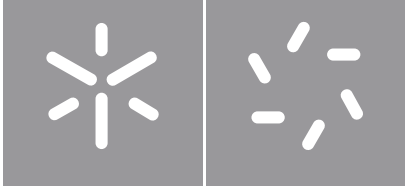
Universidade do Minho
Escola de Ciências

Carla Sofia Carneiro Gomes da Silva

Modelação Estatística na Análise em Processos Ambientais

outubro de 2019

Modelação Estatística na Análise em Processos Ambientais
Carla Sofia Carneiro Gomes da Silva
UMinho | 2019



Universidade do Minho
Escola de Ciências

Carla Sofia Carneiro Gomes da Silva

Modelação Estatística na Análise em Processos Ambientais

Dissertação de Mestrado
Mestrado em Estatística

Trabalho efetuado sob a orientação da
Professora Doutora Arminda Manuela Andrade Pereira Gonçalves
e da
Professora Doutora Susana Margarida Ferreira de Sá Faria

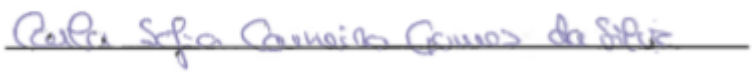
outubro de 2019

DECLARAÇÃO DE INTEGRIDADE

Declaro ter atuado com integridade na elaboração do presente trabalho académico e confirmo que não recorri à prática de plágio nem a qualquer forma de utilização indevida ou falsificação de informações ou resultados em nenhuma das etapas conducente à sua elaboração.

Mais declaro que conheço e que respeitei o Código de Conduta Ética da Universidade do Minho.

Universidade do Minho, 31 de outubro de 2019,



(Carla Sofia Carneiro Gomes da Silva)

Agradecimentos

"Os meus mestres foram todos os homens e mulheres que me deslumbraram em leitura e não só: em exemplos de vida."
(Natália Correia, 1983)

A partilha do conhecimento permite a união de pessoas, a sua cooperação faculta a realização de um trabalho maior. Agradeço a todas as pessoas que permitiram de alguma forma a conceção deste trabalho, momentânea ou continuamente.

Este trabalho não seria possível sem a orientação da Professora Doutora A. Manuela Gonçalves e da Professora Doutora Susana Faria. Toda a disponibilidade, o entusiasmo pela procura constante, a partilha de conhecimento e a tranquilidade determinaram a direção a seguir. E estou grata pela oportunidade de aprender e de vivenciar ensinamentos das "minhas" Professoras, que acompanharam todo este percurso.

Um agradecimento ao Engenheiro Vitorino José pela colaboração prestada na perceção das ferramentas informáticas, disponibilizadas pela Agência Portuguesa do Ambiente (APA). Não podendo esquecer, pelo tempo e pela brevidade na resposta, na crónica da fronteira da RH3, o Doutor Luís Margalho.

Pelo apoio incessante e pela paciência incansável durante todo este percurso, agradeço à minha família e a todos os meus amigos.

Resumo

A degradação do ambiente é atualmente um tema de grande importância, quer pela dificuldade na recuperação e na reabilitação, quer também pelas gravosas consequências sociais e económicas. Em parte, a crise ambiental é o somatório de muitos erros cometidos pelo Homem e que ainda hoje é possível observar. Investigações realizadas no intuito de minimizar ou estimar problemas ambientais têm levado a estudos mais aprofundados de métodos, que possibilitem uma melhor perceção dos dados associados a estes problemas. Neste estudo, no contexto de um problema de monitorização de Qualidade da Água de superfície de uma bacia hidrográfica, propõe-se uma abordagem baseada em modelos espaciais e temporais com o objetivo de analisar e avaliar a evolução de séries temporais de variáveis ambientais. Os dados dizem respeito à bacia hidrográfica do rio Douro localizada no Norte de Portugal. Para o processo de modelação, consideraram-se as séries temporais relativas à variável de qualidade de Oxigénio Dissolvido (OD), medido mensalmente no período de março de 2002 a fevereiro de 2013. Com o objetivo de obter estimativas de valores mensais de precipitação, em área, nas estações de amostragem de qualidade (onde não há medições de precipitação), é desenvolvida uma metodologia com recurso a processos estocásticos espaciais (*Kriging*), a ser aplicada aos dados de precipitação existentes nesta bacia. Os valores estimados vão representar o fator hidrometeorológico nas estações de qualidade, para o processo de modelação do Oxigénio Dissolvido. Para o processo de modelação do Oxigénio Dissolvido foram estabelecidos Modelos de Efeitos Mistos (ou Modelos Lineares Generalizados de Efeitos Mistos), pois mostram versatilidade e flexibilidade para a inclusão de efeitos aleatórios, incorporação de componentes de tendência e de sazonalidade, de covariáveis (como o fator hidrometeorológico e outras variáveis de Qualidade da Água de superfície), bem como da estrutura de correlação temporal própria das séries ambientais. Foi efetuado um estudo comparativo dos diversos modelos estabelecidos, considerando critérios e métricas de qualidade de ajustamento.

Palavras-chave: Bacia Hidrográfica; rio Douro; Qualidade da Água; Geoestatística; Modelos de Efeitos Mistos.

Abstract

Environmental degradation is nowadays a critical issue, both due to the difficulty of restoration and rehabilitation and to the serious social and economic consequences. The environmental crisis is partially the result of many man-made mistakes that still remain visible today. Investigations aimed at curbing or estimating environmental problems have led to more in-depth study of methods to better understand the data associated with these problems.

This study investigates a problem in the context of surface water quality monitoring in a watershed, and we propose an approach based on spatial and temporal models in order to analyze and evaluate the time series evolution of environmental variables. The data refer to the Douro watershed located in northern Portugal and for the modeling process we considered time series relative to the Dissolved Oxygen (DO) quality variable measured monthly from March 2002 to February 2013.

In order to obtain estimates of monthly precipitation values, in area, in the quality sampling stations (where there are no precipitation measurements), we developed a methodology using spatial stochastic processes (Kriging) to be applied to the precipitation data extant in this basin. The estimated values will represent the hydrometeorological factor in the quality sampling stations for the Dissolved Oxygen modeling process.

For the Dissolved Oxygen modeling process we established Mixed Effects Models (or Generalized Linear Mixed Effects Models) as they show versatility and flexibility in including random effects, in incorporating trend and seasonality components, covariates (such as the hydrometeorological factor and other surface water quality variables), as well as the temporal correlation structure typical of the environmental series. A comparative study of the various established models was performed considering criteria and quality adjustment metrics.

Key-words: Watershed; Douro River; Water quality; Geostatistics; Mixed Effects Models.

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Siglas/Acrónimos

AIC — *Akaike Information Criterion* (em português, Critério de Informação de Akaike)

APA — Agência Portuguesa do Ambiente

AR — *Autoregressive* (em português, Autorregressivo)

ARIMA — *Autoregressive Integrated Moving Average* (em português, Autorregressivo Integrado de Médias Móveis)

ARMA — *Autoregressive Moving Average* (em português, Autorregressivo de Médias Móveis)

BIC — *Bayesian Information Criterion* (em português, Critério de Informação Bayesiano)

BLE — *Best Linear Estimator* (em português, Melhor Estimador Linear)

BLUE — *Best Linear Unbiased Estimator* (em português, Melhor Estimador Linear Não Enviesado)

BLUP — *Best Linear Unbiased Prediction* (em português, Melhor Preditor Linear Não Enviesado)

CBO5 — Carência Bioquímica de Oxigénio, a 5 dias

CH — Coeficiente Hidrometeorológico

DSRH — Direção dos Serviços de Recursos Hídricos

EQM — Erro Quadrático Médio

EQMN — Erro Quadrático Médio Normalizado

ET — Estatística de Teste

FAC — Função de Autocorrelação

FACP — Função de Autocorrelação Parcial

INAG — Instituto da Água

LMM — *Linear Mixed Models* (em português, Modelos de Efeitos Mistos)

loglik — *Log-Maximum Likelihood* (em português, Logaritmo da Função de Máxima Verosimilhança)

MA — *Moving Average* (em português, Médias Móveis)

MEP — Média dos Erros de Predição

ML — *Maximum Likelihood* (em português, Função de Máxima Verosimilhança)

MLN — Modelo Linear Normal

MLG — Modelo Linear Generalizado

OD — Oxigénio Dissolvido

PGRH — Plano de Gestão da Região Hidrográfica

REML — *Restricted Maximum Likelihood* (em português, Método de Máxima Verosimilhança Restrito)

RH — Região Hidrográfica

RH1 — Região Hidrográfica do Minho e do Lima

RH2 — Região Hidrográfica do Ave e Leça

RH3 — Região Hidrográfica do Douro

RH4 — Região Hidrográfica do Vouga, do Mondego, de Lis e das Ribeiras do Oeste

RH5 — Região Hidrográfica do Tejo

RH6 — Região Hidrográfica do Sado e de Mira

RH7 — Região Hidrográfica do Guadiana

RH8 — Região Hidrográfica das Ribeiras do Algarve

SAR — *Seasonal Autoregressive* (em português, Autorregressivo Sazonal)

SARIMA — *Seasonal Autoregressive Integrated Moving Average* (em português, Autorregressivo Integrado de Médias Móveis Sazonal)

SARMA — *Seasonal Autoregressive Moving Average* (em português, Autorregressivo

de Médias Móveis Sazonal)

SMA — *Seasonal Moving Average* (em português, Médias Móveis Sazonal)

SNIRH — Sistema Nacional de Informação de Recursos Hídricos

Capítulo 1

Introdução

Nos tempos atuais, a degradação dos recursos hídricos devido à poluição por atividades antropogénicas é inegável (van Dijk *et al.*, 1994). Há falta de métodos, planos de gestão e ferramentas que modelem e prevejam eventos críticos com o objetivo de preservação do Ambiente. Os métodos de modelação são instrumentos importantes para auxiliar a tomada de decisão em diferentes áreas. Tem vindo a aumentar o número de metodologias na área da Estatística para processos de modelação em Ambiente e, em particular, em processos de modelação de variáveis de Qualidade da Água de superfície de uma bacia hidrográfica. Por exemplo, foi desenvolvido um processo de modelação para prever o estado de Qualidade da Água da bacia hidrográfica do rio Douro por um método dinâmico estocástico (Cabecinha *et al.*, 2009; Silva-Santos *et al.*, 2008), de forma holística, e um método de modelação via modelos lineares para a quantidade de Oxigénio Dissolvido (OD), uma medida de Qualidade da Água de superfície da bacia hidrográfica do rio Ave (Gonçalves & Alpuim, 2011). O modelo escolhido para este estudo é o Modelo de Efeitos Mistos, na medida em que incorpora efeitos fixos, associados à população, e efeitos aleatórios, associados às unidades de análise/indivíduos selecionados aleatoriamente na população. Para uma melhor compreensão destes modelos pode-se consultar Verbenke & Molenberghs (2000). No processo de modelação existem diversas formas de integrar as componentes determinísticas e aleatórias. Para avaliar e comparar estes processos de modelação, recorre-se a critérios e a métricas de avaliação de qualidade de ajustamento, permitindo a avaliação da eficácia no ajustamento do modelo aos dados reais.

1.1 Dados e Motivação

No âmbito da água, a União Europeia estabeleceu a Diretiva Quadro da Água (Diretiva 2000/60/CE do Parlamento Europeu, de 23 de outubro de 2000) para que exista uma uniformização de ação comunitária para a proteção das águas de superfície, de transição,

costeiras e subterrâneas. Em Portugal, o Decreto-Lei n.º 45/1994, a Lei da Água n.º 58/2005 e o Decreto-Lei n.º 77/2006 são medidas que estão na génese do planeamento e da gestão dos recursos hídricos. A partir de 2012, a Agência Portuguesa do Ambiente (APA) ficou responsável pela elaboração, revisão e atualização dos planos de gestão das regiões hidrográficas (RH). Um exemplo é o Plano de Gestão Hidrográfica do Norte de Portugal (PGRH), que engloba as regiões do Minho e do Lima (RH1), do Cávado, do Ave e do Leça (RH2) e do Douro (RH3).

A Agência Portuguesa do Ambiente (APA) está envolvida na política de monitorização, nomeadamente na implementação de redes de medição, entre outras. A informação recolhida neste domínio é compilada no repositório do Sistema Nacional de Informação de Recursos Hídricos (SNIRH).

Na presente dissertação pretende-se analisar os dados referentes a variáveis relacionadas com a Qualidade da Água de superfície. As observações recolhidas são referentes à bacia hidrográfica do rio Douro, localizada no Norte de Portugal. A sua monitorização é prioritária, na medida em que existem zonas industrializadas e zonas sujeitas a grandes períodos de seca, nesta bacia hidrográfica.

A variável Oxigénio Dissolvido (OD), em mg/l, é uma das variáveis indicadores mais importantes na determinação do grau de poluição existente na água de superfície de uma bacia hidrográfica (Costa & Gonçalves, 2011; Gonçalves & Costa, 2013).

Assim, neste estudo, o Oxigénio Dissolvido foi a variável de Qualidade da Água de superfície escolhida para estabelecer o processo de modelação com o objetivo de avaliação e monitorização da Qualidade da Água. Os dados foram obtidos a partir do Sistema Nacional de Informação de Recursos Hídricos (SNIRH), recolhidos mensalmente, no período de março de 2002 até fevereiro de 2013, em estações de monitorização da bacia. O comportamento das variáveis de Qualidade da Água será analisado através de metodologias espaciais e temporais.

Na aplicação das metodologias aos dados utiliza-se o ambiente `R` (R *Core Team*, 2017), recorrendo-se a algumas *packages* implementadas, como *nlme* e *geoR*, e são criadas funções para otimização de procedimentos.

1.2 Objetivos e Organização do Trabalho

Na realização da presente dissertação, o principal objetivo é a identificação dos métodos mais adequados para a modelação de variáveis ambientais em análise, procedendo-se, para isso, ao seu ajustamento e comparação relativamente à sua capacidade explicativa e preditiva.

A dissertação é composta por seis Capítulos e dois Anexos. No Capítulo 1 é feita

uma breve introdução, na qual se caracteriza a problemática abordada na dissertação, com a exposição das metas a atingir e a estrutura da dissertação. O Capítulo 2 visa a investigação sobre a temática da Geoestatística (análise estatística espacial), desde a sua revisão bibliográfica até alguns conceitos teóricos e respetivos métodos, que serão, posteriormente, aplicados a uma análise dos dados de precipitação.

No Capítulo 3 apresentam-se breves noções teóricas relacionadas com a análise de séries temporais: os conceitos fundamentais e os métodos que são aplicados aos dados, neste estudo.

Inicia-se o Capítulo 4 com uma revisão bibliográfica sobre os Modelos de Efeitos Mistos e aprofundam-se os conceitos essenciais e as metodologias adotadas.

No Capítulo 5 é realizado um breve resumo dos dados em estudo, efetua-se uma análise exploratória dos dados, seguindo-se para a aplicação das metodologias e do estudo comparativo da capacidade de modelação dos diferentes modelos.

O Capítulo 6 visa a apresentação das conclusões gerais da pesquisa e da análise efetuada no âmbito da Qualidade da Água e são indicados futuros desenvolvimentos.

Na Bibliografia são indicadas as referências bibliográficas, que suportam a investigação da presente dissertação.

Capítulo 2

Geoestatística

2.1 Interpolação Espacial

Os Métodos de Interpolação Espacial correspondem a procedimentos de estimação do valor de um atributo em locais onde não estão disponíveis observações do mesmo, a partir de pontos em que tenham sido registadas observações do atributo em causa. Assim, a interpolação é uma técnica cujo objetivo é a estimação de valores desconhecidos de uma função, a partir de valores conhecidos da mesma função. Quando a informação disponível, proveniente de uma amostra recolhida, não cobre todo o domínio espacial, a interpolação é uma opção para completar os valores em falta. Existem vários métodos de interpolação, como os Métodos Determinísticos e os Métodos Estocásticos.

2.1.1 Métodos Determinísticos

Os Métodos Determinísticos, que continuam a ter uma grande importância e aplicação em áreas de fénomenos espaciais, vão ser apresentados de um modo resumido.

Os Polígonos de Thiessen visam a subdivisão do domínio espacial em áreas de influência (polígonos de influência) das observações disponíveis (Thiessen, 1911). Assim, qualquer localização no espaço tem o valor estimado igual ao valor observado mais próximo, que é o do centro do polígono em que a localização está contida. Os Polígonos de Voronoy recorrem a métodos que também consistem na divisão geométrica do espaço em áreas de influência (poliedros convexos) e utilizam esta decomposição para o cálculo do peso de cada valor observado na interpolação.

O Método das Médias Móveis estima os valores numa determinada localização, pela determinação da média aritmética dos valores observados nas localizações mais próximas.

No Método da Média Aritmética, o valor estimado num local é calculado pela média aritmética de todas as observações.

A Interpolação Quadrática determina o valor estimado num local a partir da soma ponderada dos valores observados, em que a contribuição de cada valor é inversamente proporcional ao quadrado da distância ao ponto a estimar.

Na Interpolação Multiquadrática, o valor é estimado com base na ponderação das distâncias desse local aos locais de observação (mais uma constante), em que os pesos são tais que a superfície de interpolação obtida passa exatamente pelos valores observados.

O Método de Ajustamento de uma Superfície consiste em ajustar os valores observados a uma superfície polinomial (*splines*).

As dificuldades principais na aplicação dos Métodos Determinísticos traduzem-se na quantificação da estrutura espacial da grandeza em estudo e na avaliação da incerteza associada à caracterização do fenómeno espacial.

2.1.2 Métodos Estocásticos

Os Métodos Estocásticos pressupõem que os fenómenos se distribuam no espaço de uma forma aleatória, com uma determinada estrutura de correlação e, assim, com um grau de incerteza associado aos fenómenos, resultante da falta de informação disponível. Estes consideram os dados como realizações de um determinado processo aleatório e consistem na modelação da estrutura de variação do processo e utilizam-na para construir um estimador para os valores não observados.

No contexto de um processo de modelação espacial, dado um conjunto de dados provenientes das amostras experimentais, inicia-se pela conceção de um processo aleatório que caracteriza o conjunto de dados, sendo considerada a seleção de um número restrito de parâmetros que, sob determinadas hipóteses, permitem a inferência espacial.

Em 1951, o engenheiro de minas sul-africano, D. G. Krige, desenvolveu um método para estimar o teor em minério de um subsolo a partir de amostras extraídas. Com base nas ideias de D. G. Krige, Matheron (1963) estabelece o termo Geoestatística em que "a Geoestatística é a aplicação do formalismo das funções aleatórias ao reconhecimento e a estimação de fenómenos naturais". Ao mesmo tempo que são desenvolvidas as técnicas geoestatísticas na área da Engenharia Mineira com G. Matheron, as mesmas ideias são desenvolvidas na área da Meteorologia com L. S. Gandin, na União Soviética, sob o nome de "análise objetiva" e "interpolação ótima" (Lefèvre, 1997).

Mercer & Hall (1911) consideram algumas características da Geoestatística moderna a dependência espacial, a correlação e o efeito de pepita (a variabilidade à pequena escala). A representação da correlação espacial, reconhecida como variograma, é desenvolvida por Kolmogorov (1941), assim como o método de interpolação (Ripley, 1981). Matérn (1960) desenvolve algumas funções que permitem descrever a covariância espacial. Jowett (1955) também estuda e apresenta algumas funções, posteriormente denominadas como

variogramas, que expressam a dependência espacial entre amostras vizinhas.

A Geoestatística permite, também, fornecer estimativas de erros de estimação, a partir de técnicas numéricas que caracterizam atributos espaciais (Olea, 2012). A Geoestatística oferece uma forma de descrever a continuidade espacial dos fenómenos naturais, adaptando técnicas clássicas de regressão, aproveitando a continuidade espacial (Isaacs & Srivastava, 1989). Alguns exemplos são aplicações relacionadas com a Meteorologia (Cressie & Huang, 1999; Kyriakidis *et al.*, 2001) ou Hidrologia (Goovaerts, 2000).

Constata-se um grande desenvolvimento na associação da dimensão temporal à dimensão espacial, em diversos autores como, por exemplo, Cressie (1993), Cressie & Wikle (2015), Cressie *et al.* (2019), Diggle & Giorgi (2019) e Goovaerts (1997).

Bárdossy & Pegram (2009) e Gräler (2014) defendem que a covariância tem um papel preponderante na evolução da Geoestatística, através de campos aleatórios espaço-temporais, o que permite uma maior flexibilidade na modelação dos dados.

As implementações destes métodos com ferramentas computacionais são bastante recentes. Exemplos disso são a *package gstat* (Pebesma, 2004) e a *package spacetime* (Pebesma, 2012), através da extensão para a Geoestatística espaço-temporal proposta por Gräler (Pebesma & Heuvelink, 2016), em ambiente `R`. Cressie & Wikle (2015) explicam e abordam explicitamente estatísticas de dados espaço-temporais, com exemplos práticos.

2.2 Processos Aleatórios

Considerando os dados como uma série espacial associada a n localizações espaciais $\{s_1, s_2, ..., s_n\}$ e os valores de uma variável contínua $\{z(s_1), z(s_2), ..., z(s_n)\}$, observados nestas localizações. Cada valor observado $z(s_i)$, $i = 1, \ldots, n$, é considerado como uma realização particular de uma determinada variável aleatória $Z(s)$, em que s varia numa região do espaço real de dimensão finita positiva, $D \subseteq \mathbb{R}^r$, e, usualmente, $r = 2, 3$ (espaço real bidimensional ou tridimensional). Este conjunto de variáveis aleatórias (geralmente correlacionadas) é denominado por processo aleatório, campo aleatório ou função aleatória, sendo definido por

$$\{Z(s) : s \in D\} \tag{2.1}$$

e tem de satisfazer as condições de simetria e de consistência (Yaglom, 1962).

Um processo aleatório $\{Z(s) : s \in D\}$ é usualmente caracterizado através da função distribuição cumulativa

$$F_{Z(s_1),\ldots,Z(s_n)}(s_1, \ldots, s_n) = P(Z(s_1) \leq z_1, \ldots, Z(s_n) \leq z_n). \tag{2.2}$$

Para cada $s \in D$, $Z(s)$ é uma variável, assim, define-se como função valor médio ou

momento de primeira ordem

$$\forall s \in D, \qquad E[Z(s)] = \mu_Z(s), \tag{2.3}$$

quando a esperança existe.

Também se pode especificar a covariância do processo aleatório, se existir, e é tal que

$$\begin{aligned} &Cov(Z(s_j), Z(s_k)) = E[(Z(s_j) - \mu_Z(s_j)(Z(s_k) - \mu_Z(s_k)] \\ &s_j, s_k \in D, \quad j, k = 1, ..., n, \end{aligned} \tag{2.4}$$

em particular, a variância é tal que $Cov(Z(s_j), Z(s_j)) = E[(Z(s_j) - \mu_Z(s_j))^2] = Var[Z(s_j)]$, $\forall s_j \in D$, com $j = 1, \ldots, n$.

Na maioria dos casos práticos não se conhece a lei de probabilidade que define o processo aleatório. Deve-se inferir a distribuição ou alguns dos seus momentos, o que requer várias realizações do processo $Z(s)$. Na teoria dos processos aleatórios, a hipótese (restrição) usual é a da estacionaridade (ligada à noção intuitiva de homogeneidade espacial).

Um processo aleatório espacial $\{Z(s) : s \in D\}$ diz-se processo estacionário de primeira ordem ou intrinsecamente estacionário em D se para qualquer conjunto de localizações $s_1, ..., s_n \in D$, a distribuição conjunta é invariante com respeito a qualquer translação nas localizações. Isto é, para quaisquer $n \geq 1$, $h \in \mathbb{R}^r$ e $s_1 + h, \ldots, s_n + h \in D$ as distribuições de $(Z(s_1 + h), ..., Z(s_n + h))$ e $(Z(s_1), ..., Z(s_n))$ são idênticas, ou seja,

$$\forall h \in \mathbb{R}^r, \qquad F_{s_1+h,...,s_n+h}(z_1, ..., z_n) = F_{s_1,...,s_n}(z_1, ..., z_n). \tag{2.5}$$

Na prática, no entanto, a lei de distribuição não é conhecida, pois os dados são insuficientes para a inferir. Assim, em algumas situações é desejável disponibilizar-se de um conceito de estacionaridade menos restritivo, envolvendo apenas os dois primeiros momentos que são suficientes para aproximar corretamente a solução do problema.

Um processo aleatório $\{Z(s) : s \in D\}$ diz-se estacionário de segunda em $D \subseteq \mathbb{R}^r$, se

$$\forall s \in D, \qquad E[Z(s)] = \mu_Z(s) = \mu_Z \tag{2.6}$$

e, para cada duas variáveis aleatórias $Z(u)$ e $Z(v)$, a função covariância existe e apenas depende da diferença entre u e v, i.e.,

$$\forall u, v \in D, \qquad Cov(Z(u), Z(v)) = C_Z(u - v), \tag{2.7}$$

em que C_Z designa-se por covariograma ou função de covariância estacionária do processo aleatório $Z(s)$.

A estacionaridade da covariância implica que a variância $Var[Z(s)]$ existe e não de-

pende de s (implica a estacionaridade da variância). Isto é, $Var[Z(s)] = C_Z(0)$, $\forall s \in D$.

Em particular, se o processo espacial considerado é tal que $C_Z(0) > 0$ e é estacionário de segunda ordem, a função $C_Z(.)$ é designada por correlograma (ou função de correlação estacionária) e denominada por $\rho_Z(.)$, tal que

$$\forall s, u \in D, \quad \rho(s-u) = \frac{C_Z(s-u)}{C_Z(0)} \in [-1, 1]. \tag{2.8}$$

O correlograma, bem como o covariograma, modela a estrutura de dependência espacial do processo aleatório (estacionário de segunda ordem).

Um processo aleatório $\{Z(s) : s \in D\}$ diz-se intrinsecamente estacionário (ou de estacionaridade intrínseca) em $D \subseteq \mathbb{R}^r$ se o valor médio do processo existe e é constante em D, isto é,

$$\forall s \in D, \qquad E[Z(s)] = \mu_Z(s) = \mu_Z, \tag{2.9}$$

em que a variância $Var[Z(u) - Z(v)]$ existe para todo $u, v \in D$ e depende apenas da diferença $u - v$, isto é,

$$\forall u, v \in D, \qquad Var[Z(u) - Z(v)] = E[(Z(u) - Z(v))^2] = 2\gamma_Z(u - v) \tag{2.10}$$

em que se designa $2\gamma_Z$ a função variograma e γ_Z é denominada função semi-variograma do processo $Z(.)$ (salvaguarda-se a existência de autores que definem outra nomenclatura).

A definição de variograma como a variância dos acréscimos (incrementos) espaciais de um processo aleatório faz com que se verifiquem algumas propriedades. Se $\gamma_Z(.)$ é o semivariograma de um processo aleatório intrinsecamente estacionário $Z(.)$, então

$$\forall n \geq 1, \forall \lambda_1, \ldots, \lambda_n \in \mathbb{R}, \forall s_1, ..., s_n \in \mathbb{R}^r, \qquad \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j C_z(s_i - s_j) \geq 0. \tag{2.11}$$

Além disso, se o covariograma C_Z resulta dum processo aleatório estacionário de segunda ordem $Z(s)$, tem-se que

$$\forall s \in D, \qquad C_Z(0) = Var(Z(s)) \geq 0, \tag{2.12}$$

$$\forall u, v \in D, \qquad C_Z(u - v) = C_Z(v - u) \tag{2.13}$$

e, pela desigualdade de Cauchy-Schwarz, tem-se que

$$\forall u, v \in D, \qquad |C_Z(u - v)| \leq C_Z(0). \tag{2.14}$$

No caso do processo aleatório com apenas estacionaridade intrínseca, o semivariograma existe mas o covariograma pode não existir.

Se o semivariograma $\gamma_Z(.)$ de um processo intrinsecamente estacionário $Z(s)$, então apresenta algumas propriedades, tais como

$$\gamma_Z(0) = 0, \tag{2.15}$$

$$\forall u, v \in D, \qquad \gamma_Z(u - v) = \gamma_Z(v - u), \tag{2.16}$$

$$\forall u, v \in D, u \neq v, \qquad \gamma_Z(u - v) > 0 \tag{2.17}$$

e

$$\forall u, v \in D, \qquad \lim_{||u-v||\to\infty} \gamma_Z(u - v) = c_0, \tag{2.18}$$

designado por efeito de pepita.

A razão do crescimento de um semivariograma de um processo aleatório $Z(s)$ é dada por

$$\lim_{||u-v||\to\infty} \frac{\gamma_Z(u - v)}{||u - v||^2} \tag{2.19}$$

e pode ser indicador se o processo é intrinsecamente estacionário ou não. Caso o semivariograma $\gamma_Z(.)$ tiver um crescimento mais lento que $||u - v||^2$, então o limite anterior tende para zero e o processo $Z(.)$ é intrinsecamente estacionário. Se o crescimento de $\gamma_Z(.)$ for mais rápido que $||u - v||^2$, então a hipótese intrínseca não é válida.

Se a estacionaridade do processo aleatório é de segunda ordem, então o variograma e o covariograma existem e são estruturalmente equivalentes, cuja relação estrutural é dada por

$$C_Z(u - v) = C_Z(0) - \gamma_Z(u - v) \Leftrightarrow \gamma_Z(u - v) = C_Z(0) - C_Z(u - v) \tag{2.20}$$

e se for verificado que $\lim_{||u-v||\to\infty} C_Z(u - v) = 0$, então

$$\lim_{||u-v||\to\infty} \gamma_Z(u - v)) = C_Z(0) = Var[Z(.)], \tag{2.21}$$

em que $C_Z(0)$ é designado por patamar do semivariograma. O patamar parcial é determinado pela diferença entre o patamar e o efeito de pepita, $C_Z(0) - C_0$.

Um processo aleatório $\{Z(s), s \in D\}$ pode ser descrito como uma combinação de processos aleatórios não correlacionados $\{Z_i(s), s \in D\}$, com $i = 1, \ldots, k$, ou seja, formalmente

$$\forall \lambda_1, \ldots, \lambda_k \in \mathbb{R}_0^+, \qquad Z(s) = \sum_{i=1}^{k} \lambda_i Z_i(s). \tag{2.22}$$

Se $Z_i(s)$ é um processo estacionário de segunda ordem com covariograma $C_{Z_i}(s)$, para

cada $i = 1, \ldots, k$, então $Z(s)$ também é estacionário de segunda ordem com covariograma dado por $C_Z(u - v) = \sum_{i=1}^{k} \lambda_i^2 C_{Z_i}(u - v)$, quaisquer que sejam $u, v \in D$.

Se $Z_i(s)$ é um processo intrinsecamente estacionário, com semivariograma $\gamma_{Z_i}(.)$, para cada $i = 1, \ldots, k$, então $Z(s)$ é também intrinsecamente estacionário com semivariograma dado por $\gamma_Z(u - v) = \sum_{i=1}^{k} \lambda_i^2 \gamma_{Z_i}(u - v)$, quaisquer que sejam $u, v \in D$.

2.3 Continuidade Espacial

Habitualmente, o estudo de uma característica revela que se a distância entre dois valores for reduzida (pontos próximos), então são mais semelhantes do que aqueles com distâncias maiores (pontos afastados). A continuidade espacial, em maior ou menor grau, permite a descrição do quanto os valores se dispersam espacialmente, e de que forma variam com as diferentes direções do espaço (anisotropia). Os conceitos que serão abordados visam descrever e quantificar a continuidade espacial.

2.3.1 Variograma, Covariograma e Correlograma

Na Geoestatística, as funções para a modelação da dependência espacial e/ou temporal recorrentemente utilizadas são o variograma, o covariograma e o correlograma.

O variograma pode ser classificado de acordo com a sua natureza: o variograma empírico é resultante do conjunto de observações da amostra em estudo, o variograma teórico é o modelo de variograma de referência e o variograma verdadeiro é o variograma real e é desconhecido.

De forma simplista, o variograma quantifica a dispersão natural das variáveis e a variabilidade espacial entre pares de valores separados por uma distância previamente estabelecida $||d||$, em que $d = s_i - s_j$. Existem vários métodos para o cálculo do semivariograma. Por exemplo, pelo método dos momentos, o semivariograma é calculado pela média aritmética do quadrado das diferenças de todos os pares de pontos que estão separados de um vetor d (Matheron, 1963), tal que

$$\hat{\gamma}_Z(d) = \frac{1}{2|N(d)|} \sum_{(i,j) \in N(d)} \big(Z(s_i) - Z(s_j)\big)^2, \tag{2.23}$$

em que $N(d) = \big\{(i,j) : s_i - s_j = h, i, j \in 1, \ldots, n, \big\}$ e $|N(d)| = \#N(d)$ é o número de pares de pontos $||d||$ distanciados e alinhados segundo a direção do vetor d. A representação dos pares de valores $\big(||d||, \hat{\gamma}_z(d)\big)$ num sistema de eixos representa o semivariograma experimental. A aplicação destes métodos tem como pressuposto a malha amostral ser regular. No caso contrário, deve-se proceder a uma regularização angular e por classes de

distâncias. Um processo aleatório designa-se por isotrópico se o respetivo semivariograma ou o covariograma depender do vetor d apenas na sua norma, não podendo depender da direção angular desse mesmo vetor. Um processo espacial intrinsecamente estacionário diz-se isotrópico quando o seu variograma $\gamma_Z(d)$ é função apenas de $||d||$, ou seja

$$\forall d, \quad \gamma_Z(d) = \gamma_Z(||d||). \tag{2.24}$$

O processo que não verifique a condição supracitada é denominado como anisotrópico (depende de $||d||$ e da direção de d, dir), ou seja,

$$\forall d, \quad \gamma_Z(d) = \gamma_Z(||d||, dir). \tag{2.25}$$

A anisotropia pode ser entendida como a variabilidade espacial dependente das direções do espaço. A modelação de fenómenos isotrópicos tem como objetivo reduzir as estruturas de continuidade das diferentes direções a um só modelo. Esta modelação é conseguida geralmente através de um conjunto de transformadas geométricas do sistema de coordenadas, de modo a que os diferentes semivariogramas nas diferentes direções sejam equivalentes a um mesmo modelo ou representando separadamente cada um das variabilidades direcionais consideradas.

Os dois modelos mais comuns de anisotropia são a anisotropia geométrica e a anisotropia zonal. Diz-se que um processo espacial intrinsecamente estacionário $\{Z(s) : s \in D\}$ exibe anisotropia geométrica se tal anisotropia pode ser reduzida a uma isotropia através de uma transformação linear das coordenadas, ou seja, se existir uma matriz invertível $A_{b\times b}$ tal que $Z(As)$ é isotrópico. A anisotropia zonal corresponde ao caso em que os semivariogramas ajustados nas diferentes direções apresentam diferentes características de variabilidade (diferentes patamares), podendo ter valores de amplitude também diferentes (amplitude é o valor de $||d||$ para o qual o semivariograma se estabiliza, isto é, $\gamma_Z(||d||) = C_1$, onde C_1 é uma constante). Mais pormenores sobre anisotropia podem ser consultados em Soares (2000). O objetivo da correção de anisotropia é obter um único semivariograma isotrópico que possa modelar a variabilidade espacial do fenómeno em estudo. A isotropia é muito importante porque é fácil de interpretar e esta característica ajuda na compreensão do processo e na interpretação do modelo e, além disso, reduz a carga dos cálculos computacionais.

2.3.2 Modelos Teóricos de Semivariogramas

Os modelos teóricos de semivariogramas definidos positivos (modelos de transição) utilizam um número restrito de funções ou combinações de funções, definidas positivas, de

forma a satisfazerem as condições de positividade, para interpolar os valores experimentais dos variogramas.

A utilização de um número restrito de funções, definidas positivas, para interpolar os valores experimentais dos variogramas é uma das possíveis formas de satisfazer as condições de positividade.

Estes modelos são independentes da direção, isótropos, simples e são funções do escalar $||d||$, sendo classificados em modelos de transição e em modelos não estacionários.

Se $C_{Z_k}(.)$, $\forall k \in \mathbb{N}$, é um covariograma válido em $\mathbb{R}^r$ e $\lim_{k\to\infty} C_{Z_k}(d) = C_Z(d)$, $\forall d \in \mathbb{R}^r$, então $C_Z(.)$ é um covariograma válido em $\mathbb{R}^r$.

A construção de semivariogramas/covariogramas válidos e o estabelecimento de condições necessárias e suficientes para que a função seja um semivariograma/covariograma válido é um tema extensamente estudado, embora neste trabalho se limite a expor os modelos de semivariogramas mais clássicos.

Num processo aleatório intrinsecamente estacionário, $\{Z(s), s \in D\}$, com semivariograma $\gamma_Z(.)$, se $\lim_{||d||\to\infty} \gamma_Z(d) = C_Z(0) = \sigma_Z^2 < \infty$, então o processo aleatório $Z(.)$ designa-se por fenómeno de transição e a $\gamma_Z(.)$ denomina-se por modelo de transição.

$\sigma_Z^2+\tau_Z^2$ é, nesta definição, o valor do patamar do semivariograma $\gamma_Z(.)$ e é caracterizado pela altura máxima, atingida pela curva do semivariograma. O menor valor de $||\phi||$ para o qual $\gamma_Z(\phi(1+\epsilon)) = C_Z(0) = \sigma_Z^2$ é denominado por amplitude do semivariograma na direção $\dfrac{\phi}{||\phi||}$ $(\phi \in \mathbb{R})$.

Quando o semivariograma do processo aleatório $Z(.)$ possui um valor do patamar (o semivariograma é limitado), este é o valor da variância do processo $Z(.)$ e $Z(.)$ é também um processo estacionário de segunda ordem. Assim, um fenómeno de transição corresponde a um processo estacionário de segunda ordem. Neste caso, o covariograma do processo aleatório $Z(.)$ possui também um valor do patamar que é igual a zero e no caso de existência de amplitude $||\phi||$ é obviamente a mesma para o semivariograma e semicovariograma.

Vão ser apresentados os principais modelos de transição que abragem a generalidade das situações de dispersão de fenómenos espaciais nas Ciências do Ambiente, Figura 2.1.

Modelo de Efeito de Pepita

No caso de variáveis contínuas é expectável que o variograma passe na origem, contudo na maioria dos casos tal não se verifica. Esta descontinuidade representa as variações locais ou de pequena escala, como erros de amostragem ou de análise (a variância dos erros de análise contribui para este valor) e a possível existência de micro regionalizações desenvolvendo-se a uma escala não detetável pela escala de amostragem adotada.

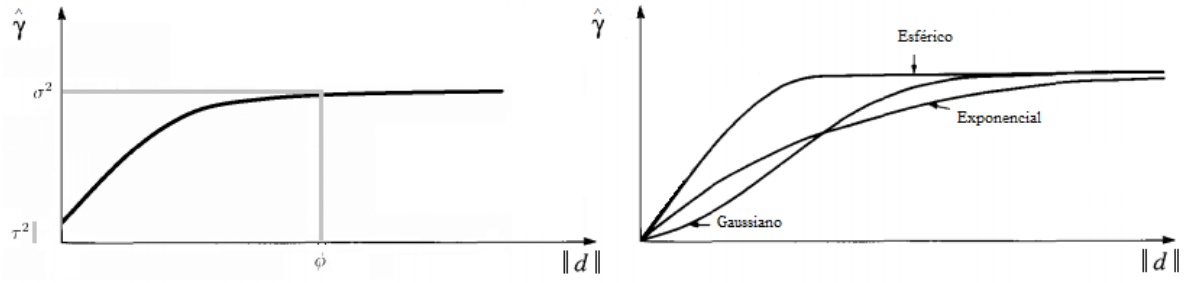


Figura 2.1: Esquerda: Representação esquemática do variograma teórico. Direita: Representação esquemática dos variogramas teóricos mais usuais: Exponencial, Esférico e Gaussiano (Matheron, 1963), adaptado.

A uma distância mínima, d_{min}, superior a zero, o semivariograma nessa distância, $\gamma_Z(d_{min})$, é relativamente elevado (grande variabilidade à pequena escala), provocando uma descontinuidade na origem. Este valor pode ser modelado por uma constante C_0 e esta discrepância designa-se por efeito de pepita. Visualmente é identificado por um salto vertical de zero à origem do variograma e a constante é determinada pela interseção da reta, resultante dos primeiros pontos do variograma, com o eixo das ordenadas.

Considerando o processo espacial $\{Z(s) : s \in D\}$, com $Z(s)$ e $Z(v)$ variáveis aleatórias não correlacionadas $\forall s \neq v$, $E[Z(s)] = 0$ e $Var(Z(s)) = \sigma_Z^2 = C_0$, $\forall s \in D$. Nestas condições, o processo $Z(s)$ é um processo estacionário de segunda ordem, designado por processo espacial de ruído branco. O semivariograma deste processo é dado por

$$\gamma_Z(d) = \begin{cases} 0 & , d = 0 \\ C_0 & , d \neq 0 \end{cases}. \tag{2.26}$$

Modelo Esférico

O Modelo Esférico, acompanhado com o efeito de pepita, também é bastante utilizado e o seu variograma é definido por

$$\gamma(d) = \begin{cases} 0 & , d \leq 0 \\ C_0 + C_1\left(\frac{3}{2}\frac{||d||}{C_2} - \frac{1}{2}\frac{||d||^3}{C_2}\right) & , 0 < d < C_2 \\ C_0 + C_1 & , d \geq C_2 \end{cases}, \tag{2.27}$$

em que C_0 é o efeito de pepita que origina uma discontinuidade na origem do semivariograma, C_1 é a amplitude a partir da qual os valores do processo espacial deixam de estar correlacionados e o patamar C_2 é o valor $\gamma_Z(d)$ para o qual o variograma se estabiliza (o limite superior para o qual tendem os valores do semivariograma com o aumento dos valores de d) que é normalmente coincidente com a variância de $Z(.)$.

Modelo Exponencial

No Modelo Exponencial, acompanhado com o efeito de pepita, o semivariograma exponencial é caracterizado por atingir assintoticamente o patamar, C_0+C_1, quando a diferença entre duas localizações tende para infinito. Formalmente tem a seguinte formulação

$$\gamma_Z(d) = \begin{cases} 0 & , d = 0, \\ C_0 + C_1\left(1 - e^{-\frac{||d||}{C_2}}\right) & , d \neq 0, \end{cases} \tag{2.28}$$

em que C_0, C_1 e C_2 representam o efeito de pepita, a amplitude a partir da qual os valores do processo deixam de estar correlacionados e o patamar é o valor $\gamma_Z(d)$ para o qual o variograma se estabiliza, respetivamente.

É bastante difícil definir a distinção do patamar assintótico do patamar efetivo, devido ao comportamento do semivariograma e à flutuação experimental. Considera-se que a amplitude ϕ^* é dada por $C_2^* = 3C_2$ e é o valor em que o modelo atinge 95 % do patamar: $\gamma_Z(C_2^*) = C_0 + C_1(1 - e^{-3}) = C_0 + 0,95C_1$.

Os Modelos Esférico e Exponencial apresentam uma evolução inicial linear, o que significa que o semivariograma não é derivável na origem e, consequentemente, a média quadrática da função aleatória é contínua e não é diferenciável. As duas estruturas revelam uma maior continuidade espacial, ou seja, as correlações mais elevadas a grandes distâncias levam a prolongamentos maiores.

Modelo Gaussiano

O Modelo Gaussiano, acompanhado com o efeito de pepita, é apropriado para a modelação de comportamentos de fenómenos regulares e contínuos, com um crescimento lento e um comportamento parabólico na origem, e é dado pela expressão

$$\gamma(d) = \begin{cases} 0 & , d = 0 \\ C_0 + C_1\left(1 - e^{-\frac{||d||^2}{C_2}}\right) & , d \neq 0, \end{cases} \tag{2.29}$$

em que C_0, C_1 e C_2 representam o efeito de pepita, a amplitude a partir da qual os valores do processo deixam de estar correlacionados e o patamar é o valor $\gamma_Z(d)$ para o qual o variograma se estabiliza, respetivamente.

O Modelo Gaussiano é um caso particular da família de "Matern" e alcança o patamar de forma assintótica. A amplitude $C_2^* = \sqrt{3C_2}$ é a distância para a qual os valores do modelo atinjam 95 % do patamar: $\gamma_Z(C_2^*) = C_0^2 + 0,95C_1$.

2.4 Estimação Estocástica

2.4.1 Estimação Linear

Considere-se $Z(s)$ um processo aleatório e $z(s_1), z(s_2), \ldots, z(s_n)$ realizações do processo, nos pontos $s_1, s_2, \ldots, s_n \in D \subseteq \mathbb{R}^r$. O problema mais simples é estimar um determinado valor desconhecido de $Z(s)$, num ponto $s = s_0 \in D$, através de $z(s_1), z(s_2), \ldots, z(s_n)$. A metodologia consiste na construção de uma função real $h(.)$, designada por função de estimação para $Z(s_0)$, do estimador $\hat{Z}(s_0) = h(Z(s_1), Z(s_2), \ldots, Z(s_n))$ e na determinação do erro de estimação tal que $\hat{Z}(s_0) - Z(s_0)$. A qualidade da estimativa $\hat{Z}(s_0)$ pode ser medida pelo erro quadrático médio que é dado por

$$EQM = E[(\hat{Z}(s_0) - Z(s_0))^2], \tag{2.30}$$

em alternativa,

$$EQM = Var\big[\hat{Z}(s_0) - Z(s_0)\big] + B^2[\hat{Z}(s_0) - Z(s_0)], \tag{2.31}$$

em que $B^2[\hat{Z}(s_0) - Z(s_0)]$ é o viés do estimador. Com base nesta métrica, a função de estimação $h(.)$ pode ser definida como a função tal que EQM seja mínimo.

A melhor função de estimação de $Z(s_0)$, construída a partir de $Z(s_1), Z(s_2), \ldots, Z(s_n)$, é determinada pela esperança de $Z(s_0)$, condicionado a $Z(s_1), Z(s_2), \ldots, Z(s_n)$, se

$$h(Z(s_1), Z(s_2), \ldots, Z(s_n)) = E[Z(s_0)|Z(s_1), Z(s_2), \ldots, Z(s_n)]. \tag{2.32}$$

Considerando $H^*(.)$ o conjunto das funções de estimação das combinações lineares de $Z(s_1), Z(s_2), \ldots, Z(s_n)$, a função $h^*(.)$ é a melhor função linear para a estimação de $Z(s_0)$ se $h^*(Z(s_1), Z(s_2), \ldots, Z(s_n)) \in H^*$, se

$$\forall \lambda_i \in \mathbb{R}, i \in 1, 2, \ldots, n, \qquad h^*(Z(s_1), \ldots, Z(s_n)) = \lambda_0 + \sum_{i=1}^{n} \lambda_i Z(s_i) \tag{2.33}$$

e se

$$E[h^*(Z(s_1), \ldots, Z(s_n)) - Z(s_0)]^2 \leq E[h^*(Z(s_1), \ldots, Z(s_n)) - Z(s_0)]^2, \tag{2.34}$$

define-se:

1. $\mathbf{Z}(s)^T = (Z(s_1), \ldots, Z(s_n))$;

2. $\mathbf{\Sigma}$ é a matriz de covariância de $\mathbf{Z}(s)$, com dimensão $n \times n$, em que $\Sigma_{ij} =$

$Cov(Z(s_i), Z(s_j)) = C_Z(s_i, s_j)$, com $i, j = 1, \ldots, n$;

3. $\mathbf{C}_0$ é o vetor de covariâncias entre $Z(s_0)$ e $Z(s_i)$, $i = 1, \ldots, n$, ou seja, $\mathbf{C}_{0i} = Cov(Z(s_0), Z(s_i))$, com $i = 1, \ldots, n$;

4. $\boldsymbol{\mu}$ é o vetor das esperanças de $Z(s)$, em que $\mu_i = E[Z(s_i)] = m(s_i)$, $i = 1, \ldots, n$;

5. $\mu_0 = E[Z(s_0)]$.

Considerando $\boldsymbol{\Sigma}$, $\mathbf{C}_0$ e $E[\mathbf{Z}(\mathbf{s})]$ conhecidos num conjunto, $\{s_0, \ldots, s_n\}$, define-se como a melhor função linear de estimação de $Z(s_0)$ (*Best Linear Estimation*, BLE), baseada em $\mathbf{Z}(s)$, como

$$h^*(\mathbf{Z}(s)) = \mu_0 + \mathbf{C}_0^T \boldsymbol{\Sigma}^{-1}[Z(s) - \boldsymbol{\mu}]. \tag{2.35}$$

A BLE de $Z(s_0)$ é a projeção ortogonal de $Z(s_0)$ em H^*, e que a BLE de $Z(s_0)$ não é enviesada. O estimador ótimo é aquele que apresenta valor de EQM menor possível e que verifica o não enviesamento, sabendo que a sua é variância mínima. Sabe-se que o BLE de $Z(s_0)$ não é enviesado, logo $h^*(\mathbf{Z}(s))$ é a melhor função linear de estimação não enviesada (*Best Linear Unbiased Estimation*, BLUE) de $Z(s_0)$ em $\mathbf{Z}(s)$.

2.4.2 Estimação Linear Geoestatística

O método de estimação linear geoestatística, *Kriging*, é um método estocástico (interpolação linear) e consiste na modelação da estrutura de variação do processo e esta é utilizada para a construção de um preditor para os valores não observados. Existem vários tipos de *Kriging* que dependem da distribuição do processo aleatório como, por exemplo, *Kriging* Simples, *Kriging* Ordinário, *Kriging* Universal, entre outros.

Goovaerts (1997) considera que os estimadores de *Kriging* são variantes do estimador $\hat{Z}$, considerando o processo aleatório $Z(s)$, $Z(s_1), \ldots, Z(s_n)$, com n observações,

$$\hat{Z}(s) - \boldsymbol{\mu}(s) = \sum_{i=1}^{n(s)} \lambda_i(s)(Z(s_i) - \mu(s_i)), \tag{2.36}$$

em que $\lambda_i(s)$ é o peso de cada observação $Z(s_i)$, com $i = 1, \ldots, n(s)$. O principal objetivo é minimizar a variância do erro, considerando que não existe enviesamento, isto é, sob a hipótese $E[\hat{Z}(s) - Z(s)] = 0$, tem-se que $min\ \sigma^2(s) = Var[\hat{Z}(s) - Z(s)]$.

Considerando um processo aleatório $\{Z(s), s \in D\}$, $D \subseteq \mathbb{R}^r$, pretende-se a estimação sobre pontos ou regiões não observados, no processo $Z(s)$, a partir dum conjunto finito de observações, localizadas em $s_1, s_2, \ldots, s_n$. A partir das observações conhecidas, estabelece-se a melhor função de estimação não enviesada para $Z(s_0)$, $s_0 \in D$, neste caso da aplicação do método de *Kriging* resulta a estimação pontual, ou para $Z(A)$, $A \subseteq D$, neste caso da

aplicação do método de *Kriging* resulta a estimação global, um subconjunto r-dimensional estritamente positivo.

Note-se que existem cenários em que não é possível admitir a estacionariedade na média, por ser desconhecida, e designam-se por fenómenos não-estacionários, cujos valores da característica a estimar apresentam um comportamento não homogéneo na amostra. Na presença destes fenómenos não é possível a utilização do método de estimação de *Kriging* Simples, na medida em que o pressuposto é de estacionaridade de primeira ordem, mas poderão ser aplicados outros tipos de *Kriging*.

Kriging Simples

O *Kriging* Simples é o algoritmo mais geral para os fenómenos não estacionários, onde se assume o conhecimento das médias do conjunto de variáveis aleatórias referentes aos valores amostrados e aos pontos no espaço não amostrado. Sabe-se que o estimador é não enviesado e a sua formulação mais geral é a combinação linear dos n dados, $i = 1, \ldots, n$, e a constante 1, dada por

$$\hat{Z}(s_0) = \sum_{i=1}^{n} \lambda_i Z(s_i) + \lambda_0 1 = \boldsymbol{\lambda}^T \mathbf{Z}(s) + \lambda_0 1. \tag{2.37}$$

Considerando o processo não estacionário de primeira ordem, com médias das variáveis aleatórias conhecidas e não constantes, sem enviesamento $E[\hat{Z}(s_0)] = E[Z(s_0)] = \mu_0$, em que $\lambda_0 = \mu_0 - \sum_{i=1}^{n} \lambda_i E[Z(s_i)] = \mu_0 - \sum_{i=1}^{n} \sum_{i=1}^{n} \lambda_i \mu_i$ e a variância de estimação mínima, tal que $Var[\hat{Z}(s_0) - Z(s_0)] = E[(\hat{Z}(s_0) - Z(s_0))^2]$, tem-se que

$$\hat{Z}(s_0) = \sum_{i=1}^{n} \lambda_i Z(s_i) + \lambda_0 1 = \sum_{i=1}^{n} \lambda_i Z(s_i) + \mu_0 - \sum_{i=1}^{n} \lambda_i \mu_i, \tag{2.38}$$

com variância de estimação dada por

$$\begin{aligned}
\sigma^2_{KS} &= Var[\hat{Z}(s_0) - Z(s_0)] \\
&= E[(\hat{Z}(s_0) - Z(s_0))^2 \\
&= Var\Big[\sum_{i=1}^{n} \lambda_i Z(s_i) + \lambda_0 - Z(s_0)\Big] + E\Big[\big(\sum_{i=1}^{n} \lambda_i Z(s_i) + \lambda_0 - Z(s_0)\big)^2\Big] \\
&= Var\Big[\sum_{i=1}^{n} \lambda_i Z(s_i) - Z(s_0)\Big] + \left(\sum_{i=1}^{n} \lambda_i \mu_i + \lambda_0 - \mu_0\right)^2 \\
&= Var\Big[\sum_{i=1}^{n} \lambda_i Z(s_i) - Z(s_0)\Big] \\
&= \boldsymbol{\lambda}^T \boldsymbol{\Sigma} \boldsymbol{\lambda} - 2\boldsymbol{\lambda}^T \mathbf{C}_0 + Var[Z(s_0)],
\end{aligned} \tag{2.39}$$

em que $\mathbf{\Sigma}$ é a matriz de covariâncias $Z(s)$, de dimensão $n \times n$, e $\mathbf{C}_0$ é o vetor de covariâncias entre $Z(s_0)$ e $Z(s_i)$, $i = 1, \ldots, n$. Maximizando a expressão anterior (ou seja, derivando em ordem a λ e igualando a zero), encontra-se um sistema linear com n equações tais que $\mathbf{\Sigma} = \boldsymbol{\lambda}^{-1}\mathbf{C}_0$ e a variância dada por $\sigma^2_{KS} = -\boldsymbol{\lambda}^T\mathbf{C}_0 + Var[Z(s_0)]$. Pelo método de *Kriging* Simples, o estimador de $Z(s_0)$ é dado por

$$\begin{aligned} \hat{Z}_{SK}(s_0) &= \boldsymbol{\lambda}^T\mathbf{Z}(s) + \lambda_0 1 \\ &= \boldsymbol{\lambda}^T\mathbf{Z}(s) + \mu_0 - \boldsymbol{\lambda}^T\boldsymbol{\mu} \\ &= \mu_0 + \boldsymbol{\lambda}^T(\mathbf{Z}(s) - \boldsymbol{\mu}) \\ &= \mu_0 + \mathbf{C}_0^T\mathbf{\Sigma}^{-1}[\mathbf{Z}(s) - \boldsymbol{\mu}]. \end{aligned} \tag{2.40}$$

Na maioria dos casos práticos, a função de covariância não é conhecida e é necessário introduzir a hipótese de estacionaridade, para que a estimação seja possível. Esta condição permite estimar a forma da função de covariância, com base nas observações da amostra. Supondo um processo estacionário de segunda ordem, para as n observações, sabe-se que o estimador de $Z(s_0)$ define-se por

$$\hat{Z}(s_0) = \sum_{i=1}^{n} \lambda_i Z(s_i) = \boldsymbol{\lambda}^T\mathbf{Z}(s), \tag{2.41}$$

em que $\boldsymbol{\lambda}^T = (\lambda_1, \ldots, \lambda_n)$. O estimador $\hat{Z}(s_0)$ tem EQM mínimo, em todas as combinações de $\mathbf{Z}(s)$ não enviesadas, e é não enviesado ($E[\hat{Z}(s_0) - Z(s_0)] = 0$).

Kriging Ordinário

O *Kriging* Ordinário é utilizado quando o processo aleatório tem média constante e desconhecida em D, $m = E[Z(s)]$, a sua formulação é dada por

$$\begin{aligned} E\big[\hat{Z}_{OK}(s_0) - Z(s_0)\big] &= E\Big[\sum_{i=1}^{n} \lambda_i Z(s_i) - Z(s_0)\Big] \\ &= \sum_{i=1}^{n} \lambda_i m - m \\ &= m\bigg(\sum_{i=1}^{n} \lambda_i - 1\bigg). \end{aligned} \tag{2.42}$$

Uma das condições é o estimador não ser enviesado e para tal limitam-se os ponderadores à condição $\sum_{i=1}^{n} \lambda_i = 1$. Se $\boldsymbol{\lambda}^T\mathbf{Z}(s)$ é uma combinação linear de $\mathbf{Z}(s)$, então o erro

quadrático médio é igual à sua variância, isto é,

$$\begin{aligned} EQM(\boldsymbol{\lambda}\mathbf{Z}(s)) &= E\Big[\big(\boldsymbol{\lambda}\mathbf{Z}(s) - Z(s_0)\big)^2\Big] \\ &= Var\big[\boldsymbol{\lambda}\mathbf{Z}(s) - Z(s_0)\big] \\ &= \boldsymbol{\lambda}^T\boldsymbol{\Sigma}\boldsymbol{\lambda} - 2\boldsymbol{\lambda}^T\mathbf{C}_0 + Var[Z(s_0)], \end{aligned} \tag{2.43}$$

em que $\boldsymbol{\Sigma}$ é a matriz de covariância de $\mathbf{Z}(s)$ e $\mathbf{C}_0$ é o vetor de dimensão n. Assim, estabelecem-se as seguintes condições

$$\sum_{i=1}^{n} \lambda_i = 1 \tag{2.44}$$

e

$$\boldsymbol{\lambda}^T\boldsymbol{\Sigma}\boldsymbol{\lambda} - 2\boldsymbol{\lambda}^T\mathbf{C}_0 \quad \text{é mínimo.} \tag{2.45}$$

Para a obtenção da minimização da expressão (2.45), recorre-se ao multiplicador de Lagrange, $\bigtriangledown$, e tem-se que

$$\begin{aligned} &\frac{\boldsymbol{\lambda}^T\boldsymbol{\Sigma}\boldsymbol{\lambda} - 2\boldsymbol{\lambda}^T\mathbf{C}_0 - 2(\mathbf{1}^T\boldsymbol{\lambda} - 1)\bigtriangledown}{d\boldsymbol{\lambda}} = 0 \\ &\Rightarrow 2\boldsymbol{\Sigma}\boldsymbol{\lambda} - 2\mathbf{C}_0 - 2\mathbf{1}\bigtriangledown = 0 \\ &\Rightarrow \boldsymbol{\Sigma}\boldsymbol{\lambda} - \mathbf{C}_0 - \mathbf{1}\bigtriangledown = 0 \\ &\Rightarrow \boldsymbol{\Sigma}\boldsymbol{\lambda} - \mathbf{1}\bigtriangledown = \mathbf{C}_0 \\ &\Rightarrow \boldsymbol{\Sigma}\boldsymbol{\lambda} = \mathbf{C}_0 + \mathbf{1}\bigtriangledown \\ &\Rightarrow \boldsymbol{\lambda} = \boldsymbol{\Sigma}^{-1}\mathbf{C}_0 + \boldsymbol{\Sigma}^{-1}\mathbf{1}\bigtriangledown \\ &\Rightarrow \mathbf{1}^T\boldsymbol{\lambda} = \mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{C}_0 + \mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}\bigtriangledown . \end{aligned} \tag{2.46}$$

Pela condição $\mathbf{1}^T\boldsymbol{\lambda} = \sum_{i=1}^{n} \lambda_i = 1$, tem-se que

$$\begin{aligned} &\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{C}_0 + \mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}\bigtriangledown = 1 \\ &\Rightarrow \bigtriangledown = \frac{1 - \mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{C}_0}{\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}} \quad . \end{aligned} \tag{2.47}$$

O vetor $\boldsymbol{\lambda}$ é estimado através de

$$\hat{\boldsymbol{\lambda}} = \boldsymbol{\Sigma}^{-1}\mathbf{C}_0 + \boldsymbol{\Sigma}^{-1}\mathbf{1}\frac{1 - \mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{C}_0}{\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}} \tag{2.48}$$

e o estimador para $Z(s_0)$ obtido por *Kriging* Ordinário é tal que

$$\hat{Z}_{KO}(s_0) = \mathbf{C}_0^T \boldsymbol{\Sigma}^{-1} \mathbf{Z}(s_0) + \frac{1 - \mathbf{1}^T \boldsymbol{\Sigma}^{-1} \mathbf{C}_0}{\mathbf{1}^T \boldsymbol{\Sigma}^{-1} \mathbf{1}} \mathbf{1}^T \boldsymbol{\Sigma}^{-1} \mathbf{Z}(s_0), \tag{2.49}$$

com a variância do erro de estimação de *Kriging* dada por

$$\begin{aligned} \sigma_{KO}^2 &= Var[\hat{Z}_{KO} - Z(s_0)] \\ &= \boldsymbol{\lambda}^T \boldsymbol{\Sigma} \boldsymbol{\lambda} - 2\boldsymbol{\lambda}^T \mathbf{C}_0 + Var[Z(s_0)] \\ &= \boldsymbol{\lambda}^T \mathbf{C}_0 + \boldsymbol{\lambda}^T \mathbf{1} \bigtriangledown - 2\boldsymbol{\lambda}^T \mathbf{C}_0 + Var[Z(s_0)]. \end{aligned} \tag{2.50}$$

Sabe-se que $\boldsymbol{\lambda}^T \mathbf{1} = 1$, então o estimador pode ser escrito através de

$$\sigma_{KO}^2(s_0) = -\boldsymbol{\lambda}^T \mathbf{C}_0 + \bigtriangledown + Var[Z(s_0)]. \tag{2.51}$$

Considerando o processo estacionário de segunda ordem (intrinsecamente estacionário), pode-se adaptar o método de estimação de *Kriging* Ordinário e exprimir as equações de covariâncias, em termos do semivariograma,

$$\Lambda_{ij} = \gamma_Z(s_i, s_j), \tag{2.52}$$

em que $\boldsymbol{\Lambda}$ é a matriz de semivariogramas de $\mathbf{Z}(s)$, dimensão $n \times n$, com $i, j = 1, \ldots, n$,

$$\upsilon_{0i} = \gamma_Z(s_0, s_i), \tag{2.53}$$

em que υ_{0i} é o vetor de semivariogramas entre $Z(s_0)$ e $Z(s_i)$, e $\mathbf{1}$ é o vetor composto por valores unitários, de dimensão n. Através da relação $\gamma_Z(d) = C_Z(0) - C_Z(d)$, tem-se que $Var[Z(s_0)] = C_Z(0)$, $\mathbf{C}_0 = C_Z(0)\mathbf{1} - \boldsymbol{\upsilon}_0$ e $\boldsymbol{\Sigma} = C_Z(0)\mathbf{1}\mathbf{1}^T - \boldsymbol{\Lambda}$. O erro quadrático médio pode ser expresso

$$\begin{aligned} EQM\big(\boldsymbol{\lambda}^T \mathbf{Z}(s)\big) &= \boldsymbol{\lambda}^T [C_Z(0)\mathbf{1}\mathbf{1}^T - \boldsymbol{\Lambda}]\boldsymbol{\lambda} - 2\boldsymbol{\lambda}^T [C_Z(0)\mathbf{1} - \boldsymbol{\upsilon}_0] + C_Z(0) \\ &= C_Z(0)\boldsymbol{\lambda}^T \mathbf{1}\mathbf{1}^T \boldsymbol{\lambda} - \boldsymbol{\lambda}^T \boldsymbol{\Lambda} \boldsymbol{\lambda} - 2\boldsymbol{\lambda}^T C_Z(0) + 2\boldsymbol{\lambda}^T \boldsymbol{\upsilon}_0 + C_Z(0), \end{aligned} \tag{2.54}$$

tendo em conta que $\boldsymbol{\lambda}\mathbf{1} = 1$, tem-se que

$$EQM\big(\boldsymbol{\lambda}^T \mathbf{Z}(s)\big) = 2\boldsymbol{\lambda}^T \boldsymbol{\upsilon}_0 + \boldsymbol{\lambda}^T \boldsymbol{\Lambda} \boldsymbol{\lambda}. \tag{2.55}$$

As condições apresentadas anteriormente podem ser reescritas

$$\sum_{i=1}^{n} \lambda_i = 1 \tag{2.56}$$

e

$$2\boldsymbol{\lambda}^T\boldsymbol{v}_0 + \boldsymbol{\lambda}^T\boldsymbol{\Lambda}\boldsymbol{\lambda} \text{ é mínimo.} \tag{2.57}$$

Para a minimização de (2.57), utilizando o multiplicador de Lagrange, $\bigtriangledown$,

$$\begin{aligned}
&\frac{2\boldsymbol{\lambda}^T\boldsymbol{v}_0 - \boldsymbol{\lambda}^T\boldsymbol{\Lambda}\boldsymbol{\lambda} - 2(\mathbf{1}^T\boldsymbol{\lambda} - 1)\bigtriangledown}{d\boldsymbol{\lambda}} = 0 \\
&\Rightarrow 2\boldsymbol{v}_0 - 2\boldsymbol{\Lambda}\boldsymbol{\lambda} - 2\mathbf{1}^T\bigtriangledown = 0 \\
&\Rightarrow \boldsymbol{v}_0 - \boldsymbol{\Lambda}\boldsymbol{\lambda} - \mathbf{1}^T\bigtriangledown = 0 \\
&\Rightarrow \boldsymbol{\Lambda}\boldsymbol{\lambda} = \boldsymbol{v}_0 - \mathbf{1}^T\bigtriangledown \\
&\Rightarrow \boldsymbol{\lambda} = \boldsymbol{\Lambda}^{-1}\boldsymbol{v}_0 - \boldsymbol{\Lambda}^{-1}\mathbf{1}^T\bigtriangledown \\
&\Rightarrow \mathbf{1}^T\boldsymbol{\lambda} = \mathbf{1}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{v}_0 - \mathbf{1}^T\boldsymbol{\Lambda}^{-1}\mathbf{1}\bigtriangledown .
\end{aligned} \tag{2.58}$$

Pela condição $\mathbf{1}^T\boldsymbol{\lambda} = \sum_{i=1}^{n}\lambda_i = 1$ tem-se que

$$\begin{aligned}
\mathbf{1}^T\boldsymbol{\lambda} = \mathbf{1}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{v}_0 - \mathbf{1}^T\boldsymbol{\Lambda}^{-1}\mathbf{1}\bigtriangledown = 1 \\
\Rightarrow \bigtriangledown = -\frac{1 - \mathbf{1}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{v}_0}{\mathbf{1}^T\boldsymbol{\Lambda}^{-1}\mathbf{1}}.
\end{aligned} \tag{2.59}$$

O vetor $\boldsymbol{\lambda}$ é estimado através de

$$\hat{\boldsymbol{\lambda}} = \boldsymbol{\Lambda}^{-1}\boldsymbol{v}_0 + \boldsymbol{\Lambda}^{-1}\mathbf{1}\frac{1 - \mathbf{1}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{v}_0}{\mathbf{1}^T\boldsymbol{\Lambda}^{-1}\mathbf{1}}, \tag{2.60}$$

o estimador para $Z(s_0)$ obtido por *Kriging* Ordinário é tal que

$$\hat{Z}_{KO} = \boldsymbol{v}_0\boldsymbol{\Lambda}^{-1}Z(s) + \frac{1 - \mathbf{1}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{v}_0}{\mathbf{1}^T\boldsymbol{\Lambda}^{-1}\mathbf{1}}\mathbf{1}^T\boldsymbol{\Lambda}^{-1}Z(s) \tag{2.61}$$

e a variância do erro de estimação é dada por

$$\begin{aligned}
\hat{\sigma}^2_{KO} &= Var\big[\hat{Z}_{KO}(s_0) - Z(s_0)\big] \\
&= EQM\big(\hat{Z}_{KO}(s_0)\big) \\
&= 2\boldsymbol{\lambda}^T\boldsymbol{v}_0 - \boldsymbol{\lambda}^T\boldsymbol{\Sigma}\boldsymbol{\lambda} \\
&= 2\boldsymbol{\lambda}^T\boldsymbol{v}_0 - \boldsymbol{\lambda}^T(\boldsymbol{v}_0 - \mathbf{1}\bigtriangledown) \\
&= 2\boldsymbol{\lambda}^T\boldsymbol{v}_0 - \boldsymbol{\lambda}^T\boldsymbol{v}_0 + \boldsymbol{\lambda}^T\mathbf{1}\bigtriangledown \\
&= \boldsymbol{\lambda}^T\boldsymbol{v}_0 + \bigtriangledown \\
&= {\boldsymbol{v}_0}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{v}_0 - \frac{(1 - \mathbf{1}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{v}_0)^2}{\mathbf{1}^T\boldsymbol{\Lambda}^{-1}\mathbf{1}}.
\end{aligned} \tag{2.62}$$

Propriedades do Estimador

Nos métodos apresentados, a variância do erro de estimação não depende apenas da esperança, nem das observações, mas depende do segundo momento do processo aleatório. Então, é possível conhecer a qualidade do estimador antes de se observar o processo.

A estimativa do estimador nos pontos observados é igual à função nesses pontos, $\hat{Z}(s_i) = Z(s_i)$, $i = 1, \ldots, n$, ou seja, é um interpolador exato.

Considerando que $Z(u)$ e $Z(w)$ são variáveis aleatórias independentes, com $u \neq w$, e $Var[Z(s)] = \sigma_Z^2$, não depende de s, ou seja, $\{Z(s), s \in D\}$ é um processo de ruído branco. Então, a estimação em qualquer ponto s_0 é a média aritmética dos valores observados, do processo $Z(.)$, ou seja,

$$\hat{Z}(s_0) = \frac{1}{n}\sum_{i=1}^{n} Z(s_i) = \bar{Z}, \qquad \forall s_0 \neq s_i,\ i = 1, \ldots, n, \tag{2.63}$$

para além disso, tem-se que

$$\hat{\lambda} = \mathbf{1}\frac{1}{n} \qquad e \qquad \bigtriangledown = \frac{\sigma_Z^2}{n} \tag{2.64}$$

e a variância de estimação é tal que

$$\sigma_{KO}^2(s_0) = Var[\hat{Z}(s_0) - Z(s_0)] = \sigma_Z^2 + \frac{\sigma_Z^2}{n}. \tag{2.65}$$

Das propriedades supracitadas pode-se concluir que a superfície dos $\hat{Z}(.)$ pode não ser contínua.

Considerando o modelo $Z(s) = m(s) + Y(s)$, $s \in D$, $Y(s)$ é um processo aleatório de covariância estacionária, média zero e função covariância conhecida.

No caso de $m(s) = m$ conhecido, então

$$\begin{aligned} \hat{Y}(s_0) &= \mathbf{C}_0^T \boldsymbol{\Sigma}^{-1} \mathbf{Y}(s) \\ &= \mathbf{C}_0^T \boldsymbol{\Sigma}^{-1} \big[\mathbf{Z}(s) - \mathbf{1}m\big] \\ &= \mathbf{C}_0^T \boldsymbol{\Sigma}^{-1} \mathbf{Z}(s) - \mathbf{C}_0^T \boldsymbol{\Sigma}^{-1} \mathbf{1}m \end{aligned} \tag{2.66}$$

e, consequentemente, o estimador é dado por

$$\begin{aligned} \hat{Z}(s_0) &= m + \hat{Y}(s_0) \\ &= \mathbf{C}_0^T \boldsymbol{\Sigma}^{-1} \mathbf{Z}(s) + (1 - \mathbf{C}_0^T \boldsymbol{\Sigma}^{-1} \mathbf{1})m, \end{aligned} \tag{2.67}$$

cujo erro quadrático médio determina-se por

$$Var[\hat{Z}(s_0) - Z(s_0)] = Var\big[\mathbf{C}_0^T\boldsymbol{\Sigma}^{-1}\mathbf{Z}(s) - Z(s_0)\big]. \tag{2.68}$$

No caso de $m(s) = m$ desconhecido, então

$$\hat{m} = \big(\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}\big)^{-1}\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{Z}(s), \tag{2.69}$$

com variância

$$Var[\hat{m}] = \big(\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}\big)^{-1}. \tag{2.70}$$

Desta forma,

$$\hat{Z}^*(s_0) = \mathbf{C}_0^T\boldsymbol{\Sigma}^{-1}\mathbf{Z}(s) + \big(1 - \mathbf{C}_0^T\boldsymbol{\Sigma}^{-1}\mathbf{1}\big)\big(\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}\big)^{-1}\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{Z}(s), \tag{2.71}$$

cuja variância é determinada por

$$Var\big[\hat{Z}(s_0) - Z(s_0)\big] = Var\big[\hat{Z}^*(s_0) - Z(s_0)\big] + \big(1 - \mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{C}_0\big)^2 Var[\hat{m}], \tag{2.72}$$

a parcela da equação reflete a perda de precisão quando se estima m.

Outros Métodos de Estimação

A restrição de estacionariedade na média desconhecida, mas constante para todo o domínio, nem sempre é fácil de cumprir, nomeadamente, nos fenómenos não estacionários e, neste caso, é utilizado o *Kriging* Universal (ou *Kriging* com deriva externa).

