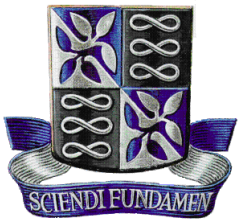
Universidade Federal da Bahia - UFBA
Instituto de Matemática e Estatística - IME
Programa de Pós-Graduação em Matemática - PGMAT

Dissertação de Mestrado



O modelo unit-Lindley autorregressivo e de médias móveis (ULARMA) aplicado no monitoramento e previsão de dados contínuos no intervalo unitário

José Guilherme Santana de Sena

Área de Concentração: Estatística

Salvador - Bahia

Abril de 2024

O modelo unit-Lindley autorregressivo e de médias móveis (ULARMA) aplicado no monitoramento e previsão de dados contínuos no intervalo unitário

José Guilherme Santana de Sena

Dissertação de Mestrado apresentada ao Colegiado da Pós-Graduação em Matemática da Universidade Federal da Bahia (UFBa), como parte dos requisitos para obtenção do título de Mestre em Matemática. Área de concentração: Estatística.

Orientador: Prof. Dr. Paulo Henrique Ferreira da Silva
Coorientadora: Profa. Dra. Rosemeire Leovigildo Fiaccone

Salvador - Bahia
Abril de 2024

O modelo unit-Lindley autorregressivo e de médias móveis (ULARMA) aplicado no monitoramento e previsão de dados contínuos no intervalo unitário

José Guilherme Santana de Sena

Dissertação apresentada ao Colegiado do Curso de Pós-graduação em Matemática da Universidade Federal da Bahia, como requisito parcial para obtenção do Título de Mestre em Matemática com área de concentração em Estatística.

Banca examinadora

Documento assinado digitalmente
gov.br
PAULO HENRIQUE FERREIRA DA SILVA
Data: 20/05/2024 02:16:49-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Paulo Henrique Ferreira da Silva (UFBA)

Documento assinado digitalmente
gov.br
PAULO JORGE CANAS RODRIGUES
Data: 14/05/2024 19:45:34-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Paulo Jorge Canas Rodrigues (UFBA)

Documento assinado digitalmente
gov.br
FABIO MARIANO BAYER
Data: 13/05/2024 16:55:40-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Fábio Mariano Bayer (UFSM)

Ficha catalográfica elaborada pela Biblioteca Universitária de Ciências e Tecnologias Prof. Omar Catunda, SIBI – UFBA.

S474 Sena, José Guilherme Santana de

O modelo unit-Lindley autorregressivo e de médias móveis (ULARMA) aplicado no monitoramento e previsão de dados contínuos no intervalo unitário / José Guilherme Santana de Sena. – Salvador, 2024.

108 f.

Orientador: Prof. Dr. Paulo Henrique Ferreira da Silva
Coorientadora: Prof.ª Dr.ª Rosemeire Leovigildo Fiaccone

Dissertação (Mestrado) – Universidade Federal da Bahia, Instituto de Matemática e Estatística, 2024.

1. Dados Autocorrelacionados. 2. Distribuição Unit-Lindley. 3. Gráfico de Controle. 4. Taxas e Proporções. 5. Estatística. I. Silva, Paulo Henrique Ferreira da. II. Fiaccone, Rosemeire Leovigildo. III. Universidade Federal da Bahia. IV. Título.

CDU:519.22

Dedico este trabalho à minha mãe, Rubinalva, que é a minha maior incentivadora e esteve presente comigo durante toda essa jornada.

Agradecimentos

Primeiramente a Deus, pelas oportunidades que me foram concedidas, pelas bençãos e graças alcançadas durante este caminho, por me guiar nos momentos de dúvidas e incertezas e, principalmente, pela resiliência nos momentos de dificuldade.

À minha mãe, por depositar toda confiança necessária em mim, por não medir esforços na hora de me auxiliar e por estar sempre ao meu lado, compartilhando momentos de alegrias e tristezas.

À minha companheira, Beatriz, pelas palavras de conforto quando necessário, apoio nos momentos difíceis, pelas palavras de incentivo e, acima de tudo, pela paciência para lidar com as situações adversas durante esse período.

A meus orientadores, Paulo Henrique Ferreira e Rosemeire Fiaccone, pela paciência, compreensão, ensinamentos e por se manterem sempre à disposição.

Aos meus amigos, com quem durante essa jornada eu tive a oportunidade de compartilhar momentos de tristeza, felicidade e experiências. Destacando aqui, Renan Bispo e Michelle Vale; sem vocês este trabalho não existiria.

Por fim, mas não menos importante, a todos os professores do Departamento de Estatística da Universidade Federal da Bahia, que contribuíram na minha formação durante todo o curso.

"É isso aí, você não pode parar
Esperar o tempo ruim vir te abraçar
Acreditar que sonhar sempre é preciso
É o que mantém os irmãos vivos."
(Racionais Mc's)

Resumo

Neste trabalho são propostos modelos estatísticos para a análise de dados que exibem variação no tempo. Em particular, quando a variável (ou característica) de interesse é contínua no intervalo $(0, 1)$, como é o caso, por exemplo, de taxas, proporções e índices. Dentre as distribuições de probabilidade no intervalo unitário que foram introduzidas na literatura recente e que possuem propriedades interessantes e úteis (e.g., um único parâmetro, versão reparametrizada em termos da média, expressões fechadas para os momentos), destaca-se a distribuição *unit-Lindley* ou Lindley unitária. Neste trabalho é proposto o modelo *unit-Lindley* autorregressivo e de médias móveis (ULARMA), como uma extensão da distribuição *unit-Lindley* para o caso de dados autocorrelacionados. Além disso, para o controle de futuras observações do processo, são apresentados também gráficos (ou cartas) de controle para monitoramento e previsão de dados desse tipo. Estudos de simulação numérica são realizados para avaliar o desempenho dos procedimentos de estimação (e.g., baseados no método de máxima verossimilhança condicional) e dos gráficos de controle (e.g., baseados no modelo temporal com variável resposta contínua em $(0, 1)$ e descrita pela distribuição *unit-Lindley*) propostos. Por fim, a metodologia aqui desenvolvida é ilustrada em um conjunto de dados reais com informação sobre valores máximos e mínimos da umidade relativa do ar diária, no deserto do Atacama, situado ao norte do Chile, a fim de verificar a sua aplicabilidade em um contexto prático, quando comparada com técnicas tradicionais/existentes.

Palavras-chave: dados autocorrelacionados; distribuição *unit-Lindley*; gráfico de controle; máxima verossimilhança condicional; taxas e proporções.

Abstract

In this work, we propose statistical models for the analysis of data that exhibit variation over time. In particular, when the variable (or characteristic) of interest is continuous in the interval $(0, 1)$, as is the case with rates, proportions, and indices. The unit-Lindley distribution stands out among the probability distributions in the unit interval that have been introduced in recent literature and have interesting and useful properties (e.g., a single parameter, reparameterized version in terms of the mean, closed expressions for moments). In this work, the unit-Lindley autoregressive and moving average (ULARMA) model is proposed, as an extension of the unit-Lindley distribution for the case of autocorrelated data. Furthermore, to take control of future observations of the process, control charts are also presented for monitoring and forecasting data of this type. Numerical simulation studies are carried out to evaluate the performance of estimation procedures (e.g., based on the conditional maximum likelihood method) and control charts (e.g., based on the proposed time series model with a continuous response variable in $(0, 1)$ described by the unit-Lindley distribution). Finally, the methodology developed here is illustrated in a set of real data with information on maximum and minimum values of daily relative air humidity in the Atacama Desert, located in the north of Chile, in order to verify its applicability in a practical context, when compared to traditional/existing techniques.

Keywords: autocorrelated data; unit-Lindley distribution; control chart; conditional maximum likelihood; rates and proportions.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

βARFIMA - *Beta Autoregressive Fractionally Integrated Moving Average*

βARMA - *Beta Autoregressive Moving Average*

βSARMA - *Beta Seasonal Autoregressive Moving Average*

AIC - *Akaike Information Criterion*

ARIMA - Autorregressivo Integrado de Médias Móveis

ARL - *Average Run Length*

BFGS - *Broyden-Fletcher-Goldfarb Shanno*

BIC - Critério de Informação Bayesiano (ou de Schwarz)

CEP - Controle Estatístico de Processos

CLR - *Conditional Likelihood Ratio*

CMLEs - *Conditional Maximum Likelihood Estimators*

CUSUM - *Cumulative Sum*

EUA - Estados Unidos da América

FDA - Função de Distribuição Acumulada

FDP - Função Densidade de Probabilidade

FGM - Função Geradora de Momentos

GAMLSS - *Generalized Additive Models for Location, Scale and Shape*

KARMA - *Kumaraswamy Autoregressive Moving Average*

LC - Linha Central

LIC - Limite Inferior de Controle

LSC - Limite Superior de Controle

MAE - *Mean Absolute Error*

MAPE - *Mean Absolute Percentage Error*

MLE - *Maximum Likelihood Estimator*

MLG - Modelos Lineares Generalizados

MMEP - Média Móvel Exponencialmente Ponderada

MRL - *Median Run Length*

MSE - *Mean Square Error*

PRL - *Percentile Run Length*

RB - *Relative Bias*

RMSE - *Root Mean Square Error*

RS - Rio Grande do Sul

SDRL - *Standard Deviation of the Run Length*

sMAPE - *Symmetric Mean Absolute Percentage Error*

UL - *Unit-Lindley*

ULARMA - *Unit-Lindley* Autorregressivo e de Médias Móveis

Capítulo 1

Introdução

A Estatística é utilizada desde a Antiguidade com a finalidade de coletar, registrar e compreender fenômenos. Os povos egípcios faziam uso para quantificar ovelhas disponíveis, gados possuídos e grãos coletados; os povos romanos utilizavam da Estatística, principalmente, na contagem da população com a finalidade de recolher impostos; durante a Idade Média, era utilizada para tecer previsões e no controle de pragas (Triola et al., 2004). Recentemente, as técnicas de modelagem têm experimentado um aumento da atividade em diversos campos do conhecimento, devido à sua versatilidade e, principalmente, aplicabilidade em situações reais. O barateamento de mecanismos de armazenamento e o aumento da velocidade na coleta da informação corroboram com a popularização e crescimento de dados disponíveis (discretos, contínuos, categóricos, assimétricos, com excesso de curtose etc.), que comumente não seguem a distribuição *Normal* ou *Gaussiana*. Isto, por sua vez, exige métodos cada vez mais específicos e eficazes que permitam extrair *insights* e direcionar para a melhor tomada de decisão. O estudo de características cujos valores são limitados pela própria natureza do fenômeno vem ganhando espaço na literatura dos últimos anos (Prataviera et al., 2021; Martínez-Flórez et al., 2020; Lemonte and Bazán, 2016).

Em particular, essa busca é intensificada quando se trata de características que assumem valores dentro do intervalo unitário padrão, isto é, entre zero e um, como por exemplo, proporções ou frações, escores, índices e taxas. Para a modelagem de variáveis aleatórias que apresentam tais restrições, têm-se utilizado frequentemente as distribuições *Beta* (Ferrari and Cribari-Neto, 2004), *Kumaraswamy* (Kumaraswamy, 1980) e *Simplex* (Barndorff-Nielsen and Jørgensen, 1991). O modelo de regressão *Beta* (Ferrari and Cribari-Neto, 2004) é muito popular na literatura devido à sua flexibilidade, que, por sua vez, incentiva seu uso empírico em uma gama de aplicações, com resultados interessantes quando o suporte está limitado ao intervalo unitário (Martínez-Flórez et al., 2020; Lemonte et al., 2013). Quando se trata do estudo de dados com origem hidrológica que

podem ser mensurados em um intervalo duplamente limitado (e.g., umidade relativa do ar), destaca-se a distribuição *Kumaraswamy*, que é considerada a melhor alternativa à distribuição *Beta*, além de possuir grande flexibilidade, podendo se aproximar de diversas distribuições de probabilidade, com a obtenção de melhores resultados (Bayer et al., 2017; Lemonte et al., 2013). Outra distribuição que merece destaque é a *Simplex*, que, diferente da *Beta* e *Kumaraswamy*, faz parte de uma classe mais geral de modelos, denominada modelos de dispersão, sendo frequentemente utilizada como alternativa ao modelo *Beta* (Altun and El-Morshedy, 2021; Lãpez, 2013). No entanto, apesar da popularidade dessas distribuições, observa-se que nos últimos anos houve um crescimento de novas propostas que são derivadas de distribuições já existentes, com suporte nos reais, para versões com suporte no intervalo unitário padrão (Altun and El-Morshedy, 2021; Sagrillo et al., 2021; Mazucheli et al., 2018a,b).

Dois métodos comuns que têm sido usados para gerar novas distribuições definidas em um intervalo unitário são as transformações logarítmica e unitária. Seguindo a última vertente, Mazucheli et al. (2019) introduziram uma distribuição chamada *unit-Lindley*, que vem ganhando espaço na literatura (Bapat and Bhardwaj, 2021; Wongrin et al., 2020) por desfrutar de propriedades interessantes que outras distribuições restritas ao intervalo unitário não possuem. Do ponto de vista teórico, essa distribuição apresenta forma fechada para a função de distribuição acumulada e função quantil, expressões simples para obtenção dos momentos e pertence à família exponencial. Do ponto de vista prático, a sua principal vantagem reside em ser uma nova distribuição unimodal, uniparamétrica (isto é, com um único parâmetro) e bastante flexível. Além disso, devido à fórmula simples para a média, a distribuição *unit-Lindley* permite incorporar diretamente as covariáveis disponíveis para quantificar sua influência na média da variável resposta, apresentando, assim, um novo modelo de regressão.

Em contextos práticos, observa-se que a grande maioria dos dados que são coletados e ordenados ao longo do tempo apresenta observações que são altamente dependentes à posição temporal em que se encontram, caracterizando uma série temporal (Sena et al., 2022; Bayer et al., 2017; Lohani et al., 2012). Sob essa ótica, a literatura apresenta técnicas que são capazes de considerar a existência da estrutura de dependência intrínseca aos dados, que normalmente é desprezada ao modelar segundo as abordagens tradicionais, e incorporá-la ao processo de estimação. Neste cenário, destaca-se a classe de modelos *autorregressivos integrados de médias móveis* (ARIMA), proposta por Box et al. (2015). No entanto, foi reconhecido que a suposição de distribuição *Gaussiana* associada a este modelo é muito restritiva para várias aplicações (Tiku et al., 2000). Como consequência, tem-se observado um interesse crescente no estudo de variáveis tempo-dependentes com comportamento não *Gaussiano* (Bayer et al., 2018; Benjamin et al., 2003), e alguns

modelos foram propostos como extensões às abordagens tradicionais, tais como: *Beta autoregressive moving average* (Rocha and Cribari-Neto, 2017), *Kumaraswamy autoregressive moving average* (Bayer et al., 2017), *Beta autoregressive fractionally integrated moving average* (Pumi et al., 2019).

Em paralelo ao desenvolvimento de novos modelos, observa-se que as aplicações do Controle Estatístico de Processos (CEP) também têm aumentado em diversos campos. Embora essas técnicas (que também são conhecidas como ferramentas de controle da qualidade) tenham sido desenvolvidas para aplicação em áreas manufatureiras e industriais (Shewhart, 1931), atualmente esse conjunto de métodos é utilizado em diversas áreas, incluindo ecologia, saúde, finanças, indústria e serviços (Sagrillo et al., 2023; Boaventura et al., 2022; Sena et al., 2022; Fonseca et al., 2021). Segundo Ho et al. (2018), os gráficos (ou cartas) de controle têm sido a ferramenta mais utilizada para monitorar a estabilidade dos parâmetros do processo, tais como média, variância ou frações não conforme, além de desempenhar um papel importante em processos de detecção.

O gráfico de controle é considerado uma ferramenta robusta, pois sua aplicação ocorre para diferentes tipos de dados, como contagens, atributos e taxas/proporções. Geralmente, para o monitoramento de taxas ou proporções dos componentes de um produto, utiliza-se os gráficos de controle p e np. Neste caso, as proporções, em sua maioria, são resultados de experimentos de *Bernoulli* e assume-se que seguem aproximadamente uma distribuição *Gaussiana* (Wang, 2009). No entanto, em muitas situações práticas, as taxas e proporções não são resultados de um experimento de *Bernoulli*, apesar de assumirem valores no intervalo unitário padrão. Segundo de Araujo Lima-Filho et al. (2019), a utilização dos gráficos p e np nestes casos possui algumas desvantagens devido à suposição de distribuição *Gaussiana*.

Assumir distribuições com suporte/domínio diferente da variável aleatória observada pode fazer com que os limites de controle apresentem resultados irreais (assumindo valores negativos ou superiores a um), comprometendo o poder para detectar melhorias do processo. Quando as proporções monitoradas não são resultados de experimentos de *Bernoulli*, os gráficos de controle podem ser construídos usando outras distribuições definidas no intervalo $(0, 1)$, como *Beta* (Sant'Anna and ten Caten, 2012), *Kumaraswamy* (Lima-Filho and Bayer, 2021), *Simplex* (Ho et al., 2018) e *unit-Lindley* (Fonseca et al., 2021).

Em alguns casos, o processo de interesse sofre influência de múltiplos fatores que apresentam independência entre si e ignorar o efeito que esses fatores têm sobre o processo pode direcionar a uma tomada de decisão enganosa. Mandel (1969) introduziu uma metodologia capaz de considerar a influência de múltiplos fatores sobre o processo de interesse, denominado gráfico de controle de regressão, cuja ideia principal consiste em

propor um modelo em que a característica de qualidade é a variável dependente e os fatores são as covariáveis, e utilizar as estimativas do modelo proposto como variável a ser monitorada. Montgomery (2020) estende essa abordagem ao considerar a presença de correlação nos dados coletados do processo, propondo o monitoramento dos resíduos como variável a ser monitorada. Tal procedimento apresenta algumas vantagens: os limites de controle obtidos são constantes, possui facilidade de interpretação, os resíduos são não correlacionados e auxilia na visualização do comportamento da série.

1.1 Justificativa do projeto

Os dados oriundos de diversas áreas, tais como ciências ambientais, biologia, ecologia, epidemiologia, sociologia e agricultura, dentre outras, são, muitas vezes, caracterizados pela variabilidade no tempo (Bicalho, 2008). Em particular, os dados limitados no intervalo contínuo $(0, 1)$, tais como taxas, proporções e índices. Nas últimas décadas, tem sido observado um aumento crescente no desenvolvimento de técnicas para a análise de processos desta natureza, devido principalmente à grande aplicabilidade dos modelos temporais. Além disso, com o surgimento de novas e interessantes (e.g., com um único parâmetro, versão reparametrizada pela média, expressões fechadas para os momentos etc.) distribuições de probabilidade para a modelagem de dados limitados no intervalo unitário, como é o caso, por exemplo, da distribuição *unit-Lindley* (Mazucheli et al., 2019), torna-se possível desenvolver, com base nessas distribuições de probabilidade, modelos temporais inéditos, úteis e menos complexos (ou ainda, parcimoniosos). E, também desenvolver, com base nos modelos temporais propostos, gráficos de controle (também inéditos) para o monitoramento e previsão de processos com dados contínuos no intervalo $(0, 1)$.

1.2 Objetivos

1.2.1 Objetivo geral

O objetivo principal deste trabalho consiste em propor um novo modelo para variáveis aleatórias com domínio contido no intervalo unitário padrão e mostrar as suas propriedades inferenciais, que possibilitam a generalização e descrição de padrões ocultos nos dados. Esse modelo estatístico, denominado *unit-Lindley* autorregressivo e de médias móveis, ou ainda, ULARMA, é capaz de considerar a existência de estruturas de dependência temporal nos dados. Além disso, é apresentado neste trabalho um novo gráfico de controle para dados unitários (que não são resultados de um processo de *Bernoulli*), baseado no modelo ULARMA.

1.2.2 Objetivos específicos

São objetivos específicos deste trabalho:

1. Realizar um levantamento bibliográfico dos principais modelos temporais para dados contínuos no intervalo unitário padrão, em especial aqueles que foram introduzidos na literatura recente, bem como possíveis propostas existentes de gráficos de controle baseados nesses modelos;

2. Avaliar o desempenho do modelo proposto e do gráfico de controle associado (inéditos na literatura) por meio de estudos de simulação numérica;

3. Aplicar tais procedimentos a conjuntos de dados reais encontrados na literatura.

1.3 Organização do trabalho

O restante deste trabalho está subdividido em três capítulos e dez seções. O segundo capítulo é composto por cinco seções, sendo que: a Seção 2.1 contém o levantamento teórico e bibliográfico realizado sobre modelos de séries temporais para dados contidos no intervalo unitário; na Seção 2.2 é discutida brevemente a distribuição *unit-Lindley*; na Seção 2.3 é apresentada a proposta deste trabalho, o modelo ULARMA; a Seção 2.4 possui uma aplicação a dados reais do modelo ULARMA; e, por fim, a Seção 2.5 apresenta as conclusões deste capítulo. O terceiro capítulo contém cinco seções, sendo que: a Seção 3.1 contém o levantamento teórico e bibliográfico realizado sobre gráficos de controle aplicados a dados contidos no intervalo unitário; na Seção 3.2 é apresentada uma proposta de gráfico de controle para o modelo ULARMA; a Seção 3.3 contém a descrição e os resultados de um estudo de simulação utilizado para avaliar a performance do gráfico proposto; a Seção 3.4 possui uma aplicação a dados reais do gráfico proposto baseado no modelo ULARMA; e, por fim, a Seção 3.5 apresenta as conclusões deste capítulo. O quarto capítulo contém as considerações finais deste trabalho.

