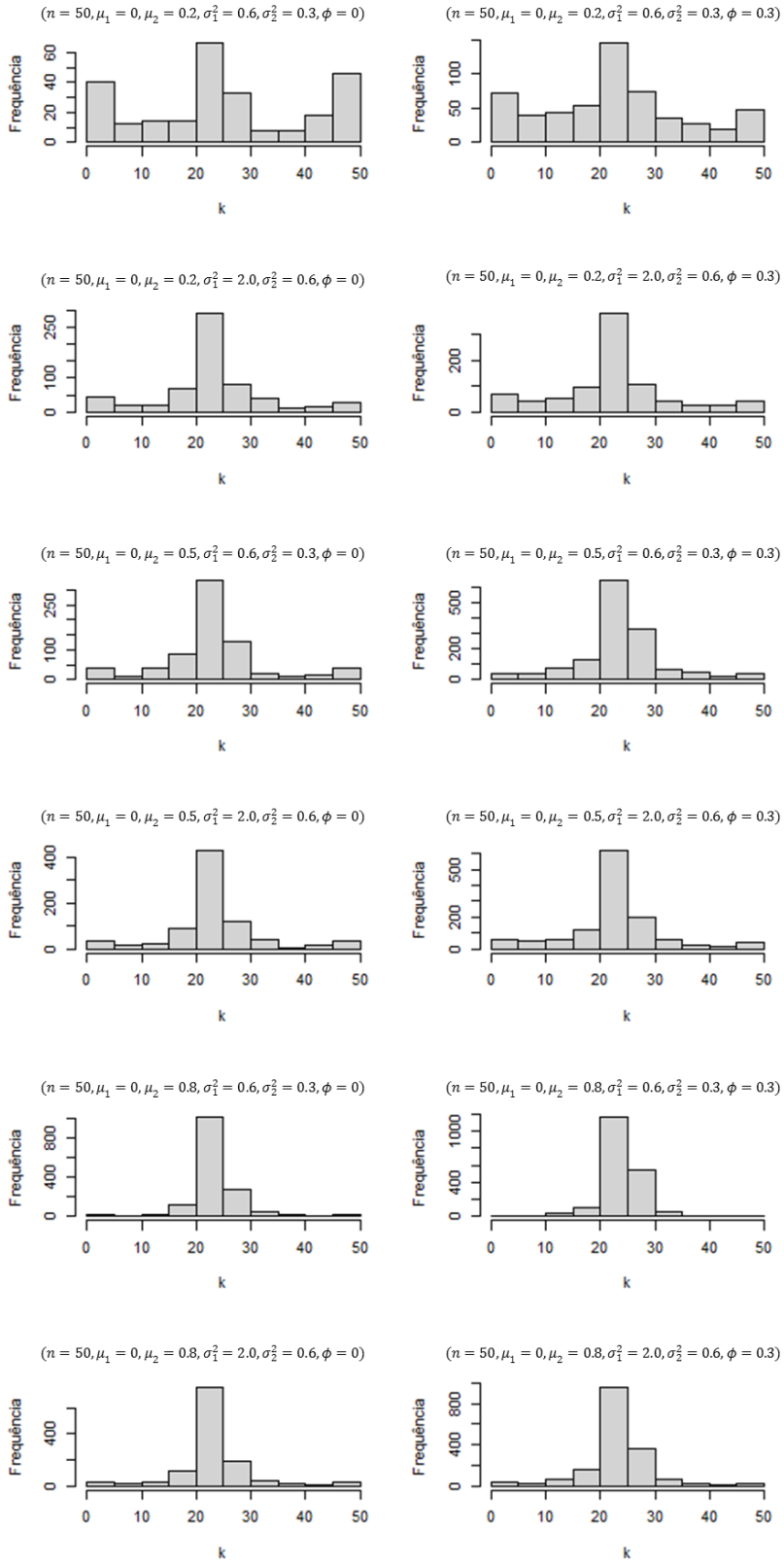


Figura A.7: Histogramas dos *change-points* identificados considerando os erros com distribuição Normal e $n = 50$.