O *Kriging* é um método que permite a estimação de um processo $\{Z(s), s \in D\}$, $D \subseteq \mathbb{R}^r$, utilizando os valores observados deste processo, nas localizações $s_1, s_2, \ldots, s_n$. Porém, é possível melhorar a qualidade da estimação através de observações de outro processo, que tem em conta observações de processos que estão correlacionados com o que se pretende estudar. Neste sentido, é possível aplicar um método de estimação de um processo $\{Z_1(s), s \in D\}$, $D \subseteq \mathbb{R}^r$, recorrendo aos valores observados, nas localizações $s_1, s_2, \ldots, s_n$, e a outros processos, como $Z_2(s), s \in D$, $D \subseteq \mathbb{R}^r$, $Z_3(s), s \in D$, $D \subseteq \mathbb{R}^r$, entre outros. A técnica é uma solução possível quando se tem um número reduzido de observações, no processo em análise, e outros processos têm um maior número de observações, designando-se por *Cokriging* (que não vai ser aplicado nesta dissertação).

2.5 Estimação Global

Nas Secções anteriores descreveram-se os conceitos básicos para o método de estimação de *Kriging* pontual (de um valor do atributo do processo). Mas, quando se pretende a estimação do valor médio $Z(.)$ numa determinada região $A \subseteq D$, em que A é um subconjunto com volume r-dimensional estritamente positivo ($|A| > 0$), o método denomina-se por *Kriging* Global. Este pode ser obtido pela média dos valores pontuais estimados pelo *Kriging* que compõem A ou pode ser estimado diretamente.

O valor médio do processo aleatório $Z(.)$, numa determinada região A é dado por

$$Z(A) = \frac{1}{|A|}\int_A Z(v)dv. \tag{2.73}$$

De acordo com a teoria de processos aleatórios, define-se o integral de um processo aleatório de uma variável aleatória como o limite de uma soma de Riemann, tendo como suporte a definição supracitada, deduzindo-se as expressões seguintes

$$E\big[Z(A)\big] = \frac{1}{|A|}\int_A E[Z(v)]dv, \tag{2.74}$$

$$Var\big[Z(A)\big] = Cov\big(Z(A), Z(A)\big) = \frac{1}{|A|^2}\int_A\int_A Cov\big(Z(v), Z(w)\big)dw\,dv, \tag{2.75}$$

$$Cov(Z(A), Z(v)) = \frac{1}{|A|}\int_A Cov\big(Z(v), Z(u)\big)du, \forall v \in A. \tag{2.76}$$

Na presença de um processo intrinsecamente estacionário, com semivariograma $\gamma_Z(.)$, podem-se escrever as equações anteriores como

$$E[Z(A)] = \mu_Z, \tag{2.77}$$

$$Var\big[Z(A)\big] = \frac{1}{|A|}\int_A\int_A Var\big[Z(v)\big]dv - \frac{1}{|A|^2}\int_A\int_A \gamma_Z(v-w)dw\,dv \tag{2.78}$$

e

$$\frac{1}{2}Var\big[Z(A) - Z(v)\big] = \int_A \gamma_Z(v-w)dw - \frac{1}{|A|^2}\int_A\int_A \gamma_Z(w-u)du\,dv, \forall v \in A. \tag{2.79}$$

Devido à complexidade das expressões e dispêndio no cálculo numérico, usualmente utilizam-se aproximações. Neste sentido, numa região A com uma amostra finita de valores de atributos, localizados nessa região, os integrais das expressões anteriores podem ser substituídos por médias e gerando, desta forma, aproximações

$$Var[Z(B)] \simeq \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} Cov\big[Z(s_i) - Z(s_j)\big] \tag{2.80}$$

ou

$$Var[Z(B)] \simeq \frac{1}{n}\sum_{i=1}^{n} Var\big[Z(s_i)\big] - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\gamma_Z(Z(s_i) - Z(s_j)), \tag{2.81}$$

$$Cov[Z(B), Z(s)] \simeq \frac{1}{n}\sum_{i=1}^{n} Cov\big[Z(s_i) - Z(s)\big], \forall s \in B, \tag{2.82}$$

e

$$\frac{1}{2}Var\big[Z(B) - Z(v)\big] = \frac{1}{n}\sum_{i=1}^{n}\gamma_Z(s_i - s) - \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\gamma_Z(s_i - s_j), \forall v \in A. \tag{2.83}$$

Na estimação global, o estimador $\hat{Z}(A)$, $A \subseteq D$, obtido a partir das observações pontuais, é semelhante ao descrito na estimação pontual. O estimador de *Kriging* Ordinário é dado por

$$\hat{Z}_{KO}(A) = \sum_{i=1}^{n}\lambda_{A,i}Z(s_i) = \boldsymbol{\lambda}_A^T\mathbf{Z}(s), \tag{2.84}$$

em que $\boldsymbol{\lambda}_A^T = (\lambda_{A,1}, \lambda_{A,2}, \ldots, \lambda_{A,n})$, $\mathbf{Z}(s) = (Z(s_1), Z(s_2), \ldots, Z(s_n))$ é o vetor das observações do processo. Relativamente às covariâncias, um processo estacionário de segunda ordem, em D, tem variância finita e a função covariância existe e o vetor $\boldsymbol{\lambda}$ é dado por

$$\hat{\boldsymbol{\lambda}}_A = \boldsymbol{\Sigma}^{-1}\mathbf{C}_A + \boldsymbol{\Sigma}^{-1}\mathbf{1}\frac{1 - \mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{c}_A}{\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}}, \tag{2.85}$$

em que $\boldsymbol{\Sigma}$ é a matriz de covariâncias de $\mathbf{Z}(s)$, $\mathbf{C}_A$ é o vetor de covariâncias entre $\mathbf{Z}(B)$ e $\mathbf{Z}(s_i)$, com $i = 1, \ldots, n$, tal que $\mathbf{C}_{A_i} = Cov(\mathbf{Z}(B), \mathbf{Z}(s_i))$, $\mathbf{1}$ é o vetor de valores unitários. O estimador de *Kriging* Ordinário é dado por

$$\hat{Z}_{KO}(A) = \mathbf{C}_A^T\boldsymbol{\Sigma}^{-1}\mathbf{Z}(s) + \frac{1 - \mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{C}_A}{\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}}\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{Z}(s) \tag{2.86}$$

e a sua variância do erro de estimação é dada por

$$\hat{\sigma}_{KO}(A) = -\boldsymbol{\lambda}_A^T\mathbf{C}_A + \bigtriangledown + Var\big[Z(A)\big]. \tag{2.87}$$

As equações de covariância podem ser formuladas, em termos de semivariograma. Considerando um processo estacionário de segunda ordem, intrinsecamente estacionário, o método de estimação de *Kriging* Ordinário é caraterizado pela matriz de semivariogramas

de $\mathbf{Z}(s)$, tal que

$$\Lambda_{ij} = \gamma_Z(s_i, s_j), \qquad i, j = 1, \ldots, n, \tag{2.88}$$

pelo vetor de semivariogramas entre $\mathbf{Z}(B)$ e $\mathbf{Z}(s_i)$, $i = 1, \ldots, n$,

$$\upsilon_A = \gamma_Z(A, s_i), \qquad i = 1, \ldots, n, \tag{2.89}$$

e pelo vetor composto por n valores unitários. A variância do erro da estimação é dada por

$$\begin{aligned}
\hat{\sigma}^2_{KO} &= Var\big[\hat{Z}_{KO}(A) - Z(A)\big] \\
&= EQM(\hat{Z}_{KO}(A)) \\
&= Var\Big[\sum_{i=1}^{n} \lambda_{A_i} Z(s_i) - Z(B)\Big] \\
&= 2\sum_{i=1}^{n} \lambda_{A,i} \frac{1}{2} Var\big[Z(A) - Z(s_i)\big] - \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_{A_j}\lambda_{A_i} \frac{1}{2} Var\big[Z(s_j) - Z(s_i)\big]
\end{aligned} \tag{2.90}$$

e, sob a forma matricial,

$$\hat{\sigma}^2_{KO} = 2\boldsymbol{\lambda}_A^T \boldsymbol{\upsilon}_A - \boldsymbol{\lambda}_A^T \boldsymbol{\Lambda} \boldsymbol{\lambda}_A. \tag{2.91}$$

Assim, as condições impostas anteriormente reformulam-se para as seguintes

$$\sum_{i=1}^{n} \lambda_{A,i} = 1 \tag{2.92}$$

e

$$2\boldsymbol{\lambda}_A^T \boldsymbol{\upsilon}_A + \boldsymbol{\lambda}_A^T \boldsymbol{\Lambda} \boldsymbol{\lambda}_A \text{ é mínimo.} \tag{2.93}$$

Recorrendo ao multiplicador de Lagrange, tem-se que

$$\hat{\boldsymbol{\lambda}}_A = \boldsymbol{\Lambda}^{-1}\boldsymbol{\upsilon}_A + \boldsymbol{\Lambda}^{-1}\mathbf{1}\frac{1 - \mathbf{1}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{\upsilon}_A}{\mathbf{1}^T\boldsymbol{\Lambda}^{-1}\mathbf{1}}, \tag{2.94}$$

em que

$$\bigtriangledown_A = -\frac{1 - \mathbf{1}^T\boldsymbol{\Lambda}^{-1}\upsilon_A}{\mathbf{1}^T\boldsymbol{\Lambda}^{-1}\mathbf{1}}, \tag{2.95}$$

o estimador de *Kriging* Ordinário de $Z(A)$ é

$$\hat{\mathbf{Z}}(A) = \boldsymbol{\upsilon}_A^T\boldsymbol{\Lambda}^{-1}\mathbf{Z}(s) + \frac{1 - \mathbf{1}^T\boldsymbol{\Lambda}^{-1}\upsilon_A}{\mathbf{1}^T\boldsymbol{\Lambda}^{-1}\mathbf{1}}\mathbf{1}^T\boldsymbol{\Lambda}^{-1}\mathbf{Z}(s) \tag{2.96}$$

e a variância do erro de estimação é

$$\begin{aligned}\hat{\sigma}^2_{KO} &= 2\boldsymbol{\lambda}_A^T \boldsymbol{v}_A - \boldsymbol{\lambda}_A^T \boldsymbol{\Lambda} \boldsymbol{\lambda}_A \\ &= \boldsymbol{\lambda}_A^T \boldsymbol{v}_A + \bigtriangledown_A \\ &= \boldsymbol{\lambda}_A^T \boldsymbol{v}_A - \frac{1 - \mathbf{1}^T \boldsymbol{\Lambda}^{-1} \boldsymbol{v}_A}{\mathbf{1}^T \boldsymbol{\Lambda}^{-1} \mathbf{1}}.\end{aligned} \tag{2.97}$$

Uma abordagem alternativa ao método exposto para a estimação global consiste na determinação do estimador pontual para todos os pontos e na obtenção da média dos valores estimados obtidos. Sob a hipótese de estacionariedade na região A, pode-se provar que estas duas metodologias são equivalentes. A abordagem utiliza uma grelha de pontos para a aproximação numérica do vetor v_A e determina o estimador global, com base nessa aproximação, o que é equivalente a calcular todos os estimadores pontuais na grelha e o valor estimado para $Z(A)$ ser considerado como a média de todos os estimadores pontuais. No entanto, as variâncias dos erros de estimação pontuais e globais não têm uma relação tão simplista, porque a variância não se traduz em combinações lineares.

2.6 Validação Cruzada

O método de estimação requer que sejam validados os modelos de variograma e as hipóteses de homogeneidade espacial para todo o campo em análise, prevenindo dificuldades do ajustamento, como a presença de valores discrepantes (*outliers*).

Um método bastante utilizado é a Validação Cruzada (*cross-validation*), que utiliza a informação disponível e compara os valores observados e os estimados, nas localizações das observações. Com os valores reais e os valores estimados $(Z(s_i), \hat{Z}(s_i))$, $i = 1, \ldots, n$, pode-se determinar estatísticas das distribuições univariadas das estimativas e dos erros, para se determinar a qualidade do modelo do variograma adotado.

O procedimento inicia-se com o cálculo do preditor, através de *Kriging*, e a sua variância de precisão, ou seja, determina-se $\hat{Z}(s_i)$, $i = 1, \ldots, n$, em função das variáveis $Z(s_1), \ldots, Z(s_{i-1}), Z(s_{i+1}), \ldots, Z(s_n)$, $i = 1, \ldots, n$, através do semivariograma ajustado e do valor da variância da estimação.

Posteriormente, para cada i, calcula-se a diferença normalizada entre o estimado e o observado, ou seja, calcula-se $\dfrac{Z(s_i) - \hat{Z}(s_i)}{\sigma(s_i)}$. Para a avaliação da qualidade do ajustamento, utilizam-se os histogramas das diferenças normalizadas para deteção de valores discordantes, o valor médio das diferenças normalizadas, que, no cenário ideal, é zero ou

próximo de zero. Assim, a média do erro de predição é dada por

$$MEP = \frac{1}{n}\sum_{i=1}^{n}\frac{Z(s_i) - \hat{Z}(s_i)}{\sigma(s_i)} \tag{2.98}$$

e o erro quadrático médio normalizado, que no cenário ideal é um ou próximo de um, é dado por

$$EQMN = \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{Z(s_i) - \hat{Z}(s_i)}{\sigma_(s_i)}\Big)^2. \tag{2.99}$$

O processo de estimação da Validação Cruzada é influenciado por três fatores, que estão bastante relacionados, que tornam difícil a perceção da sua influência nos valores das estatísticas globais dos desvios: as hipóteses de estacionaridade/homogeneidade espacial, o modelo de variograma (que se pretende validar) e o próprio processo de estimação. Antes de verificar qual o variograma mais apropriado, deve ser primeiramente validada a hipótese de estacionariedade/ homogeneidade espacial, em segundo lugar há que julgar se os desvios não têm a ver com o tipo de estimador usado. O imbricamento destes fatores torna qualquer destas análises extremamente difícil.

Se se considerar uma pequena área com elevada variabilidade local, então vão ser gerados grandes desvios entre os valores reais e os estimados. Esta dificuldade pode ser ultrapassada pelo aumento do Efeito de Pepita e, assim, conseguir uma atenuação dos desvios locais, com base no aumento da média global, relativamente aos valores locais.

No entanto, um variograma que tem boas métricas na Validação Cruzada não é condição suficiente para que o modelo seja o mais adequado.

Capítulo 3

Séries Temporais

Neste Capítulo apresentam-se conceitos importantes sobre séries temporais, fulcrais para a compreensão da metodologia adotada no Capítulo 4.

3.1 Conceitos

Uma série temporal pode ser descrita simplesmente como um conjunto de observações medidas de forma sequencial ao longo do tempo. Chatfield (2000, 2004) mostra que estas medições podem ser feitas continuamente (série temporal contínua) ou pontualmente (série temporal discreta). As séries temporais podem representar uma variável (série temporal univariada) ou mais do que uma variável (série temporal multivariada).

3.1.1 Componentes

A variação de uma série temporal pode ser decomposta em quatro componentes: a tendência (T), a sazonalidade (S), a componente cíclica (C) e a componente irregular/residual (E), (Kirchgässner & Wolters, 2008; Jebb *et al.*, 2015; Alpuim, 1998).

A tendência pode ser modelada como a inclinação da série temporal, ao longo do período de observação, e pode ser linear, quadrática ou polinomial de grau superior a 3, crescente ou decrescente. A tendência pode ser consequência dos valores observados dependerem de uma componente determinística ou, para alguns autores, de uma componente de natureza estocástica.

A sazonalidade corresponde a um padrão de crescimento e decrescimento, em determinados períodos de tempo, originando oscilações que se repetem. Os períodos de cada oscilação têm duração fixa e devem-se a fatores sazonais. Chatfield (2000) classifica a sazonalidade como aditiva ou multiplicativa. A primeira não varia com o nível da série e a segunda depende do nível da série.

A componente cíclica é um padrão de flutuação, que não apresenta qualquer periodicidade definida, nem efeito associado à sazonalidade. Ou seja, os ciclos são componentes não sazonais que se caracterizam por um padrão. A componente cíclica tem um carácter irregular que se pode prolongar durante um período de tempo. Por esse motivo, quando se trata de séries curtas não é tido em conta. Se for difícil a dissociação da componente cíclica da tendência, considera-se a componente de tendência cíclica (a aglomeração das duas anteriores).

A aleatoriedade exprime a variação não explicada pelos componentes anteriores e representa o ruído aleatório. O ruído chama-se ruído branco, se for modelado por um processo estocástico de variáveis aleatórias independentes (não correlacionadas) de média nula, por regra, e identicamente distribuídas.

3.1.2 Decomposição

Geralmente, para a análise da série temporal, consideram-se as componentes mencionadas e efetua-se a decomposição da série em análise. Considerando o tempo t, Y_t é o valor da série temporal no tempo t, T_t é a componente de tendência, no tempo t, S_t é a componente de sazonalidade no tempo t e E_t é a componente de aleatoriedade (irregular), no tempo t.

Quando cada valor da série temporal é obtido a partir da soma das suas componentes, designa-se pelo modelo de decomposição aditivo e descreve-se por

$$Y_t = T_t + S_t + E_t. \tag{3.1}$$

Quando cada valor da série temporal é obtido a partir da multiplicação das suas componentes, designa-se pelo modelo de decomposição multiplicativo e é dado por

$$Y_t = T_t \times S_t \times E_t. \tag{3.2}$$

É aconselhável que, quando a magnitude das oscilações sazonais não varia com o nível da série, se utilizem os modelos aditivos. Se estas magnitudes forem proporcionais à tendência, é usual recorrer-se ao modelo multiplicativo (Makridakis *et al.*, 1998).

Em muitos casos utiliza-se a transformação logarítmica aplicada aos dados de forma a converter o modelo multiplicativo num modelo aditivo,

$$logY_t = logT_t + logS_t + logE_t. \tag{3.3}$$

Existem também outros tipos de modelos que misturam os dois modelos, como, por

exemplo, o modelo multiplicativo com erros aditivos

$$Y_t = T_t \times S_t + E_t. \tag{3.4}$$

3.2 Processos Estocásticos

Dado um processo estocástico $\{Y(t),\ t \in \mathcal{T}\}$, uma série temporal é um conjunto de observações do processo estocástico em instante $t_1, \ldots, t_n$. Geralmente considera-se t inteiro e as observações são feitas em intervalos de tempo regulares, ou seja, com a mesma amplitude (Alpuim, 1998). Define-se um processo estocástico como qualquer família ou coleção de variáveis aleatórias $\{Y(t),\ t \in \mathcal{T}\}$, em que $\mathcal{T}$ é um conjunto de índices representando o tempo. Ao conjunto de índices $\mathcal{T}$ designa-se por espaço de parâmetros e ao contradomínio das variáveis aleatórias $Y(t)$ dá-se o nome de espaço de estados, representado por S. Usualmente, o processo é de tempo contínuo se $\mathcal{T} = \mathbb{R}^+$, mas pode ser considerado um processo de tempo discreto se $\mathcal{T} = \mathbb{Z}$ ou $\mathcal{T} = \mathbb{N}$.

Os processos estocásticos podem dividir-se em estacionários ou não estacionários. Nesta Secção introduzem-se os dois tipos de estacionariedade (forte e fraca), alguns procedimentos que permitem transformar processos não estacionários em estacionários e outras ferramentas essenciais para a posterior modelação das séries temporais (como as funções de autocorrelação, FAC, e de autocorrelação parcial, FACP, e o processo de ruído branco).

Um processo estocástico $\{Y(t),\ t \in \mathcal{T}\}$ diz-se estritamente estacionário ou fortemente estacionário se a distribuição conjunta de $(Y(t_1), \ldots, Y(t_n))$ e de $(Y(t_1+h), \ldots, Y(t_n+h))$ forem iguais, $\forall h \in \mathbb{R}$ e qualquer que seja o n-úplo $(t_1, \ldots, t_n)$, ou seja,

$$F_{(Y(t_1),\ldots,Y(t_n))}(y_1, \ldots, y_n) = F_{(Y(t_1+h),\ldots,Y(t_n+h))}(y_1, \ldots, y_n). \tag{3.5}$$

Esta propriedade revela que a distribuição de um qualquer conjunto de margens é igual, independentemente das translações no tempo (Alpuim, 1998). O processo fortemente estacionário é difícil de verificar. Na prática, recorre-se aos processos fracamente estacionários ou processos estacionários de 2.^{a} ordem que obedecem a uma propriedade mais fraca mas que descreve "quase" o mesmo tipo de comportamento (Murteira *et al.*, 1993).

Um processo estocástico diz-se estacionário de 2.^{a} ordem ou fracamente estacionário se todos os momentos até à 2.^{a} ordem de $(Y(t_1), \ldots, Y(t_n))$ e de $(Y(t_1+h), \ldots, Y(t_n+h))$ existirem e forem iguais, $\forall h \in \mathbb{R}$ e qualquer que seja o n-úplo $(t_1, \ldots, t_n)$.

Num processo fracamente estacionário verifica-se que o valor médio e a variância não dependem de t e a covariância de $Y(t_1)$ e $Y(t_2)$, depende apenas do desfasamento $t_2 - t_1$, ou seja,

$$\forall t, \qquad \mu(t) = \mu, \tag{3.6}$$

$$\forall t, \qquad \sigma^2(t) = \sigma^2, \tag{3.7}$$

$$\forall t_1, t_2, \qquad Cov(Y(t_1), Y(t_2)) = \gamma(|t_2 - t_1|). \tag{3.8}$$

Nestas condições, os momentos de segunda ordem são finitos. Quando Y_t representa um processo estocástico, contínuo ou discreto, possui certas funções para melhor perceção e análise da série, seguidamente apresentadas.

A função de autocovariância, γ_k, de um processo estacionário permite avaliar a intensidade com que covariam pares de valores do processo, separados por um intervalo (*lag*) de amplitude k, em que $k \in \mathbb{R}$ no processo contínuo ou $k \in \mathbb{N}$ (ou $\mathbb{Z}$) no processo discreto, e define-se por

$$\forall k \in \mathbb{R}(ou \in \mathbb{N}\ ou\ \mathbb{Z}) \qquad \gamma_k = Cov(Y_t, Y_{t+k}) = E[(Y_t - \mu)(T_{t+k} - \mu)]. \tag{3.9}$$

Esta função tem as seguintes propriedades

$$\gamma_0 = Cov(Y_t, Y_t) = Var(Y_t) = \sigma^2, \tag{3.10}$$

$$\gamma_k = \gamma_{-k} \text{ a função é par} \tag{3.11}$$

e, como consequência da Desigualdade de Cauchy-Schwarz, $|E[XY]| \leq \sqrt{E[X^2]E[Y^2]}$,

$$|\gamma_k| \leq \gamma_0, \tag{3.12}$$

e

$$\forall \alpha_1, \ldots, \alpha_n \in \mathbb{R}, \ \ \forall t_1, \ldots, t_n \in \mathcal{T}, \qquad \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \gamma(|t_i - t_j|) \geq 0, \tag{3.13}$$

isto é, a função γ_k é semidefinida positiva.

A função de autocorrelação de um processo estacionário permite medir a correlação entre pares de valores do processo separados por um intervalo (*lag*) de amplitude k, tal que

$$\rho_k = Cor(Y_t, Y_{t+k}) = \frac{Cov(Y_t, Y_{t+k})}{\sqrt{Var(Y_t)Var(Y_{t+k})}} = \frac{Cov(Y_t, Y_{t+k})}{Var(Y_t)} = \frac{\gamma_k}{\gamma_0} \tag{3.14}$$

e a sua representação gráfica é denominada por correlograma. O comportamento do correlograma dá indicações sobre certas características da série para a identificação do modelo mais adequado. Na maioria dos casos práticos, o aumento de k, traduz-se num decréscimo de γ_k. Assim, é expectável que, com o aumento da amplitude do intervalo, k, exista uma perda de memória no processo, ou seja, quando a amplitude k é elevada, é natural que o valor no instante $t+k$ não seja afetado pelo valor no instante t (Murteira *et al.*, 1993). Quando k tende para infinito, a correlação temporal diminui tendendo para zero. À semelhança da função de autocovariância, a função de autocorrelação,γ_k, pode ser definida no tempo contínuo ou discreto e tem as seguintes propriedades

$$\rho_0 = Cor(Y_t, Y_t) = 1, \tag{3.15}$$

$$\rho_k = \rho_{-k} \text{ a função é par}, \tag{3.16}$$

como consequência da Desigualdade de Cauchy-Schwarz, $|E[XY]| \leq \sqrt{E[X^2]E[Y^2]}$,

$$|\rho_k| \leq 1 \tag{3.17}$$

e

$$\forall \alpha_1, \ldots, \alpha_n \in \mathbb{R}, \quad \forall t_1, \ldots, t_n \in \mathcal{T}, \qquad \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \rho(|t_i - t_j|) \geq 0, \tag{3.18}$$

isto é, a função ρ_k é semidefinida positiva.

Considerando um conjunto de observações de um dado processo aleatório estacionário, num determinado período de tempo, definem-se os estimadores clássicos dos parâmetros (Alpuim, 1998)

a média μ estimada por

$$\hat{\mu} = \bar{Y} = \frac{1}{n} \sum_{t=1}^{n} Y_t, \tag{3.19}$$

a autocovariância γ_k estimada por

$$\hat{\gamma_k} = \frac{1}{n} \sum_{t=1}^{n-k} (Y_t - \bar{Y})(Y_{t+k} - \bar{Y}), \tag{3.20}$$

e a autocorrelação ρ_k estimada por

$$\hat{\rho_k} = \frac{\hat{\gamma_k}}{\hat{\gamma_0}} = \frac{\sum_{t=1}^{n-k}(Y_t - \bar{Y})(Y_{t+k} - \bar{Y})}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2}. \tag{3.21}$$

Para além da correlação global, também se pode recorrer à correlação parcial entre Y_t e Y_{t+k}, quando são fixadas as variáveis intermédias $Y_{t+1}, \ldots, Y_{t+k-1}$. É possível obter a correlação parcial com base na regressão linear, tal que

$$Y_{t+k} = \phi_{k1} Y_{t+k-1} + \cdots + \phi_{kk} Y_t + \epsilon_{t+k}, \tag{3.22}$$

em que ϕ_{kj} são os coeficientes do modelo, com $j = 1, \ldots, k$, uma vez que os erros seguem uma distribuição Normal. O valor de ϕ_{kk} é o coeficiente de correlação do Modelo de Regressão Linear, onde $\{\epsilon_t,\ t \in \mathbb{Z}\}$, são independentes e Gaussianos, com média nula e variância σ^2. Este coeficiente exprime a variação entre t e $t + k$, quando os restantes coeficientes são constantes. A partir da equação (3.22), multiplicando por Y_{t+k-j}, $j = 1, \ldots, k$, aplicando o valor esperado e dividindo por ρ_0, tem-se que

$$\rho_j = \phi_{k1}\rho_{j-1} + \cdots + \phi_{kk} Y_{k-j} \qquad j = 1, \ldots, k, \tag{3.23}$$

resolvendo em ordem a ϕ_{kj}, recorrendo à regra de Cramer, consegue-se obter a função de autocorrelação parcial, ϕ_{kk}.

A função de autocorrelação parcial pode-se definir por

$$\phi_{kk} = Cor[Y_t, Y_{t+k} | Y_{t+1}, Y_{t+2}, \ldots, Y_{t+k-1}] = \frac{|P_k^*|}{|P_k|}, \tag{3.24}$$

em que P_k é a matriz $k \times k$ dada por

$$P_k = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \ldots & \rho_{k-1} \\ \rho_1 & 1 & \rho_1 & \ldots & \rho_{k-2} \\ \rho_2 & \rho_1 & 1 & \ldots & \rho_{k-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \ldots & 1 \end{bmatrix} \tag{3.25}$$

e P_k^* é a matriz $k \times k$ de autocorrelações em que a última coluna é substituída por $[\rho_1, \rho_2, \ldots, \rho_k]^T$. Sabe-se que

$$\phi_{11} = \rho_1, \qquad \phi_{22} = \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2} \quad \text{e} \qquad \phi_{33} = \frac{\rho_3(1 - \rho_1^2) + \rho_1(\rho_1^2 + \rho_2^2 - 2\rho_2)}{(1 - \rho_2)(1 - \rho_2 - 2\rho_1^2)}. \tag{3.26}$$

Um processo de ruído branco é um processo estocástico caracterizado pela sucessão

de variáveis aleatórias independentes e identicamente distribuídas com média e variância constantes, com as seguintes propriedades

$$\forall t, \qquad E[\epsilon_t] = \mu_t, \tag{3.27}$$

$$\forall t, \qquad Var(\epsilon_t) = \sigma_\epsilon^2, \tag{3.28}$$

$$\forall t, \quad \forall k = \pm 1, \pm 2, \ldots, \qquad Cov(\epsilon_t, \epsilon_{t+k}) = \gamma_k = 0. \tag{3.29}$$

Se, além disso, as variáveis aleatórias seguirem uma distribuição Normal, designa-se o processo por ruído branco gaussiano. Um ruído branco é um processo estacionário cujas funções de autocorrelação (FAC) e autocorrelação parcial (FACP) são nulas para todo o $k \neq 0$.

3.2.1 Processo Estocástico Não Estacionário

No contexto ambiental, as séries geralmente são não estacionárias. Um processo pode ser não estacionário, na medida em que a média e/ou a variância são funções do tempo e não constantes.

Numa primeira análise, a série pode ser transformada de forma a obter-se uma série estacionária (estabilizar a média e/ou a variância). No caso de uma série não estacionária em média e em variância, deve-se estabilizar a variância e só depois a média (Murteira *et al.*, 1993; Caiado, 2016). Mas também existem métodos que extraem a tendência e a sazonalidade na série temporal, fazendo com que haja estacionariedade, com base na decomposição das suas componentes.

Transformações para a Estacionariedade

Em muitos processos, quando se pretende estabilizar a média, utiliza-se a diferenciação, através da aplicação do operador diferença Δ. Assim, a série, Y_t, não estacionária pode ser sujeita a uma diferenciação de primeira ordem, tal que

$$\Delta Y_t = Y_t - Y_{t-1}, \qquad t = 2, 3, \ldots, n. \tag{3.30}$$

Se, após a aplicação da diferenciação de primeira ordem, não se atingir a estacionariedade, aplica-se a diferenciação de segunda ordem, isto é,

$$\Delta^2 Y_t = \Delta(\Delta Y_t) = \Delta(Y_t - Y_{t-1}) = Y_t - 2Y_{t-1} + Y_{t-2}, \qquad t = 3, 4, \ldots, n. \tag{3.31}$$

O operador de diferenciação de ordem d define-se como

$$\Delta^d Y_t = \Delta(\Delta^{d-1} Y_t), \qquad d \geq 1 \text{ e } \qquad t = d+1, \ldots, n. \tag{3.32}$$

É de evitar a sobrediferenciação na medida em que a variância aumenta com a ordem de diferenciação. Uma boa prática é a diferenciação até à primeira ou à segunda ordem para a obtenção de uma série estacionária.

Para estabilizar a variância de uma série não estacionária pode recorrer-se a transformações paramétricas, como é exemplo a transformação de Box-Cox, dada pela seguinte expressão

$$Z_t = T(Y_t) = \begin{cases} \dfrac{Y_t^{\lambda}}{\lambda}, & \lambda \neq 0, \\ log(Y_t), & \lambda = 0 \end{cases} \qquad , \lambda \in [-1, 1]. \tag{3.33}$$

Habitualmente, este tipo de transformações estão definidas para séries temporais de valores positivos. Para ultrapassar esta dificuldade, nas séries que apresentam valores negativos, utiliza-se a adição de uma constante c que a torne positiva e só depois se recorre às transformações, como o logaritmo. De notar que, após a transformação dos dados, os valores ajustados pelo modelo estarão nas unidades transformadas, isto significa que é necessário reverter as transformações de modo a obter as previsões nas unidades originais (Jebb *et al.*, 2015).

Passeio Aleatório

Um passeio aleatório é caracterizado por movimentos de tendência crescente ou decrescente, em períodos longos, seguidos de mudanças abruptas imprevisíveis e define-se

$$Y_t = Y_{t-1} + \epsilon_t, \tag{3.34}$$

com ϵ_t é um ruído branco. No caso em que Y_0 é conhecido, o modelo de passeio aleatório representa-se por

$$Y_t = Y_0 + \sum_{i=1}^{t} \epsilon_i. \tag{3.35}$$

Porém, o modelo de passeio aleatório é caracterizado por ser um processo não estacionário, na medida em que a sua variância depende do tempo t (Enders, 2015).

Passeio Aleatório com *Drift*

Um passeio aleatório com *drift* é uma extensão do modelo de passeio aleatório com a adição de um termo constante, a_0, podendo ser formulado por

$$Y_t = a_0 + Y_{t-1} + \epsilon_t, \tag{3.36}$$

em que ϵ_t é o ruído branco. A tendência é descrita por parcelas determinísticas e parcelas estocásticas. Assim, o valor médio depende do tempo t, no caso em que Y_0 é conhecido, tem-se o modelo de passeio aleatório com *drift* formulado por

$$Y_t = Y_0 + a_0 t + \sum_{i=1}^{t} \epsilon_i. \tag{3.37}$$

Ao calcular-se o valor esperado de Y_t, $Y_0 + a_0 t$, é percetível que depende de t e, assim, este processo é não estacionário.

Estacionariedade

A análise prévia da estacionariedade pode ser realizada pela representação gráfica da série, ao longo do tempo. Posteriormente, é necessário utilizar testes estatísticos de forma a realizar um estudo formal. Existem vários testes que permitem a avaliação da estacionariedade da série, nomeadamente, o teste de Dickey-Fuller, o teste de Dickey-Fuller Aumentado (*Augmented Dickey Fuller*), o teste Phillips-Perron e o teste de Kwiatkowski-Phillips-Schmidt-Shin. Os três primeiros testes têm como hipótese testar a presença de uma raiz unitária (não estacionariedade) e na sua não rejeição os testes fornecem informação sobre o número de diferenciações necessárias para atingir a estacionariedade. A hipótese nula do último teste considera que a série temporal é estacionária. Mais detalhes sobre estes testes podem ser consultados em Dickey & Fuller (1979); Said & Dickey (1984); Phillips & Perron (1988); Kwiatkowski *et al.* (1992).

3.3 Metodologia Box-Jenkins

Nesta Secção apresenta-se uma breve súmula sobre as metodologias a adotar no estudo de séries temporais, no âmbito deste estudo. Dar-se-á ênfase aos modelos AR, utilizados, posteriormente, no Capítulo 4.

Em 1970, Box & Jenkins desenvolveram o seu trabalho sobre os modelos SARIMA (*Seasonal Autoregressive Integrated Moving Average*), com o objetivo da modelação e da previsão de séries temporais estacionárias e não estacionárias. Estes modelos descrevem a série Y_t como uma função dos seus valores passados e como combinação linear de uma

sucessão de choques aleatórios. Dentro destes modelos, os mais simples são os Modelos Autorregressivos (AR), os Modelos de Médias Móveis (MA) e os Modelos Autorregressivos de Médias Móveis (ARMA). O primeiro descreve o comportamento da série à custa dos seus valores passados, o segundo através de uma sucessão de choques aleatórios, ao longo do tempo. O modelo ARMA é a combinação dos dois modelos anteriores. Estes modelos são úteis para séries estacionárias.

Todavia, os modelos mencionados, quando aplicados a processos não estacionários, não revelam um bom ajustamento, neste caso é aconselhável recorrer aos Modelos Autorregressivos Integrados de Médias Móveis (ARIMA). Este tipo de modelos são designados por modelos integrados, uma vez que o modelo estacionário, que é ajustado aos dados diferenciados, deve ser somado ou integrado para fornecer um modelo para os dados não estacionários. À semelhança dos modelos ARMA, estes modelos podem ser generalizados para incluir termos sazonais dando origem aos Modelos Autorregressivos Integrados de Médias Móveis Sazonais (SARIMA).

Quando se utiliza este tipo de modelos recorre-se à metodologia Box-Jenkins para a sua seleção. Esta metodologia implica um processo iterativo constituído por três fases: identificação do modelo, estimação dos parâmetros e análise de diagnóstico. A ideia base da identificação do modelo é que se uma série temporal é gerada a partir de um processo SARIMA, então deve ter algumas propriedades teóricas de autocorrelação. Box & Jenkins (1970) propuseram, então, usar a função de autocorrelação (FAC) e a função de autocorrelação parcial (FACP) como ferramentas básicas para identificar as ordens do Modelo Autorregressivo Integrado de Médias Móveis Sazonal (SARIMA). A estimação permite a obtenção dos parâmetros do modelo escolhido (pelo Método de Máxima Verosimilhança e pelo Método dos Mínimos Quadrados) e é feita a sua avaliação na análise diagnóstico.

A fase de diagnóstico engloba duas etapas: a avaliação da qualidade das estimativas obtidas e a avaliação da qualidade do ajustamento do modelo às observações da série em estudo (deve-se proceder à análise dos resíduos que devem ter um comportamento semelhante a um ruído branco). No contexto de dados ambientais existem vários estudos que mostraram que o processo para melhor descrever o comportamento dos resíduos é o processo AR.

3.3.1 Processo Autorregressivo de ordem p, AR(p)

Considerando o processo autorregressivo de ordem p, AR(p), sabe-se que este tem como suporte o facto de que a observação da variável no instante t estar relacionada linearmente com as observações nos instantes anteriores. O processo Y_t diz-se um processo autorregressivo de ordem p, AR(p), quando satisfaz a equação

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \epsilon_t, \tag{3.38}$$

em que ϵ_t é um ruído branco com média nula e independente de Y_{t-k}, $\forall k \geq 1$. A variável Y_t pode ser vista como uma variável dependente que é explicada através de uma regressão linear múltipla, em que as observações em p instantes anteriores funcionam como variáveis explicativas e ϕ_i são os coeficientes de cada Y_{t-i}. Existe outra forma de representação deste processo através do operador atraso

$$\Phi_p(B) Y_t = \epsilon_t, \tag{3.39}$$

em que $\Phi_p(B) = 1 - \phi_1 B - \cdots - \phi_p B^P$ é o polinómio autorregressivo de ordem p. A fatorização deste polinómio é possível através das p raízes, $G_1^{-1}, \ldots, G_p^{-1}$, da característica, $\Phi_p(B) = 0$, tornando-se possível fatorizar o polinómio autorregressivo do seguinte modo

$$\Phi_p(B) Y_t = \prod_{i=1}^{p} (1 - G_i B). \tag{3.40}$$

Se o módulo de cada raíz da equação característica for inferior a um, ou de forma equivalente, $|G_i| < 1$ com $i = 1, \ldots, p$, então é condição necessária e suficiente para que o processo seja estacionário. Garantida esta condição, sabe-se que o processo é invertível, o que significa que a dependência do passado vai sendo menor à medida que o passado se torna mais remoto. Graficamente, a FACP de um processo AR(p) revela uma queda brusca para zero a partir do *lag* $p + 1$ e a FAC apresenta um decaimento exponencial ou sinusoidal amortecido para zero.

3.3.2 Sazonalidade

Os dados ambientais tipicamente apresentam uma forte sazonalidade e esta é possível ser integrada no processo de modelação.

Indicadores Sazonais

Nas séries temporais com sazonalidade, a sazonalidade pode ser modelada através da especificação de um modelo de regressão que inclua uma variável indicatriz para representar cada um dos s períodos sazonais, isto é,

$$Y_t = t\beta + D_1\gamma_1 + \cdots + D_s\gamma_s + \epsilon_t, \qquad t = 1, \ldots, n, \tag{3.41}$$

em que $t\beta$ representa a tendência, $\gamma_1, \ldots, \gamma_s$ são os coeficientes que representam os s efeitos sazonais e D_k são as variáveis indicatrizes, que representam os diferentes períodos

sazonais: tomam o valor 1 quando o tempo t pertence ao período k e 0 nos restantes casos. Por exemplo, para dados mensais, se D_1 corresponder às ocorrências no mês de janeiro (ou seja, 1, se t ocorre em janeiro e 0, caso contrário), então γ_1 só é tido em consideração para observações registadas nesse mês.

Sazonalidade Harmónica

Na integração da sazonalidade no modelo, pode ser considerada uma variável indicatriz por cada período sazonal considerado. Na prática, os efeitos sazonais são refletidos de forma contínua e suave, o que leva a considerar outros tipos de representação da sazonalidade. Cowpertwait & Melcalfe (2009) apresentam uma alternativa através de um modelo sazonal harmónico, recorrendo a funções trigonométricas (seno e cosseno), para incorporar as oscilações observadas. De forma simples, pode-se descrever uma onda sinusoidal por

$$Asen(2\pi ft + \phi) = \alpha_c cos(2\pi ft) + \alpha_s sen(2\pi ft), \tag{3.42}$$

em que f é a frequência dos ciclos, A representa a amplitude, ϕ é a constante de fase, $\alpha_s = Acos(\phi)$ e $\alpha_c = Asen(\phi)$.

O modelo sazonal harmónico pode ser definido por

$$Y_t = t\beta + \sum_{k=1}^{s/2} \left[\alpha_{1k} cos(\frac{2\pi kt}{s}) + \alpha_{2k} sen(\frac{2\pi kt}{s}) \right] + \epsilon_t, \tag{3.43}$$

em que $t\beta$ representa a tendência, α_{1k} e α_{2k} são os parâmetros desconhecidos de interesse, s é o período sazonal ($s = 12$ para dados mensais), k é um índice que varia entre 1 e $s/2$, e t é uma variável codificada que representa o tempo (por exemplo, no caso em estudo, $t = 1, \ldots, 132$ para 132 observações igualmente espaçadas).

Capítulo 4

Modelos

Muitos estudos estatísticos têm como objetivo principal o estudo da relação entre variáveis ou, em particular, a análise da influência que uma ou mais variáveis (explicativas), medidas em indivíduos ou objetos, têm sobre uma variável de interesse, que se denomina por variável resposta (Turkman & Silva, 2000). O modo como o estatístico aborda tal problema é através do modelo de regressão.

4.1 Modelos Lineares Generalizados

O Modelo Linear Normal (MLN) é o mais usado na modelação estatística. Este modelo tem várias limitações: a relação é descrita através de uma função linear; exige a independência das respostas e a variável dependente condicionada aos valores das variáveis explicativas segue a Distribuição Normal, com variância constante (condicionada aos valores das variáveis explicativas). O MLN pode ser expresso pela seguinte equação

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon, \tag{4.1}$$

em que Y é a variável resposta, $X_1, \ldots, X_p$ são as variáveis explicativas, p é o número de variáveis explicativas, $\beta_0, \beta_1, \ldots, \beta_p$ são os parâmetros do modelo e ϵ é o erro aleatório e não observável e assume-se que $E[\epsilon] = 0$ e $Var[\epsilon] = \sigma^2$. Outra possível representação do MLN é

$$\begin{gathered} Y|X \sim N(\mu, \sigma^2), \\ E[Y|X] = \mu = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p. \end{gathered} \tag{4.2}$$

Em dados reais, os pressupostos do MLN são difíceis de verificar e muitas vezes recorre-se a transformações. Box & Cox (1964) propõem uma transformação que tem como objetivo verificar os pressupostos da normalidade, da variância constante e da linearidade.

Esta transformação faz com que a variável resposta seja alterada, podendo até deixar de ser definida no espaço amostral original.

Com a evolução dos modelos, associada ao desenvolvimento computacional, foi possível estabelecer uma extensão do MLN a distribuições não normais, os Modelos Lineares Generalizados (MLG) apresentados por Nelder & Wedderburn (1972). Os MLG tiveram um impacto considerável na evolução da estatística aplicada, mais concretamente, nas últimas duas décadas, existiu um grande avanço que permitiu a acessibilidade e o dinamismo destes modelos. Turkman & Silva (2000) descrevem esta importância: "Do ponto de vista teórico a sua importância advém, essencialmente, do facto de a metodologia destes modelos constituir uma abordagem unificada de muitos procedimentos estatísticos correntemente usados nas aplicações e promover o papel central da verosimilhança na teoria da inferência".

As vantagens principais dos MLG são a possibilidade de admitir várias distribuições para a variável resposta, através da família exponencial de distribuições, e a flexibilidade para a relação funcional entre o valor esperado da variável resposta (μ) e o preditor linear ($\eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$). Algumas das limitações dos MLG são o facto de manterem uma estrutura de linearidade, das distribuições da variável resposta se restringirem à família exponencial e de exigirem a independência das respostas.

4.1.1 Notação e Terminologia

Os indivíduos são considerados as unidades do estudo. O valor da variável resposta, do indivíduo i, denomina-se por y_i e é uma realização da variável resposta (ou variável dependente) Y_i, em que i varia de 1 a n, em que n é o número total dos indivíduos em estudo. A variável resposta, para um indivíduo i, representa-se pelo vetor das variáveis resposta, resultante das medições, isto é,

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \tag{4.3}$$

ou, em alternativa, $\mathbf{Y}^T = (Y_1, Y_2, \ldots, Y_n)$.

Cada variável aleatória Y_i associada a cada i está o vetor das p covariáveis (ou variáveis explicativas ou variáveis independentes) com dimensão $p \times 1$, isto é, $\mathbf{X^T} = (X_1, ..., X_p)$. Assim, para $\mathbf{Y}$, a matriz da variável resposta, de ordem $n \times p$ é dada por

$$\mathbf{X} = \begin{bmatrix} x_{11} & \dots & x_{1p} \\ x_{21} & \dots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{np} \end{bmatrix}, \tag{4.4}$$

associado ao vetor de parâmetros desconhecidos $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)^T$.

4.1.2 Família Exponencial

Nos Modelos Lineares Generalizados, a variável resposta segue uma distribuição que pertence à família exponencial.

Uma variável aleatória $\mathbf{Y}$ tem distribuição pertencente à família exponencial de dispersão (ou simplesmente família exponencial) se a sua função densidade de probabilidade (f.d.p.) ou a sua função massa de probabilidade (f.m.p.) se puder escrever da seguinte forma

$$f(y|\theta, \phi) = exp\left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\}, \tag{4.5}$$

em que θ é o parâmetro de localização e ϕ é o parâmetro de dispersão. $a(.)$, $b(.)$, $c(.)$ são funções reais específicas para cada distribuição. Uma descrição mais pormenorizada desta família pode ser consultada em Cox & Hinkley (1974).

Turkman & Silva (2000) apontam que quando ϕ for conhecido, tem-se uma distribuição da família exponencial com parâmetro canónico θ. No caso de ϕ ser desconhecido, a distribuição pode ou não fazer parte da família exponencial, a função $b(.)$ é diferenciável e que o suporte da distribuição não depende dos parâmetros. Habitualmente, tem-se que $a(\phi) = \dfrac{\phi}{w}$, em que w é uma constante conhecida e revela o peso da observação.

Considerando $\ell(\theta, \phi, y) = log(f(y|\theta, \phi))$, em que ℓ representa o logaritmo da função de verosimilhança, define-se por função *Score*

$$S(\theta) = \frac{\partial \ell(\theta, \phi, y)}{\partial \theta}. \tag{4.6}$$

Sabe-se que no caso das famílias regulares, tem-se que

$$E[S(\theta)] = 0, \tag{4.7}$$

$$E[S^2(\theta)] = E\left[\left(\frac{\partial \ell(\theta, \phi, y)}{\partial \theta} \right)^2 \right], \tag{4.8}$$

e obtém-se

$$E[Y] = \mu = a(\phi)E[S(\theta)] + b'(\theta) = b'(\theta) \tag{4.9}$$

e

$$Var[Y] = a^2(\phi)Var[S(\theta)] = a^2(\phi)\frac{b''(\theta)}{a(\phi)} = a(\phi)b''(\theta). \tag{4.10}$$

Assim, a variância de Y é o produto de duas funções, $b''(\theta)$, que depende do parâmetro canónico θ (e consequentemente do valor médio μ), designada por função de variância, $V(\mu) = \frac{d\mu}{d\theta}$, e $a(\phi)$ que depende do parâmetro de dispersão ϕ.

A função de variância, $V(\mu)$, desempenha um papel preponderante na família exponencial porque é responsável pela caracterização da distribuição. Ou seja, cada distribuição pertencente à família exponencial apresenta uma e só uma função de variância e vice-versa (unicidade). Por exemplo, a função de variância tal que $V(\mu) = \mu(1-\mu)$, em que $0 < \mu < 1$, caracteriza a classe de distribuições binomiais, com probabilidades de sucesso μ (Turkman & Silva, 2000).

4.1.3 Formulação do Modelo

O Modelo Linear Generalizado pode ser expresso como

$$Y = Z\beta + \epsilon, \tag{4.11}$$

em que Z é uma matriz de covariáveis, de dimensão $n \times (p+1)$ (em geral igual à matriz de covariáveis $\mathbf{X}$ com um primeiro vetor unitário), $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ é o vetor de parâmetros, e ϵ é um vetor de erros aleatórios com distribuição $N(0, \sigma^2 I)$. O valor esperado da variável resposta é uma função linear das covariáveis, isto é,

$$E(Y|Z) = \mu = Z\beta. \tag{4.12}$$

Um MLG é definido por uma distribuição de probabilidade, membro da família exponencial (para a variável resposta), um conjunto de variáveis explicativas descrevendo a estrutura linear do modelo e uma função de ligação diferenciável que relaciona o valor esperado da variável resposta e a estrutura linear. O MLG assenta em três componentes fundamentais:

1) Componente aleatória

   Dado o vetor de covariáveis $\mathbf{Z}_i$ as variáveis aleatórias Y_i são condicionalmente independentes com distribuição pertencente à família exponencial, e portanto o seu

valor médio é dado por

$$E[Y_i|\mathbf{z}_i] = \mu_i = b'(\theta_i) \quad i = 1, \ldots, n; \tag{4.13}$$

2) Componente estrutural ou sistemática

Define-se o preditor linear η_i como combinação linear das variáveis explicativas, dado por

$$\eta_i = \mathbf{z}_i^T \beta; \tag{4.14}$$

associado a cada valor da variável resposta, Y_i, tem-se o vetor $(p+1)\times 1$ de covariáveis $\mathbf{z}_i^T = (z_{i1}, \ldots, z_{ip})$, $i = 1, \ldots, n$, em que $\mathbf{z}_{ik}$, $k = 1 \ldots, p$, representa a k-ésima covariável para o i-ésimo indivíduo e $\beta = (\beta_0, \ldots, \beta_p)^T$ um vetor $(p+1) \times 1$ de parâmetros desconhecidos.

3) Função de ligação

A função de ligação para relacionar o valor esperado de Y_i com o preditor linerar η_i é uma função $g(.)$ tal que $g(\mu_i) = \eta_i$.

$$\begin{aligned} & g(\mu_i) = \eta_i \Leftrightarrow \\ & \Leftrightarrow g(\mu_i) = z_i^T \beta \\ & \Leftrightarrow g(\mu_i) = \sum_{i=1}^{p} z_{ij}\beta_j \\ & \Leftrightarrow \mu_i = g^{-1}(z_i^T \beta). \end{aligned} \tag{4.15}$$

Agresti (1990) define MLG como um modelo linear para uma transformação do valor esperado duma variável aleatória cuja distribuição pertence à família exponencial. Comparativamente ao Modelo Linear, os erros aleatórios não são adicionados explicitamente à equação, mas como uma flutuação aleatória da variável resposta através da distribuição de probabilidades.

Quando o preditor linear coincide com o parâmetro canónico, ou seja,

$$\theta_i = \mu_i = z_i^T \beta, \tag{4.16}$$

a função de ligação denomina-se por função de ligação canónica. A utilização das ligações canónicas garantem a concavidade da função de verosimilhança, o que leva mais facilmente aos resultados assintóticos.

A metodologia adotada para a construção de um MLG segue várias etapas. A primeira, a formulação dos modelos, engloba a escolha da distribuição para a variável resposta, das covariáveis e da função de ligação. A segunda etapa, o ajustamento do modelo, permite a estimação dos parâmetros e dos respetivos intervalos de confiança, testes de hipóteses e seleção de covariáveis. A terceira etapa, a seleção e validação dos modelos, averigua as discrepâncias entre os valores observados e os valores preditos.

4.1.4 Estimação

Existem muitos métodos para a estimação dos parâmetros de um MLG, nomeadamente através da verosimilhança. O Método da Máxima Verosimilhança é o método mais usado na estimação dos parâmetros de regressão. A vantagem da utilização dos métodos baseados na verosimilhança é o facto de fornecerem estimadores consistentes, assintoticamente eficientes, com distribuição assintoticamente Normal. As estimativas através deste método são obtidas pela maximização do logaritmo da função verosimilhança, que podem ser resultantes algebricamente, em casos simples, ou através de otimização numérica, em casos mais complexos.

De notar que nem sempre se recorre ao Método da Máxima Verosimilhança, por exemplo, se existir o parâmetro de dispersão ϕ a sua estimação é feita pelo Método dos Momentos.

Considerando que os dados são da forma (y_i, x_i), com $i = 1, \ldots, n$, em que y_i é o valor observado da variável resposta e x_i é o correspondente vetor de covariáveis. Designa-se por $z_i = z_i(x)$ o vetor de especificação, com dimensão $p+1$, associado ao vetor de covariáveis, e é tal que $z_i = (1, x_{i1}, \ldots, x_{ip})^T$. Para simplificação de notação admite-se que a matriz Z é dada por

$$Z = (z_1, z_2, \ldots, z_n)^T = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1(p+1)} \\ z_{21} & z_{22} & \cdots & z_{2(p+1)} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{n(p+1)} \end{bmatrix}, \tag{4.17}$$

e tem característica igual à ordem $p+1$ (característica completa) e, consequentemente, $Z^T Z$ também tem característica $p+1$.

O MLG é definido por

$$f(y_i|\theta_i, \phi, w_i) = exp\Big\{ \frac{w_i}{\phi}\big(y_i\theta_i - b(\theta_i)\big) + c(y_i, \phi, w_i)\Big\}, \quad i = 1, \ldots, n, \tag{4.18}$$

em que y_i é uma observação do indivíduo i, com função ligação dada por $g(\mu_i) = \eta_i = z_i^T \boldsymbol{\beta}$

e y_i são variáveis aleatórias independentes. A função de verosimilhança pode ser escrita como função de $\boldsymbol{\beta}$,

$$\begin{aligned} L(\boldsymbol{\beta}) &= \prod_{i=1}^{n} f(y_i|\theta_i, \phi, w_i) \\ &= \prod_{i=1}^{n} exp\Big\{ \frac{w_i}{\phi}\big(y_i\theta_i - b(\theta_i)\big) + c(y_i, \phi, w_i)\Big\} \\ &= exp\Big\{ \frac{1}{\phi}\sum_{i=1}^{n} w_i\big(y_i\theta_i - b(\theta_i)\big) + \sum_{i=1}^{n} c(y_i, \phi, w_i)\Big\}. \end{aligned} \tag{4.19}$$

Aplicando o logaritmo à função de verosimilhança, obtém-se a log-verosimilhança (*loglik*), dada por

$$\begin{aligned} \ell(\boldsymbol{\beta}) &= logL(\boldsymbol{\beta}) \\ &= \sum_{i=1}^{n} \frac{w_i\big(y_i\theta_i - b(\theta_i)\big)}{\phi} + c(y_i, \phi, w_i) \\ &= \sum_{i=1}^{n} \ell_i(\boldsymbol{\beta}), \end{aligned} \tag{4.20}$$

em que ℓ_i é a contribuição de cada observação y_i para a verosimilhança e é expressa por

$$\ell_i(\boldsymbol{\beta}) = \frac{w_i\big(y_i\theta_i - b(\theta_i)\big)}{\phi} + c(y_i, \phi, w_i). \tag{4.21}$$

Os estimadores de máxima verosimilhança dos parâmetros $\boldsymbol{\beta}$ são obtidos da solução de um sistema de equações de verosimilhança. Esta solução determina-se derivando a expressão (4.21), em ordem a cada parâmetro β_j, com $j = 0, \ldots, p$, e igualando a zero. Ou seja,

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ell_i(\boldsymbol{\beta})}{\partial \beta_j} = 0, \quad j = 0, \ldots, p. \tag{4.22}$$

Aplicando a regra de cadeia, obtém-se que

$$\frac{\partial \ell_i(\boldsymbol{\beta})}{\partial \beta_j} = \frac{\partial \ell_i(\theta_i)}{\partial \theta_i}\frac{\partial \theta_i(\mu_i)}{\partial \mu_i}\frac{\partial \mu_i(\eta_i)}{\partial \eta_i}\frac{\partial \eta_i(\boldsymbol{\beta})}{\partial \beta_j}, \quad j = 0, \ldots, p, \tag{4.23}$$

em que

$$\frac{\partial \ell_i(\theta_i)}{\partial \theta_i} = \frac{w_i(y_i - b'(\theta_i))}{\phi} = \frac{w_i(y_i - \mu_i)}{\phi}, \tag{4.24}$$

$$\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) = \frac{w_i(y_i - \mu_i)}{\phi} \tag{4.25}$$

e

$$\frac{\partial \eta_i(\boldsymbol{\beta})}{\partial \beta_j} = z_{ij}. \tag{4.26}$$

Assim, a equação (4.23) pode ser escrita como

$$\begin{aligned} \frac{\partial \ell_i(\boldsymbol{\beta})}{\partial \beta_j} &= \frac{w_i(y_i - \mu_i)}{\phi} \frac{\phi}{w_i Var(Y_i)} \frac{\partial \mu_i(\eta_i)}{\partial \eta_i} z_{ij} \\ \frac{\partial \ell_i(\boldsymbol{\beta})}{\partial \beta_j} &= \frac{(y_i - \mu_i) z_{ij}}{Var(Y_i)} \frac{\partial \mu_i(\eta_i)}{\partial \eta_i}, \qquad \begin{array}{l} i = 1, \ldots, n, \\ j = 0, \ldots, p. \end{array} \end{aligned} \tag{4.27}$$

As equações de verosimilhança para $\boldsymbol{\beta}$ são determinadas pelo somatório das n observações, igualando a zero, dadas por

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i) z_{ij}}{Var(Y_i)} \frac{\partial \mu_i(\eta_i)}{\partial \eta_i} = 0, \qquad j = 0, \ldots, p. \tag{4.28}$$

Em muitas situações, as equações de verosimilhança para $\boldsymbol{\beta}$ não são lineares e é necessário recorrer a Métodos Numéricos, com processos iterativos, como é o caso do Método Newton-Raphson. A solução pode não ser necessariamente o máximo global da função $\ell(\boldsymbol{\beta})$. Porém, em alguns modelos, a função $\ell(\boldsymbol{\beta})$ é côncava, o que permite que os seus máximos global e local sejam coincidentes.

A resolução das equações de verosimilhança por Métodos Numéricos utiliza o Método de *Scores* de Fisher. No caso da função de ligação ser canónica, este método é igual ao Método de Newton-Raphson. A diferença do Método de *Scores* de Fisher e o Método de Newton-Raphson é o facto do primeiro resolver sistemas de equações não lineares a partir da matriz de informação de Fisher, $I(\boldsymbol{\beta})$,

$$I(\boldsymbol{\beta}) = E\Big[- \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \Big], \tag{4.29}$$

ao invés da matriz Hessiana, constituída pelas derivadas parciais de segunda ordem de $\ell(\boldsymbol{\beta})$, dada por

$$H(\boldsymbol{\beta}) = \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}. \tag{4.30}$$

Esta diferença prende-se com o facto de ser mais fácil o cálculo da matriz de informação de Fisher, $I(\boldsymbol{\beta})$, e também por esta matriz ser sempre semi-definida positiva.

Para se obter a matriz de informação de Fisher, inicia-se com a determinação da função *Score*, $S(\boldsymbol{\beta})$, que é o vetor p-dimensional, constituído pelas derivadas parciais de primeira ordem da função logaritmo da função verosimilhança, e dada por

$$S(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} s_i(\boldsymbol{\beta}), \tag{4.31}$$

em que $s_i(\boldsymbol{\beta})$ é o vetor de componentes $\dfrac{\partial \ell_j(\boldsymbol{\beta})}{\partial \beta_j}$. O elemento genérico de ordem j da função *Score* é

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i) z_{ij}}{var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i}. \tag{4.32}$$

Assim, para obter a matriz de informação de Fisher, considera-se o valor esperado das segundas derivadas de $\ell_i(\boldsymbol{\beta})$, com $i = 1, \ldots, n$, e pode-se escrever como

$$\begin{aligned} E\Big[-\frac{\partial^2 \ell_i(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k}\Big] &= -E\Big[\frac{\partial^2 \ell_i(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k}\Big] \\ &= E\Big[\frac{\partial \ell_i(\boldsymbol{\beta})}{\partial \beta_j} \frac{\partial \ell_i(\boldsymbol{\beta})}{\partial \beta_k}\Big] \\ &= \frac{z_{ij} z_{ik}}{var(Y_i)} \Big(\frac{\partial \mu_i}{\partial \eta_i}\Big), \quad j, k = 0, \ldots, p \end{aligned} \tag{4.33}$$

e, portanto, o elemento genérico de ordem (j, k) da matriz de informação de Fisher é dado por

$$-\sum_{i=1}^{n} E\Big[\frac{\partial^2 \ell_i}{\partial \beta_j \partial \beta_k}\Big] = \sum_{i=1}^{n} \frac{z_{ij} z_{ik}}{var(Y_i)} \Big(\frac{\partial \mu_i}{\partial \mu_i}\Big)^2. \tag{4.34}$$

Na forma matricial tem-se que

$$I(\boldsymbol{\beta}) = Z^T W Z, \tag{4.35}$$

em que W é a matriz diagonal de ordem n. O i-ésimo elemento da matriz W é dado por

$$W_i = \frac{\Big(\dfrac{\partial \mu_i}{\partial \eta_i}\Big)^2}{var(Y_i)} = \frac{w_i \Big(\dfrac{\partial \mu_i}{\partial \eta_i}\Big)^2}{\phi V(\mu_i)}. \tag{4.36}$$

Parâmetros do Modelo

Como anteriormente foi mencionado, os estimadores de máxima verosimilhança dos parâmetros do modelo $\boldsymbol{\beta}$ são determinados a partir da solução das equações de verosimilhança.

O cálculo das estimativas de Máxima Verosimilhança de $\boldsymbol{\beta}$ é dado por um processo iterativo de duas etapas:

1. Dado $\hat{\boldsymbol{\beta}}^{(k)}$ (com $k = 0$) determina-se $\mathbf{u}_i^{(k)}$, que é um vetor com elemento genérico

$$u_i^{(k)} = \eta_i^{(k)} + (y_i - \mu_i^{(k)})\frac{\partial \eta_i(k)}{\partial \mu_i^{(k)}} \tag{4.37}$$

   e calcula-se $W^{(k)}$ utilizando a equação (4.36);

2. A nova iteração $\boldsymbol{\beta}^{(k+1)}$ é calculada usando

$$\hat{\boldsymbol{\beta}}^{(k+1)} = (Z^T W^{(k)} Z)^{-1} Z^T W^{(k)} \mathbf{u}^{(k)}. \tag{4.38}$$

   O processo iterativo termina quando é atingido o seguinte critério

$$\frac{||\hat{\boldsymbol{\beta}}^{(k+1)} - \hat{\boldsymbol{\beta}}^{(k)}||}{||\hat{\boldsymbol{\beta}}^{(k)}||} \leq \epsilon, \tag{4.39}$$

   para algum valor $\epsilon > 0$ previamente definido.

Salienta-se que a equação (4.37) é idêntica à que se obteria para os estimadores de Método dos Mínimos Quadrados Ponderados se, em cada iteração, se calculasse a regressão linear de $\mathbf{u}^{(k)}$ em vez de Z, em que $W^{(k)}$ é uma matriz de pesos.

Na expressão (4.38), a matriz W contém o parâmetro de dispersão ϕ, mas este último não faz parte dos cálculos de $\boldsymbol{\beta}^{(k+1)}$ e torna-se irrelevante para a determinação deste parâmetro. Por este motivo, assumindo $\phi = 1$ não há perda de generalidade, na determinação das estimativas para $\boldsymbol{\beta}$.

Parâmetro de Dispersão

Para a estimação do parâmetro de dispersão é possível recorrer ao Método de Máxima Verosimilhança. Mas existem métodos mais simples, que também apresentam bons resultados, como o método que assenta na distribuição de amostragem para grandes valores de n, da Estatística de Pearson Generalizada,

$$\hat{\phi} = \frac{1}{n-p}\sum_{i=1}^{n}\frac{w_i(y_i - \hat{\mu}_i)^2}{V(\mu_i)}, \tag{4.40}$$

em que $\hat{\phi}$ é um estimador consistente de ϕ. Quando a amostra é muito grande (elevado valor n), esta estatística pode ser aproximada à distribuição Qui-quadrado, com $n - p$ graus de liberdade.

Propriedades dos Estimadores de Máxima Verosimilhança

O estimador máxima verosimilhança, o $\hat{\boldsymbol{\beta}}$, é um estimador assintoticamente centrado, ou seja,

$$E[\hat{\boldsymbol{\beta}}] \simeq \boldsymbol{\beta}. \tag{4.41}$$

A matriz de covariância de $\hat{\boldsymbol{\beta}}$ é aproximadamente igual ao inverso da matriz de informação de Fisher, isto é,

$$cov(\hat{\boldsymbol{\beta}}) \simeq E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T] = I^{-1}(\boldsymbol{\beta}). \tag{4.42}$$

Além disso, a distribuição assintótica de $\hat{\boldsymbol{\beta}}$ é normal $(p+1)$-variada, com valor esperado $\boldsymbol{\beta}$ e matriz de covariância $I^{-1}(\boldsymbol{\beta})$,

$$\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, I^{-1}(\boldsymbol{\beta})). \tag{4.43}$$

A distribuição assintótica da estatítica de Wald $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T I(\boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ tem uma distribuição assintótica de um χ^2, com $p + 1$ graus de liberdade. Estes resultados são importantes para fazer inferência sobre $\boldsymbol{\beta}$, nomeadamente, nos testes de hipóteses e intervalos de confiança.

4.1.5 Inferência

A escolha das variáveis apropriadas é uma etapa extremamente importante na modelação estatística. A análise da relação entre a variável resposta e as covariáveis permite ter uma ideia das variáveis que serão importantes para um modelo final.

Testes de Hipóteses

Os testes de hipóteses são procedimentos de validação estatística que possibilitam avaliar hipóteses sobre determinadas características da população, sujeitos a um determinado nível de risco de falharem associado (nível de confiança). Os testes são divididos em paramétricos ou não paramétricos. Os testes paramétricos baseiam-se em parâmetros ou características quantitativas da variável dependente, sob o pressuposto de que as variáveis seguem uma Distribuição Normal. Os testes não paramétricos são utilizados quando as variáveis são ordinais ou categóricas e os pressupostos paramétricos não se verifiquem. Estes testes não são tão potentes comparativamente com os testes paramétricos.

Teste de Wald

O Teste de Wald é apropriado para testar uma componente do vetor parâmetro, com as hipóteses de teste

$$\begin{aligned} H_0 &: C\boldsymbol{\beta} = \xi \\ H_1 &: C\boldsymbol{\beta} \neq \xi, \end{aligned} \tag{4.44}$$

em que C é uma matriz $q \times p$ de característica completa q, o vetor ξ tem dimensão q e é especificado pelo investigador. A estatística de teste (ET), denominada por Estatística de Wald, é baseada na normalidade assintótica do estimador de MV, é dada por

$$W = (C\hat{\boldsymbol{\beta}} - \xi)^T [CI^{-1}(\hat{\boldsymbol{\beta}})C^T]^{-1}(C\hat{\boldsymbol{\beta}} - \xi) \tag{4.45}$$

e tem distribuição assintótica χ^2, com q graus de liberdade. No caso de $ET > \chi^2_{1-\alpha,q}$, em que α é o nível de significância, então a hipótese nula é rejeitada.

Teste da Razão de Verosimilhança

O Teste da Razão de Verosimilhança é apropriado para testar

$$\begin{aligned} H_0 &: C\boldsymbol{\beta} = \xi \\ H_1 &: C\boldsymbol{\beta} \neq \xi, \end{aligned} \tag{4.46}$$

em que C é uma matriz $q \times p$ de característica completa q, o vetor ξ tem dimensão q e é especificado pelo investigador. A estatística de teste (ET), denominada por estatística de Wilks ou estatística da Razão de Verosimilhanças, é baseada na distribuição assintótica da razão do máximo das verosimilhanças sob as hipóteses H_0 e $H_0 \cup H_1$, é dada por

$$ET = -2\big[\ell(\tilde{\boldsymbol{\beta}}) - \ell(\hat{\boldsymbol{\beta}})\big] \sim \chi^2_q. \tag{4.47}$$

No caso de $ET > \chi^2_{1-\alpha,q}$, em que α é o nível de significância, então a hipótese nula é rejeitada.

De notar que também existem outras estatísticas que não são mencionadas neste trabalho, como a estatística de Rao (ou estatística *Score*), baseada nas propriedades assintóticas da função *Score*.

Seleção de Variáveis

Existem três procedimentos habitualmente utilizados na seleção de variáveis: *backward elimination*, *forward selection* e *stepwise selection*. Estes têm como base um algoritmo que

permite perceber se a variável é importante ou não para o modelo em estudo, recorrendo à significância estatística do coeficiente da variável em análise.

No procedimento *backward elimination*, ou processo de seleção regressiva, começa-se com um modelo que inclui todas as covariáveis, modelo completo. Com base no nível de confiança do teste estatístico, elimina-se a variável menos significativa. Ajusta-se novamente o modelo, excluindo a variável definida, e repete-se o procedimento até existirem no modelo covariáveis todas significativas.

O procedimento *forward selection*, ou processo de seleção progressiva, é o inverso do processo anterior. Inicia-se com o modelo sem covariáveis e adiciona-se a variável com o menor valor de prova. Repete-se o processo até não existirem mais covariáveis significativas.

O procedimento *stepwise selection* é a combinação das duas metodologias anteriores. Uma determinada covariável pode ser adicionada ou removida, a cada passo testa-se se as restantes são significativas para verificar se deve ser removida.

4.1.6 Qualidade de Ajustamento

Após a elaboração de um modelo, é necessário verificar a qualidade do seu ajustamento. Para avaliar essa qualidade, pode-se recorrer à função desvio e à Estatística de Pearson Generalizada.

Função Desvio

Como já foi mencionado, o logaritmo da função de verosimilhança de um MLG é dado por

$$\ell(\boldsymbol{\beta}) = logL(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{w_i\big(y_i\theta_i - b(\theta_i)\big)}{\phi} + c(y_i, \phi, w_i). \tag{4.48}$$

Para se comparar o modelo em estudo com o modelo completo (ou saturado), recorre-se à estatística de razão de verosimilhanças, obtendo-se

$$D^*(y;\hat{\mu}) = -2\big(\ell_M(\hat{\boldsymbol{\beta}}_M) - \ell_S(\hat{\boldsymbol{\beta}}_S)\big) = \frac{D(y;\hat{\mu})}{\phi}, \tag{4.49}$$

em que $D^*(y;\hat{\mu})$ designa-se por desvio reduzido, $\hat{\boldsymbol{\beta}}_M$ é o vetor de parâmetros, para o modelo em investigação, $\hat{\boldsymbol{\beta}}_S$ é o vetor de parâmetros, para o modelo saturado (modelo com todas as covariáveis em estudo), e $D(y;\hat{\mu})$ é o desvio para o modelo em estudo e é dado por

$$D(y;\hat{\mu}) = \sum_{i=1}^{n} 2w_i \Big\{ y_i\big(q(y_i) - q(\hat{\mu}_i)\big) - \big(q(y_i)\big) + b\big(q(\hat{\mu}_i)\big) \Big\} = \sum_{i=1}^{n} d_i, \tag{4.50}$$

em que d_i é a diferença entre os logaritmos das funções de verosimilhança da observada e ajustada em cada observação.

A estatística da razão de verosimilhança para comparar dois modelos, $M1$ e $M2$, é dada por

$$\frac{D(y;\hat{\mu}_1) - D(y;\hat{\mu}_2)}{\phi} \sim \chi^2_{p_1-p_2}, \tag{4.51}$$

em que p_1 e p_2 é a dimensão do vetor dos parâmetros $\boldsymbol{\beta}$, para os Modelos 1 e 2, respetivamente. A comparação de modelos encaixados (modelos em que um é submodelo do outro) pode ser realizada através da diferença dos desvios.

Estatística de Pearson Generalizada

Como anteriormente visto, a Estatística de Pearson Generalizada é dada por

$$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{w_i(y_i - \hat{\mu}_i)^2}{V(\mu_i)}, \tag{4.52}$$

em que $\hat{\phi}$ é um estimador consistente de ϕ. Quando a amostra é bastante grande (elevado valor n), esta estatística pode ser aproximada à distribuição Qui-quadrado, com $n-p$ graus de liberdade.

Seleção e Comparação de Modelos

A seleção e a comparação de modelos permite que se encontre o modelo mais parcimonioso, ou seja, o modelo que explique o comportamento da variável resposta, com o mínimo de parâmetros possíveis.

A comparação da qualidade do ajustamento de dois modelos aninhados é geralmente realizada a partir de testes de hipóteses, como o Teste da Razão de Verosimilhança. O Teste da Razão de Verosimilhança é apropriado para testar dois modelos, M_p e M_q, com p e q número de variáveis respetivamente, desde que os mesmos sejam modelos aninhados, com as hipóteses de teste

$$\begin{array}{ll} H_0: & \text{As } q-p \text{ variáveis no modelo não são significativas} \\ H_1: & \text{As } q-p \text{ variáveis no modelo não são significativas} \end{array}, \tag{4.53}$$

sob a hipótese nula, a estatística de teste e distribuição respetiva é

$$-2log\Big[\frac{L_{M_p}(\beta)}{L_{M_q}(\beta)}\Big] \sim \chi^2_{q-p}, \tag{4.54}$$

em que $\big(L_{M_p}(\beta)\big)$ é a função verosimilhança do modelo M_p e $\big(L_{M_q}(\beta)\big)$ é a função verosimilhança do modelo M_q.

