UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS DA COMPUTAÇÃO

DAILYS MAITE ALIAGA REYES

Predição para Dados Simbólicos Multi-valorados de Tipo Quartis: Caso Especial Dados Representados por Boxplots

Recife

2022

DAILYS MAITE ALIAGA REYES

Predição para Dados Simbólicos Multi-valorados de Tipo Quartis: Caso Especial Dados Representados por Boxplots

Trabalho apresentado ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco, como requisito parcial para obtenção do grau de Doutor em Ciência da Computação.

Área de Concentração: Inteligência Computacional

Orientador: Prof. Dr. Adriano Lorena Inacio de Oliveira.

Coorientadora: Profa. Dra. Renata Maria Cardoso Rodrigues de Souza.

Recife

2022

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

R457p Reyes, Dailys Maite Aliaga
Predição para dados simbólicos multi-valorados de tipo quartis: caso especial dados representados por boxplots / Dailys Maite Aliaga Reyes. – 2022.
116 f.: il., fig., tab.

Orientador: Adriano Lorena Inacio de Oliveira.
Tese (Doutorado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2022.
Inclui referências.

1. Inteligência computacional. 2. Séries temporais. I. Oliveira, Adriano Lorena Inacio de (orientador). II. Título.

006.31 CDD (23. ed.) UFPE - CCEN 2022 – 88

Dailys Maite Aliaga Reyes

“Predição para Dados Simbólicos Multi-valorados de Tipo Quartis: Caso Especial Dados Representados por Boxplots”

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Doutor em Ciência da Computação. Área de Concentração: Inteligência Computacional

Aprovado em: 23/02/2022.

_____

Orientador: Prof. Dr. Adriano Lorena Inacio de Oliveira

BANCA EXAMINADORA

_____

Prof. Dr. Paulo Salgado Gomes de Mattos Neto
Centro de Informática /UFPE

_____

Prof. Dr. Nivan Roberto Ferreira Junior
Centro de Informática /UFPE

_____

Prof. Dr. Francisco Cribari Neto
Departamento de Estatística / UFPE

_____

Prof. Dr. Telmo de Menezes e Silva Filho
University of Bristol, Reino Unido

_____

Prof. Dr. Leandro Carlos de Souza
Departamento de Informática / UFPB

Á minha irmã Daliannis Maité Aliaga Reyes (in memoriam), que já se foi, mas se faz presente em todos os dias da minha vida. Sei que, de algum lugar, ela olha por mim e fica orgulhosa.

Aos meus queridos meninos Dylan e Bastian que são a minha inspiração.

A Deus, que sempre foi o autor da minha vida e do meu destino.

AGRADECIMENTOS

Desejo exprimir os meus agradecimentos a todos aqueles que, de alguma forma, permitiram que esta tese se concretizasse.

Em primeiro lugar quero agradecer à Professora Dra. Renata Maria Cardoso Rodrigues de Souza, o ter-me deixado fazer parte do seu grupo de trabalho e, ter acreditado em mim e nas minhas capacidades. Agradeço ainda o trato simples, correto e científico, com que sempre abordou as nossas reuniões de trabalho, sem nunca ter permitido que o desalento se instalasse, mesmo quando as coisas não corriam bem.

Ao meu orientador Dr.Adriano Lorena Inácio de Oliveira, por me acolher quando precisei e pelos auxílios diversos.

Ao Professor Leandro Carlos de Souza, pela ajuda preciosa dada ao desenvolvimento do meu artigo e capítulo de esta tese.

Agradeço a todos os professores do programa de pós graduação em ciências da computação, principalmente aqueles que tive contato na sala de aula. A todos do CIn que me deram a oportunidade de estudar aqui apesar de ser estrangeira e à FACEPE pelo apoio econômico.

Agradeço a todos que participaram direta ou indiretamente da pesquisa nos diversos experimentos realizados. Fica aqui minha gratidão aos colegas do grupo de análise de dados simbólicos.