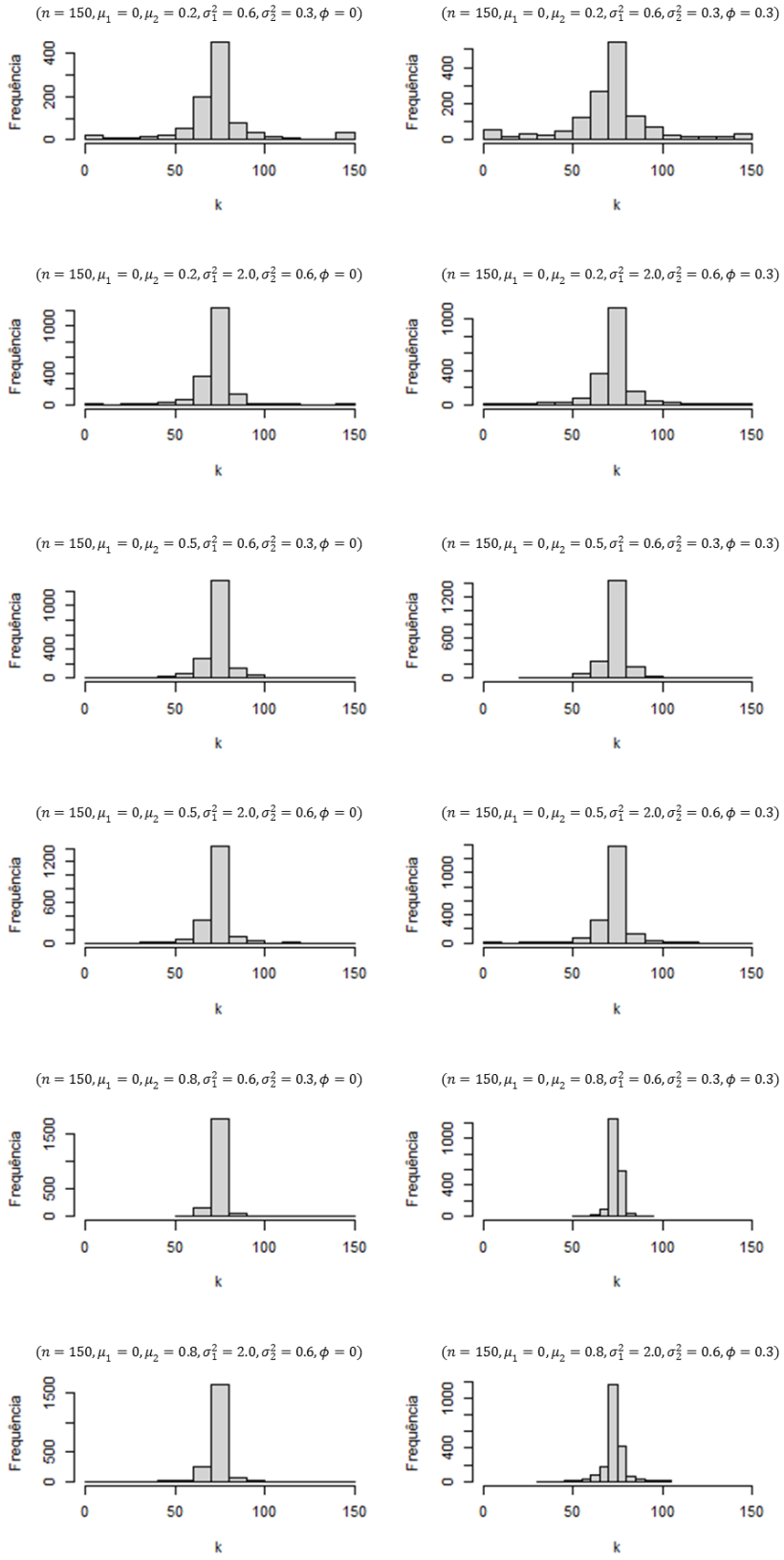


Figura A.8: Histogramas dos *change-points* identificados considerando os erros com distribuição Normal e $n = 150$.