Na maioria dos casos práticos, existem várias variáveis que são potenciais candidatas para explicar a variabilidade da variável resposta. Tal tem como consequência a possibilidade de diversos modelos com combinações diferentes das covariáveis para explicar o fenómeno em estudo, o que leva a que o processo da seleção seja difícil e moroso. Usualmente recorre-se ao método de seleção *stepwise* porque facilita a seleção dos modelos.

Quando se pretende comparar modelos não encaixados (isto é, no caso em que o Teste da Razão de Verosimilhança não é aplicável) usam-se os critérios de informação. Os critérios mais utilizados para a seleção de modelos são o Critério de Informação de Akaike (AIC) e o Critério Bayesiano de Schwarz (BIC). Nestes critérios, a comparação dos modelos é feita a partir da maximização do logaritmo da verosimilhança, acrescentando uma penalidade ao número de parâmetros de modelo. Neste sentido, o modelo selecionado é o modelo de menor valor do critério de informação.

Akaike (1974) propõe o Critério de Informação de Akaike que tem a seguinte fórmula

$$AIC = -2\ell(\hat{\boldsymbol{\beta}}, \hat{\alpha}) + 2n_{par}, \tag{4.55}$$

em que ℓ é a função logaritmo de verosimilhança, p é o número de parâmetros, n_{par} é o número de parâmetros do modelo em estudo.

Schwarz (1978) propõe o Critério de Informação Bayesiano, estabelecendo que

$$BIC = -2\ell(\hat{\boldsymbol{\beta}}, \hat{\alpha}) + 2n_{par}ln(N), \tag{4.56}$$

em que n_{npar} é o número de parâmetros do modelo em estudo e N é o número total de observações.

4.1.7 Análise Diagnóstico

A análise de diagnóstico tem como objetivo averiguar a existência de desvios isolados do modelo, ou seja, a existência de uma ou mais observações mal ajustadas, não tendo o mesmo comportamento que as restantes observações. Os desvios sistemáticos podem ser provocados pela seleção inadequada da função de variância, da função de ligação e da matriz do modelo, ou pela definição da escala da variável resposta ou das covariáveis. As discrepâncias isoladas podem ser resultantes dos pontos estarem nos extremos da amplitude de validade da covariável ou dos pontos estarem errados, devido a uma leitura

errada ou a uma transcrição mal realizada ou fatores não controlados.

As metodologias de análise de resíduos do MLG são semelhantes às dos Modelos Lineares, sofrendo algumas modificações. A variância residual é alterada por uma estimativa consistente do parâmetro ϕ. A análise da variância requer um cuidado especial, pois deve ser adequada à distribuição em estudo. O seu comportamento deve ser constante e unitário.

O conceito de resíduos, $r_i = \mathbf{Y}_i - \hat{\mathbf{Y}}_i$, do Modelo Linear tem que ser adaptado à variável dependente ajustada e ao preditor linear. Um resíduo tem como objetivo descrever a discrepância entre o valor observado e o valor ajustado pelo modelo e pode ser calculado de várias formas.

Os resíduos ordinários são definidos por

$$r_i = y_i - \hat{\mu}_i. \tag{4.57}$$

Os resíduos de Pearson determinam-se por

$$r_{pi} = \frac{(y_i - \hat{\mu}_i)\sqrt{w_i}}{\sqrt{V(\hat{\mu}_i)}}. \tag{4.58}$$

Os resíduos de Pearson padronizados calculam-se por

$$r_{pi}^P = \frac{r_{pi}}{\sqrt{\hat{\phi}(1 - h_{ii})}}, \tag{4.59}$$

em que h_{ii} é o i-ésimo elemento da diagonal principal da matriz H, dada por

$$H = W^{1/2} Z (Z^T W Z)^{-1} Z^T W^{1/2}. \tag{4.60}$$

A matriz H depende das variáveis explicativas, da função de ligação e da função de variância, tornando mais difícil a interpretação da medida de alavanca.

Os resíduos *deviance* (desvio residual) são determinados por

$$r_{Di} = sinal(y_i - \hat{\mu}_i)\sqrt{d_i}, \tag{4.61}$$

em que $d(y, \hat{\mu}) = \sum_{i=1}^{n} d_i$. Os resíduos *deviance* padronizados são dados por

$$r_{Di}^P = \frac{r_{Di}}{\sqrt{\hat{\phi}(1 - h_{ii})}}. \tag{4.62}$$

Se os resíduos de Pearson e *deviance* revelam a variância aproximadamente constante, então é indício que o modelo é adequado.

Turkman & Silva (2000) sugerem a representação gráfica dos resíduos padronizados *vs* preditores lineares ou alguma transformação adequada dos preditores lineares. Algumas transformações são apresentadas por McCullagh & Nelder (1989). Se não existirem anomalias, os resíduos devem predispor-se em torno de zero, sem ordem, para os diferentes valores de $\hat{\mu}$. Se existirem anomalias, estas podem dever-se à escolha errada da função de ligação, à escolha errada da escala de uma ou mais covariáveis ou à omissão de um termo quadrático.

Os gráficos adequados são as representações dos resíduos padronizados *vs* preditores lineares ou função dos valores ajustados $\hat{\mu}$ ou índices. Nesta avaliação deve-se encontrar um padrão nulo, os resíduos devem estar dispostos em torno de zero com uma amplitude constante para diferentes valores de $\hat{\mu}$. Para avaliar a presença de correlação entre observações deve-se recorrer a gráficos dos resíduos do modelo *vs* a ordem da observação.

4.2 Modelo de Efeitos Mistos

A evolução da ciência dos dados permitiu o aumento da complexidade na análise estatísticas e das metodologias estatística (Kass *et al.*, 2016).

Quando cada indivíduo apresenta medidas repetidas ao longo do tempo duma determinada característica, sendo o próprio tempo um fator de interesse, diz-se que se tem dados longitudinais (Molenberghs & Verbeke, 2005). A sua obtenção poderá ser prospetiva (os registos são obtidos através dos indíviduos, em estudo, ao longo do tempo) ou retrospetiva, através de um histórico de onde são extraídas as medições dos indivíduos em estudo (Diggle *et al.*, 2002).

Os estudos longitudinais analisam medições repetidas ao longo do tempo sobre o mesmo indivíduo, que permite separar o que no contexto de um estudo populacional se chama o efeito de coorte do efeito da idade (Diggle *et al.*, 2002).

Outra particularidade dos dados longitudinais é o facto de serem agrupados, ou seja, cada grupo é o resultado das medições repetidas de um determinado indivíduo, ao longo do tempo. Estas medições têm uma ordem cronológica e, consequentemente, existe uma correlação entre as observações de um determinado indivíduo. Cada indivíduo tem um vetor resposta com todas as observações ao longo do tempo que usualmente estão correlacionadas, o que leva a que a estrutura de autocorrelação tenha um papel fulcral na modelação e estimação de cada parâmetro do modelo (Diggle *et al.*, 2002).

Um estudo longitudinal tem como objetivo caracterizar a alteração da variável resposta, ao longo do tempo, assim como a relação entre as covariáveis e a variável resposta. Consideram-se alguns fatores que fazem com que esta análise estatística seja complexa, nomeadamente, a estrutura de autocorrelação revela uma especial importância no ajusta-

mento, a existência de variabilidade entre indivíduos distintos, o número de observações de cada indivíduo pode ser diferente e as covariáveis também se podem modificar ao longo do tempo. A principal vantagem é a utilização da totalidade dos dados que leva a evidenciar alterações dentro do mesmo indivíduo, o aumento da potência estatística (é possível distinguir os erros de medição dos erros aleatórios) e a redução do enviesamento.

Existem vários tipos de modelos para a análise de dados longitudinais, nomeadamente o Modelo Marginal e o Modelo de Efeitos Aleatórios. O Modelo Marginal (*Population-average*) tem como objetivo inferir sobre o valor médio populacional. Num Modelo Marginal, o valor esperado marginal é modelado como função das covariáveis. Como as medições são repetidas, em cada indivíduo, não têm tendência a ser independentes, a análise marginal tem que incluir pressupostos em relação à correlação. O Modelo Marginal tem a vantagem do valor esperado da variável resposta e a covariância serem modelados separadamente (Diggle *et al.*, 2002)

Os Modelos de Efeitos Aleatórios (*Subject-specific*) permitem descrever as alterações da resposta média da variável resposta de cada indivíduo e a relação destas com as covariáveis possibilitam realizar inferências sobre o indivíduo, a modelação da sobredispersão e da correlação intrínseca a cada indivíduo, através da incorporação dos efeitos aleatórios. Estes modelos podem ser aplicados a variáveis resposta Distribuição Normal ou não.

Os efeitos aleatórios incorporam a heterogeneidade entre indivíduos e são representados por variáveis aleatórias que usualmente seguem a Distribuição Normal (Diggle *et al.*, 1994).

4.2.1 Terminologia

Consideram-se algumas notações e conceitos básicos utilizados no presente Capítulo. Os indivíduos são considerados unidades do estudo. As observações no mesmo indivíduo i são medições, ao longo do tempo, em que i varia de 1 a n. O valor da variável resposta, do indivíduo i no tempo t, denomina-se por y_{it} e é uma realização da variável resposta Y_{it}, em que t varia de 1 a m_i. A variável resposta de um indivíduo i representa-se pelo vetor das variáveis resposta, resultante das medições, isto é,

$$\mathbf{Y_i} = \begin{bmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{im_i} \end{bmatrix} \tag{4.63}$$

ou, em alternativa, $\mathbf{Y}_i^T = [Y_{i1}, Y_{i2}, \ldots, Y_{im_i}]$, sendo as medições obtidas no indivíduo i designadas por $\mathbf{y}_i^T = (y_{i1}, y_{i2}, \ldots, y_{im_i})$. Cada variável aleatória $\mathbf{Y}_{it}$ tem associado o vetor das p covariáveis com dimensão $p \times 1$, isto é, $\mathbf{x}_{it}^T = (x_{it}^1, ..., x_{im_i}^p)$. Assim, para cada $\mathbf{Y}_i$,

está associada a matriz de ordem $m_i \times p$ e é dada por

$$\mathbf{X_i} = \begin{bmatrix} x_{i1}^1 & \dots & x_{i1}^p \\ x_{i2}^1 & \dots & x_{i2}^p \\ \vdots & \vdots & \vdots \\ x_{im_i}^1 & \dots & x_{im_i}^p \end{bmatrix}. \tag{4.64}$$

O valor esperado e a variância de Y_i são designados por $E(\mathbf{Y}_i) = \mu_i$ e $Var(\mathbf{Y}_i) = \mathbf{v}_i$ e o vetor $p \times 1$ dos parâmetros desconhecidos, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)$. A variável resposta é dada pelo vetor de dimensão $M \times 1$,

$$\mathbf{Y} = \begin{bmatrix} Y_1^T \\ Y_2^T \\ \vdots \\ Y_{m_i}^T \end{bmatrix} = \begin{bmatrix} Y_{11} \\ \vdots \\ Y_{1m_1} \\ Y_{21} \\ \vdots \\ Y_{2m_2} \\ \vdots \\ Y_{n1} \\ \vdots \\ Y_{nm_n} \end{bmatrix}, \tag{4.65}$$

em que $M = \sum_{i=1}^{n} m_i$. A respetiva matriz de desenho é dada por

$$\mathbf{X} = \begin{bmatrix} x_{11}^1 & \dots & x_{11}^p \\ x_{12}^1 & \dots & x_{12}^p \\ \vdots & \vdots & \vdots \\ x_{1m_1}^1 & \dots & x_{1m_1}^p \\ x_{21}^1 & \dots & x_{21}^p \\ x_{22}^1 & \dots & x_{22}^p \\ \vdots & \vdots & \vdots \\ x_{2m_2}^1 & \dots & x_{2m_2}^p \\ \vdots & \vdots & \vdots \\ x_{n1}^1 & \dots & x_{n1}^p \\ x_{n2}^1 & \dots & x_{n2}^p \\ \vdots & \vdots & \vdots \\ x_{nm_n}^1 & \dots & x_{nm_n}^p \end{bmatrix}, \tag{4.66}$$

associado ao vetor de parâmetros desconhecidos $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)^T$.

4.2.2 Formulação do Modelo

Os Modelos de Efeitos Mistos (LMM) permitem descrever as relações lineares entre uma variável resposta e uma ou mais covariáveis, num determinado conjunto de dados, como é o caso das medições repetidas. Estes modelos são uma extensão dos Modelos de Regressão Linear, com a vantagem de incluir efeitos aleatórios. Os efeitos aleatórios são inseridos no modelo de forma a introduzir correlação entre as medidas de um mesmo indivíduo, pois permitem retratar modificações dentro de cada indivíduo, ao longo do tempo, e representar a estrutura de variância-covariância de forma mais flexível, de acordo com os dados. Os Modelos de Efeitos Mistos, tal como o nome indica, incorporam os efeitos aleatórios, integrando a aleatoriedade, resultante dos indivíduos, e os efeitos fixos, isto é, os parâmetros associados a toda a população. O modelo permite que a base de dados não seja balanceada, ou seja, a base de dados pode não ser completa ou conter ausência de informação (Fausto *et al.*, 2008). A possibilidade de utilização destes modelos, na presença de dados omissos, é outra das vantagens dos Modelos de Efeitos Mistos (Wu, 2009).

Laird & Ware (1982) definem o Modelo de Efeitos Mistos para um único nível de agrupamento como

$$\mathbf{Y}_{it} = \underbrace{\beta_1 X_{1it} + \beta_2 X_{2it} + \cdots + \beta_p X_{pit}}_{Componente\ Fixa} + \underbrace{u_1 Z_{1it} + u_2 Z_{2it} + \cdots + u_q Z_{qit}}_{Componente\ Aleatória} + \epsilon_{it}, \tag{4.67}$$

em que $X_{1it}, \ldots, X_{pit}$ representam as covariáveis da parte fixa, $Z_{1it}, \ldots, Z_{qit}$ são as covariáveis da parte aleatória, com $t = 1, \ldots, m_i$ e $i = 1, \ldots, n$, ϵ_{it} trata-se do erro aleatório associado ao indivíduo i, no tempo t.

Alternativamente, o modelo pode ser expresso na forma matricial

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i + \epsilon_i \qquad\qquad i = 1, \ldots, n, \tag{4.68}$$

em que $\mathbf{Y}_i^T = (Y_{i1}, \ldots, Y_{i,m_i})$ é o vetor da variável resposta para o indivíduo i, $\boldsymbol{\beta}$ é o vetor dos efeitos fixos, $\mathbf{X_i}$ é a matriz das covariáveis dos efeitos fixos. $\mathbf{u_i}$ é o vetor dos efeitos aleatórios, $\mathbf{Z_i}$ é a matriz das covariáveis dos efeitos aleatórios e $\epsilon_\mathbf{i}$ é o vetor dos erros aleatórios dentro do grupo. Na utilização dos Modelos de Efeitos Mistos é necessário verificar alguns pressupostos, quer na inferência, quer na estimação de parâmetros (Bolker *et al.*, 2009)

$$\begin{aligned}&\mathbf{u}_i \sim N(\mathbf{0}, \mathbf{D}),\\ &\epsilon_i \sim N(\mathbf{0}, \mathbf{\Sigma}_i),\\ &\mathbf{u}_i \ \ e \ \ \epsilon_i \text{ são independentes para diferentes grupos } i \text{ e entre si,}\end{aligned} \tag{4.69}$$

em que $\mathbf{D}$ e $\mathbf{\Sigma_i}$ são matrizes definidas positivas. Sob as hipóteses supracitadas e o modelo (4.68), condicionado ao efeito aleatório $\mathbf{u_i}$, tem-se que a distribuição de $\mathbf{Y_i}$ segue uma Distribuição Multivariada Normal com o valor médio $\mathbf{X_i}\boldsymbol{\beta} + \mathbf{Z_i u_i}$ e matriz variância-covariância $\sum_{\mathbf{i}}$, isto é,

$$\mathbf{Y_i}|\mathbf{u_i} \sim MVN(\mathbf{X_i}\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i, \mathbf{\Sigma_i}), \tag{4.70}$$

em que $\mathbf{\Sigma_i}$ é a matriz da variação intragrupo. A função densidade de probabilidade pode ser escrita

$$\begin{aligned}f(\mathbf{y_i}|\mathbf{u_i}) &= (2\pi)^{-\frac{m_i}{2}}|\mathbf{\Sigma_i}|^{-\frac{1}{2}}\\ &\times exp\left\{ -\frac{1}{2}(\mathbf{Y_i} - \mathbf{X_i}\boldsymbol{\beta} - \mathbf{Z_i u_i})^T \mathbf{\Sigma_i}^{-1}(\mathbf{Y_i} - \mathbf{X_i}\boldsymbol{\beta} - \mathbf{Z_i u_i})\right\}.\end{aligned} \tag{4.71}$$

Considerando a distribuição de $\mathbf{u_i}$, a função densidade de probabilidade é dada por

$$f(\mathbf{u_i}) = (2\pi)^{-\frac{q}{2}}|\mathbf{D}|^{-\frac{1}{2}} \times exp\left\{ -\frac{1}{2}\mathbf{u_i}^T\mathbf{D}^{-1}\mathbf{u_i}\right\}. \tag{4.72}$$

Deste modo, a função de densidade de probabilidade marginal de $\mathbf{Y_i}$ é dada por

$$\begin{aligned}f(\mathbf{y_i}) &= \int f(\mathbf{y_i}|\mathbf{u_i})f(\mathbf{u_i})d\mathbf{u_i}\\ &= (2\pi)^{-\frac{m_i}{2}}|\mathbf{V_i}|^{-\frac{1}{2}} \times exp\left\{ -\frac{1}{2}(\mathbf{Y_i} - \mathbf{X_i}\boldsymbol{\beta})^T\mathbf{V_i}^{-1}(\mathbf{Y_i} - \mathbf{X_i}\boldsymbol{\beta})\right\},\end{aligned} \tag{4.73}$$

ou seja, a função densidade de $f(\mathbf{y_i})$ é a função densidade de uma variável aleatória Normal m_i dimensional, com valor médio $\mathbf{X_i}\boldsymbol{\beta}$ e matriz de variância-covariância $\mathbf{V_i} = \mathbf{Z_i D Z_i^T} + \mathbf{\Sigma_i}$, ou seja,

$$\mathbf{Y}_i \sim N(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{V}_i), \tag{4.74}$$

em que a matriz $\mathbf{V_i}$ representa a variação entre grupos definida positiva. Molenberghs & Verbeke (2005) designam a equação (4.73) por Formulação Marginal do Modelo.

Para cada grupo, tendo em conta todos os modelos (4.68), tem-se que

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \epsilon, \tag{4.75}$$

em que $\mathbf{Y}$ é o vetor de todas as observações da variável resposta, com dimensão $N \times 1$, $\mathbf{X}$ é matriz dimensão $N \times p$, $\mathbf{Z}$ é matriz de dimensão $N \times q$, $\mathbf{u}$ e ϵ vetores de dimensão $qn \times 1$ e $N \times 1$, respetivamente. O modelo (4.75) tem valor médio $E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$ e matriz variância-covariância $Var(\mathbf{Y}) = \mathbf{ZFZ^T} + \mathbf{\Sigma} = \mathbf{V}$, em que Z, F, Σ e V são matrizes diagonais compostas por blocos de dimensão $N \times qn$, $qn \times qn$, $N \times N$ e $N \times N$ dadas por

$$\mathbf{Z} = \begin{bmatrix} Z_1 & 0 & \dots & 0 \\ 0 & Z_2 & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & Z_n \end{bmatrix}, \tag{4.76}$$

$$\mathbf{F} = \begin{bmatrix} D & 0 & \dots & 0 \\ 0 & D & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & D \end{bmatrix}, \tag{4.77}$$

$$\mathbf{\Sigma} = \begin{bmatrix} \Sigma_1 & 0 & \dots & 0 \\ 0 & \Sigma_2 & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & \Sigma_n \end{bmatrix} \tag{4.78}$$

e

$$\mathbf{V} = \begin{bmatrix} V_1 & 0 & \dots & 0 \\ 0 & V_2 & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & V_n \end{bmatrix}. \tag{4.79}$$

A distribuição da variável aleatória $\mathbf{Y}$ é Multivariada Normal, isto é,

$$\mathbf{Y} \sim MVN(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}). \tag{4.80}$$

Assim, o modelo (4.68) pode ser definido através das funções densidade de probabilidade de $\mathbf{y}_i|\mathbf{u}_i$ e de $\mathbf{u}_i$. A esta definição chama-se Formulação Hierárquica do Modelo de Efeitos Mistos.

4.2.3 Estimação dos Efeitos Fixos

Na estimação dos parâmetros do Modelo Marginal, os métodos mais utilizados para a sua estimação são o Método de Máxima Verosimilhança (ML) e o Método de Máxima Verosimilhança Restrita (REML). O objetivo é a estimação dos coeficientes $\boldsymbol{\beta}$ associados

à parte fixa do Modelo de Efeitos Mistos e as matrizes $\mathbf{D}$ e $\mathbf{\Sigma_i}$, que representam as componentes de variância. Para simplificação das formulações apresentadas, considera-se que $\mathbf{\Sigma_i} = \sigma^2\mathbf{I}_{m_i}$, $\mathbf{D} = \sigma^2\mathbf{G}$, $\mathbf{V}_i = \sigma^2(\mathbf{Z}_i\mathbf{G}\mathbf{Z}_i^T + \mathbf{I}_{m_i}) = \sigma^2\mathbf{M}_i$, $\mathbf{V} = \sigma^2(\mathbf{Z}\ diag(\mathbf{G})\mathbf{Z}^T + \mathbf{I}) = \sigma^2\mathbf{M}$.

Método de Máxima Verosimilhança

Assumindo a independência entre os indivíduos, tem-se que a função verosimilhança é dada por

$$\begin{aligned} L_{ML}(\mathbf{y};\boldsymbol{\beta},\boldsymbol{\alpha}) = &\prod_{i=1}^{n}(2\pi)^{-\frac{m_i}{2}}|\mathbf{V_i}(\alpha)^{-\frac{1}{2}}| \\ &\times exp\Big\{-\frac{1}{2}(\mathbf{y_i}-\mathbf{X_i}\boldsymbol{\beta})^{\mathbf{T}}\mathbf{V_i}^{-\mathbf{1}}(\alpha)(\mathbf{y_i}-\mathbf{X_i}\boldsymbol{\beta})\Big\}, \end{aligned} \tag{4.81}$$

ou, outra forma de escrita desta função é

$$\begin{aligned} L_{ML}(\mathbf{y};\boldsymbol{\beta},\theta,\sigma^{\mathbf{2}}) = &\prod_{i=1}^{n}(2\pi\sigma^2)^{-\frac{m_i}{2}}|\mathbf{M_i}(\theta)|^{-\frac{1}{2}} \\ &\times exp\Big\{-\frac{1}{2\sigma^2}(\mathbf{y_i}-\mathbf{X_i}\boldsymbol{\beta})^{\mathbf{T}}\mathbf{M_i}(\theta)^{-\mathbf{1}}(\mathbf{y_i}-\mathbf{X_i}\boldsymbol{\beta})\Big\}. \end{aligned} \tag{4.82}$$

Aplicando o logaritmo nas equações (4.81) e (4.82), obtém-se as seguintes equações

$$\begin{aligned} \ell_{ML}(\mathbf{y};\boldsymbol{\beta},\alpha) = &-\frac{N}{2}log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}log|\mathbf{V_i}(\alpha)| \\ &-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{y_i}-\mathbf{X_i}\boldsymbol{\beta})^{\mathbf{T}}\mathbf{V_i}^{-\mathbf{1}}(\alpha)(\mathbf{y_i}-\mathbf{X_i}\boldsymbol{\beta}) \end{aligned} \tag{4.83}$$

e

$$\begin{aligned} \ell(\mathbf{y};\boldsymbol{\beta},\theta,\sigma^{\mathbf{2}}) = &-\frac{N}{2}log(2\pi\sigma^2) - \frac{1}{2}\sum_{i=1}^{n}log|\mathbf{M_i}(\theta)| \\ &-\sum_{i=1}^{n}\frac{1}{2\sigma^2}(\mathbf{y_i}-\mathbf{X_i}\boldsymbol{\beta})^{\mathbf{T}}\mathbf{M_i}(\theta)^{-\mathbf{1}}(\mathbf{y_i}-\mathbf{X_i}\boldsymbol{\beta}). \end{aligned} \tag{4.84}$$

Para a obtenção dos estimadores dos parâmetros, observam-se vários cenários: α conhecido, θ conhecido ou α desconhecido. No primeiro caso, deriva-se a equação (4.83) em função do parâmetro pretendido, $\boldsymbol{\beta}$, iguala-se a zero e resolve-se em ordem ao parâmetro que se pretende estimar

$$\hat{\boldsymbol{\beta}}_{ML}(\alpha) = \left(\sum_{i=1}^{n}\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i\right) = \big(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\big)^{-1}\big(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\big), \tag{4.85}$$

obtendo-se o estimador ML para $\boldsymbol{\beta}$. Este é o estimador dos Mínimos Quadrados generalizados de $\boldsymbol{\beta}$, em que α é conhecido. No segundo caso, deriva-se a equação (4.84) em função dos parâmetros pretendidos, $\boldsymbol{\beta}$ e σ^2, iguala-se a zero e resolve-se em ordem aos parâmetros supracitados, (4.86) e (4.87), obtendo-se

$$\hat{\boldsymbol{\beta}}_{ML}(\theta) = \left(\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{X_i^T M_i^{-1} X_i} \right)^{-1} \left(\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{X_i^T M_i^{-1} Y_i} \right) \tag{4.86}$$

e

$$\hat{\sigma}_{\mathbf{ML}}(\theta) = \frac{\mathbf{1}}{\mathbf{N}} \sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{Y_i} - \mathbf{X_i}\hat{\boldsymbol{\beta}}(\theta))^{\mathbf{T}} \mathbf{M_i^{-1}} (\mathbf{Y_i} - \mathbf{X_i}\hat{\boldsymbol{\beta}}(\theta)). \tag{4.87}$$

O terceiro caso é o mais usual, α desconhecido, inicia-se por substituir o $\boldsymbol{\beta}$ obtido em (4.81) e maximiza-se a equação resultante, em ordem a α. O estimador para $\boldsymbol{\beta}$ obtido é dado por

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{ML} &= \left(\sum_{i=1}^{n} \mathbf{X_i^T \hat{V}_i^{-1}}(\hat{\alpha}_{\mathbf{ML}}) \mathbf{X_i} \right)^{-1} \left(\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{X_i^T \hat{V}_i^{-1}}(\hat{\alpha}_{\mathbf{ML}}) \mathbf{Y_i} \right) \\ &= \left(\mathbf{X^T \hat{V}^{-1}}(\hat{\alpha}_{\mathbf{ML}}) \mathbf{X} \right)^{-1} \mathbf{X^T \hat{V}^{-1}}(\hat{\alpha}_{\mathbf{ML}}) \mathbf{Y}. \end{aligned} \tag{4.88}$$

Nos três casos é necessário recorrer a métodos de otimização numérica para a maximização das equações anteriores. Uma das desvantagens deste método prende-se com o facto dos estimadores das componentes de variância resultantes não serem centrados. Outra refere-se à escolha inapropriada da matriz de desenho $\mathbf{X}$ que poderá levar a estimadores das componentes de variância não consistentes. Tal deve-se ao método não ter em conta a perda de graus de liberdade e, consequentemente, originar estimadores das componentes de variância enviesados. Uma proposta de resolução é o aumento das covariáveis, contudo poderá enviesar ainda mais os estimadores.

Método de Máxima Verosimilhança Restrita

Patterson & Thompson (1971) e Harville (1974, 1977) propõem um método alternativo que soluciona o enviesamento dos estimadores das componentes da variância. Este método tem como suporte uma linearização dos dados tal que $\mathbf{W} = \mathbf{B^T Y}$, em que $\mathbf{B}$ é uma matriz de característica completa ortogonal às colunas da matriz de desenho $\mathbf{X}$, ou seja, por exemplo $\mathbf{B} = \mathbf{I} - \mathbf{X}(\mathbf{XX})^{-1}\mathbf{X^T}$ (existem outras formas de obtenção).

A distribuição de $\mathbf{W}$ segue a Distribuição Normal tal que $\mathbf{W} \sim \mathbf{N}(\mathbf{0}, \mathbf{B^T V}(\alpha)\mathbf{B})$, qualquer que seja $\boldsymbol{\beta}$ e a cada elemento designado de W por contraste de erro (Molenberghs & Verbeke, 2005).

Considerando o $\hat{\boldsymbol{\beta}}$, Harville (1974) propôs para a função verosimilhança para os con-

trastes de erros (ou função de verosimilhança restrita) a seguinte fórmula

$$\begin{aligned} L_{REML}(\mathbf{y};\alpha) = (2\pi)^{-\frac{N-p}{2}} \left|\sum_{i=1}^{n} \mathbf{X}_i^T\mathbf{X}_i\right|^{\frac{1}{2}} \times \left|\sum_{i=1}^{n} \mathbf{X}_i^T\mathbf{V}_i^{-1}(\alpha)\mathbf{X}\right|^{\frac{1}{2}} \\ \times \prod_{i=1}^{n} |\mathbf{V}_i^{-1}(\alpha)|^{-\frac{1}{2}} \\ \times exp\Big\{-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})^T\mathbf{V}_i^{-1}(\alpha)(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})\Big\}, \end{aligned} \tag{4.89}$$

em que $\hat{\boldsymbol{\beta}}$ é dado por (4.88).

Aplicando o logaritmo à equação (4.89) tem-se que

$$\begin{aligned} \ell_{REML} = logL_{REML}(\mathbf{y};\alpha) = constante - \frac{1}{2}log\left|\sum_{i=1}^{n} \mathbf{X_i^T}\mathbf{V_i^{-1}}(\alpha)\mathbf{X}\right| \\ - \frac{1}{2}\sum_{i=1}^{n} log|\mathbf{V_i}(\alpha)| \\ \times \sum_{i=1}^{n}(\mathbf{y_i} - \mathbf{X_i}\hat{\boldsymbol{\beta}})^{\mathbf{T}}\mathbf{V_i}(\alpha)^{-\mathbf{1}}(\alpha)(\mathbf{y_i} - \mathbf{X_i}\hat{\boldsymbol{\beta}}). \end{aligned} \tag{4.90}$$

De forma análoga ao referido anteriormente, maximiza-se a equação (4.90) em ordem ao parâmetro pretendido, α e obtém-se os estimadores REML de $\boldsymbol{\alpha}$. Confrontando as equações (4.83) e (4.90) constata-se que divergem no termo $-\frac{1}{2}log\Big|\sum_{i=1}^{n} \mathbf{X_i^T}\mathbf{V_i^{-1}}(\alpha)\mathbf{X}\Big|$, a menos de uma constante. Pode-se reescrever a equação (4.90) da seguinte forma

$$\begin{aligned} \ell_R = logL_{REML}(\mathbf{y};\alpha) = constante - \frac{1}{2}log\left|\sum_{i=1}^{n} \mathbf{X_i^T}\mathbf{V_i^{-1}}(\alpha)\mathbf{X}\right| \\ + logL_{ML}(\mathbf{y};\hat{\boldsymbol{\beta}}(\alpha),\alpha). \end{aligned} \tag{4.91}$$

O termo $\Big|\mathbf{X_i^T}\mathbf{V_i^{-1}}(\alpha)\mathbf{X}\Big|$ não depende de $\boldsymbol{\beta}$, então é possível obter os estimadores de REML através da seguinte função de verosimilhança restrita

$$L_{REML}(\mathbf{y};\boldsymbol{\beta},\alpha) = \mathbf{L_{ML}}(\mathbf{y};\boldsymbol{\beta},\alpha)\Big|\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{X_i^T}\mathbf{V_i^{-1}}(\alpha)\mathbf{X}\Big|^{-\frac{1}{2}}, \tag{4.92}$$

com respeito a todos os parâmetros de $\boldsymbol{\alpha}$ e $\boldsymbol{\beta}$, assim, o estimador $\boldsymbol{\beta}$ pode ser obtido pela

seguinte equação

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{REML} &= \Big(\sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{V}_i^{-1}(\alpha_{REML})\mathbf{X_i}\Big)^{-\mathbf{1}} \Big(\sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{X_i^T V_i^{-1}}(\alpha_{\mathbf{REML}})\mathbf{Y_i}\Big) \\ &= \big(\mathbf{X}^T \mathbf{V}_{REML}^{-1} \mathbf{X}_i\big)^{-1} \mathbf{X}^T \mathbf{V}_{REML}^{-1} \mathbf{Y}.\end{aligned} \tag{4.93}$$

Relativamente aos métodos apresentados, verifica-se que os estimadores resultantes não dependem de $\mathbf{B}$ e $\boldsymbol{\beta}_{REML}$ é diferente de $\boldsymbol{\beta}_{ML}$. A partir de $\mathbf{W}$, este método permite que não haja perda de informação sobre α (Patterson & Thompson, 1971).

O método REML depende do estimador $\boldsymbol{\beta}$ obtido através do método ML. O termo $-\frac{1}{2} log\Big|\sum_{i=1}^{n} \mathbf{X_i^T V_i^{-1}}(\alpha)\mathbf{X}\Big|$ revela que quaisquer modificações na matriz de desenho originam reparametrizações nos efeitos fixos e, como consequência, não é possível comparar diferentes Modelos de Efeitos Mistos com diferentes componentes fixas (Pinheiro & Bates, 2000).

Os métodos conduzem a resultados semelhantes (Zuur *et al.*, 2009), contudo as estimativas obtidas por ML das componentes da variância são menores do que as obtidas por REML.

4.2.4 Predição dos Efeitos Aleatórios

Os efeitos aleatórios $\mathbf{u_i}$ representam a variabilidade dentro do mesmo indivíduo (evolução do indivíduo), relativamente à média da população $\mathbf{X_i}\boldsymbol{\beta}$. Nos Modelos de Efeitos Mistos é importante perceber quais os efeitos a incluir. A distinção dos efeitos aleatórios dos fixos, a seleção e a inferência não é fácil. Um cuidado adicional é a especificação dos modelos aninhados, integrando os efeitos aleatórios. Na prática, os efeitos aleatórios serão determinados pelo estudo definido ou pela amostra recolhida (Schielzeth & Nakagawa, 2013). Bates *et al.* (2015) fornecem um bom guia para determinar, de forma iterativa, a complexidade ideal da estrutura de efeitos aleatórios.

Considerando que $\mathbf{u}_i$ são variáveis aleatórias, recorre-se à abordagem Bayesiana para obter a melhor predição. Anteriormente, viu-se que

$$\mathbf{Y_i}|\mathbf{u_i} \sim \mathbf{N}(\mathbf{X_i}\boldsymbol{\beta} + \mathbf{Z_i u_i}, \boldsymbol{\Sigma_i}) \tag{4.94}$$

e que a distribuição *a priori* é dada por

$$\mathbf{u_i} \sim \mathbf{N}(\mathbf{0}, \mathbf{D}). \tag{4.95}$$

A sua distribuição *a posteriori* pode ser determinada a partir da distribuição de $\mathbf{u_i}$, condicionada a $\mathbf{Y_i} = \mathbf{y_i}$. Assim, a distribuição *a posteriori* é dada por

$$f(\mathbf{u_i}|\mathbf{y_i}) = \mathbf{f}(\mathbf{u_i}|\mathbf{y_i}) = \frac{\mathbf{f}(\mathbf{y_i}|\mathbf{u_i})\mathbf{f}(\mathbf{u_i})}{\int \mathbf{f}(\mathbf{y_i}|\mathbf{u_i})\mathbf{f}(\mathbf{u_i})\mathbf{du_i}}. \tag{4.96}$$

Neste sentido, a distribuição conjunta é Normal Multivariada (Azzalini, 1996),

$$\begin{pmatrix} \mathbf{u_i} \\ \mathbf{Y_i} \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{X_i}\boldsymbol{\beta} \end{pmatrix}, \begin{bmatrix} \mathbf{D} & \mathbf{DZ_i^T} \\ \mathbf{Z_iD} & \mathbf{Z_iDZ_i^T} + \mathbf{\Sigma_i} \end{bmatrix} \right) \tag{4.97}$$

e, consequentemente,

$$\mathbf{u_i}|\mathbf{y_i} \sim \mathbf{N}\big(\mathbf{DZ_i^T}(\mathbf{Z_iDZ_i^T} + \mathbf{\Sigma_i})^{-\mathbf{1}}(\mathbf{y_i} - \mathbf{X_i}\boldsymbol{\beta}), \mathbf{V}\big), \tag{4.98}$$

em que $\mathbf{V} = \mathbf{D} - \mathbf{DZ_i^T}(\mathbf{Z_iDZ_i^T} + \mathbf{\Sigma_i})^{-\mathbf{1}}\mathbf{Z_iD}$. O valor esperado para a distribuição *a posteriori* pode ser escrito da seguinte forma

$$\begin{aligned} E(\mathbf{u_i}|\mathbf{Y_i} = \mathbf{y_i}) &= \int \mathbf{u_i}\mathbf{f}(\mathbf{u_i}|\mathbf{y_i})\mathbf{du_i} \\ &= \mathbf{DZ_i^T}(\mathbf{Z_iDZ_i^T} + \mathbf{\Sigma_i})^{-\mathbf{1}}(\mathbf{y_i} - \mathbf{X_i}\boldsymbol{\beta}) \\ &= \mathbf{DZ_i^T}\mathbf{V_i^{-1}}(\mathbf{y_i} - \mathbf{X_i}\boldsymbol{\beta}). \end{aligned} \tag{4.99}$$

Considerando que $\boldsymbol{\alpha}$ é conhecido, o melhor preditor linear centrado para a variável que representa os efeitos aleatórios é determinado com base na equação (4.99), em que $\boldsymbol{\beta}$ é aproximado por $\hat{\boldsymbol{\beta}}(\alpha) = (\mathbf{XV}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$ (McCulloch & Searle, 2001), em que se tem

$$\mathbf{u}_{BLUP,i}(\alpha) = \mathbf{DZ_i^T}\mathbf{V_i^{-1}}(\mathbf{y_i} - \mathbf{X_i}\hat{\boldsymbol{\beta}}(\alpha)), \tag{4.100}$$

ou, alternativamente,

$$\mathbf{u}_{BLUP}(\alpha) = \mathbf{DZ^T}\mathbf{V^{-1}}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}(\alpha)). \tag{4.101}$$

O preditor da combinação linear de $\mathbf{v} = \mathbf{v}_B^T\boldsymbol{\beta} + \mathbf{v}_u^T\mathbf{u}_i$, condicionado a α é dado por

$$\mathbf{v}_{BLUP}(\alpha) = v_{\boldsymbol{\beta}}^T\hat{\boldsymbol{\beta}}(\alpha) + v_u^T\hat{\mathbf{u}}_{BLUP}(\alpha), \tag{4.102}$$

em que $v_{\boldsymbol{\beta}}$ representa o vetor $p \times 1$ dos efeitos fixos e v_u representa o vetor $p \times 1$ dos efeitos aleatórios. Harville (1976) e Searle *et al.* (1992) provam que $\mathbf{v}_{BLUP}(\alpha)$ é o melhor preditor linear centrado para $\mathbf{v}$.

Equações

A solução do sistema de equações lineares, às quais se dá o nome de equações do Modelo de Efeitos Mistos, resultam em estimadores dos efeitos fixos $\boldsymbol{\beta}$ estimados e preditores dos efeitos aleatórios $\mathbf{u}$. Sabe-se que $\mathbf{Y}|\mathbf{u} \sim \mathbf{N}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Zu}, \mathbf{\Sigma})$ e que $\mathbf{u} \sim \mathbf{N}(\mathbf{0}, \mathbf{F})$, assim a

densidade conjunta de $\mathbf{Y}$ e $\mathbf{u}$ é dada por

$$\begin{aligned} f(\mathbf{y},\mathbf{u}) &= \mathbf{f}(\mathbf{y}|\mathbf{b})\mathbf{f}(\mathbf{b}) \\ &= (2\pi)^{-\frac{N}{2}}|\boldsymbol{\Sigma}^{-\frac{1}{2}}| \\ &\times exp\Big\{ -\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})\Big\} \\ &\times (2\pi)^{-\frac{qn}{2}}|\mathbf{F}^{-\frac{1}{2}}|exp\Big\{ -\frac{1}{2}\mathbf{u}^T\mathbf{F}^{-1}\mathbf{u}\Big\} \\ &= (2\pi)^{-\frac{N+qn}{2}}|\boldsymbol{\Sigma}^{-\frac{1}{2}}||\mathbf{F}^{-\frac{1}{2}}| \\ &\times exp\Big\{ -\frac{1}{2}\Big[(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})+\mathbf{u}^{\mathbf{T}}\mathbf{F}^{-1}\mathbf{u}\Big]\Big\}. \end{aligned} \tag{4.103}$$

Aplicando o logaritmo na equação anterior tem-se que

$$\begin{aligned} logf(\mathbf{y},\mathbf{u}) = &-\frac{N+qn}{2}log(2\pi)-\frac{1}{2}log|\boldsymbol{\Sigma}|-\frac{1}{2}log|\mathbf{F}| \\ &-\frac{1}{2}\Big[(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})^{\mathbf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})+\mathbf{u}^{\mathbf{T}}\mathbf{F}^{-1}\mathbf{u}\Big]. \end{aligned} \tag{4.104}$$

Derivando em relação aos parâmetros de interesse e igualando a zero, tem-se o seguinte sistema de equações

$$\begin{cases} \dfrac{\partial logf(\mathbf{y},\mathbf{b})}{\partial\boldsymbol{\beta}} &= 0 \\ \dfrac{\partial logf(\mathbf{y},\mathbf{b})}{\partial\mathbf{u}} &= 0 \end{cases} \Leftrightarrow \begin{cases} \mathbf{X}^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u}) &= 0 \\ \\ \mathbf{Z}^T\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\mathbf{u})-\mathbf{F}\mathbf{u} &= 0 \end{cases}, \tag{4.105}$$

alternativamente,

$$\begin{bmatrix} \mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X} & \mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{Z} \\ \mathbf{Z}^T\boldsymbol{\Sigma}^{-1}\mathbf{X} & \mathbf{Z}^T\boldsymbol{\Sigma}^{-1}\mathbf{Z}+\mathbf{F} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{y} \\ \mathbf{Z}^T\boldsymbol{\Sigma}^{-1}\mathbf{y} \end{bmatrix}. \tag{4.106}$$

Considerando $\mathbf{F}$ e $\boldsymbol{\Sigma}$ conhecidos, tem-se que

$$\begin{aligned} \hat{\boldsymbol{\beta}}(\alpha) &= (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} \\ &= \Big(\sum_{i=1}^{n}\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i\Big)^{-1}\sum_{i=1}^{n}\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{y}_i \end{aligned} \tag{4.107}$$

e

$$\mathbf{u}_{BLUP} = \mathbf{F}\mathbf{Z}^T\mathbf{V}^{-1}\Big(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}(\alpha)\Big), \tag{4.108}$$

em alternativa,

$$\mathbf{u}_{BLUP,i} = \mathbf{F}\mathbf{Z}_i^T\mathbf{V}_i^{-1}\Big(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta}(\alpha)\Big). \tag{4.109}$$

O estimador $\boldsymbol{\beta}$, resultante das equações do Modelo de Efeitos Mistos, é igual ao obtido pelo método ML.

Métodos Numéricos

As equações de maximização do logaritmo da função verosimilhança exigem Métodos Numéricos para a sua otimização (Pinheiro & Bates, 1995). Existem vários Métodos Numéricos de otimização, como são exemplo o Algoritmo *Expected-Maximization* e o Método de Newton-Raphson. Estes dois estão implementados no `R`.

O Algoritmo *Expected-Maximization* é um método iterativo e cada iteração consiste em dois passos. O primeiro determina o valor esperado do logaritmo da função verosimilhança e o segundo a sua maximização. Para garantir a existência de convergência no algoritmo, é necessário ter atenção aos valores iniciais. A cada iteração existe um aumento da verosimilhança (Dempster *et al.*, 1977; David e Giltinian, 1993).

O Método de Newton-Raphson recorre à derivada do logaritmo da função verosimilhança, função *Score*, que, igualando a zero, resulta num sistema de equações, geralmente não lineares. Em cada iteração é necessário o cálculo da função *Score* e da sua derivada. O cálculo desta última é bastante complexo, existem estudos que tentam reduzir essa complexidade, nomeadamente o Método Quasi-Newton (Thisted, 1988).

No `R` está implementada uma metodologia em que se inicia com as estimativas utilizando o algoritmo EM e, quando perto do ótimo, utiliza-se o Método de Newton-Raphson.

4.2.5 Matriz Variância-Covariância dos Erros Aleatórios

O Modelo Misto permite flexibilizar a estrutura dos efeitos aleatórios, sob a condição da estrutura dos erros aleatórios $\boldsymbol{\Sigma}_i = \sigma^2 \mathbf{I}_{m_i}$. Contudo, esta condição nem sempre se verifica, pois, na maioria dos casos, as observações estão correlacionadas. A heterocedasticidade dos erros aleatórios pode ser modelada se for considerado outro tipo de estrutura para $\boldsymbol{\Sigma}_i$. Considerando a equação do Modelo de Efeitos Mistos, a estrutura para os erros aleatórios é dada por

$$\epsilon_i \sim N(0, \sigma^2 \boldsymbol{\Lambda}_i) \qquad i = 1, \ldots, n, \tag{4.110}$$

em que $\boldsymbol{\Lambda}_i$ é uma matriz definida positiva parametrizada por um número geralmente pequeno de parâmetros que se designa por $\boldsymbol{\lambda}$. Dado que admite raiz quadrada invertível, $\boldsymbol{\Lambda}_i^{-1/2}$, tem-se que

$$\boldsymbol{\Lambda}_i = (\boldsymbol{\Lambda}_i^{1/2})^T \boldsymbol{\Lambda}_i^{1/2} \qquad \boldsymbol{\Lambda}_i^{-1} = \boldsymbol{\Lambda}_i^{-1/2} (\boldsymbol{\Lambda}_i^{-1/2})^T. \tag{4.111}$$

Considerando a reparametrização $\mathbf{Y}_i^* = (\mathbf{\Lambda}_i^{-1/2})^T\mathbf{Y}_i$, $\mathbf{X}_i^* = (\mathbf{\Lambda}_i^{-1/2})^T\mathbf{X}_i$, $\mathbf{Z}_i^* = (\mathbf{\Lambda}_i^{-1/2})^T\mathbf{Z}_i$ e $\epsilon_i^* = (\mathbf{\Lambda}_i^{-1/2})^T\epsilon_i$, o modelo pode ser escrito da seguinte forma

$$\mathbf{Y}_i^* = \mathbf{X}_i^*\boldsymbol{\beta} + \mathbf{Z}_i^*\mathbf{u}_i^* + \epsilon_i^* \qquad i = 1, \ldots, n, \tag{4.112}$$

em que $\epsilon_i^* \sim N(0, \sigma^2\mathbf{I})$ e $\mathbf{u}_i^* \sim N(0, \mathbf{D})$. De notar que $E[\epsilon_i^*] = (\mathbf{\Lambda}_i^{-1/2})^T E[\epsilon_i]$ e $var(\epsilon_i^*) = \mathbf{\Lambda}_i^{-1/2} var[\epsilon_i](\mathbf{\Lambda}_i^{-1/2})^T = \sigma^2\mathbf{I}$.

A função de verosimilhança para o modelo é dado por

$$\begin{aligned} L_{ML}(\mathbf{y}; \boldsymbol{\beta}, \theta, \sigma^2, \lambda) &= \prod_{i=1}^{n} f(\mathbf{y}_i; \boldsymbol{\beta}, \theta, \sigma^2, \lambda) \\ &= \prod_{i=1}^{n} f(\mathbf{y}_i^*; \boldsymbol{\beta}, \theta, \sigma^2, \lambda)\left|\mathbf{\Lambda}_i^{-1/2}\right| \\ &= L_{ML}(\mathbf{y}^*; \boldsymbol{\beta}, \theta, \sigma^2, \lambda) \prod_{i=1}^{n} \left|\mathbf{\Lambda}_i^{-1/2}\right|, \end{aligned} \tag{4.113}$$

e a função de verosimilhança restrita é dada por

$$L_{REML}(\mathbf{y}; \boldsymbol{\beta}, \theta, \sigma^2, \lambda) = L_{REML}(\mathbf{y}^*; \boldsymbol{\beta}, \theta, \sigma^2, \lambda) \prod_{i=1}^{n} \left|\mathbf{\Lambda}_i^{-1/2}\right|. \tag{4.114}$$

A matriz de variância-covariância $\mathbf{\Lambda}_i$ pode ser decomposta num produto de matrizes

$$\mathbf{\Lambda}_i = \mathbf{W}_i\mathbf{C}_i\mathbf{W}_i, \tag{4.115}$$

em que $\mathbf{W}_i$ e $\mathbf{C}_i$ são uma matriz diagonal (variância) e uma matriz de correlação, respetivamente. Para que a matriz $\mathbf{W}_i$ seja única, é necessário impor que W_i tenha todos os elementos da diagonal principal positivos, por outro lado, que $var(\epsilon_{it}) = \sigma^2[\mathbf{W}_i]_{tt}^2$ e que $corr(\epsilon_{it}, \epsilon_{il}) = [\mathbf{C}_i]_{tl}^2$ pelo que $\mathbf{W}_i$ descreve a variância e $\mathbf{C}_i$ descreve a correlação dos erros $\boldsymbol{\epsilon}_i$ dentro do grupo. Esta decomposição permite a flexibilidade nestas estruturas e, consequentemente, é possível definir uma estrutura de correlação e modelar a dependência. Cressie (1993) mostra que a estrutura da modelação da dependência dos erros aleatórios dentro do grupo é isotrópica. Considerando dois erros, a sua correlação depende da distância entre eles, isto é,

$$corr(\epsilon_{it}, \epsilon_{il}) = h[d(\mathbf{p}_{it}, \mathbf{p}_{il}), \boldsymbol{\rho}] \quad i = 1, \ldots, n \quad t = 1, \ldots, t_{m_i}, \tag{4.116}$$

em que ρ é um vetor de parâmetros de correlação e $h(.)$ é a função de correlação (ou função de autocorrelação) e varia entre -1 e 1 e tal que $h(0, \boldsymbol{\rho}) = 1$.

A função de autocorrelação empírica é uma estimativa não paramétrica da função $h(.)$.

Supondo que $r_{it} = (y_{it} - \hat{y}_{it})/\hat{\sigma}_{it}$, em que $\hat{\sigma}_{it}$ é o estimador da variância do erro aleatório do indivíduo i, no tempo t. Define-se a função de autocorrelação empírica no espaçamento l, dado por

$$\hat{\rho}(l) = \frac{\sum_{i=1}^{n} \sum_{t=1}^{t_{m_i}} r_{it} r_{i(t+l)}/N(l)}{\sum_{i=1}^{n} \sum_{t=1}^{t_{m_i}} r_{it}^2/N(0)}, \tag{4.117}$$

em que $N(l)$ representa o número de pares de resíduos que são utilizados no somatório do numerador da função. De notar que quando os valores se aproximam gradualmente de zero, o processo pode ser identificado como autoregressivo e, se a função supracitada for consistente no intervalo $\pm\frac{z_{1-\alpha/2}}{\sqrt{N(l)}z_{1-\alpha/2}}$, após o espaçamento 2, pode ser um possível processo de médias móveis 1 ou 2. Existem várias estruturas de correlação temporal, como a correlação não estruturada, a simetria composta e a autorregressiva.

Não estruturada

Na estrutura de correlação não estruturada é assumido um parâmetro para cada correlação entre as observações, dada por

$$h(k, \rho) = \rho_k, \qquad k = 1, 2, \ldots \tag{4.118}$$

Esta correlação é útil apenas quando se tem poucos dados.

Simetria Composta

Na estrutura de correlação simétrica composta assume-se uma correlação igual entre todos os erros do mesmo grupo, dada por

$$corr(\epsilon_{it}, \epsilon_{il}) = \rho, \quad \forall t \neq l, \quad h(k, \rho) = \rho, \quad k = 1, 2, \ldots, \tag{4.119}$$

em que ρ representa o coeficiente de correlação intragrupo e é único.

Autorregressiva

Na estrutura de correlação autorregressiva é assumido um parâmetro que depende do espaçamento, isto é,

$$\epsilon_t = \phi_1 \epsilon_{t-1} + \cdots + \phi_p \epsilon_{t-p} + a_t, \tag{4.120}$$

em que a_t é um processo de ruído branco gaussiano. Este processo denota-se por $AR(p)$, em que p é a ordem do processo, tal que $\mathbf{\Phi} = (\phi_1, \ldots, \phi_p)$.

Um exemplo é o processo autorregressivo de ordem 1, AR(1), em que os erros no tempo t são modelados em função dos erros no tempo $t-1$, tal que

$$\epsilon_t = \phi_1 \epsilon_{t-1} + a_t, \qquad |\phi| < 1. \tag{4.121}$$

A função de correlação é dada por

$$h(k, \phi) = \phi^k, \quad k = 0, 1, \ldots, \tag{4.122}$$

ou seja, decresce exponencialmente em valor absoluto com o espaçamento (*lag*). Este processo é descrito com maior detalhe no Capítulo 2.

4.2.6 Inferência Estatística

O ajustamento do modelo marginal aos dados permite a inferência dos parâmetros do modelo de modo a que os resultados obtidos possam ser generalizados para a população a partir da qual a amostra foi obtida.

Distribuição Assintótica

Considerando o modelo marginal (4.74), condicionada a α conhecido, sabe-se que o estimador $\boldsymbol{\beta}$

$$\hat{\boldsymbol{\beta}}(\alpha) = \left(\sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{X}_i \right)^{-1} \sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{Y}_i \tag{4.123}$$

tem Distribuição Normal Multivariada com valor esperado

$$E(\hat{\boldsymbol{\beta}}(\alpha)) = \left(\sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{X}_i \right)^{-1} \sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{V}_i^{-1} E[\mathbf{Y}_i] = \boldsymbol{\beta} \tag{4.124}$$

e matriz variância-covariância dada por

$$\begin{aligned} var(\hat{\boldsymbol{\beta}}(\alpha)) &= \left(\sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{X}_i \right)^{-1} \sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{V}_i^{-1} var(\mathbf{Y}_i) \mathbf{V}_i^{-1} \mathbf{X}_i \\ &\times \left(\sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{X}_i \right)^{-1} \\ &= \left(\sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{X}_i \right)^{-1}. \end{aligned} \tag{4.125}$$

No caso em que α é desconhecido, é necessário considerar a sua distribuição assintótica.

Sob os pressupostos apresentados por Pinheiro & Bates (1995, 2000), os estimadores ML são consistentes e a sua distribuição assintótica é Normal Multivariada. A inversa da matriz de informação de Fisher aproxima a matriz de variância-covariância dos estimadores (Cox & Hinkley, 1974; Pinheiro & Bates, 2000). Tendo em conta que

$$E\left[\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \theta^{\mathbf{T}}}\right] = 0 \quad E\left[\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \sigma^{\mathbf{2}}}\right] = 0, \tag{4.126}$$

os estimadores dos efeitos fixos, através de Método de Máxima Verosimilhança, não são assintoticamente correlacionados com os estimadores ML de θ e de σ^2. A sua distribuição assintótica é dada por

$$\hat{\boldsymbol{\beta}} \sim N\Big(\boldsymbol{\beta}, \sigma^2\big(\mathbf{X}^T \mathbf{M}^{-1}(\theta)\mathbf{X}\big)^{-1}\Big) \tag{4.127}$$

e

$$\begin{pmatrix} \hat{\theta} \\ log\hat{\sigma} \end{pmatrix} \sim N\left(\begin{pmatrix} \theta \\ log\sigma \end{pmatrix}, \mathbf{I}^{-1}(\theta, \sigma)\right), \tag{4.128}$$

em que $\ell(\theta, \sigma^2)$ é o logaritmo da função verosimilhança marginal dos efeitos fixos e $\mathbf{I}(\theta, \sigma)$ é a matriz empírica de Informação de Fisher e é dada por

$$\mathbf{I}(\theta, \sigma) = \begin{bmatrix} \dfrac{\partial^2 \ell(\theta, \sigma^2)}{\partial \theta \partial \theta^T} & \dfrac{\partial^2 \ell(\theta, \sigma^2)}{\partial log\sigma \partial \theta^T} \\ \dfrac{\partial^2 \ell(\theta, \sigma^2)}{\partial \theta \partial log\sigma} & \dfrac{\partial^2 \ell(\theta, \sigma^2)}{\partial^2 log\sigma} \end{bmatrix}. \tag{4.129}$$

Recorre-se a $log\ \sigma$ em vez de σ para simplificação de parametrização. Tal transformação faz com que a aproximação à Distribuição Normal seja mais evidente. À semelhança dos estimadores de Máxima Verosimilhança, os obtidos pelo REML são também consistentes e com Distribuição Assintótica Normal Multivariada. Pinheiro (1994) prova esta conclusão a partir de equações semelhantes às anteriores, com a diferença no logaritmo da função verosimilhança, a qual passa a ser restrita. Na maioria dos casos, os parâmetros são desconhecidos e procede-se à sua implementação por aproximação (4.128). Esta constatação é importante, na medida em que suporta os testes e intervalos de confiança para os efeitos fixos.

Efeitos fixos

O Teste da Razão de Verosimilhança é aplicável na comparação de modelos encaixados, apenas definindo a estrutura de efeitos fixos. A estatística de teste é dada por

$$2log\Big(\frac{L_1}{L_0}\Big) = 2(logL_1 - logL_0), \tag{4.130}$$

em que L_1 é a verosimilhança do modelo geral (com mais parâmetros) e L_0 é a verosimilhança do modelo encaixado. Considerando a qualidade do ajustamento, para os dois modelos, a estatística de teste segue assintoticamente a distribuição do Qui-Quadrado com $k_1 - k_0$ graus de liberdade, em que $k_1 - k_0$ é a diferença entre o número de parâmetros de cada modelo. O presente teste só é possível ser efetuado se os estimadores dos modelos forem obtidos a partir do Método de Máxima Verosimilhança, na medida em que o logaritmo da função verosimilhança restrita se altera se as condições dos efeitos fixos forem alterados. Dado que este teste conduz a valor de prova inferiores ao verdadeiro, esta imprecisão faz com que autores citem outros testes para avaliação da significância dos efeitos fixos, como o teste-t e o teste-F aproximado.

Através do teste-t aproximado pretende-se testar

$$H_0 : \beta_j = 0 \qquad H_1 : \beta_j \neq 0 \quad , j = 1, \ldots, p, \tag{4.131}$$

com base na estatística de teste

$$ET = \frac{\hat{\beta}_j}{\hat{\sigma}_{REML}\sqrt{\left[\left(\mathbf{X}_i^T\mathbf{M}_i^{-1}(\hat{\theta})\mathbf{X}_i\right)^{-1}\right]_{jj}}}, \tag{4.132}$$

sob hipótese nula, a estatística de teste tem distribuição assintótica à t de Student com gl_j. Este teste permite a avaliação do coeficiente do efeito fixo j, quando os restantes estão presentes no modelo.

Através do teste-F aproximado pretende-se testar

$$H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{0} \qquad H_1 : \mathbf{L}\boldsymbol{\beta} \neq \mathbf{0}, \tag{4.133}$$

com base na estatística de teste

$$ET = \frac{\hat{\boldsymbol{\beta}}^T\mathbf{L}^T\left[\mathbf{L}\hat{\sigma}_{REML}\left(\mathbf{X}_i^T\mathbf{M}_i^{-1}(\hat{\theta})\mathbf{X}_i\right)^{-1}\mathbf{L}^T\right]\mathbf{L}\hat{\boldsymbol{\beta}}}{c(\mathbf{L})}, \tag{4.134}$$

sob hipótese nula, a estatística de teste tem distribuição assintótica F de Snedcor com (l, k) graus de liberdade, em que l representa o número de graus de liberdade do numerador dado pela característica da matriz $\mathbf{L}$, $c(\mathbf{L})$. Este teste permite a avaliação dos coeficientes dos efeitos fixos, presentes no modelo.

A construção dos intervalos de confiança aproximados para os efeitos fixos é baseada nos testes-t aproximados. Considerando gl_j o número de graus de liberdade do denominador teste-t aproximado, relativamente ao efeito fixo j, ao nível de confiança $1 - \alpha$, o

intervalo de confiança para β_j é dado por

$$\left]\hat{\beta}_j - t_{\beta_j,1-\alpha/2}\hat{\sigma}_{REML}\sqrt{\left[\left(\mathbf{X}_i^T\mathbf{M}_i^{-1}(\hat{\theta})\mathbf{X}_i\right)^{-1}\right]_{jj}},\right.$$
$$\left.\hat{\beta}_j + t_{\beta_j,1-\alpha/2}\hat{\sigma}_{REML}\sqrt{\left[\left(\mathbf{X}_i^T\mathbf{M}_i^{-1}(\hat{\theta})\mathbf{X}_i\right)^{-1}\right]_{jj}}\right[, \tag{4.135}$$

em que $t_{\beta_j,1-\alpha/2}$ representa o quantil $1-\alpha/2$ da distribuição t de Student com graus de liberdade gl_j.

Componentes de Variância

Para além de estudar o comportamento médio da população, existe a necessidade de clarificar quais os efeitos aleatórios que devem ser incluídos no modelo e a estrutura de correlação a adotar.

O Teste da Razão de Verosimilhança é utilizado quando os parâmetros são estimados, usando o método REML, já que a estrutura fixa é a mesma nos dois modelos a comparar. Uma das condições exigidas é o facto da estatística de teste ter distribuição assintótica Qui-Quadrado com graus de liberdade igual à diferença entre o especificado nas hipóteses alternativa e nula. No `R` está implementada uma metodologia sugerida por Pinheiro & Bates (2000). Esta solução calcula os valores de prova de forma sobrevalorizada, o que conduz a uma análise conservativa.

O intervalo de confiança aproximado ao nível α para o desvio padrão σ é dado por

$$\left]\hat{\sigma}exp\Big\{-z_{1-\alpha/2}\sqrt{[\mathbf{I}^{-1}]_{\sigma\sigma}}, \hat{\sigma}exp\Big\{z_{1-\alpha/2}\sqrt{[\mathbf{I}^{-1}]_{\sigma\sigma}}\Big\}\right[, \tag{4.136}$$

em que $z_{1-\alpha/2}$ representa o quartil $1-\alpha/2$ da distribuição Normal padrão. Para as componentes da matriz de variância-covariância dos efeitos aleatórios, os intervalos de confiança aproximados são um pouco mais difícieis de construir e são estimados com menor precisão do que os dos efeitos fixos e o do desvio padrão dentro dos grupos. O aumento da precisão na estimação dos primeiros só é possível com o aumento do número de grupos estudados (Pinheiro & Bates, 2000).

Efeitos Aleatórios

O valor esperado do BLUP de $\mathbf{u}_i$ é dado por

$$E[\mathbf{u}_{BLUP}(\alpha)] = E[\mathbf{D}\mathbf{Z}_i^T\mathbf{V}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}})] \ = \mathbf{D}\mathbf{Z}_i^T\mathbf{V}_i^{-1}(\mathbf{X}_i\boldsymbol{\beta} - \mathbf{X}_i\hat{\boldsymbol{\beta}})] = 0 \tag{4.137}$$

e a sua variância é dada por

$$var[\mathbf{u}_{BLUP}(\alpha)] = \mathbf{D}\mathbf{Z}_i^T \left[\mathbf{V}_i^{-1} - \mathbf{V}_i^{-1}\mathbf{X}_i \left(\sum_{i=1}^{n} \mathbf{X}_i^T \mathbf{V}_i^{-1} \mathbf{X}_i \right)^{-1} \mathbf{X}_i^T \mathbf{V}_i^{-1} \right] \mathbf{Z}_i \mathbf{D}. \tag{4.138}$$

Laird & Ware (1982) provam que existe uma subestimação na variabilidade da equação anterior e propõem

$$var[\mathbf{u}_{BLUP}(\alpha) - \mathbf{u}_i] = \mathbf{D} - var[\mathbf{u}_{BLUP}(\alpha)] \tag{4.139}$$

como estimador da variação de $\mathbf{u}_{BLUP}(\alpha) - \mathbf{u}_i$. As inferências sobre os efeitos aleatórios são baseadas em aproximações de teste-t e de teste-F aproximados, com processos idênticos de estimação do número de graus de liberdade (Verbeke & Molenberghs, 2000).

4.2.7 Qualidade de Ajustamento

Para o caso dos modelos não serem aninhados, a avaliação da qualidade de ajustamento recorre às métricas definidas na Secção 4.16, nas equações (4.55) e (4.56), nas Secções 4.1.7 e 4.2.3. De notar, com base na máxima verosimilhança restrita, que só é possível a sua aplicação a modelos que tenham a mesma estrutura média. O *AIC* tende a selecionar modelos excessivamente complexos (Burnham & Anderson, 2002), o critério *BIC* por selecionar modelos excessivamente simplistas porque admite que o verdadeiro modelo é o modelo em estudo. Brewer *et al.* (2016) mostram que a otimização dos critérios para a previsão são influenciados pela dimensão da amostra (Burnham & Anderson, 2004).

4.2.8 Análise de Diagnóstico

A análise dos resíduos é uma ferramenta útil para a verificação dos pressupostos dos métodos utilizados para o ajustamento do modelo aos dados (Secção 4.17).

No caso dos Modelos de Efeitos Mistos deve-se verificar se:

- Os erros aleatórios são independentes e identicamente distribuídos, com distribuição Normal, de valor médio nulo e variância constante, e se são independentes dos efeitos aleatórios;

- Os efeitos aleatórios seguem uma distribuição Normal, com valor médio igual a zero e matriz de variância-covariância (não dependente do grupo) e são independentes para diferentes grupos.

A análise de resíduos dentro do grupo permite a validação dos pressupostos exigidos. As representações gráficas utilizadas são os resíduos padronizados *vs* os valores ajustados;

valores observados *vs* valores estimados; diagramas de extremo-quartis dos resíduos por grupo.

Capítulo 5

Aplicação aos Dados Ambientais

A água é um elemento essencial à vida e é parte integrante do planeta, desde a crosta à atmosfera. Muitas das atividades humanas e estratégias de desenvolvimento na qualidade de vida têm como impulsionador a água.

Atualmente, observa-se um conjunto de ações humanas com consequências nefastas para a Natureza, as quais obrigam a uma consciencialização por parte do indivíduo. Uma dessas ações passa pela poluição dos rios, o que gera uma séria discussão e preocupação acerca da gestão dos recursos hídricos e da imposição de normativas para a sua prevenção.

Neste contexto, este estudo recai sobre a análise da Qualidade da Água de superfície da bacia hidrográfica do rio Douro, no território português. Os dados utilizados são mensais e foram recolhidos a 31 de maio de 2019, a partir da plataforma *online* do Sistema Nacional de Informação de Recursos Hídricos (SNIRH) e correspondem a medições da precipitação e de variáveis de Qualidade da Água de superfície.

A região hidrográfica do rio Douro, designada por RH3, é transfronteiriça e encontra-se dividida entre Portugal e Espanha. A região hidrográfica tem como limites RH2, RH4 e RH5, a Este, a fronteira com Espanha, e Oeste, o oceano Atlântico. A área desta região é 79000 km^2 aproximados, em que cerca de 19 % está em território português, 18600 km^2, e os restantes 81 % estão em território espanhol, 78960 km^2 (Figuras 5.1 e 5.2).

O Decreto-Lei n.^{o} 347/2007, de 19 de outubro, define que a região hidrográfica do rio Douro é constituída pela bacia hidrográfica do rio Douro, pela bacia hidrográfica das ribeiras costeiras entre o Douro e o Vouga, incluindo as massas de águas adjacentes. A região é constituída pela bacia de Águeda, pela bacia de Côa, pelas zonas costeiras entre o Douro e o Vouga, pela bacia do Douro, pela bacia de Paiva, pelas bacias de Rabaçal/Tuela, pela bacia do Sabor, pelas bacias do Tâmega e do Tua.

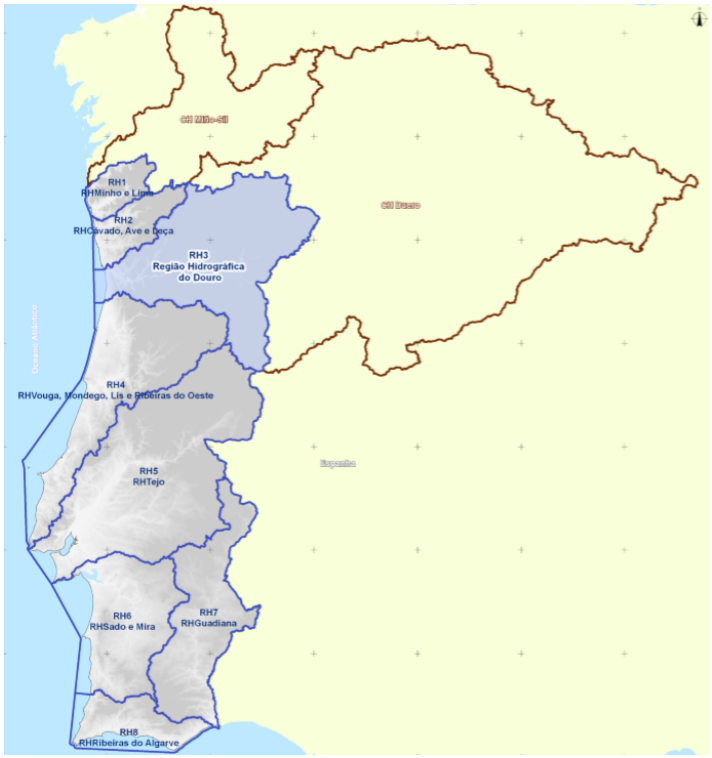


Figura 5.1: Enquadramento geográfico das regiões hidrográficas em Portugal Continental. Fonte: Relatório Técnico do PGRH-Douro, APA (mapa modificado).

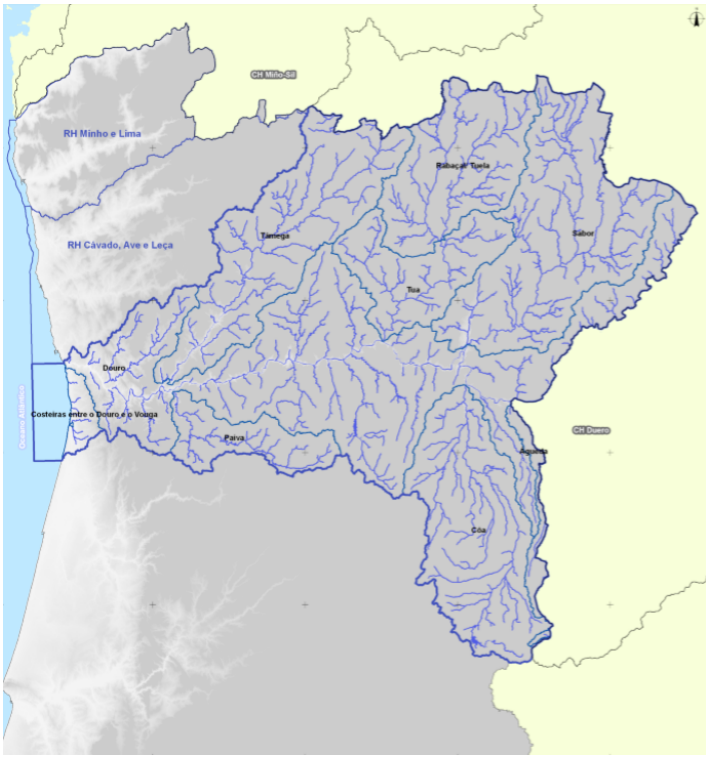


Figura 5.2: Região Hidrográfica do rio Douro (RH3). Fonte: Relatório Técnico do PGRH-Douro, APA (mapa modificado).

O rio Douro é o terceiro maior rio da Península Ibérica e a bacia do rio Douro é uma das maiores bacias em área e engloba cerca de 55 concelhos. Nasce na Serra de Úrbion, a 1700 m de altitude, e percorre aproximadamente 597 km em Espanha e 122 km, em Portugal, até à foz, no oceano Atlântico, no Porto. Os seus afluentes são muitos, nomeadamente, os rios Aguiar, Côa, Corgo, Paiva, Sabor, Sousa, Tâmega, Távora e Tua, entre outros.

A área da bacia do rio Sabor é cerca de 3297 km^2 e engloba 12 concelhos. O rio Sabor tem a sua génese em Espanha (Zamora) e percorre aproximadamente 200 km até desaguar no rio Douro, em Portugal (Bragança). A área da bacia do rio Tâmega é aproximadamente 2646 km^2 e engloba 18 concelhos. O rio Tâmega nasce em Espanha (Ourense), percorrendo cerca 150 km até desaguar no rio Douro, em Portugal (Penafiel). A área da bacia do rio Côa é cerca de 2521 km^2 e engloba 8 concelhos. O rio Côa tem a sua origem na Serra da Malcata, em Portugal, e desagua no rio Douro (perto de Vila Nova de Foz Côa). A área da bacia do rio Rabaçal/Tuela é cerca de 1867 km^2 e engloba 7 concelhos. O rio Rabaçal tem a sua génese em Espanha e percorre 65 km até ao rio Tuela, próximo de Mirandela. O rio Tuela nasce em Espanha (Castela e Leão) e passa a fronteira no concelho de Vinhais, em Portugal. A afluência do rio Rabaçal no rio Tuela dá origem ao rio Tua. Esta bacia é onde existem as menores concentrações populacionais, juntamente com a bacia de Águeda. A área da bacia do rio Tua tem 1255 km^2 e engloba 10 concelhos. O rio Tua nasce em Portugal e desagua no rio Douro, na aldeia do Tua.