Capítulo 2

Modelagem de dados unitários autocorrelacionados

Neste capítulo são apresentadas técnicas relacionadas ao processo de modelagem de dados autocorrelacionados cujo domínio é restrito ao intervalo unitário padrão. Mais especificamente, é abordada a construção do modelo ULARMA, descrevendo sua formulação, método de estimação, estudo de simulação para avaliar a performance de estimadores, e, por fim, o modelo proposto é aplicado a um conjunto de dados reais sobre umidade relativa do ar no deserto do Atacama, Chile.

2.1 Revisão de literatura

Esta seção apresenta uma breve revisão de literatura a respeito da modelagem de variáveis aleatórias cujo domínio está contido no intervalo unitário $(0, 1)$, considerando dados independentes e dados correlacionados, isto é, que apresentam estrutura de dependência entre si.

Rocha and Cribari-Neto (2017) apresentaram o modelo de regressão para séries temporais de dados unitários baseado na distribuição *Beta*, denominado *Beta autoregressive moving average* (βARMA), com estimação via máxima verossimilhança condicional. Esse modelo consiste em uma extensão da regressão *Beta*, sob a parametrização descrita previamente por Ferrari and Cribari-Neto (2004), em que são incluídos os termos autorregressivos e/ou de médias móveis no preditor linear. Tal abordagem se assemelha à utilizada por Benjamin et al. (2003), que desenvolveram modelos para séries temporais cuja distribuição condicional, dado seu passado histórico, pertence à família exponencial canônica. A performance do modelo proposto foi avaliada em uma aplicação a um conjunto de dados sobre taxas de desemprego oculto devido às condições de trabalho precárias em São Paulo (capital), em que o modelo βARMA$(4, 0)$ foi ajustado aos dados, definido

com base no critério de informação de Akaike (AIC; Akaike et al., 1977).

Segundo Bayer et al. (2017), uma alternativa à distribuição *Beta* consiste na utilização da distribuição *Kumaraswamy*, que é bastante popular em aplicações voltadas a processos hidrológicos e climatológicos, visto que a distribuição *Beta*, mesmo sendo bastante flexível, pode não apresentar ajustes satisfatórios (Lemonte et al., 2013). Posto isto, foi então proposta a construção do modelo *Kumaraswamy autoregressive moving average* (KARMA), que estende o modelo proposto por Mitnik and Baek (2013) para o caso da existência de dependência entre as observações, de forma similar ao que fora descrito por Benjamin et al. (2003). Além disso, foi também discutida a estimação por máxima verossimilhança condicional, inferência por teste de hipóteses, análise de diagnóstico e previsão. Uma aplicação a dados reais sobre umidade relativa do ar em Brasília, capital do Brasil, foi utilizada para comparar a performance do modelo proposto com a do βARMA (Rocha and Cribari-Neto, 2017), em que notou-se que o modelo KARMA$(5, 4)$ apresentou melhor desempenho preditivo, com base nas métricas MSE (do inglês "*mean square error*") e MAPE (do inglês "*mean absolute percentage error*").

Na literatura é descrito que, no contexto de séries temporais, é comum a ocorrência de padrões no comportamento que acontecem de forma periódica, caracterizando a presença do efeito sazonal (Basawa et al., 2004). Bayer et al. (2018) então propuseram o modelo *Beta seasonal autoregressive moving average* (βSARMA), que é uma extensão ao βARMA (Rocha and Cribari-Neto, 2017), capaz de captar flutuações sazonais no processo de modelagem que são impulsionadas por um mecanismo estocástico. Além disso, foram apresentados também os estimadores de máxima verossimilhança condicional para os parâmetros do modelo, testes de hipóteses e ferramentas para análise de diagnóstico. Uma aplicação a dados reais sobre umidade relativa do ar em Santa Maria, Rio Grande do Sul, foi apresentada e discutida.

Como apresentado em Rocha and Cribari-Neto (2017), no modelo βARMA todo processo de inferência realizado baseia-se na abordagem de inferência condicional, que permite a inclusão de covariáveis de origem determinística no modelo. Pumi et al. (2019) ampliaram esse processo após introduzir a classe de modelos *Beta autoregressive fractionally integrated moving average* (βARFIMA), que aborda a verossimilhança parcial, assim possibilitando não somente a inclusão de covariáveis determinísticas, mas também de covariáveis aleatórias (e dependentes no tempo), ou ainda uma interação entre elas. Uma grande vantagem do modelo βARFIMA é a capacidade de acomodar dependência de longo alcance; para isto, assume-se que a estrutura de dependência temporal segue uma estrutura ARFIMA (Brockwell and Davis, 1991), ao invés de ARMA, que pode não ser suficiente em determinadas situações. Além disso, neste trabalho também foram discutidos os testes de hipóteses, ferramentas de diagnóstico e previsões, considerando este modelo.

A performance do modelo foi avaliada em uma aplicação a dados sobre umidade relativa do ar em Manaus, capital do Amazonas.

Mazucheli et al. (2019) apresentaram uma nova distribuição de probabilidade uniparamétrica para variáveis aleatórias contínuas no intervalo unitário, denominada *unit-Lindley*. Essa distribuição considera uma transformação apropriada na classe de distribuições *Lindley* (Lindley, 1958). Várias propriedades estatísticas da distribuição proposta foram estudadas, incluindo o método dos momentos e a estimação por máxima verossimilhança, bem como a expressão analítica para correção do viés do estimador de máxima verossimilhança (MLE, do inglês "*maximum likelihood estimator*"). Além disso, essa distribuição permite incorporar covariáveis diretamente na média e, consequentemente, quantificar suas influências na média da variável resposta (através de um modelo de regressão), sendo vista como uma alternativa mais parcimoniosa em relação ao modelo de regressão *Beta*. Foi apresentada uma aplicação a um conjunto de dados sobre o acesso de pessoas em domicílios com abastecimento inadequado de água e esgotamento sanitário nas cidades das regiões do Sudeste e Nordeste do Brasil, em que o modelo proposto resultou em um ajuste melhor que o modelo *Beta*.

Wongrin et al. (2020) desenvolveram o modelo de regressão Bayesiana *unit-Lindley* baseado em uma priori *Normal*, que frequentemente é utilizada neste cenário. Além disso, foi investigada a priori específica para todas as variáveis exploratórias padronizadas e, para fins de comparação, os modelos de regressão *unit-Lindley* (Mazucheli et al., 2019) e o proposto foram aplicados a dois conjuntos de dados reais, em que as variáveis de interesse eram o rendimento de gasolina e a porcentagem do ativo total, respectivamente. Notavelmente, a regressão Bayesiana *unit-Lindley* apresentou melhores resultados com base nas estimativas produzidas e nos valores da log-verossimilhança.

Akdur (2021) apresentou o modelo *unit-Lindley* de efeitos mistos como alternativa aos modelos de regressão *Beta* e *unit-Lindley* (Mazucheli et al., 2019), para modelagem de variáveis aleatórias contínuas e hierárquicas no intervalo unitário. Um estudo de simulação de Monte Carlo foi conduzido para investigar o desempenho dos métodos desenvolvidos para estimação dos parâmetros do modelo proposto, que incluem aproximação de Laplace e quadratura Gaussiana adaptativa. Além disso, uma aplicação a dados reais sobre proporção de domicílios com abastecimento insuficiente de água e esgoto foi apresentada usando o modelo *unit-Lindley* com intercepto aleatório para os estados do Brasil. Os resultados obtidos indicaram que o modelo proposto forneceu um melhor ajuste quando comparado às regressões *Beta* e *unit-Lindley* tradicionais, em termos de log-verossimilhança e AIC.

Bapat and Bhardwaj (2021) introduziram uma versão da distribuição *unit-Lindley* que considera cenários em que ocorre inflação de zeros e/ou uns, denominada *inflated unit-Lindley*. Foram discutidas propriedades relacionadas a essa distribuição, que incluem a

função de densidade, expressões para os momentos e métodos de estimação pontual e intervalar. Um estudo de simulação de Monte Carlo foi apresentado, comparando o desempenho da distribuição proposta ao de outras já existentes, como a *Beta* e a *Kumaraswamy* inflacionadas, em que observou-se que a *inflated unit-Lindley* obteve melhor ajuste. Foi ressaltado que esse modelo é adequado para situações em que o parâmetro θ da distribuição assume valores superiores a 1,3, o que, em termos práticos, pode ser de rara ocorrência. Além disso, duas aplicações foram apresentadas: na primeira, o interesse residia na porcentagem de elefantes recém-nascidos que tiveram suas cabeças formadas até a metade da gestação, entre diferentes tamanhos de rebanho; e, na segunda, utilizou-se um conjunto de dados fictícios, gerados apenas para ilustrar o método considerando uma amostra maior.

Bayer et al. (2023) propuseram os modelos autorregressivos e de médias móveis *Beta* inflacionada (IβARMA) para modelar a média condicional da variável beta inflacionada condicionalmente distribuída observada ao longo do tempo. Neste trabalho foi apresentada uma estrutura de regressão em função da média, estimadores de máxima verossimilhança parcial, expressões de forma fechada para o vetor de pontuação, matriz de informação parcial cumulativa, testes de hipóteses, intervalos de confiança e algumas ferramentas de diagnóstico e previsão. Os estimadores propostos apresentaram bom desempenho avaliado por meio de um estudo de simulação de Monte Carlo. Além disso, foram apresentadas duas aplicações a dados de contextos hidroambientais.

2.2 Distribuição *unit-Lindley*

Introduzida por Mazucheli et al. (2019), a distribuição *unit-Lindley* surge por meio de uma transformação na classe de distribuições *Lindley* (Lindley, 1958). Segundo Mazucheli et al. (2019), essa distribuição apresenta propriedades interessantes que outras distribuições restritas ao intervalo unitário não possuem, tais como: (i) distribuição unimodal com apenas um parâmetro; (ii) grande flexibilidade, permitindo aplicação em diferentes cenários; (iii) possui forma fechada para as funções de distribuição acumulada e quantil; e (iv) expressões simples para os momentos.

Considere que Y é uma variável aleatória com distribuição *Lindley* com função densidade de probabilidade (FDP) e função de distribuição acumulada (FDA) definidas, respectivamente, por:

$$f(y|\theta) = \frac{\theta^2}{1+\theta}(1+y)\exp\{-\theta y\}, \quad y > 0,\, \theta > 0$$

e

$$F(y|\theta) = 1 - \left(1 + \frac{\theta y}{1+\theta}\right)\exp\{-\theta y\}, \quad y > 0,\, \theta > 0.$$

Mazucheli et al. (2019) propuseram então utilizar a transformação $X = Y/(1+Y)$, definindo uma nova distribuição cujo domínio é restrito ao intervalo unitário, denominada *unit-Lindley* (UL). A FDP e a FDA resultantes desta transformação são apresentadas, respectivamente, nas Equações (2.1) e (2.2):

$$f(x|\theta) = \frac{\theta^2}{1+\theta}(1-x)^{-3}\exp\left\{-\frac{\theta x}{1-x}\right\}, \quad 0 < x < 1,\ \theta > 0 \tag{2.1}$$

e

$$F(x|\theta) = 1 - \left(1 - \frac{\theta x}{(1+\theta)(x-1)}\right)\exp\left\{-\frac{\theta x}{1-x}\right\}, \quad 0 < x < 1,\ \theta > 0. \tag{2.2}$$

Na Figura 2.1 é possível verificar diferentes comportamentos para a FDP da distribuição UL ao considerar diferentes valores do parâmetro θ.

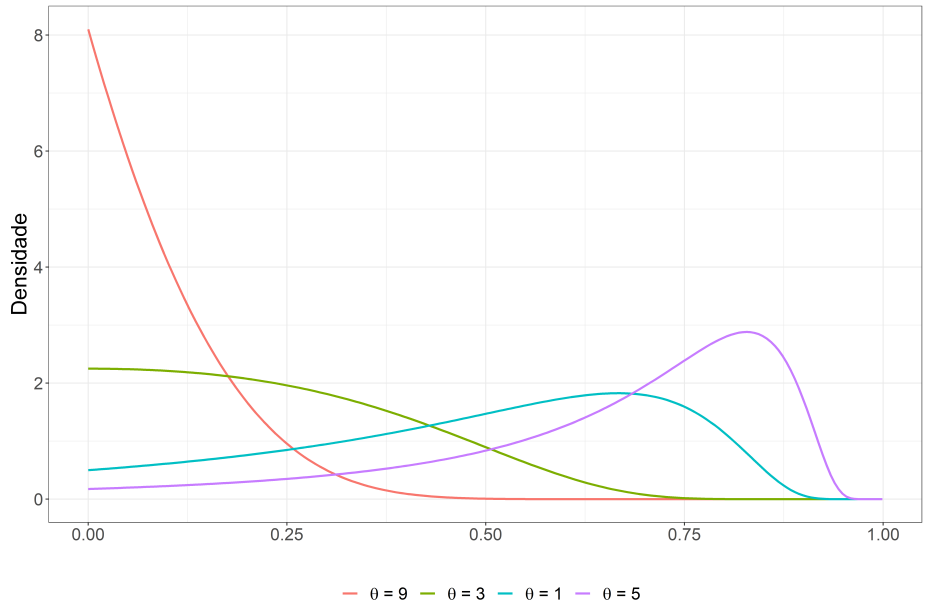


Figura 2.1: Função densidade de probabilidade da distribuição *unit-Lindley*, para diferentes valores de θ.

2.2.1 Função quantil

A função quantil da distribuição UL, $Q(p|\theta) = F^{-1}(p|\theta)$, pode ser escrita conforme apresentado na Equação (2.3):

$$Q(p|\theta) = \frac{1+\theta+W_{-1}[(1+\theta)(p-1)\exp\{-(1+\theta)\}]}{1+W_{-1}[(1+\theta)(p-1)\exp\{-(1+\theta)\}]}, \tag{2.3}$$

em que $0 < p < 1$ e $W_{-1}(\cdot)$ denota o ramo negativo da função $W(\cdot)$ de Lambert (Knuth, 1996). No *software* R, $W_{-1}(\cdot)$ pode ser calculada através da função $lambertWm1(\cdot)$ do pacote `lamW` (Borchers, 2019).

2.2.2 Função geradora de momentos

A função geradora de momentos (FGM) é uma importante ferramenta, pois permite, dentre outras, caracterizar distribuições de probabilidade. Considerando a distribuição UL, o k-ésimo momento em relação à origem da distribuição é apresentado na Equação (2.4):

$$M_k = \mathbb{E}[X^k] = \frac{k}{(1+\theta)} \int_0^1 \frac{x^{k-1}(1-\theta+x)}{(1-x)} \exp\left\{-\frac{\theta x}{1-x}\right\} dx, \quad k = 1, 2, \ldots \tag{2.4}$$

Portanto, dado a Equação (2.4), a média e a variância de X são, respectivamente:

$$\mathbb{E}[X] = \frac{1}{1+\theta} \qquad \text{e} \qquad \mathbb{V}\text{ar}[X] = \frac{1}{1+\theta}\left(\theta^2 \exp\{\theta\} E_i(1,\theta) - \theta + 1 - \frac{1}{1+\theta}\right),$$

em que $E_i(a,z) = \int_1^\infty z^{-a} \exp\{-xz\} dx$ é a função exponencial integral (Abramowitz and Stegun, 1964).

2.2.3 Família exponencial

A família exponencial é uma classe composta por distribuições de probabilidade, capaz de incorporar dados com diferentes comportamentos (assimétricos, discretos, contínuos e intervalares). Diz-se que uma distribuição pertence à família exponencial se sua FDP pode ser escrita conforme a Equação (2.5):

$$f(x|\theta) = \exp\{Q(\theta)T(x|\theta) + D(\theta) + S(x|\theta)\}, \tag{2.5}$$

em que θ é o parâmetro natural, $Q(\cdot)$ e $D(\cdot)$ são funções apenas do parâmetro θ, e $T(\cdot)$ e $S(\cdot)$ são funções da amostra. É possível ver que a distribuição UL pertence à família exponencial e, conforme a Equação (2.5), sua FDP é reescrita como:

$$f(x|\theta) = \exp\left\{-\frac{\theta x}{1-x}\right\} \exp\left\{\log\left(\frac{\theta^2}{1+\theta}\right)\right\} \exp\left\{\log(1-x)^{-3}\right\},$$

em que $Q(\theta) = \theta$, $T(x|\theta) = -x/(1-x)$, $D(\theta) = \log\left(\theta^2/(1+\theta)\right)$ e $S(x|\theta) = \log(1-x)^{-3}$.

2.2.4 Estimação

Conforme descrito em Mazucheli et al. (2019), a estimação do parâmetro θ da distribuição UL pode ser realizada pelo método da máxima verossimilhança. Considerando que $X_1, \ldots, X_n$ é uma amostra aleatória de tamanho n da distribuição UL, então o logaritmo da função de verossimilhança (também chamado de função de log-verossimilhança, denotada por $\ell(\cdot)$), pode ser escrito como:

$$\ell(\theta|\mathbf{x}) \propto n\log(\theta) - n\log(1+\theta) - \theta t(\mathbf{x}),$$

em que $t(\mathbf{x}) = \sum_{i=1}^{n} x_i/(1-x_i)$ é uma estatística suficiente e completa para θ. Dessa forma, tem-se que:

$$\frac{\partial}{\partial\theta}\ell(\theta|\mathbf{x}) = \frac{2n}{\theta} - \frac{n}{1+\theta} - t(\mathbf{x}).$$

Portanto, fazendo $\dfrac{\partial}{\partial\theta}\ell(\theta|\mathbf{x}) = 0$, tem-se o MLE para θ, dado por:

$$\widehat{\theta} = \frac{1}{2t(\mathbf{x})}\left[n - t(\mathbf{x}) + \sqrt{[t(\mathbf{x})]^2 + 6nt(\mathbf{x}) + n^2}\right].$$

2.2.5 Modelo de regressão *unit-Lindley*

Em análise de regressão, tradicionalmente existe o interesse em modelar a média da variável de interesse (variável resposta) em função de outras variáveis, também chamadas de variáveis explicativas ou covariáveis (Ferreira et al., 2022). A distribuição UL possui forma fechada para a média e, portanto, pode ser utilizada nesse contexto, como apresentado por Mazucheli et al. (2019). A reparametrização da distribuição UL em termos da média é apresentada na Equação (2.6):

$$f(y|\mu) = \frac{(1-\mu)^2}{\mu(1-y)^3}\exp\left\{-\frac{y(1-\mu)}{\mu(1-y)}\right\}, \quad 0 < y < 1,\ 0 < \mu < 1. \tag{2.6}$$

Então, se $Y \sim \text{UL}(\mu)$, a média e a variância de Y são dadas, respectivamente, por:

$$\mathbb{E}[Y] = \mu \qquad \text{e} \qquad \mathbb{V}\text{ar}[Y] = \mu\left[\left(\frac{1}{\mu}-1\right)^2\exp\left\{\frac{1}{\mu}-1\right\}E_i\left(1,\left(\frac{1}{\mu}-1\right)\right) - \frac{1}{\mu} + 2\right] - \mu^2.$$

A função quantil para a distribuição $\text{UL}(\mu)$ é definida conforme apresentado na Equação (2.7):

$$Q(p|\mu) = \frac{\frac{1}{\mu} + W_{-1}\left[\frac{(p-1)}{\mu}\exp\left\{-\frac{1}{\mu}\right\}\right]}{1 + W_{-1}\left[\frac{(p-1)}{\mu}\exp\left\{-\frac{1}{\mu}\right\}\right]}, \quad 0 < p < 1. \tag{2.7}$$

Considere que $Y_1, \ldots, Y_n$ é uma amostra aleatória, em que $Y_i \sim \text{UL}(\mu_i)$, para $i = 1, 2, \ldots, n$. Assume-se, então, que μ_i satisfaz à seguinte relação funcional:

$$g(\mu_i) = \mathbf{x}_i^\top\boldsymbol{\beta},$$

em que $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ denota o vetor dos coeficientes de regressão, com $p < n$, e $\mathbf{x}_i^\top = (x_{i1}, \ldots, x_{ip})$ é o vetor de covariáveis. Assume-se que $g(\cdot)$ é uma função monótona e diferenciável que mapeia o intervalo $(0, 1)$ em $\mathbb{R}$, tendo como possíveis candidatas as inversas das FDAs das distribuições *Normal*, *Logística*, entre outras (McCullagh and Nelder, 1989). Segundo Mazucheli et al. (2019), uma forma clássica para obter estimativas

para o vetor de parâmetros $\boldsymbol{\beta}$ consiste na maximização da função de log-verossimilhança (estimação via máxima verossimilhança):

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ell(\mu_i),$$

em que:

$$\ell(\mu_i) = 2\log(1-\mu_i) - \log(\mu_i) - 3\log(1-y_i) - \frac{y_i(1-\mu_i)}{\mu_i(1-y_i)}.$$

O MLE de $\boldsymbol{\beta}$ não pode ser obtido de forma fechada e, por isso, deve-se recorrer a métodos numéricos e iterativos, como os algoritmos de *Newton-Raphson* e *Broyden-Fletcher-Goldfarb Shanno* (BFGS) (Nocedal and Wright, 2006).