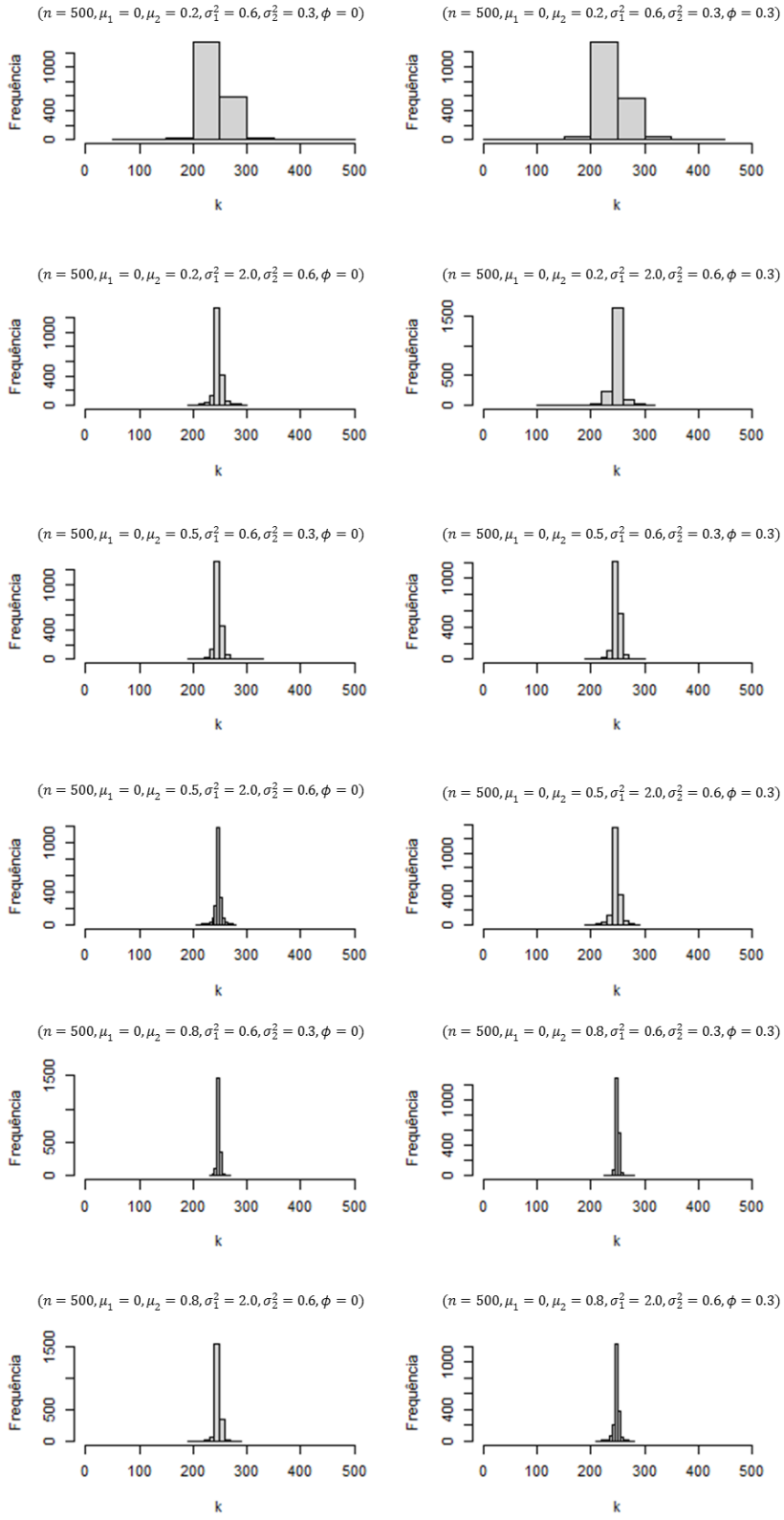


Figura A.9: Histogramas dos *change-points* identificados considerando os erros com distribuição Normal e $n = 500$.