Agradeço aos meus pais Alberto e Lérida, que me deram o privilégio sagrado da vida. Pelo seus esforços, carinho, dedicação e conselhos. Agradeço pelo seu apoio apesar das lágrimas que estarão sempre gravadas no meu coração quando decidi deixar meu país.

A meus avós Ramón (in memoriam) e Rafaela, responsáveis pela formação de meu caráter. Agradeço pela sua sabedoria, determinação e por me apoiar nas escolhas boas e ruins que fiz na minha vida.

Agradecimento especial a meu marido Juan Alberto Fajardo Barrera presença constante, compreensão máxima e companheiro de todas as horas. Muito Obrigada meu amor.

Aos meus amigos, em especial Juan González e Yaicel Gé, companheiros de vida que sempre me ajudam resolvendo problemas que eu arrumo para eles, pelas conversas filosóficas sobre o futuro e pelo incentivo.

Enfim, agradeço a todos os que participaram direta ou indiretamente contribuindo para mais uma conquista alcançada até aqui. Muitíssimo obrigada!

"A tarefa não é tanto ver aquilo que ninguém viu, mas pensar o que ninguém ainda pensou sobre aquilo que todo mundo vê." (CORDEIRO, 2017, p. 1).

RESUMO

Um dado simbólico de tipo $boxplot$ pode ser considerado como um caso particular das variáveis numéricas multi-valoradas no contexto da Análises de Dados Simbólicos (ADS). Este tipo de dado tem uma estrutura simples que permite resumir informações de unidades agregadas, chamadas de classes. No entanto, esse tipo de estrutura tem sido pouco explorada na literatura de ADS. Este trabalho apresenta duas novas abordagens de predição com o objetivo de extrair conhecimento e fazer inferência usando dados de $boxplot$. A primeira abordagem considera um modelo de regressão para $boxplot$ através da equação paramétrica da reta. Esta parametrização permite o ajuste dos pontos nas variáveis regressoras que permite melhorar a qualidade da variável resposta. Nessa direção, um critério é também proposto para verificar a coerência matemática da predição. Se a coerência não é garantida, uma nova estratégia, através de transformações $Box - Cox$ é aplicada sobre a variável resposta de tipo $boxplot$. A segunda abordagem proposta nesse trabalho consiste de um modelo que combina agregação, seleção de protótipos e previsão de séries temporais. Inicialmente, as séries temporais são agregadas em classes de entidades e representadas por $boxplots$. Um processo de seleção de protótipos baseado na informação mútua é aplicado para mitigar ruídos no conjunto de dados. Por último, um modelo multivariado para previsão de $boxplots$ é construído. Ambos modelos são avaliados com conjuntos de dados sintéticos e reais. Uma comparação entre as abordagens propostas e outros métodos de predição da literatura de ADS é também descrita. Os resultados obtidos reforçam que para os conjuntos de dados usados, o poder preditivo das abordagens propostas é superior aos métodos da literatura usados para comparar. Além disso, este trabalho apresenta uma aplicação do mundo real no Setor Elétrico Brasileiro para fazer predição da temperatura dos motores usando a abordagem de regressão paramétrica para dados de $boxplot$.

Palavras-chaves: análise de dados simbólicos; boxplots; seleção de protótipos; informação mútua; múltiplas séries temporais; regressão linear.