2.3 Modelo ULARMA

Neste trabalho é introduzido o modelo *unit-Lindley* autorregressivo e de médias móveis (ULARMA) para variáveis aleatórias cujo domínio pertence ao intervalo unitário, como taxas, índices e proporções, capaz de considerar a existência de dependência temporal entre os dados observados. A abordagem empregada baseia-se na proposta de Benjamin et al. (2003), cuja ideia consiste em modelar a distribuição condicional da série de interesse dado o seu passado histórico, através de uma distribuição que pertence à família exponencial canônica. Variáveis contidas no intervalo unitário tradicionalmente são modeladas considerando as distribuições *Beta*, *Simplex* ou *Kumaraswamy*. Quando existe uma estrutura de dependência entre os dados coletados, suas extensões são consideradas, tais como: βARMA (Ferrari and Cribari-Neto, 2004) e KARMA (Bayer et al., 2017). Neste trabalho é considerada a distribuição *unit-Lindley* (Mazucheli et al., 2019) que, embora seja uma proposta recente, vem ganhando espaço na literatura (Bapat and Bhardwaj, 2021; Akdur, 2021; Fonseca et al., 2021; Wongrin et al., 2020). Vale ressaltar que essa distribuição desfruta de propriedades teóricas e práticas interessantes, como, por exemplo: possui forma analítica para as funções de distribuição e quantil; expressões simples para obtenção dos momentos; e, além disso, é uma distribuição com apenas um parâmetro (uniparamétrica) e unimodal.

Inicialmente, deve-se assumir que a série observada é uma variável aleatória contínua que assume valores dentro do intervalo unitário padrão $(0, 1)$. Seja Y_t, com $t = 1, \ldots, n$, variáveis aleatórias em que a distribuição condicional para cada Y_t, dado um conjunto de informações prévias (isto é, dado o seu passado histórico), segue uma distribuição UL. Como mencionado anteriormente, no contexto de análise de regressão comumente tem-se o interesse em modelar a média da variável sob investigação como função de outras variáveis. Portanto, para o desenvolvimento deste trabalho, utilizou-se a expressão

apresentada na Equação (2.8):

$$f(y_t|\Im_{t-1}) = \frac{(1-\mu_t)^2}{\mu_t(1-y_t)^3}\exp\left\{-\frac{y_t(1-\mu_t)}{\mu_t(1-y_t)}\right\}, \tag{2.8}$$

em que $\Im_{t-1}$ representa o passado histórico da série (ou ainda, o conjunto de informação prévia da série). Ademais,

$$\mathbb{E}[Y_t|\Im_{t-1}] = \mu_t$$

e

$$\mathbb{V}\text{ar}[Y_t|\Im_{t-1}] = \mu_t\left[\left(\frac{1}{\mu_t}-1\right)^2\exp\left\{\frac{1}{\mu_t}-1\right\}E_i\left(1,\left(\frac{1}{\mu_t}-1\right)\right)-\frac{1}{\mu_t}+2\right]-\mu_t^2$$

representam, respectivamente, a média e a variância condicionais da série. As funções de distribuição acumulada e quantil são dadas como apresentado nas Equações (2.9) e (2.10), respectivamente:

$$F(y_t|\Im_{t-1}) = 1-\left(\frac{\mu_t y_t-1}{y_t-1}\right)\exp\left\{-\frac{y_t(1-\mu_t)}{\mu_t(1-y_t)}\right\} \tag{2.9}$$

e

$$Q(y_t|\Im_{t-1}) = \frac{\frac{1}{\mu_t}+W_{-1}\left[\left(\frac{1}{\mu_t}\right)(y_t-1)\exp\left\{-\frac{1}{\mu_t}\right\}\right]}{1+W_{-1}\left[\left(\frac{1}{\mu_t}\right)(y_t-1)\exp\left\{-\frac{1}{\mu_t}\right\}\right]}. \tag{2.10}$$

Assim como ocorre nos Modelos Lineares Generalizados (MLG) (Nelder and Wedderburn, 1972), a média é relacionada ao preditor linear por meio de uma função monótona e diferenciável $g(\cdot)$ que mapeia o intervalo $(0,1)$ para $\mathbb{R}$, denominada função de ligação. Neste contexto, as funções de ligação mais utilizadas são: *logito*, *probito* e *complementar log-log* (Rocha and Cribari-Neto, 2017). Nota-se a adição do termo τ_t ao preditor linear, que representa a inclusão dos termos autorregressivos e/ou de médias móveis inseridos no modelo. Desta forma, o modelo geral para μ_t é dado por:

$$g(\mu_t) = \eta_t = \mathbf{x}_t^\top\boldsymbol{\beta} + \tau_t,$$

em que $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_\kappa)^\top$ denota o vetor dos coeficientes desconhecidos da regressão, com $\kappa < n$; $\mathbf{x}_t^\top = (x_{t1},\ldots,x_{t\kappa})$ representa o vetor de covariáveis no instante t; e η_t é o preditor linear no instante t.

Para introduzir formalmente o componente τ_t, assim como descrito em Rocha and Cribari-Neto (2017), considera-se inicialmente um modelo ARMA(p,q) como função do termo ε_t que, por sua vez, representa o erro. Sabe-se, por definição, que $\varepsilon_t = g(y_t) - \mathbf{x}_t^\top\boldsymbol{\beta}$. Desta forma,

$$\varepsilon_t = \alpha + \sum_{j=1}^{p}\phi_j\varepsilon_{t-j} + \sum_{l=1}^{q}\theta_l r_{t-l} + r_t,$$

em que r_t representa um erro aleatório; α é uma constante, tal que $\alpha \in \mathbb{R}$; p e q são as ordens dos termos autorregressivos e de médias móveis, respectivamente, tais que $p, q \in \mathbb{N}$;

ϕ_j e θ_l representam os coeficientes autorregressivos e de médias móveis, respectivamente. É assumido que a esperança condicional de r_t dado o passado histórico é igual a zero. Portanto, $\mathbb{E}[\varepsilon_t|\Im_{t-1}] \approx \tau_t$.

$$\mathbb{E}[\varepsilon_t|\Im_{t-1}] \approx \alpha + \sum_{j=1}^{p} \phi_j \varepsilon_{t-j} + \sum_{l=1}^{q} \theta_l r_{t-l} = \alpha + \sum_{j=1}^{p} \phi_j [g(y_{t-j}) - \mathbf{x}_{t-j}^{\top}\boldsymbol{\beta}] + \sum_{l=1}^{q} \theta_l r_{t-l} = \tau_t.$$

Finalmente, define-se o modelo geral para μ_t, apresentado na Equação (2.11):

$$g(\mu_t) = \eta_t = \alpha + \mathbf{x}_t^{\top}\boldsymbol{\beta} + \sum_{j=1}^{p} \phi_j [g(y_{t-j}) - \mathbf{x}_{t-j}^{\top}\boldsymbol{\beta}] + \sum_{l=1}^{q} \theta_l r_{t-l}. \tag{2.11}$$

2.3.1 Inferência

Nesta seção é discutida a obtenção de estimadores pontuais e intervalos de confiança para os parâmetros do modelo proposto, considerando propriedades assintóticas desses estimadores. O processo de estimação dos parâmetros do modelo ULARMA pode ser realizado pelo método da máxima verossimilhança condicional. Os estimadores de máxima verossimilhança condicional (CMLEs, do inglês "*conditional maximum likelihood estimators*") são obtidos ao maximizar o logaritmo da função de verossimilhança condicional. É observado que a função de log-verossimilhança para o vetor de parâmetros, condicional a $\Im_{t-1}$, é nula (ou aproximadamente nula) para as m primeiras observações, em que $m = \max\{p, q\}$. Então, com base na Equação (2.8), pode-se escrever a função de log-verossimilhança como:

$$\ell = \sum_{t=m+1}^{n} \log\left(f(y_t|\Im_{t-1})\right) = \sum_{t=m+1}^{n} \ell(\mu_t),$$

sendo:

$$\ell(\mu_t) = 2\log(1-\mu_t) - \log(\mu_t) - \frac{y_t(1-\mu_t)}{\mu_t(1-y_t)} - 3\log(1-y_t).$$

2.3.1.1 Vetor escore condicional

Seja $\Theta = (\alpha, \boldsymbol{\lambda}^{\top}, \boldsymbol{\gamma}^{\top}, \boldsymbol{\beta}^{\top})^{\top}$ o vetor de parâmetros do modelo ULARMA, com dimensão $(p+q+\kappa+1)$, em que $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)^{\top}$ e $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)^{\top}$. Define-se, então, o vetor composto pelas derivadas parciais de primeira ordem do logaritmo da verossimilhança como **vetor escore**, aqui denotado por $\boldsymbol{U}$. Como $\ell(\cdot)$ não é função direta dos parâmetros do modelo, cada elemento U_j do vetor é obtido pela regra da cadeia para diferenciação. Isto é,

$$U_j = \frac{\partial \ell}{\partial \Theta_j} = \sum_{t=m+1}^{n} \frac{\partial \ell(\mu_t)}{\partial \Theta_j} = \sum_{t=m+1}^{n} \frac{\partial \ell(\mu_t)}{\partial \mu_t} \frac{\partial \mu_t}{\partial \eta_t} \frac{\partial \eta_t}{\partial \Theta_j}.$$

Note que $\partial\mu_t/\partial\eta_t = 1/g'(\mu_t)$. Além disso, tem-se que:

$$\begin{aligned}\frac{\partial\ell(\mu_t)}{\partial\mu_t} &= \frac{2(-1)}{(1-\mu_t)} - \frac{1}{\mu_t} + \frac{-y_t[\mu_t(1-y_t)] - y_t(1-\mu_t)(1-y_t)}{[\mu_t(1-y_t)]^2} \\ &= \frac{-2}{(1-\mu_t)} - \frac{1}{\mu_t} + \frac{-y_t(1-y_t)[\mu_t + 1 - \mu_t]}{[\mu_t(1-y_t)]^2} \\ &= \frac{-2}{(1-\mu_t)} - \frac{1}{\mu_t} - \frac{y_t}{(\mu_t)^2(1-y_t)}.\end{aligned}$$

Será denotado aqui, por simplicidade, que $\partial\ell(\mu_t)/\partial\mu_t = \Psi(\mu_t, y_t)$. Desta forma, para obter a equação de estimação associada à j-ésima posição do vetor Θ, basta resolver a expressão apresentada na Equação (2.12):

$$U_j = \sum_{t=m+1}^{n} \Psi(\mu_t, y_t)\frac{1}{g'(\mu_t)}\frac{\partial\eta_t}{\partial\Theta_j}. \tag{2.12}$$

Então, considerando os parâmetros do modelo ULARMA e a Equação (2.12), para $\Theta_j = \alpha$, tem-se que:

$$\frac{\partial\ell}{\partial\alpha} = \sum_{t=m+1}^{n} \Psi(\mu_t, y_t)\frac{1}{g'(\mu_t)}\frac{\partial\eta_t}{\partial\alpha} = \sum_{t=m+1}^{n} \Psi(\mu_t, y_t)\frac{1}{g'(\mu_t)}\left[1 + \sum_{h=1}^{q}\theta_h\frac{\partial r_{t-h}}{\partial\alpha}\right].$$

Adicionalmente, para $\Theta_j = \beta_l$, considerando $l = 1, \ldots, \kappa$,

$$\frac{\partial\ell}{\partial\beta_l} = \sum_{t=m+1}^{n} \Psi(\mu_t, y_t)\frac{1}{g'(\mu_t)}\frac{\partial\eta_t}{\partial\beta_l} = \sum_{t=m+1}^{n} \Psi(\mu_t, y_t)\frac{1}{g'(\mu_t)}\left[\mathbf{x}_{tl}^{\top} - \sum_{j=1}^{p}\phi_j\mathbf{x}_{(t-j)l}^{\top} + \sum_{h=1}^{q}\theta_h\frac{\partial r_{t-h}}{\partial\beta_l}\right].$$

Para $\Theta_j = \phi_i$, considerando $i = 1, \ldots, p$,

$$\frac{\partial\ell}{\partial\phi_i} = \sum_{t=m+1}^{n} \Psi(\mu_t, y_t)\frac{1}{g'(\mu_t)}\frac{\partial\eta_t}{\partial\phi_i} = \sum_{t=m+1}^{n} \Psi(\mu_t, y_t)\frac{1}{g'(\mu_t)}\left\{[g(y_{t-i}) - \mathbf{x}_{(t-i)}^{\top}\boldsymbol{\beta}] + \sum_{h=1}^{q}\theta_h\frac{\partial r_{t-h}}{\partial\phi_i}\right\}.$$

Para $\Theta_j = \theta_h$, considerando $h = 1, \ldots, q$,

$$\frac{\partial\ell}{\partial\theta_h} = \sum_{t=m+1}^{n} \Psi(\mu_t, y_t)\frac{1}{g'(\mu_t)}\frac{\partial\eta_t}{\partial\theta_h} = \sum_{t=m+1}^{n} \Psi(\mu_t, y_t)\frac{1}{g'(\mu_t)}\left[r_{t-h} + \sum_{h=1}^{q}\theta_h\frac{\partial r_{t-h}}{\partial\theta_h}\right].$$

Os CMLEs $\widehat{\Theta}$ são obtidos ao encontrar a solução do sistema dado por $\boldsymbol{U} = 0$. Em geral, essas equações são não lineares e o sistema deve ser resolvido de forma numérica por processos iterativos, como, por exemplo, via algoritmos de *Newton-Raphson* e BFGS.

2.3.1.2 Matriz de informação observada

Um componente importante na etapa de inferência estatística é a matriz *Hessiana*, aqui denotada por $H(\cdot)$, utilizada na construção de intervalos de confiança para os parâmetros do modelo proposto. Essa matriz é composta pelas derivadas parciais de segunda ordem de $\ell(\cdot)$ com relação aos parâmetros do modelo. Essa matriz é utilizada no cálculo

das variâncias associadas às estimativas de máxima verossimilhança, que por sua vez, assumem um papel importante na formulação de estatísticas de teste. Assim, tem-se que a expressão para cada componente pode ser obtida a partir do seguinte cálculo:

$$\frac{\partial^2 \ell(\mu_t)}{\partial\Theta_i\partial\Theta_j} = \sum_{t=m+1}^{n}\left[\frac{\partial \ell^2(\mu_t)}{\partial\mu_t^2}\left(\frac{\partial\mu_t}{\partial\eta_t}\right)^2\frac{\partial\eta_t}{\partial\Theta_j}\frac{\partial\eta_t}{\partial\Theta_i} + \frac{\partial\ell(\mu_t)}{\partial\mu_t}\frac{\partial}{\partial\mu_t}\left(\frac{\partial\mu_t}{\partial\eta_t}\frac{\partial\eta_t}{\partial\Theta_j}\right)\frac{\partial\mu_t}{\partial\eta_t}\frac{\partial\eta_t}{\partial\Theta_i}\right].$$

Então, generalizando a expressão anterior para os parâmetros do modelo ULARMA, a matriz *Hessiana* é dada por:

$$H(\Theta) = \begin{bmatrix} \frac{\partial^2\ell(\mu_t)}{\partial\Theta_\alpha\partial\Theta_\alpha} & \frac{\partial^2\ell(\mu_t)}{\partial\Theta_\alpha\partial\Theta_\beta} & \frac{\partial^2\ell(\mu_t)}{\partial\Theta_\alpha\partial\Theta_\lambda} & \frac{\partial^2\ell(\mu_t)}{\partial\Theta_\alpha\partial\Theta_\gamma} \\ \frac{\partial^2\ell(\mu_t)}{\partial\Theta_\beta\partial\Theta_\alpha} & \frac{\partial^2\ell(\mu_t)}{\partial\Theta_\beta\partial\Theta_\beta} & \frac{\partial^2\ell(\mu_t)}{\partial\Theta_\beta\partial\Theta_\lambda} & \frac{\partial^2\ell(\mu_t)}{\partial\Theta_\beta\partial\Theta_\gamma} \\ \frac{\partial^2\ell(\mu_t)}{\partial\Theta_\lambda\partial\Theta_\alpha} & \frac{\partial^2\ell(\mu_t)}{\partial\Theta_\lambda\partial\Theta_\beta} & \frac{\partial^2\ell(\mu_t)}{\partial\Theta_\lambda\partial\Theta_\lambda} & \frac{\partial^2\ell(\mu_t)}{\partial\Theta_\lambda\partial\Theta_\gamma} \\ \frac{\partial^2\ell(\mu_t)}{\partial\Theta_\gamma\partial\Theta_\alpha} & \frac{\partial^2\ell(\mu_t)}{\partial\Theta_\gamma\partial\Theta_\beta} & \frac{\partial^2\ell(\mu_t)}{\partial\Theta_\gamma\partial\Theta_\lambda} & \frac{\partial^2\ell(\mu_t)}{\partial\Theta_\gamma\partial\Theta_\gamma} \end{bmatrix}.$$

O negativo da matriz de valores esperados das derivadas parciais de segunda ordem, isto é, $-\mathbb{E}[H(\Theta)]$, é denominado matriz de informação de *Fisher*, denotada por $I(\Theta)$. Sob condições de regularidade, o vetor $\widehat{\Theta}$ é consistente e segue assintoticamente uma distribuição normal com vetor de médias Θ e matriz de covariâncias $\Sigma(\Theta)$. Além disso, sabe-se que para amostras de tamanho grande, isto é, assintoticamente, tem-se que: $\Sigma(\Theta) = I^{-1}(\Theta)$. Em particular, para o modelo ULARMA, a matriz de informação de *Fisher* não possui expressões simples, e, portanto, será utilizada a matriz de informação observada. Nesta matriz, os CMLEs $\widehat{\Theta}$ são aplicados aos componentes da matriz *Hessiana*, o que resulta numa boa aproximação da matriz de informação de *Fisher*.

2.3.1.3 Teste de hipóteses e construção de intervalos de confiança

Os resultados que foram apresentados na Seção 2.3.1.1 permitem a construção de intervalos de confiança assintóticos e estatísticas de teste para avaliar hipóteses. Considere que as hipóteses nula ($\mathcal{H}_0$) e alternativa ($\mathcal{H}_1$) são, respectivamente:

$$\mathcal{H}_0 : \nu\Theta = \mathbf{0} \quad \text{e} \quad \mathcal{H}_1 : \nu\Theta \neq \mathbf{0}, \tag{2.13}$$

em que ν é uma matriz com dimensão $r \times (p+q+\kappa+1)$ de posto r.

Seja $\widehat{\Theta}$ os CMLEs para Θ sob a hipótese nula e $\widetilde{\Theta}$ os CMLEs para Θ sob a hipótese alternativa. Comumente, para testar as hipóteses apresentadas em (2.13), é utilizada a estatística da razão de verossimilhança condicional (CLR, do inglês "*conditional likelihood ratio*") (Rocha and Cribari-Neto, 2009):

$$\omega = 2[\ell(\widehat{\Theta}) - \ell\widetilde{\Theta}],$$

em que $\ell(\cdot)$ denota o logaritmo da função de verossimilhança condicional.

Sob condições de regularidade, ω converge em distribuição para uma distribuição *Qui-Quadrado* com r graus de liberdade (isto é, $\omega \xrightarrow{\mathcal{D}} \chi^2_r$). Outra forma de testar as hipóteses apresentadas em (2.13) consiste em considerar a estatística Z, definida como a raiz quadrada da estatística CLR, que assintoticamente segue a distribuição *Normal* padrão sob $\mathcal{H}_0$. Isto é,

$$Z = \frac{\nu\widehat{\Theta} - \nu\Theta}{\sqrt{\nu\,\mathrm{diag}(H(\widehat{\Theta}))}} \xrightarrow{\mathcal{D}} N_p(\mathbf{0}, \mathbf{1}),$$

em que $\mathrm{diag}(H(\widehat{\Theta}))$ denota um vetor com os elementos da diagonal de $H(\widehat{\Theta})$.

Ao utilizar a normalidade assintótica dos CMLEs de Θ, pode-se facilmente construir intervalos de confiança para os elementos de Θ como segue:

$$\left(\widehat{\Theta}_j - z_{\alpha/2}\sqrt{h_{jj}(\widehat{\Theta})};\ \widehat{\Theta}_j + z_{\alpha/2}\sqrt{h_{jj}(\widehat{\Theta})}\right),$$

em que $\widehat{\Theta}_j$ é o CMLE para a j-ésima componente de Θ, h^{-1}_{jj} é o j-ésimo componente da diagonal da matriz de informação observada, e $z_{\alpha/2}$ denota o quantil da distribuição *Normal* padrão que deixa uma probabilidade $\alpha/2$ na cauda direita.