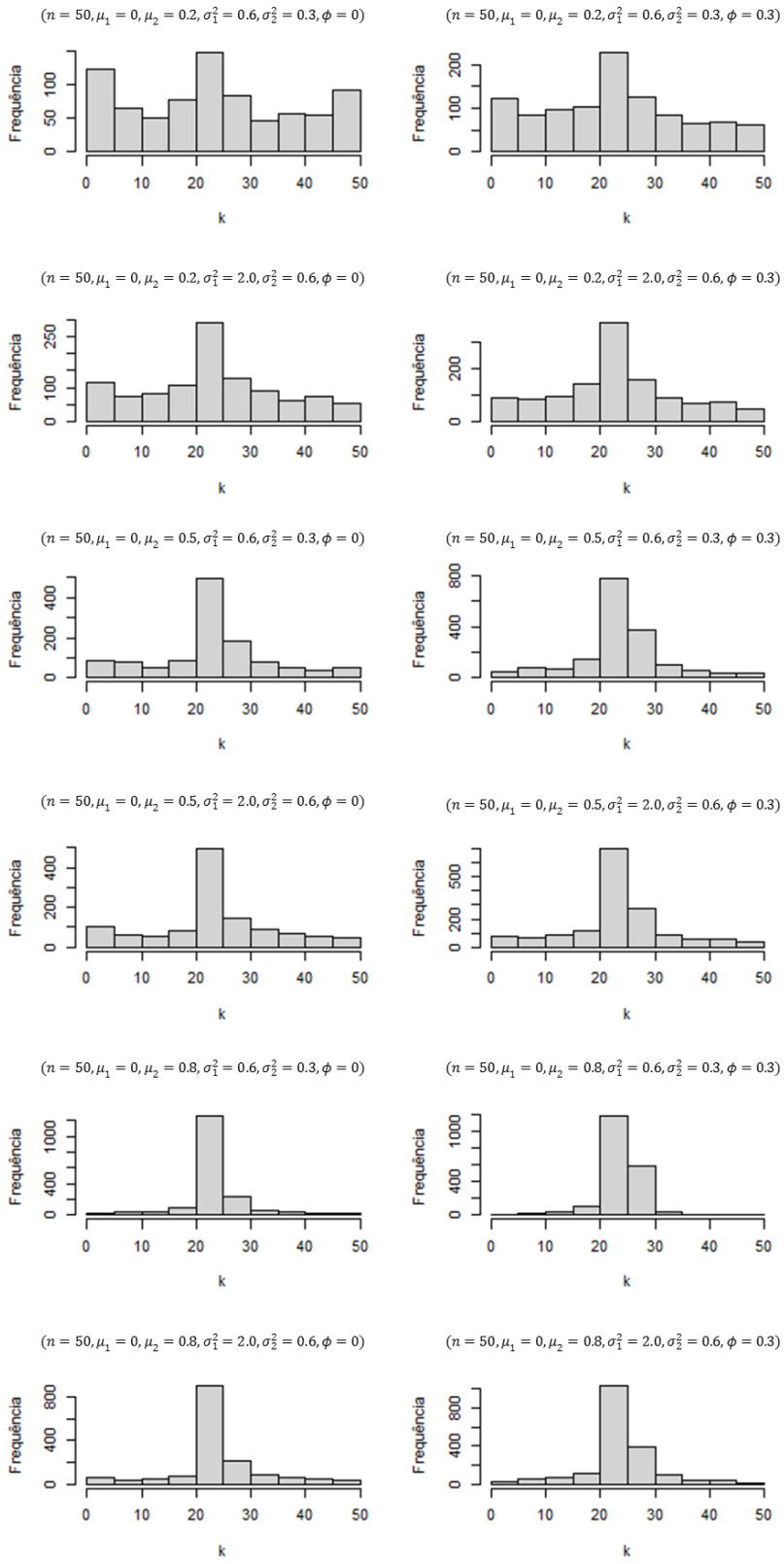


Figura A.10: Histogramas dos *change-points* identificados considerando os erros com distribuição Exponencial e $n = 50$.