A área da bacia do rio Águeda tem cerca de 248 km^2 e abrange 3 concelhos. O rio Águeda nasce na Serra de Gata, em Espanha, após percorrer 130 km e desagua no rio Douro, perto de Barca de Alva.

A menor bacia hidrográfica da RH3 é a bacia das ribeiras costeiras entre o Douro e o Vouga. A sua área é 207 km^2 e engloba 4 concelhos. Esta bacia é onde existem as maiores concentrações populacionais.

5.1 Precipitação

A precipitação é um dos componentes mais importantes no ciclo hidrológico (Chow *et al.*, 1988). Tendo em conta o aquecimento global, a estabilidade do sistema global do ciclo da água reduziu (D′ Odorico *et al.*, 2019) e existem relatos de alteração da precipitação e da precipitação extrema em todo o mundo (Trigo *et al.*, 2004; Min *et al.*, 2011; Allan & Liu, 2019).

Existem vários tipos de precipitação, nos estados líquido ou sólido (como a chuva e a neve), que dependem do processo dentro da nuvem e as temperaturas do ar entre a base da nuvem e o solo. Por exemplo, se no seu caminho até ao solo os cristais de

gelo encontrarem temperaturas positivas, os flocos de neve derretem e transformam-se em chuva. A variação da precipitação depende da região, como a exposição geográfica, a altitude, a proximidade do mar e a distribuição de zonas de pressão e de frente polar. A precipitação é uma variável que apresenta uma enorme variabilidade espacial e temporal.

A precipitação é um fenómeno espacialmente distribuído de natureza contínua que apenas é avaliado em localizações pontuais através das estações climatológicas. A medição da precipitação tem como intuito a recolha de dados sobre a quantidade de precipitação, durante um determinado intervalo de tempo, quando existe queda de água. A amostra recolhida deve ser representativa da quantidade real de precipitação, que ocorreu na região em estudo. A unidade de medida da intensidade de precipitação é o milímetro (mm) que corresponde à altura de água de 1 litro por metro quadrado.

Para determinar a intensidade de precipitação utilizam-se aparelhos para a medição de precipitação, como os udómetros (Figura 5.3). Como qualquer sistema de medição, os udómetros estão sujeitos a certos tipos de erros que afetam a fiabilidade da estimação de precipitação.

Os udómetros são constituídos por um cilindro com uma abertura superior e um recipiente de recolha de água, que originam usualmente boas estimativas pontuais. No entanto, sob fatores extremos, como precipitação intensa ou ventos fortes, as medições podem conter erros de medição. A sua localização é bastante importante, normalmente localizam-se em locais descampados e abrigados do vento, parcialmente enterrados no solo ou fixos por estruturas metálicas. A proteção do aparelho é importante para reduzir os efeitos aerodinâmicos que levam a uma subestimação ou sobreestimação da medição. Em Portugal, a rede udométrica é bastante limitada, na medida em que a cobertura de uma região por udómetros é bastante dispendiosa e de manutenção e operação complexas.

A quantificação espacial e temporal da precipitação sobre uma área geográfica, em particular, é imprescindível no cálculo dos balanços hídricos para a estimação indireta de caudais em cursos de água e para o desenvolvimento do estudo de carga dos aquíferos. É ainda essencial à modelação de diversos fenómenos ambientais, entre os quais a variação da Qualidade da Água numa bacia hidrográfica fluvial.

A Qualidade da Água, num determinado local, é o reflexo das condições dominantes da bacia de alimentação desse local, nomeadamente de fatores hidrometeorológicos. A variação espaço-temporal de uma variável de qualidade está associada à variação do caudal e esta acompanha geralmente a variação sazonal da precipitação. Daí a necessidade de estimar um fator hidrometeorológico (via precipitação) nas estações de amostragem de Qualidade da Água na bacia do rio Douro, onde não há nem valores de caudais nem valores de precipitação.

A bacia hidrográfica do rio Douro é monitorizada por várias estações de amostragem

Figura 5.3: Esquerda: Udómetro, utilizado para medir a precipitação total que caiu num determinado período de tempo, 28 de setembro de 1992, em Vale dos Camelos (DSRH/I-NAG). Direita: Udómetro entupido e com água acumulada da Estação Automática sem telemetria, 28 de dezembro de 2012, em Laranjal, Ponte Sôr (DMSDIH/APA).

de precipitação distribuídas pelo rio Douro e pelos seus principais afluentes. No entanto, só foram consideradas 18 estações de amostragem devido à enorme falta de dados (dados omissos) em grande parte delas. As estações selecionadas apresentam uma percentagem de dados omissos inferior a cerca de 20 %. Nas Tabelas 5.1 e 5.2 apresentam-se as características e o período de observação destas estações de amostragem e a sua localização na Figura 5.4. O acesso aos dados mensais foi realizado a partir do Sistema Nacional de Informação de Recursos Hídricos (SNIRH).

5.1.1 Análise Descritiva

A Tabela 5.2 sintetiza algumas estatísticas descritivas da precipitação no período observado. É possível verificar que quanto maior a média, maior é o valor do desvio padrão. É também visível que o valor mínimo é zero em todas as estações, o que significa que existiu pelo menos um mês em que não ocorreu precipitação em cada estação de amostragem.

O maior valor de precipitação observado ($408,50\ mm$) é identificado na estação de amostragem de Santa Marta da Montanha, em novembro de 2002. A estação de amostragem de Vilar Formoso é a que apresenta menor desvio padrão ($31,74\ mm$) e menor coeficiente de variação (84 %). A estação de amostragem de Santa Maria da Montanha é a que apresenta maior desvio padrão ($93,34\ mm$). A estação de amostragem de Fonte Longa é a que revela maior coeficiente de variação (112 %). No Apêndice A procedeu-se à representação gráfica da evolução da precipitação ao longo do tempo observado em cada

Tabela 5.1: As estações de amostragem da precipitação, na Bacia Hidrográfica do rio Douro, e o respetivo período observado.

	Código	Nome	Latitude (^oN)	Longitude (^oO)	Período Observado
1	10O/02UG	Almeidinha	40,59400	-7,12900	12/2003 - 1/2013
2	04R/01UG	Argozelo	41,63800	-6,60100	1/2003 - 2/2013
3	07O/05UG	Castelo Melhor	41,01546	-7,06808	3/2002 - 11/2009
4	08J/06G	Castro Daire (Lamelas)	40,92400	-7,94000	3/2002 - 9/2011
5	02R/02G	Deilão	41,84700	-6,58600	10/2002 - 2/2013
6	08P/02G	Escalhão	40,94800	-6,92400	3/2002 - 2/2013
7	06N/03UG	Fonte Longa	41,23200	-7,26800	3/2002 - 10/2012
8	08I/01UG	Mosteiro de Cabril	40,94700	-8,10000	3/2002 - 2/2012
9	04R/02G	Pinelo	41,63500	-6,55200	3/2002 - 2/2013
10	09O/01G	Pinhel	40,77100	-7,06100	3/2002 - 2/2013
11	04K/02G	Santa Maria da Montanha	41,50075	-7,74599	3/2002 - 9/2011
12	04O/01G	Torre de Dona Chama	41,65654	-7,11589	3/2002 - 9/2010
13	08K/01UG	Touro	40,89700	-7,74800	3/2002 - 12/2009
14	03N/01G	Travancas	41,82797	-7,30561	3/2002 - 6/2012
15	05M/04UG	Vales (Valpaços)	41,46547	-7,35143	3/2002 - 5/2012
16	05L/01UG	Vila Pouca de Aguiar	41,49806	-7,63600	6/2003 - 5/2012
17	10Q/01UG	Vilar Formoso	40,60900	-6,83100	3/2002 - 5/2012
18	02O/02UG	Vinhais	41,82798	-6,99384	11/2003 - 2/2013

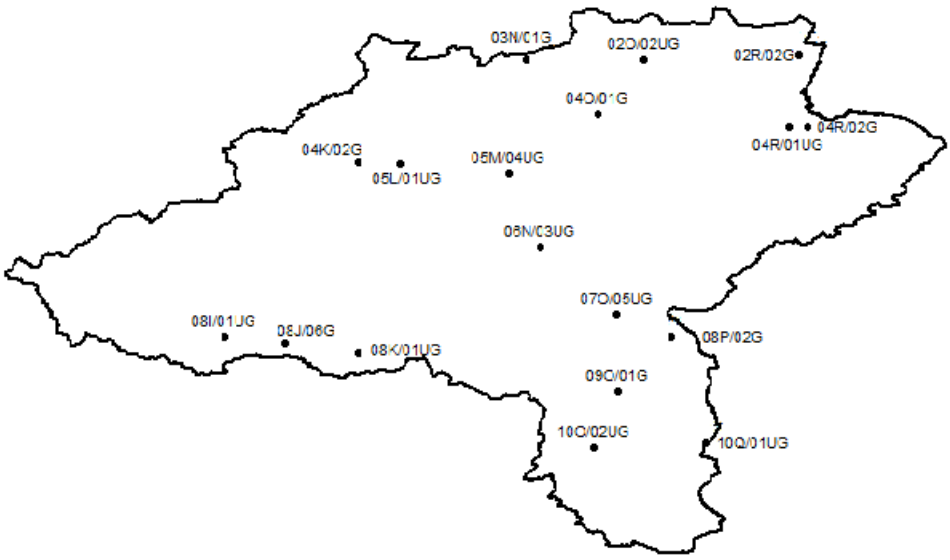


Figura 5.4: Representação da bacia hidrográfica do rio Douro e localizações das estações de medição de precipitação.

estação de amostragem, bem como os diagramas em caixa de bigodes e os histogramas.

Pode observar-se que existem estações com *outliers*, valores discrepantes que representam meses com grande intensidade de precipitação. Estas observações não são excluídas, na medida em que mostram o comportamento extremo de precipitação que foi atingido.

Tabela 5.2: Medidas descritivas da variável da precipitação, no período observado.

Código	Número	Dados Omissos	Mínimo (mm)	Máximo (mm)	1.º Quartil (mm)	Mediana (mm)	3.º Quartil (mm)	Média (mml)	Desvio Padrão (mm)	Coeficiente de Variação (%)
10O/02UG	110	24	0,00	196,00	11,70	32,15	54,88	41,30	39,96	0,97
04R/01UG	98	36	0,00	219,70	11,15	34,25	66,45	46,99	46,17	0,98
07O/05UG	95	39	0,00	157,40	11,40	25,70	44,00	34,87	33,88	0,97
08J/06G	107	27	0,00	366,90	25,20	63,00	139,10	93,25	91,34	0,98
02R/02G	99	35	0,00	227,20	12,35	33,60	69,55	49,49	48,08	0,97
08P/02G	126	8	0,00	239,00	14,50	33,15	59,35	42,84	39,41	0,92
06N/03UG	106	28	0,00	233,00	7,50	26,65	51,45	39,57	44,42	1,12
08I/01UG	102	32	0,00	329,80	23,98	52,30	107,40	78,20	74,76	0,96
04R/02G	124	10	0,00	201,00	7,30	29,30	58,20	43,53	47,44	1,09
09O/01G	130	4	0,00	166,20	7,85	25,50	52,92	35,68	36,40	1,02
04K/02G	105	29	0,00	408,50	23,50	57,50	129,90	92,98	93,57	1,01
04O/01G	104	30	0,00	139,70	12,55	26,60	49,95	36,63	33,30	0,91
08K/01UG	95	39	0,00	223,10	12,30	28,40	70,95	47,73	48,82	1,02
03N/01G	103	31	0,00	269,00	13,90	37,20	74,95	56,08	59,08	1,05
05M/04UG	109	25	0,00	172,90	7,80	23,70	48,10	36,26	39,31	1,08
05L/01UG	96	38	0,00	354,60	16,85	40,65	100,55	67,21	71,17	1,06
10Q/01UG	112	22	0,00	128,00	13,75	30,65	50,70	37,72	31,74	0,84
02O/02UG	109	25	0,00	211,20	11,40	29,40	56,60	41,72	44,40	1,06

5.1.2 Análise da Continuidade Espacial

No presente estudo existe a necessidade de dispor de medições mensais médias em área, dependentes da quantidade de precipitação que cai sobre uma certa região geográfica, que influenciam um determinado local da bacia hidrográficado rio Douro e que funcionam como um fator hidrometeorológico na modelação (espacial e temporal) da Qualidade da Água. Pretende-se a identificação de modelos que possibilitem a obtenção dessas medições, construindo a distribuição espacial da precipitação, na bacia do rio Douro, em locais onde não há valores observados (nas estações de amostragem de Qualidade, onde foram efetuadas medições das variáveis de Qualidade da Água de superfície).

Foi realizado o estudo dos dados de precipitação espaciais através de gráficos que resumem as observações de acordo com os seus principais atributos (gráfico precipitação/Latitude e gráfico Longitude/precipitação). A análise da variabilidade espacial da precipitação é realizada separadamente para cada mês do ano e foi necessário transformar as coordenadas devido à presença de uma componente de tendência polinomial. Após a transformação, as observações transformadas apresentam-se já de forma aleatória e os gráficos também já indicam estacionaridade na média e distribuições gaussianas da precipitação mensal, ao longo dos anos observados.

Falta verificar a condição de isotropia dos processos subjacentes à precipitação. A isotropia verifica-se quando os valores do processo estocástico dependem apenas do módulo do vetor da distância entre eles, não podendo depender da direção angular desses mesmos vetores. Foram calculados os variogramas direcionais e estes apresentam uma forma semelhante de conduta, dependendo da direção angular, até aproximadamente à distân-

cia máxima considerada. As direções 0^o e 45^o têm valores de semicovariância um pouco superiores a zero e as direções 90^o e 135^o têm valor aproximadamente de um. Os dados expõem-se de forma aleatória, na superfície em estudo; nos variogramas direcionais, a magnitude da semicovariância é bastante reduzida, comparativamente ao variograma com tendência na média. Considera-se que o pressuposto de isotropia é cumprido. Assumem-se as hipóteses de homogeneidade do processo (na região em estudo, o processo é estacionário de segunda ordem, logo intrinsecamente estacionário) e a isotropia. Os dados são analisados através de diferentes medidas de continuidade espacial (o covariograma, o semivariograma e o correlograma empíricos).

Modelação da Continuidade Espacial

O variograma empírico é bastante usado na análise exploratória de dados, para além de poder ser utilizado para descobrir o modelo de correlação espacial. Para a construção de um variograma definem-se duas regras muito utilizadas na prática: a distância máxima é considerada cerca de 60 % da distância máxima das localizações e cada ponto do semivariograma tem que ser calculado através de, pelo menos, 30 pares de observações.

A noção de continuidade espacial advém da hipótese de que as observações próximas no espaço tendem a ser mais semelhantes do que as mais afastadas. Para a análise desta associação entre as medições existem diversos métodos como o correlagrama, o covariograma e o semivariograma. No presente trabalho recorre-se ao semivariograma empírico determinado pelas equações mencionadas no Capítulo 2, supondo que cada medição ao longo do tempo são réplicas independentes do mesmo processo.

De notar que a estimação do variograma pode ser prejudicada pela presença de *outliers*, mas opta-se por manter os *outliers*. O variograma de Matheron estima os valores com base no cálculo de mais de 30 pares de pontos. Após a determinação do variograma empírico para cada mês do ano, o objetivo é ajustar o modelo teórico que mais se adeque. Esta etapa consiste em estimar os parâmetros desconhecidos dos modelos teóricos. No ajustamento das observações aos modelos de transição foram considerados o Modelo Exponencial, o Modelo Esférico e o Modelo Gaussiano. Apresentam-se graficamente o semivariograma estimado, bem como os ajustamentos para cada um dos modelos de transição anteriores (Figuras 5.5 e 5.6).

O ajustamento de todos os modelos de transição considerados foi efetuado pelo Método dos Mínimos Quadrados.

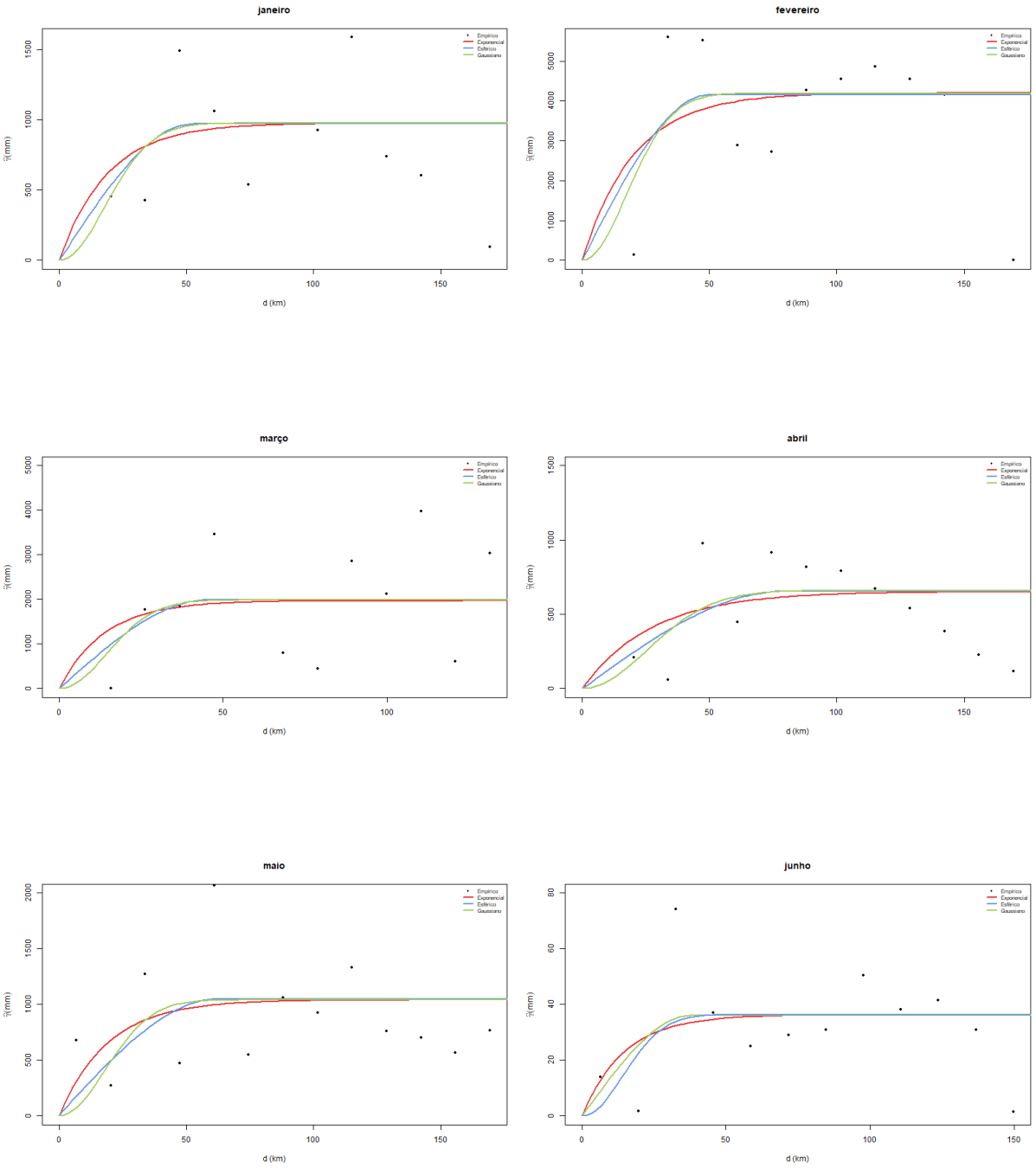


Figura 5.5: Representações gráficas dos semivariogramas empíricos e dos ajustamentos aos modelos teóricos, pelo Método dos Mínimos Quadrados.

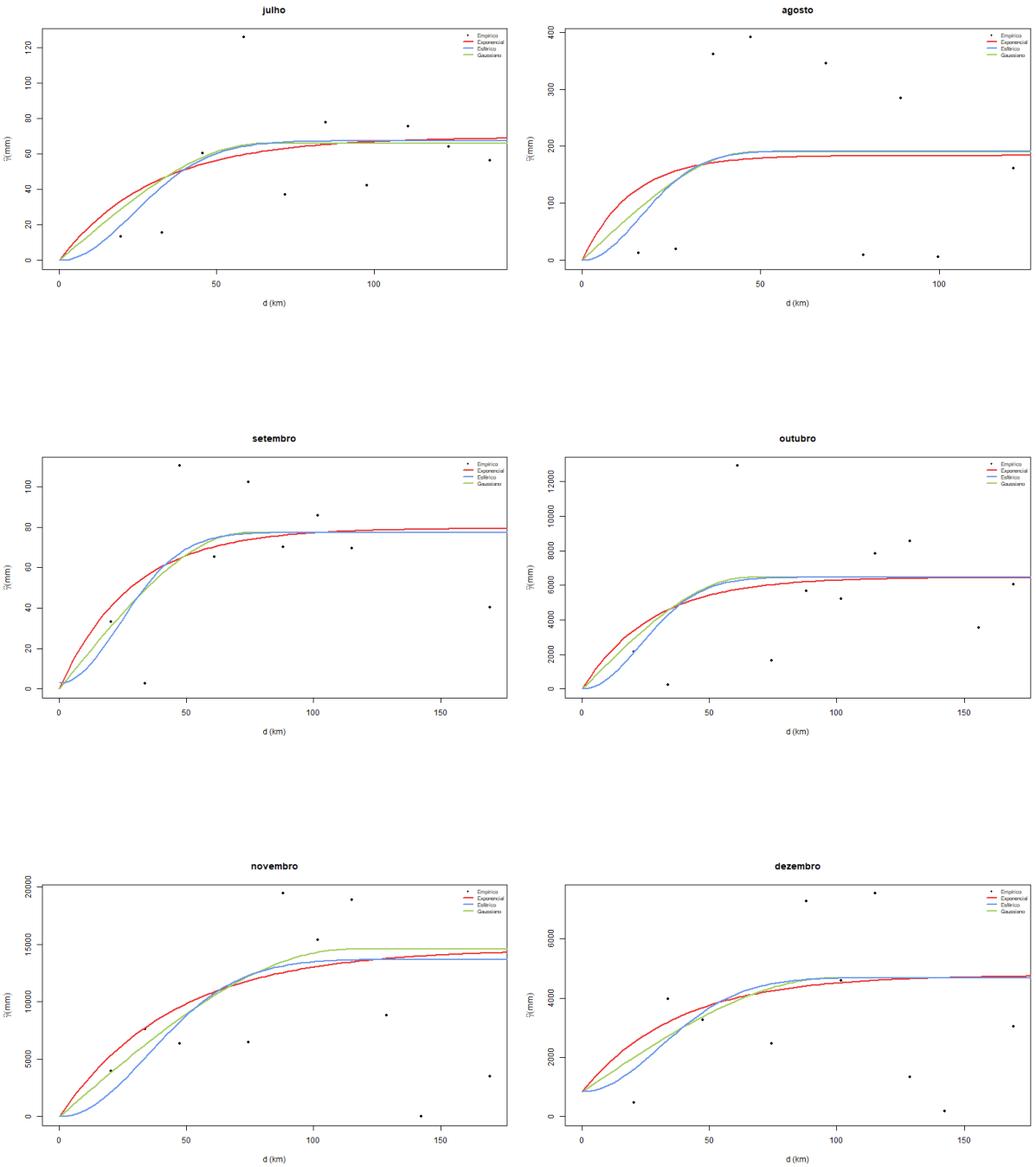


Figura 5.6: Representações gráficas dos semivariogramas empíricos e dos ajustamentos aos modelos teóricos, pelo Método dos Mínimos Quadrados.

Tabela 5.3: Valores estimados para cada semivariograma e métricas utilizadas para a Validação Cruzada.

Mês	Modelo	$\hat{C}_0$	$\hat{C}_1$	$\hat{C}_2$	$\hat{MEP}$	$\hat{EQMN}$
janeiro	Exponencial	2694,0494	3755,9527	1,4050	-0,7073	0,9935
	Esférico	3188,9815	808,3802	0,0065	15,9762	0,9976
	Gaussiano	2969,6354	31464,2498	3,8027	-2,7856	0,9937
fevereiro	Exponencial	1680,5434	1713681,7978	1901,9569	-2,3646	0,9947
	Esférico	2103,4303	130,2867	0,0100	2,3457	0,9940
	Gaussiano	1811,9229	1685,9462	1,2240	-4,1105	0,9955
março	Exponencial	1623,2178	2662,1474	2,0588	0,1131	0,9937
	Esférico	1624,4941	1924,5520	2,5823	-0,0198	0,9937
	Gaussiano	1759,6503	1759,6503	1,0821	-0,7046	0,9938
abril	Exponencial	1414,0899	1949226,7488	2771,4596	-0,0128	0,9937
	Esférico	343,8506	1493,8159	0,0000	7,3619	0,9959
	Gaussiano	1500,5562	1364,5816	1,2145	0,2645	0,9937
maio	Exponencial	772,1211	350925,3785	2283,3108	-0,3200	0,9940
	Esférico	801,4512	68,2032	0,0096	-0,0286	0,9940
	Gaussiano	795,5256	65286,5162	20,7456	-0,6595	0,9940
junho	Exponencial	540,7291	282418,4835	2827,5728	-1,1848	0,9944
	Esférico	601,1948	0,1673	0,0100	-2,1623	0,9946
	Gaussiano	552,1412	22443,4841	14,4906	-1,3728	0,9943
julho	Exponencial	156,6149	87427,6303	3284,5601	0,5631	0,9942
	Esférico	156,8556	17,0238	0,0100	0,6902	0,9944
	Gaussiano	156,7661	17,1760	0,0430	0,7238	0,9945
agosto	Exponencial	1369,6930	1194,9599	2,5556	1,6822	0,9944
	Esférico	1424,5478	203,8984	0,0071	1,7706	0,9945
	Gaussiano	1440,1172	116374,8005	17,4713	2,1366	0,9944
setembro	Exponencial	853,8326	386767,6486	2267,6718	-0,9458	0,9940
	Esférico	167,0694	789,8636	0,0010	-1,1823	0,9939
	Gaussiano	742,3707	214,5623	0,0010	-1,1823	0,9939
outubro	Exponencial	3967,3306	2333780,5287	1058,1762	2,6941	0,9941
	Esférico	4961,8748	356,8896	0,0105	11,1138	0,9971
	Gaussiano	4253,6523	384908,2541	12,9947	2,0860	0,9937
novembro	Exponencial	4267,3646	2065591,9849	1455,6130	1,8141	0,9937
	Esférico	4818,1061	324,7365	0,0103	8,7028	0,9963
	Gaussiano	4481,9407	157796,2363	10,6039	-0,5441	0,9936
dezembro	Exponencial	3280,7657	3133396,5178	1620,6543	-1,5792	0,9935
	Esférico	3291,6828	1231,9447	0,0007	3,1170	0,9943
	Gaussiano	3484,0075	2622,0005	0,9267	-2,0149	0,9935

Validação Cruzada

Na presença de vários modelos teóricos para ajustar um variograma teórico aos dados empíricos, o Método de Validação Cruzada permite diagnosticar eventuais problemas com o ajustamento dos modelos escolhidos, indicando qual será o mais adequado. Assim sendo, para avaliar a qualidade de ajustamento de cada um dos variogramas obtidos, analisam-se os valores da média dos erros de predição (MEP) e do erro quadrático médio normalizado ($EQMN$). Os dois primeiros valores deverão ser aproximadamente zero e o último deverá ser aproximadamente um.

Por exemplo, no mês de janeiro, na Tabela 5.3, verifica-se que no modelo exponencial, a medida de avaliação MEP é próximo de zero e o $EQMN$ é próximo de um, relativamente aos restantes modelos. Com base nestas métricas, o modelo teórico que melhor se ajusta ao variograma empírico é o Modelo Exponencial, através do Método dos Mínimos Quadrados, para o mês de janeiro. Nos restantes meses, o Modelo Exponencial, é também o modelo teórico que melhor se ajusta ao variograma empírico. Assim, opta-se pelo Modelo Exponencial para a modelação espacial, em cada mês.

5.1.3 Predição Pontual e Global

Na predição global (em área), o processo que se segue diz respeito à previsão da precipitação numa certa área geográfica. Uma vez que a predição se refere a inferências sobre valores do processo estocástico que não foram observados, após a escolha do modelo teórico que melhor se ajusta ao variograma empírico, será aplicado o modelo *Kriging* sobre uma grelha de 200 por 200 (aproximadamente 40000 pontos), delimitada pela fronteira de $RH3$. Considerando a Latitude e a Longitude onde se localiza a Região Hidrográfica 3, sabe-se que 1 grau de Longitude é aproximadamente 1,11 km, ou seja, o espaçamento entre dois pontos da grelha é aproximadamente $1,11$ km.

Na aplicação da técnica de *Kriging* é necessário ter em conta o modelo teórico escolhido (neste estudo, o Exponencial), as estimativas dos valores dos parâmetros e o tipo de *Kriging* utilizado (neste caso, o Universal). O ajustamento do variograma empírico e a interpolação realizada pelo método de *Kriging* Universal possibilitaram a construção de dois gráficos, os quais indicam os valores estimados da precipitação na área da bacia hidrográfica e os desvios padrão associados a essas estimativas.

Os resultados da aplicação do método de *Kriging* Universal em área são apresentados para o mês de janeiro na Figura 5.7 e para os restantes meses nas Figuras A.11 a A.21 (Apêndice A) e os desvios padrão das estimativas da precipitação na bacia hidrográfica do rio Douro são apresentados para o janeiro na Figura 5.8 e para os restantes meses nas Figuras A.22 a A.32 (Apêndice A).

Estes gráficos sugerem que os níveis de precipitação estimados mais baixos se encontram na zona Noroeste, assinalada pela cor verde e, à medida que se avança para Este, surgem níveis de precipitação mais elevados, estando os valores máximos assinalados pela cor cinzenta. As linhas desenhadas a preto representam as localidades "contornadas" pelas estimativas do mesmo nível, ou seja, as zonas interiores às linhas de contorno assumem estimativas de concentração de precipitação na ordem do valor apresentado nessa mesma linha. As superfícies são bastante semelhantes e a estimação parece bem conseguida.

O processo de *Kriging* utilizado permitiu, também, o mapeamento dos desvios padrão, resultando na representação dos pontos verde como sendo as zonas de menor desvio padrão, as quais correspondem às zonas onde foram amostrados mais pontos.

Note-se que a existência de desvios padrão relativamente mais elevados pode dever-se ao facto de existirem observações mais extremas, que resultaram numa estimação menos precisa e, consequentemente, num erro de estimação maior.

O objetivo final da análise espacial efetuada foi a estimação pontual da precipitação, por mês, nas 36 estações de amostragem de Qualidade da Água de superfície da bacia hidrográfica do rio Douro. Estes valores, em número elevado, não são apresentados na dissertação, pois são estimativas mensais para os anos de 2002 até 2013.

O processo de estimação para a precipitação deu origem a uma nova "variável", CH_t, que representa o fator hidrometeorológico e pretende "traduzir" o valor aproximado da medida de um "volume médio" de água que passa, por mês, na secção de determinada estação de amostragem de Qualidade da Água de superfície da bacia hidrográfica do rio Douro, e que será utilizado nos instantes t, $t-1$ e $t-2$ (isto é, nos dois meses antes, no mês anterior e no próprio mês). Consideram-se os três instantes de tempo, porque existem estudos que comprovam que os meses anteriores em que houve precipitação têm maior influência na Qualidade da Água do que o próprio mês.

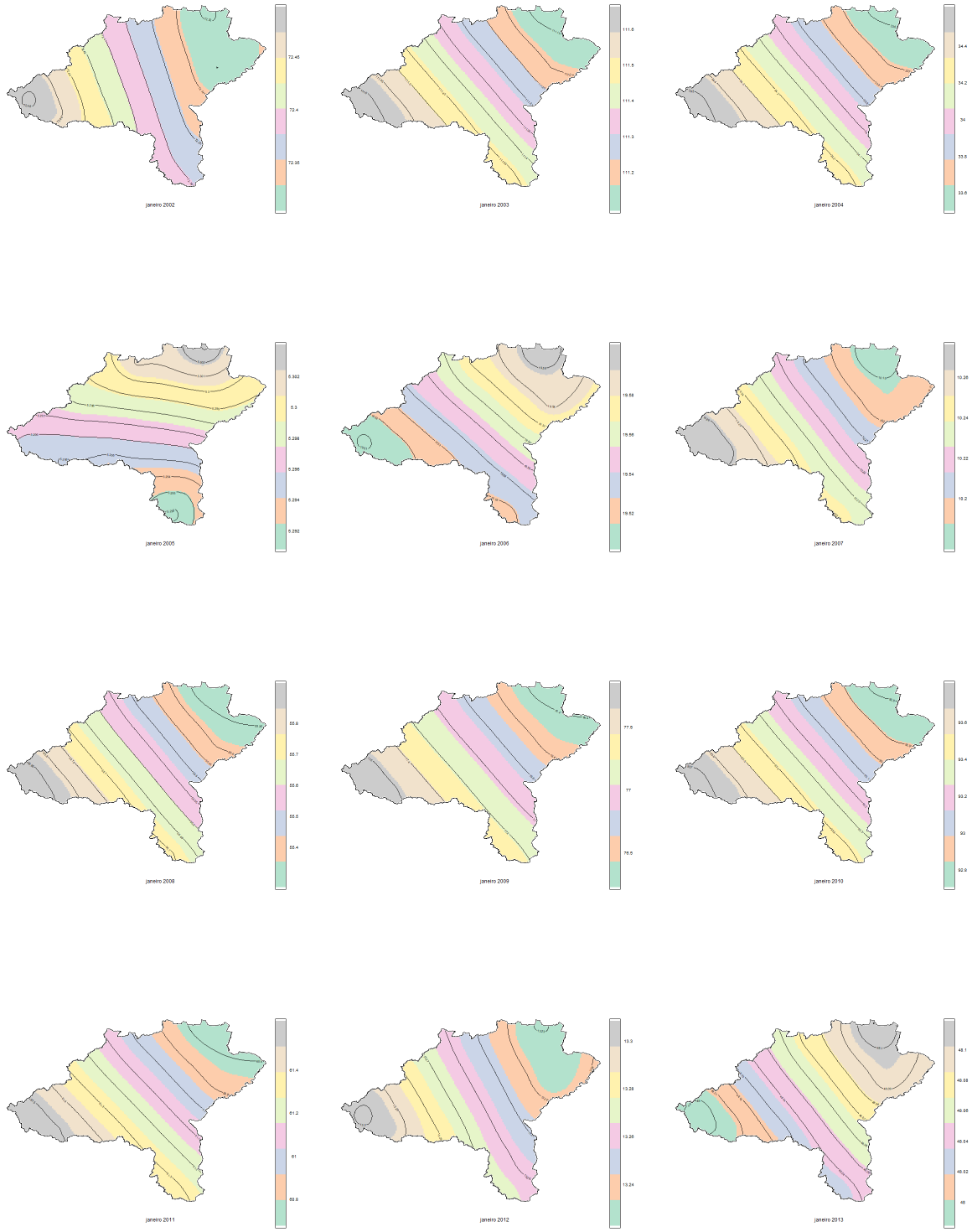


Figura 5.7: Mapa das superfícies estimadas de precipitação para o mês de janeiro, nos anos de 2002 até 2013, através da metodologia *Kriging* Universal.

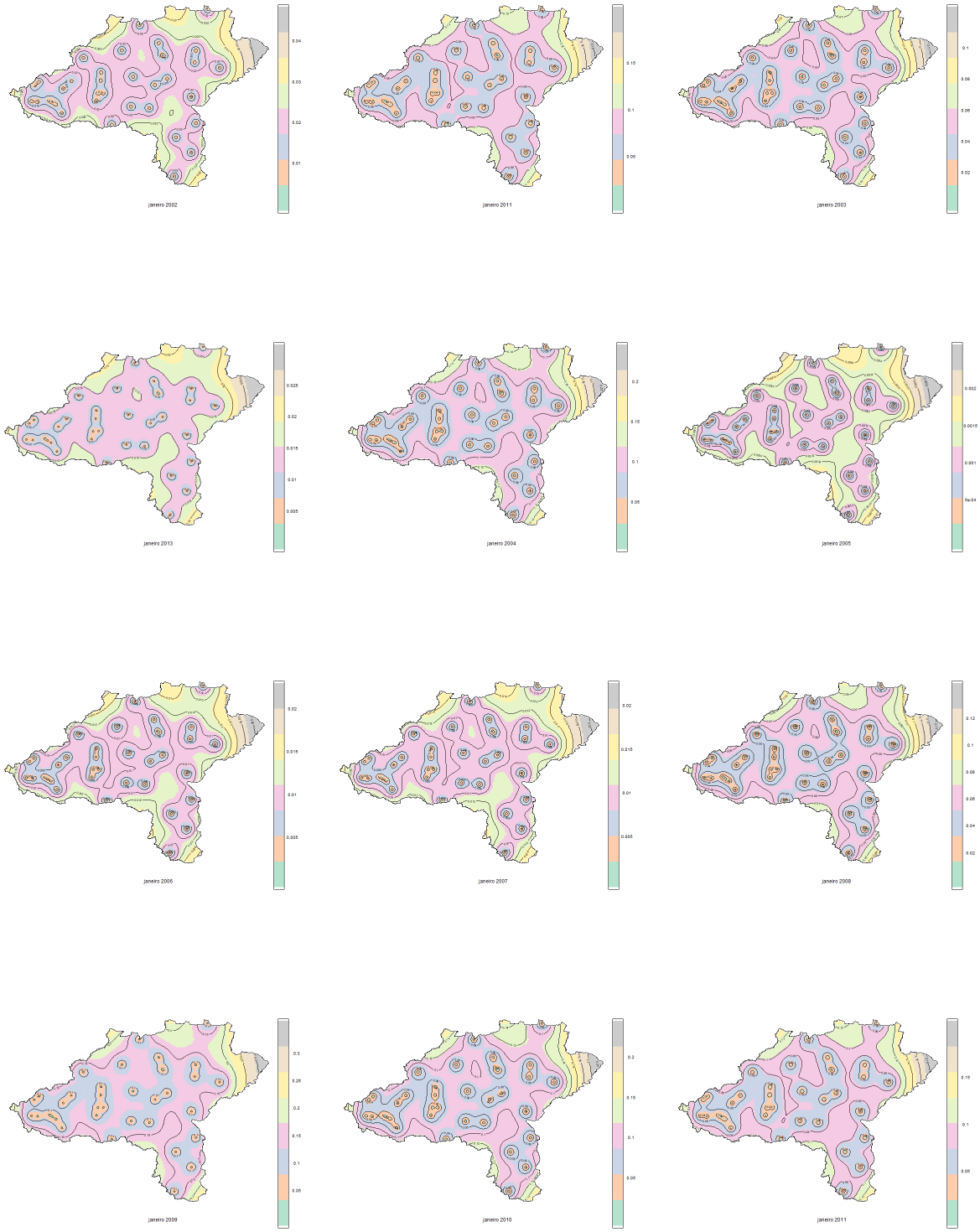


Figura 5.8: Mapa das superfícies dos desvio padrão estimados de precipitação para o mês de janeiro, nos anos de 2002 até 2013, através da metodologia *Kriging* Universal.

5.2 Qualidade da Água

A bacia hidrográfica do rio Douro é monitorizada por várias estações de amostragem de Qualidade da Água de superfície distribuídas pelo rio Douro e pelos seus principais afluentes. As estações selecionadas apresentam uma percentagem de dados omissos inferior a 20 %. Na recolha de uma amostra de água podem ser analisadas diversas variáveis analíticas, físico-químicas e microbiológicas para que seja possível uma caracterização global da Qualidade da Água e respetiva evolução. Para a recolha das variáveis, utilizam-se aparelhos para a sua medição de Qualidade da Água, como as sondas. À semelhança dos udómetros, as sondas constituem sistemas de medição, as quais estão sujeitas a determinados erros.

Figura 5.9: Esquerda: Estação hidro-qualidade automática, de dezembro de 2003, em Ermida-Corgo (DSRH/INAG). Direita: Sonda de Qualidade da Água e de nível danificada (DSRH/INAG).

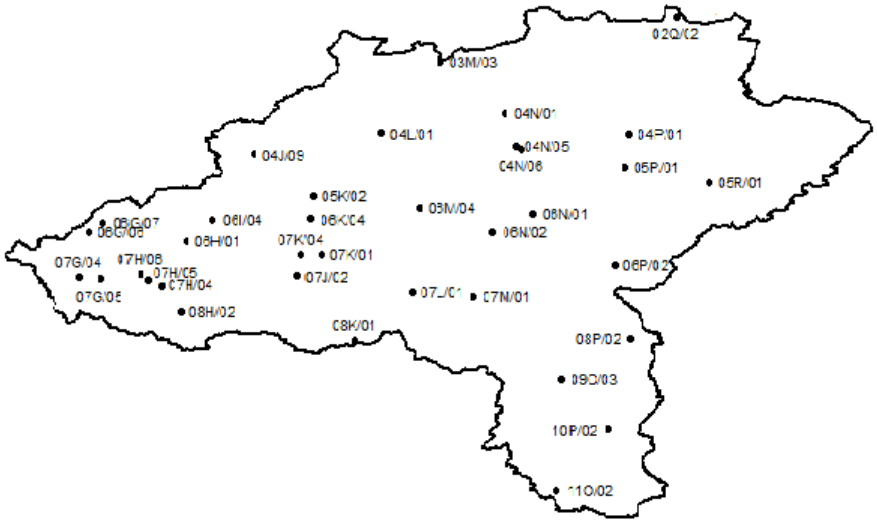


Figura 5.10: Representação da Região Hidrográfica do Douro e as localizações das estações de medição de Qualidade da Água selecionadas para o estudo.

5.2.1 Base de Dados

Na área da bacia hidrográfica do rio Douro existe uma rede de monitorização de Qualidade da Água, onde foram selecionadas 36 estações de amostragem cujas características e período de observação se encontram nas Tabelas 5.4 e 5.6 e a sua localização na Figura 5.10. O acesso aos dados mensais foi realizado a partir do Sistema Nacional de Informação de Recursos Hídricos (SNIRH).

Tabela 5.4: Localizações das estações de amostragem de Qualidade da Água e respetivo período observado, na bacia hidrográfica do rio Douro.

	Código	Nome	Latitude (^oN)	Longitude (^oO)	Período observado
1	05K/02	Alb. Alvão - V. Real	41,35098	-7,79780	3/2002 - 2/2013
2	04P/01	Alb. Azibo	41,56043	-6,88966	3/2002 - 2/2013
3	05R/01	Alb. Bastelo	41,39464	-6,66014	3/2002 - 2/2013
4	05P/01	Alb. Camba	41,44596	-6,90104	3/2002 - 2/2013
5	07G/04	Alb. Crestuma-Lever	41,07223	-8,46692	3/2002 - 2/2013
6	06N/02	Alb. Fonte Longa	41,22931	-7,28058	3/2002 - 2/2013
7	06N/01	Alb. Peneireiro	41,28956	-7,16376	4/2002 - 2/2013
8	10P/02	Alb. Porto São Miguel	40,58094	-6,93188	3/2002 - 2/2013
9	07M/01	Alb. Ranhados	41,00804	-7,33498	3/2002 - 2/2013
10	11O/02	Alb. Sabugal	40,35032	-7,09565	3/2002 - 2/2013
11	08P/02	Alb. Santa Maria Aguiar	40,86371	-6,88765	3/2002 - 2/2013
12	02Q/02	Alb. Serra Serrada	41,95846	-6,77515	3/2002 - 2/2013
13	06K/04	Alb. Sordo	41,27101	-7,80244	3/2002 - 2/2013
14	06H/01	Alb. Torrão (Semealho)	41,19505	-8,16060	3/2002 - 2/2013
15	06P/02	Alb. Vale Ferreiro	41,11584	-6,93167	3/2002 - 2/2013
16	09O/03	Alb. Vascoveiro	40,72751	-7,08067	3/2002 - 2/2013
17	06M/04	Alb. Vila Chã	41,30875	-7,49100	3/2002 - 2/2013
18	08K/01	Azenha	40,84923	-7,69649	3/2003 - 2/2013
19	03M/03	Aç. Vila Verde da Raia	41,80590	-7,42888	3/2003 - 2/2013
20	07H/05	Castelo (Alb. Crestuma)	41,06471	-8,26708	3/2002 - 2/2013
21	04N/05	Eixes	41,51875	-7,21356	3/2002 - 2/2013
22	07K/01	Foz Corgo	41,15277	-7,77453	3/2002 - 2/2013
23	07H/06	Foz Tâmega (Alb. Crestuma)	41,08212	-8,29317	3/2002 - 2/2013
24	08H/02	Fragas Torre	40,95736	-8,17367	3/2002 - 2/2013
25	07G/05	Melres	41,06709	-8,40614	3/2002 - 2/2013
26	06G/07	Modelos	41,25970	-8,39531	3/2002 - 2/2013
27	07L/01	Moinho Ponte Nova	41,02345	-7,51167	3/2002 - 2/2013
28	07K/04	Moledo	41,15150	-7,82934	3/2002 - 2/2013
29	04L/01	Pedras	41,56202	-7,59963	3/2002 - 2/2013
30	07J/02	Penude	41,07676	-7,84331	3/2002 - 2/2013
31	07H/04	Ponte Bateira	41,04268	-8,23123	3/2002 - 2/2013
32	04N/01	Ponte Vale Telhas	41,63213	-7,24665	3/2002 - 2/2013
33	06I/04	Praia Aurora (Alb. Torrão)	41,26820	-8,08180	3/2002 - 2/2013
34	04N/06	Quinta Maravilha	41,50939	-7,19766	3/2002 - 2/2013
35	06G/06	Souto	41,22645	-8,43941	3/2002 - 2/2013
36	04J/09	Vau	41,49102	-7,96709	3/2002 - 2/2013

O objetivo principal deste estudo é o estabelecimento de modelos na área de Modelos de Efeitos Mistos para explicar o comportamento da variável OD, no período observado.

Na Qualidade da Água, a variável que representa o Oxigénio Dissolvido (OD), medido em mg/l, é uma das variáveis consideradas como fulcral na verificação do nível de poluição presente num recurso hídrico. Esta poluição orgânica deriva dos processos de transformação do curso de água, nomeadamente, a oxidação da matéria orgânica, a fotossíntese e a respiração. Entre duas medições de quantidade de OD, a amostra menos poluída é aquela que apresenta o valor maior na quantidade de Oxigénio Dissolvido.

Na construção dos modelos são consideradas covariáveis qualitativas (EST, ALB, pH, Mês) e covariáveis quantitativas ($CBO5$, CH, $Clorofila$, $Temperatura$ e $Tempo_{it}$), Tabela 5.5. As covariáveis de Qualidade da Água pH, $CBO5$, CH, $Clorofila$ e $Temperatura$ foram selecionadas com base em estudos anteriores (Ho *et al.*, 2018; Moshogianis, 2015), que concluiram que estão associadas, isto é, influenciam o comportamento do OD.

Tabela 5.5: Variáveis em estudo e respetiva descrição.

Variável	Descrição	Classificação	Níveis/Unidades
EST	Identificação da estação de amostragem	Qualitativa nominal	05K/02, ..., 04J/09
ALB	Identificação das albufeiras	Qualitativa nominal	0 - não é albufeira, 1 - é albufeira
$CBO5$	Carência Bioquímica de Oxigénio aos 5 dias	Quantitativa	mg/l
CH	Fator Hidrometeorológico	Qantitativa	mm
$Clorofila$	Clorofila A	Quantitativa	ug/l
pH	Especificação sobre a acidez ou basicidade de uma solução aquosa	Quantitativa	0, ..., 14
$Temperatura$	Temperatura da amostra	Quantitativa	^oC
$Tempo$	Tempo em que é medida a observação t, na estação de amostragem i	Quantitativa	1, ..., 132
OD	Oxigénio Dissolvido	Quantitativa	mg/l

A variável EST identifica a estação de amostragem em estudo, com 36 níveis (segunda coluna da Tabela 5.4). A variável ALB identifica se a estação de amostragem em estudo está localizada numa albufeira ou não. A albufeira usualmente é uma área coberta de água retida pela construção de uma represa ou barragem num rio, resultando num lago artificial, com forte tendência para a sedimentação elevada e poucas correntes.

A variável $CBO5$ representa a medição de carência Bioquímica de Oxigénio aos 5 dias, em mg/l. É uma das variáveis mais utilizadas na avaliação do grau de poluição de águas de superfície, uma vez que um valor elevado de carência pode ser indício da existência de uma quantidade elevada de matéria orgância que utiliza o Oxigénio Dissolvido na

água para a sua decomposição. Em caso extremo, a falta de Oxigénio Dissolvido na água origina a morte da vida aquática e a eutrofização da água (crescimento excessivo de plantas aquáticas).

A variável CH representa o fator hidrometeorológico, determinado na Secção 5.1 e que é considerada nos tempos $t-2$, $t-1$ e t. Já foi mencionada a importância deste fator no processo de modelação.

A variável $Clorofila$ representa a quantidade de clorofila A, pigmentos fotossintéticos presentes nos cloroplastos das plantas grupo A, em ug/l. A concentração da clorofila é um indicador do estado trófico da água, o que faz com que seja possível o cálculo da biomassa e da atividade fotossintética potencial de microorganismos, como as cianobactérias.

A variável pH é o valor da escala numérica adimensional utilizada para especificar a acidez ou basicidade de uma solução aquosa. A variável química pH da água depende da sua origem e características naturais, mas é influenciada pela introdução de resíduos. Um valor de pH baixo significa que a água é corrosiva e um valor elevado leva a que existam incrustações no contacto com infraestruturas criadas pelo Homem. O valor recomendável é um valor entre 6 e 9, para que a vida aquática não seja afetada.

A variável $Temperatura$ é o valor registado da temperatura da amostra de água, em ^oC. É uma variável física muito importante porque influencia diretamente as propriedades da água e, consequentemente, a qualidade desta. A variação da temperatura pode ser alterada devido a fontes naturais (energia solar) e/ou fontes antropogénicas (como efluentes industriais).

5.2.2 Análise Descritiva

Pretende-se realizar uma breve análise descritiva da variável OD, em cada estação de amostragem da Qualidade da Água. A Tabela 5.6 sintetiza algumas estatísticas descritivas da Oxigénio Dissolvido no período observado.

O maior valor de Oxigénio Dissolvido observado ($13,80\ mg/l$) é identificado na estação de amostragem de Eixes, em outubro de 2002. O menor valor de Oxigénio Dissolvido observado ($3,10\ mg/l$) é identificado na estação de amostragem de Pedras, em julho de 2002. A estação de amostragem de Albufeira de Sordo é a que apresenta menor desvio padrão ($0,93\ mg/l$) e menor coeficiente de variação ($10,11\ \%$). No Apêndice B está representada graficamente a evolução do OD em cada estação de amostragem, da bacia do Douro, ao longo do período em estudo, bem como o respetivo diagrama em caixa de bigodes e histograma.

Pode observar-se que existem estações com *outliers*, ou seja, valores discrepantes relativamente aos restantes. Estas observações não são excluídas, na medida em que mostram os valores extremos que o Oxigénio Dissolvido pode atingir, na maioria das vezes, devido

Tabela 5.6: Medidas descritivas da variável do Oxigénio Dissolvido, no período observado.

Código	Número	Dados Omissos	Mínimo (mg/l)	Máximo (mg/l)	1.º Quartil (mg/l)	Mediana (mg/l)	3.º Quartil (mg/l)	Média (mg/l)	Desvio Padrão (mg/l)	Coeficiente de Variação (%)
05K/02	125	7	6,50	11,90	8,30	9,20	10,20	9,15	1,21	13,22
04P/01	127	5	6,00	12,20	8,60	9,20	10,00	9,24	1,06	11,45
05R/01	126	6	6,90	12,20	8,30	8,95	9,70	9,08	1,10	12,11
05P/01	126	6	7,10	12,10	8,40	9,20	10,00	9,18	1,03	11,24
07G/04	130	2	4,70	11,40	7,87	8,70	9,88	8,82	1,33	15,11
06N/02	129	3	6,60	11,10	8,50	9,10	9,80	9,16	1,01	11,02
06N/01	127	5	6,00	12,30	8,30	9,20	10,00	9,16	1,11	12,16
10P/02	116	16	3,30	11,70	7,89	8,75	9,83	8,67	1,50	17,33
07M/01	130	2	6,00	11,90	8,20	8,96	9,70	9,04	1,03	11,42
11O/02	117	15	6,20	12,00	8,20	8,70	9,60	8,77	1,07	12,22
08P/02	116	16	5,37	11,00	7,70	8,80	9,43	8,64	1,24	14,31
02Q/02	126	6	5,80	12,70	8,60	9,22	10,14	9,32	1,13	12,17
06K/04	126	6	7,20	11,30	8,50	9,20	10,09	9,25	0,93	10,11
06H/01	129	3	5,80	12,50	8,20	9,20	10,00	9,10	1,20	13,20
06P/02	126	6	7,00	12,20	8,31	9,20	10,10	9,21	1,12	12,20
09O/03	116	16	5,20	12,00	7,80	8,62	9,60	8,62	1,30	15,13
06M/04	128	4	6,80	11,70	8,20	8,90	9,70	8,99	1,08	12,00
08K/01	115	17	4,90	12,40	8,30	9,20	10,10	9,08	1,35	14,89
03M/03	115	17	4,78	12,00	7,10	8,30	9,55	8,25	1,53	18,54
07H/05	130	2	5,10	11,70	8,12	9,05	10,00	9,10	1,26	13,81
04N/05	120	12	5,30	13,80	8,54	9,50	10,16	9,39	1,32	14,03
07K/01	128	4	4,80	12,80	8,40	9,40	10,20	9,30	1,26	13,55
07H/06	129	3	5,40	12,50	7,80	8,90	9,70	8,73	1,34	15,34
08H/02	129	3	4,10	12,30	8,60	9,50	10,30	9,43	1,22	12,98
07G/05	129	3	6,30	11,20	8,00	8,90	9,90	8,92	1,22	13,66
06G/07	118	14	6,80	11,40	8,43	9,30	10,08	9,25	0,99	10,73
07L/01	125	7	5,60	12,70	8,20	9,00	9,90	9,02	1,27	14,08
07K/04	128	4	6,00	12,40	7,97	8,75	10,00	8,88	1,34	15,11
04L/01	112	20	3,10	12,50	7,95	9,25	9,94	8,81	1,76	20,01
07J/02	128	4	5,40	11,90	8,10	9,15	10,01	9,07	1,30	14,38
07H/04	129	3	3,60	12,20	8,80	9,60	10,40	9,58	1,17	12,21
04N/01	127	5	6,40	13,00	8,45	9,60	10,30	9,39	1,24	13,19
06I/04	130	2	6,70	12,80	8,40	9,40	10,30	9,40	1,20	12,76
04N/06	128	4	5,30	12,50	8,40	9,35	10,40	9,33	1,42	15,22
06G/06	127	5	6,70	11,80	8,60	9,30	10,00	9,31	0,96	10,30
04J/09	131	1	7,40	13,50	8,90	9,80	10,40	9,73	1,11	11,37

à contaminação das massas de água por poluição de origem urbana, industrial e agrícola e à contaminação de águas subterrâneas.

No PGRH3 constata-se que nas zonas junto às localidades de Bragança e Régua a falta de água no Verão é mais acentuada e que as sub-bacias, onde as necessidades de água para a indústria são mais elevadas, são as do Douro e Costeiras entre o Douro e o Vouga.

As massas de água em mau estado/incumprimento, na sua maioria, situam-se nas zonas médias e inferiores das principais bacias hidrográficas da RH3, designadamente perto do litoral, nas sub-bacias do Douro, do Tâmega e do Côa. Nestas localizações verificam-se as maiores densidades populacionais e áreas de ocupação urbana, o que se reflete no valor médio do OD observado nas estações de amostragem perto do litoral e na região Nordeste.

Tabela 5.7: Medidas Descritivas sobre Oxigénio Dissolvido, em função do mês, no período observado.

Código	Número	Dados Omissos	Mínimo (mg/l)	Máximo (mg/l)	1.ž Quartil (mg/l)	Mediana (mg/l)	3.ž Quartil (mg/l)	Média (mg/l)	Desvio Padrão (mg/l)	Coeficiente de Variação (%)
janeiro	372	24	8,10	12,70	10.00	10.40	10.72	10.38	0.63	9.67
janeiro	372	24	8,10	12,70	10,00	10,40	10,72	10,38	0,63	9,67
fevereiro	381	15	8,30	13,50	10,00	10,40	10,80	10,47	0,72	0,04
março	384	12	8,10	12,80	9,60	10,00	10,40	9,99	0,70	0,01
abril	369	27	6,30	11,70	9,10	9,60	10,00	9,54	0,76	8,35
maio	388	8	6,50	11,53	8,50	8,90	9,30	8,90	0,78	8,45
junho	382	14	4,78	11,30	7,60	8,30	8,90	8,25	0,92	10,18
julho	375	21	3,10	10,80	7,40	8,00	8,50	7,90	0,94	10,21
agosto	353	43	4,00	10,60	7,50	8,00	8,50	7,97	0,86	9,79
setembro	376	20	3,30	11,20	7,60	8,20	8,60	8,02	1,01	11,01
outubro	374	22	3,60	13,80	8,00	8,50	9,00	8,46	1,05	11,42
novembro	389	7	6,40	11,32	8,90	9,30	9,90	9,30	0,80	9,27
dezembro	355	41	7,31	12,80	9,50	10,00	10,40	10,00	0,83	9,13

Para o estudo da sazonalidade da série recorre-se à representação das médias mensais da variável OD, por mês (Figura 5.11). A evolução do Oxigénio Dissolvido em função dos meses é notória: os valores mais elevados são observados em dezembro, janeiro, fevereiro e março e os valores mais baixos nos meses de julho, agosto e setembro. Existem trabalhos que apontam para a existência de sazonalidade para este tipo de variáveis ambientais (Cabecinha *et al.*, 2009; Costa & Gonçalves, 2011; Costa & Gonçalves, 2012; Gonçalves e Costa, 2013). No período em estudo, as séries temporais apresentaram valores em falta, mas não foi realizada qualquer ação para imputar dados.

Na Figura 5.13 são apresentadas as séries temporais de cada estação numa única representação para a concentração de Oxigénio Dissolvido, o perfil temporal da variável OD, onde se verifica um padrão ao longo do tempo.

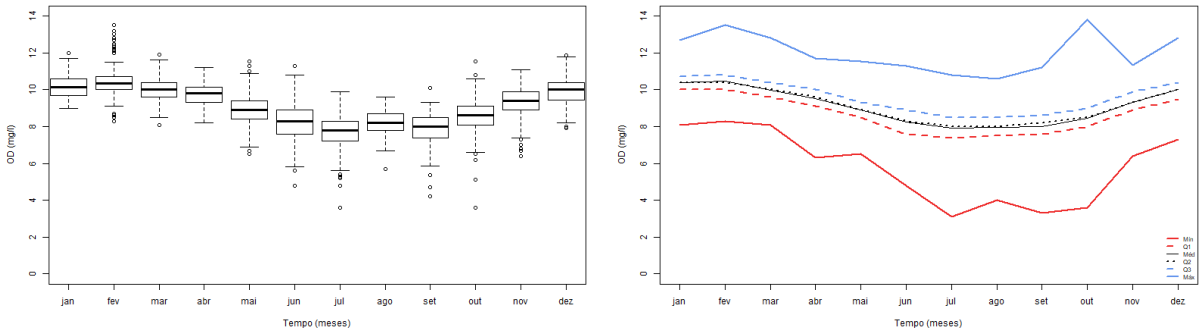


Figura 5.11: Esquerda: Diagramas em caixa de bigodes; Direita: As principais métricas da variável OD.

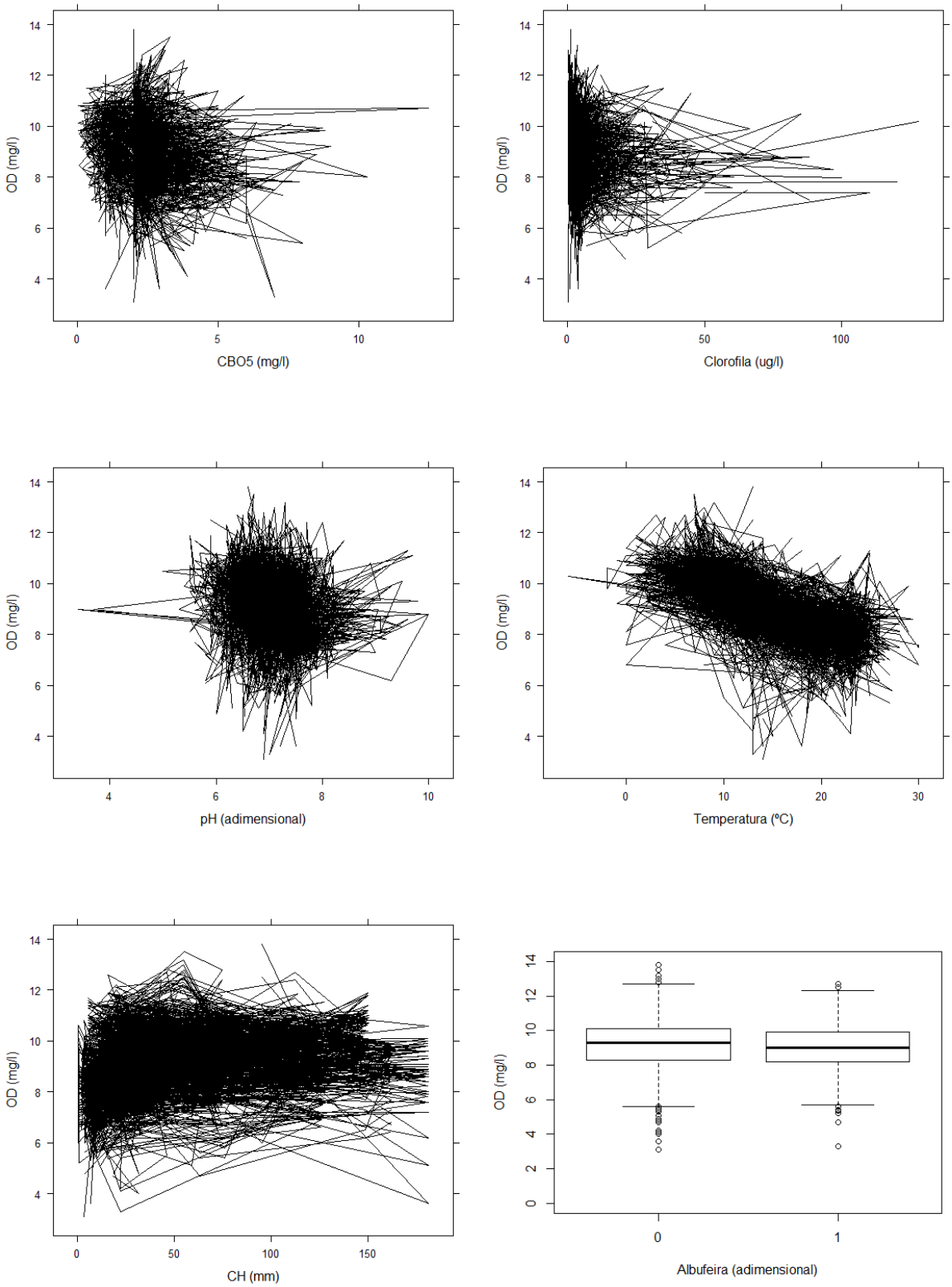


Figura 5.12: Representação gráfica do OD, em função das covariáveis $CBO5$, $Clorofila$, pH, $Temperatura$, CH e ALB.

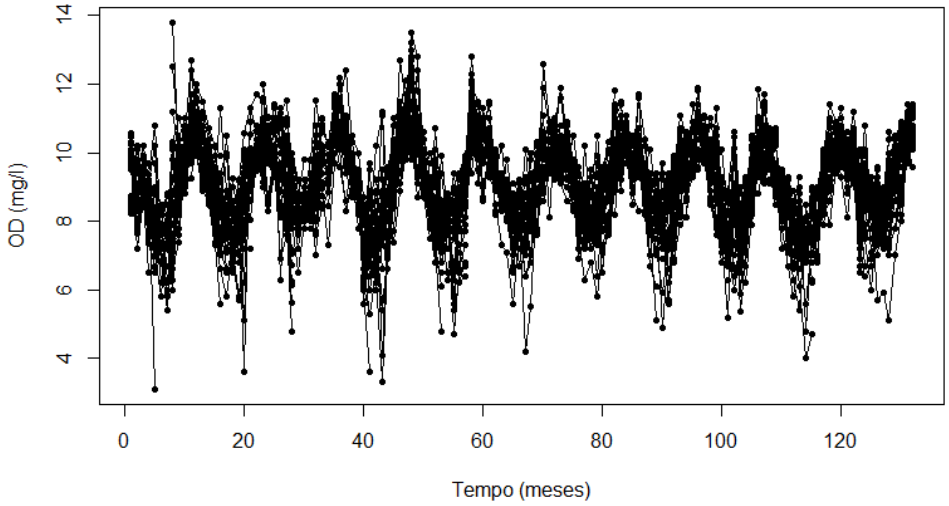


Figura 5.13: Perfil temporal da variável resposta, OD, em função do $Tempo$, no período observado.

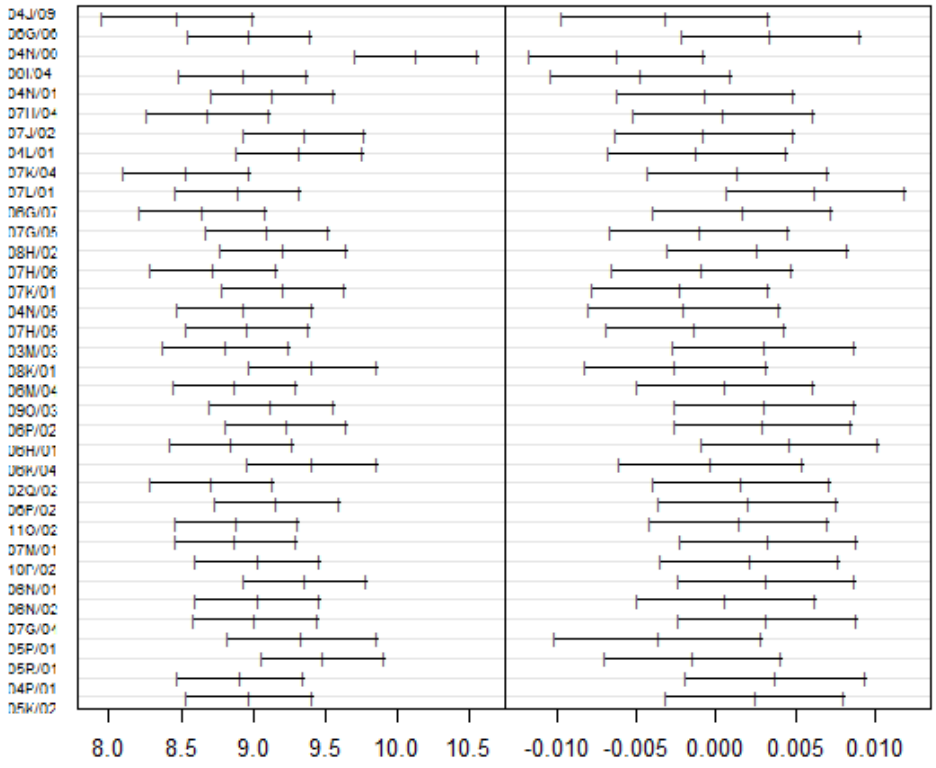


Figura 5.14: Representação gráfica dos intervalos de confiança para os coeficientes relativamente à constante (esquerda) e à variável tempo (direita), relativamente ao ajustamento linear para cada de amostragem.

Na Figura 5.12 representa-se graficamente a relação entre a concentração de OD e as

outras covariáveis. A inspeção gráfica sugere que o aumento da $Temperatura$ diminui a concentração da variável OD. Não é clara uma existência de uma associação entre as variáveis ALB, $CBO5$, $Clorofila$, pH, $Temperatura$, CH e o OD.

Com o objetivo de identificar os efeitos aleatórios a considerar, faz-se um ajustamento linear individual e analisam-se quais os parâmetros que mais variam de estação para estação. Esta análise é feita com base no gráfico da estimativa dos intervalos de confiança para os parâmetros do modelo ajustado a cada estação (Figura 5.14). Como se pode verificar, há variabilidade nas estimativas quer da intersecção quer do declive. Na construção dos modelos serão considerados os efeitos aleatórios no termo constante e no termo da covariável $Tempo$.

5.2.3 Formulação dos Modelos

Alpuim & El-Shaarawi (2008) e Gonçalves & Alpuim (2011) sugerem que uma boa opção para modelar a estrutura de correlação temporal dos erros aleatórios é um processo autorregressivo de ordem 1 (AR(1)), no contexto de dados ambientais. Este pressuposto prende-se com o facto da Qualidade da Água ser influenciada por condições que dependem do mês anterior. Os modelos em estudo consideram esta estrutura e as representações do comportamento da FAC e da FACP dos resíduos mostrarem que existe uma correlação temporal fraca na série, ou seja, os resíduos comportam-se como um processo autorregressivo AR(1).

Na parte fixa consideram-se dois casos: sem interação entre as covariáveis em estudo e a variável $Tempo$ e com interação entre as covariáveis em estudo e a variável $Tempo$. No segundo caso, se a interação for significativa, então o efeito da covariável no valor esperado da concentração de OD depende do tempo em estudo. Os efeitos sazonais usualmente variam de forma suave e contínua, pelo que faz sentido a consideração de funções de suavização, recorrendo a funções seno e cosseno para descrever as oscilações observadas ao longo do tempo, ajustando um modelo sazonal harmónico. Os Modelos Sazonais Harmónicos consideram um somatório, que descreve a variação sazonal dos dados, ao longo de cada período (s = 12, para dados mensais):

- cos^x é o cosseno de período ^{x}, onde ^{x} assume os valores 12; 6; 4; 3; 2,4 e 2. Por exemplo, cos^6 é a codificação para o cosseno de período 6;
- sen^x é o seno de período ^{x}, onde ^{x} assume os valores 12; 6; 4; 3 e 2,4.

De notar que nos Modelos Sazonais Harmónicos (com erros correlacionados), quando uma curva cosseno com um certo período é significativa, esta deve ser incluída juntamente com o correspondente seno do mesmo período e vice-versa.