2.3.2 Previsão

A previsão é uma etapa importante no processo de extrapolação, isto é, de estender as estimativas do modelo proposto para dados não observados em pontos posteriores no tempo. Conforme descrito em Palm (2016), o processo para previsão de valores futuros (y_{n+h}), em que h representa o número de passos à frente, é dado por:

$$\widehat{Y_t}(h) = g^{-1}\left(\widehat{\alpha} + \mathbf{x}_t^\top\widehat{\boldsymbol{\beta}} + \sum_{j=1}^{p}\widehat{\phi_j}[g(y_{t+h-j}) - \mathbf{x}_{t-j}^\top\widehat{\boldsymbol{\beta}}] + \sum_{l=1}^{q}\widehat{\theta_l}r_{t+h-l}\right).$$

Nota-se, então, que:

$$g(Y_{t+h-j}) = \begin{cases} g(\widehat{Y_t}(h-j)), \text{ se } j < h, \\ g(Y_{t+h-j}), \text{ se } j \geq h \end{cases}$$

e

$$r_{t+h-l} = \begin{cases} 0, \text{ se } l < h, \\ g(Y_{t+h-l}) - g(\widehat{\mu}_{t+(h-l)}), \text{ se } l \geq h. \end{cases}$$

2.3.3 Estudo de simulação

Nesta seção são apresentados e discutidos os resultados obtidos por meio de um estudo de simulação de Monte Carlo que avalia a performance dos CMLEs, desenvolvidos

nas Seções 2.3.1.1 e 2.3.1.2. Para avaliação numérica foram consideradas 10.000 réplicas em amostras de tamanho $n = \{70, 100, 200, 300\}$. Conforme apontado por Schaffer and Kim (2007), esse número de réplicas é suficiente para obter resultados precisos. Em cada réplica de Monte Carlo, foram gerados n valores para uma variável restrita ao intervalo unitário padrão de um modelo ULARMA$(1, 1)$ e ULARMA$(2, 2)$, com função de ligação *logito*. Ao todo, foram analisados sete cenários distintos.

Considerando o modelo ULARMA$(1, 1)$, foram simulados cenários: i) sem covariáveis e parâmetros $\alpha = -1{,}00$, $\phi = -0{,}50$ e $\theta = 0{,}25$; ii) com uma covariável contínua seguindo a distribuição *Normal* padrão e parâmetros $\alpha = -1{,}00$, $\phi = -0{,}50$, $\theta = 0{,}25$ e $\beta = 0{,}50$; iii) com uma covariável binária seguindo a distribuição *Binomial* (com $n = 1$ e $p = 0{,}4$) e parâmetros $\alpha = -1{,}00$, $\phi = -0{,}50$, $\theta = 0{,}25$ e $\beta = 10{,}5$. Para o modelo ULARMA$(2, 2)$, foram simulados cenários: i) sem covariáveis e parâmetros $\alpha = 0{,}50$, $\boldsymbol{\phi} = (0{,}50; -0{,}30)$ e $\boldsymbol{\theta} = (0{,}40; 0{,}15)$; ii) com uma covariável contínua seguindo a distribuição *Normal* padrão e parâmetros $\alpha = 0{,}50$, $\boldsymbol{\phi} = (0{,}50; -0{,}30)$, $\boldsymbol{\theta} = (0{,}40; 0{,}15)$ e $\beta = 1{,}20$; iii) com uma covariável binária seguindo a distribuição *Binomial* (com $n = 1$ e $p = 0{,}7$) e parâmetros $\alpha = 0{,}50$, $\boldsymbol{\phi} = (0{,}50; -0{,}30)$, $\boldsymbol{\theta} = (0{,}40; 0{,}15)$ e $\beta = -4{,}50$; iv) com duas covariáveis com distribuições *Normal* padrão e *Gama* (com $a = 1/2$ e $b = 3$) e parâmetros $\alpha = 0{,}50$, $\boldsymbol{\phi} = (0{,}50; -0{,}30)$, $\boldsymbol{\theta} = (0{,}40; 0{,}15)$ e $\boldsymbol{\beta} = (8{,}4; -6{,}9)$. Para gerar valores das distribuições *Normal*, *Binomial* e *Gama*, foram utilizadas, respectivamente, as funções *rnorm(·)*, *rbinom(·)* e *rgamma(·)* do pacote `MASS` do *software* R.

Para avaliação de desempenho, foram calculados: a média das estimativas, o MSE e o viés relativo (RB, do inglês "*relative bias*") percentual. O RB é definido como a razão entre o viés e o verdadeiro valor do parâmetro multiplicada por 100%. Todas as maximizações da função de log-verossimilhança condicional foram realizadas usando o método BFGS. Para gerar amostras de tamanho n do modelo ULARMA(p, q), foi considerado o algoritmo apresentado em Bayer et al. (2017), desenvolvido para o modelo *Kumaraswamy* e adaptado aqui para o modelo ULARMA. O primeiro passo do algoritmo consiste em definir $r_t = 0$ e $\mu_t = g^{-1}(\alpha)$, para $t = 1, \ldots, m$. No segundo passo, obtém-se η_t, para $t = m + 1$, como apresentado na Equação (2.11), e, então, tem-se que $\mu_t = g^{-1}(\eta_t)$. No terceiro e último passo, y_t é gerado por meio da função de densidade mostrada na Equação (2.8).

Nas Tabelas 2.1,2.2 e 2.3 são apresentados os resultados do estudo de simulação para os modelos ULARMA$(1, 1)$ e ULARMA$(2, 2)$, respectivamente. Observa-se que, em geral, a performance dos CMLEs é bastante satisfatória, mesmo para o cenário em que o tamanho de amostra é pequeno ($n = 70$). A precisão das estimativas melhora com o aumento da amostra e, consequentemente, ocorre a redução do RB. Os estimadores que possuem o menor RB são $\widehat{\alpha}$ e $\widehat{\beta}$, enquanto $\widehat{\theta}$ possui o maior. As estimativas do componente

autorregressivo se mostram mais precisas quando comparado ao componente de médias móveis; tal achado é apresentado e discutido em outros trabalhos (Ansley and Newbold, 1980) sobre modelos autoregressivos e de médias móveis (ARMA). Isto implica em dizer que a inferência realizada sobre o componente de médias móveis é menos precisa (ou ainda, assertiva), em comparação aos demais parâmetros. Em todos os cenários, o MSE obtido é bem pequeno.

2.3.4 Critérios para seleção de modelos

Para avaliar o ajuste do modelo aos dados serão utilizados como critérios de seleção/discriminação de modelos: o AIC e o critério de informação de Schwarz ou Bayesiano.

- **Critério de Informação de Akaike (AIC):** definida por Akaike et al. (1977), esta métrica penaliza a função de verossimilhança obtida através dos dados, pelo número de parâmetros a serem estimados pelo modelo proposto. Pode ser calculado pela seguinte expressão:

$$\text{AIC} = -2\ell(\widehat{\boldsymbol{\varphi}}) + 2k,$$

  em que k denota o número de parâmetros a serem estimados e $\ell(\widehat{\boldsymbol{\varphi}})$ representa o máximo da função de log-verossimilhança. O modelo selecionado deve ser aquele cujo AIC seja o mínimo.

- **Critério de Informação Bayesiano (BIC)**: consiste, assim como o AIC, em uma métrica de penalização da verossimilhança, que caracteriza-se por indicar modelos com menor número de parâmetros, visto que penaliza mais fortemente modelos com mais parâmetros Schwarz, 1978. É obtido pela seguinte expressão:

$$\text{BIC} = -2\ell(\widehat{\boldsymbol{\varphi}}) + k\log(n),$$

  em que n representa o número de observações. O modelo que apresentar o menor valor de BIC deve ser escolhido.

A avaliação da performance preditiva dos modelos será feita por meio das métricas: Raiz do Erro Quadrático Médio (RMSE), Erro Absoluto Médio (MAE) e *Symmetric Mean Absolute Percentage Error* (sMAPE). Tais métricas são amplamente utilizadas no estudo e análise de séries temporais, para comparar previsões realizadas por diferentes modelos.

- **Raiz do Erro Quadrático Médio**: denotada por RMSE (do inglês "*root mean square error*"). Em algumas áreas, tais como meteorologia e geociências, esta métrica é intitulada como padrão para avaliação de desempenho entre modelos (Savage et al., 2013). No entanto, alguns trabalhos não sugerem a utilização dessa métrica,

Tabela 2.1: Resultados da simulação de Monte Carlo para os CMLEs baseados no modelo ULARMA$(1,1)$.

		Cenário 1			
Parâmetro	Métrica	$n = 70$	$n = 100$	$n = 200$	$n = 300$
$\alpha = -1{,}00$	Média	-0,9579	-0,9726	-0,9863	-0,9915
	RB (%)	-4,2109	-2,7430	-1,3719	-0,8539
	MSE	0,0530	0,0350	0,0146	0,0089
$\phi = -0{,}50$	Média	-0,4035	-0,4338	-0,4660	-0,4787
	RB (%)	-19,3014	-13,2464	-6,7930	-4,2639
	MSE	0,0995	0,0643	0,0249	0,0144
$\theta = 0{,}25$	Média	0,1515	0,1814	0,2152	0,2286
	RB (%)	-39,4141	-27,4297	-13,9204	-8,5617
	MSE	0,1155	0,0737	0,0303	0,0178
		Cenário 2			
Parâmetro	Métrica	$n = 70$	$n = 100$	$n = 200$	$n = 300$
$\alpha = -1{,}00$	Média	-0,9636	-0,9759	-0,9893	-0,9927
	RB (%)	-3,6396	-2,4071	-1,0739	-0,7329
	MSE	0,0544	0,0355	0,0141	0,0090
$\phi = -0{,}50$	Média	-0,3973	-0,4331	-0,4685	-0,4789
	RB (%)	-20,5461	-13,3740	-6,3034	-4,2225
	MSE	0,1017	0,0633	0,0237	0,0146
$\theta = 0{,}25$	Média	0,1405	0,1807	0,2177	0,2287
	RB (%)	-43,7849	-27,7366	-12,9302	-8,5197
	MSE	0,1210	0,0737	0,0292	0,0180
$\beta = 0{,}50$	Média	0,4994	0,4994	0,4998	0,4999
	RB (%)	-0,1248	-0,1121	-0,0430	-0,0246
	MSE	0,0102	0,0079	0,0035	0,0021
		Cenário 3			
Parâmetro	Métrica	$n = 70$	$n = 100$	$n = 200$	$n = 300$
$\alpha = -1{,}00$	Média	-1,0137	-1,0055	-0,9994	-0,9975
	RB (%)	1,3705	0,5531	-0,0590	-0,2488
	MSE	0,0360	0,0251	0,0126	0,0089
$\phi =$ -0,50	Média	-0,4892	-0,4868	-0,4896	-0,4900
	RB (%)	-2,1601	-2,6500	-2,0833	-2,0022
	MSE	0,0086	0,0088	0,0069	0,0058
$\theta = 0{,}25$	Média	0,2417	0,2368	0,2393	0,2404
	RB (%)	-3,3308	-5,2861	-4,2917	-3,8219
	MSE	0,0190	0,0165	0,0105	0,0083
$\beta = 10{,}5$	Média	10,4985	10,4999	10,4998	10,4990
	RB (%)	-0,0145	-0,0009	-0,0022	-0,0098
	MSE	0,0352	0,0231	0,0109	0,0073

Tabela 2.2: Resultados da simulação de Monte Carlo para os CMLEs baseados no modelo ULARMA$(2, 2)$.

		Cenário 1			
Parâmetro	Métrica	$n = 70$	$n = 100$	$n = 200$	$n = 300$
$\alpha = 0{,}50$	Média	0,4219	0,4426	0,4719	0,4809
	RB (%)	-15,6111	-11,4704	-5,6265	-3,8199
	MSE	0,0570	0,0413	0,0196	0,0123
$\phi_1 = 0{,}50$	Média	0,6263	0,5901	0,5471	0,5322
	RB (%)	25,2644	18,0225	9,4253	6,4434
	MSE	0,1545	0,1178	0,0573	0,0352
$\phi_2 = -0{,}30$	Média	-0,3413	-0,3285	-0,3145	-0,3101
	RB (%)	13,7667	9,4896	4,8470	3,3753
	MSE	0,0218	0,0156	0,0077	0,0051
$\theta_1 = 0{,}40$	Média	0,2555	0,2992	0,3486	0,3660
	RB (%)	-36,1169	-25,1953	-12,8486	-8,4972
	MSE	0,1889	0,1368	0,0631	0,0378
$\theta_2 = 0{,}15$	Média	0,0740	0,0948	0,1216	0,1317
	RB (%)	-50,6607	-36,7670	-18,9241	-12,1726
	MSE	0,0896	0,0627	0,0268	0,0161
		Cenário 2			
Parâmetro	Métrica	$n = 70$	$n = 100$	$n = 200$	$n = 300$
$\alpha = 0{,}50$	Média	0,4299	0,4470	0,4707	0,4795
	RB (%)	-14,0294	-10,5930	-5,8640	-4,1082
	MSE	0,0559	0,0392	0,0192	0,0123
$\phi_1 = 0{,}50$	Média	0,5936	0,5707	0,5443	0,5282
	RB (%)	18,7213	14,1494	8,8536	5,6309
	MSE	0,1435	0,1064	0,0541	0,0342
$\phi_2 = -0{,}30$	Média	-0,3351	-0,3243	-0,3140	-0,3088
	RB (%)	11,6911	8,0870	4,6643	2,9368
	MSE	0,0226	0,0159	0,0077	0,0051
$\theta_1 = 0{,}40$	Média	0,2998	0,3256	0,3546	0,3718
	RB (%)	-25,0503	-18,5931	-11,3441	-7,0618
	MSE	0,1739	0,1212	0,0590	0,0363
$\theta_2 = 0{,}15$	Média	0,0968	0,1064	0,1224	0,1331
	RB (%)	-35,4976	-29,0592	-18,3935	-11,2714
	MSE	0,0860	0,0554	0,0254	0,0150
$\beta_1 = 1{,}20$	Média	1,1985	1,1999	1,2005	1,2003
	RB (%)	-0,1257	-0,0076	0,0392	0,0242
	MSE	0,0068	0,0040	0,0016	0,0010

Tabela 2.3: Resultados da simulação de Monte Carlo para os CMLEs baseados no modelo ULARMA$(2, 2)$. *(continuação)*

		Cenário 3			
Parâmetro	Métrica	$n = 70$	$n = 100$	$n = 200$	$n = 300$
$\alpha = 0{,}50$	Média	0,4045	0,4184	0,4550	0,4672
	RB (%)	-19,0987	-16,3115	-9,0000	-6,5671
	MSE	0,0804	0,0564	0,0279	0,0181
$\phi_1 = 0{,}50$	Média	0,5549	0,5726	0,5467	0,5353
	RB (%)	10,9807	14,5104	9,3472	7,0542
	MSE	0,1626	0,1209	0,0576	0,0372
$\phi_2 = -0{,}30$	Média	-0,3239	-0,3273	-0,3181	-0,3135
	RB (%)	7,9548	9,1111	6,0343	4,5082
	MSE	0,0277	0,0192	0,0089	0,0058
$\theta_1 = 0{,}40$	Média	0,3401	0,3241	0,3519	0,3643
	RB (%)	-14,9791	-18,9769	-12,0142	-8,9166
	MSE	0,1921	0,1348	0,0616	0,0394
$\theta_2 = 0{,}15$	Média	0,1179	0,1064	0,1246	0,1310
	RB (%)	-21,3764	-29,0529	-16,9492	-12,6661
	MSE	0,0924	0,0590	0,0256	0,0161
$\beta_1 = -4{,}50$	Média	-4,5028	-4,5010	-4,5001	-4,4990
	RB (%)	0,0621	0,0224	0,0031	-0,0217
	MSE	0,0271	0,0198	0,0093	0,0059
		Cenário 4			
Parâmetro	Métrica	$n = 70$	$n = 100$	$n = 200$	$n = 300$
$\alpha = 0{,}50$	Média	0,4497	0,4658	0,4791	0,4848
	RB (%)	-10,0636	-6,8450	-4,1828	-3,0448
	MSE	0,0464	0,0319	0,0163	0,0112
$\phi_1 = 0{,}50$	Média	0,5047	0,5050	0,5091	0,5098
	RB (%)	0,9310	1,0088	1,8264	1,9588
	MSE	0,0550	0,0444	0,0307	0,0219
$\phi_2 = -0{,}30$	Média	-0,3092	-0,3087	-0,3046	-0,3040
	RB (%)	3,0577	2,9145	1,5451	1,3472
	MSE	0,0138	0,0108	0,0065	0,0045
$\theta_1 = 0{,}40$	Média	0,4043	0,3996	0,3935	0,3915
	RB (%)	1,0755	-0,1055	-1,6127	-2,1335
	MSE	0,0713	0,0527	0,0338	0,0240
$\theta_2 = 0{,}15$	Média	0,1501	0,1495	0,1443	0,1441
	RB (%)	0,0561	-0,3275	-3,7714	-3,9148
	MSE	0,0448	0,0275	0,0151	0,0100
$\beta_1 = 8{,}40$	Média	8,4031	8,4022	8,4005	8,4006
	RB (%)	0,0371	0,0266	0,0062	0,0067
	MSE	0,0081	0,0052	0,0021	0,0012
$\beta_2 = -6{,}90$	Média	-6,9113	-6,9061	-6,9003	-6,9011
	RB (%)	0,1637	0,0886	0,0049	0,0163
	MSE	0,1621	0,1286	0,0412	0,0275

pois em seu cálculo atribui maior importância aos erros com valores absolutos maiores (Chai and Draxler, 2014). A expressão para sua obtenção é apresentada a seguir:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=m}^{n}(Y_t - \widehat{Y}_t)^2},$$

em que Y_t representa o valor observado da variável resposta no t-ésimo instante de tempo; $\widehat{Y}_t$ corresponde ao valor previsto pelo modelo para a variável resposta no t-ésimo instante de tempo; e n é o tamanho da amostra.

- **Erro Absoluto Médio (MAE)**: denotado por MAE (do inglês "*mean absolute error*"). Muitas vezes, esta métrica é preterida pelo RMSE, sendo recomendada principalmente para dados em que se espera uma distribuição uniforme dos erros. A expressão para sua obtenção é apresentada a seguir:

$$\text{MAE} = \frac{1}{n}\sum_{t=m}^{n}|Y_t - \widehat{Y}_t|.$$

- ***Symmetric Mean Absolute Percentage Error* (sMAPE)**: surge como uma correção da métrica MAPE (do inglês "*mean absolute percentage error*"), com características de invariância por escala e simetria, que são interessantes para a comparação de modelos (Fiorucci and Louzada, 2020). Esta métrica é bastante popular em competições de previsão, como, por exemplo, nas competições de Makridakis (Makridakis et al., 2018). É calculada pela expressão a seguir:

$$\text{sMAPE} = \frac{100}{n}\sum_{i=m}^{n}\frac{|Y_{n+h} - \widehat{Y}_{n+h}|}{(Y_{n+h} + \widehat{Y}_{n+h})/2},$$

em que n representa o tamanho da série; Y_{n+h} é o valor observado na série h passos à frente; e $\widehat{Y}_{n+h}$ é a previsão do modelo proposto h passos à frente.

2.4 Aplicação a dados reais

Nesta seção são apresentados e discutidos os principais resultados obtidos a partir da aplicação das técnicas estatísticas descritas na Seção 2.3 a um conjunto de dados reais com informações sobre valores máximos e mínimos da umidade relativa do ar diária, no deserto do Atacama, situado ao norte do Chile. O conjunto adotado neste estudo foi adquirido da *Dirección General De Aeronáutica Civil, Dirección Meteorológica de Chile - Servicios Climáticos*, que fornece informações relacionadas a fontes naturais no Chile. Além disso, as informações acerca da velocidade do vento e da radiação solar estão disponíveis em: `https://www.kaggle.com/datasets/dnstata/atacamahumidity`. A umidade relativa do ar expressa a razão entre a quantidade de água disponível no ar e a quantidade

máxima que poderia haver na mesma temperatura, para atingir um equilíbrio na pressão de vapor. Devido à sua formulação, a umidade relativa do ar está restrita ao intervalo unitário padrão.

Fonseca et al. (2021) destacaram fatores climáticos que se relacionam diretamente com a precipitação da água e, consequentemente, com a umidade relativa do ar. Aqui, serão considerados: i) o movimento do vento, que desloca massas de ar e, portanto, influencia em seu conteúdo de água; e ii) a radiação solar, que afeta a temperatura e, consequentemente, o equilíbrio na pressão de vapor. Vale ressaltar que a região em estudo é afetada pelo fenômeno da *Camanchaca*, que consiste na formação de bancos de nuvens que não produzem chuva, caracterizada por neblina densa junto ao deserto mais seco da Terra (Atacama). Será adotado o nível de significância estatística de 5% para as conclusões neste trabalho.

2.4.1 Análise descritiva

Para compor a série temporal da umidade relativa do ar no deserto do Atacama (Chile), foram coletados os valores máximos e mínimos diários, durante o período de 01/01/2019 a 01/07/2021. Portanto, a série é composta por 871 observações [1]. Inicialmente, foi realizada uma análise de caráter descritivo, com a finalidade de identificar possíveis inconsistências, pontos atípicos e, além disso, estudar padrões no comportamento dos dados coletados.