ABSTRACT

A symbolic boxplot data can be considered as a particular case of the numerical multi-valued variables in the context of Simbolic Data Analysis (SDA). This data type has a simple structure that allows to summarize information from aggregated units, called classes. However, this type of structure has been little explored in the SDA literature. This work presents two new prediction approaches with the objective of extracting knowledge and making inferences using boxplot data. The first approach considers a regression model for boxplot through the parametric equation of the line. This parameterization allows the adjustment of the points in the regressors, which improves the quality of the response variable. In this direction, a criterion is also proposed to verify the mathematical coherence of the prediction. If coherence is not guaranteed, a new strategy, through Box-Cox transformations, is applied on the response variable of type boxplot. The second approach proposed in this work consists of a model that combines aggregation, prototype selection and time series prediction. Initially, time series are aggregated into entity classes and represented by boxplots. A prototype selection process based on mutual information is applied to mitigate noise in the dataset. Finally, a multivariate model for forecasting boxplots is built. Both models are evaluated with synthetic and real data sets. A comparison between the proposed approaches and other prediction methods from the SDA literature is also described. In addition, this work presents a real-world application in the Brazilian Electricity Sector to predict the temperature of motors using the parametric regression approach for boxplot data.

Keywords: symbolic data analysis; boxplots; prototype selection; mutual information; multiple time series; linear regression.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ADS	Análises de Dados Simbólicos
AIC	*Akaike Information Criterion*
AID	*Automatic Interation Detector*
ANEEL	Agência Nacional de Energia Elétrica
AR	*Auto-regressive Model*
ARIMA	*Auto-regressive Integrated Moving Average*
ARMA	*Auto-regressive Moving Average Models*
CA	*Cellular Automata*
CNN	*Condensed Nearest Neighbor Rule*
CO	Monóxido de Carbono
ENN	*Edited Nearest Neighbor Rule*
HT	Alta Temperatura
IFF	*International Institute of Forecasters*
IM	Informação Mútua
K-NN	*k-Neighbors Neighbours*
LT	Baixa Temperatura
LVQ	*Learning Vector Quantization*
MA	*Moving Average Model*
MC	Método do Centro
MCR	Método do Centro e Range
MinMax	Método do Mínimo e o Máximo
MLP	*Multilayer Perceptron*
MMRE	Magnitude Média dos Erros Relativos
MP	Método parametrizado
MRLC	Modelo de Regressão Linear Clássico

MRPB	Método de Regressão Linear Parametrizada para Boxplots
MSE	*Mean Square Error*
NCR	*Neighborhood Cleaning Rule*
OLS	*Ordinary Least Squares*
OSS	*One-Sided Selection*
PCA	*Principal Component Analysis*
SIN	Sistema Interligado Nacional
SVR	*Support Vector Regression*
UTE	Usina Termo-eléctrica
VAR	Vetores Auto-Regressivos
VECMN	*Vector Error Correction Model*

LISTA DE SÍMBOLOS

γ	Letra grega Gama
Λ	Letra grega Lambda
α	Letra grega alpha
$\in$	Pertence
$\geq$	Maior ou Igual
δ	Letra grega Delta
θ	Letra grega Teta
σ	Letra grega Sigma
μ	Letra grega Mi
ω	Letra grega Omega
β	Letra grega Beta

SUMÁRIO

1 INTRODUÇÃO

Este primeiro capítulo fundamenta a utilização da análises de dados simbólicos (ADS), expõem-se os objetivos em relação ao trabalho proposto e por fim, descreve-se a organização dos demais capítulos.

1.1 CONTEXTO E MOTIVAÇÃO

Nos últimos tempos, a capacidade de armazenamento de dados disparou e as bases de dados podem se tornar excessivamente grandes e complexas. Consequentemente, os conjuntos de dados que são extraídos podem facilmente ter centenas de variáveis e milhares de registros, o que os torna difíceis de processar com as técnicas tradicionais de análise de dados. Apesar do alto poder de processamento dos computadores atuais, o esforço computacional necessário para a manipulação dessas bases de dados em um tempo razoável ainda é um problema (DIDAY; NOIRHOMME-FRAITURE, 2008; BILLARD; DIDAY, 2019; SILVA; SOUZA; CYSNEIROS, 2022).