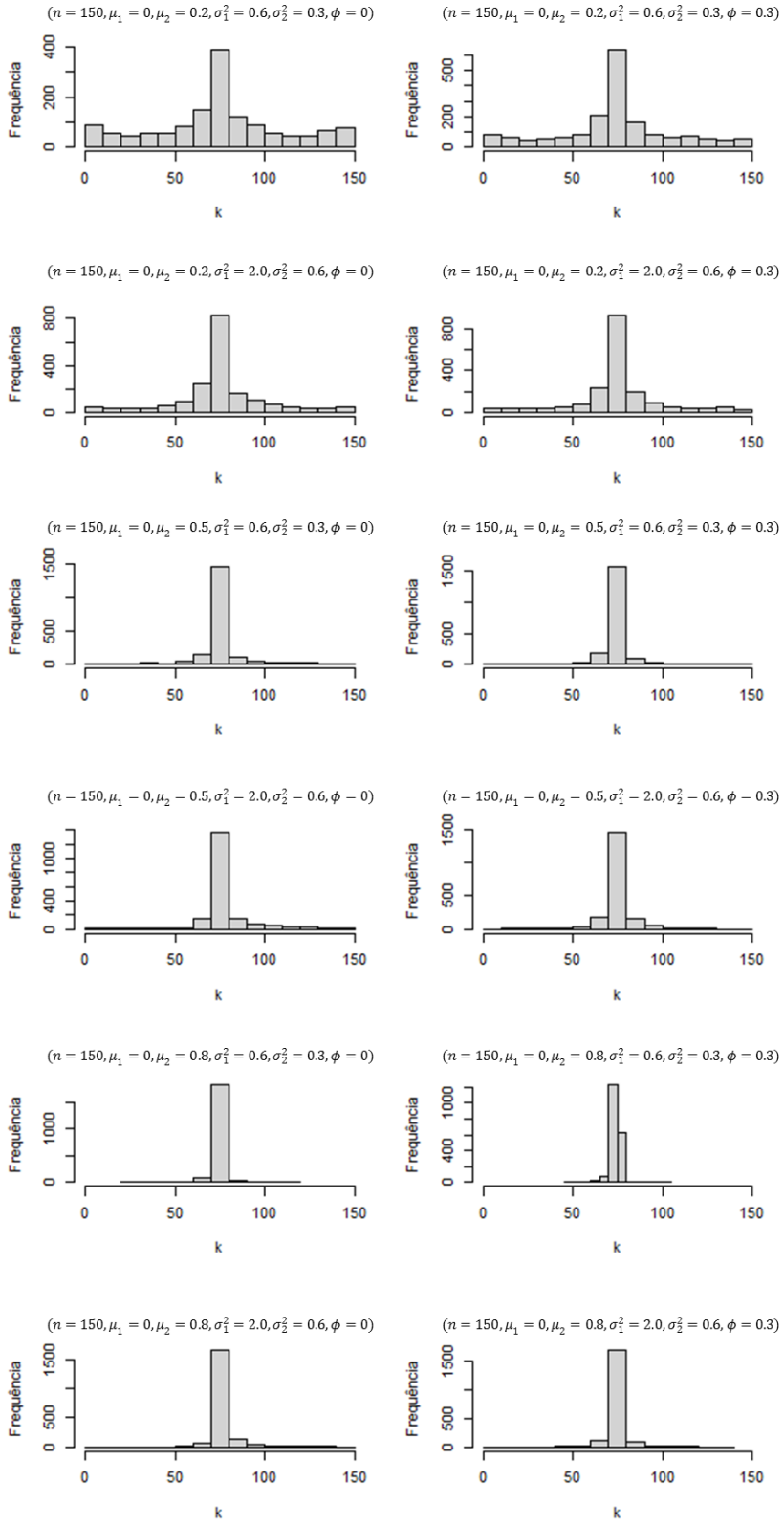


Figura A.11: Histogramas dos *change-points* identificados considerando os erros com distribuição Exponencial e $n = 150$.