Na Tabela 2.4 são apresentadas algumas métricas descritivas para as séries de mínimos e máximos da umidade relativa do ar. Analisando todo o período, nota-se que, na série de mínimos, a média da umidade relativa do ar foi de 0,367, enquanto que, para a série de máximos, foi de 0,871. Observa-se uma menor amplitude na série de máximos (0,483), evidenciada pela distância entre o maior e o menor valores observados. No entanto, a série de mínimos apresenta maior oscilação, refletida pelo desvio-padrão.

Avaliando cada ano que compõe o período isoladamente, nota-se que, para a série de máximos, em média, os anos possuem comportamento similar, sendo observado em 2020 o maior registro (0,875). O menor valor registrado ocorreu em 2019, enquanto que em 2020 registrou-se o pico, com 0,492 e 0,974, respectivamente. É interessante observar que as oscilações na umidade, para a série de máximos, vêm reduzindo com o passar dos anos. Quando avaliada a série de mínimos, é observada uma similaridade com o comportamento da série de máximos; em média, os anos possuem valores próximos, sendo a menor média registrada em 2019 (0,362). O menor e o maior valores registrados ocorreram em 2019: 0,013 e 0,760, respectivamente. Nesta série, também existe redução nas oscilações da

[1] Vale ressaltar que, durante este período, 42 dias apresentaram problemas na coleta dos dados para as varáveis: umidade relativa do ar, movimento do vento e radiação solar.

umidade com o passar dos anos.

Tabela 2.4: Medidas descritivas da série temporal de umidade relativa do ar no deserto do Atacama, Chile, durante o período de 01/01/2019 a 30/06/2021. DP = desvio-padrão.

	Período	Mínimo	1° Quartil	Mediana	3° Quartil	Máximo	Média	DP
Máximo	2019	0,492	0,840	0,878	0,911	0,974	0,867	0,075
	2020	0,588	0,839	0,879	0,923	0,975	0,875	0,068
	2021	0,680	0,829	0,878	0,909	0,969	0,870	0,055
	2019-2021	0,492	0,838	0,878	0,917	0,975	0,871	0,069
Mínimo	2019	0,013	0,321	0,371	0,425	0,760	0,362	0,113
	2020	0,046	0,321	0,377	0,434	0,751	0,370	0,112
	2021	0,064	0,329	0,371	0,417	0,734	0,372	0,089
	2019-2021	0,013	0,322	0,373	0,428	0,760	0,367	0,108

A Tabela 2.5 informa o máximo e o mínimo mensais, registrados entre os meses de janeiro de 2019 e junho de 2021, para a série de umidade relativa do ar no deserto do Atacama, Chile. Nota-se que, durante o período de inverno (que, no Chile, ocorre de junho a setembro), acontece as maiores oscilações na umidade relativa do ar, isto é, ocorre os maiores e menores valores ao longo de todo o período.

Tabela 2.5: Máximo e mínimo mensais registrados durante o período de 01/01/2019 a 30/06/2021, para a série de umidade relativa do ar no deserto do Atacama, Chile.

	Máximo			Mínimo		
Mês	2019	2020	2021	2019	2020	2021
Janeiro	0,945	0,919	0,962	0,276	0,295	0,267
Fevereiro	0,925	0,959	0,936	0,290	0,313	0,266
Março	0,944	0,965	0,943	0,315	0,267	0,259
Abril	0,969	0,969	0,954	0,290	0,249	0,282
Maio	0,971	0,975	0,960	0,021	0,092	0,195
Junho	0,973	0,972	0,969	0,026	0,068	0,064
Julho	0,973	0,975	-	0,028	0,046	-
Agosto	0,974	0,972	-	0,013	0,105	-
Setembro	0,968	0,974	-	0,199	0,176	-
Outubro	0,970	0,960	-	0,190	0,173	-
Novembro	0,933	0,958	-	0,189	0,173	-
Dezembro	0,948	0,921	-	0,262	0,306	-

As Figuras 2.2 e 2.3 apresentam, respectivamente, as séries temporais dos valores máximos e mínimos, além de suas correspondentes funções de autocorrelação (ACF) e autocorrelação parcial (PACF, ou ainda, *Partial* ACF) amostrais. Observa-se que, na

série de máximos (Figura 2.2(A)), os valores oscilam ao redor de um valor central, sendo mais intenso entre os anos de 2019 e 2020. Além disso, o comportamento apresentado pelas funções de autocorrelação e autocorrelação parcial (Figura 2.2(B)-(C)) indica a existência de dependência temporal nos dados, e ainda, fornece indícios para o ajuste de um modelo com parâmetro autorregressivo de ordem 2, ou seja, um modelo ULARMA$(2, 0)$. É possível também identificar a presença de comportamento sazonal que não será considerado neste trabalhado. Quando analisada a série de mínimos (Figura 2.3(A)), nota-se um comportamento que possui oscilações ao redor de um valor central. As funções de autocorrelação e autocorrelação parcial (Figura 2.3(B)-(C)) sugerem o ajuste de um modelo com parâmetro autorregressivo de ordem 1, ou seja, o modelo ULARMA$(1, 0)$. Vale ressaltar que oscilações ao redor de um valor central, no contexto de séries temporais, são desejáveis, pois trata-se de um comportamento similar ao de um ruído branco. Tal característica indica a não violação do pressuposto de estacionariedade, que comumente está associado às técnicas de modelagem clássicas.

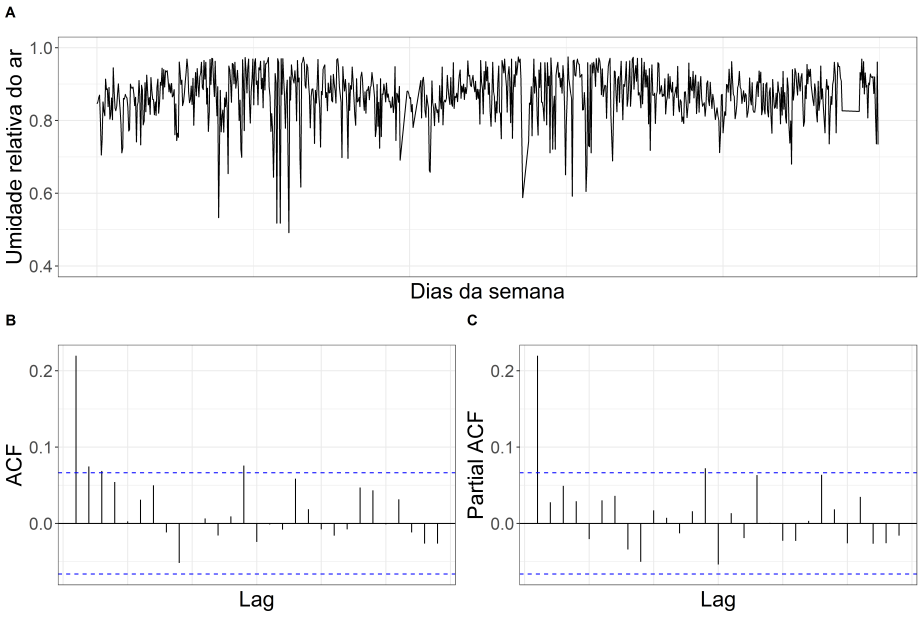


Figura 2.2: (A) Série temporal; (B) Gráfico de autocorrelação; e (C) Gráfico de autocorrelação parcial, dos valores máximos da umidade relativa do ar no deserto do Atacama, Chile.

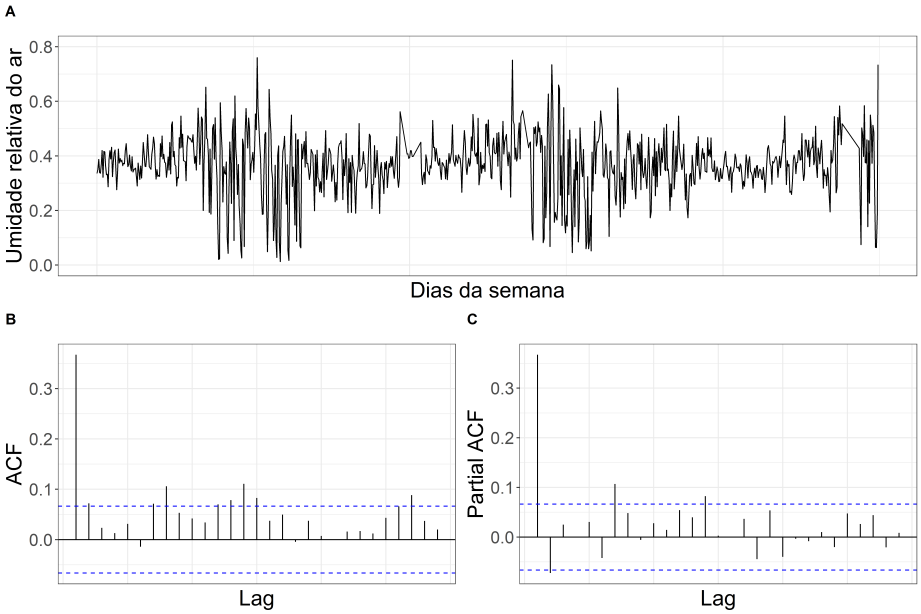


Figura 2.3: (A) Série temporal; (B) Gráfico de autocorrelação; e (C) Gráfico de autocorrelação parcial, dos valores mínimos da umidade relativa do ar no deserto do Atacama, Chile.

2.4.2 Modelagem de séries temporais

Para comparar a performance dos modelos, foi empregada a técnica de validação cruzada *out-of-time* (Maldonado et al., 2022). Nessa abordagem os dados coletados são divididos em dois conjuntos, denominados treino e teste, considerando a relevância das informações no tempo. No primeiro, o conjunto de treino, os dados são utilizados para estimar os parâmetros do modelo proposto, e após definido o modelo mais adequado, este será aplicado ao conjunto de teste para avaliar o seu desempenho na predição dos dados. Técnicas como esta têm ganhado espaço na literatura devido à sua importância no processo de construção, validação e generalização de modelos (Tantithamthavorn et al., 2017). Para avaliação da capacidade preditiva, foram utilizadas as métricas descritas na Seção 2.3.4, considerando diferentes desenhos na composição das amostras de treino e teste, aqui, intitulados como: curto prazo (previsão de 7 dias - 25/06/2021 a 01/07/2021), médio prazo (previsão de 15 dias - 17/06/2021 a 01/07/2021) e longo prazo (previsão de 30 dias - 14/05/2021 a 01/07/2021).

A Tabela 2.6 contém as medidas descritivas para as variáveis utilizadas como preditoras no modelo: velocidade do vento e radiação solar. Nota-se que, para a série de máximos, há maior flutuação dos valores nas duas variáveis, refletida pelo desvio-padrão. No entanto, para a série de mínimos, destaca-se a ocorrência de muitos valores iguais a zero (aproximadamente 75% dos dados observados); no contexto de análise de dados, este fenômeno é conhecido como inflação ou excesso de zeros. Na literatura, existem técnicas

específicas que abordam dados com essa característica. Este trabalho não irá se aprofundar neste assunto, no entanto, as variáveis serão mantidas com o intuito de investigar o seu impacto na performance do modelo preditivo. Em ambas as séries, foi observado que existe um comportamento assimétrico das variáveis, sendo que na série de máximos há assimetria negativa, ou seja, existe uma maior dispersão dos dados para valores abaixo da mediana e maior concentração para valores acima dela, enquanto que, para a série de mínimos, há assimetria positiva, exibindo um comportamento oposto ao descrito para a série de máximos. A correlação de *Spearman* entre as variáveis foi avaliada e não se mostrou relevante; portanto, elas foram testadas simultaneamente no modelo final.

Tabela 2.6: Medidas descritivas da velocidade do vento e da radiação solar, no deserto do Atacama, Chile, de 2019 a 2021.

	Máximo		Mínimo	
Métrica	Velocidade do vento	Radiação solar	Velocidade do vento	Radiação solar
Mínimo	3,400	0,000	0,000	0,000
1° Quartil	12,600	859,500	0,000	0,000
Mediana	13,800	1.069,800	0,000	0,000
3° Quartil	14,900	1.215,250	0,000	0,000
Máximo	20,700	1.605,500	4,500	300,900
Média	13,734	1.036,835	0,016	0,345
DP	1,800	228,591	0,216	10,196

Após a análise descritiva, foram investigados os termos autorregressivos e/ou de médias móveis que serão incorporados ao modelo. Utilizando como critérios o AIC e o BIC, definiu-se que: na série de máximos, existe o efeito do componente autorregressivo de ordem 1; e, na série de mínimos, existe o efeito dos componentes autorregressivo e de médias móveis, ambos de ordem 1. Para auxiliar na identificação e comparação dos termos temporais, utilizou-se a função *auto.arima(·)* do pacote `forecast` (Hyndman and Khandakar, 2008). Foram então ajustados os modelos ULARMA, KARMA e βARMA, considerando as diferentes composições de amostras de treino e teste, definidas pelo *hold-dout (out-of-time) simples*.

Na Figura 2.4 são apresentados os modelos de regressão ajustados às séries (de máximos nos painéis superiores e de mínimos nos painéis inferiores) no conjunto de treino, para configuração de curto prazo. Nota-se que o modelo ULARMA (Figura 2.4(A) e (D)) apresentou bom ajuste nas duas séries, assim como os modelos KARMA (Figura 2.4(B)) e βARMA (Figura 2.4(C) e (E)). O modelo KARMA, para a série de máximos, obteve melhor performance segundo as métricas MAE (0,048) e sMAPE (5,629), enquanto o modelo βARMA apresentou melhor performance sob o critério RMSE (0,067). Ao

considerar a série de mínimos, o modelo βARMA obteve melhor desempenho em todas as métricas consideradas. Os resultados encontrados para o cenário de curto prazo são similares aos observados nos demais cenários (médio e longo prazo) e, por isso, não serão abordados com mais detalhes. Não foi apresentado o ajuste do modelo KARMA para a série de mínimos devido a problemas de convergência.

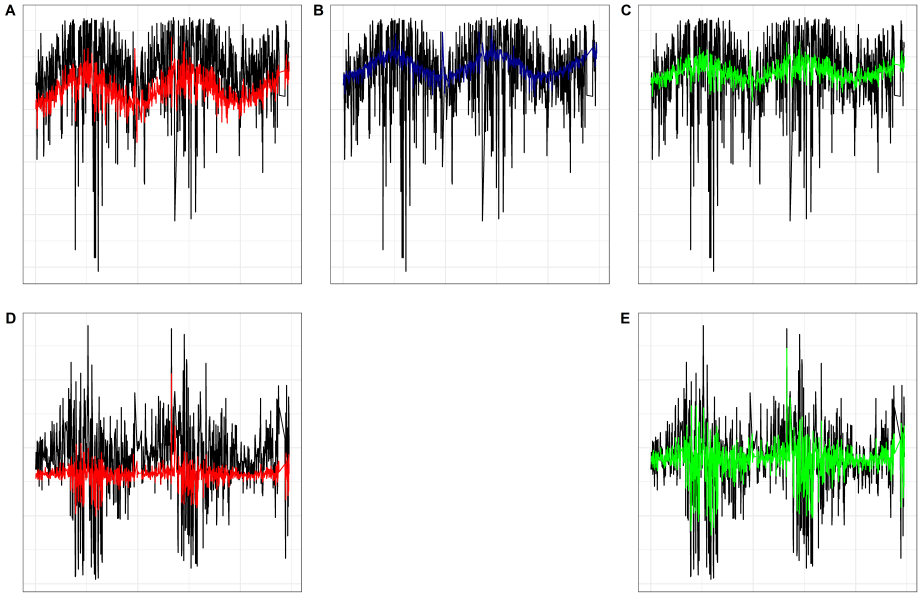


Figura 2.4: (A) e (D) Modelos ULARMA; (B) KARMA; e (C) e (E) βARMA, ajustados às séries de valores máximos (painéis superiores) e mínimos (painéis inferiores) da umidade relativa do ar, no deserto do Atacama, Chile, considerando a amostra de treino a curto prazo.

Na Tabela 2.7 são apresentados os resultados dos modelos propostos no conjunto de teste, considerando os diferentes cenários para previsão (curto, médio e longo prazo). Nota-se que o modelo ULARMA apresentou excelente desempenho no cenário de curto prazo na série de valores máximos, e também na série de valores mínimos, obtendo a melhor performance em todas as métricas consideradas. É evidente que o aumento da janela de previsões impacta na performance dos modelos, principalmente do ULARMA, que possui menor quantidade de parâmetros, quando comparado aos modelos βARMA e KARMA. No entanto, os resultados entre os modelos se mantêm próximos. Vale ressaltar que o modelo KARMA se destacou quando avaliado a longo prazo para a série de máximos; enquanto o βARMA se destacou no médio e longo prazo, quando avaliada a série de mínimos, em ambos os casos obtendo a melhor performance em todas as métricas consideradas.

Na Tabela 2.8 são apresentadas as estimativas dos parâmetros dos modelos ULARMA, βARMA e KARMA, ajustados às séries de valores máximos e mínimos, considerando o

Tabela 2.7: Métricas para avaliação de desempenho das previsões dos modelos ULARMA, βARMA e KARMA, ajustados às séries de valores máximos e mínimos.

Modelo	Máximo			Modelo	Mínimo		
	RMSE	MAE	sMAPE		RMSE	MAE	sMAPE
	7 dias				7 dias		
ULARMA(1,0)	**0,083**	**0,074**	**8,743**	ULARMA(1,1)	**0,235**	**0,207**	**71,131**
βARMA(1,0)	0,090	0,074	8,758	βARMA(1,1)	0,238	0,216	71,950
KARMA(1,0)	0,101	0,078	9,229	KARMA(1,1)	-	-	-
	15 dias				15 dias		
ULARMA(1,0)	0,068	0,060	6,999	ULARMA(1,1)	0,195	0,170	54,731
βARMA(1,0)	**0,066**	0,050	5,870	βARMA(1,1)	**0,189**	**0,161**	**51,060**
KARMA(1,0)	0,073	**0,048**	**5,649**	KARMA(1,1)	-	-	-
	30 dias				30 dias		
ULARMA(1,0)	0,073	0,066	7,560	ULARMA(1,1)	0,187	0,166	49,631
βARMA(1,0)	0,061	0,050	5,690	βARMA(1,1)	**0,170**	**0,144**	**42,686**
KARMA(1,0)	**0,059**	**0,041**	**4,695**	KARMA(1,1)	-	-	-

cenário de curto prazo. Quando avaliada a série de valores máximos, nota-se que apenas no modelo ULARMA todos os parâmetros foram estatisticamente significativos, enquanto que nos demais modelos a velocidade do vento perde relevância. Ao analisar a série de valores mínimos, é observado que as estimativas para os parâmetros do modelo não são estatisticamente significativas, para o modelo ULARMA, e nos demais (modelo βARMA) são significativas as estimativas para o intercepto, médias móveis e, também, a velocidade do vento e a precisão. Vale ressaltar que o critério de escolha dos modelos baseou-se em seu desempenho preditivo e, portanto, não foram removidas as variáveis sem relevância estatística, desde que melhorassem o desempenho do modelo ao tecer previsões.

2.5 Conclusões

Neste capítulo foi apresentada a construção de um modelo estatístico inédito para variáveis aleatórias contínuas, com domínio restrito ao intervalo unitário padrão $(0, 1)$, considerando uma estrutura de dependência temporal, isto é, para dados unitários que são observados ao longo do tempo. Mais especificamente, foi mostrada a formulação teórica, propriedades, estudo de simulação e, ainda, uma aplicação a um conjunto de dados reais sobre umidade relativa do ar diária no deserto do Atacama, Chile.

O modelo proposto foi desenvolvido sob a suposição de que a distribuição condicional da variável de interesse, dado o seu histórico, é a *unit-Lindley*, assim como introduzido por Benjamin et al. (2003). Denominado de *unit-Lindley* autorregressivo e de

Tabela 2.8: Estimativas dos parâmetros dos modelos ajustados à base de treino, no cenário de curto prazo. EP = erro padrão.

	Máximo			Mínimo		
ULARMA	Estimativa	EP	p-valor	Estimativa	EP	p-valor
Intercepto	1,8913	0,0003	<0,0001	-0,9520	1,0000	0,3411
ϕ_1	0,2078	0,0162	<0,0001	-0,2140	1,0000	0,8305
θ_1	-	-	-	0,4286	1,0000	0,6682
Radiação solar	-0,0006	0,0001	<0,0001	-0,0006	1,0000	0,9995
Velocidade do vento	-0,0141	0,0047	0,0027	0,2328	1,0000	0,8159
	Máximo			Mínimo		
βARMA	Estimativa	EP	p-valor	Estimativa	EP	p-valor
Intercepto	1,8729	0,0822	<0,0001	-0,5125	0,0441	<0,0001
ϕ_1	0,1797	0,0155	<0,0001	0,0690	0,0678	0,3091
θ_1	-	-	-	0,3203	0,0681	<0,0001
Radiação solar	-0,0004	0,0001	<0,0001	-0,0006	0,0017	0,7183
Velocidade do vento	0,0000	0,0064	0,9956	0,2342	0,0922	0,0111
Precisão	26,4517	1,2753	<0,0001	18,4637	0,8677	<0,0001
	Máximo			Mínimo		
KARMA	Estimativa	EP	p-valor	Estimativa	EP	p-valor
Intercepto	2,5412	0,1622	<0,0001	-	-	-
ϕ_1	0,0765	0,0284	0,0071	-	-	-
θ_1	-	-	-	-	-	-
Radiação solar	-0,0008	0,0001	<0,0001	-	-	-
Velocidade do vento	0,0101	0,0116	0,3838	-	-	-
Precisão	15,4971	0,5196	<0,0001	-	-	-

médias móveis, e denotado pela sigla ULARMA, este modelo se mostrou uma alternativa interessante para a modelagem de dados que assumem valores no intervalo $(0, 1)$, como taxas e proporções.