Uma forma de contornar esse problema é resumir os grandes conjuntos de dados em menores e mais gerenciáveis sem perda significativa de informações. Nesse contexto, destaca-se a ADS, um paradigma que abre um vasto domínio de pesquisa e aplicações possibilitando a agregação de bases de dados clássicos no nível de granularidade escolhido pelo usuário, mantendo as informações sobre a variabilidade intrínseca e, em seguida, analisando os dados resultantes (simbólicos). Neste contexto, os indivíduos são agrupados em classes que obedecem a um determinado critério e que são consideradas como novas unidades. Essas unidades são exemplificadas por variáveis simbólicas, que podem assumir distribuições de probabilidade, intervalos, histogramas, listas de categorias ou números (DIDAY, 2016).

ADS surgiu através da influência simultânea da Análise Exploratória de Dados (BEATON; TUKEY, 1974; BOCK, 1974; SAPORTA, 1990), Inteligência Artificial (BEATON; TUKEY, 1974; RUSSELL et al., 2003; LUGER, 2005) e Taxonomia Numérica (SNEATH; SOKAL, 1962). As primeiras tentativas para obter dados simbólicos a partir de dados clássicos foram realizadas por Belson (1959), seguido por Morgan e Sonquist (1963) com o método *Automatic Interation Detector* (AID). A partir do final da década dos 80, ADS deixou de ser restrita a um pequeno grupo de pesquisadores para ser uma área de pesquisa bastante relevante marcada por muitas publicações e conferências (NOIRHOMME-FRAITURE; BRITO, 2011).

Como a estrutura dos dados clássicos é diferente, é necessário generalizar os métodos tradicionais para métodos com dados simbólicos através de desenvolvimentos exploratórios, estatísticos e representações gráficas para esses novos tipos de dados (BILLARD; DIDAY, 2006; DIDAY; NOIRHOMME-FRAITURE, 2008). Sendo uma área com objetivos tão ambiciosos, há muitas possibilidades de desenvolvimento dentro dela. Inicialmente, as contribuições na área concentraram-se na análise de $cluster$ e nos dados simbólicos de tipo intervalo, uma vez que estes envolvem menos complexidade. Ao longo dos anos, o catálogo de métodos foi expandido e é mais comum encontrar métodos propostos para outros dados simbólicos.

Apesar dessa evolução, como afirmado por Billard e Diday (2003), é necessário e desejável que o repertório de métodos que lida com esse tipo de dados continue a se expandir. Ainda existem algumas áreas dentro da mineração de dados e do aprendizado estatístico em que há pouco desenvolvimento. No início desta pesquisa, uma dessas áreas pouco explorada era a predição de séries temporais simbólicas multi-valoradas, que é um dos temas explorados na presente teses.

Os dados simbólicos multi-valorados podem surgir em duas situações. Em primeiro lugar, podem ser originalmente observados e coletados do mundo real. Um exemplo desse caso é a variação de uma ação na Bolsa de Valores ao longo do dia. O resultado é uma faixa de valores em uma determinada frequência (em segundos), mas que precisam ser analisados em uma frequência menor (em minutos ou horas).

Em segundo lugar, os dados simbólicos de múltiplos valores ou multi-valorados, são gerados a partir dos dados de pontos usuais (COSTA; PIMENTEL; SOUZA, 2013). Para ilustrar essa situação, é possível imaginar um país que tem suas temperaturas medidas em um determinado número de estações meteorológicas. Esses dados podem ser agrupados para obter a variação das temperaturas por dia em todo o país. Nesse caso, todos os dados podem ser agregados para produzir dados de quartil diários. Para fazer isso, os dados de ponto são agregados para obter um conjunto de classes. Deste modo, cada classe é descrita por uma lista de valores numéricos que são os quartis de essa classe. Os quartis são excelente para apresentar uma classe, porque são estatísticas sólidas que descrevem distribuições distorcidas ou com observações extremas. Além, são úteis para construir $boxplots$, que são uma ferramenta exploratória extremamente útil, com uma estrutura simples que resume grupos de dados numéricos usando estatísticas de resumo robustas (TUKEY, 1977; BENJAMINI, 1988).