Considerando que i representa a estação de amostragem em estudo; $Tempo_{it}$ designa o tempo desde março de 2002 até fevereiro de 2013, em meses; $CBO5_{it}$ refere-se à variável de Carência Bioquímica de Oxigénio, em mg/l, na estação de amostragem i no instante t; $Clorofila_{it}$ retrata o valor dos níveis de Clorofila, em ug/l, na estação de amostragem i no instante t; ALB_{it} designa se a estação de amostragem i é ou não uma albufeira; pH_{it} representa o valor de pH, na estação de amostragem i no instante t; $Temperatura_{it}$ designa o valor de temperatura da amostra, em $^{\circ}C$, na estação de amostragem i no instante t; CH_{it}, $CH_{i(t-1)}$ e $CH_{i(t-2)}$ referem-se ao fator hidrometeorológico, na estação de amostragem i, nos instantes t, $t-1$ e $t-2$, respetivamente; s é o período de sazonalidade ($s = 12$). Os quatro modelos analisados são:

(i) O Modelo 1 consiste no modelo com um efeito aleatório no termo constante (associado ao parâmetro β_0),

$$\begin{aligned} OD_{it} &= \beta_0 + \beta_1 Tempo_{it} + \beta_2 CBO5_{it} + \beta_3 Clorofila_{it} + \beta_4 pH_{it} + \beta_5 ALB_{it} \\ &\quad + \beta_6 Temperatura_{it} + \beta_7 CH_{it} + \beta_8 CH_{i(t-1)} + \beta_9 CH_{i(t-2)} \\ &\quad + \sum_{k=1}^{s/2} \left[\alpha_{1i} cos\Big(\frac{2\pi k Tempo_{it}}{s}\Big) + \alpha_{2i} sen\Big(\frac{2\pi k Tempo_{it}}{s}\Big) \right] + u_{1it} + \epsilon_{it} \\ i &= 1, \ldots, 36, \quad t = 1, \ldots, 132 \end{aligned} \tag{5.1}$$

em que ϵ_{it} é o erro aleatório tal que $\epsilon_{it} = \phi_1 \epsilon_{it-1} + a_{it}$, com $a_{it} \sim N(0, \sigma^2)$; u_{1i} representa o efeito aleatório e $u_{1it} \sim N(0, d_{11}^2)$;

(ii) O Modelo 2 consiste no modelo com efeito aleatório no termo constante e na variável tempo, isto é,

$$\begin{aligned} OD_{it} &= \beta_0 + \beta_1 Tempo_{it} + \beta_2 CBO5_{it} + \beta_3 Clorofila_{it} + \beta_4 pH_{it} + \beta_5 ALB_{it} \\ &\quad + \beta_6 Temperatura_{it} + \beta_7 CH_{it} + \beta_8 CH_{i(t-1)} + \beta_9 CH_{i(t-2)} \\ &\quad + \sum_{k=1}^{s/2} \left[\alpha_{1i} cos\Big(\frac{2\pi k Tempo_{it}}{s}\Big) + \alpha_{2i} sen\Big(\frac{2\pi k Tempo_{it}}{s}\Big) \right] \\ &\quad + u_{1it} + u_{2it} Tempo_{it} + \epsilon_{it} \\ i &= 1, \ldots, 36, \quad t = 1, \ldots, 132 \end{aligned} \tag{5.2}$$

em que ϵ_{it} é o erro aleatório tal que $\epsilon_{it} = \phi_1 \epsilon_{it-1} + a_{it}$, com $a_{it} \sim N(0, \sigma^2)$; u_{1it}, u_{2it} representam os efeitos aleatórios e

$$\begin{pmatrix} u_{1i} \\ u_{2i} \end{pmatrix} \sim N(\mathbf{0}, \mathbf{D}), \qquad \mathbf{D} = \begin{bmatrix} d_{11}^2 & d_{12} \\ d_{21} & d_{22}^2 \end{bmatrix};$$

(iii) O Modelo 3 consiste no modelo com um efeito aleatório no termo constante e um termo de interação

$$\begin{aligned} OD_{it} &= \beta_0 + \beta_1 Tempo_{it} + \beta_2 CBO5_{it} + \beta_3 Clorofila_{it} + \beta_4 pH_{it} + \beta_5 ALB_{it} \\ &+ \beta_6 Temperatura_{it} + \beta_7 CH_{it} + \beta_8 CH_{i(t-1)} + \beta_9 CH_{i(t-2)} \\ &+ \beta_{10} CBO5_{it} : Tempo_{it} + \beta_{11} Clorofila_{it} : Tempo_{it} + \beta_{12} pH_{it} : Tempo_{it} \\ &+ \beta_{10} ALB_{it} : Tempo_{it} + \beta_{13} Temperatura_{it} : Tempo_{it} + \beta_{14} CH_{it} : Tempo_{it} \\ &+ \beta_{15} CH_{i(t-1)} : Tempo_{it} + \beta_{16} CH_{i(t-2)} : Tempo_{it} \\ &+ \sum_{k=1}^{s/2} \left[\alpha_{1i} cos\Big(\frac{2\pi k Tempo_{it}}{s}\Big) + \alpha_{2i} sen\Big(\frac{2\pi k Tempo_{it}}{s}\Big) \right] \\ &+ u_{1it} + \epsilon_{it} \\ i &= 1, \ldots, 36, \quad t = 1, \ldots, 132 \end{aligned} \tag{5.3}$$

em que ϵ_{it} é o erro aleatório tal que $\epsilon_{it} = \phi_1 \epsilon_{it-1} + a_{it}$, com $a_{it} \sim N(0, \sigma^2)$; u_{1it} representa o efeito aleatório e $u_{1it} \sim N(0, d_{11}^2)$;

(iv) O Modelo 4 consiste no modelo com um efeito aleatório no termo constante e na variável tempo e um termo de interação

$$\begin{aligned} OD_{it} &= \beta_0 + \beta_1 Tempo_{it} + \beta_2 CBO5_{it} + \beta_3 Clorofila_{it} + \beta_4 pH_{it} + \beta_5 ALB_{it} \\ &+ \beta_6 Temperatura_{it} + \beta_7 CH_{it} + \beta_8 CH_{i(t-1)} + \beta_9 CH_{i(t-2)} \\ &+ \beta_{10} CBO5_{it} : Tempo_{it} + \beta_{11} Clorofila_{it} : Tempo_{it} \\ &+ \beta_{12} pH_{it} : Tempo_{it} + \beta_{10} ALB_{it} : Tempo_{it} \\ &+ \beta_{13} Temperatura_{it} : Tempo_{it} + \beta_{14} CH_{it} : Tempo_{it} \\ &+ \beta_{15} CH_{i(t-1)} : Tempo_{it} + \beta_{16} CH_{i(t-2)} : Tempo_{it} \\ &+ \sum_{k=1}^{s/2} \left[\alpha_{1i} cos\Big(\frac{2\pi k Tempo_{it}}{s}\Big) + \alpha_{2i} sen\Big(\frac{2\pi k Tempo_{it}}{s}\Big) \right] \\ &+ u_{1it} + u_{2it} Tempo_{it} + \epsilon_{it} \\ i &= 1, \ldots, 36, \quad t = 1, \ldots, 132 \end{aligned} \tag{5.4}$$

em que ϵ_{it} é o erro aleatório tal que $\epsilon_{it} = \phi_1 \epsilon_{it-1} + a_{it}$, com $a_{it} \sim N(0, \sigma^2)$; u_{1it}, u_{2it} representam os efeitos aleatórios e

$$\begin{pmatrix} u_{1i} \\ u_{2i} \end{pmatrix} \sim N(\mathbf{0}, \mathbf{D}), \qquad \mathbf{D} = \begin{bmatrix} d_{11}^2 & d_{12} \\ d_{21} & d_{22}^2 \end{bmatrix}.$$

Com base na formulação dos modelos estimaram-se os modelos completos. As estimativas dos efeitos fixos, os erros padrão, os valores da estatística de teste e os valores de prova são apresentados nas Tabelas 5.8 a 5.11.

O Método de Máxima Verosimilhança Restrito (REML) foi utilizado para o ajustamento dos diferentes modelos. Os modelos finais são apresentados na Tabela 5.13.

Com base no Teste da Razão de Verosimilhança apresentado na Tabela 5.12, onde se comparam os modelos encaixados (Modelo 1 *vs* Modelo 3 e Modelo 2 *vs* Modelo 4), pode-se verificar que os valores de prova são inferiores a 5 %, rejeitando-se a hipótese nula.

A comparação entre o Modelo 3 e o Modelo 4 é feita com base nos critérios de informação AIC e BIC, uma vez que estes não são aninhados.

O Modelo 4 é o modelo que apresenta maior valor de AIC e BIC (Tabela 5.12) e, por essa razão, o Modelo 3 é o modelo selecionado.

Tabela 5.8: Estimativas dos coeficientes da parte fixa, do Modelo Completo 1.

Variável	Estimativa	Erro Padrão	Estatística t	Valor de Prova
Constante	7,7822	0,3145	24,7482	**<0,0001**
$Tempo_{it}$	0,0018	0,0005	3,2851	**0,0010**
CBO5	0,0607	0,0181	3,3605	**0,0008**
$Clorofila$	0,0007	0,0018	0,3586	0,7200
pH	0,2732	0,0408	6,6895	**<0,0001**
Temperatura	-0,0591	0,0061	-9,6186	**<0,0001**
ALB	-0,1393	0,1164	-1,1968	0,2399
CH_t	-0,0002	0,0005	-0,3102	0,7565
CH_{t-1}	0,0018	0,0005	3,3020	**0,0010**
CH_{t-2}	0,0007	0,0005	1,2282	0,2195
cos^{12}	0,8606	0,0553	15,5681	**<0,0001**
cos^{6}	0,0997	0,0262	3,8072	**<0,0001**
cos^4	-0,0373	0,0209	-1,7866	0,0741
cos^{3}	0,0414	0,0189	2,1899	**0,0286**
$cos^{2,4}$	0,0053	0,0174	0,3028	0,7621
cos^2	0,0203	0,0132	1,5305	0,1260
sen^{12}	-0,0434	0,0352	-1,2356	0,2167
sen^6	-0,0273	0,0264	-1,0348	0,3009
sen^{4}	-0,0415	0,0204	-2,0374	**0,0417**
sen^3	0,0119	0,0191	0,6229	0,5334
$sen^{2,4}$	-0,0175	0,0183	-0,9566	0,3388

Tabela 5.9: Estimativas dos coeficientes da parte fixa, do Modelo Completo 2.

Variável	Estimativa	Erro Padrão	Estatística t	Valor de Prova
Constante	7,7822	0,3145	24,7482	**<0,0001**
$\mathbf{Tempo_{it}}$	0,0018	0,0005	3,2851	**0,0010**
CBO5	0,0607	0,0181	3,3605	**0,0008**
$Clorofila$	0,0007	0,0018	0,3586	0,7200
pH	0,2732	0,0408	6,6895	**<0,0001**
Temperatura	-0,0591	0,0061	-9,6186	**<0,0001**
$Albufeira$	-0,1393	0,1164	-1,1968	0,2399
CH_t	-0,0002	0,0005	-0,3102	0,7565
$\mathbf{CH_{t-1}}$	0,0018	0,0005	3,3020	**0,0010**
CH_{t-2}	0,0007	0,0005	1,2282	0,2195
$\mathbf{cos^{12}}$	0,8606	0,0553	15,5681	**<0,0001**
$\mathbf{cos^{6}}$	0,0997	0,0262	3,8072	**0,0001**
cos^4	-0,0373	0,0209	-1,7866	0,0741
$\mathbf{cos^{3}}$	0,0414	0,0189	2,1899	**0,0286**
$cos^{2,4}$	0,0053	0,0174	0,3028	0,7621
cos^2	0,0203	0,0132	1,5305	0,1260
sen^{12}	-0,0434	0,0352	-1,2356	0,2167
sen^6	-0,0273	0,0264	-1,0348	0,3009
$\mathbf{sen^{4}}$	-0,0415	0,0204	-2,0374	**0,0417**
sen^3	0,0119	0,0191	0,6229	0,5334
$sen^{2,4}$	-0,0175	0,0183	-0,9566	0,3388

Tabela 5.10: Estimativas dos coeficientes da parte fixa, do Modelo Completo 3.

Variável	Estimativa	Erro Padrão	Estatística t	Valor de Prova
Constante	7,9377	0,4665	17,0165	**<0,0001**
$Tempo_{it}$	0,0007	0,0065	0,1112	0,9115
$CBO5$	0,0222	0,0334	0,6669	0,5049
$Clorofila$	0,0048	0,0033	1,4753	0,1402
pH	0,3082	0,0631	4,8802	**<0,0001**
Temperatura	-0,0781	0,0085	-9,1897	**<0,0001**
ALB	-0,2068	0,1305	-1,5840	0,1227
CH_t	-0,0002	0,0008	-0,1981	0,8430
CH_{t-1}	0,0015	0,0009	1,6558	0,0979
CH_{t-2}	0,0002	0,0009	0,2239	0,8229
$\mathbf{cos^{12}}$	0,8652	0,0557	15,5285	**<0,0001**
$\mathbf{cos^{6}}$	0,0950	0,0265	3,5863	**0,0003**
cos^4	-0,0346	0,0209	-1,6535	0,0983
cos^3	0,0412	0,0190	2,1677	**0,0303**
$cos^{2,4}$	0,0059	0,0177	0,3338	0,7386
cos^2	0,0201	0,0135	1,4951	0,1350
sen^12	-0,0441	0,0361	-1,2210	0,2222
sen^6	-0,0382	0,0266	-1,4342	0,1516
sen^4	-0,0451	0,0204	-2,2051	0,0275
sen^3	0,0142	0,0192	0,7382	0,4605
$sen^2, 4$	-0,0186	0,0184	-1,0123	0,3115
$CBO5 : Tempo_{it}$	0,0007	0,0006	1,3007	0,1935
$Clorofila : Tempo_{it}$	-0,0001	0,0001	-1,4800	0,1390
pH	-0,0008	0,0009	-0,8707	0,3840
Temperatura : $\mathbf{Tempo_{it}}$	0,0003	0,0001	3,1540	**0,0016**
$ALB : Tempo_{it}$	0,0011	0,0011	1,0459	0,2957
$CH_t : Tempo_{it}$	0,0000	0,0000	0,0790	0,9371
$CH_{t-1} : Tempo_{it}$	0,0000	0,0000	0,0642	0,9488
$CH_{t-2} : Tempo_{it}$	0,0000	0,0000	0,3525	0,7245

Tabela 5.11: Estimativas dos coeficientes da parte fixa, do Modelo Completo 4.

Variável	Estimativa	Erro Padrão	Estatística t	Valor de Prova
Constante	7,9281	0,4733	16,7521	**<0,0001**
$Tempo_{it}$	0,0008	0,0067	0,1194	0,9049
$CBO5$	0,0218	0,0334	0,6521	0,5144
$Clorofila$	0,0048	0,0033	1,4708	0,1415
pH	0,3112	0,0641	4,8562	**<0,0001**
Temperatura	-0,0787	0,0085	-9,2484	**<0,0001**
ALB	-0,2080	0,1337	-1,5559	0,1293
CH_t	-0,0002	0,0008	-0,2041	0,8383
CH_{t-1}	0,0015	0,0009	1,6470	0,0997
CH_{t-2}	0,0002	0,0009	0,2132	0,8312
$\mathbf{cos^{12}}$	0,8635	0,0557	15,4974	**<0,0001**
$\mathbf{cos^{6}}$	0,0952	0,0265	3,5980	**0,0003**
cos^4	-0,0345	0,0209	-1,6497	0,0991
cos^3	0,0410	0,0190	2,1569	0,0311
$cos^{2,4}$	0,0059	0,0177	0,3345	0,7380
cos^2	0,0202	0,0135	1,5023	0,1331
sen^{12}	-0,0434	0,0361	-1,2021	0,2294
sen^6	-0,0384	0,0266	-1,4412	0,1496
$\mathbf{sen^{4}}$	-0,0452	0,0204	-2,2136	**0,0269**
sen^3	0,0143	0,0192	0,7420	0,4582
$sen^{2,4}$	-0,0187	0,0184	-1,0162	0,3096
$CBO5 : Tempo_{it}$	0,0007	0,0006	1,3183	0,1875
$Clorofila : Tempo_{it}$	-0,0001	0,0001	-1,4727	0,1410
$pH : Tempo_{it}$	-0,0008	0,0009	-0,8859	0,3757
$\mathbf{Temperatura : Tempo_{it}}$	0,0003	0,0001	3,2109	**0,0013**
$ALB : Tempo_{it}$	0,0011	0,0011	1,0168	0,3093
$CH_t : Tempo_{it}$	0,0000	0,0000	0,0889	0,9292
$CH_{t-1} : Tempo_{it}$	0,0000	0,0000	0,0768	0,9388
$CH_{t-2} : Tempo_{it}$	0,0000	0,0000	0,3598	0,7190

Tabela 5.12: Teste da Razão de Verosimilhança aplicado aos modelos em análise.

Modelo	AIC	BIC	$loglik$	Teste	Valor de Prova
Modelo 1	6492,439	6592,938	-3229,220		
Modelo 3	6496,439	6608,761	-3229,220	1 vs 3	0,0178
Modelo 2	6497,727	6604,130	-3230,863		
Modelo 4	6501,467	6619,694	-3230,734	2 vs 4	<0,0001

Tabela 5.13: Estimativas dos coeficientes dos Modelos 1, 2, 3 e 4.

Modelo	**Modelo 1**		**Modelo 2**		**Modelo 3**		**Modelo 4**	
	Estimativa	Erro Padrão	Estimativa	Erro Padrão	Estimativa	Erro Padrão	Estimativa	Erro Padrão
Efeitos Fixos								
$Constante$	7,7283	0,3045	7,7283	0,3045	8,0208	0,3141	8,0208	0,3146
$Tempo$	0,0016	0,0005	0,0016	0,0005	-0,0024	0,0012	-0,0025	0,0012
$CBO5$	0,0657	0,0177	0,0657	0,0177	0,0640	0,0176	0,0641	0,0176
pH	0,2743	0,0402	0,2743	0,0402	0,2700	0,0402	0,2712	0,0402
$Temperatura$	-0,0589	0,0061	-0,0589	0,0061	-0,0760	0,0076	-0,0766	0,0076
CH_{t-1}	0,0016	0,0005	0,0016	0,0005	0,0014	0,0005	0,0014	0,0005
cos^{12}	0,8777	0,0524	0,8777	0,0524	0,8776	0,0522	0,8759	0,0522
cos^{6}	0,0983	0,0251	0,0983	0,0251	0,0969	0,0250	0,0972	0,0250
cos^{4}	-0,0416	0,0203	-0,0416	0,0203	-0,0385	0,0203	-0,0384	0,0203
cos^{3}	0,0401	0,0186	0,0401	0,0186	0,0406	0,0186	0,0404	0,0186
sen^{12}	-0,0415	0,0327	-0,0415	0,0327	-0,0416	0,0325	-0,0408	0,0325
sen^{6}	-0,0442	0,0230	-0,0442	0,0230	-0,0519	0,0230	-0,0521	0,0230
sen^{4}	-0,0434	0,0201	-0,0434	0,0201	-0,0449	0,0201	-0,0450	0,0201
sen^{3}	0,0170	0,0183	0,0170	0,0183	0,0166	0,0183	0,0166	0,0183
$Temperatura : Tempo$	-	-	-	-	0,0002	0,0001	0,0002	0,0001
Efeitos Aleatórios								
σ^2	0,7861		0,7862		0,7831		0,7821	
ϕ	0,2602		0,2602		0,2548		0,2531	
d_{11}	0,3303		0,3303		0,3328		0,3431	
$d_{12} = \hat{d}_{21}$	-		-0,0020		-		-0,2680	
d_{22}	-		0,8455		-		0,7821	

5.2.4 Análise dos Resíduos

A análise dos resíduos é uma ferramenta útil para a verificação dos pressupostos dos modelos, relativamente à variável de concentração de Oxigénio Dissolvido.

Na Figura 5.15 representam-se os valores ajustados *vs* os resíduos padronizados (à esquerda), os valores ajustados *vs* os valores observados (ao centro) e o gráfico *Q-Q plot* (à direita), do Modelo 3. No gráfico da esquerda verifica-se um padrão de homocedasticidade e poucos *outliers*. Na representação gráfica dos valores observados e valores ajustados verifica-se que se dispõe linearmente e é percetível que existem poucos *outliers*. A representação dos quantis teóricos e quantis empíricos sugere que os resíduos seguem aproximadamente uma distribuição Normal.

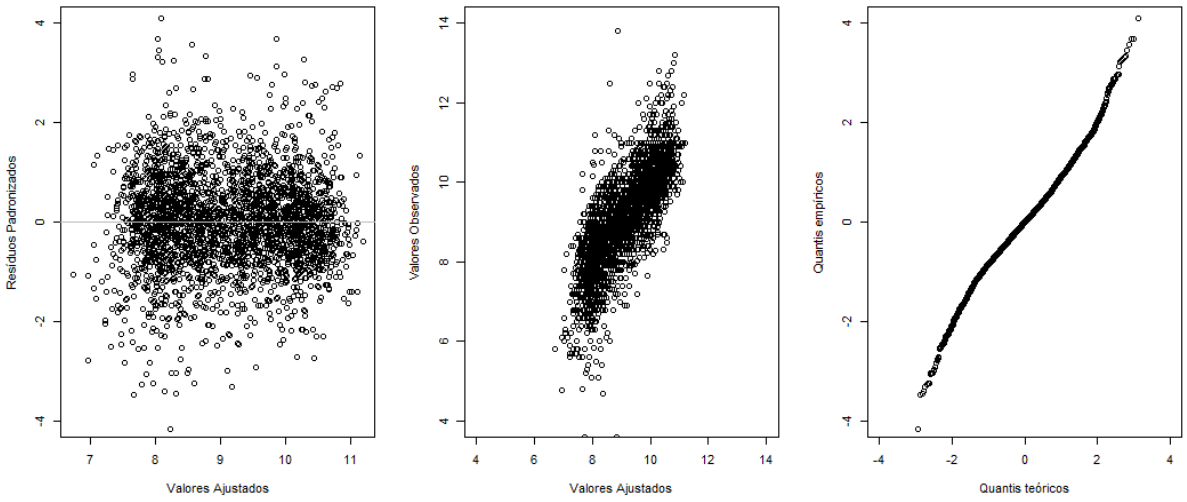


Figura 5.15: Esquerda: Resíduos padronizados do Modelo *vs* Valores Ajustados; Centro: Valores observados *vs* Valores Ajustados; Direita: *Q-Q plot* dos resíduos normalizados, Modelo 3.

Nas Figuras 5.16 e B.20 a B.35 (Apêndice B) são representadas a série original e os valores estimados, a FAC, a FACP e *Q-Q plot* dos resíduos padronizados, para cada estação de amostragem em análise. De acordo com a análise das figuras, os pressupostos de normalidade e de média nula parecem ser válidos. No entanto, há casos de estações de amostragem em que os resíduos não têm um bom comportamento.

A maioria das representações da FAC apresenta um decaimento exponencial para zero. As FACP das séries dos resíduos padronizados têm uma queda brusca para zero a partir *lag* 1.

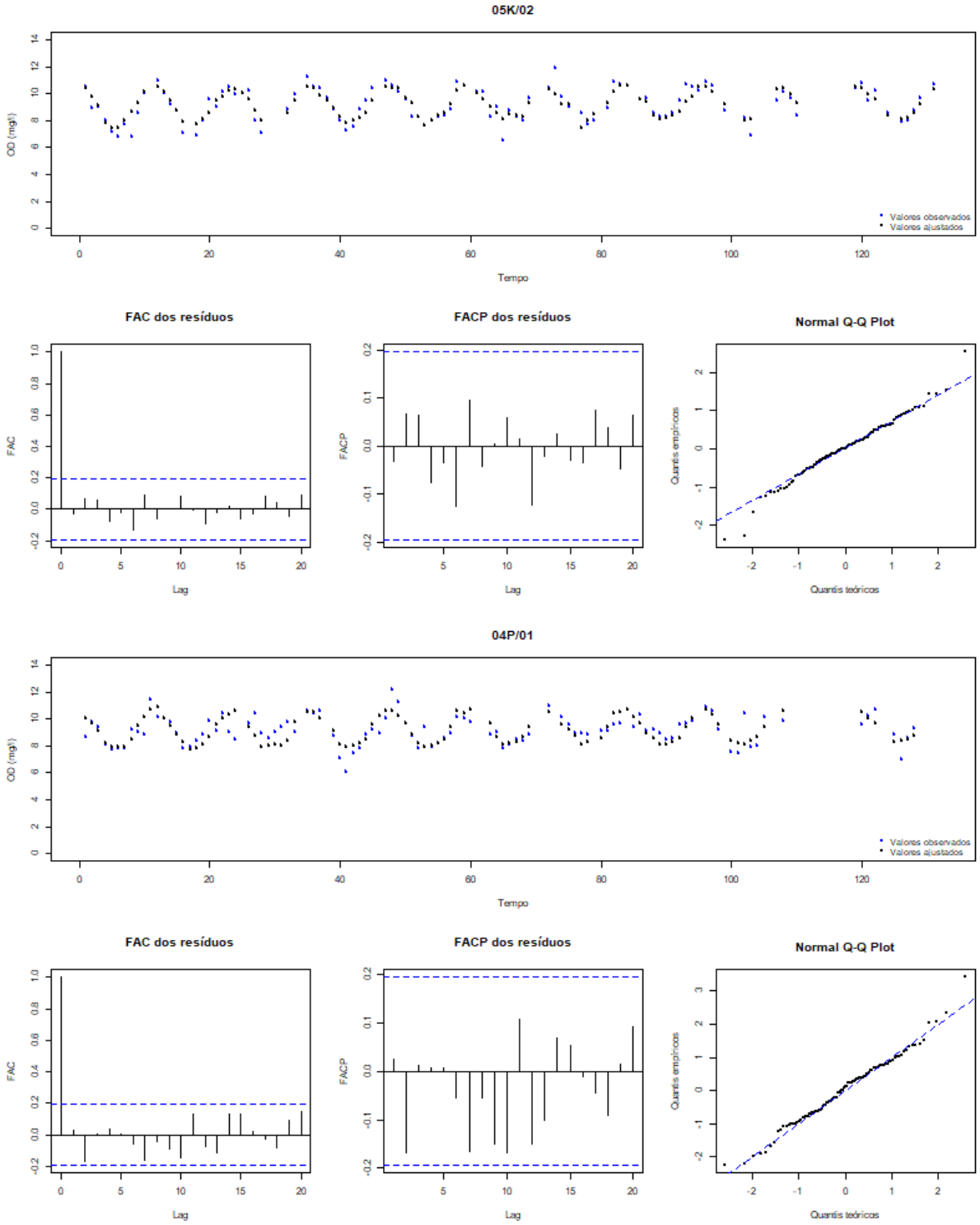


Figura 5.16: Representações da série original e dos os valores estimados, da FAC, da FACP e do *Q-Q plot* dos resíduos do modelo, nas estações.

Capítulo 6

Conclusão

Este trabalho foi dedicado à análise de séries temporais de variáveis ambientais, nomeadamente de precipitação e de Qualidade da Água de superfície, observadas em estações de amostragem da bacia hidrográfica do rio Douro. A modelação das séries de precipitação por aplicação de técnicas Geoestatísticas (em particular, *Kriging*) foi realizada com o intuito de se obter um fator hidrometeorológico importante para descrever e explicar o comportamento da variável de Qualidade da Água nos Modelos de Efeitos Mistos, neste caso de estudo na modelação do Oxigénio Dissolvido. Sabe-se que a variação temporal da maioria das variáveis de Qualidade da Água está naturalmente correlacionada com a variação da precipitação. Assim, para isso foram aplicadas técnicas de modelação espacial e temporal por forma a obterem-se estimativas do fator hidrometeorológico no espaço (em cada estação de amostragem de qualidade) e no tempo (em cada mês do período observado das variáveis qualidade).

Em relação às séries temporais de Qualidade da Água, dada a sua estrutura temporal e para ter em consideração também a sua estrutura espacial (variação espacial da sua localização, estação de amostragem de qualidade, na bacia hidrográfica do rio Douro), foram estabelecidos Modelos de Efeitos Mistos. Estes modelos revelaram-se adequados, uma vez que se existem medidas repetidas ao longo do tempo em unidades experimentais com grande variabilidade entre si. A combinação de efeitos fixos com os efeitos aleatórios permitiu a modelação do comportamento ao longo do tempo do Oxigénio Dissolvido. A estrutura aleatória foi identificada através das estimativas dos intervalos de confiança para os parâmetros do ajustamento individual e a significância das covariáveis incluídas foi avaliada com base no Teste da Razão de Verosimilhança, no caso de modelos encaixados, e pelo critério AIC/BIC, em modelos não encaixados. Na modelação da estrutura de correlação dos erros aleatórios recorreu-se a um modelo autorregressivo de ordem 1.

O modelo final selecionado foi o modelo com apenas um efeito aleatório no termo constante e observou-se uma associação significativa e positiva entre as variáveis carência

bioquímica de oxigénio ($CBO5$), valor de PH (pH), fator hidrometeorológico (CH) e a concentração de oxigénio (OD), mas uma associação significativa e negativa entre a temperatura ($Temperatura$) e a concentração de Oxigénio Dissolvido (OD).

A limitação principal deste trabalho foi a falta de dados. A não existência de séries completas da precipitação e das variáveis de qualidade nas estações de amostragem da rede de monitorização, que são em grande número mas com fraca qualidade. Foi necessário apenas selecionarem-se as estações de amostragem com um maior número de observações, aquelas com, no máximo, 20 % de dados omissos, no período de março de 2002 a fevereiro de 2013. Seria muito interessante trabalhar com dados mais recentes, mas a qualidade de monitorização tem vindo a piorar nos últimos anos.

6.1 Trabalho Futuro

O trabalho desenvolvido abre um conjunto de possíveis novas investigações, em particular em relação a alguns problemas que surgiram, mas que não puderam ser resolvidos no âmbito deste trabalho. Assim, vão-se indicar alguns tópicos que se consideram importantes para futura investigação. Como trabalho futuro sugere-se:

- Selecionar outro período temporal e/ou bacia hidrográfica e analisar de forma os resultados obtidos no processo de modelação das séries temporais de Oxigénio Dissolvido nas estações de qualidade dos rios;

- Avaliar o impacto de outras covariáveis no comportamento do Oxigénio Dissolvido como o do caudal (se houvesse essas medições) que, do ponto de vista hidrológico, seria talvez o mais adequado. Também seria importante avaliar o impacto de covariáveis de natureza antropomorfa como o número/tipo de indústrias poluentes ou o número/tipo de atividades agrícolas localizadas a montante de cada uma das estações de amostragem, ou a densidade populacional na proximidade ou a montante de cada estação de amostragem;

- Analisar a aplicação de modelos espaço-temporais. O foco deste trabalho foi direcionado mais para a modelação espacial da série de precipitação, apesar de ter sido efetuada considerando o tempo (modelação espacial da precipitação em cada mês do ano, no período observado). Também foi considerada a modelação temporal da série da variável de Qualidade da Água, apesar (da localização) da estação de amostragem ter sido uma covariável do modelo. O estudo da aplicação de modelos num contexto espaço-temporal seria uma possibilidade a desenvolver para ambos os processos de modelação.

Bibliografia

[1] Agresti, A. (1990). *Categorical Data Analysis*. New York: John Wiley and Sons.

[2] Akaike, H. (1974). A new look at statistical model identification. *IEEE transactions on Automatic Control*, 19, 716 –723.

[3] Allan, R. P., & Liu, C. (2019). Evaluating Large-Scale Variability and Change in Tropical Rainfall and Its Extremes. *Elsevier: In Tropical Extremes*, 139 –163.

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

[5] Alpuim, T., & El-Shaarawi, A. (2008). On the efficiency of regression analysis with AR (p) errors. *Journal of Applied Statistics*, 35(7), 717 –737.

[6] Bárdossy, A. & Pegram, G. G. S. (2009). Copula based multisite model for daily precipitation simulation. *Hydrology Earth System Science*, 13, 2299 –2314

[7] Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. arXiv preprint arXiv:1506.04967.

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

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

[10] Box, G., & Jenkins, G. (1970). *Time Series Analysis, Forecasting and Control*. San Francisco: Holden-Day.

[11] Brewer, M. J., Butler, A., & Cooksley, S. L. (2016) The relative performance of AIC, AICC and BIC in the presence of unobserved heterogeneity. *Methods in Ecology and Evolution*, 7(6), 679 –692.

[12] Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods & Research*, 33(2), 261 –304.

[13] Burnham, K. P., & Anderson, D.R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. New York: Springer-Verlag, 2a ed.

[14] Cabecinha, E., Cortes, R., Pardal, M. Â., & Cabral, J. A. (2009). A Stochastic Dynamic Methodology (StDM) for reservoirs water quality management: Validation of a multi-scale approach in a south european basin (Douro, Portugal). *Ecological Indicators*, 9(2), 329 –345.

[15] Caiado, J. (2016). *Métodos de Previsão em Gestão - Com Aplicações em Excel*. Lisboa: Edições Sílabo, 2a ed.

[16] Chatfield, C. (2000). *Time-Series Forecasting*. Chapman & Hall/CRC, 1a ed.

[17] Chatfield, C. (2004). *The Analysis of Time Series: An Introduction*. Chapman & Hall/CRC, 5a ed.

[18] Chow, V. T., Maidment, D. R., & Mays, L. W. (1988). *Applied Hidrology*. McGraw-Hill Series in Water Resources and Environmental Engineering, McGraw-Hill, 1a ed.

[19] Costa, M., & Gonçalves, A. M. (2011). Clustering and forecasting of dissolved oxygen concentration on a river basin. *Stochastic Environmental Research and Risk Assessment*, 25(2), 151 –163.

[20] Costa, M., & Gonçalves, A. M. (2012). Combining statistical methodologies in water quality monitoring in a hydrological Basin-Space and time approaches. *Water Quality Monitoring and Assessment*, 121 –142.

[21] Cowpertwait, P., & Metcalfe, A. (2009). *Introductory Time Series with R*. New York: Springer, 1a ed.

[22] Cox, D. R., & Hinkley, D. V. (1974). *Theoretical statistics*. New York: Chapman&Hall/CRC, 1a ed.

[23] Cressie, N. A. C. (1993). *Statistics for Spatial Data*. New York: Wiley, 2a ed.

[24] Cressie, N., & Huang, H. C. (1999). Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical Association*, 94(448), 1330 –1339.

[25] Cressie, N., & Wikle, C. K. (2015). *Statistics for spatio-temporal data*. John Wiley & Sons.

[26] Cressie, N., Zammit-Mangion, A., & Wikle, C. K. (2019). *Spatio-Temporal Statistics with R*. Chapman & Hall/CRC.

[27] D′Odorico, P., Carr, J., Dalin, C., DellAngelo, J., Konar, M., Laio, F., & Tuninetti, M. (2019). Global virtual water trade and the hydrological cycle: patterns, drivers, and socio-environmental impacts. *Environmental Research Letters*, 14(5), 053001.

[28] Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1), 1 –22.

[29] Dickey, D., & Fuller, W. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74, 427 –431.

[30] Diggle, P. J., & Giorgi, E. (2019). *Model-based Geostatistics for Global Public Health: Methods and Applications*. Chapman & Hall/CRC Interdisciplinary Statistics, 1a ed.

[31] Diggle, P. J., Heagerty, P., Liang, K. Y., Heagerty, P. J., & Zeger, S. (2002). *Analysis of longitudinal data*. Oxford University Press.

[32] Diggle, P. J., Liang, K. Y., & Zeger, S. L. (1994). *Analysis of Longitudinal Data*, Oxford: Oxford University Press.

[33] Enders, W. (2015). *Applied Econometric Time Series*. Alabama: John Wiley and Sons, 4a ed.

[34] Fausto, M., Carneiro, M., Antunes, C., Pinto, J., & Colosimo, E. (2008). O modelo de regressão linear misto para dados longitudinais: uma aplicação na análise de dados antropométricos desbalanceados. *Cadernos de Saúde Pública*, 24.

[35] Gonçalves, A. M., & Alpuim, T. (2011). Water quality monitoring using cluster analysis and linear models. *Environmetrics*, 22(8), 933 –945.

[36] Gonçalves, A. M., & Costa, M. (2013). Predicting seasonal and hydro-meteorological impact in environmental variables modelling via Kalman filtering. *Stochastic Environmental Research and Risk Assessment*, 27(5), 1021 –1038.

[37] Goovaerts, P. (1997). *Geostatistics for natural resources evaluation*. Oxford University Press on Demand.

[38] Gräler, B. (2014). Modelling skewed spatial random fields through the spatial vine copula. *Spatial Statistics*, 10, 87 –102.

[39] Harville, D. (1976). Extension of the Gauss-Markov theorem to include the estimation of random effects. *The Annals of Statistics*, 4(2), 384 –395.

[40] Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika*, 61(2), 383 –385.

[41] Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American statistical association*, 72(358), 320 –338.

[42] Ho, L., Pham, D., Van Echelpoel, W., Muchene, L., Shkedy, Z., Alvarado, A., Espinoza-Palacios, J., Arevalo-Durazno, M., Thas, O. & Goethals, P. (2018). A closer look on spatiotemporal variations of dissolved oxygen in waste stabilization ponds using mixed models. *Water*, 10(2), 201.

[43] Isaacs, E. H., & Srivastava, R. M. (1989), *An Introduction to Applied Geostatistics*, Oxford University Press.

[44] Jebb, A., Tay, L., Wang, W., & Huang, Q. (2015). Time series analysis for psychological research: examining and forecasting change. *Frontiers in Psychology*, 6, 727.

[45] Jowett, G. H. (1955). Sampling properties of local statistics in stationary stochastic series. *Biometrika*, 42(1/2), 160 –169.

[46] Kass, R. E., Caffo, B. S., Davidian, M., Meng, X. L., Yu, B., Reid, N. (2016). Ten simple rules for effective statistical practice. *PLOS Computational Biology*, 12(6).

[47] Kirchgässner, G., & Wolters, J. (2008). *Introduction to modern time series analysis*. Springer-Verlag.

[48] Kolmogorov, A. N. (1941). Interpolation and Extrapolation of Stationary Sequences. *Izvestiya the Academy of Sciences of the USSR Series Mathematic*, 5, 3-14.

[49] Krige, D. G. (1951). A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy*, 52(6), 119 –139.

[50] Kwiatkowski, D., Phillips, P., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, 54, 159 –178.

[51] Kyriakidis, P. C., Kim, J., & Miller, N. L. (2001). Geostatistical mapping of precipitation from rain gauge data using atmospheric and terrain characteristics. *Journal of Applied Meteorology*, 40(11), 1855 –1877.

[52] Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4), 963 –974.

[53] Lefêvre, C. (1997). *"Kriging"e "Cokringing"e aplicações ao Radar Meteorológico*. Master Thesis, University of Lisbon, Lisbon.

[54] Makridakis, S., Wheelwright, S., & Hyndman, R. (1998). *Forecasting: Methods and Applications*. New York: John Wiley and Sons, 3a ed.

[55] Matern, B. (1960). Spatial Variation. *Meddelanden Statens från Skogsforskningsinstitut Stockholm*, 49 (5), 1 –144. [2nd edition (1986). Spatial variation. *Lecture Notes in Statistics*, No. 36, Springer, New York, 2a ed.]

[56] Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, 58(8), 1246 –1266.

[57] Moshogianis, A. (2015). *A Statistical Model for the Prediction of Dissolved Oxygen Dynamics and the Potential for Hypoxia in the Mississippi Sound and Bight*. Master Thesis. The University of Southern Mississippi, Mississippi.

[58] McCulloch, C. E., & Searle, S. R. (2001). *Generalized, Linear, and Mixed Models*. Wiley Series in Probability and Statistics.

[59] McCulloch, P., Nelder, J.A. (1989) *Generalized Linear Models*. London: Chapman and Hall, 2a Ed.

[60] Mercer, W. B., & Hall, A. D. (1911). The experimental error of field trials. *The Journal of Agricultural Science*, 4(2), 107 –132.

[61] Min, S. K., Zhang, X., Zwiers, F. W., & Hegerl, G. C. (2011). Human contribution to more-intense precipitation extremes. *Nature*, 470(7334), 378.

[62] Molenberghs, G., & Verbeke, G. (2005). *Model for Discrete Longitudinal Data*. New York: Springer.

[63] Murteira, B., Muller, D., & Turkman, K. (1993). *Análise de Sucessões Cronológicas*. Lisboa: McGraw-Hill.

[64] Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3), 370 –384.

[65] Olea, R. A. (2012). *Geostatistics for engineers and earth scientists*. Springer Science & Business Media.

[66] Patterson, H. D., & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3), 545 –554.

[67] Pebesma, E. J. (2004). Multivariable geostatistics in S: the gstat package. *Computers & Geosciences*, 30(7), 683 –691.

[68] Pebesma, E. J. (2012). spacetime: Spatio-temporal data in R. *Journal of Statistical Software*, 51(7), 1 –30.

[69] Pebesma, E., & Heuvelink, G. (2016). Spatio-temporal interpolation using gstat.*RFID Journal*, 8(1), 204 –218.

[70] Phillips, P., & Perron, P. (1988). Testing for unit roots in time series regression. *Biometrika*, 75, 335 –346.

[71] Pinheiro, J. C. (1994). *Topics in mixed effects models*. Ph. D. Thesis, University of Wisconsin, Madison.

[72] Pinheiro, J. C., & Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of computational and Graphical Statistics*, 4(1), 12 –35.

[73] Pinheiro, J. C., & Bates, D. M. (2000). Linear mixed-effects models: basic concepts and examples. *Mixed-effects models in S and S-Plus*, 3 –56.

[74] R Core Team (2017). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*, Vienna, Austria. URL https://www.R-project.org/

[75] Ripley, B. D. (1981). Spatial Statistics. *New York: Wiley & Sons*, 252 Rouhani, S., & Wackernagel, H. (1990). Multivariate geostatistical approach to spacetime data analysis. *Water Resources Research*, 26(4), 585 –591.

[76] Said, S., & Dickey, D. (1984). *Testing for unit roots in autoregressive moving-average models with unknown order*. *Biometrika*, 71, 599 –607.

[77] Schielzeth, H., Nakagawa, S. (2013). Nested by design: model fitting and interpretation in a mixed model era. *Methods in Ecology Evolution*, 4(1), 14 –24.

[78] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461 –464.

[79] Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components*. Hoboken: John Wiley & Sons.

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

[81] Silva-Santos, P., Pardal, M. Â., Lopes, R. J., Múrias, T., & Cabral, J. A. (2008). Testing the Stochastic Dynamic Methodology (StDM) as a management tool in a shallow temperate estuary of south Europe (Mondego, Portugal). *Ecological Modelling*, 210(4), 377 –402.

[82] Thiessen, A. (1911). Precipitation Averages for Large Areas. *Monthly Weather Review*, 39(7), 1082 –1084.

[83] Thisted, R. A. (1988). *Elements of Statistical Computing*. London: Chapman & Hall.

[84] Trigo, R. M., PozoVázquez, D., Osborn, T. J., CastroDíez, Y., GámizFortis, S., & EstebanParra, M. J. (2004). North Atlantic Oscillation influence on precipitation, river flow and water resources in the Iberian Peninsula. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 24(8), 925 –944.

[85] Turkman, M. A. A., & Silva, G. L. (2000). Modelos Lineares Generalizados da teoria a prática. Lisboa: VIII Congresso Anual da Sociedade Portuguesa de Estatística.

[86] van Dijk, G. M., van Liere, L., Admiraal, W., Bannink, B. A., & Cappon, J. J. (1994). Present state of the water quality of european rivers and implications for management. *Science of the Total Environment*, 145(1-2), 187 –195.

[87] Verbeke, G., & Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*, New York: Springer.

[88] Voronoi, G. (1908). Nouvelles applications des parametres continus a la theorie des formes quadratiques. Deuxieme Memoire: Recherche sur les paralleloedres primitifs. *Journal Reine Angew. Math.*, 134, 198 –287.

[89] Wu, L. (2009). *Mixed effects models for complex data*. Chapman & Hall/CRC.

[90] Yaglom, A. M. (1962), *An introduction to the theory of stationary random functions: Englewood Cliffs*, Pretice-Hall, Inc.

[91] Zuur A. F., Ieno E. N., Walker N. J., Saveliev A. A., Smith G.M. (2009). *Mixed Effects Models and Extensions in Ecology with R*. New York: Springer.

Apêndice A

Geoestatística

A.1 Representações Gráficas da Precipitação em cada Estação de Amostragem

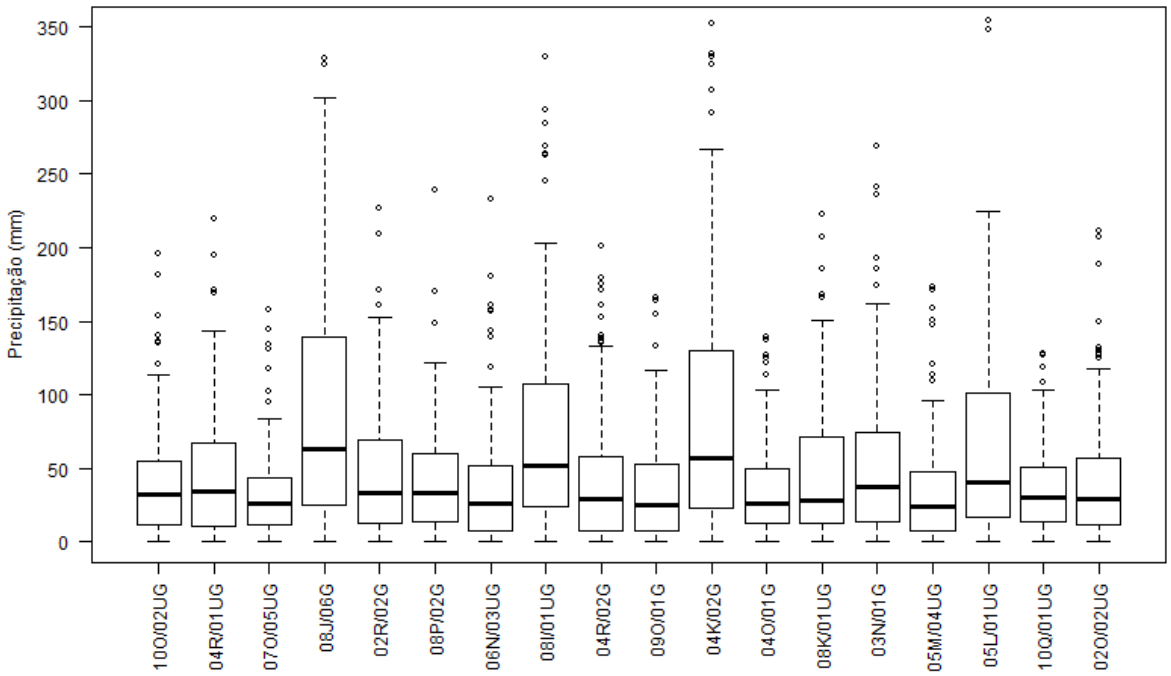


Figura A.1: Diagrama em caixa de bigodes das série de Precipitação, nas 18 estações de amostragem, no período observado.

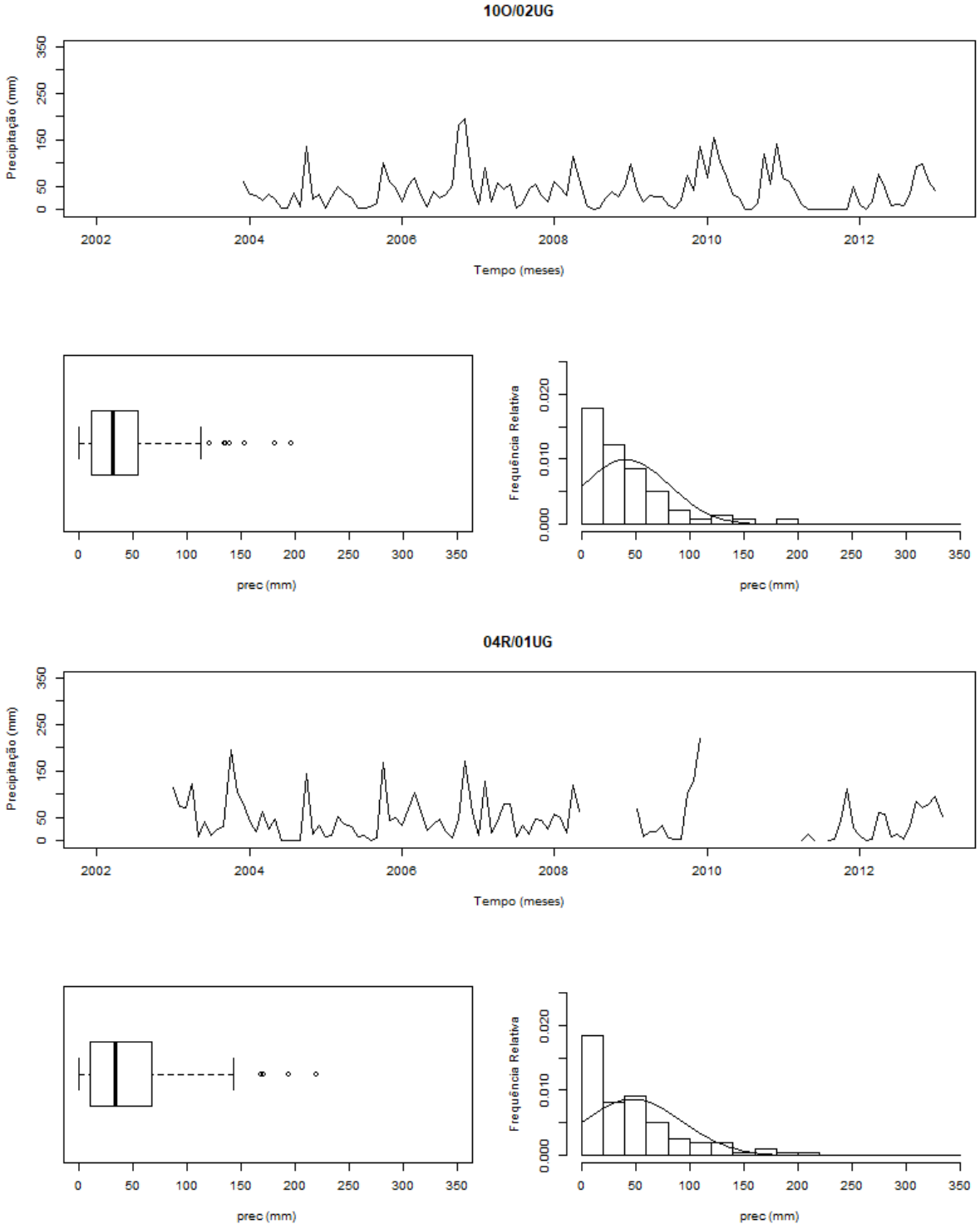


Figura A.2: Representações gráficas das séries temporais da precipitação e dos respetivos diagramas em caixas de bigodes e histogramas, no período observado.

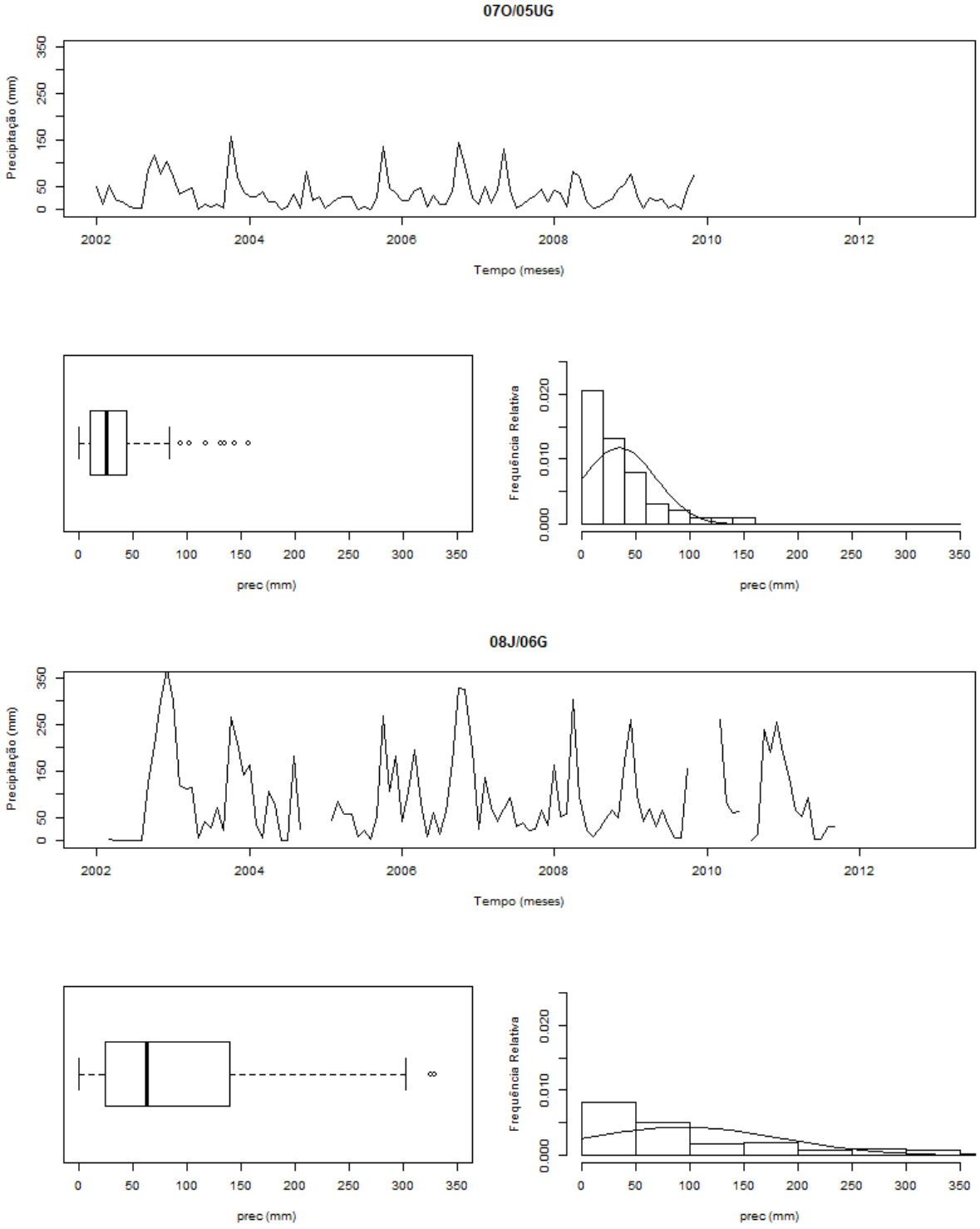


Figura A.3: Representações gráficas das séries temporais da precipitação e dos respetivos diagramas em caixas de bigodes e histogramas, no período observado.

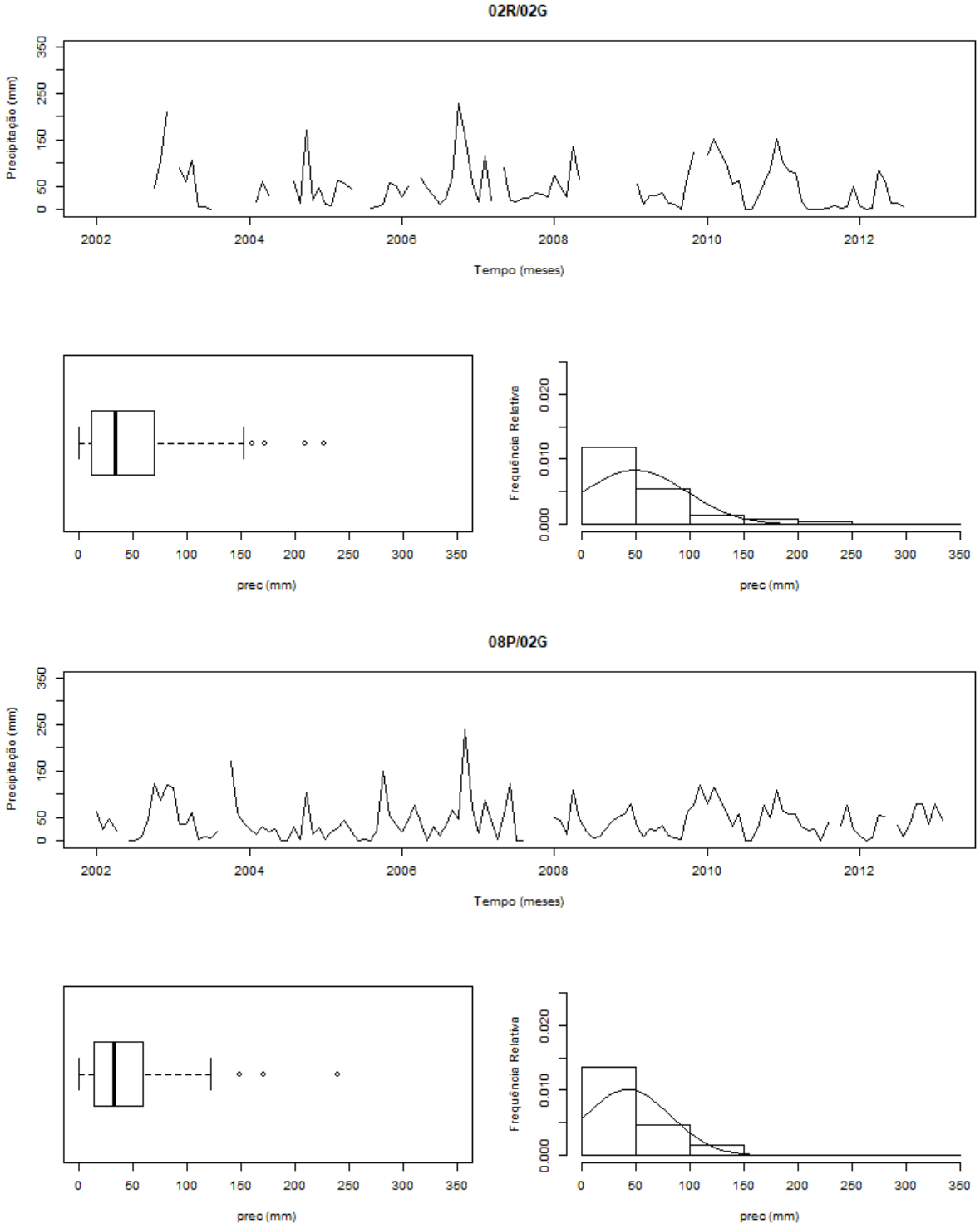


Figura A.4: Representações gráficas das séries temporais da precipitação e dos respetivos diagramas em caixas de bigodes e histogramas, no período observado.

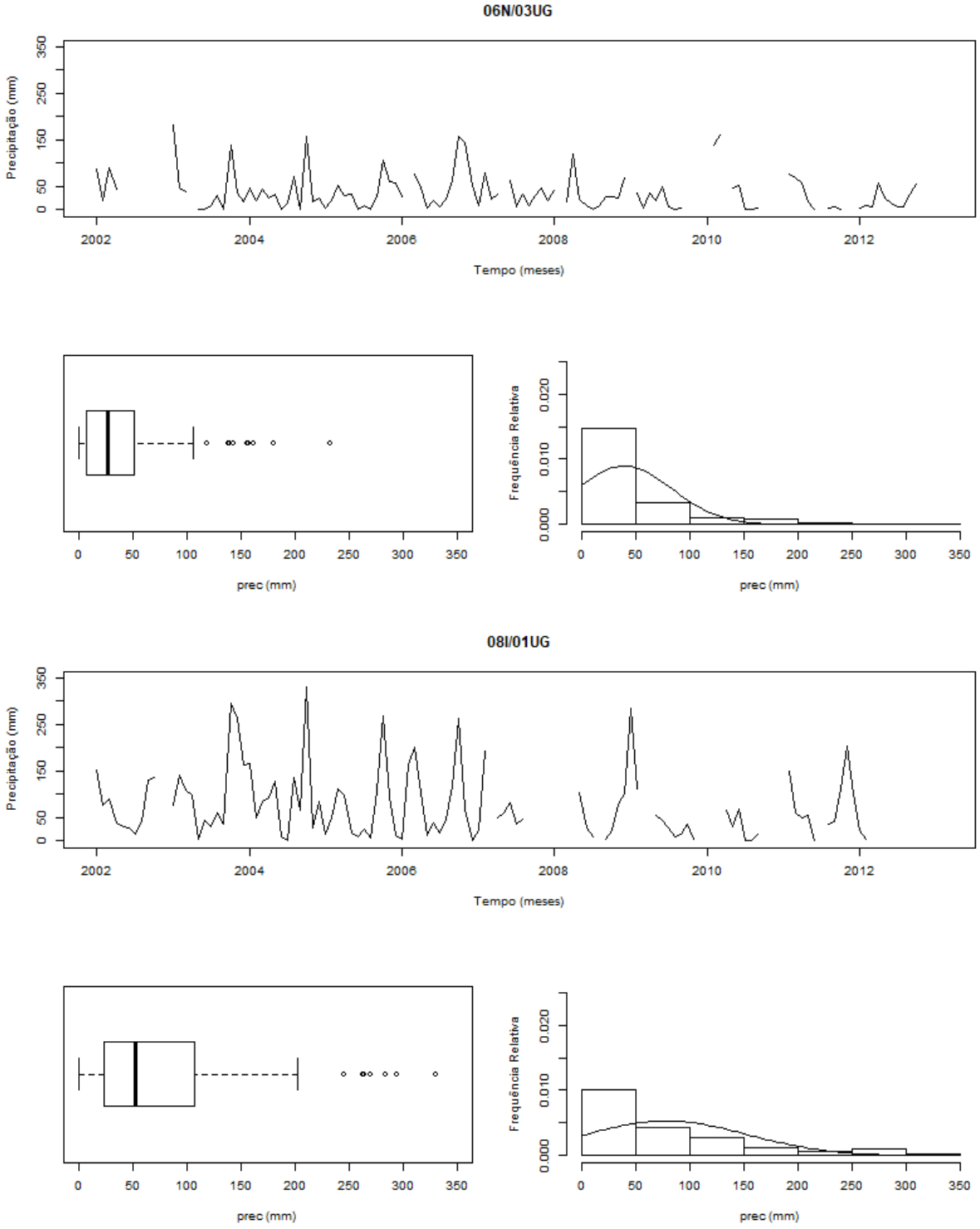


Figura A.5: Representações gráficas das séries temporais da precipitação e dos respetivos diagramas em caixas de bigodes e histogramas, no período observado.

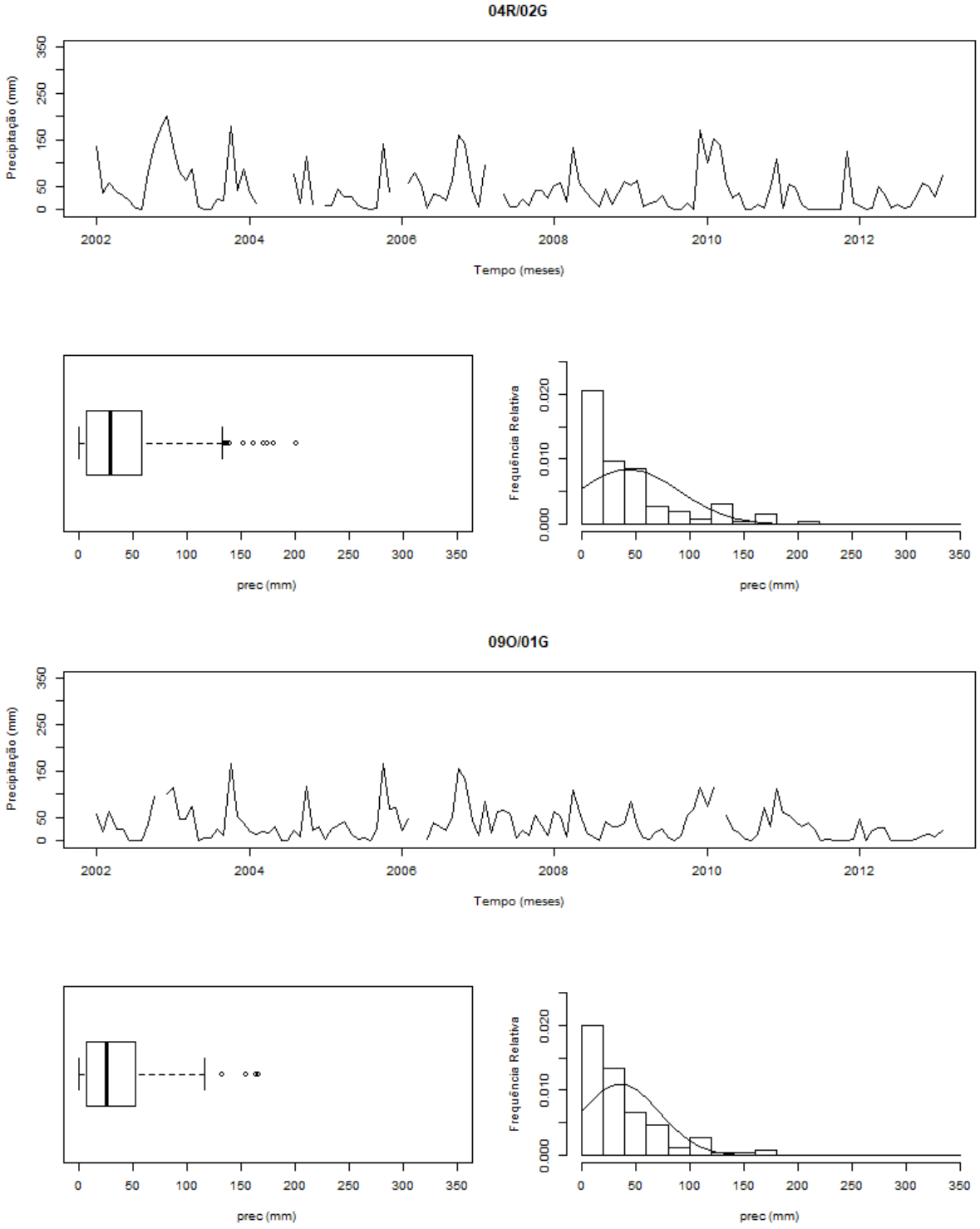


Figura A.6: Representações gráficas das séries temporais da precipitação e dos respetivos diagramas em caixas de bigodes e histogramas, no período observado.

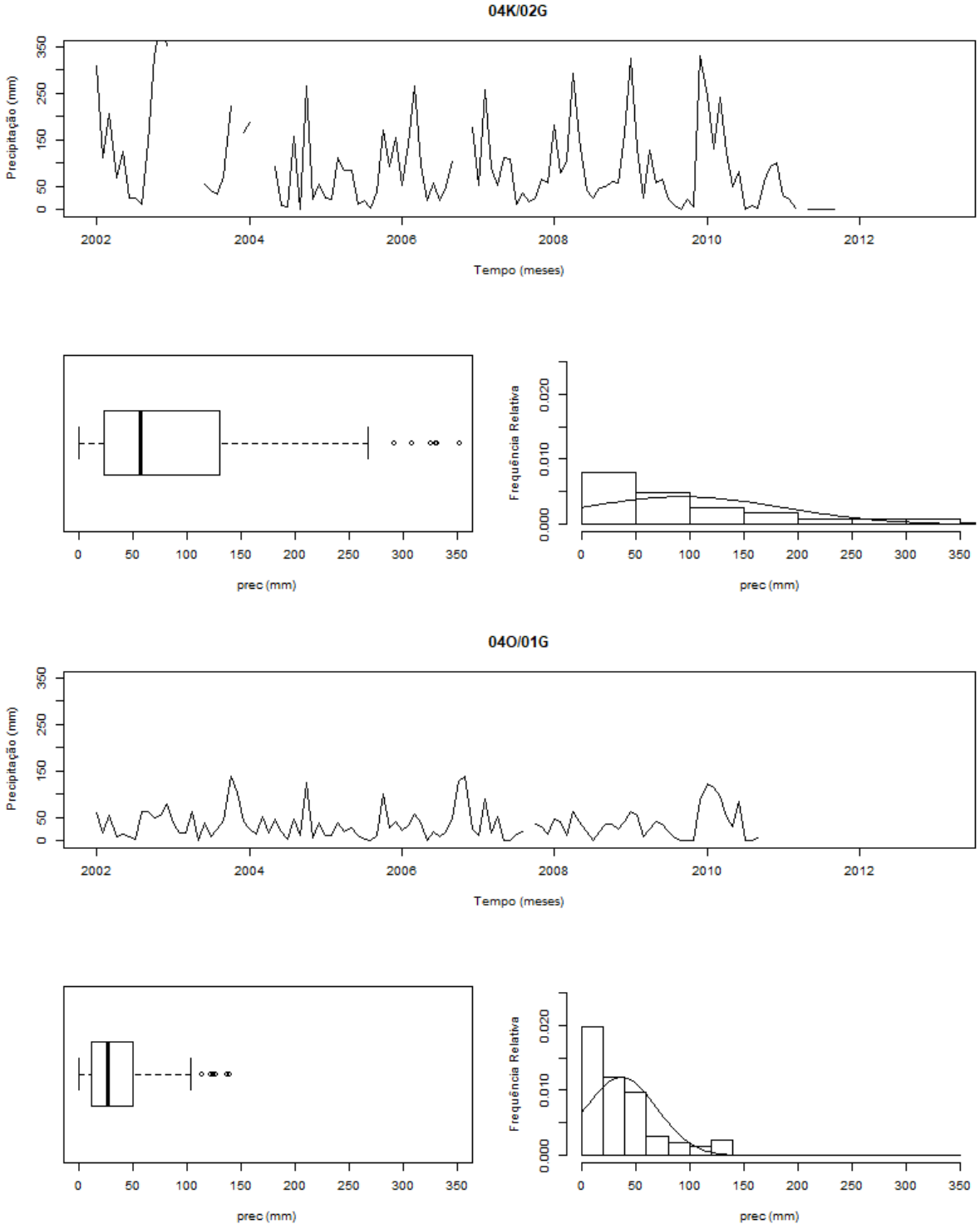


Figura A.7: Representações gráficas das séries temporais da precipitação e dos respetivos diagramas em caixas de bigodes e histogramas, no período observado.

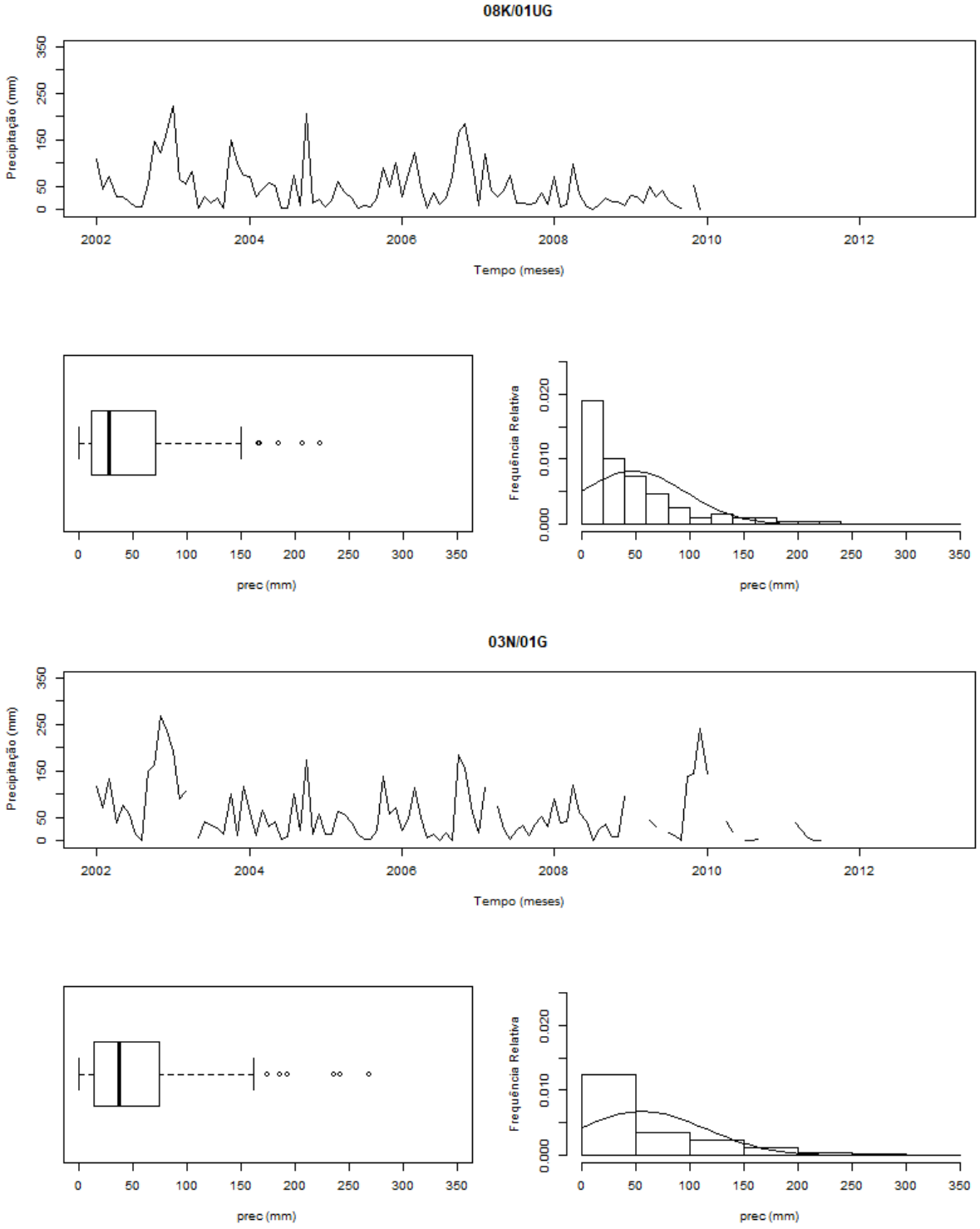


Figura A.8: Representações gráficas das séries temporais da precipitação e dos respetivos diagramas em caixas de bigodes e histogramas, no período observado.

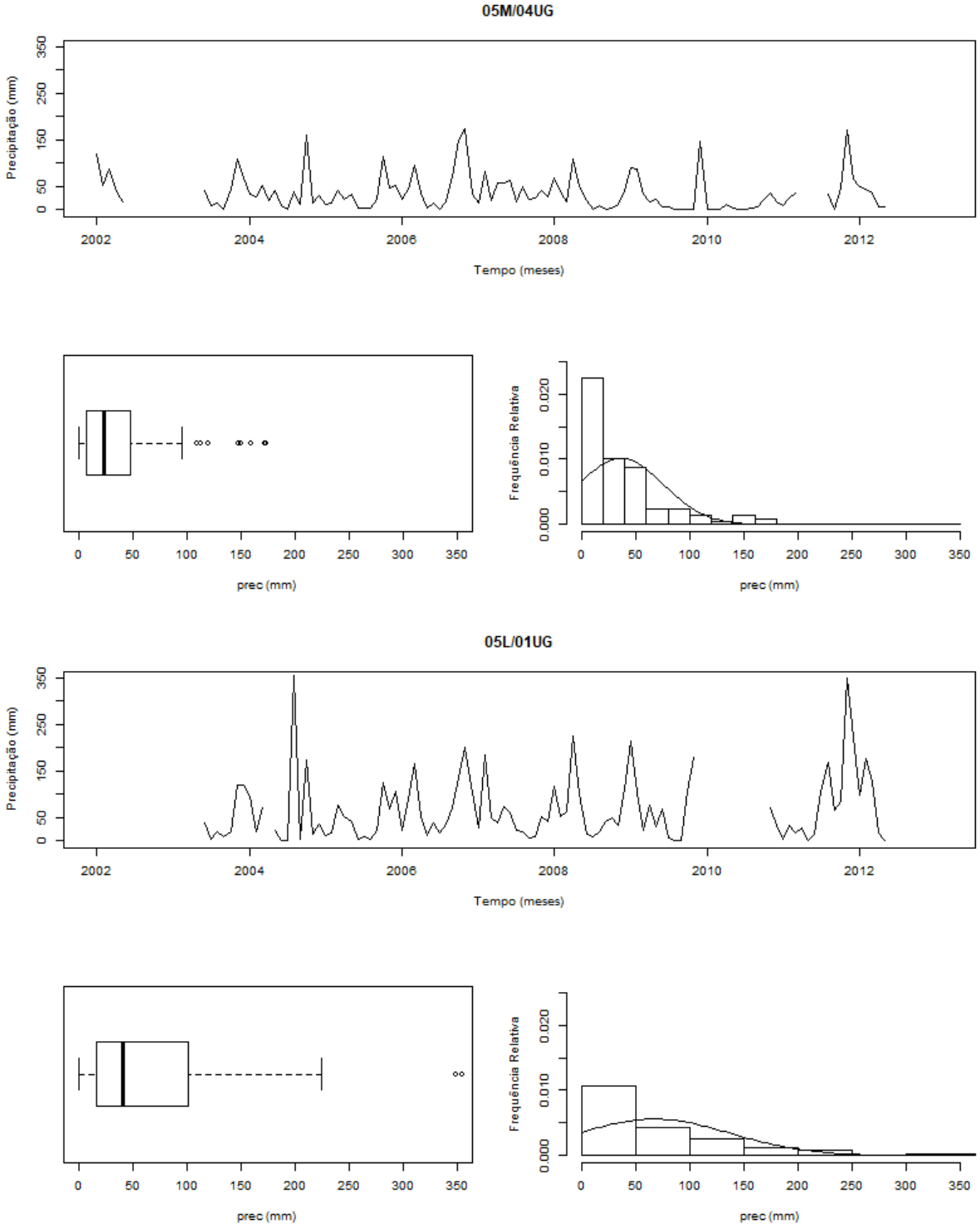


Figura A.9: Representações gráficas das séries temporais da precipitação e dos respetivos diagramas em caixas de bigodes e histogramas, no período observado.