Na literatura existem outros modelos que são comumente utilizados para descrever dados que possuem essas características/restrições, os quais são baseados em distribuições mais conhecidas, como: βARMA (baseado na distribuição *Beta*) e KARMA (baseado na distribuição *Kumaraswamy*). Vale ressaltar que, apesar de ser uma distribuição de probabilidade recente, a *unit-Lindley* dispõe de propriedades que são interessantes, quando comparada às demais distribuições. Do ponto de vista teórico, possui forma fechada para a FDA e função quantil, expressões simples para a obtenção dos momentos, e pertence à família exponencial. Do ponto de vista prático, sua principal vantagem reside em ser uma distribuição recente, isto é, que ainda é pouco explorada na literatura, além de ser unimodal, uniparamétrica e bastante flexível.

O estudo de simulação realizado revelou uma performance satisfatória dos CMLEs para os parâmetros do modelo, mesmo ao considerar tamanhos de amostra pequenos (aqui, avaliou-se a partir de 70 observações). Com o aumento do tamanho da amostra, as estimativas tornavam-se mais precisas; e as estimativas do componente autorregressivo se mostravam mais precisas quando comparado ao componente de médias móveis. Na aplica-

ção, notou-se que o modelo proposto possui performance parecida com a das abordagens tradicionais (βARMA e KARMA), na maioria dos cenários analisados. Pois, foram consideradas diferentes composições para as amostras de treino e teste (*out-of-time*) e, em um cenário específico, ele apresentou desempenho superior ao dos demais modelos, com base nas métricas de performance preditiva.

Como sugestão de trabalhos futuros, seria interessante: (i) propor um novo modelo espaço-temporal para variáveis aleatórias contínuas que assumem valores no intervalo unitário padrão, com base na distribuição *unit-Lindley*; (ii) propor uma extensão que seja capaz de modelar dados sujeitos a flutuações sazonais; (iii) considerar a aplicação do modelo ULARMA a outros conjuntos de dados reais e avaliar a sua performance quando comparado às abordagens tradicionais.

Capítulo 3

Gráfico de controle para dados unitários autocorrelacionados

Neste capítulo são apresentadas as técnicas do CEP, por meio de sua ferramenta mais popular, o gráfico (ou carta) de controle, aplicado a variáveis em que há uma estrutura de dependência entre os dados coletados. Mais especificamente, é proposto um gráfico de controle para monitoramento do modelo ULARMA, considerando diferentes tipos de resíduos. Além disso, é ilustrada a aplicação do gráfico proposto a um conjunto de dados reais sobre umidade relativa do ar no deserto do Atacama, Chile.

3.1 Revisão de literatura

Esta seção contém uma breve revisão de literatura a respeito da construção e desenvolvimento de gráficos de controle para variáveis aleatórias cujo domínio está contido no intervalo unitário $(0, 1)$, considerando dados independentes e dados correlacionados, isto é, que apresentam estrutura de dependência entre si.

Shewhart (1931) desenvolveu um conjunto de técnicas com o intuito de melhorar a qualidade, aumentar a produtividade e ampliar o mercado consumidor de itens manufatureiros e industriais, na época conhecidas como técnicas ou ferramentas de controle da qualidade. Dentre elas, destacam-se os gráficos de controle que, em homenagem a seu criador, também eram chamados de gráficos de *Shewhart*. Posteriormente, as técnicas de controle da qualidade viriam a integrar a área do Controle Estatístico de Processos (CEP), que possui aplicação em diversas áreas do conhecimento, sendo objeto de estudo até os dias atuais.

Mandel (1969) propôs uma nova abordagem na construção de gráficos de controle, capaz de considerar a influência de um fator sobre o processo de interesse, denominado *gráfico de controle de regressão*. Posteriormente, Haworth (1996) estendeu essa proposta

considerando a existência de múltiplos fatores independentes entre si (ou seja, um modelo de regressão múltipla). Sob essa metodologia, a variável a ser monitorada são as estimativas (ou previsões) do modelo proposto para a variável de interesse. Logo depois, essa proposta foi aprimorada para o monitoramento dos resíduos do ajuste, prática comumente utilizada no contexto de séries temporais.

Sant'Anna and ten Caten (2012) propuseram o gráfico de controle baseado na distribuição *Beta*, para o monitoramento de taxas e proporções. Foi discutido que o gráfico de controle *Beta* torna-se mais adequado para tais dados quando estes não são gerados por um processo de *Bernoulli*, uma vez que variáveis como essas geralmente exibem assimetria. Além disso, a distribuição *Beta* é definida no intervalo $(0, 1)$, o que garante que os limites de controle sempre pertencerão a este intervalo.

Tondolo et al. (2016) fizeram uma revisão de literatura abordando os principais gráficos de controle desenvolvidos para o monitoramento de taxas e proporções, considerando dados que possuem estrutura de dependência entre si. Investigaram o desempenho do gráfico de controle para o modelo βARMA, considerando quatro tipos de resíduos: ordinário padronizado, ordinário padronizado na escala do preditor, ponderado padronizado e *deviance*. Além disso, o gráfico proposto foi ilustrado em duas aplicações a dados reais, em que foram monitorados o volume de energia armazenada na Região Sul do Brasil e os níveis dos mananciais do Sistema Cantareira (São Paulo). Foi ressaltado o bom desempenho dos gráficos, apresentando melhor poder de detecção de causas especiais, quando comparado às alternativas usuais.

Ho et al. (2018) discutiram a construção dos gráficos de controle para situações em que a variável a ser monitorada está contida no intervalo unitário, sendo gerada por processos que não seguem uma distribuição de *Bernoulli* (como taxas e proporções). É ressaltado que, neste cenário, comumente faz-se o uso da distribuição *Beta*, porém existem outras distribuições alternativas que podem ser consideradas. O escopo desse trabalho consistiu em estudar o impacto na velocidade das mudanças de sinal na proporção média em termos de métricas de performance usuais, como ARL (do inglês "*average run length*"), MRL (do inglês "*median run length*") e SDRL (do inglês "*standard deviation of the run length*"), quando os limites de controle da distribuição *Beta* são inapropriados, visto que as taxas e proporções são provenientes de outras distribuições, como *Simplex* ou *unit-Gama*. Foi evidenciado que a utilização de limites de controle equivocados provoca uma grande variedade de impactos, como antecipação/adiamento de alarmes falsos ou mesmo ausência de alarmes falsos, principalmente no caso de pequenos deslocamentos.

de Araujo Lima-Filho et al. (2019) propuseram um gráfico de controle para processos de limite duplo no intervalo unitário e que frequentemente contêm zeros e/ou uns em situações práticas, capaz de acomodar situações dos tipos $(0, 1]$, $[0, 1)$, $[0, 1]$. O gráfico de

controle proposto baseia-se em uma extensão da distribuição *Beta*, chamada distribuição *inflated Beta* (Ospina and Ferrari, 2010), para determinar os limites de controle. Foi realizado um estudo de simulação comparando o desempenho do gráfico proposto ao do gráfico de controle *Beta* (Sant'Anna and ten Caten, 2012), com base nas métricas ARL, MRL, SDRL e PRL (do inglês "*percentile run length*"). Observou-se uma melhor performance do gráfico proposto em todos os cenários considerados, para todas as métricas utilizadas ($PRL_{0,5\%}$, MRL_0, ARL_0 e $PRL_{0,95\%}$). Além disso, foram destacadas pelos autores as seguintes vantagens do método proposto: apresentou bons resultados mesmo quando os parâmetros foram considerados desconhecidos; o desempenho do gráfico de controle não diminuiu com o aumento da proporção de zeros e/ou uns; e mostrou-se útil em situações práticas.

Fonseca et al. (2021) apresentaram uma proposta de gráfico de controle para dados contidos no intervalo unitário, desenvolvida com base na distribuição *unit-Lindley* (Mazucheli et al., 2019). O desempenho da ferramenta proposta foi avaliado por meio de um extensivo estudo de simulação de Monte Carlo, utilizando métricas como ARL, MRL e SDRL. Além disso, foi feita e discutida uma aplicação a um conjunto de dados reais sobre partículas de água (umidade relativa do ar) na bacia hidrográfica de Copiapó e região do Atacama no Chile.

Lima-Filho and Bayer (2021) propuseram um novo gráfico de controle que considera a modelagem dos dados segundo a distribuição *Kumaraswamy*, bastante útil na modelagem de dados ecológicos restritos a intervalos duplamente limitados. O desempenho do gráfico de controle proposto foi avaliado e comparado com o do gráfico de controle *Beta* (Sant'Anna and ten Caten, 2012), por meio de um extensivo estudo de simulação de Monte Carlo, que evidenciou a superioridade do gráfico de controle *Kumaraswamy* em termos de ARL, MRL e SDRL. Além disso, foi considerada uma aplicação a dados reais referentes à umidade relativa do ar em Des Moines (Iowa, EUA), em que o gráfico *Kumaraswamy* foi capaz de detectar um ponto fora de controle no ano de ocorrência de um El Niño, enquanto que no gráfico *Beta* não houve detecção. Os autores sugeriram considerar o gráfico desenvolvido quando o interesse residir em monitorar fenômenos naturais, como hidrologia e dados ambientais.

3.2 Gráfico de controle ULARMA

Após a Segunda Guerra Mundial houve, entre as principais potências da época, a necessidade de ampliar mercados consumidores, manter clientes satisfeitos e reduzir o número de reclamações. A identificação de características que afetavam diretamente a performance/qualidade final do produto se tornou essencial para a redução de custos

(Louzada et al., 2013). Deste modo, aumentou-se a preocupação com o monitoramento de características que alteram a qualidade do produto, dando ênfase às técnicas de CEP, também chamado de controle da qualidade. O termo "qualidade", no âmbito do CEP, significa uma ou mais características que permitem distinguir um artigo do outro (Shewhart, 1929). O CEP consiste em um conjunto de métodos comprovados que visam monitorar, através da comparação contínua dos resultados de um processo, o comportamento de uma ou mais características, a fim de identificar tendências para variações significativas, permitindo distinguir as fontes de variabilidade, com o objetivo de controlar, aprimorar e, assim, obter produtos com um maior grau de qualidade (Montgomery, 2020). Conforme discutido em Mukherjee (2016), existem duas principais fontes de variabilidade em uma linha de processo produtivo: a variabilidade inerente ao processo e a variabilidade ocasionada por variáveis específicas. No primeiro caso, a variabilidade é natural e incontrolável, e surge de forma aleatória pelo fenômeno gerador do processo; neste caso, diz-se que o processo opera sob controle estatístico. No segundo caso, a variabilidade é afetada por um conjunto de variáveis específicas e diz-se que o processo opera fora de controle estatístico. É neste cenário que as técnicas de CEP são amplamente utilizadas para detecção da não conformidade do processo (Montgomery, 2020).

Originalmente, as técnicas de CEP, em especial os gráficos de controle, que é a ferramenta de maior destaque devido à sua facilidade de implementação e interpretação (Montgomery, 2020), foram desenvolvidas para aplicação em áreas manufatureiras e industriais (Shewhart, 1929). Atualmente, essa ferramenta é utilizada em diversas áreas, desde serviços a controle de doenças (Sena et al., 2022; Braz et al., 2006). Mandel (1969) introduziu uma nova abordagem para aplicação dos gráficos de controle, capaz de considerar a influência de múltiplos fatores, independentes entre si, sobre a característica de interesse, denominado *gráfico de controle de regressão*. Aqui, a ideia principal consiste em propor um modelo em que a característica de qualidade é a variável dependente e os fatores são as covariáveis, e utilizar as estimativas do modelo proposto como variável a ser monitorada. Os gráficos de controle tradicionais, também conhecidos como gráficos de *Shewhart*, baseiam-se no pressuposto de independência dos dados coletados, que na prática é violado com frequência, pois efeitos temporais são substanciais em alguns processos (Alwan, 1992). Montgomery (2020) estendeu essa abordagem ao considerar a presença de correlação nos dados coletados do processo, propondo o uso dos resíduos como variável a ser monitorada. Conforme apresentado em Sena et al. (2022), esse procedimento tem algumas vantagens: (i) os resíduos são não correlacionados; (ii) os limites de controle obtidos não apresentam variação, isto é, são constantes; (iii) possui fácil interpretação; e (iv) auxilia na visualização do comportamento da série apresentada.

Neste trabalho é considerada a abordagem proposta por Montgomery (2020), em

que os resíduos de um modelo ULARMA é a variável a ser monitorada. Neste cenário em particular, a linha central é igual a zero, pois é o valor esperado para os resíduos de um modelo de regressão em que não há violações das suposições associadas, e os limites de controle são calculados com base na distribuição *Normal*. As expressões para o limite superior de controle (LSC), linha central (LC) e limite inferior de controle (LIC) são:

$$\text{LSC} = +w\sigma_r\,, \quad \text{LC} = 0 \quad \text{e} \quad \text{LIC} = -w\sigma_r,$$

em que w representa a amplitude entre os limites de controle e a linha central, e σ_r é o desvio-padrão dos resíduos.

3.2.1 Resíduos

Segundo Espinheira et al. (2008), os resíduos são métricas que indicam a similaridade entre os dados coletados e o modelo ajustado. Investigar os resíduos de um modelo é uma etapa essencial para inferir sobre a qualidade do ajuste, pois permite identificar possíveis violações de pressupostos associados ou a presença de observações que diferem do padrão (Tondolo et al., 2016). Aqui, foram considerados quatro tipos de resíduos - ordinário, *Pearson*, *deviance* e quantílico aleatorizado - para avaliar a performance do gráfico de controle proposto.

- **Resíduo ordinário**: é o tipo de resíduo mais simples em processos de modelagem, definido pela diferença entre o valor observado e o valor ajustado pelo modelo proposto. Logo, pode-se expressá-lo por:

$$r_t = y_t - \widehat{\mu}_t.$$

- **Resíduo de *Pearson***: ou ainda, resíduo de *Pearson* padronizado, consiste no resíduo ordinário dimensionado pelo desvio-padrão estimado de Y (McCullagh, 2019). Pode-se obtê-lo pela seguinte expressão:

$$r_t = \frac{y_t - \widehat{\mu}_t}{\sqrt{\widehat{\mathbb{V}\text{ar}}\left[Y_t|\Im_{t-1}\right]}}.$$

- **Resíduo *deviance***: é usado como uma medida de discrepância de um MLG, então cada unidade contribui com uma quantidade para essa medida (McCullagh, 2019). A expressão para a obtenção deste resíduo é dada por:

$$r_t = \text{sign}(y_t - \widehat{\mu}_t)\{2[\ell_t(y_t, \widehat{\boldsymbol{\varphi}}) - \ell_t(\widehat{\mu}_t, \widehat{\boldsymbol{\varphi}})]\},$$

em que: $\text{sign}(b)$ representa a função sinal, definida por -1 se $b < 0$, 0 se $b = 0$ e $+1$ se $b > 0$; $\ell_t(y_t, \widehat{\boldsymbol{\varphi}})$ e $\ell_t(\widehat{\mu}_t, \widehat{\boldsymbol{\varphi}})$ representam, respectivamente, as contribuições para a função de log-verossimilhança das observações y_t e das estimativas $\widehat{\mu}_t$.

- **Resíduo quantílico**: proposto por Dunn and Smyth (1996), possui ampla utilização em modelos aditivos generalizados para localização, escala e forma (GAMLSS, do inglês "*Generalized Additive Models for Location, Scale and Shape*"). Sua expressão é dada por:

$$r_t = \Phi^{-1}\left(F(y_t, \widehat{\Theta})\right),$$

em que $\Phi^{-1}(\cdot)$ é a inversa da FDA da distribuição *Normal* padrão, $F(\cdot)$ é a FDA da distribuição ULARMA e $\widehat{\Theta}$ são as estimativas de máxima verossimilhança condicional de Θ.

3.2.2 Critérios para seleção de resíduos

Como forma de avaliar o desempenho estatístico do gráfico de controle proposto para o modelo ULARMA, serão considerados como critérios para seleção de resíduos as seguintes métricas: comprimento médio da sequência (ARL), comprimento mediano da sequência (MRL) e desvio-padrão do comprimento da sequência (SDRL).

- **Comprimento médio da sequência (ARL)**: definido pelo número médio de amostras até que uma causa especial seja detectada, é amplamente utilizado para avaliar a qualidade de gráficos de controle (Montgomery, 2020). Quando avaliado em um processo que esteja sob controle estatístico, representa o número de amostras observadas até a ocorrência de um alarme falso (denotado por ARL_0), enquanto que, em um processo que esteja fora de controle estatístico, representa o número de amostras observadas até que um alarme verdadeiro seja detectado (denotado por ARL_1). Em situações casuais, é desejável que o ARL seja grande quando não há interferência de causas específicas no processo, indicando a baixa frequência na ocorrência de alarmes falsos, e seja pequeno quando houve tal interferência, demonstrando que o gráfico de controle possui desempenho eficaz na detecção de mudanças no processo (Chen et al., 2017). As expressões para o cálculo do ARL_0 e ARL_1 são definidas por:

$$\text{ARL}_0 = \frac{1}{\alpha}, \quad \text{ARL}_1 = \frac{1}{(1-\beta)},$$

em que $\alpha = \mathbb{P}\left(Y_t \notin [\text{LIC}, \text{LSC}] \mid \mu_t = \mu_w\right)$ é a probabilidade de alarme falso, $\beta = \mathbb{P}\left(Y_t \in [\text{LIC}, \text{LSC}] \mid \mu_t = \mu' = \mu_w + \delta\right)$, $1-\beta$ é a probabilidade de alarme verdadeiro, e δ representa uma mudança na média do processo.

- **Comprimento mediano da sequência (MRL)**: considerado uma das medidas mais confiáveis para avaliar o desempenho de gráficos de controle, pois é menos afetada pela assimetria da distribuição (Fonseca et al., 2021). É denotado por

MRL_0, quando avaliado em um processo sob controle estatístico, e MRL_1, em um processo fora de controle estatístico. As expressões para o cálculo do MRL_0 e MRL_1 são definidas por:

$$\text{MRL}_0 = \frac{\log(0{,}5)}{\log(1-\alpha)}, \quad \text{MRL}_1 = \frac{\log(0{,}5)}{\log(\beta)}.$$

- **Desvio-padrão do comprimento da sequência (SDRL)**: medida utilizada para avaliar a dispersão da distribuição do comprimento da sequência (ou corrida) (Fonseca et al., 2021). É denotado por SDRL_0, quando avaliado em um processo sob controle estatístico, e SDRL_1, em um processo fora de controle estatístico. As expressões para o cálculo do SDRL_0 e SDRL_1 são definidas por:

$$\text{SDRL}_0 = \sqrt{\frac{1-\alpha}{\alpha^2}}, \quad \text{SDRL}_1 = \sqrt{\frac{\beta}{(1-\beta)^2}}.$$

A literatura propõe a utilização de diferentes valores para α, que, por sua vez, determinam valores nominais (ou teóricos) associados às métricas citadas anteriormente. Na Tabela 3.1 são apresentados os valores de α mais utilizados (Pereira et al., 2023).

Tabela 3.1: Valores nominais de ARL_0, MRL_0 e SDRL_0, para diferentes valores de α.

α	ARL_0	MRL_0	SDRL_0
0,01	100,0	69,0	99,5
0,005	200,0	138,3	199,5
0,0027	370,0	256,1	369,5

3.3 Estudo de simulação

Nesta seção são apresentados a descrição e os resultados de um estudo de simulação de Monte Carlo com o objetivo de avaliar a performance dos diferentes tipos de resíduos (ver Seção 3.2.1) no gráfico de controle proposto para o modelo ULARMA.

O procedimento de avaliação numérica foi construído visando estudar o comportamento do processo sob controle estatístico, aqui denominado de cenário 0, em que $g(\mu_t) = \beta_0 + \phi_1 g(y_{t-1}) + \theta_1 \epsilon_{t-1}$, e também fora de controle estatístico, denominado de cenário 1, em que $g(\mu_t) = \beta_0 + \delta + \phi_1 g(y_{t-1}) + \theta_1 \epsilon_{t-1}$. O parâmetro $\delta = \{\pm 1\%, \pm 10\%, \pm 20\%, \pm 30\%\}$ representa uma oscilação adicionada à média do processo, que permite avaliar a detecção via ARL_1; note que, quando $\delta = 0$, o processo está sob controle estatístico. Na primeira etapa foram consideradas as fases I e II durante a aplicação dos gráficos de controle. A fase I é caracterizada pela análise retrospectiva, em que se analisa o passado do processo

até o instante de interesse, estimando os limites de controle, para determinar se o processo estava sob controle estatístico durante o período de tempo em que os dados foram coletados. Por sua vez, a fase II é caracterizada pela análise prospectiva, em que o gráfico de controle é utilizado para monitorar as futuras observações do processo, tendo como base os limites estabelecidos na fase I.