Um dos problemas que tem atraído o interesse da comunidade científica na área de ADS, é a regressão de variáveis simbólicas. Onde o principal objetivo é encontrar uma equação linear

nos parâmetros que represente a variável resposta a partir das variáveis regressoras (RENCHER; SCHAALJE, 2008; MONTGOMERY; PECK; VINING, 2001; DRAPER; SMITH, 1998; SEBER; LEE, 2012).

Esta tese aborda os métodos de predição (regressão e séries temporais) para dados simbólicos multi-valorados de tipo quartis: especificamente $boxplots$. Os $boxplots$ são gráficos que exibem a distribuição dos dados com base em um resumo de cinco valores. Um motivo importante para usar $boxplots$ é que podem mostrar estruturas de longo prazo em que resumem uma grande quantidade de dados ao longo do tempo, descrevendo as suas características essenciais com uma precisão razoável. Além disso, permitem a descrição das distribuições por meio de uma tela concisa com uma representação gráfica clara que pode ser facilmente explicada a não-estatísticos (TUKEY, 1977; BENJAMINI, 1988). São flexíveis o suficiente para refletir qualquer estimador de densidade com base em intervalos e não requerem nenhuma distribuição a priori. Sua estrutura simples simplifica seu tratamento computacional e sua versatilidade permite ao analista focar nas características que mais lhe interessam, como: um conjunto de quartis ou uma parte do intervalo da variável (WICKHAM; STRYJEWSKI, 2011).

Na literatura científica atual, os métodos de predição para variáveis simbólicas, tanto em regressão quanto em séries temporais, abordam maiormente os dados simbólicos de tipo intervalos. Embora tenha havido um interesse crescente na análise de outras variáveis simbólicas, ainda mais pesquisas são desenvolvidas para variáveis de tipo intervalo. Isso se deve ao fato de que a formulação de modelos não é trivial para as variáveis simbólicas de natureza mais complexa. Portanto, ainda são necessários métodos de regressão para outras variáveis simbólicas que possam ser aplicados para resolver problemas de regressão em diferentes domínios (CHACÓN; RODRÍGUEZ, 2021; CORDEIRO, 2017).

Um trabalho envolvendo dados simbólicos de tipo $boxplot$ foi apresentado por Drago (2015). Este é, sem dúvida, um bom sinal que indica que a disciplina continua seu progresso e que os pesquisadores trabalham para ampliar seus limites. Sem embargo, esta abordagem considera modelos independentes para previsão de dados simbólicos apesar de que poderia ser considerado um modelo multivariado. Pode-se esperar que o uso de um modelo multivariado gere previsões mais precisas do que vários modelos univariados, uma vez que as curvas são ajustadas conjuntamente.

Outra questão importante que motivou o desenvolvimento desta pesquisa, é a escassez de trabalhos que fazem aplicações reais com dados simbólicos na área de ADS. Dentre as aplicações mas relevantes encontradas na literatura, está a utilização de técnicas de filtragem

de informações para recomendar itens (BEZERRA; CARVALHO, 2011). Nessa abordagem, o perfil do usuário é desenhado por meio de estruturas de dados simbólicos e as correlações de usuário e item são calculadas por meio de funções de distância adaptadas do domínio de ADS. Outro trabalho que aplica dados simbólicos em problemas reais é o apresentado por Angadia e Kagawadeb (2017). Neste trabalho, os autores usam dados simbólicos de tipo intervalo para extrair conhecimento de diferentes expressões faciais. Por último, cabe ressaltar o trabalho de Silva, Souza e Cysneiros (2021), aqui os autores aplicam um modelo de regressão linear para dados simbólicos polinomiais no Sistema Brasileiro de Avaliação da Educação Básica (SAEB), dando uma nova perspectiva aos gestores dos municípios para concretizar a política pública no sistema educacional brasileiro.

Neste contexto, esperamos com esta pesquisa, não apenas estender a metodologia de regressão lineal e previsão de séries temporais para análise de dados multi-valorados de tipo $boxplots$, mas também contribuir com novas metodologias e aplicações reais no domínio de ADS.