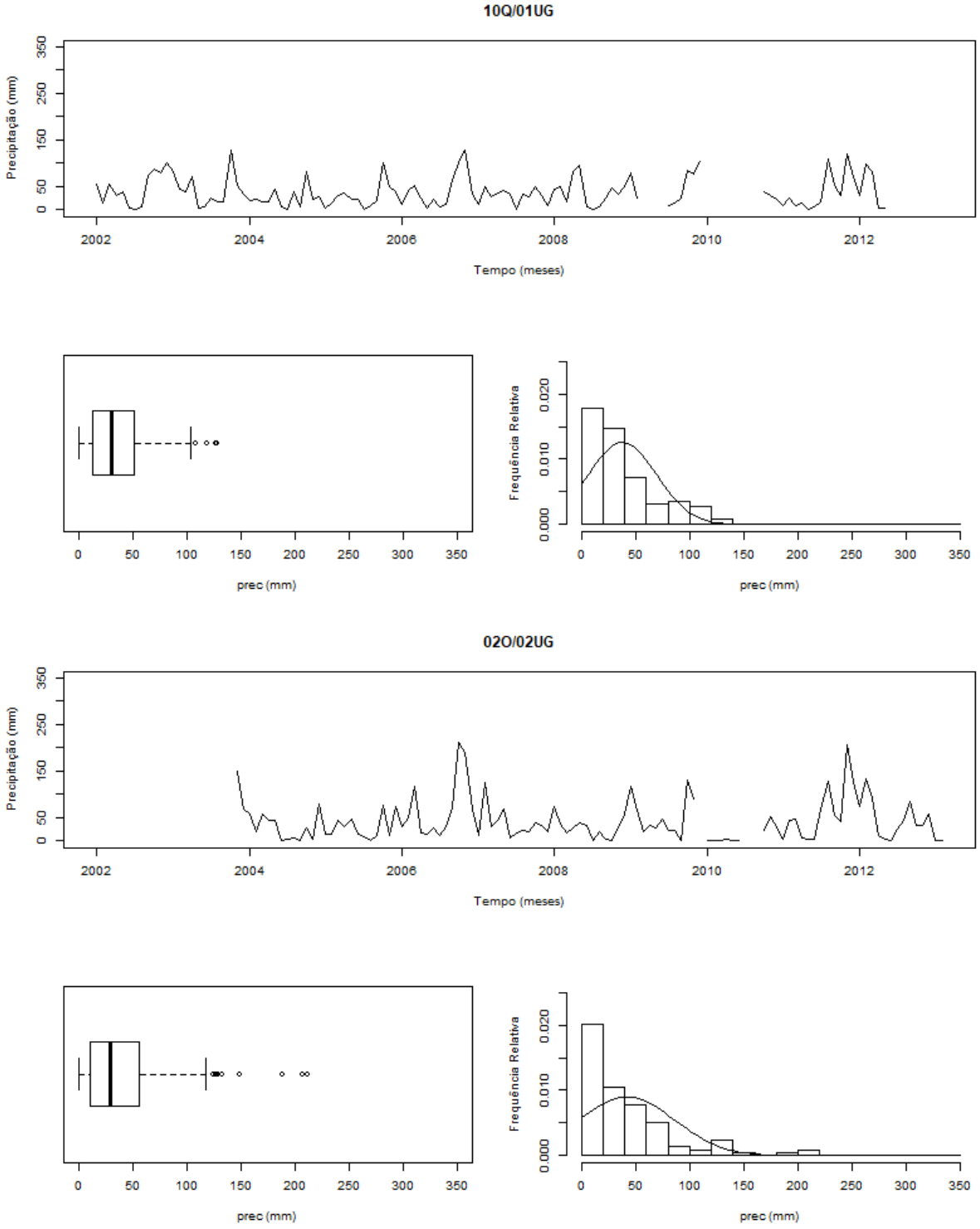


Figura A.10: Representações gráficas das séries temporais da precipitação e dos respetivos diagramas em caixas de bigodes e histogramas, no período observado.

A.2 Predição

Figura A.11: Representações das superfícies estimadas da precipitação, no mês de fevereiro, nos anos de 2002 até 2013.

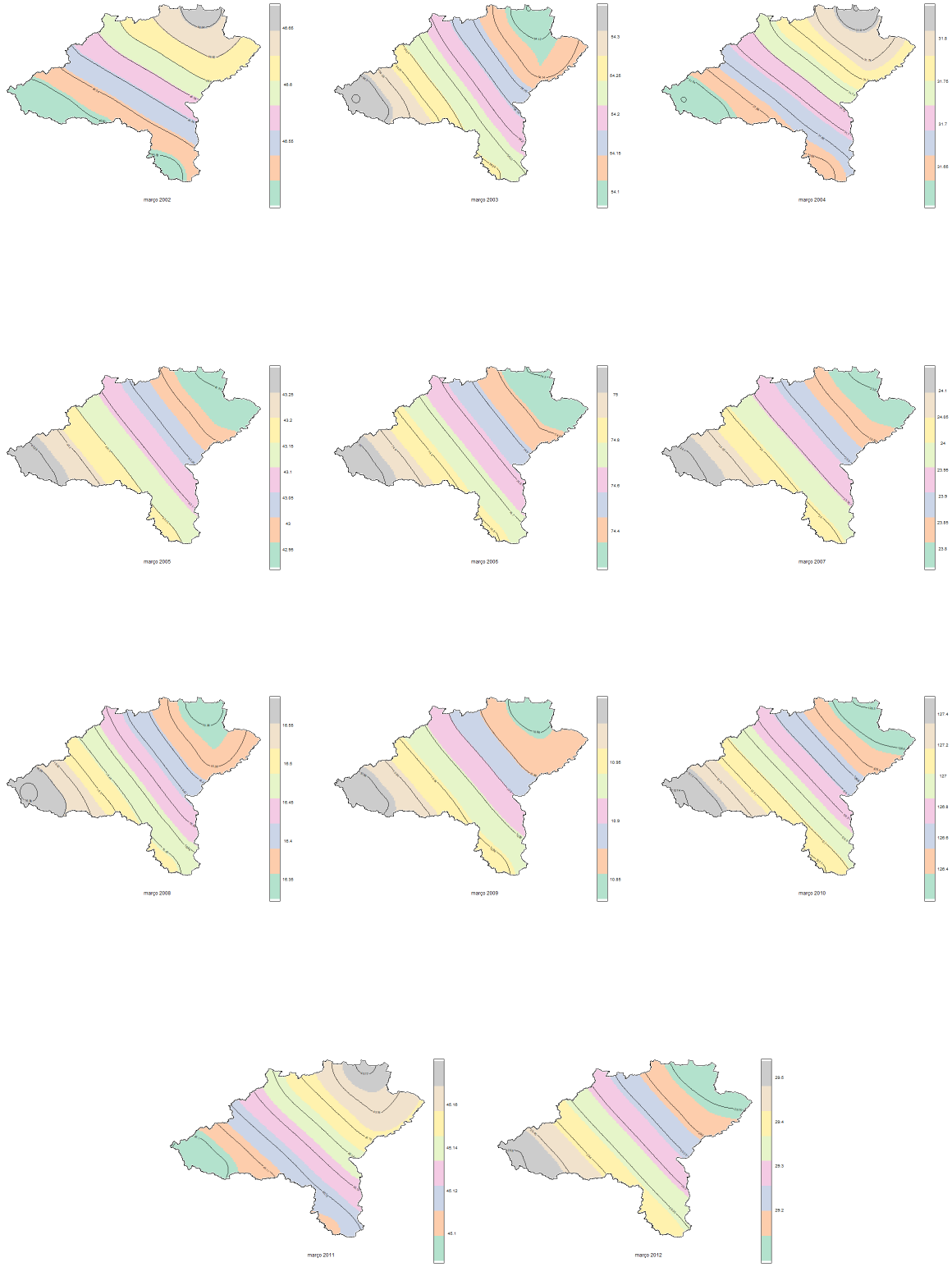


Figura A.12: Representações das superfícies estimadas da precipitação, no mês de março, nos anos de 2002 até 2012.

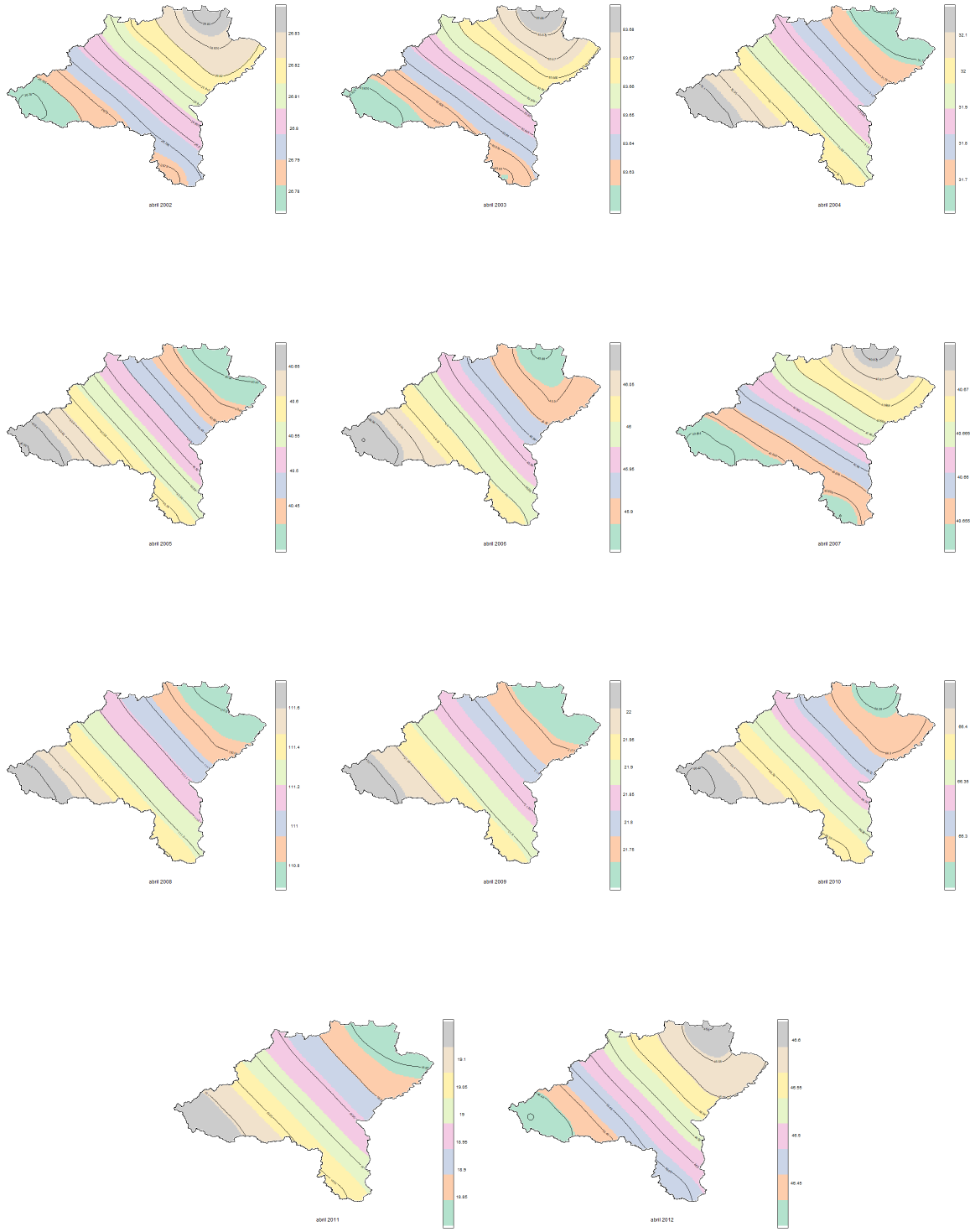


Figura A.13: Representações das superfícies estimadas da precipitação, no mês de abril, nos anos de 2002 até 2012.

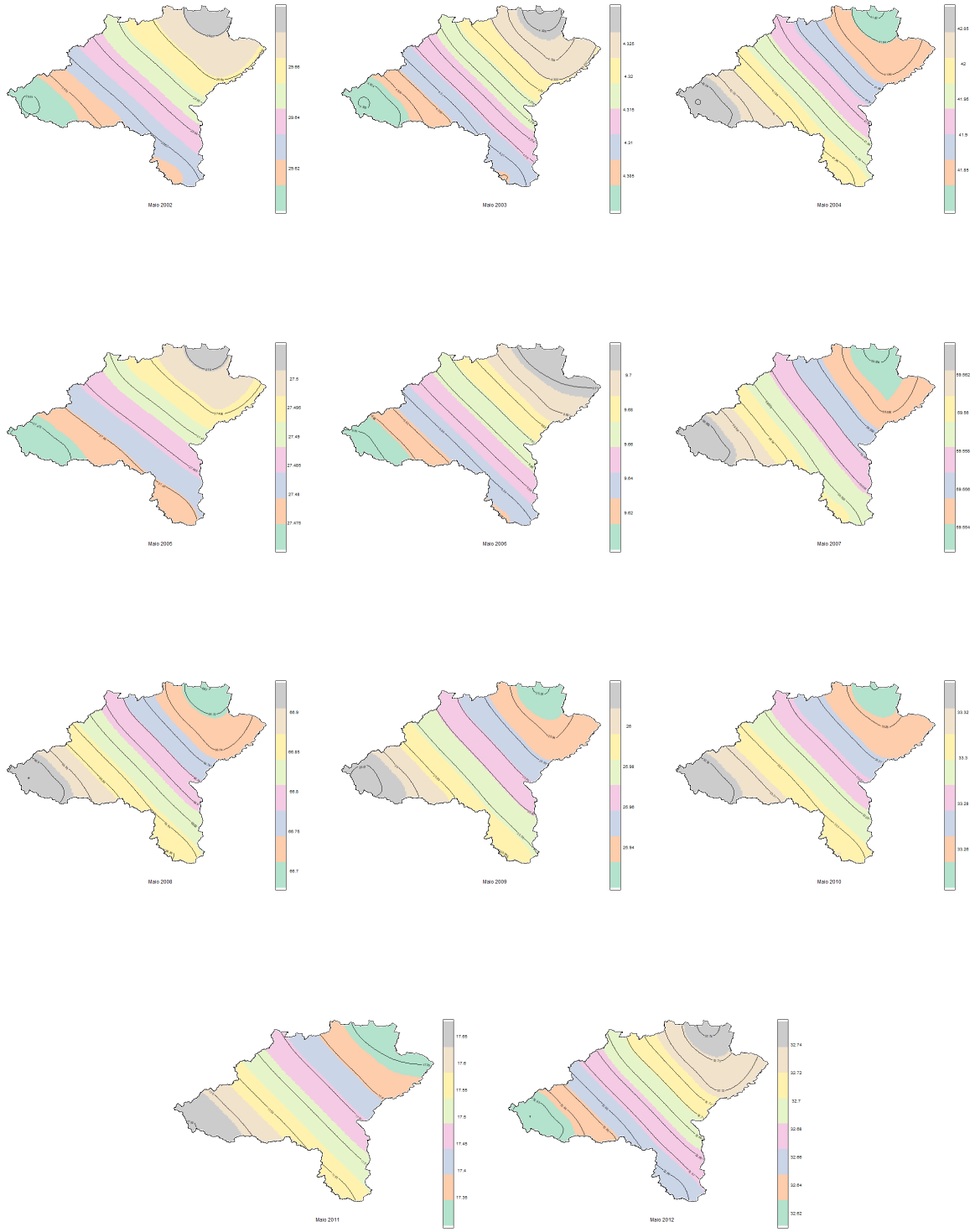


Figura A.14: Representações das superfícies estimadas da precipitação, no mês de maio, nos anos de 2002 até 2012.

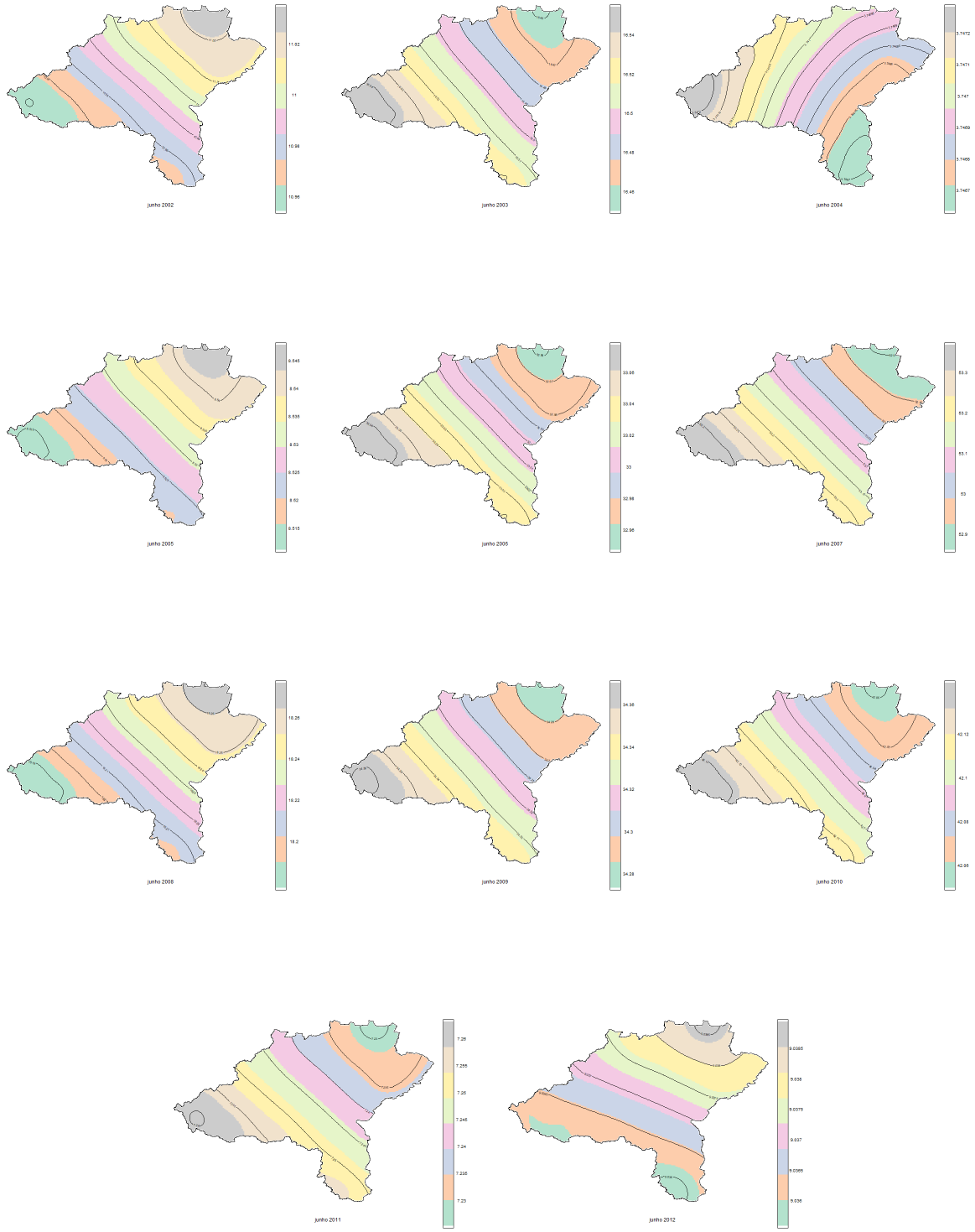


Figura A.15: Representações das superfícies estimadas da precipitação, no mês de junho, nos anos de 2002 até 2012.

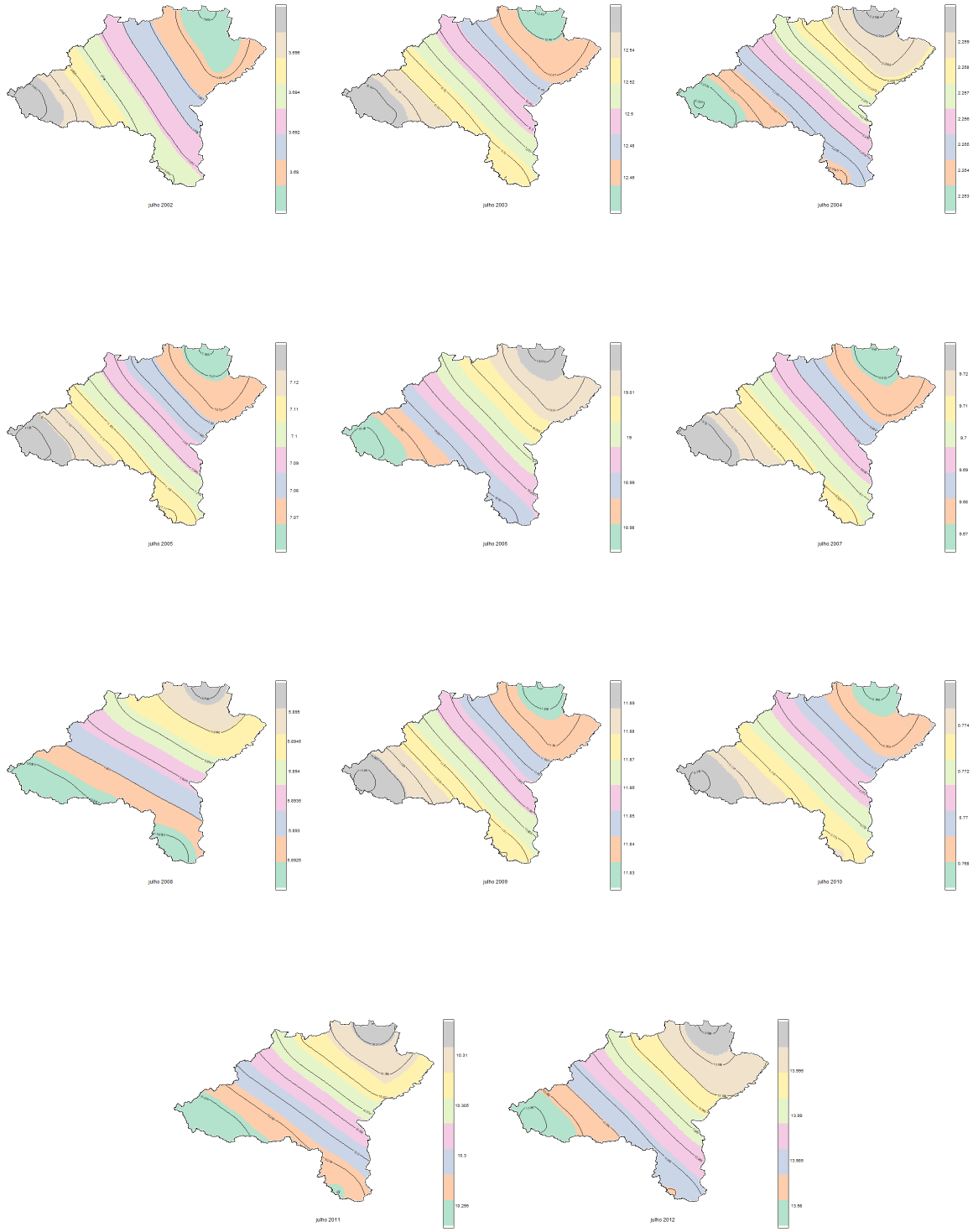


Figura A.16: Representações das superfícies estimadas da precipitação, no mês de julho, nos anos de 2002 até 2012.

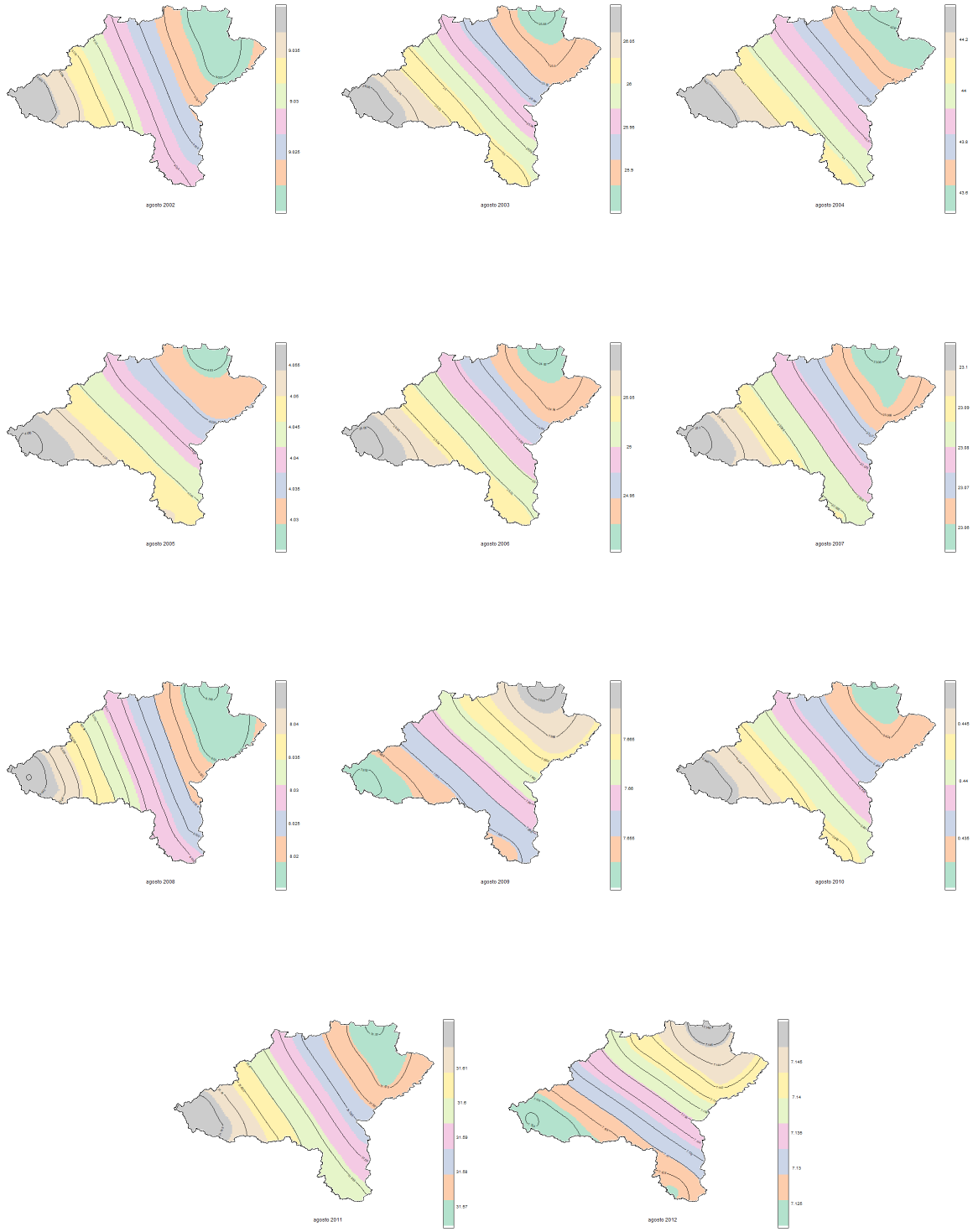


Figura A.17: Representações das superfícies estimadas da precipitação, no mês de agosto, nos anos de 2002 até 2012.

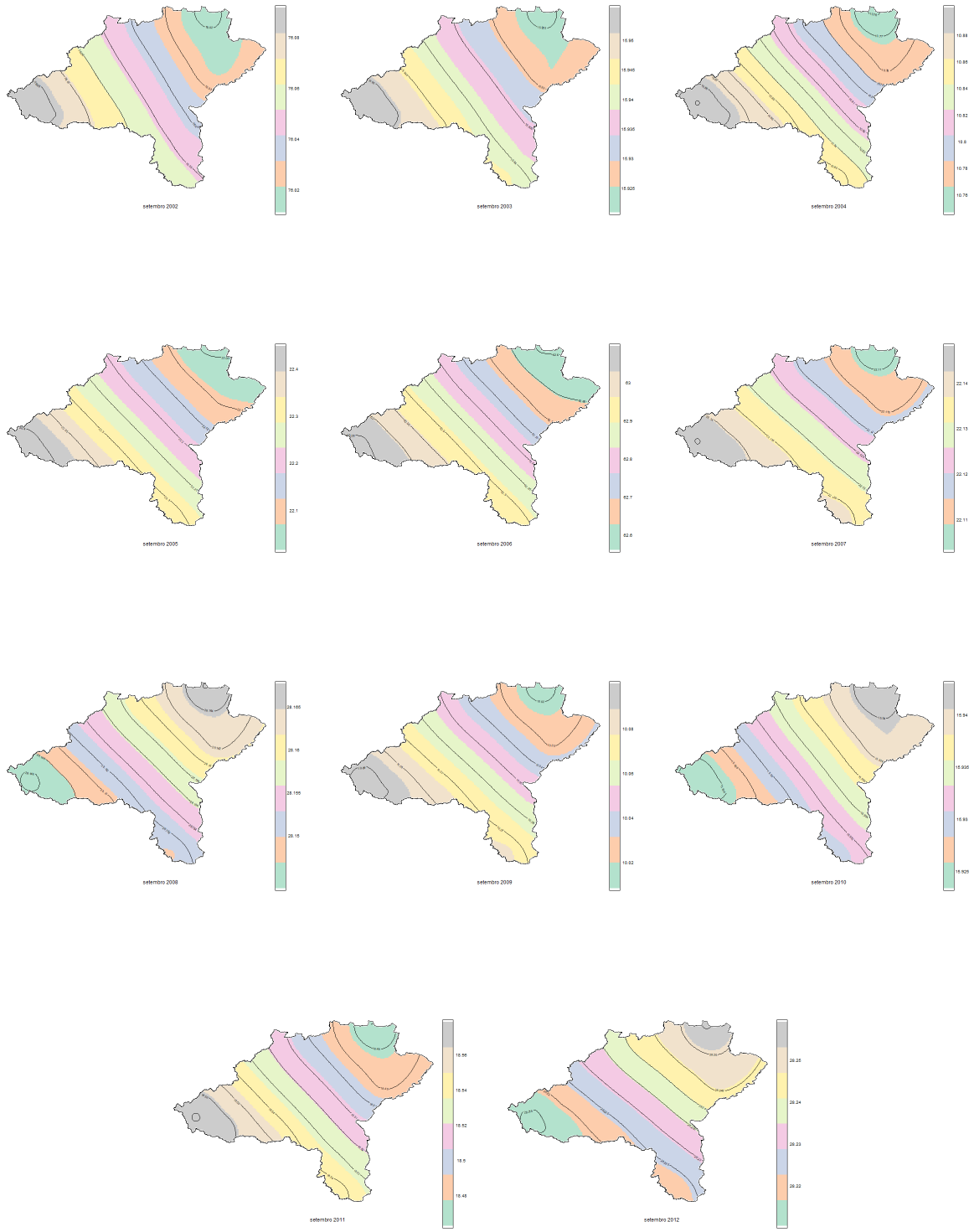


Figura A.18: Representações das superfícies estimadas da precipitação, no mês de setembro, nos anos de 2002 até 2012.

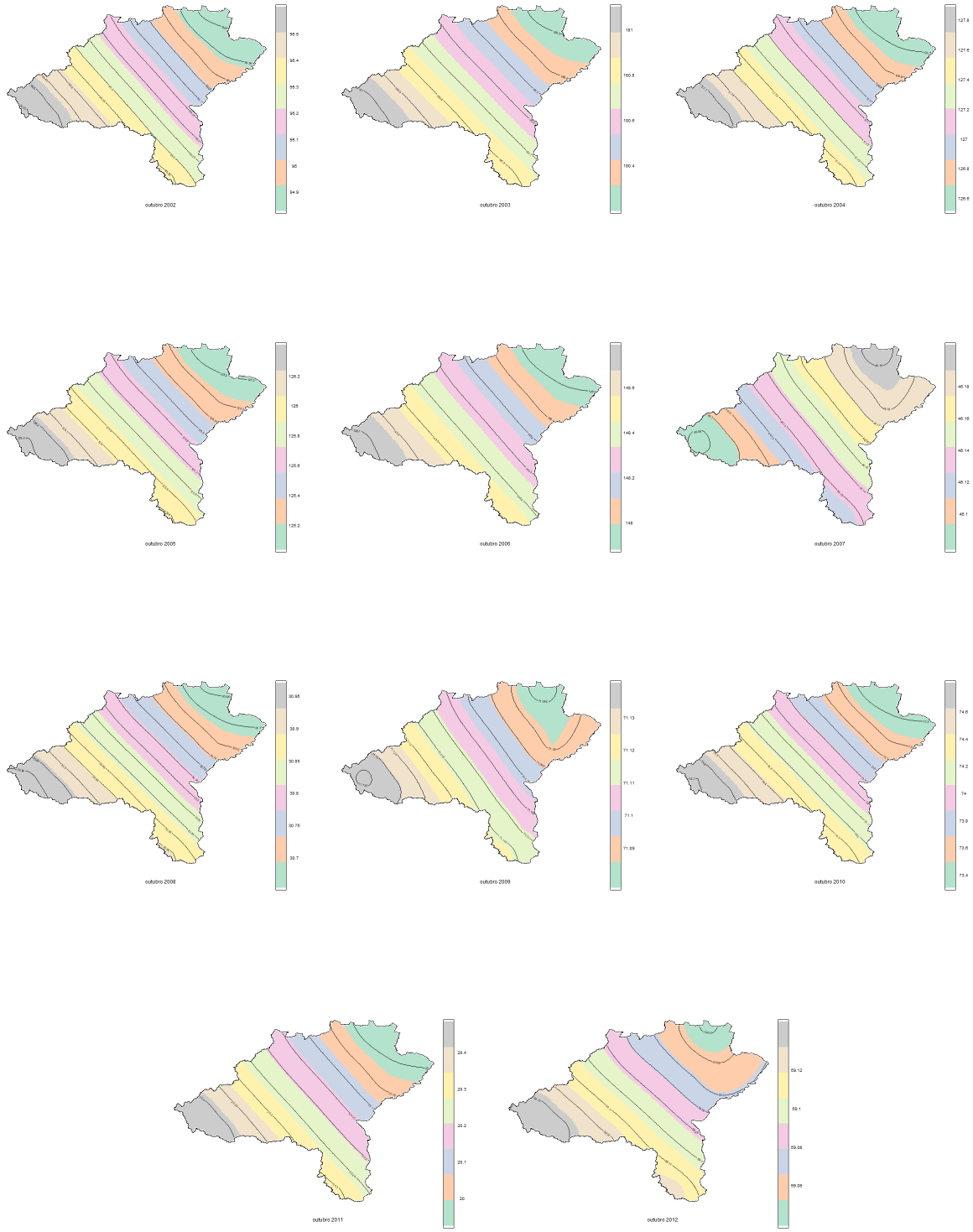


Figura A.19: Representações das superfícies estimadas da precipitação, no mês de outubro, nos anos de 2002 até 2012.

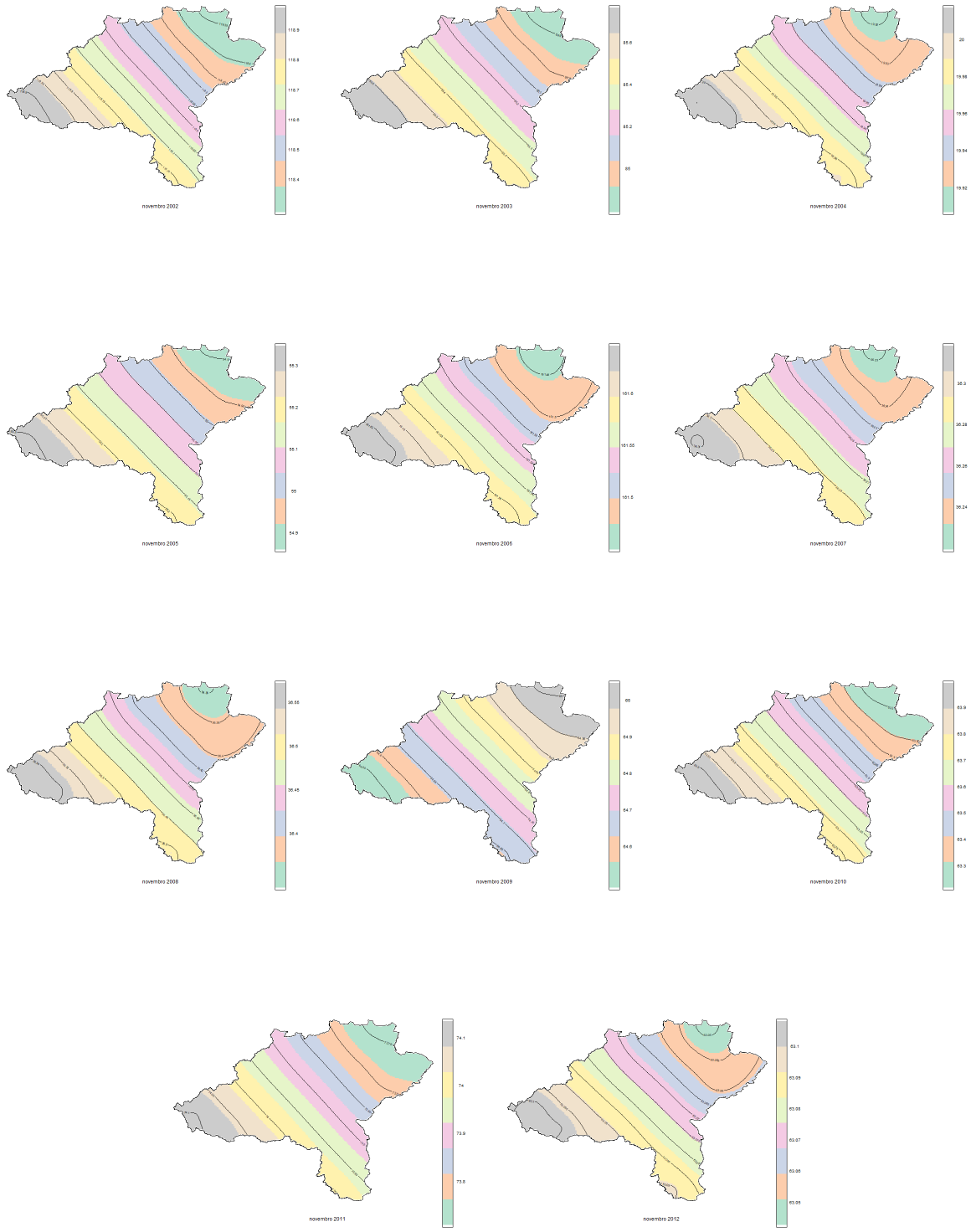


Figura A.20: Representações das superfícies estimadas da precipitação, no mês de novembro, nos anos de 2002 até 2012.

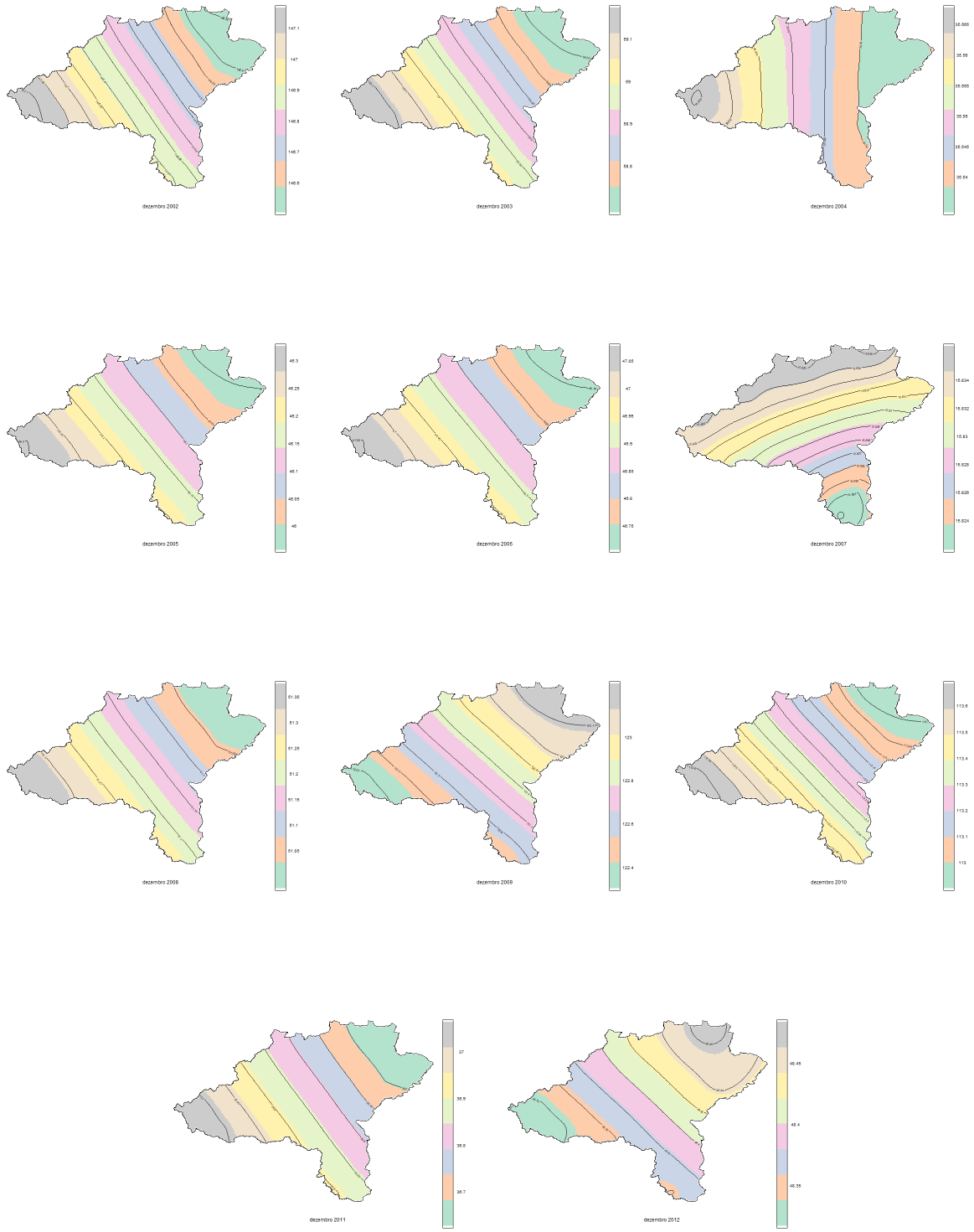


Figura A.21: Representações das superfícies estimadas da precipitação, no mês de dezembro, nos anos de 2002 até 2012.

A.3 Erros Estimados

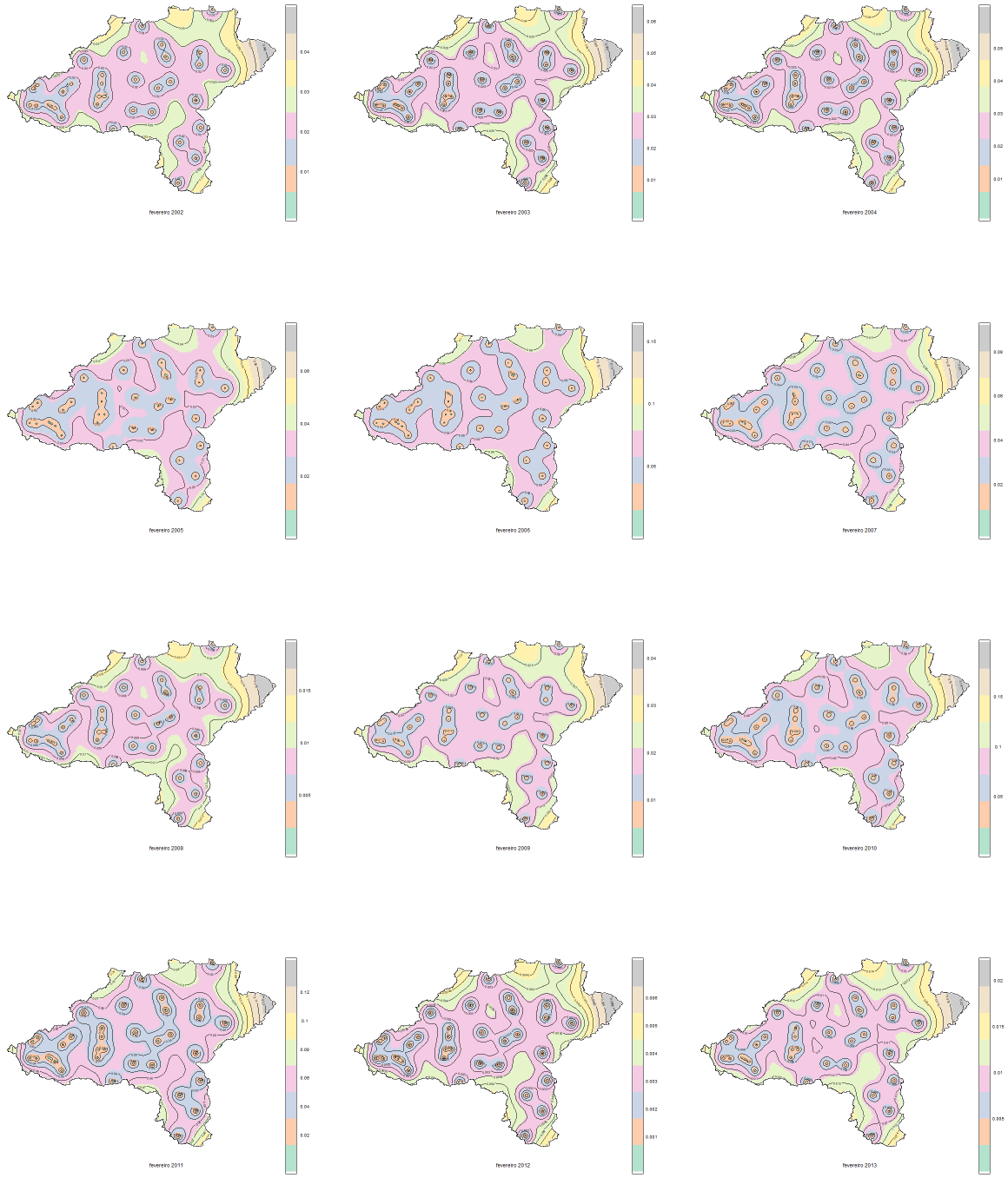


Figura A.22: Representações das superfícies de erros estimados da precipitação, no mês de fevereiro, nos anos de 2002 até 2013.

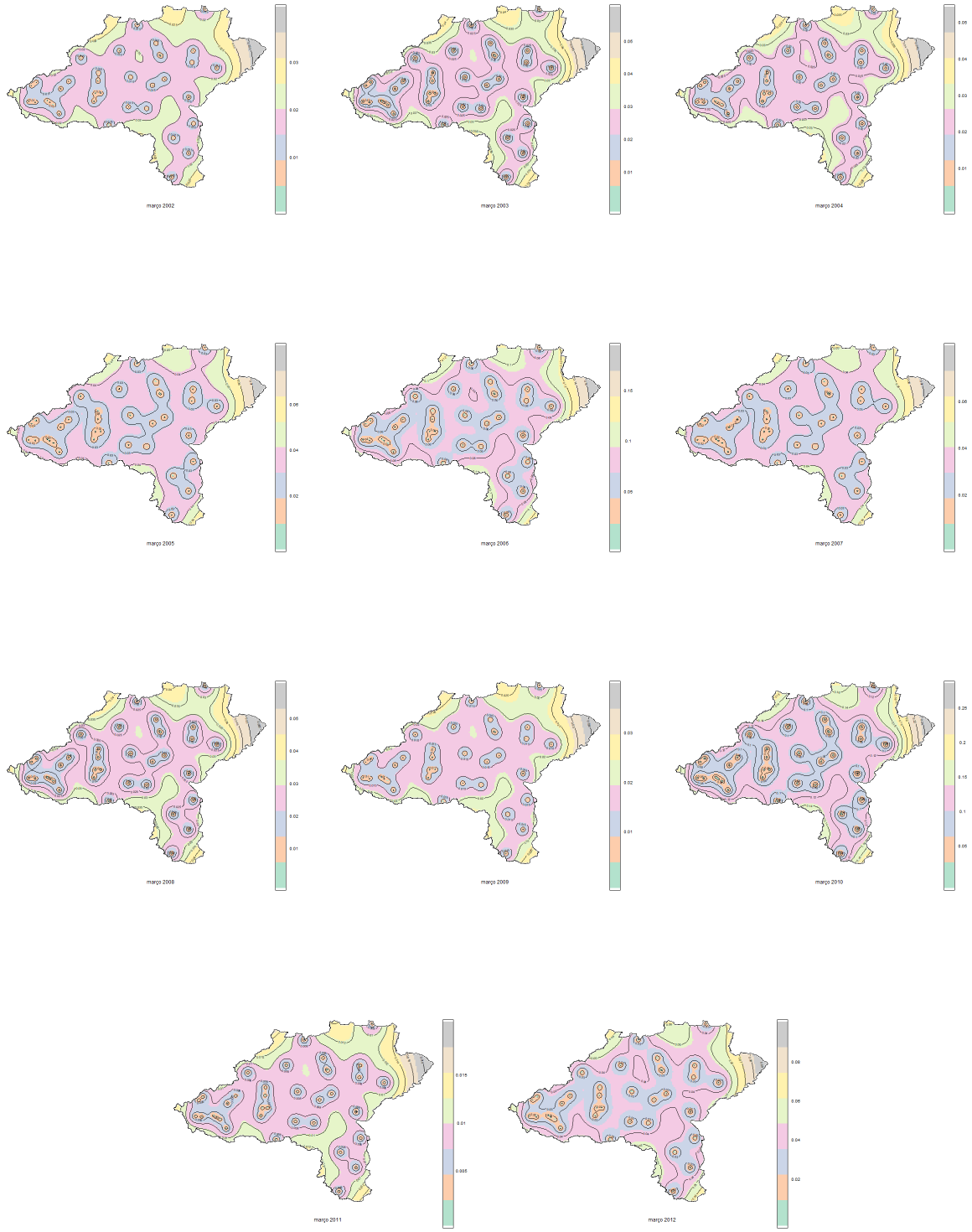


Figura A.23: Representações das superfícies de erros estimados da precipitação, no mês de março, nos anos de 2002 até 2012.

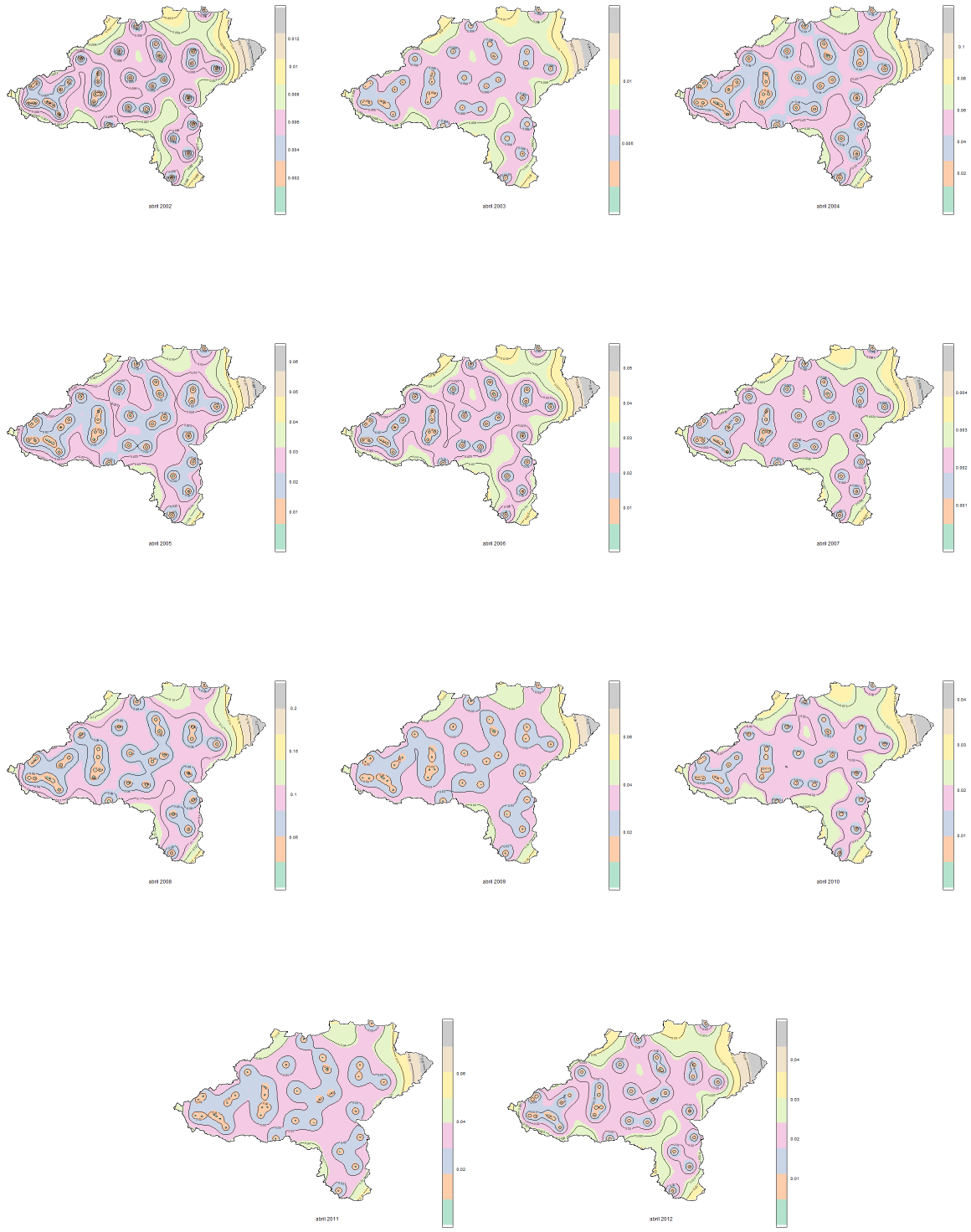


Figura A.24: Representações das superfícies de erros estimados da precipitação, no mês de abril, nos anos de 2002 até 2012.

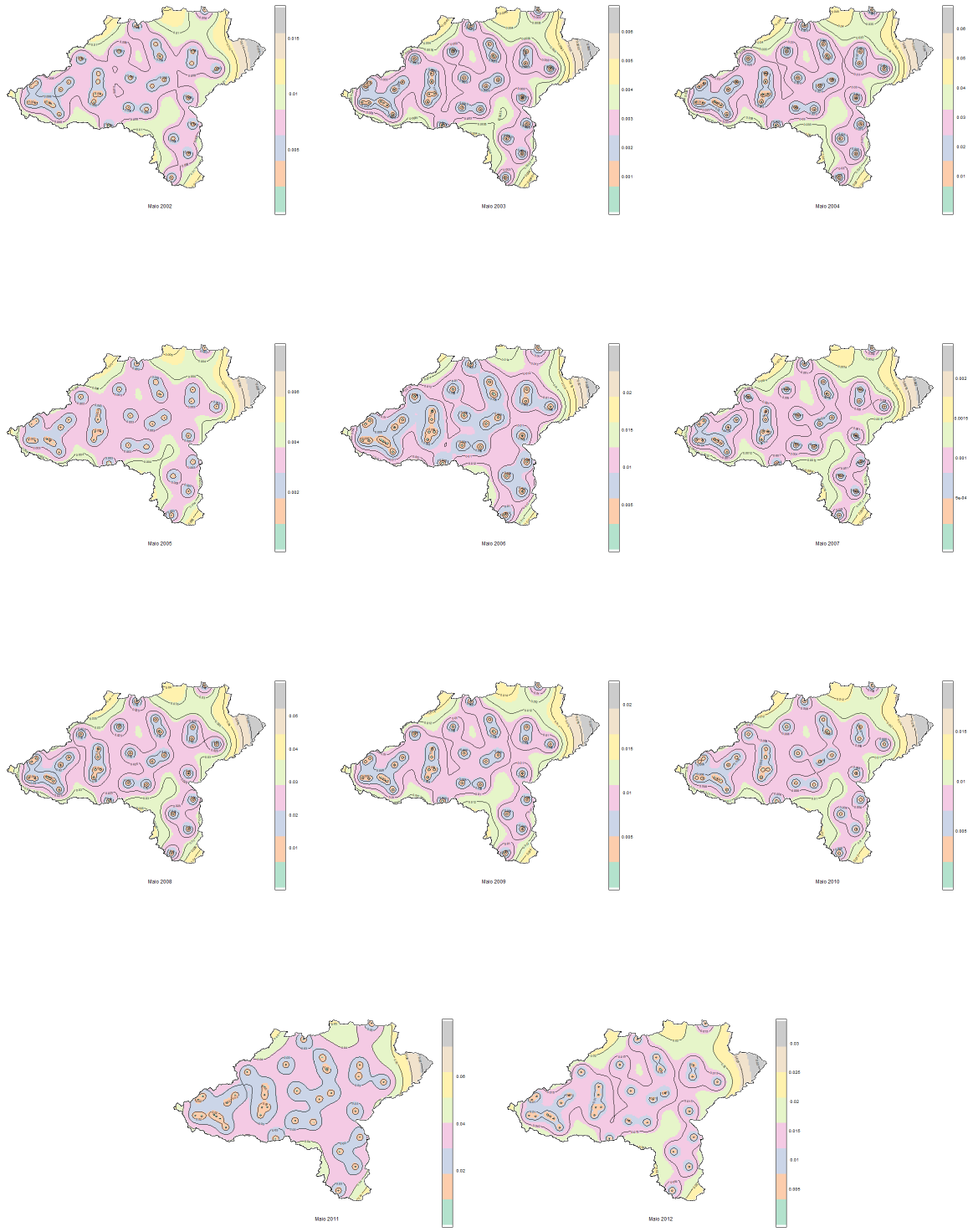


Figura A.25: Representações das superfícies de erros estimados da precipitação, no mês de maio, nos anos de 2002 até 2012.

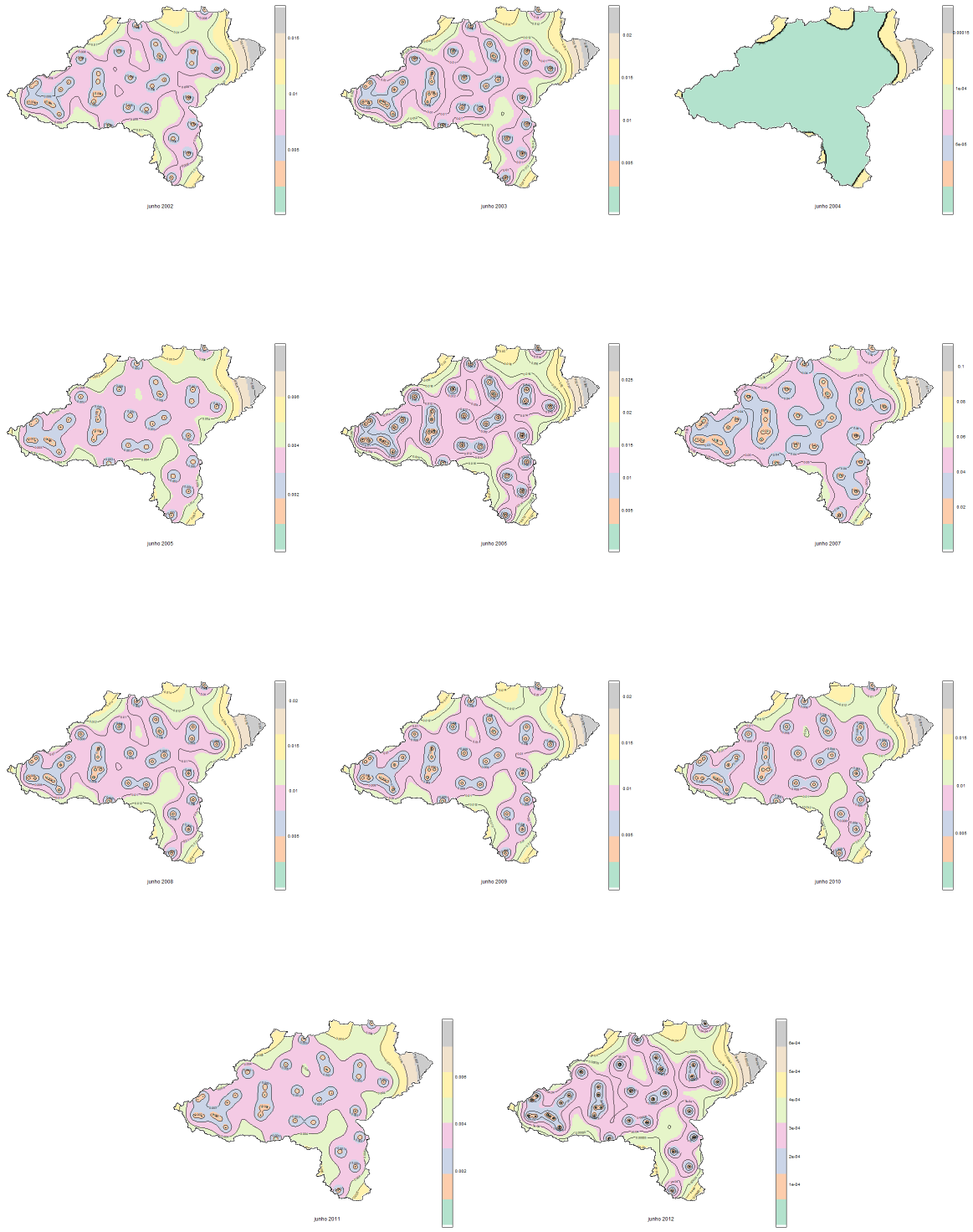


Figura A.26: Representações das superfícies de erros estimados da precipitação, no mês de junho, nos anos de 2002 até 2012.

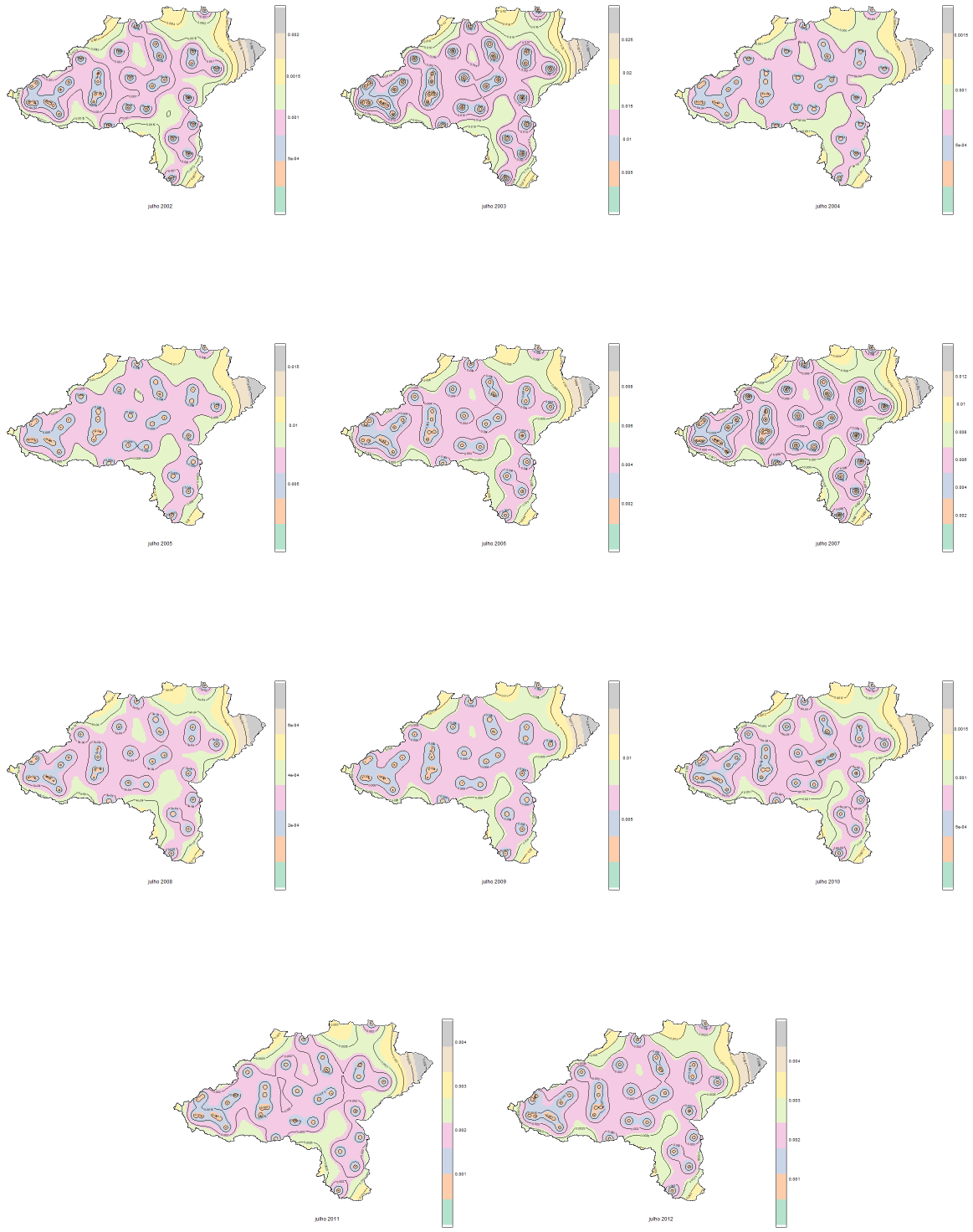


Figura A.27: Representações das superfícies de erros estimados da precipitação, no mês de julho, nos anos de 2002 até 2012.

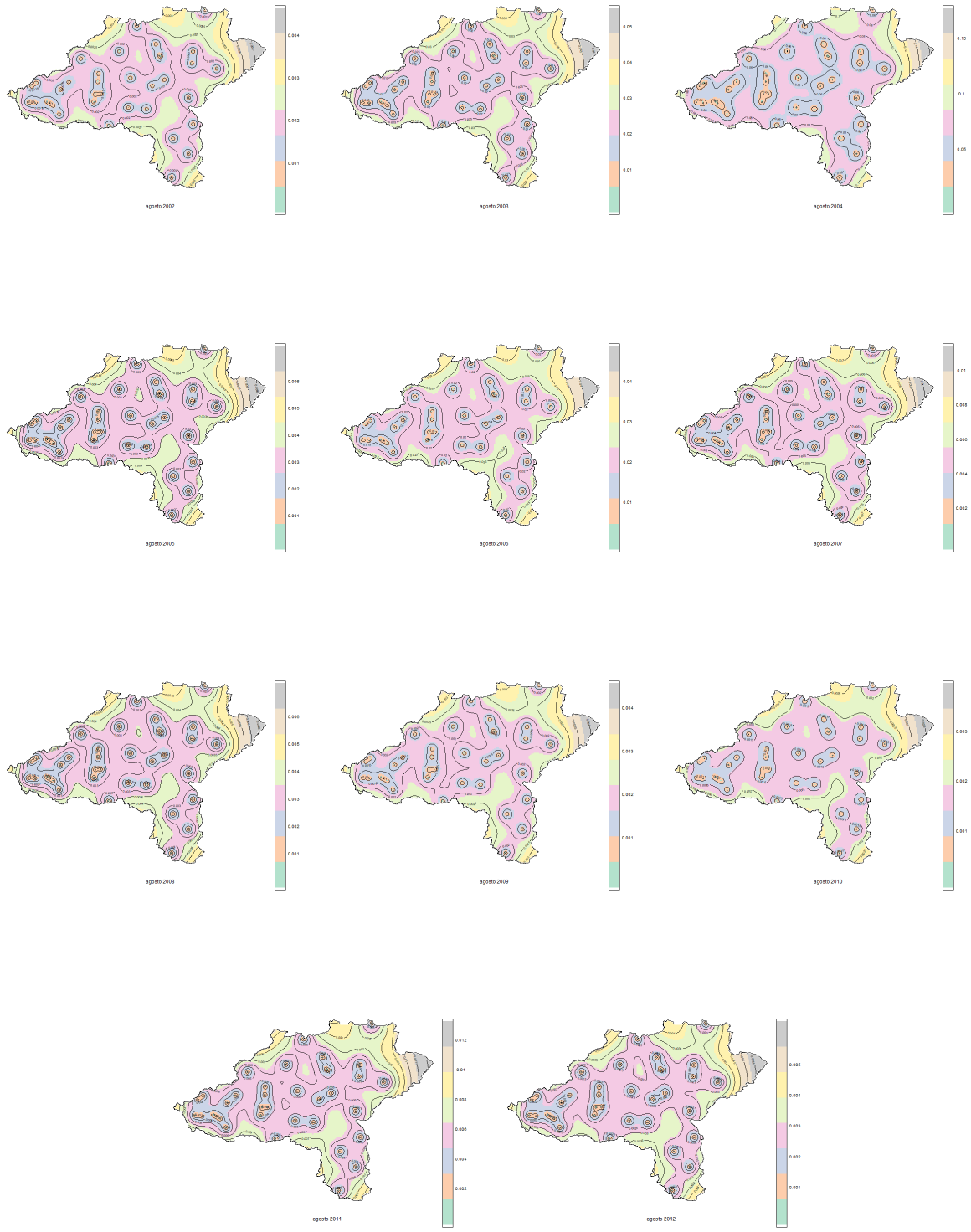


Figura A.28: Representações das superfícies de erros estimados da precipitação, no mês de agosto, nos anos de 2002 até 2012.

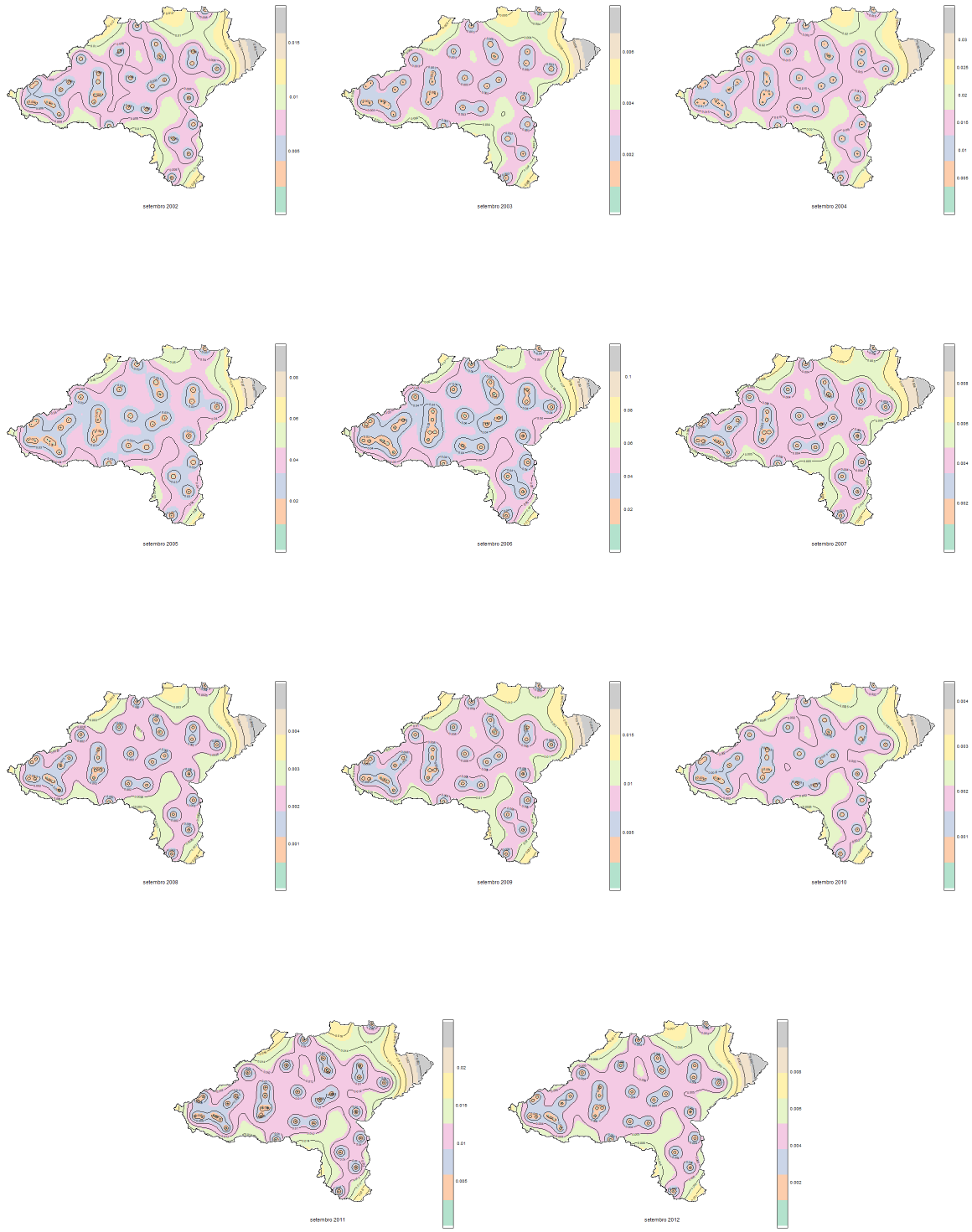


Figura A.29: Representações das superfícies de erros estimados da precipitação, no mês de setembro, nos anos de 2002 até 2012.