Ainda na primeira etapa, foi necessária a calibração dos limites de controle em termos de um ARL_0 nominal de 100 e 200. Neste procedimento, foram consideradas 2.000 réplicas de Monte Carlo em amostras de tamanho $n = 871$ na fase I e $n = 5.000$ na fase II, para uma variável restrita ao intervalo unitário padrão de um modelo ULARMA$(1, 0)$ e ULARMA$(1, 1)$, com função de ligação *logito* e parâmetros apresentados na Tabela 2.8. Foram considerados 16 valores entre 2 e 5 para a amplitude entre os limites de controle (LIC e LSC) e a linha central, e em cada valor foi computado o ARL. Então, dois métodos foram considerados para determinar o valor de w que minimiza a diferença entre o resultado obtido e o valor nominal estabelecido: a regressão linear (Tondolo et al., 2016) e a interpolação linear (Moraes et al., 2014). Posteriormente, 10.000 réplicas de Monte Carlo foram simuladas considerando o valor de w obtido por cada método, e a interpolação linear apresentou melhores resultados. Na segunda etapa, o valor de w resultante do procedimento de calibração foi utilizado para verificar a capacidade de identificar a não conformidade do processo, depois de propor incrementos na média do mesmo. Para a série de máximos foi considerada a presença de duas covariáveis, com distribuição similar à da velocidade do vento e da radiação solar, enquanto que, para a série de mínimos, não foram consideradas covariáveis, pois como observado anteriormente, existe o fenômeno de inflação de zeros para as covariáveis (ver Tabela 2.6). Os parâmetros utilizados durante o processo de simulação são apresentados na Tabela 2.8.

3.3.1 Resultados

Os valores de w que minimizaram a diferença entre o ARL_0 simulado/empírico e o nominal, obtidos pelo procedimento de calibração (utilizando o método da interpolação linear) no cenário 0, para as séries de máximos e mínimos, são apresentados na Tabela 3.2. Nota-se que o método escolhido/aplicado possui alta precisão nos resultados obtidos, visto que a distância máxima entre o ARL_0 simulado e o nominal é de 26,4089. Além disso, é observado que os resíduos *deviance* e quantílico necessitam de maior amplitude, quando comparados aos demais resíduos, para atingir o valor nominal proposto considerando a série de mínimos, enquanto que, na série de máximos, este comportamento ocorre nos resíduos ordinário e padronizado. Os valores de w obtidos para o resíduo quantílico possuem valores próximos, com diferença após a quinta casa decimal.

A Figura 3.1 apresenta o desempenho dos gráficos de controle dos diferentes tipos

Tabela 3.2: Valores de w obtidos após o procedimento de calibração, para as séries de máximos e mínimos.

	Máximos		Mínimos	
Resíduo	$ARL_0 = 100$	$ARL_0 = 200$	$ARL_0 = 100$	$ARL_0 = 200$
Ordinário	4,0000 (-16,3000)	4,7030 (-5,0966)	2,1114 (26,4089)	2,3057 (2,2765)
Padronizado	4,0000 (-14,3869)	4,7030 (-16,5055)	2,1587 (16,3106)	2,3733 (1,3362)
Deviance	3,1953 (3,8657)	3,5000 (25,0944)	2,7560 (3,4962)	3,0027 (0,0843)
Quantílico	2,5425 (-7,0331)	2,7828 (-4,8936)	2,5425 (6,5977)	2,7828 (4,1466)

Nota: Entre parênteses é apresentada a distância entre o ARL_0 simulado e o nominal.

de resíduos, obtidos do modelo ULARMA$(1, 1)$. Neste cenário simulado, a média do processo é de aproximadamente 0,367 (isto é, $\mu \approx 0{,}367$). Para $ARL_0 = 100$ (Figura 3.1(A)), apenas o resíduo quantílico detém bom poder de detecção durante todo o intervalo de variação na média do processo. Quando analisados isoladamente os valores de δ, tem-se que: para $\delta < 0$, o resíduo de *Pearson* possui bom poder de detecção; e, para valores de $\delta > 0$, o resíduo *deviance* começa a apresentar bom desempenho. Nesses casos, os resíduos de *Pearson* e *deviance* possuem desempenho superior ao do resíduo quantílico, pois identificam mais rapidamente a não conformidade no processo.

Para $ARL_0 = 200$ (Figura 3.1(B)), mantém-se apenas o resíduo quantílico com bom poder de detecção durante todo o intervalo de variação na média do processo. Quando analisados isoladamente os valores de δ, tem-se que: para $\delta < 0$, o resíduo de *Pearson* também apresenta bom poder de detecção; e, para valores de $\delta > 0$, o resíduo ordinário também apresenta bom desempenho. Nesses casos, os resíduos de *Pearson* e ordinário possuem desempenho superior, quando comparados ao resíduo quantílico, pois identificam mais rapidamente a não conformidade no processo. As Figuras 3.1(C) e (D) destacam o comportamento do resíduo quantílico, considerando ARL_0 de 100 e 200, respectivamente. Observa-se que, em todo o intervalo, o ARL_1 é impactado pelos incrementos propostos, mantendo-se sempre inferior ao ARL_0. Vale ressaltar que, nas Figuras 3.1(C) e (D), existe baixa variação nos valores de ARL e, portanto, estão sendo apresentadas numa escala que não começa no valor zero, a fim de facilitar a visualização no comportamento do resíduo (a saber, o eixo y começa em: 93,0125 para $ARL_0 = 100$; e 194,0195 para $ARL_0 = 200$).

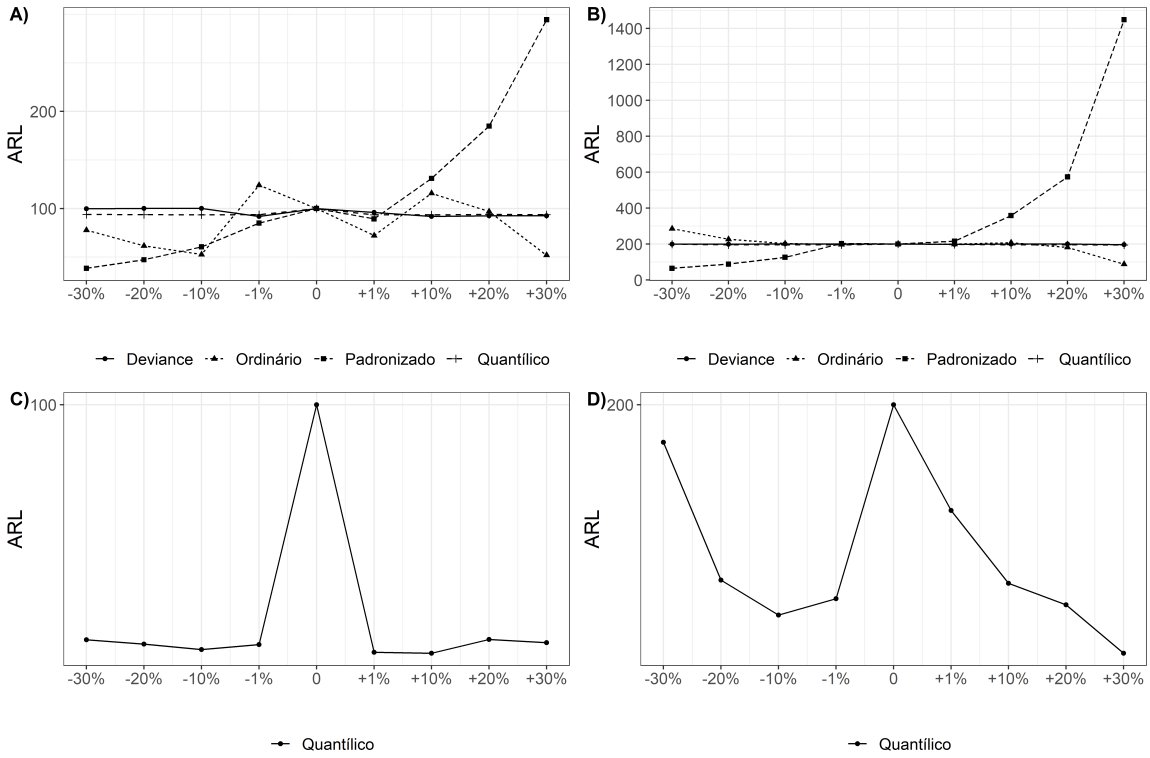


Figura 3.1: Desempenho dos gráficos de controle dos resíduos do modelo ULARMA$(1,1)$, considerando: (A) e (C) $\text{ARL}_0 = 100$; e (B) e (D) $\text{ARL}_0 = 200$.

A Figura 3.2 apresenta o desempenho dos gráficos de controle aplicados aos diferentes resíduos do modelo ULARMA$(1,0)$. Neste cenário simulado, a média do processo é de aproximadamente 0,871 (isto é, $\mu \approx 0{,}871$), e além disso, existe a presença de covariáveis incorporadas ao modelo. Para $\text{ARL}_0 = 100$ (Figura 3.2(A)), apenas o resíduo quantílico detém bom poder de detecção durante todo o intervalo de variação na média do processo. Quando analisados isoladamente os valores de δ, tem-se que: para $\delta < 0$, o resíduo ordinário possui bom poder de detecção; e, para valores de $\delta > 0$, o resíduo de *Pearson* começa a apresentar bom desempenho. Nesses casos, os resíduos ordinário e de *Pearson* possuem desempenho superior ao do resíduo quantílico, pois identificam mais rapidamente a não conformidade no processo.

Considerando agora $\text{ARL}_0 = 200$ (Figura 3.2(B)), nenhum resíduo apresenta bom poder de detecção durante todo o intervalo de variação. Quando analisados isoladamente os valores de δ, tem-se que: para $\delta < 0$, o resíduo ordinário possui bom desempenho; e, para valores de $\delta > 0$, os resíduos de *Pearson* e quantílico também apresentam bom desempenho. As Figuras 3.1(C) e (D) apresentam o comportamento do resíduo quantílico; nota-se, em (C), o bom poder de detecção para $\text{ARL}_0 = 100$, com o ARL_1 sendo impactado pelas alterações na média do processo. No entanto, em (D), observa-se um comportamento atípico e incorreto quando há alteração de $\delta = -30\%$ na média do processo, em que $\text{ARL}_1 > \text{ARL}_0$. Vale ressaltar que, nas Figuras 3.2(C) e (D), existe baixa variação nos

valores de ARL e, portanto, estão sendo apresentadas numa escala que não começa no valor 0, a fim de facilitar a visualização no comportamento do resíduo (a saber, o eixo y começa em: 92,0261 para $\text{ARL}_0 = 100$; e 192,0700 para $\text{ARL}_0 = 200$).

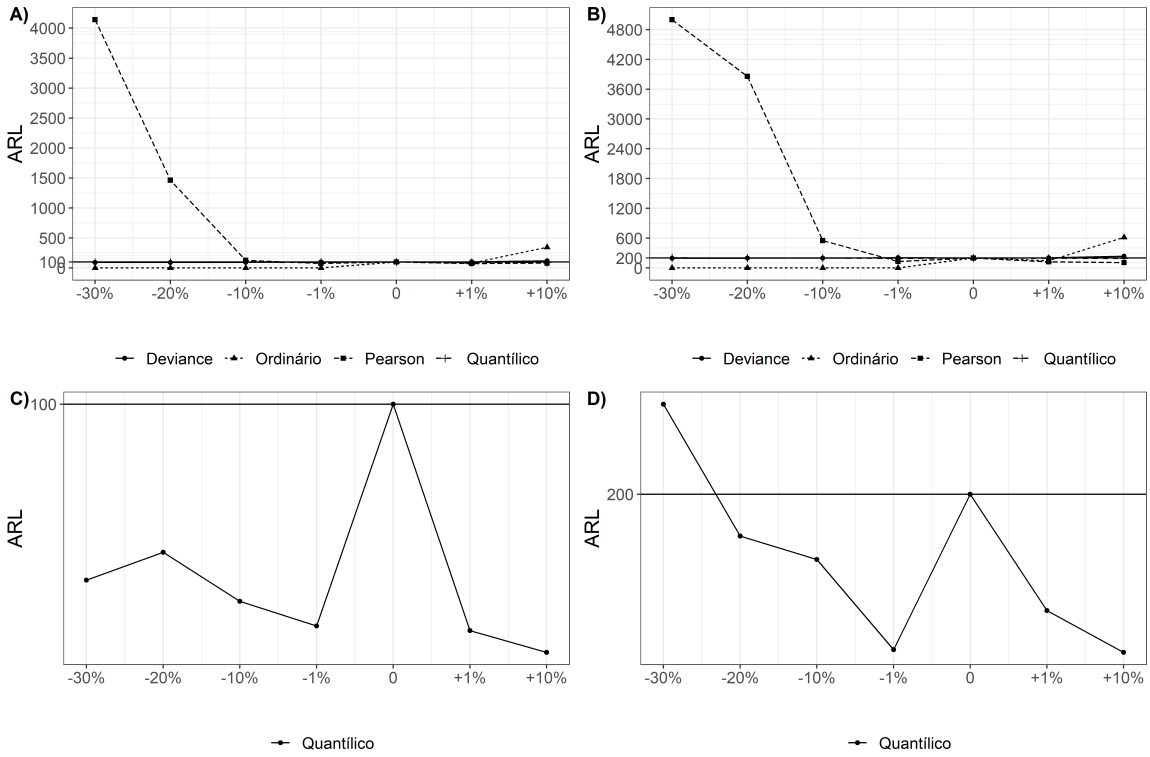


Figura 3.2: Desempenho dos gráficos de controle dos resíduos do modelo ULARMA$(1, 0)$, considerando: (A) e (C) $\text{ARL}_0 = 100$; e (B) e (D) $\text{ARL}_0 = 200$.

Em alguns casos, é observado um comportamento atípico e irregular para o ARL ($\text{ARL}_1 > \text{ARL}_0$). Gráficos de controle com esses comportamento usualmente são chamados de viesados. Geralmente, esse fenômeno ocorre quando a média de Y diminui à medida que também ocorre redução na variância de Y. Isso implica dizer que, nesses cenários, a média e a variância de Y para o processo fora de controle são menores do que a média e a variância de Y para o processo sob controle estatístico. Portanto, os resíduos que consideram de forma mais explícita a diferença entre y_t e $\widehat{\mu}_t$ (como os resíduos de *Pearson* e ordinário), também possuem menor variabilidade; mesmo com a alteração em seu valor médio, a grande maioria dos valores dos resíduos se distribui entre os limites de controle (Tondolo et al., 2016). Avaliando a performance geral dos resíduos, entre os cenários considerados, o resíduo quantílico apresenta melhor desempenho no processo de detecção.

3.4 Aplicação a dados reais

Nesta seção são apresentados e discutidos os resultados obtidos a partir da aplicação das técnicas estatísticas descritas na Seção 3.2, ao mesmo conjunto de dados utilizado no capítulo anterior, contendo informações sobre valores máximos e mínimos da umidade relativa do ar, no deserto do Atacama, situado ao norte do Chile.

Com o intuito de ilustrar a aplicação do gráfico de controle para o monitoramento dos resíduos do modelo ULARMA, considerou-se as primeiras 864 observações para a fase I, isto é, na qual são estimados os limites de controle, e as 7 observações subsequentes foram consideradas 7 novas amostras, caracterizando a fase II. Optou-se por essa configuração, pois fora visto anteriormente que o modelo ULARMA apresentou melhor desempenho em previsões a curto prazo. Utilizou-se o resíduo quantílico como variável a ser monitorada, por ter apresentado boa performance na detecção da não conformidade do processo (ver Seção 3.3.1), e além disso, considerou-se as constantes calibradas apresentadas na Tabela 3.2, para atingir os valores nominais de ARL_0 de 100 e 200. Os resultados são apresentados nas Figuras 3.3 e 3.4, que exibem os gráficos de controle aplicados aos resíduos do modelo ULARMA, para as séries de valores máximos e mínimos da umidade relativa do ar no deserto do Atacama (Chile), assim como suas respectivas funções de autocorrelação e autocorrelação parcial.

Ao analisar a série de máximos, é observado que não existe qualquer correlação serial que não tenha sido controlada pelo modelo, conforme mostram os gráficos de autocorrelação e autocorrelação parcial dos resíduos (Figura 3.3(A) e (B)). Além disso, os gráficos de controle aplicados aos resíduos do modelo (Figura 3.3(C) e (D)) não indicaram a presença de fontes de variação externa ao processo; deste modo, o processo se encontra sob controle estatístico.

Ao considerar a série de mínimos, nota-se que a dependência temporal foi contida pelo modelo, conforme mostram os gráficos de autocorrelação e autocorrelação parcial dos resíduos (Figura 3.4(A) e (B)). E, diferentemente do que foi observado na série de máximos, os gráficos de controle aplicados aos resíduos do modelo (Figura 3.4(C) e (D)) indicam a presença de fontes de variação externa ao processo; deste modo, o processo se encontra fora de controle estatístico.

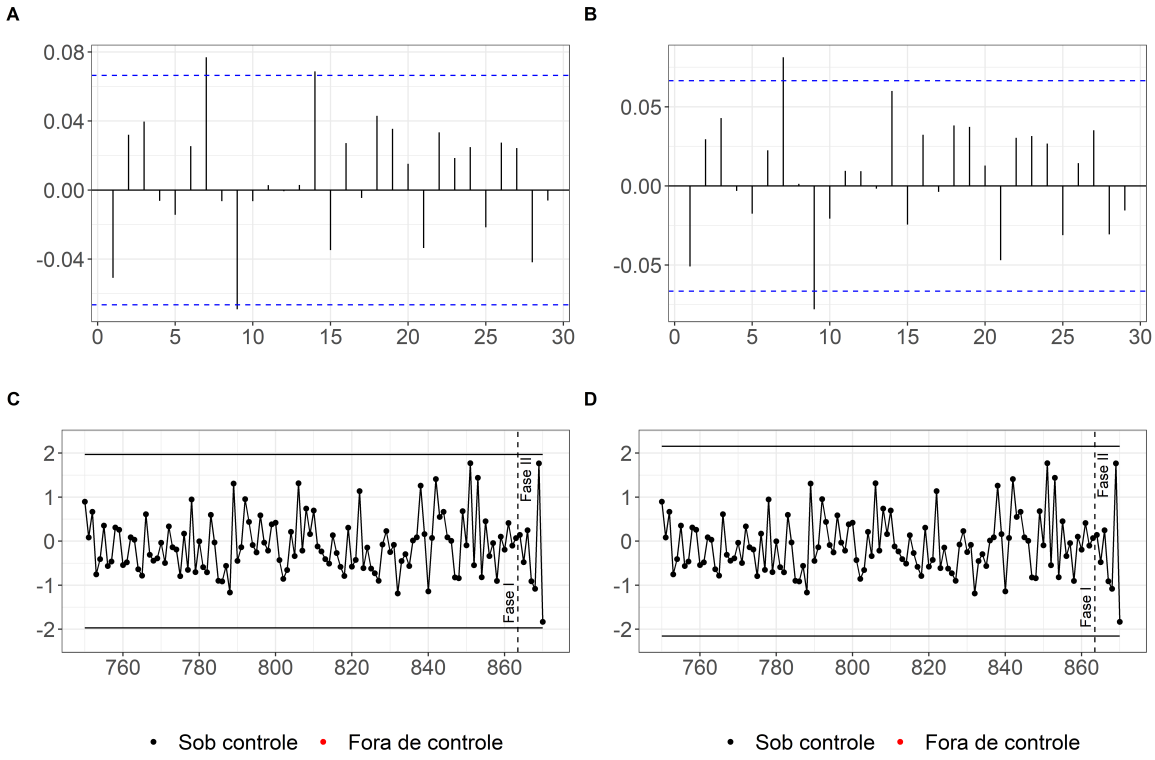


Figura 3.3: (A) Gráfico de autocorrelação; (B) Gráfico de autocorrelação parcial; (C) Gráfico de controle, com $\text{ARL}_0 = 100$; (D) Gráfico de controle, com $\text{ARL}_0 = 200$, para o resíduo quantílico do modelo ULARMA$(1, 0)$ ajustado à série de máximos da umidade relativa do ar no deserto do Atacama, Chile.

3.5 Conclusões

Neste capítulo foi abordada a construção de um gráfico de controle para o monitoramento dos resíduos do modelo ULARMA. Foram considerados quatro tipos de resíduos em dois cenários distintos, representando as séries de mínimos e máximos, com os parâmetros definidos no capítulo anterior (ver Tabela 2.8).