1.2 OBJETIVOS

O objetivo principal desta tese é desenvolver um conjunto de soluções teóricas e aplicadas na área de regressão linear para dados simbólicos multi-valorados de tipo quartis: especificamente dados simbólicos de $boxplots$, com o intuito de reduzir grandes conjuntos de dados e extrair novos conhecimentos. De modo específico se propõe:

1. Desenvolver um novo método de regressão linear por meio de uma representação paramétrica para dados simbólicos de tipo $boxplot$. Esta parametrização permite que o próprio modelo de regressão encontre os melhores pontos representativos nas variáveis regressoras, sem que eles sejam fixados previamente, o que torna a modelagem mais geral;

2. Aplicar o método de regressão linear paramétrica desenvolvido nos dados do Setor Elétrico Brasileiro como parte do projeto de P&D "PD-02901-0003 2019: Mapa de risco em tempo real baseado em aprendizado de máquina aplicado em manutenção preditiva", regulamentado pela Agência Nacional de Energia Elétrica (ANEEL), Brasil;

3. Desenvolver modelos de previsão para séries temporais simbólicas de tipo $boxplots$ ba-

seados na representação com menos parâmetros e o uso de seleção de protótipos, com a finalidade de resumir o conjunto de dados de entrada para mitigar dados ruidosos e melhorar a acurácia dos modelos existentes na literatura.

1.3 ORGANIZAÇÃO DA TESE

Além deste capítulo introdutório, esta tese é dividida em mais cinco capítulos:

Capítulo 2 - Fundamentação Teórica

A finalidade deste capítulo é apresentar os conceitos fundamentais de ADS e os trabalhos mais relevantes realizados pela comunidade científica nesta área. Também são apresentados os principais resultados sobre regressão linear simbólica e séries temporais simbólicas que existem na literatura.

Capítulo 3 - Regressão Linear Parametrizada para Dados Simbólicos de Boxplots.

Este capítulo descreve a representação dos $boxplots$ através da equação paramétrica da reta, bem como o método de regressão linear que determina os melhores pontos a serem utilizados na construção dos modelos. Também é mostrada uma análise que verifica a coerência matemática dos $boxplots$ preditos. A comparação entre os métodos de regressão é realizada por meio de simulações de Monte Carlo, utilizando o método $Hold\ Out$ tanto para os dados sintéticos como para os reais.

Capítulo 4 - Previsão de Séries Temporais Simbólicas de Boxplots.

Descreve as séries temporais simbólicas multi-valoradas de tipo quartis e como construir as séries temporais de $boxplots$. Propõe uma definição, explica como podem ser obtidas e apresenta as diferentes abordagens para a predição de séries temporais de $boxplots$.

Capítulo 5 - Conclusões.

Este capítulo apresenta os principais resultados e conclusões referentes à pesquisa realizada, bem como, as contribuições na área de regressão para dados simbólicos multi-valorados de tipo $boxplots$. Por fim, são apresentadas algumas perspectivas em aberto para trabalhos futuros.

2 EMBASAMENTO TEÓRICO: ANÁLISES DE DADOS SIMBÓLICOS

Este capítulo divide-se em três partes: inicialmente será abordada a ADS, suas principais características, aplicações e uma revisão dos trabalhos mais relevantes desenvolvidos nesta área, os quais fundamentaram esta tese. Na continuação, apresenta-se brevemente uma revisão da literatura de ADS relativa aos principais modelos de regressão linear para dados simbólicos. Por fim, uma seção relativa as séries temporais simbólicas e as principais ferramentas utilizadas atualmente na previsão de esse tipo de series. Além disso, essa seção também descreve o método de seleção de protótipos usando informação mútua e os principais trabalhos encontrados na literatura para seleção de protótipos tanto em problemas de classificação como em regressão.