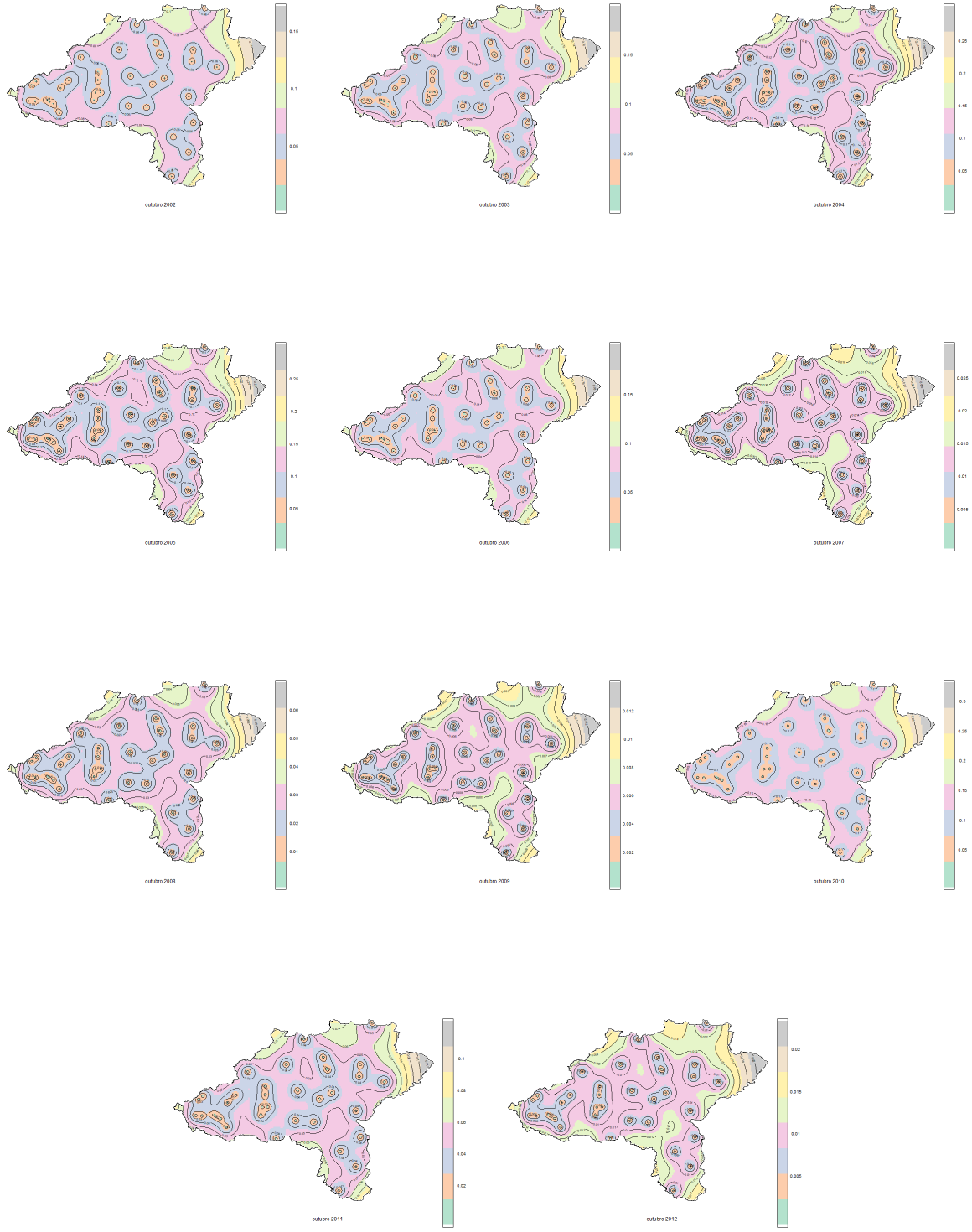


Figura A.30: Representações das superfícies de erros estimados da precipitação, no mês de outubro, nos anos de 2002 até 2012.

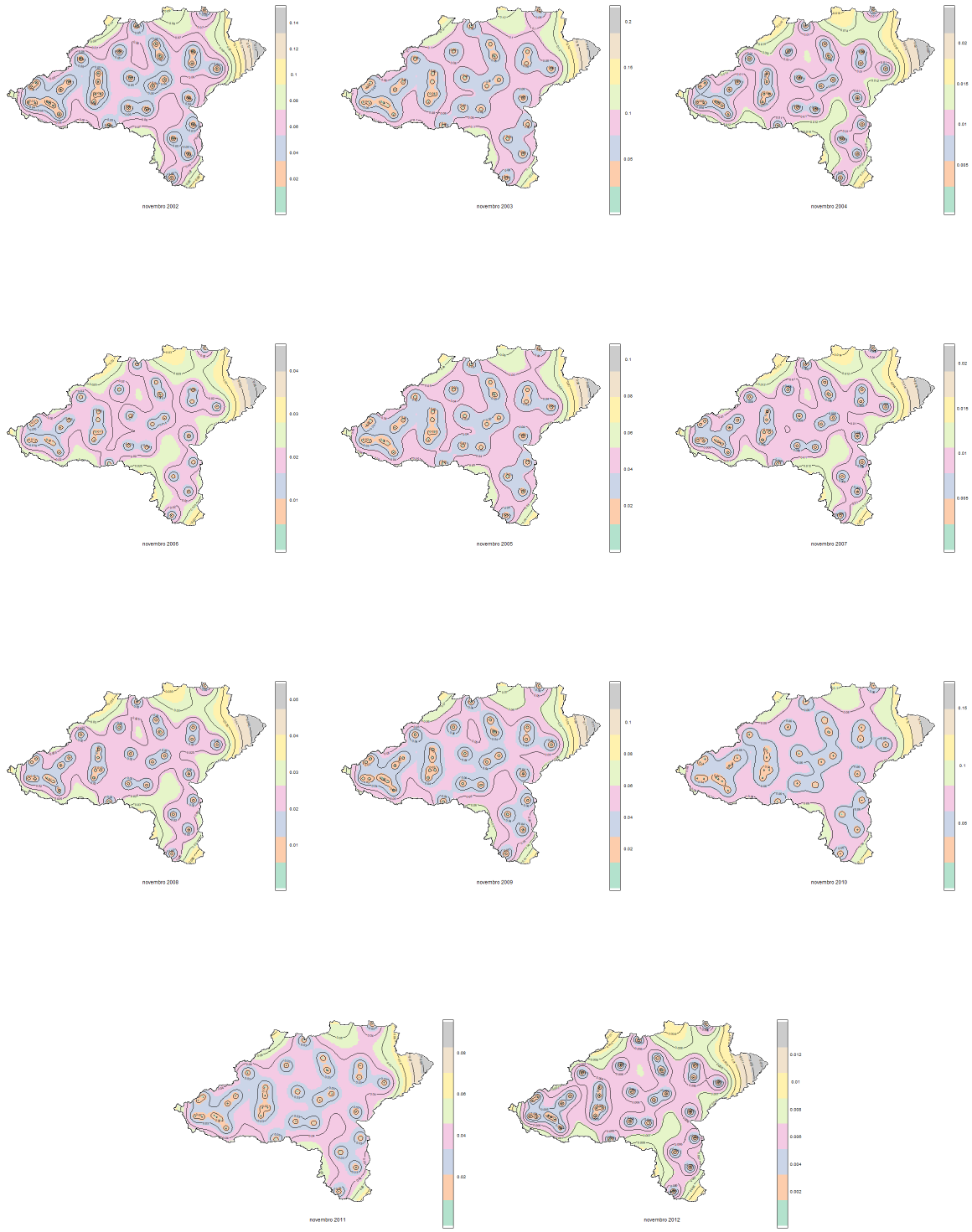


Figura A.31: Representações das superfícies de erros estimados da precipitação, no mês de novembro, nos anos de 2002 até 2012.

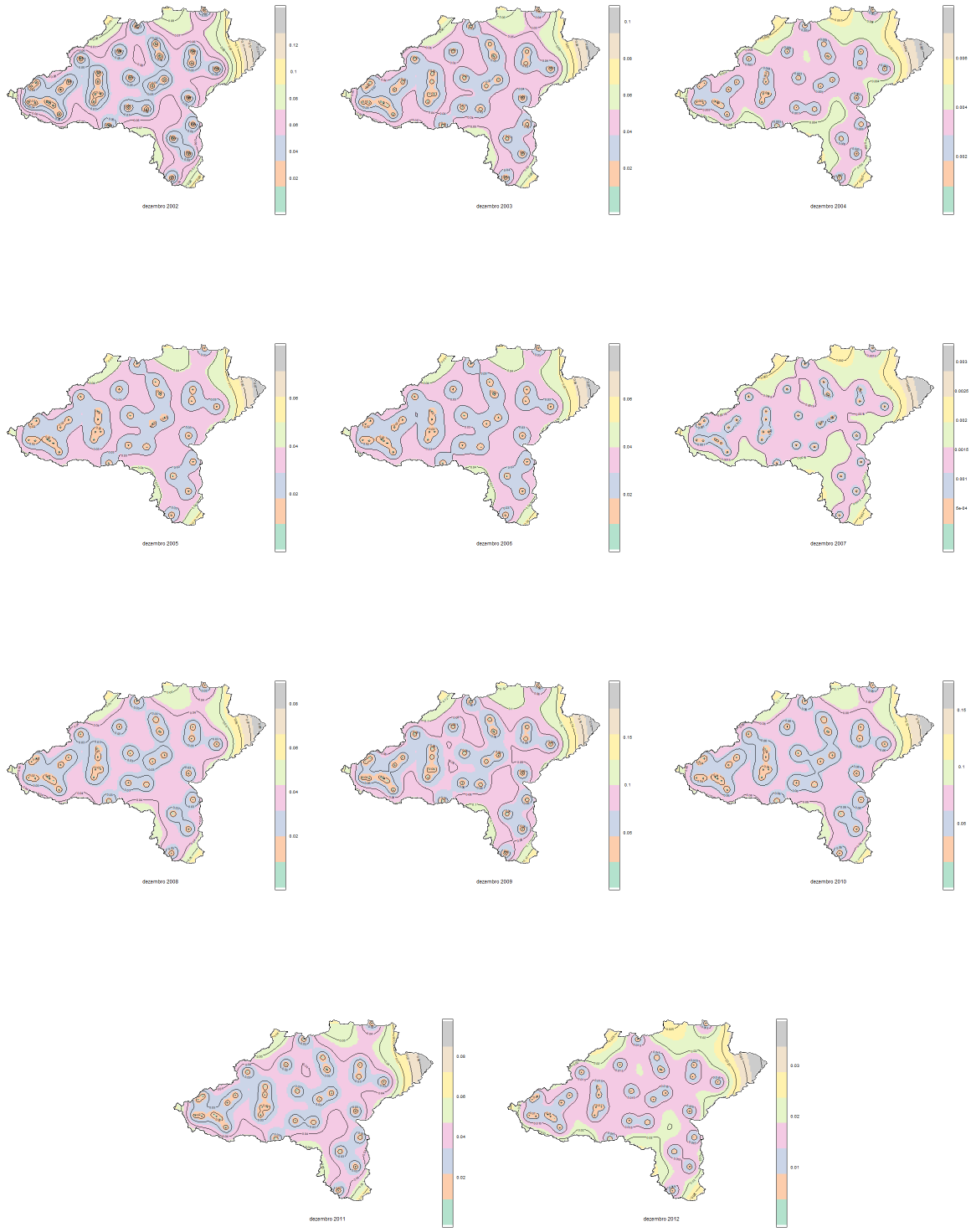


Figura A.32: Representações das superfícies de erros estimados da precipitação, no mês de dezembro, nos anos de 2002 até 2012.

Apêndice B

Qualidade da Água

B.1 Representações Gráficas do Oxigénio Dissolvido em cada Estação de Amostragem

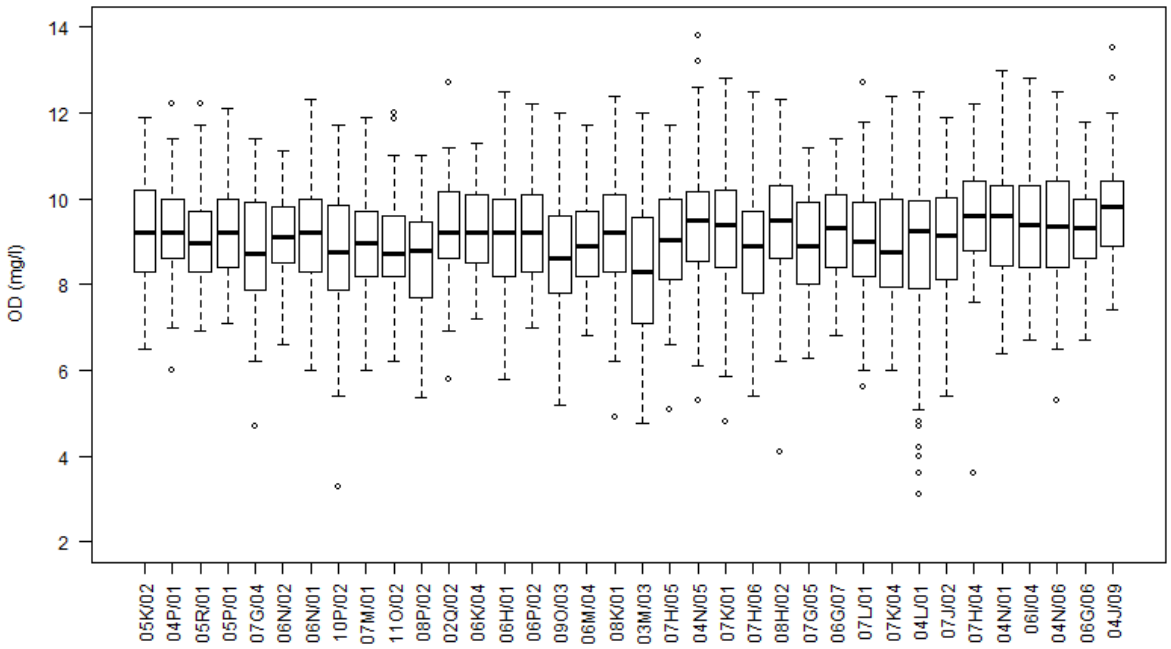


Figura B.1: Diagrama em caixa de bigodes das série de Oxigénio Dissolvido, nas 36 estações de amostragem, no período observado.

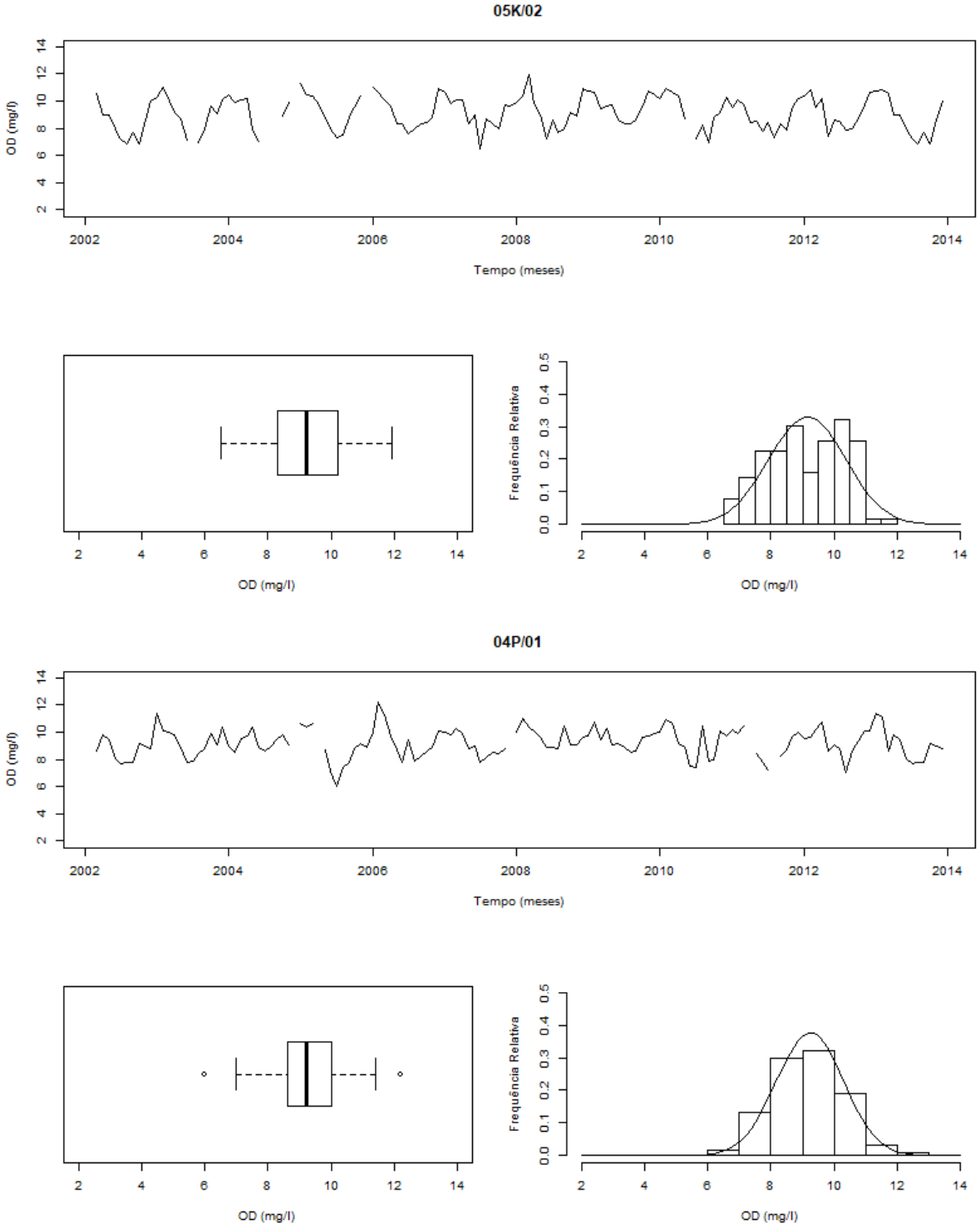


Figura B.2: Representações gráficas das séries temporais da Oxigénio Dissolvido e dos respetivos diagramas em caixas de bigodes e histogramas, no período observado.

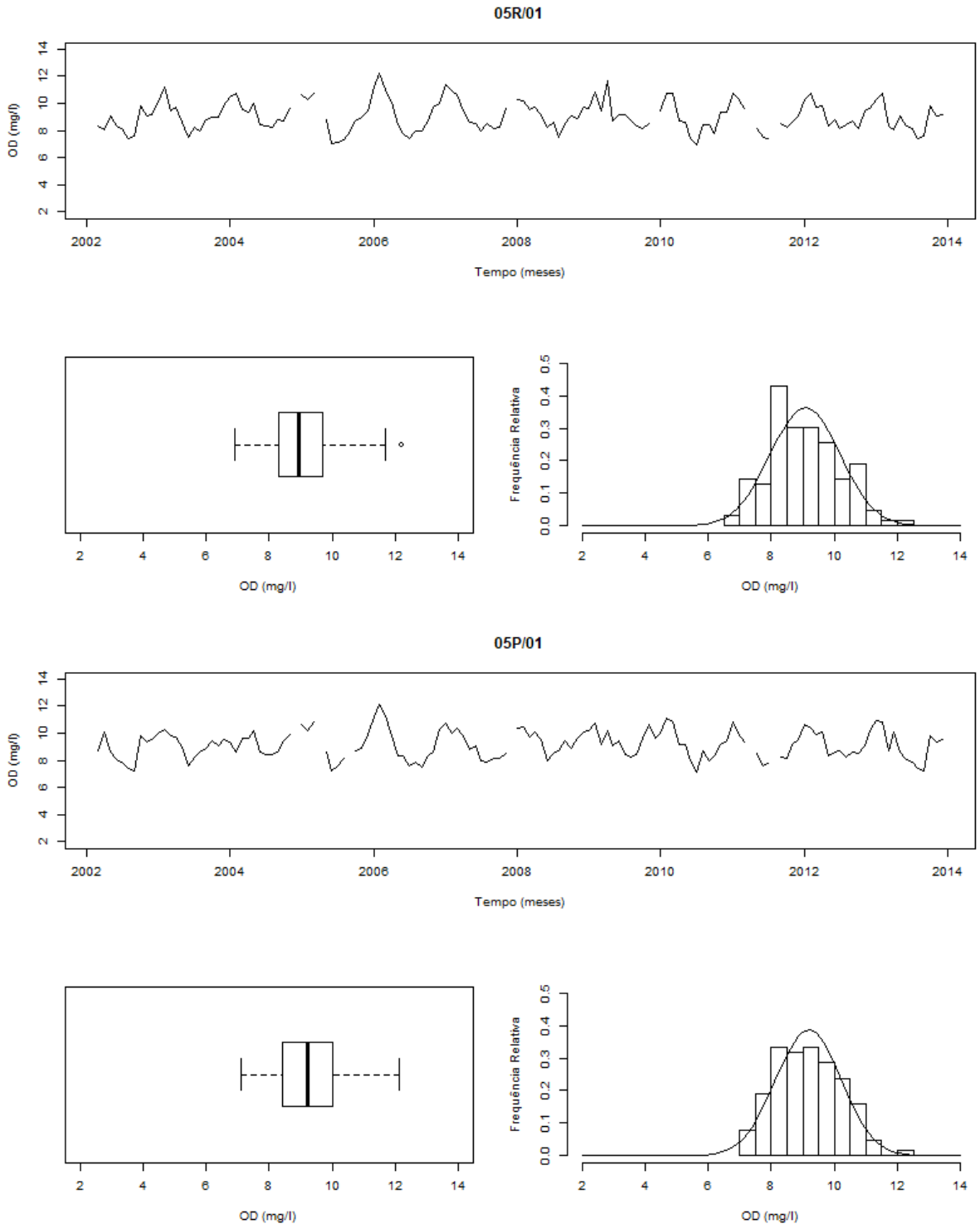


Figura B.3: Representações gráficas das séries temporais da Oxigénio Dissolvido e dos respetivos diagramas em caixas de bigodes e histogramas, no período observado.

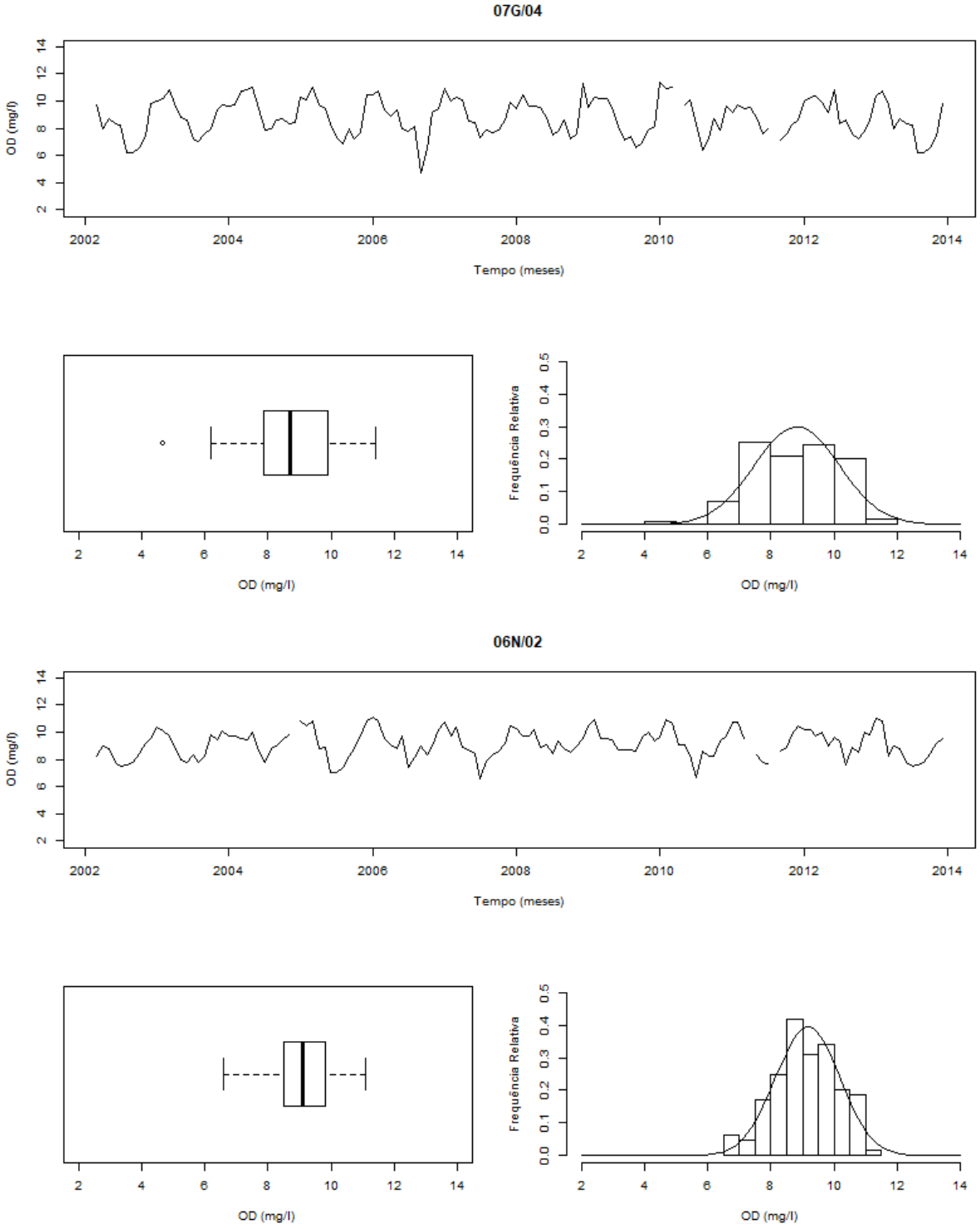


Figura B.4: Representações gráficas das séries temporais da Oxigénio Dissolvido e dos respetivos diagramas em caixas de bigodes e histogramas, no período observado.

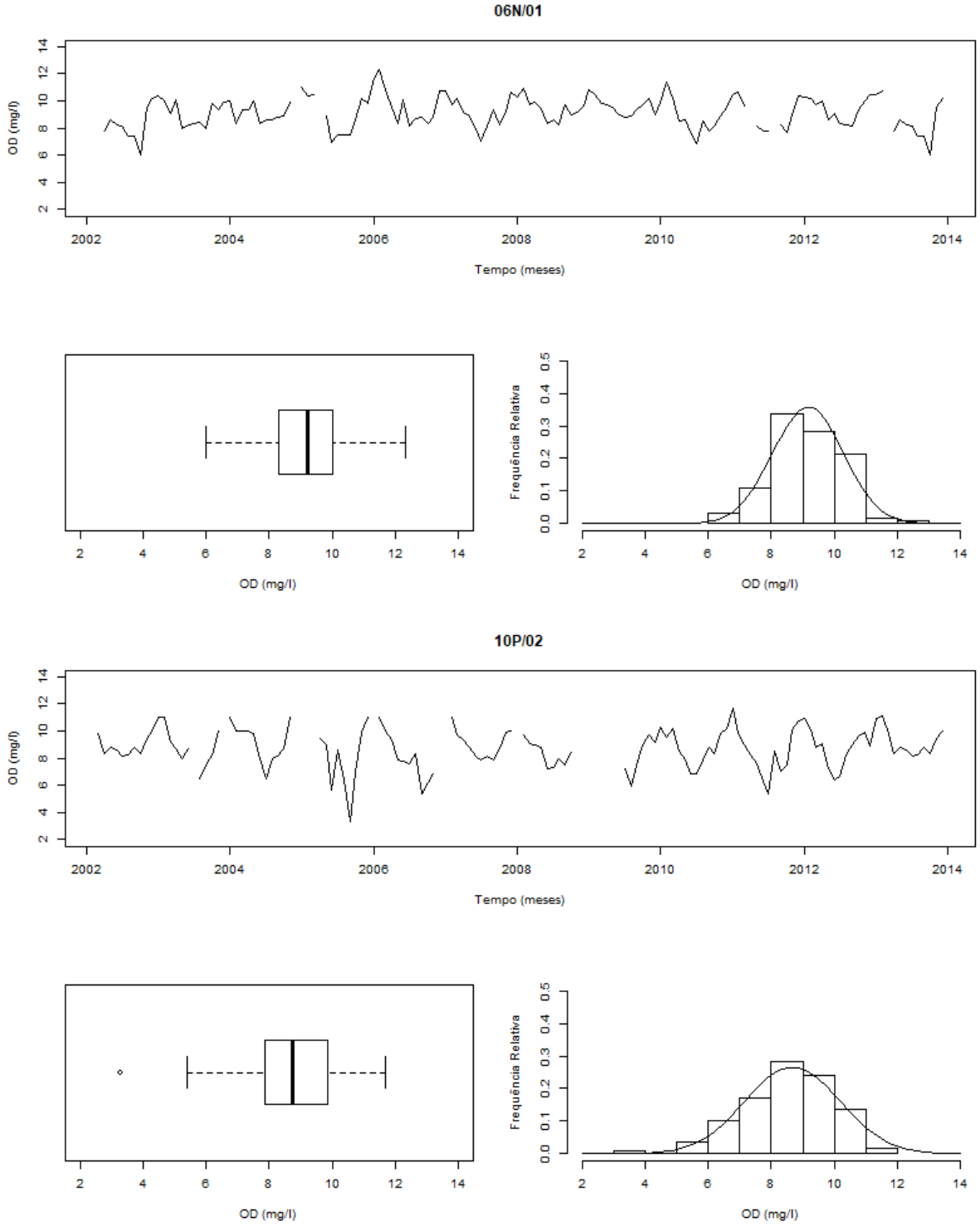


Figura B.5: Representações gráficas das séries temporais da Oxigénio Dissolvido e dos respetivos diagramas em caixas de bigodes e histogramas, no período observado.

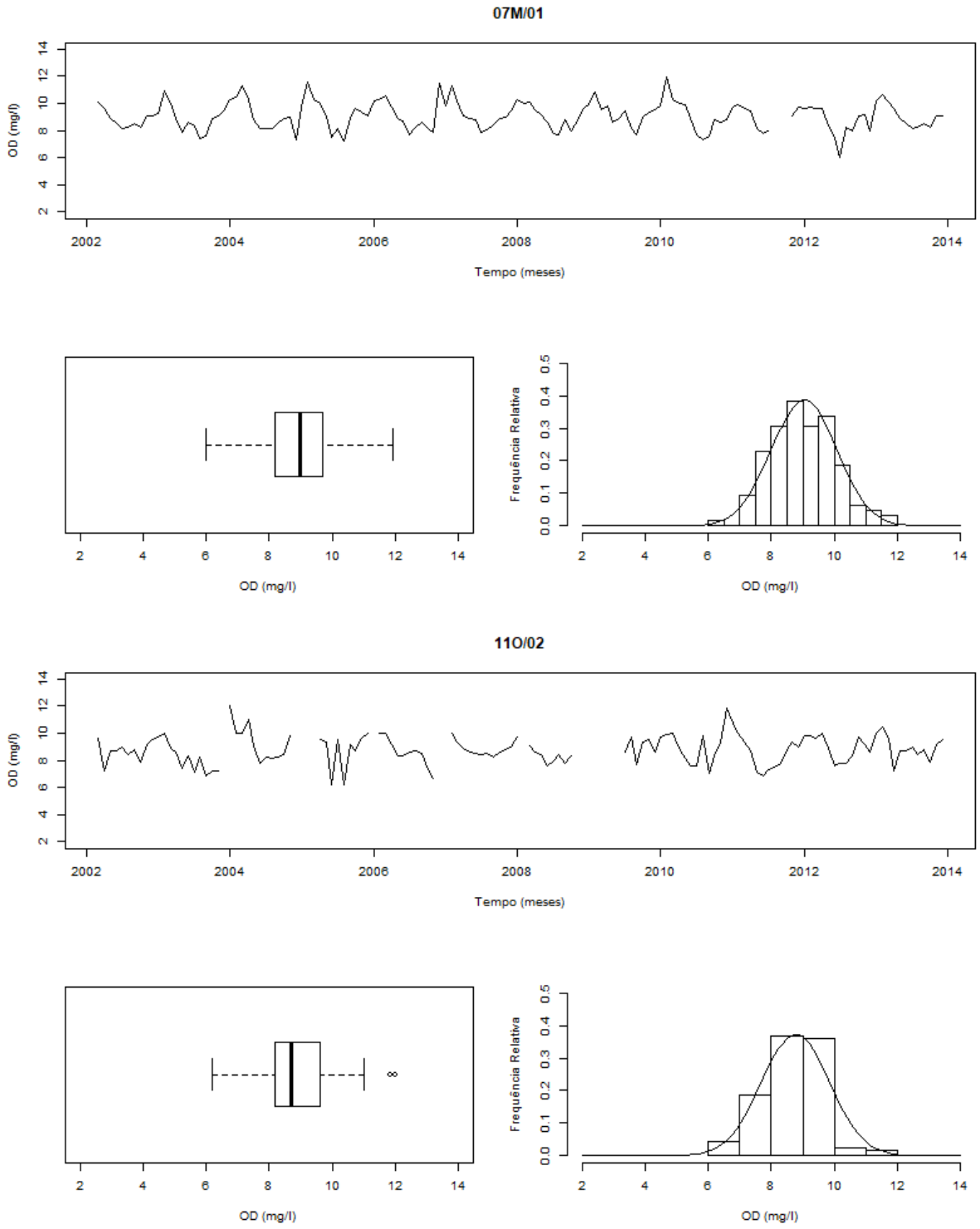


Figura B.6: Representações gráficas das séries temporais da Oxigénio Dissolvido e dos respetivos diagramas em caixas de bigodes e histogramas, no período observado.

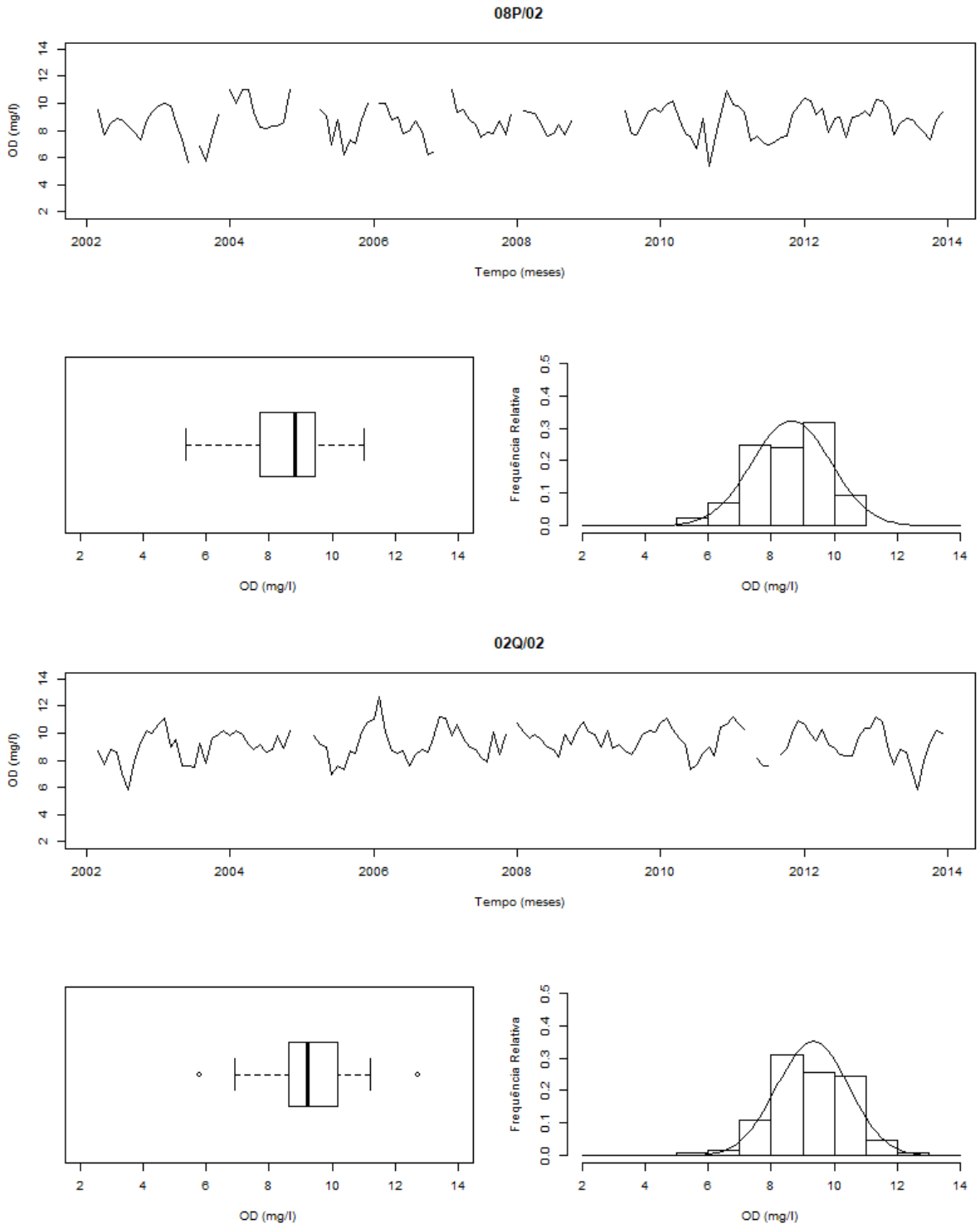


Figura B.7: Representações gráficas das séries temporais da Oxigénio Dissolvido e dos respetivos diagramas em caixas de bigodes e histogramas, no período observado.

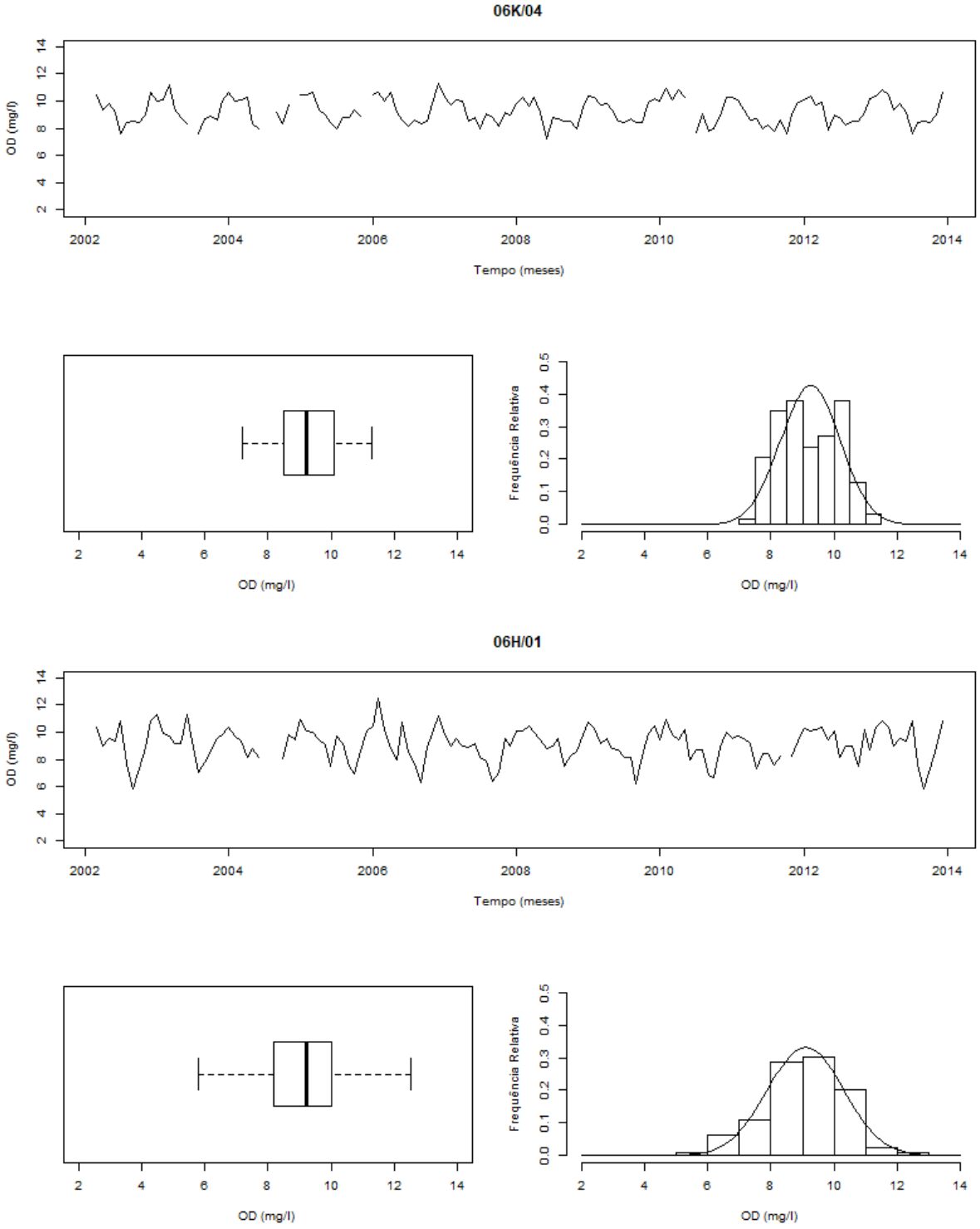


Figura B.8: Representações gráficas das séries temporais da Oxigénio Dissolvido e dos respetivos diagramas em caixas de bigodes e histogramas, no período observado.

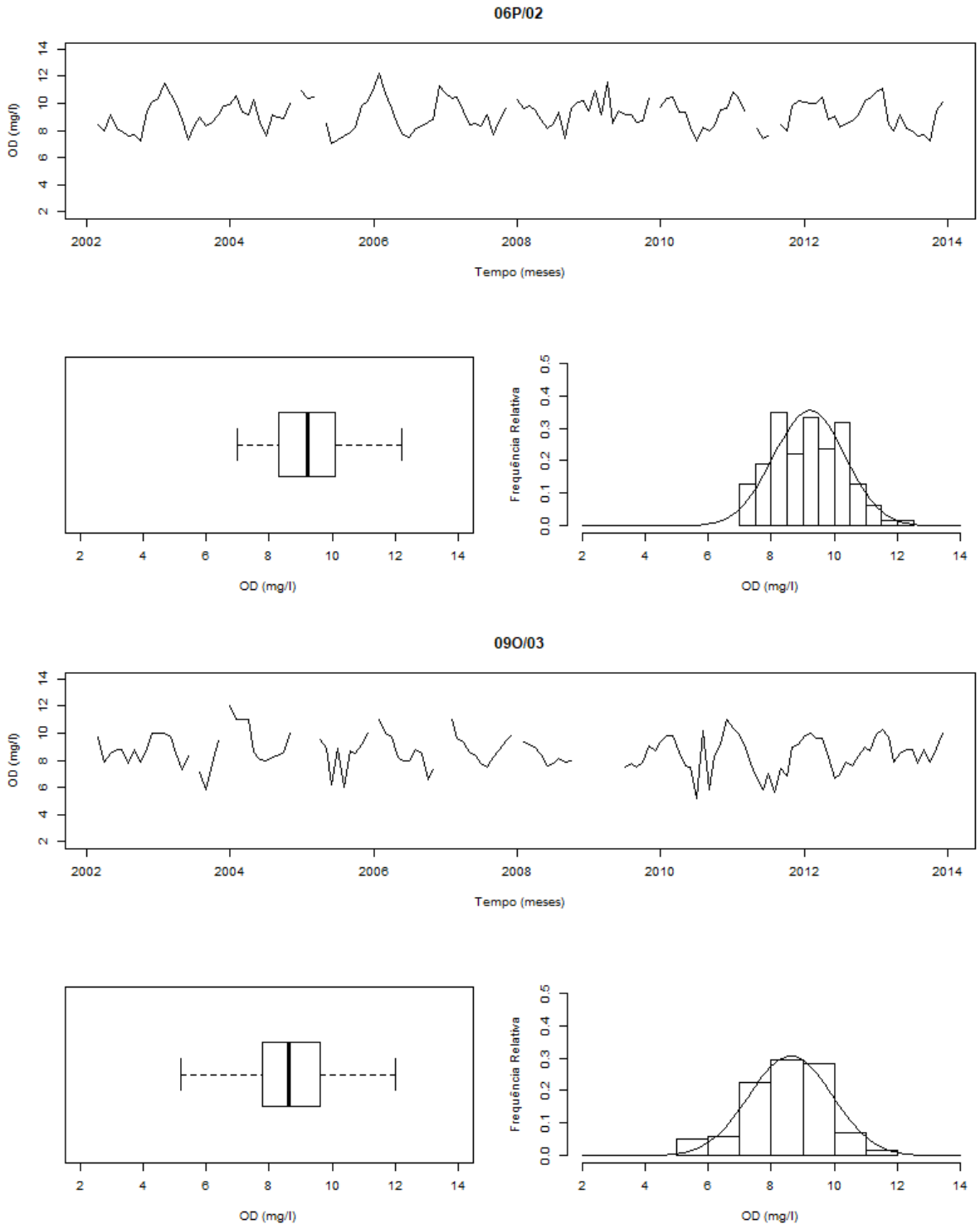


Figura B.9: Representações gráficas das séries temporais da Oxigénio Dissolvido e dos respetivos diagramas em caixas de bigodes e histogramas, no período observado.

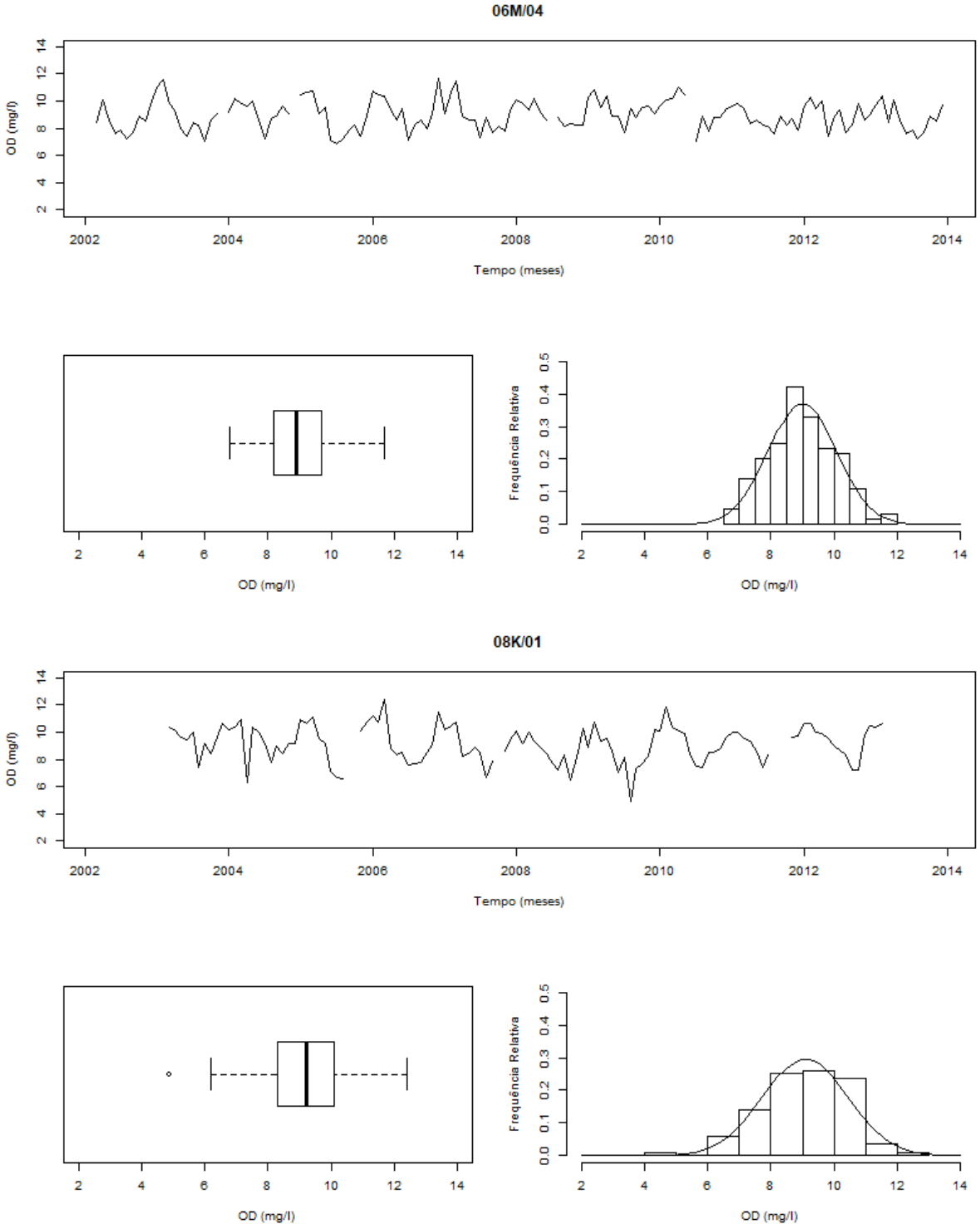


Figura B.10: Representações gráficas das séries temporais da Oxigénio Dissolvido e dos respetivos diagramas em caixas de bigodes e histogramas, no período observado.

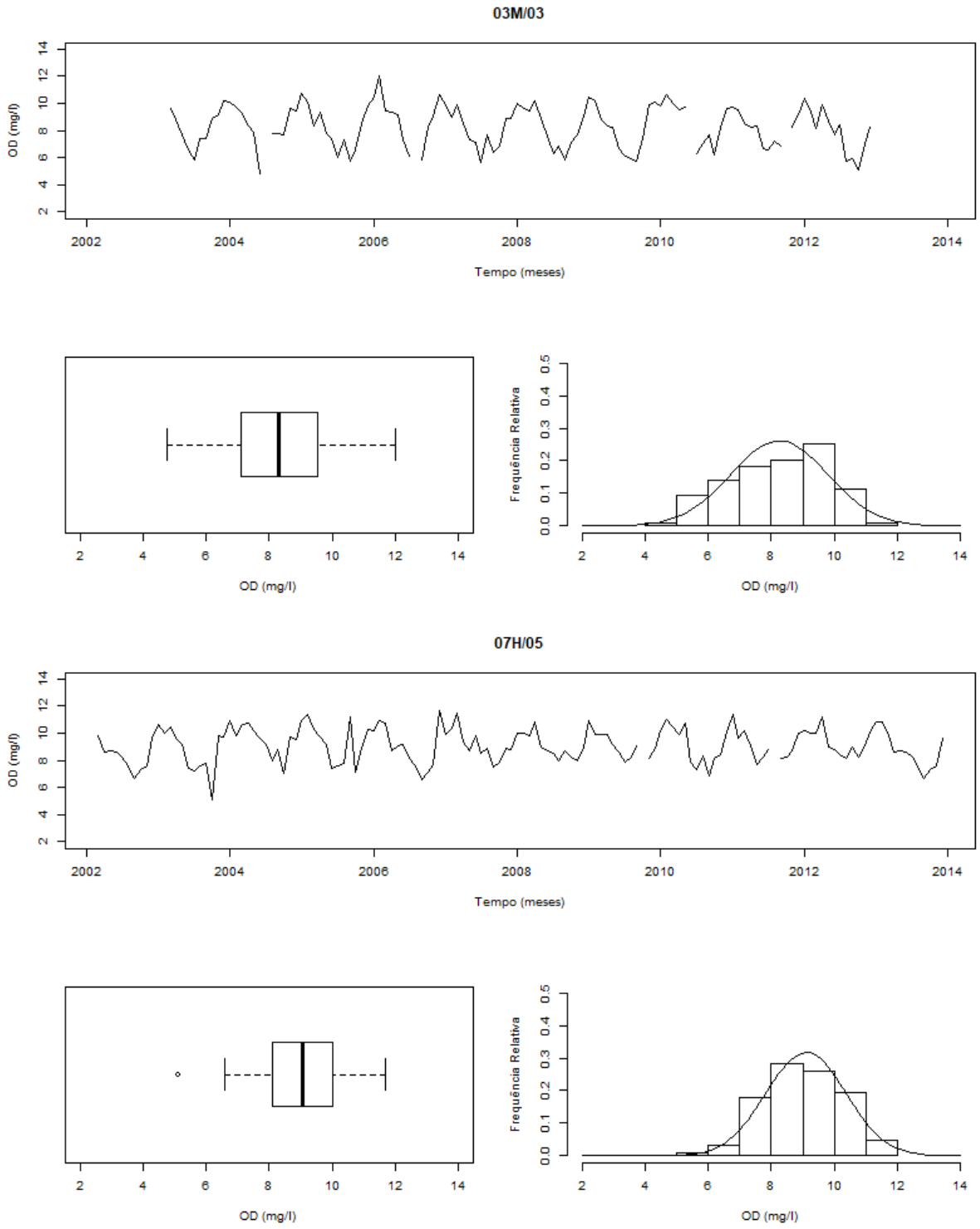


Figura B.11: Representações gráficas das séries temporais da Oxigénio Dissolvido e dos respetivos diagramas em caixas de bigodes e histogramas, no período observado.

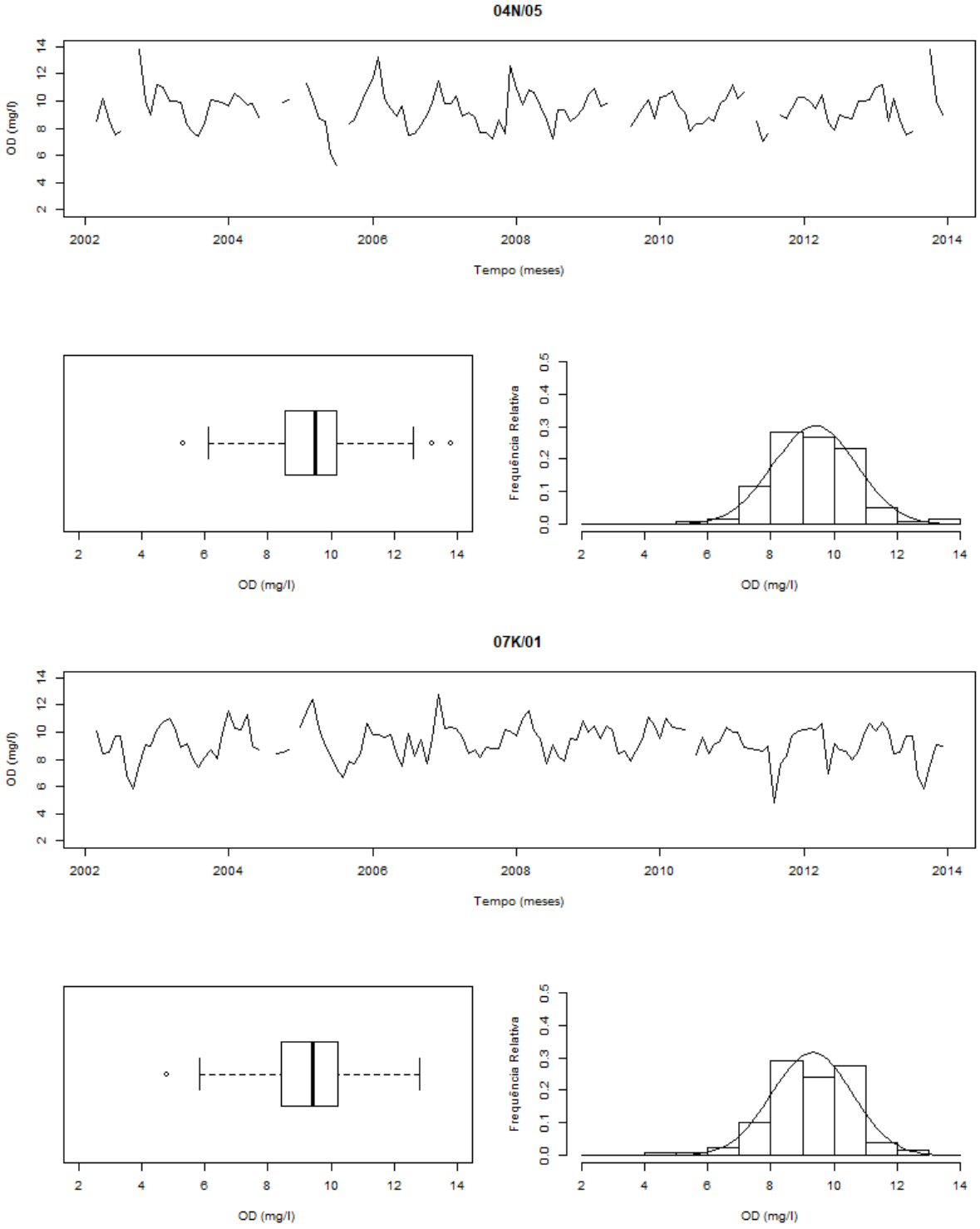


Figura B.12: Representações gráficas das séries temporais da Oxigénio Dissolvido e dos respetivos diagramas em caixas de bigodes e histogramas, no período observado.

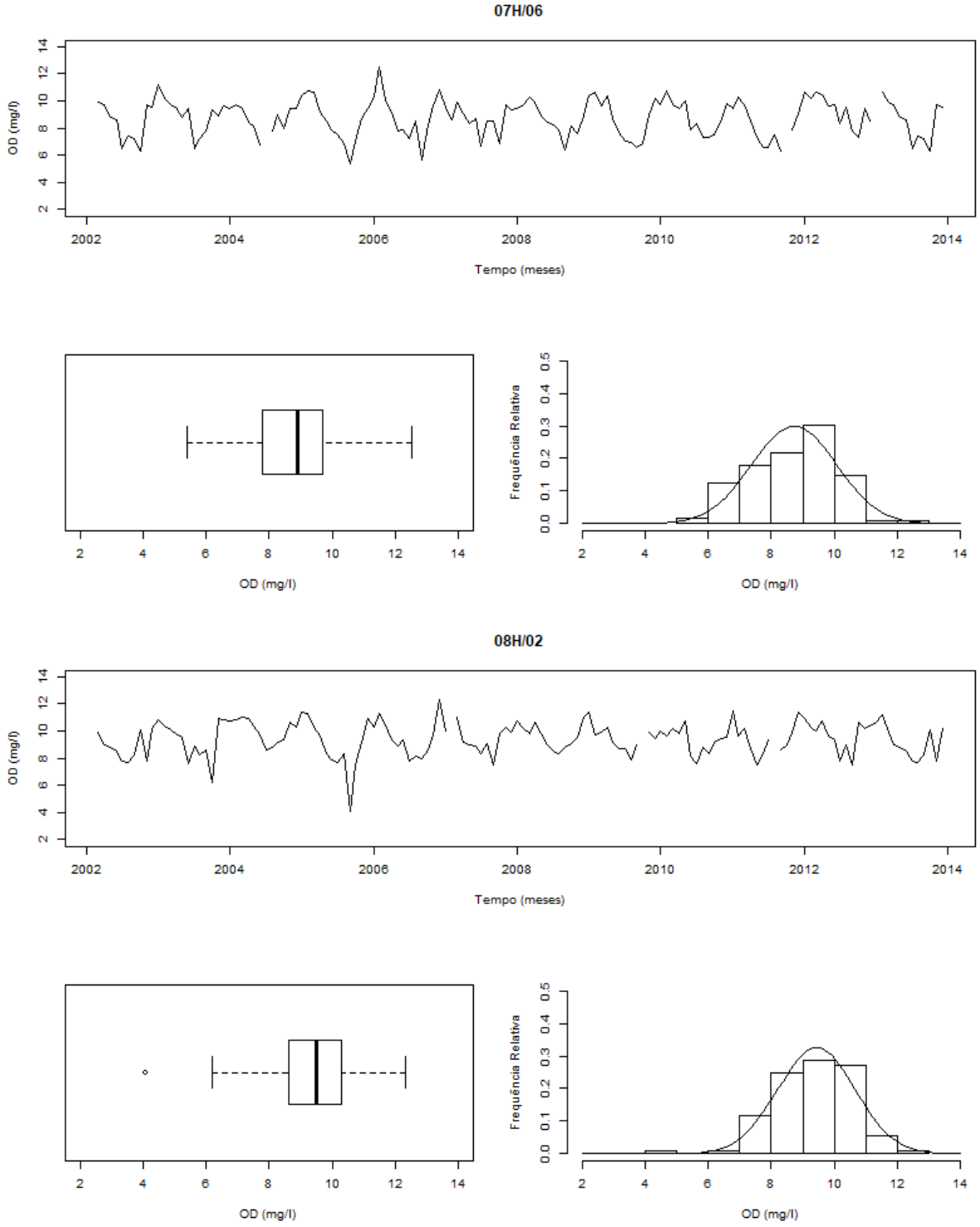


Figura B.13: Representações gráficas das séries temporais da Oxigénio Dissolvido e dos respetivos diagramas em caixas de bigodes e histogramas, no período observado.

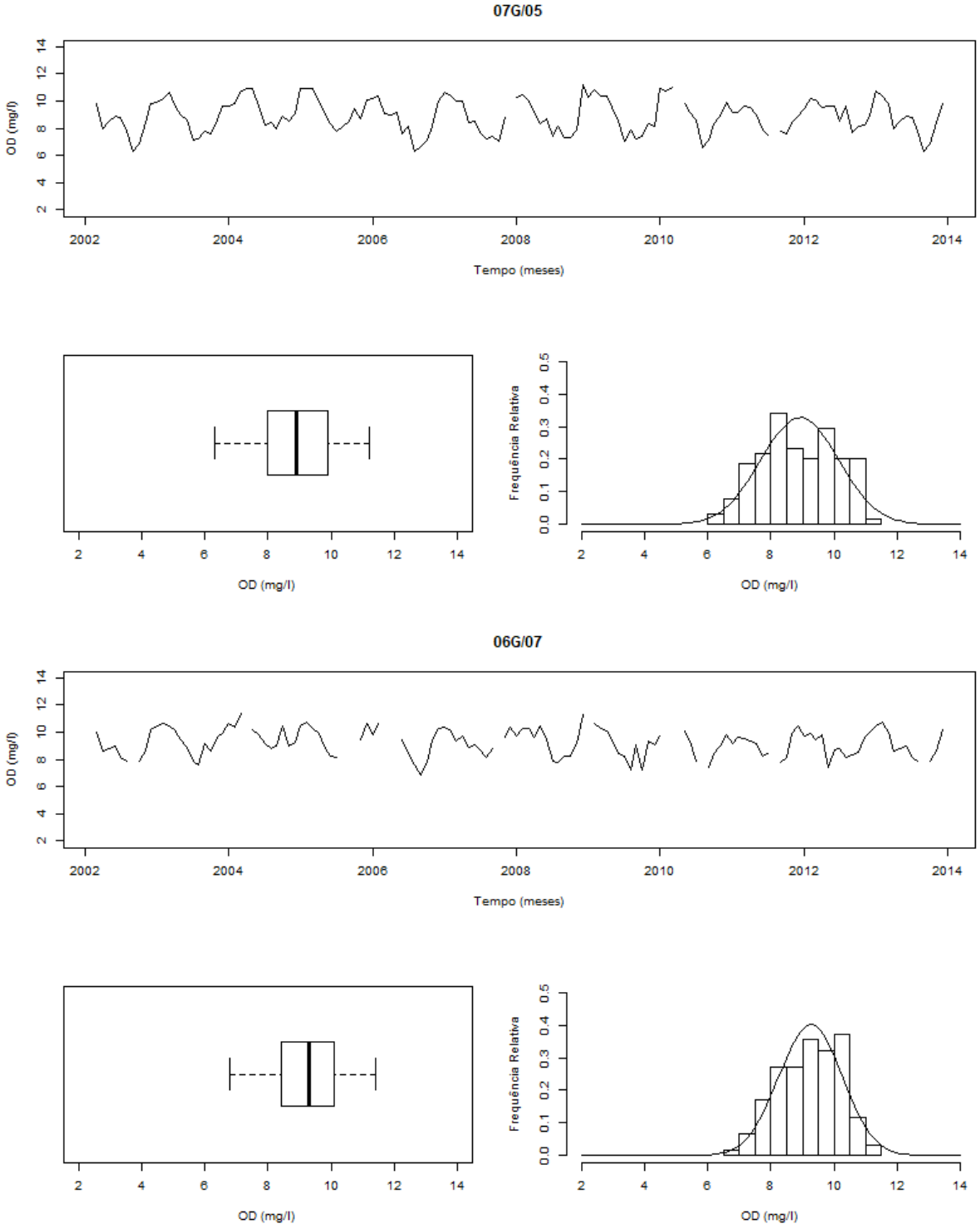


Figura B.14: Representações gráficas das séries temporais da Oxigénio Dissolvido e dos respetivos diagramas em caixas de bigodes e histogramas, no período observado.

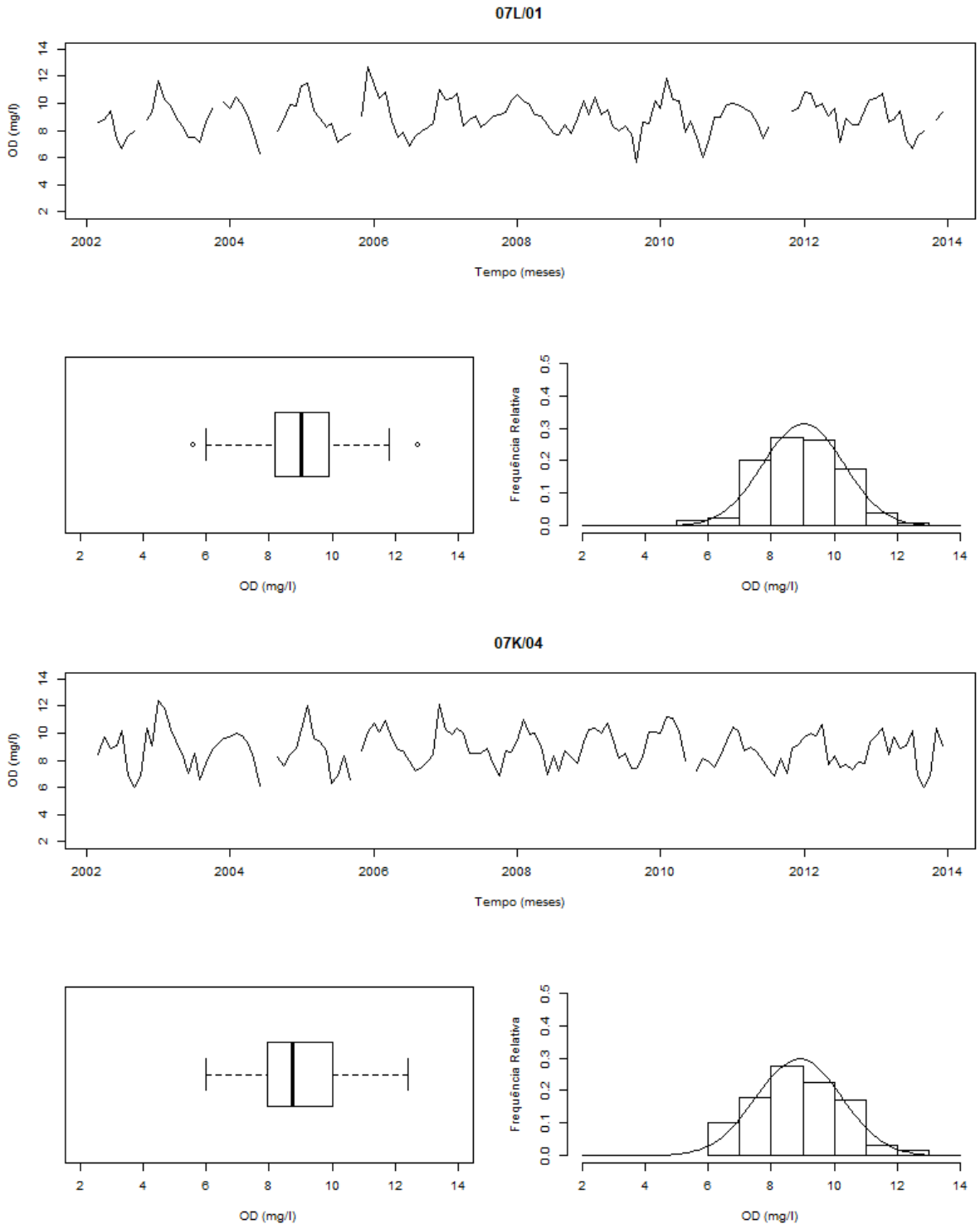


Figura B.15: Representações gráficas das séries temporais da Oxigénio Dissolvido e dos respetivos diagramas em caixas de bigodes e histogramas, no período observado.

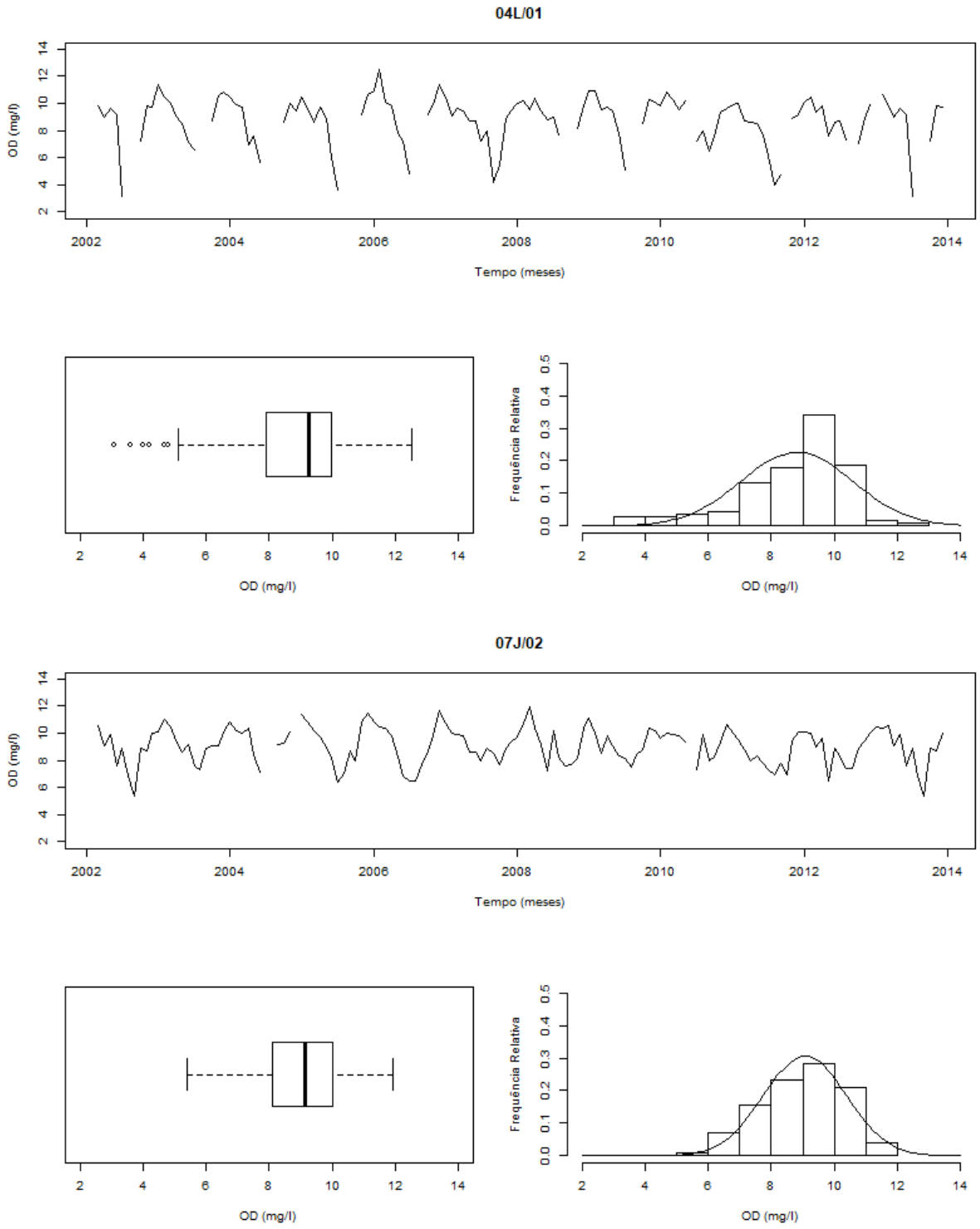


Figura B.16: Representações gráficas das séries temporais da Oxigénio Dissolvido e dos respetivos diagramas em caixas de bigodes e histogramas, no período observado.