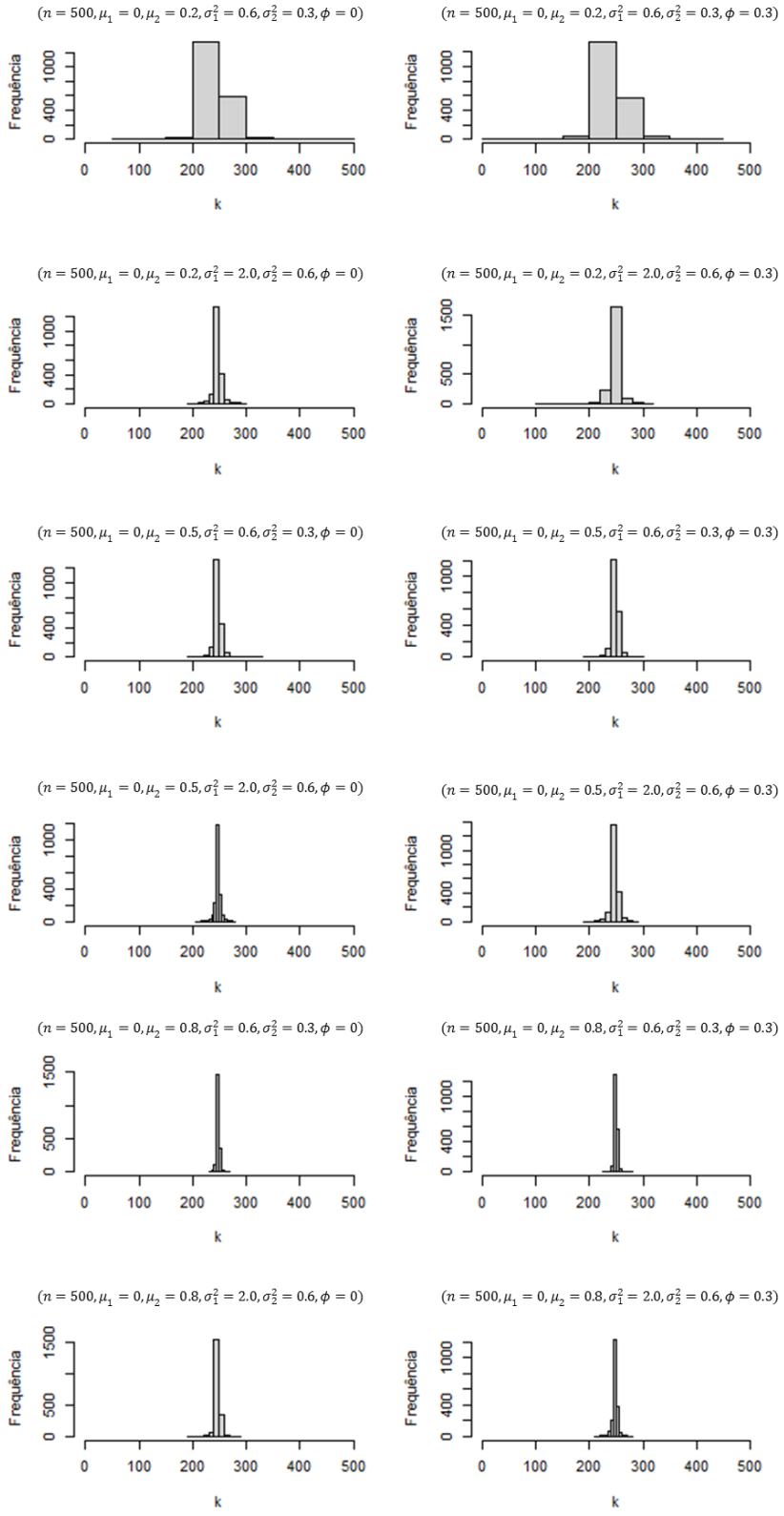


Figura A.12: Histogramas dos *change-points* identificados considerando os erros com distribuição Exponencial e $n = 500$.