2.1 ANÁLISE DE DADOS SIMBÓLICOS

A ADS é um campo de pesquisa relacionado à análise multivariada, reconhecimento de padrões e inteligência artificial, que oferece técnicas adequadas para resumir grandes conjuntos de dados em menores e mais gerenciáveis sem perda significativa de informações. Nesse sentido, a ADS propõe um caminho alternativo para construir, descrever, analisar e extrair novos conhecimentos a partir de conjuntos de dados simbólicos (DIDAY, 2016).

Como foi introduzido anteriormente, a ADS surgiu através da influência simultânea da Análise Exploratória de Dados (BEATON; TUKEY, 1974; BOCK, 1974; SAPORTA, 1990), Inteligência Artificial (BEATON; TUKEY, 1974; RUSSELL et al., 2003; LUGER, 2005) e Taxonomia Numérica (SNEATH; SOKAL, 1962). As primeiras tentativas para obter dados simbólicos a partir de dados clássicos foram realizadas por Belson (1959), seguido por Morgan e Sonquist (1963) com o método AID. Os primeiros algoritmos, chamados de *Conceptual Clustering*, foram apresentados por Diday e Simon (1980) e Michalski, Stepp e Diday (1981). Trabalhos pioneiros como Diday (1987), Diday (1989) e Diday (1991) apresentam os princípios básicos de ADS. Com isso, vários outros trabalhos foram realizados em diversas direções.

A partir do final da década dos 80, ADS deixou de ser restrita a um pequeno grupo de pesquisadores para ser uma área de pesquisa bastante relevante marcada por muitas publicações e conferências (NOIRHOMME-FRAITURE; BRITO, 2011). Bock e Diday (2000) apresentam de maneira sólida os conceitos de ADS e os principais métodos estatísticos desenvolvidos

para manipular dados desta natureza. A seguir serão comentados alguns métodos clássicos já estendidos para o tratamento de problemas que envolvem dados simbólicos.

Na estatística descritiva por exemplo: Carvalho (1995) introduziu a construção de histogramas para dados simbólicos booleanos; Noirhomme-Fraiture e Rouard (1997) apresentaram o $ZoomStar$, um método gráfico para visualizar objetos simbólicos. No caso univariado $(p = 1)$, conceitos como a média amostral, a variância amostral e a distribuição de frequência foram desenvolvidos para variáveis simbólicas (BERTRAND; GOUPIL, 2000). Posteriormente, esses conceitos foram estendidos para o caso multivariado, o seja, quando $p > 1$ (BILLARD; DIDAY, 2006; BILLARD, 2004).

Para a análise de componentes principais (*Principal Component Analysis* (PCA)) simbólica, Cazes et al. (1997), Douzal-Chouakria (1998) e Chouakria, Diday e Cazes (1998) desenvolveram métodos para reduzir um conjunto de variáveis simbólicas de natureza intervalar $p-dimensional$ para um conjunto $s-dimensional$. O objetivo é encontrar um conjunto de s componentes que juntos expliquem ao máximo a estrutura de variação das p variáveis originais. Métodos com o mesmo propósito para dados simbólicos intervalares também foram propostos por Ichino e Yaguchi (1994) utilizando a métrica de $Minskowsky$, e por Nagabhushan, Gowda e Diday (1995) usando princípios de séries de Taylor. Pouco depois, Lauro e Palumbo (2000) apresentaram técnicas novas de PCA para variáveis intervalares baseadas nos limites inferior e superior dos intervalos. Os autores asseguram que sua abordagem está em conformidade com alguns métodos desenvolvidos focados apenas no centro dos intervalos e que consideram o range como um erro de mensuração ou uma perturbação aos dados. Uma extensão mais geral de PCA para dados simbólicos intervalares foi apresentada por Irpino (2006), incluindo a dependência temporal dos dados, por exemplo, considerando os preços de abertura e fechamento de uma ação negociada no mercado financeiro.