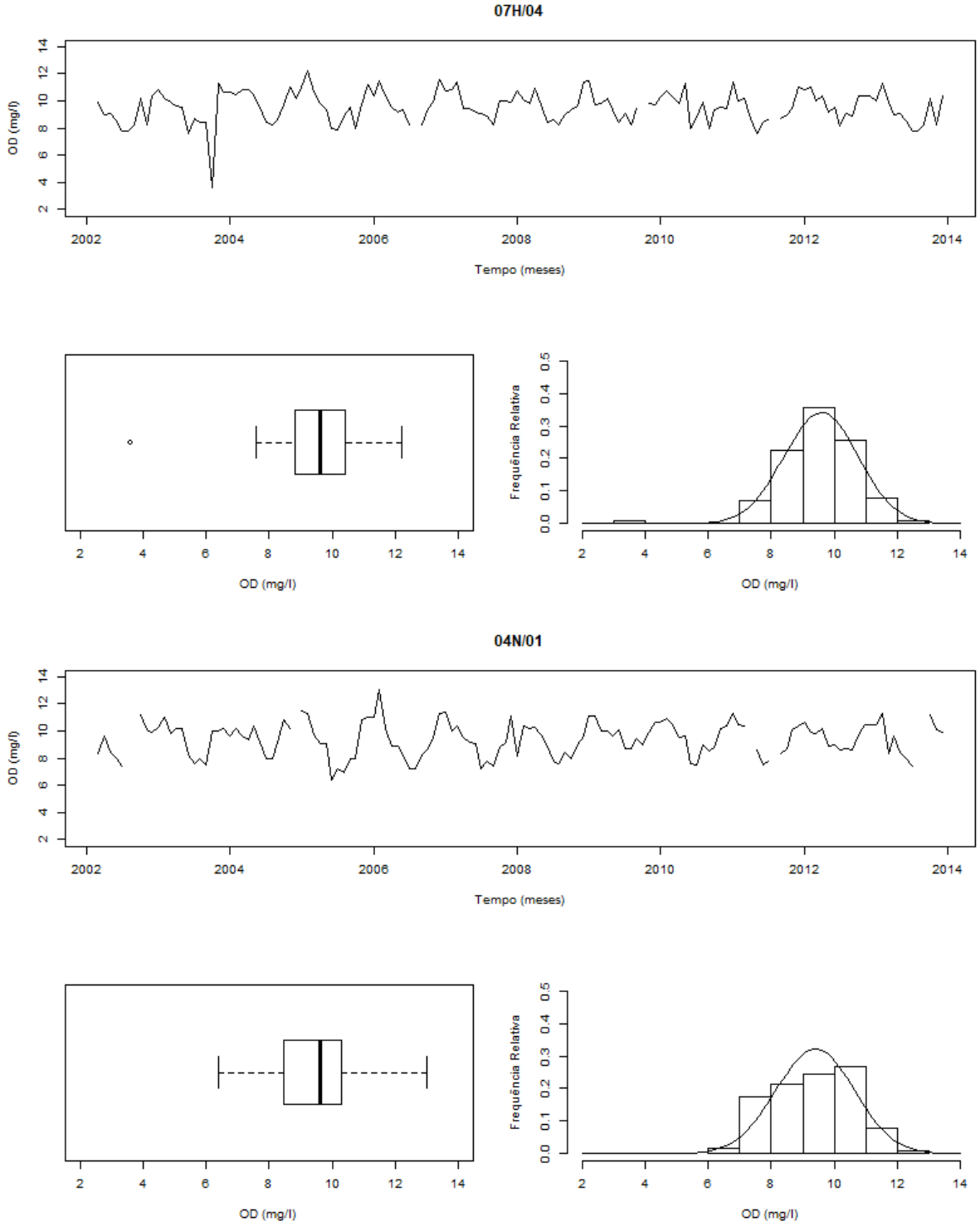


Figura B.17: Representações gráficas das séries temporais da Oxigénio Dissolvido e dos respetivos diagramas em caixas de bigodes e histogramas, no período observado.

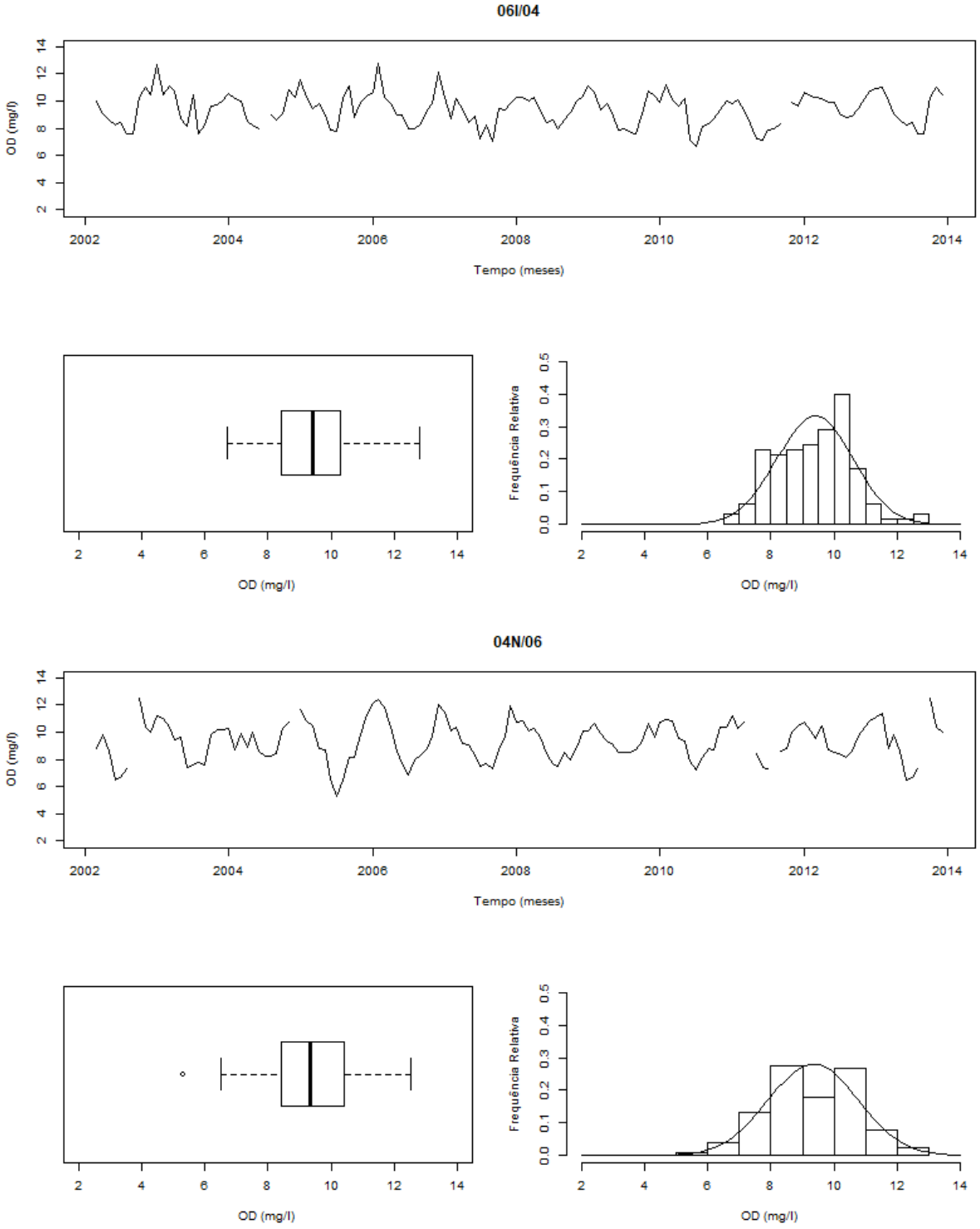


Figura B.18: Representações gráficas das séries temporais da Oxigénio Dissolvido e dos respetivos diagramas em caixas de bigodes e histogramas, no período observado.

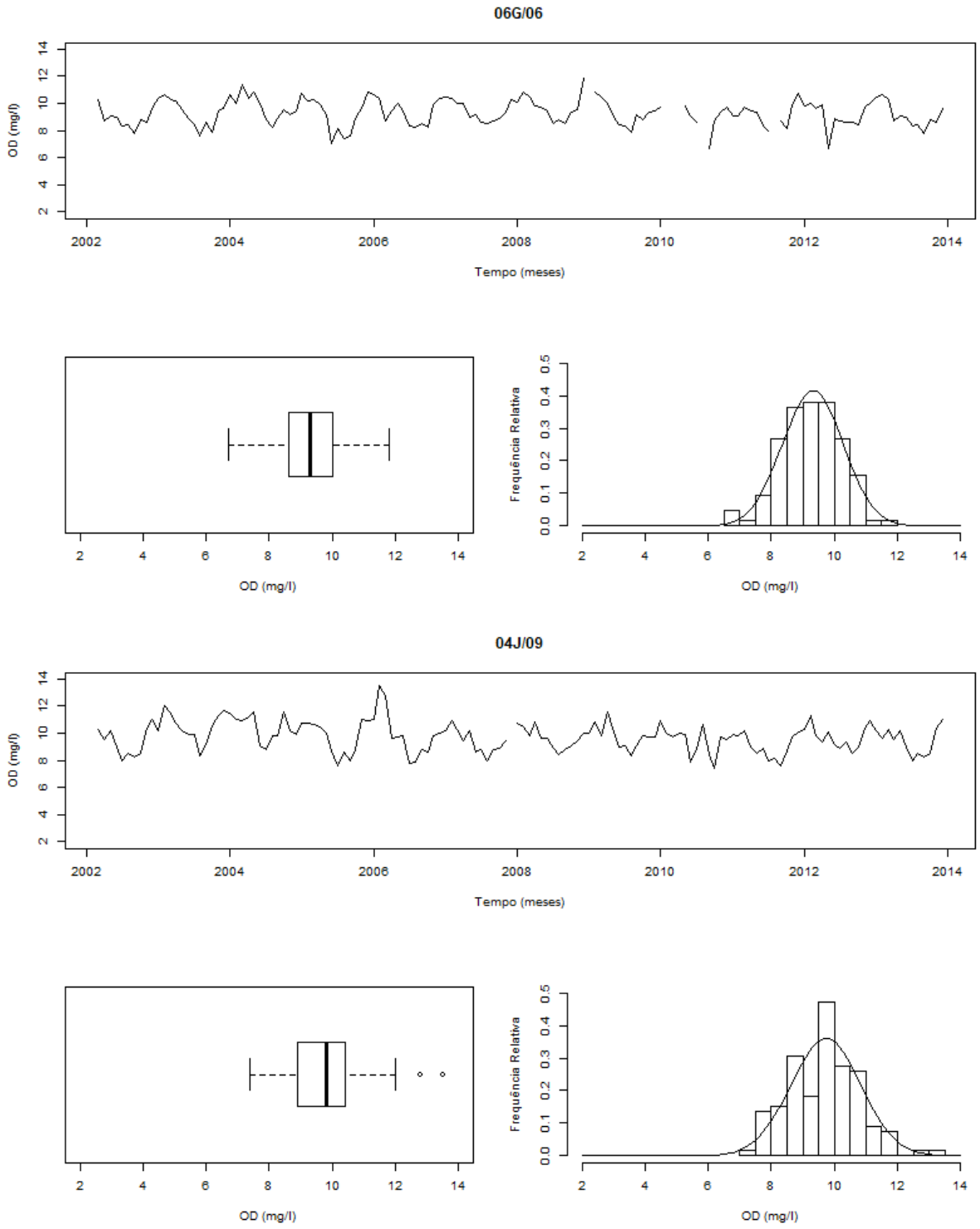


Figura B.19: Representações gráficas das séries temporais da Oxigénio Dissolvido e dos respetivos diagramas em caixas de bigodes e histogramas, no período observado.

B.2 Representações Gráficas do Ajustamento e dos Resíduos do Oxigénio Dissolvido (Modelo 3)

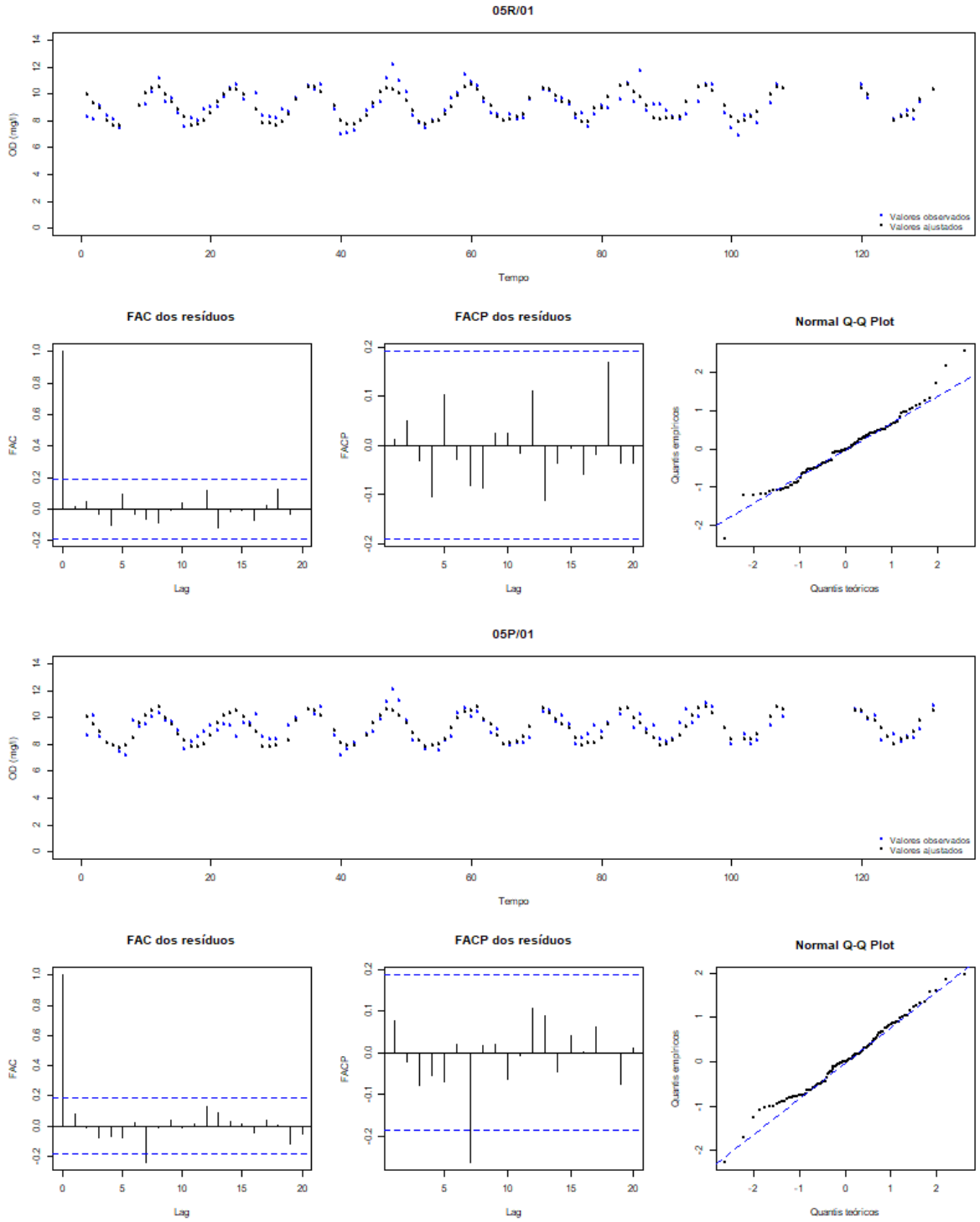


Figura B.20: Representações da série original e dos os valores estimados, da FAC, da FACP e do *Q-Q plot* dos resíduos do modelo, nas estações.

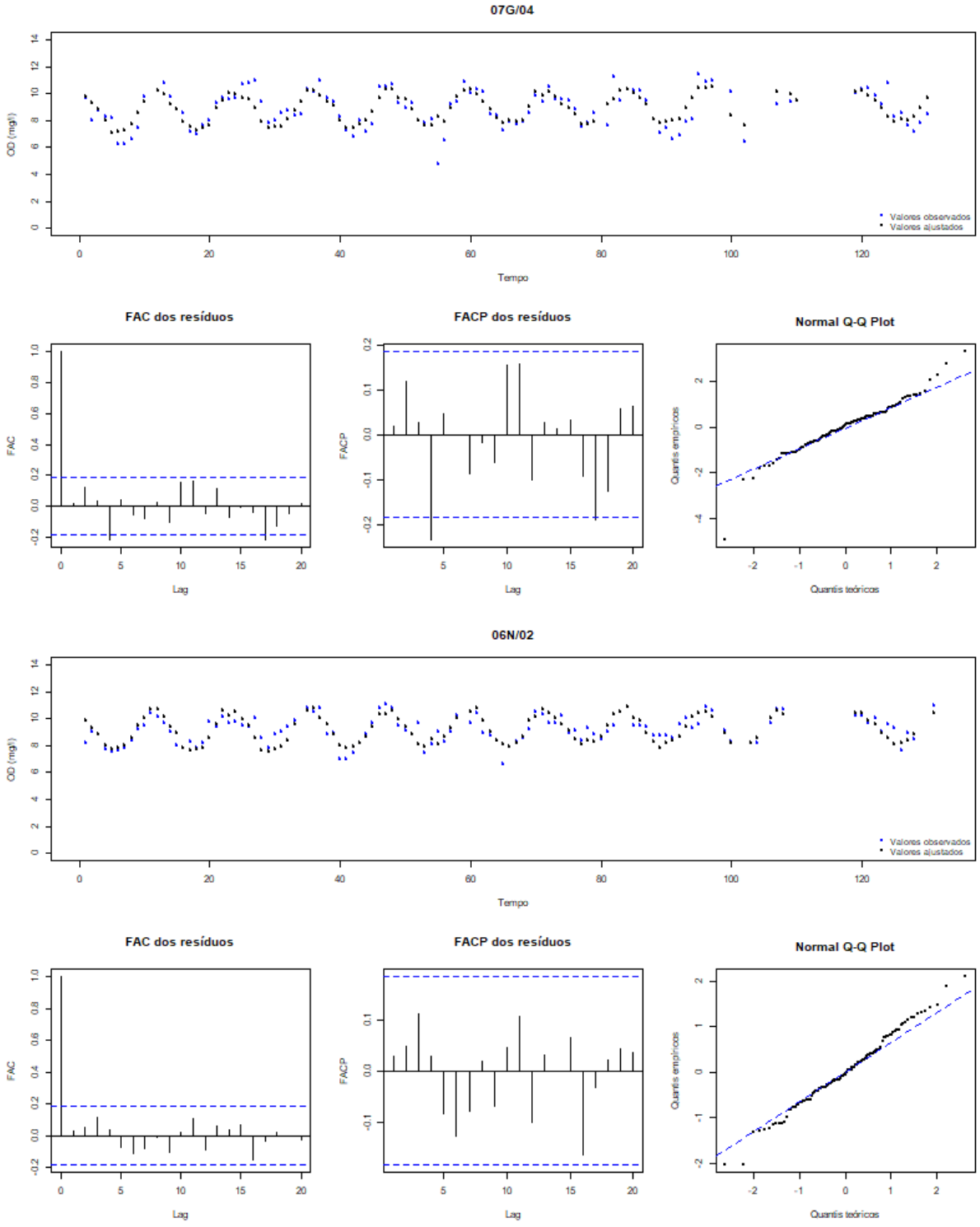


Figura B.21: Representações da série original e dos os valores estimados, da FAC, da FACP e do *Q-Q plot* dos resíduos do modelo, nas estações.

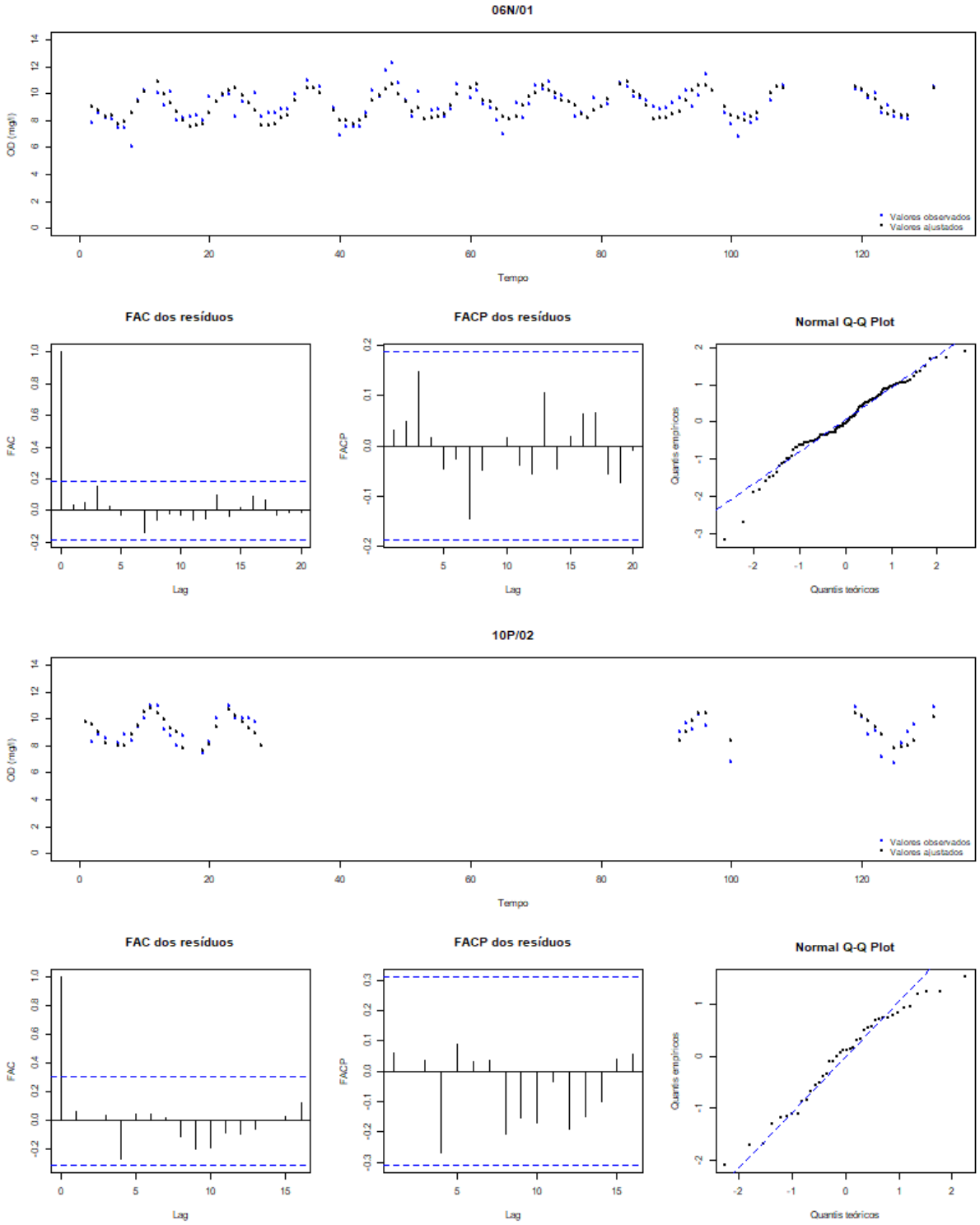


Figura B.22: Representações da série original e dos os valores estimados, da FAC, da FACP e do *Q-Q plot* dos resíduos do modelo, nas estações.

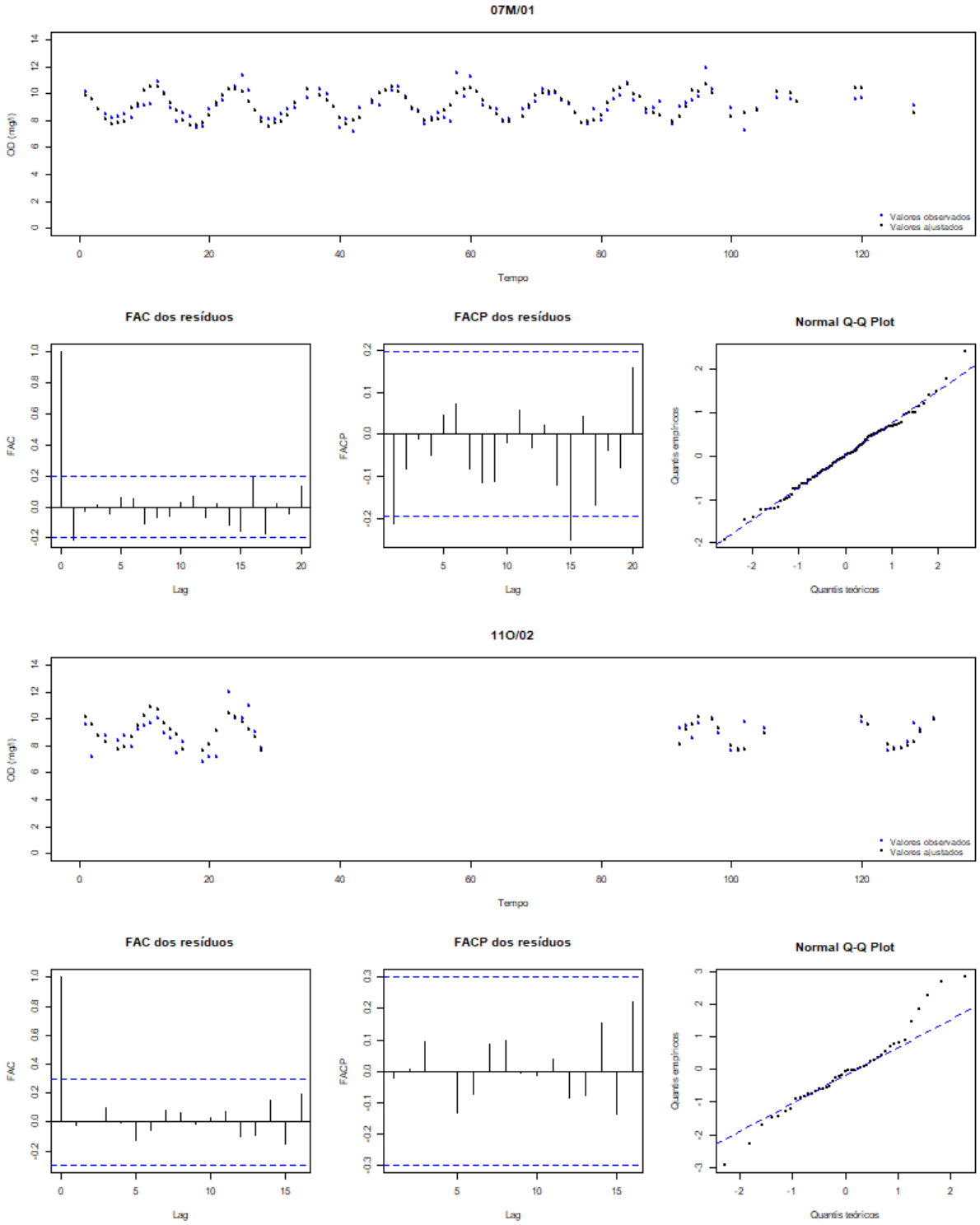


Figura B.23: Representações da série original e dos os valores estimados, da FAC, da FACP e do *Q-Q plot* dos resíduos do modelo, nas estações.

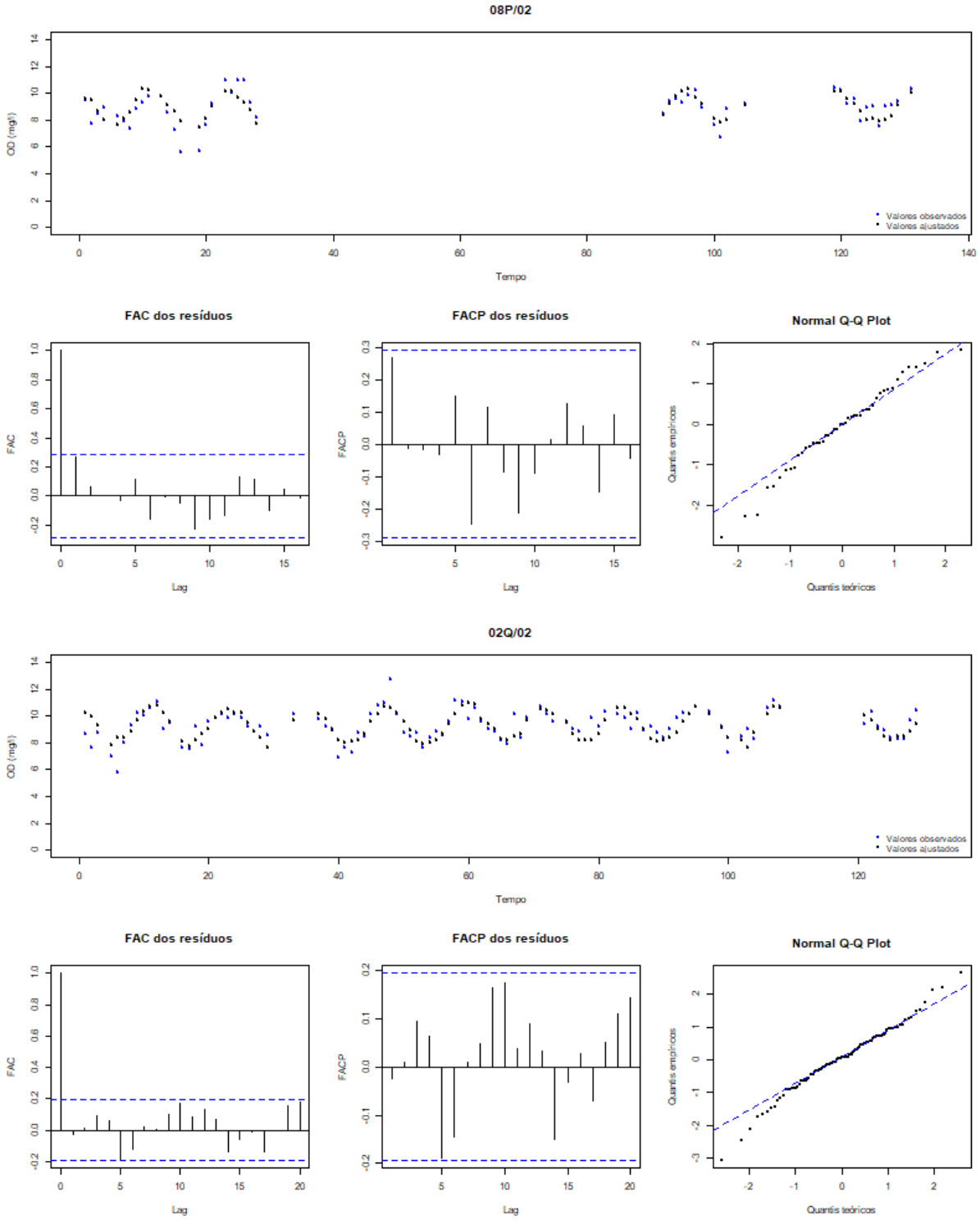


Figura B.24: Representações da série original e dos os valores estimados, da FAC, da FACP e do *Q-Q plot* dos resíduos do modelo, nas estações.

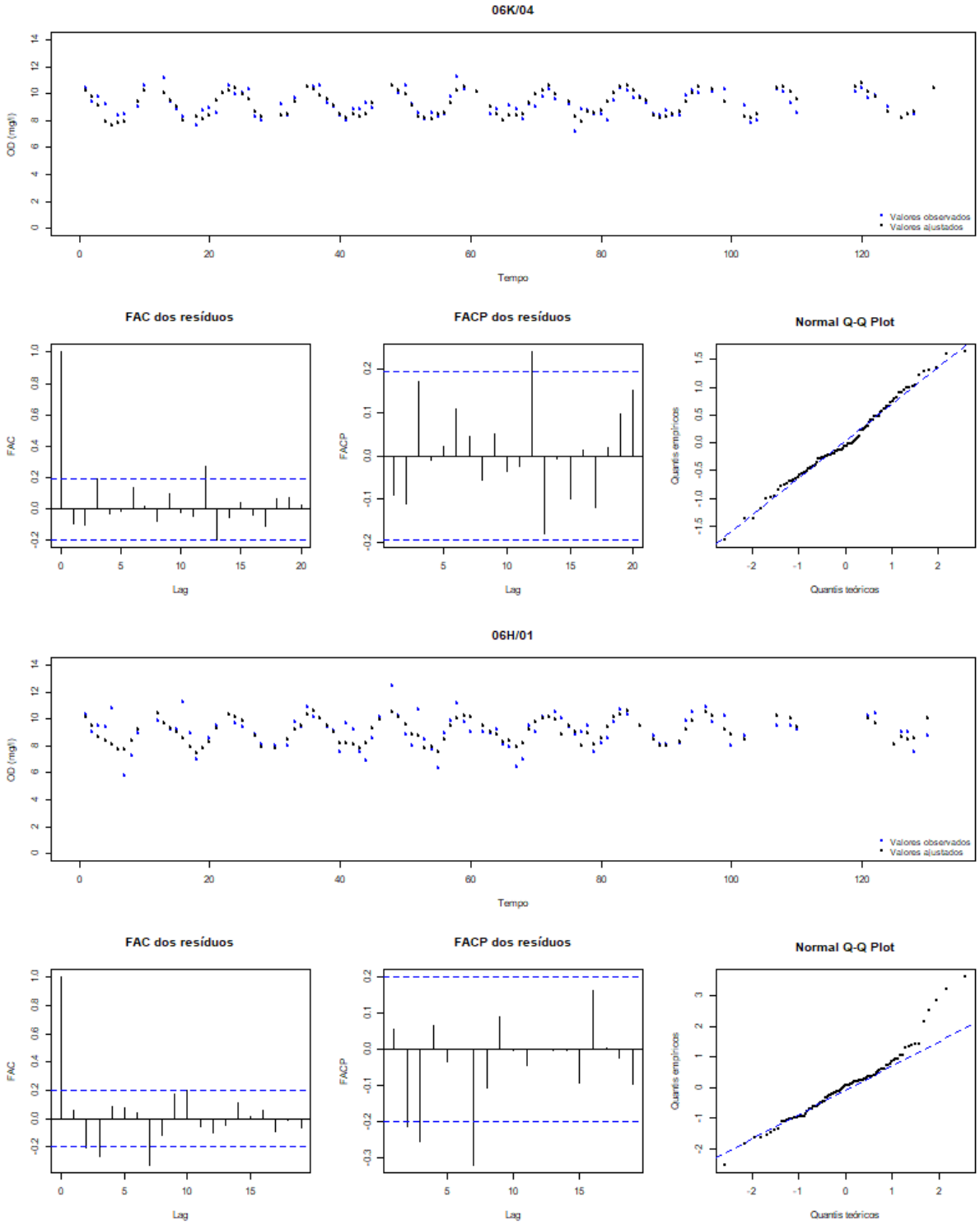


Figura B.25: Representações da série original e dos os valores estimados, da FAC, da FACP e do *Q-Q plot* dos resíduos do modelo, nas estações.

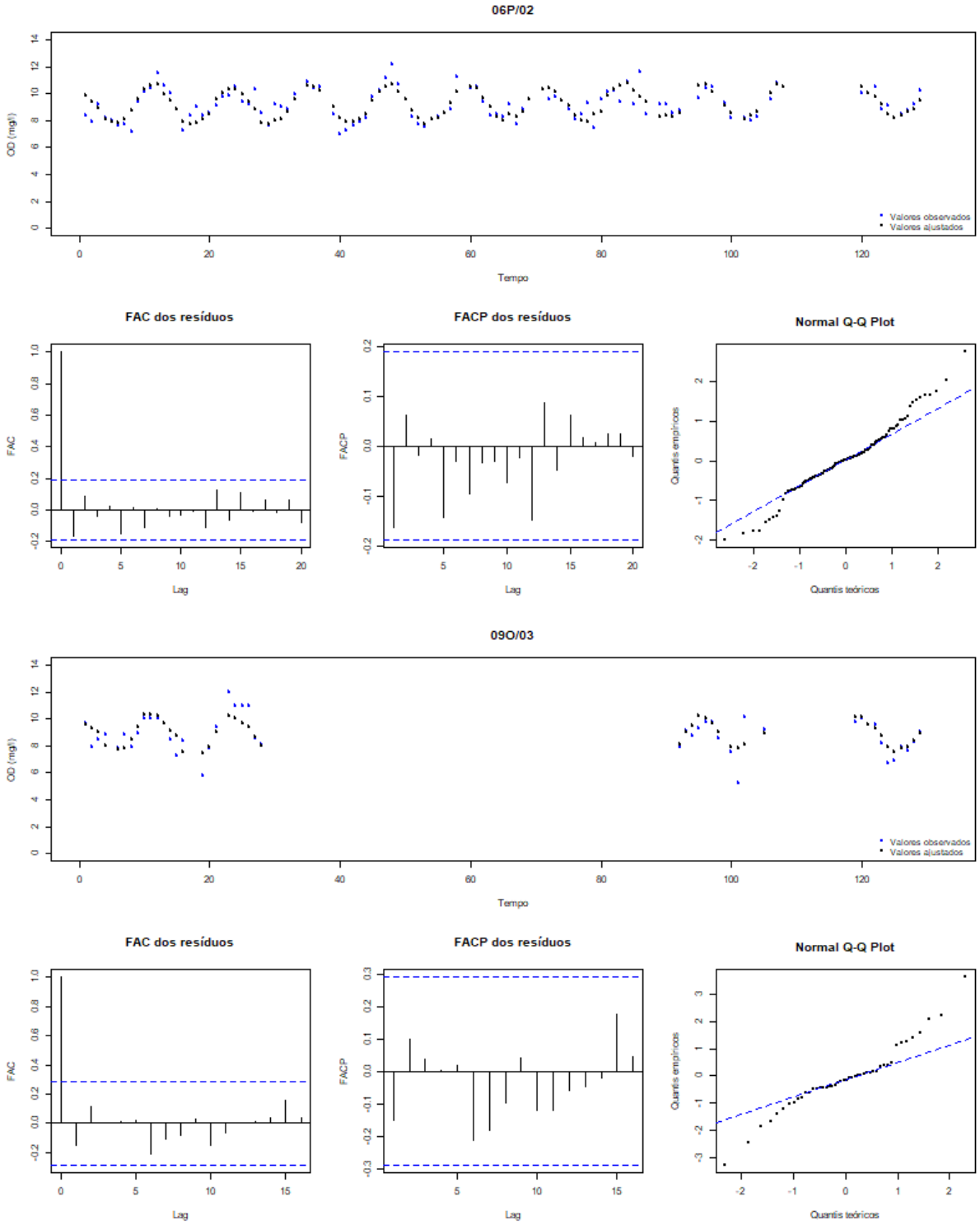


Figura B.26: Representações da série original e dos os valores estimados, da FAC, da FACP e do *Q-Q plot* dos resíduos do modelo, nas estações.

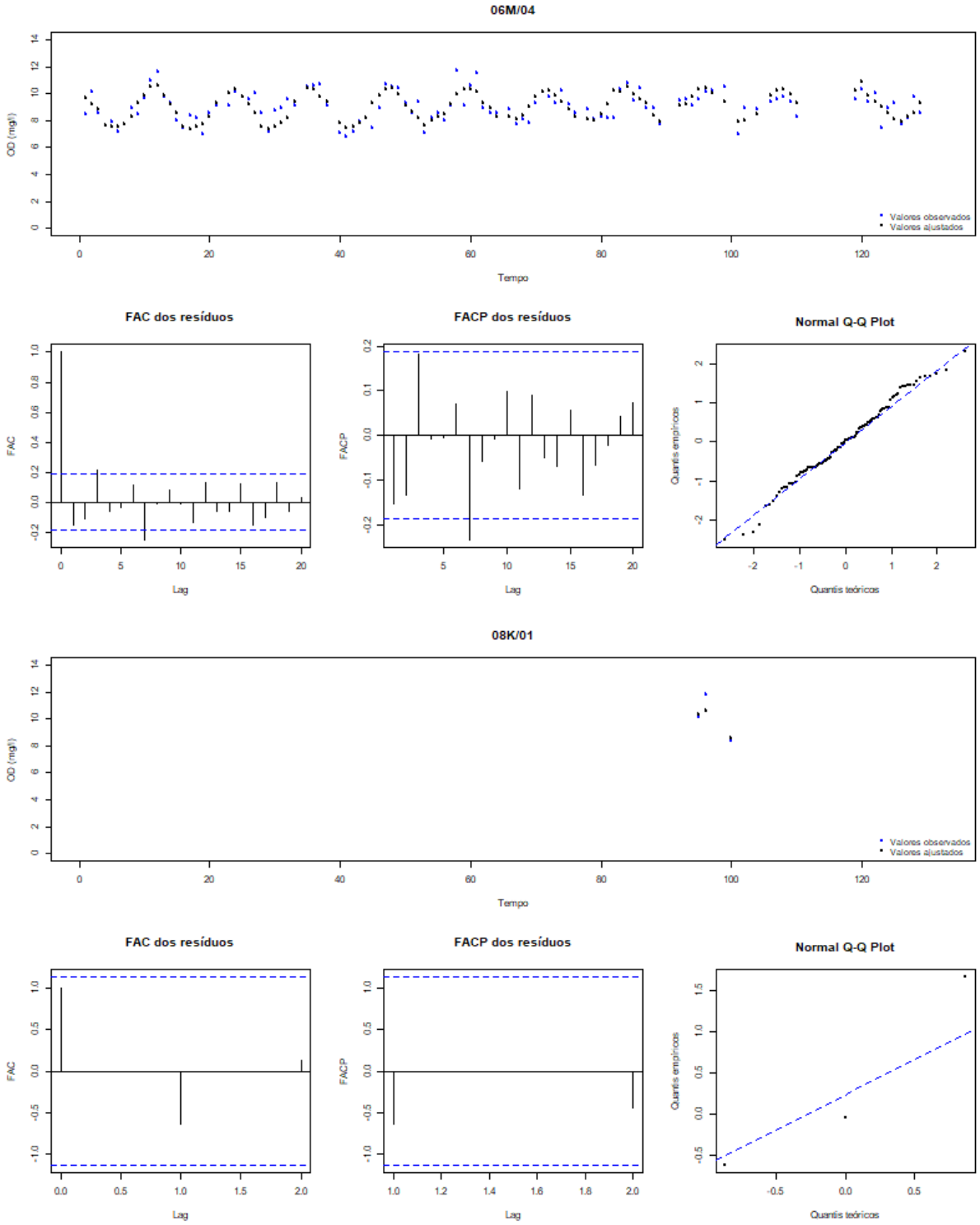


Figura B.27: Representações da série original e dos os valores estimados, da FAC, da FACP e do *Q-Q plot* dos resíduos do modelo, nas estações.

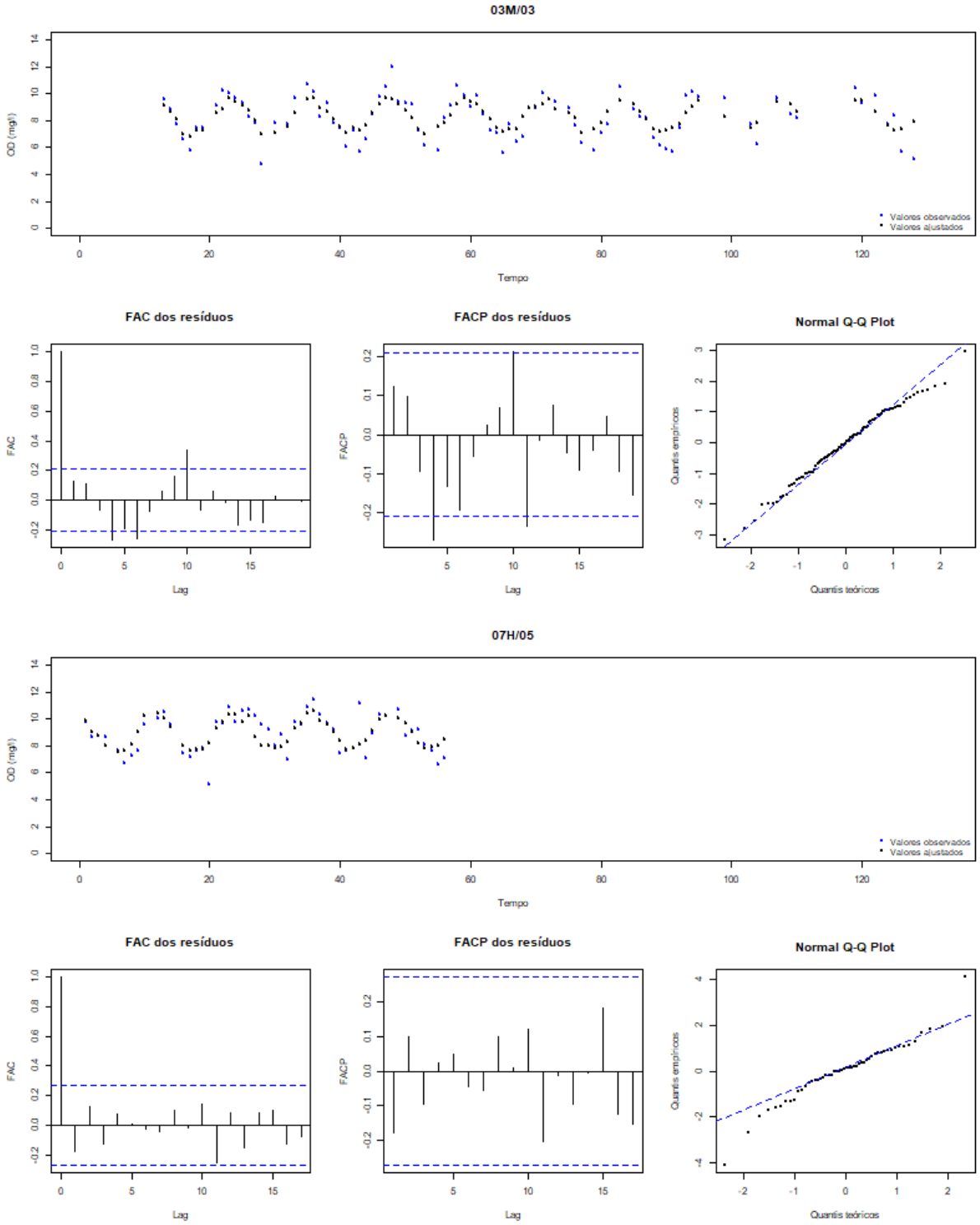


Figura B.28: Representações da série original e dos os valores estimados, da FAC, da FACP e do *Q-Q plot* dos resíduos do modelo, nas estações.

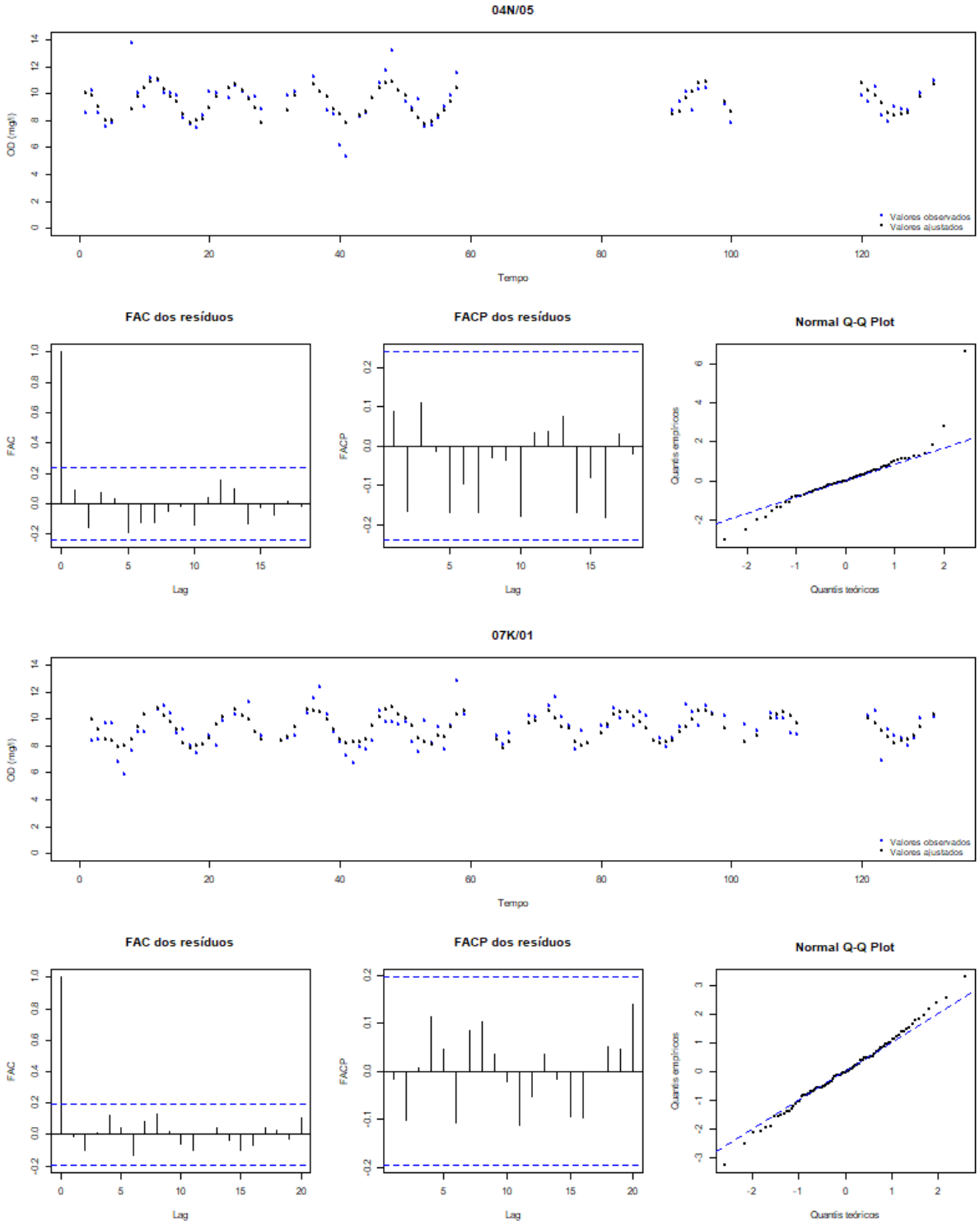


Figura B.29: Representações da série original e dos os valores estimados, da FAC, da FACP e do *Q-Q plot* dos resíduos do modelo, nas estações.

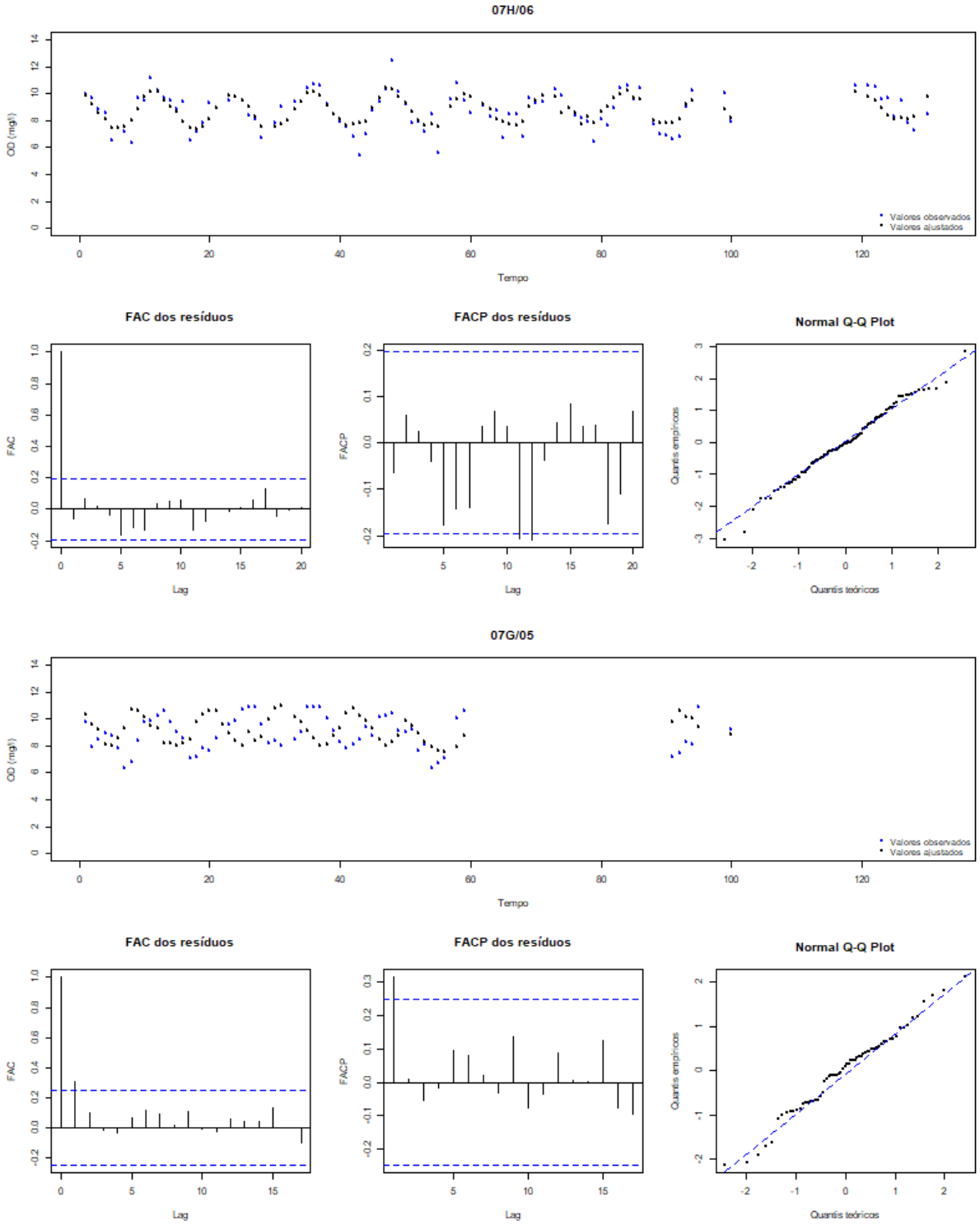


Figura B.30: Representações da série original e dos os valores estimados, da FAC, da FACP e do *Q-Q plot* dos resíduos do modelo, nas estações.

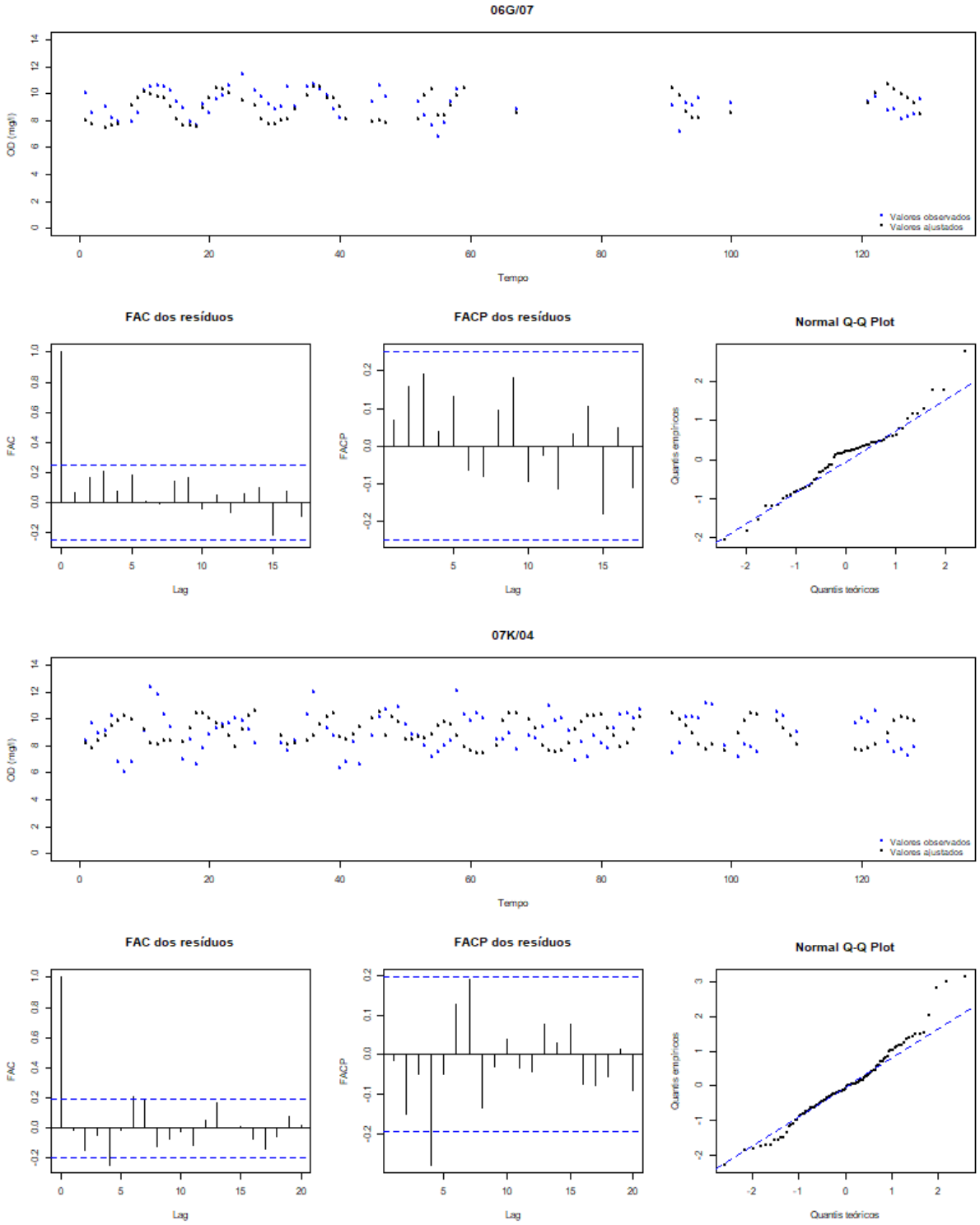


Figura B.31: Representações da série original e dos os valores estimados, da FAC, da FACP e do *Q-Q plot* dos resíduos do modelo, nas estações.

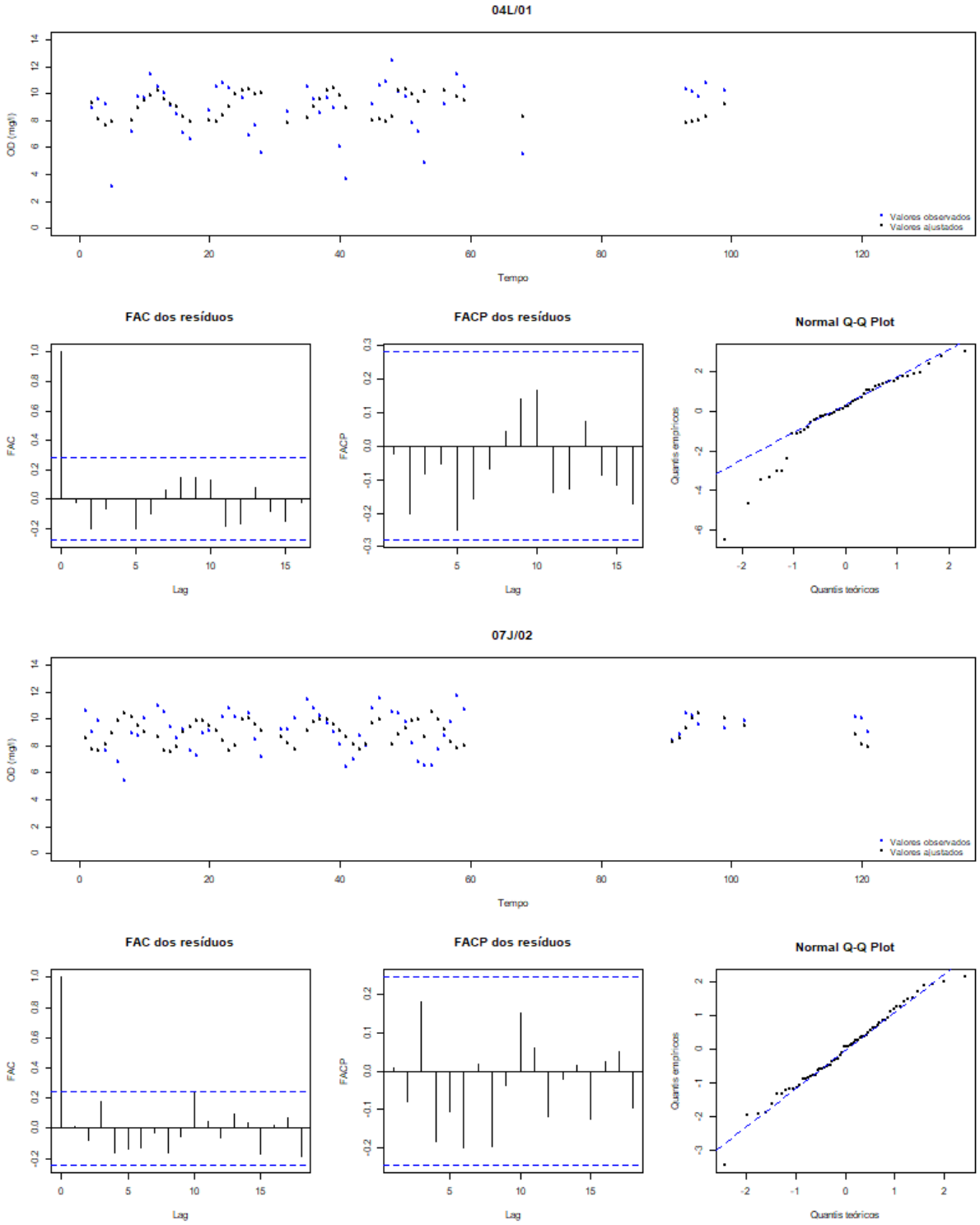


Figura B.32: Representações da série original e dos os valores estimados, da FAC, da FACP e do *Q-Q plot* dos resíduos do modelo, nas estações.

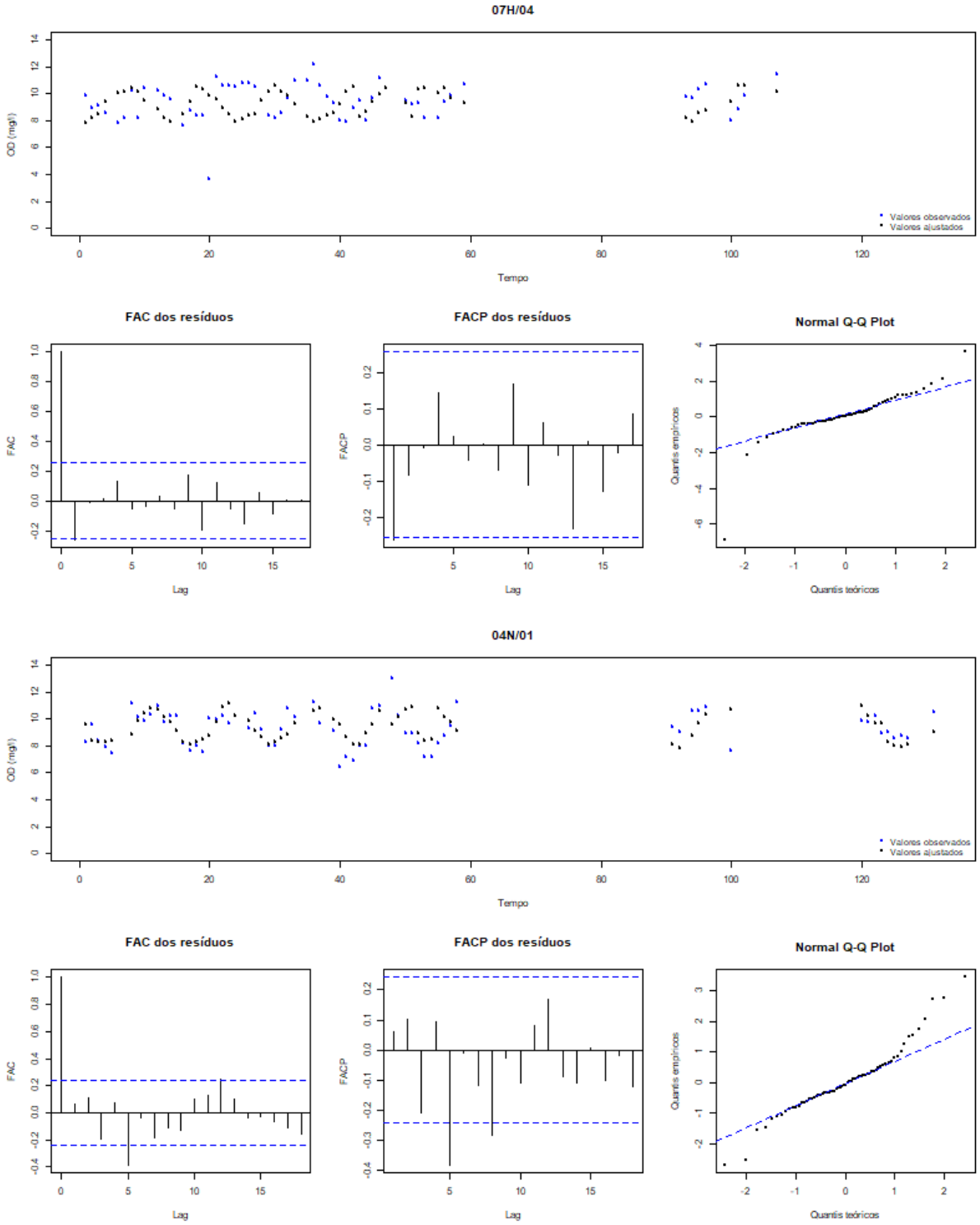


Figura B.33: Representações da série original e dos os valores estimados, da FAC, da FACP e do *Q-Q plot* dos resíduos do modelo, nas estações.

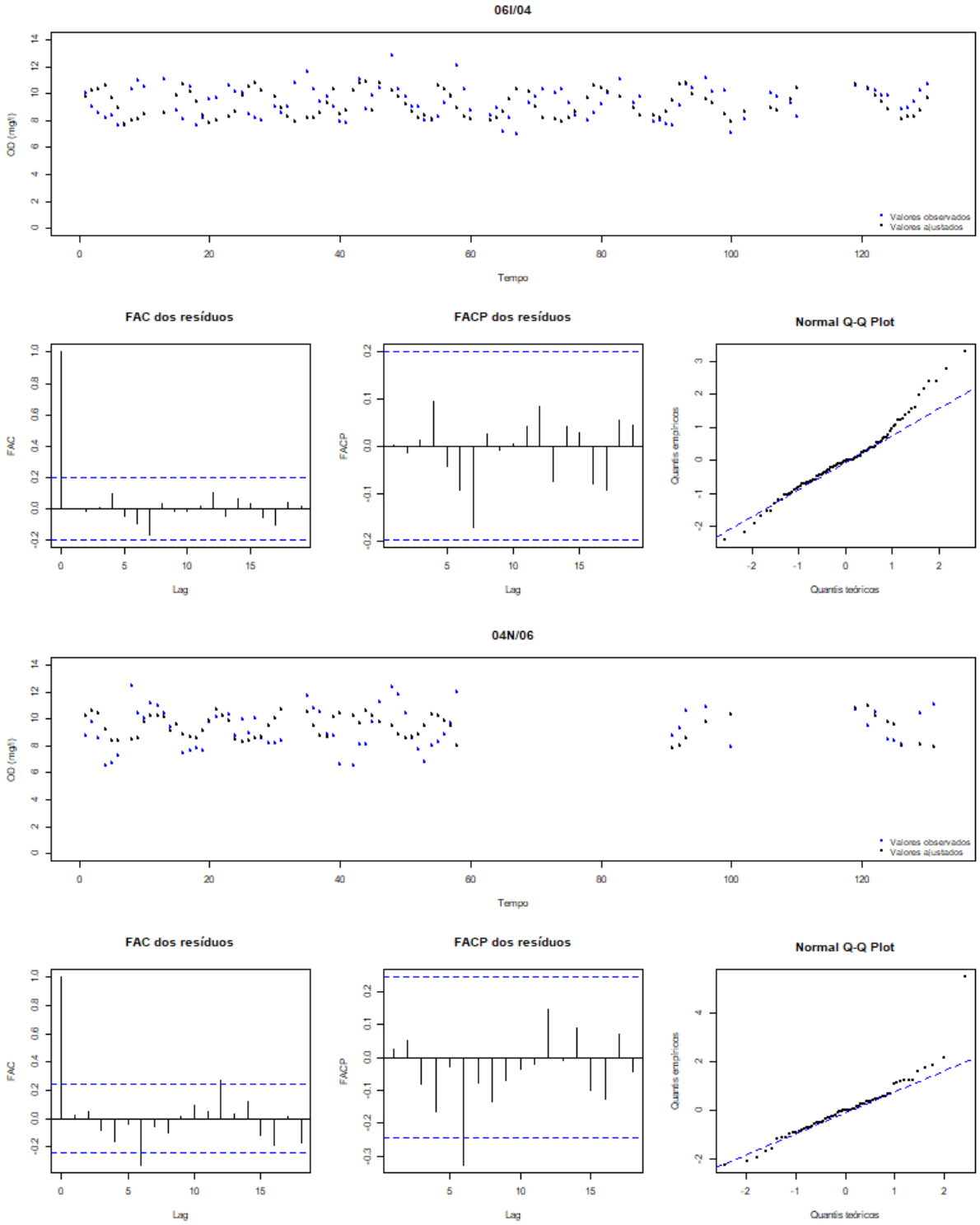


Figura B.34: Representações da série original e dos os valores estimados, da FAC, da FACP e do *Q-Q plot* dos resíduos do modelo, nas estações.

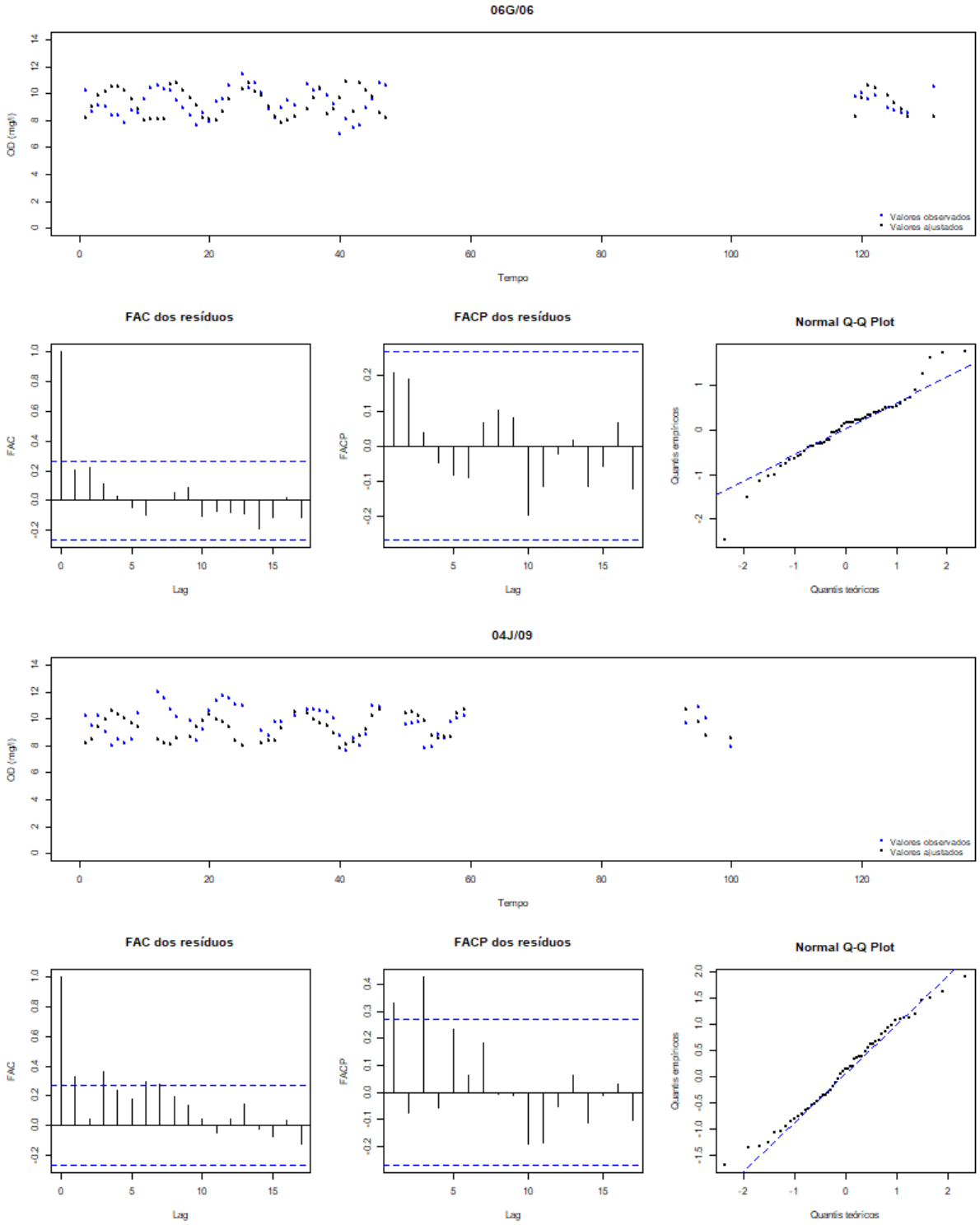


Figura B.35: Representações da série original e dos os valores estimados, da FAC, da FACP e do *Q-Q plot* dos resíduos do modelo, nas estações.