Inicialmente, foi necessário realizar uma etapa de calibração dos limites de controle, e para isto, considerou-se duas propostas: os métodos de regressão linear e de interpolação linear. O segundo método apresentou melhor desempenho e, portanto, foi escolhido para obter os valores de w a serem usados na definição dos limites de controle na fase I.

Posteriormente, foram simuladas amostras com alteração em seu valor médio e observou-se que nenhum resíduo apresentou desempenho excelente na detecção da não conformidade do processo, isto é, nenhum dos resíduos considerados aqui foi capaz de identificar a não conformidade do processo durante todo o intervalo de variação. No entanto, vale ressaltar que o gráfico de controle aplicado ao resíduo quantílico produziu resultados interessantes nos dois cenários.

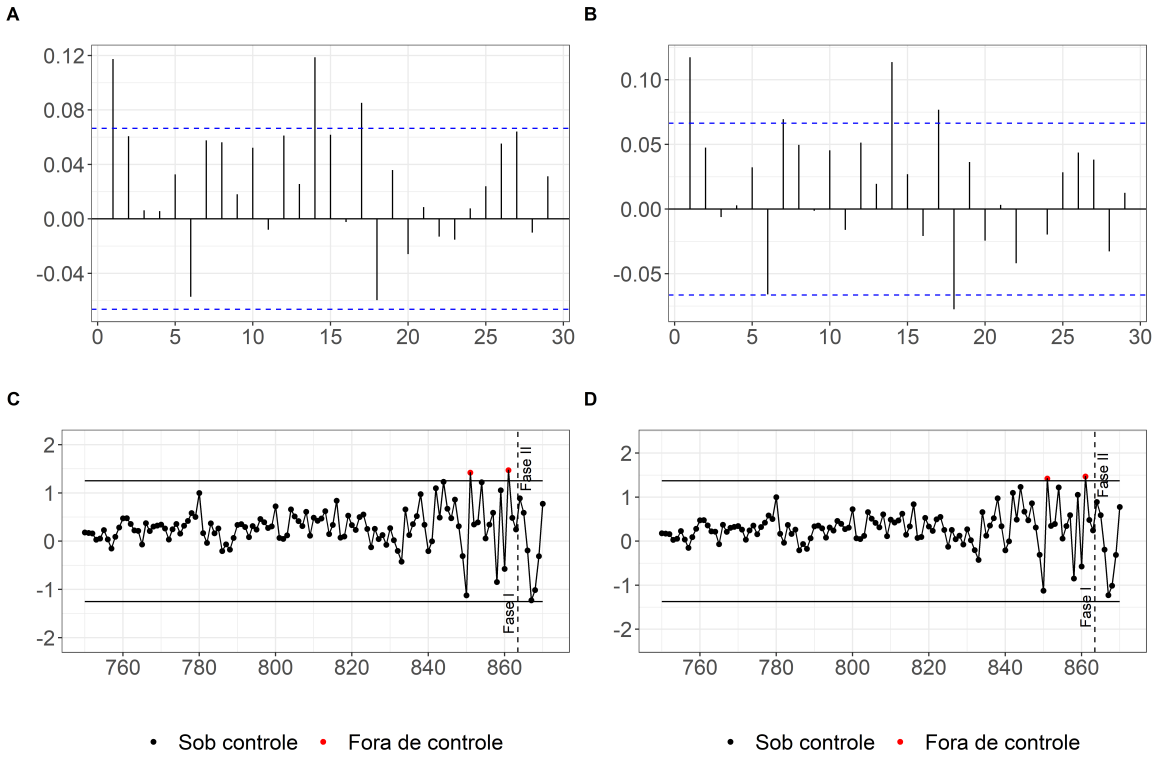


Figura 3.4: (A) Gráfico de autocorrelação; (B) Gráfico de autocorrelação parcial; (C) Gráfico de controle, com $ARL_0 = 100$; (D) Gráfico de controle, com $ARL_0 = 200$, para o resíduo quantílico aleatorizado do modelo ULARMA$(1, 1)$ ajustado à série de mínimos da umidade relativa do ar no deserto do Atacama, Chile.

Como sugestão de trabalhos futuros, seria importante: (i) investigar a performance dos gráficos de controle aplicados aos diferentes tipos de resíduos, considerando outros valores de ARL nominal, como, por exemplo, 300 e 370, que também são bastante utilizados na literatura; (ii) considerar outras propostas para construção de gráficos de controle, que sejam mais sensíveis a pequenas alterações na média do processo, tais como: o gráfico de controle da soma acumulada (CUSUM) e o gráfico de controle da média móvel exponencialmente ponderada (MMEP); (iii) propor um gráfico de controle para uma extensão do modelo ULARMA, que considere a existência de dependência espacial nos dados coletados.

Capítulo 4

Considerações finais

Neste trabalho foi abordada, num primeiro momento, a construção de um modelo estatístico para variáveis aleatórias contínuas que assumem valores dentro do intervalo unitário padrão, isto é, entre zero e um, como, por exemplo, proporções ou frações, escores, índices e taxas, capaz de considerar a existência de uma estrutura de dependência entre dados que são coletados e ordenados ao longo do tempo, caracterizando uma série temporal.

Foi então proposto o modelo *unit-Lindley* autorregressivo e de médias móveis, denotado pela sigla ULARMA. O modelo foi desenvolvido sob a suposição de que a distribuição condicional da variável de interesse, dado o seu histórico, é a *unit-Lindley* (Mazucheli et al., 2018b), assim como introduzido por Benjamin et al. (2003). Foram mostradas a formulação teórica, as propriedades, um estudo de simulação e, ainda, uma aplicação a um conjunto de dados reais sobre umidade relativa do ar diária no deserto do Atacama, Chile.

Na literatura existem outros modelos que são comumente utilizados para descrever dados que possuem essas características/restrições, os quais são baseados em distribuições mais conhecidas, como: βARMA (baseado na distribuição *Beta*) e KARMA (baseado na distribuição *Kumaraswamy*). Apesar de ser uma distribuição de probabilidade recente, a *unit-Lindley* dispõe de propriedades que são interessantes, quando comparada às demais distribuições. Do ponto de vista teórico, possui forma fechada para a FDA e função quantil, expressões simples para a obtenção dos momentos, e pertence à família exponencial. Do ponto de vista prático, sua principal vantagem reside em ser uma distribuição recente, isto é, que ainda é pouco explorada na literatura, além de ser unimodal, uniparamétrica e bastante flexível.

O estudo de simulação realizado revelou uma performance satisfatória dos CMLEs para os parâmetros do modelo, mesmo ao considerar tamanhos de amostra pequenos (avaliou-se n a partir de 70 observações). Com o aumento do tamanho da amostra, as

estimativas tornavam-se mais precisas; e as estimativas do componente autorregressivo se mostravam mais precisas quando comparado ao componente de médias móveis. Na aplicação, notou-se que o modelo proposto possui performance parecida com a das abordagens tradicionais (βARMA e KARMA), na maioria dos cenários analisados. Pois, foram consideradas diferentes composições para as amostras de treino e teste (*out-of-time*) e, em um cenário específico, ele apresentou desempenho superior ao dos demais modelos, com base nas métricas de performance preditiva.

Em seguida, foram apresentadas as técnicas do CEP, por meio de sua ferramenta mais popular, o gráfico (ou carta) de controle, aplicado a variáveis em que há uma estrutura de dependência entre os dados coletados. O escopo consistiu em apresentar o monitoramento do modelo ULARMA, como forma de investigar o comportamento temporal de fenômenos climáticos sobre umidade relativa do ar no deserto do Atacama, Chile.

Foram considerados quatro tipos de resíduos em dois cenários distintos, que representavam as séries de mínimos e máximos, com os parâmetros definidos com base nos resultados da aplicação anterior (Tabela 2.8). Inicialmente, foi necessário realizar uma etapa de calibração dos limites de controle, e para isto, considerou-se duas propostas: os métodos de regressão linear e de interpolação linear. O segundo método apresentou melhor desempenho e, portanto, foi escolhido para obter os valores de w a serem usados na definição dos limites de controle na fase I.

Posteriormente, foram simuladas amostras com alteração em seu valor médio e observou-se que nenhum resíduo apresentou desempenho excelente na detecção da não conformidade do processo, isto é, nenhum dos resíduos considerados neste trabalho foi capaz de identificar a não conformidade do processo durante todo o intervalo de variação. No entanto, o gráfico de controle aplicado ao resíduo quantílico produziu resultados interessantes nos dois cenários estudados.

Como proposta de trabalhos futuros, deseja-se propor um novo modelo espaço-temporal para variáveis aleatórias contínuas que assumem valores no intervalo unitário padrão, com base na distribuição *unit-Lindley*; propor uma extensão que seja capaz de modelar dados sujeitos a flutuações sazonais; considerar a aplicação do modelo ULARMA a outros conjuntos de dados reais e avaliar a sua performance quando comparado às abordagens tradicionais; investigar a performance dos gráficos de controle aplicados aos diferentes tipos de resíduos, considerando outros valores de ARL nominal, como, por exemplo, 300 e 370, que também são bastante utilizados na literatura; considerar outras propostas para construção de gráficos de controle, que sejam mais sensíveis a pequenas alterações na média do processo, tais como: o gráfico de controle da soma acumulada (CUSUM) e o gráfico de controle da média móvel exponencialmente ponderada (MMEP); propor um gráfico de controle para uma extensão do modelo ULARMA, que considere a

existência de dependência espacial nos dados coletados.

Referências Bibliográficas

Abramowitz, M. and Stegun, I. A. (1964). *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55. US Government printing office.

Akaike, H. et al. (1977). On entropy maximization principle.

Akdur, H. T. K. (2021). Unit-lindley mixed-effect model for proportion data. *Journal of Applied Statistics*, 48(13-15):2389–2405.

Altun, E. and El-Morshedy, M. (2021). Simbetareg web-tool: The easiest way to implement the beta and simplex regression models. *Symmetry*, 13(12).

Alwan, L. C. (1992). Effects of autocorrelation on control chart performance. *Communications in statistics-Theory and Methods*, 21(4):1025–1049.

Ansley, C. F. and Newbold, P. (1980). Finite sample properties of estimators for autoregressive moving average models. *Journal of Econometrics*, 13(2):159–183.

Bapat, S. R. and Bhardwaj, R. (2021). On an inflated unit-lindley distribution. *arXiv preprint arXiv:2102.04687*.

Barndorff-Nielsen, O. E. and Jørgensen, B. (1991). Some parametric models on the simplex. *Journal of multivariate analysis*, 39(1):106–116.

Basawa, I., Lund, R., and Shao, Q. (2004). First-order seasonal autoregressive processes with periodically varying parameters. *Statistics & probability letters*, 67(4):299–306.

Bayer, F. M., Bayer, D. M., and Pumi, G. (2017). Kumaraswamy autoregressive moving average models for double bounded environmental data. *Journal of Hydrology*, 555:385–396.

Bayer, F. M., Cintra, R. J., and Cribari-Neto, F. (2018). Beta seasonal autoregressive moving average models. *Journal of Statistical Computation and Simulation*, 88(15):2961–2981.

Bayer, F. M., Pumi, G., Pereira, T. L., and Souza, T. C. (2023). Inflated beta autoregressive moving average models. *Computational and Applied Mathematics*, 42(4):183.

Benjamin, M. A., Rigby, R. A., and Stasinopoulos, D. M. (2003). Generalized autoregressive moving average models. *Journal of the American Statistical association*, 98(461):214–223.

Bicalho, B. d. C. D. (2008). Modelos espaço-temporais: estudo de caso.

Boaventura, L. L., Ferreira, P. H., and Fiaccone, R. L. (2022). On flexible statistical process control with artificial intelligence: Classification control charts. *Expert Systems with Applications*, page 116492.

Borchers, H. W. (2019). pracma: practical numerical math functions. r package version 2.2. 9.

Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

Braz, R. M., Andreozzi, V. L., and Kale, P. L. (2006). Detecção precoce de epidemias de malária no brasil: uma proposta de automação. *Epidemiologia e Serviços de Saúde*, 15(2):21–33.

Brockwell, P. J. and Davis, R. A. (1991). *Time series: theory and methods*. Springer-Verlag.

Chai, T. and Draxler, R. R. (2014). Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature. *Geoscientific model development*, 7(3):1247–1250.

Chen, H., Goldsman, D., Schmeiser, B. W., and Tsui, K.-L. (2017). Symmetric-charts: Sensitivity to nonnormality and control-limit estimation. *Communications in Statistics-Simulation and Computation*, 46(1):358–378.

de Araujo Lima-Filho, L. M., Pereira, T. L., de Souza, T. C., and Bayer, F. M. (2019). Inflated beta control chart for monitoring double bounded processes. *Computers & Industrial Engineering*, 136:265–276.

Dunn, P. K. and Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and graphical statistics*, 5(3):236–244.

Espinheira, P. L., Ferrari, S. L., and Cribari-Neto, F. (2008). On beta regression residuals. *Journal of Applied Statistics*, 35(4):407–419.

Ferrari, S. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of applied statistics*, 31(7):799–815.

Ferreira, P. H., Fonseca, A. O., Nascimento, D. C., Bonnail, E., and Louzada, F. (2022). Unraveling water monitoring association towards weather attributes for response proportions data: A unit-lindley learning. *Plos one*, 17(10):e0275841.

Fiorucci, J. A. and Louzada, F. (2020). Groec: combination method via generalized rolling origin evaluation. *International Journal of Forecasting*, 36(1):105–109.

Fonseca, A., Ferreira, P. H., Nascimento, D. C. d., Fiaccone, R., Ulloa-Correa, C., García-Piña, A., and Louzada, F. (2021). Water particles monitoring in the atacama desert: Spc approach based on proportional data. *Axioms*, 10(3):154.

Haworth, D. A. (1996). Regression control charts to manage software maintenance. *Journal of Software Maintenance: Research and Practice*, 8(1):35–48.

Ho, L., Fernandes, F., and Bourguignon, M. (2018). Control charts to monitor rates and proportions. *Quality and Reliability Engineering International*, 35.

Hyndman, R. J. and Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, 26(3):1–22.

Knuth, D. E. (1996). On the lambert w function. *Advances in Computational Mathematics*, 5(1):329–359.

Kumaraswamy, P. (1980). A generalized probability density function for double-bounded random processes. *Journal of hydrology*, 46(1-2):79–88.

Lemonte, A. J., Barreto-Souza, W., and Cordeiro, G. M. (2013). The exponentiated kumaraswamy distribution and its log-transform. *Brazilian Journal of Probability and Statistics*, 27(1):31–53.

Lemonte, A. J. and Bazán, J. L. (2016). New class of johnson distributions and its associated regression model for rates and proportions. *Biometrical Journal*, 58(4):727–746.

Lima-Filho, L. M. and Bayer, F. (2021). Kumaraswamy control chart for monitoring double bounded environmental data. *Communications in Statistics-Simulation and Computation*, 50(9):2513–2528.

Lindley, D. V. (1958). Fiducial distributions and bayes' theorem. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 102–107.

Lohani, A., Kumar, R., and Singh, R. (2012). Hydrological time series modeling: A comparison between adaptive neuro-fuzzy, neural network and autoregressive techniques. *Journal of Hydrology*, 442:23–35.

Louzada, F., Diniz, C., Ferreira, P., and Ferreira, E. (2013). *Controle estatístico de processos: uma abordagem prática para cursos de engenharia e administração*. Grupo Gen-LTC.

Lãpez, F. O. (2013). A Bayesian Approach to Parameter Estimation in Simplex Regression Model: A Comparison with Beta Regression. *Revista Colombiana de EstadÃstica*, 36:1 – 21.

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. (2018). The m4 competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, 34(4):802–808.

Maldonado, S., López, J., and Iturriaga, A. (2022). Out-of-time cross-validation strategies for classification in the presence of dataset shift. *Applied Intelligence*, 52(5):5770–5783.

Mandel, B. (1969). The regression control chart. *Journal of Quality Technology*, 1(1):1–9.

Martínez-Flórez, G., Leiva, V., Gómez-Déniz, E., and Marchant, C. (2020). A family of skew-normal distributions for modeling proportions and rates with zeros/ones excess. *Symmetry*, 12(9):1439.

Mazucheli, J., Menezes, A., and Ghitany, M. (2018a). The unit-weibull distribution and associated inference. *J. Appl. Probab. Stat*, 13(2):1–22.

Mazucheli, J., Menezes, A. F., and Dey, S. (2018b). The unit-birnbaum-saunders distribution with applications. *Chilean Journal of Statistics*, 9(1):47–57.

Mazucheli, J., Menezes, A. F. B., and Chakraborty, S. (2019). On the one parameter unit-lindley distribution and its associated regression model for proportion data. *Journal of Applied Statistics*, 46(4):700–714.

McCullagh, P. (2019). *Generalized linear models*. Routledge.

McCullagh, P. and Nelder, J. (1989). Generalized linear models ii.

Mitnik, P. A. and Baek, S. (2013). The kumaraswamy distribution: median-dispersion re-parameterizations for regression modeling and simulation-based estimation. *Statistical Papers*, 54(1):177–192.

Montgomery, D. C. (2020). *Introduction to statistical quality control*. John Wiley & Sons.

Moraes, D., Oliveira, F. L. P. d., Quinino, R. d. C., and Duczmal, L. H. (2014). Self-oriented control charts for efficient monitoring of mean vectors. *Computers & Industrial Engineering*, 75:102–115.

Mukherjee, P. S. (2016). On phase ii monitoring of the probability distributions of univariate continuous processes. *Statistical Papers*, 57(2):539–562.

Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384.

Nocedal, J. and Wright, S. J. (2006). *Numerical Optimization*. Springer Science & Business Media, 2 edition.

Ospina, R. and Ferrari, S. L. (2010). Inflated beta distributions. *Statistical papers*, 51(1):111–126.

Palm, B. G. (2016). Intervalos de predição no modelo beta autorregressivo de médias móveis.

Pereira, M. S. et al. (2023). *Novos modelos estatísticos para processamento e monitoramento de imagens de radar de abertura sintética*. PhD thesis, Universidade Federal de Santa Maria.

Prataviera, F., Cordeiro, G., Ortega, E., Hashimoto, E., and Cancho, V. (2021). A new regression model for rates and proportions data with applications. *Journal of Applied Statistics*, pages 1–25.

Pumi, G., Valk, M., Bisognin, C., Bayer, F. M., and Prass, T. S. (2019). Beta autoregressive fractionally integrated moving average models. *Journal of Statistical Planning and Inference*, 200:196–212.

Rocha, A. V. and Cribari-Neto, F. (2009). Beta autoregressive moving average models. *Test*, 18(3):529–545.

Rocha, A. V. and Cribari-Neto, F. (2017). Erratum to: Beta autoregressive moving average models. *Test*, 26:451–459.

Sagrillo, M., Guerra, R. R., and Bayer, F. M. (2021). Modified kumaraswamy distributions for double bounded hydro-environmental data. *Journal of Hydrology*, 603:127021.

Sagrillo, M., Guerra, R. R., Machado, R., and Bayer, F. M. (2023). A generalized control chart for anomaly detection in sar imagery. *Computers & Industrial Engineering*, 177:109030.

Sant'Anna, Â. M. O. and ten Caten, C. S. (2012). Beta control charts for monitoring fraction data. *Expert Systems with Applications*, 39(11):10236–10243.

Savage, N., Agnew, P., Davis, L., Ordóñez, C., Thorpe, R., Johnson, C., O'Connor, F., and Dalvi, M. (2013). Air quality modelling using the met office unified model (aqum os24-26): model description and initial evaluation. *Geoscientific Model Development*, 6(2):353–372.

Schaffer, J. R. and Kim, M.-J. (2007). Number of replications required in control chart monte carlo simulation studies. *Communications in Statistics—Simulation and Computation®*, 36(5):1075–1087.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, pages 461–464.

Sena, J. G. S. d., Ferreira, P. H., and Fiaccone, R. L. (2022). Statistical process control as a tool to control and prevent malaria epidemics in the legal amazon region. *Brazilian Journal of Biometrics*, 40(1).

Shewhart, W. A. (1929). Control of quality of manufactured product.

Shewhart, W. A. (1931). *Economic control of quality of manufactured product*. Macmillan And Co Ltd, London.

Tantithamthavorn, C., McIntosh, S., Hassan, A. E., and Matsumoto, K. (2017). An empirical comparison of model validation techniques for defect prediction models. *IEEE Transactions on Software Engineering*, 43(1):1–18.

Tiku, M. L., Wong, W.-K., Vaughan, D. C., and Bian, G. (2000). Time series models in non-normal situations: Symmetric innovations. *Journal of Time Series Analysis*, 21(5):571–596.

Tondolo, C. M. et al. (2016). Gráficos de controle para dados do tipo taxas e proporções autocorrelacionados.

Triola, M. F., Goodman, W. M., Law, R., and Labute, G. (2004). *Elementary statistics*. Pearson/Addison-Wesley Boston.

Wang, H. (2009). Comparison of p control charts for low defective rate. *Computational statistics & data analysis*, 53(12):4210–4220.

Wongrin, W., Srianomai, S., and Klomwises, Y. (2020). Bayesian unit-lindley model: Applications to gasoline yield and risk assessment data. *Naresuan University Journal: Science and Technology (NUJST)*, 28(2):41–51.