Outra técnica estatística muito importante desenvolvida para dados simbólicos é a análise de $cluster$. Neste sentido, cabe destacar os trabalhos de Gowda e Diday (1991), Gowda e Diday (1992) e Guru, Kiranagi e Nagabhushan (2004), que apresentaram as principais medidas de similaridade ou dissimilaridade para mensurar a distância entre objetos simbólicos; bem como Silva (2005) e Billard e Diday (2006), que propuseram as principais medidas de distância para objetos simbólicos booleanos e modais. Para dados simbólicos de natureza intervalar, Bock (2002) apresentou métodos de partição e visualização mediante os mapas de $Kohone$. Carvalho e Souza (2003) desenvolveram novos métodos de $cluster$ utilizando algoritmos do tipo nuvens dinâmicas. Souza e Carvalho (2004) introduziram métodos de partição

baseados na distância $city-block$. Carvalho et al. (2006) propuseram um método dinâmico de partição baseado na distância de $Hausdorff$ e Carvalho et al. (2007) propuseram o agrupamento de dados simbólicos intervalares também baseado na distância $Hausdorff$ adaptada.

No contexto da análise fatorial para dados simbólicos a primeira abordagem foi apresentada por Cazes et al. (1997). Eles introduziram um método geométrico de classificação não supervisionado em que indivíduos são descritos por vetores de intervalos numéricos. Morineau et al. (1994) também apresentaram contribuições nesta área. Logo, foi proposta uma generalização da análise fatorial discriminante para dados simbólicos (LAURO; VERDE; PALUMBO, 2000). Uma extensão da tabela bidimensional foi proposta por Gettler-Summa e Pardoux (2000), em que os autores abordaram a análise de dados simbólicos em tabelas com três entradas, sendo o tempo ou espaço a terceira dimensão.

As árvores de decisão também foram estendidas a dados simbólicos, exemplo disso foi a generalização dos conceitos desta técnica não-paramétrica por Ciampi et al. (2000). Adicionalmente, Llatas e M. (2000) estudou o uso de árvores de decisão considerando que os objetos simbólicos fornecem uma amostra estratificada. Segundo eles, isto permite detectar a influência dos estratos nas regras de predição. Mballo e Diday (2005) propuseram também a utilização do critério de $Kolmogorov-Smirnov$ como medida em árvores de decisão simbólica.

Os modelos de regressão também têm sido estendidos a dados simbólicos, Billard e Diday (2000) foram os primeiros a propor um modelo de regressão para dados simbólicos de natureza intervalar. A abordagem proposta por eles consiste em minimizar a soma dos quadrados dos erros para os centros dos intervalos. Dois anos mais tarde, os mesmos autores propõem uma outra abordagem ajustando dois Modelo de Regressão Linear Clássico (MRLC) independentes para os limites inferiores e superiores dos intervalos (BILLARD; DIDAY, 2002). Billard e Diday (2006) também incluíram variáveis explicativas, bem como estruturas hierárquicas das variáveis no âmbito da regressão simbólica. Maia e Carvalho (2008) estenderam o modelo de regressão L_1 para dados simbólicos intervalares, considerando a soma dos desvios absolutos como critério de minimização para a estimativa dos parâmetros. Neto e Carvalho (2008) propuseram um novo modelo para dados intervalares baseado no centro e na amplitude dos intervalos, representação que mostrou melhor desempenho do que os métodos apresentados em Billard e Diday (2000) e Billard e Diday (2002). Neto e Carvalho (2010) propuseram uma nova abordagem para ajustar o modelo de regressão linear com restrição no centro e nas amplitudes dos intervalos, a fim de assegurar a coerência matemática entre os valores previstos dos limites inferior e superior do intervalo. Uma melhoria sobre os métodos existentes foi proposta por Souza et

al. (2017). Os autores apresentam o método parametrizado de regressão linear para variáveis simbólicas de tipo intervalo. Neste modelo, propõe-se a representação dos intervalos através da equação paramétrica da reta. Dois modelos são propostos para a estimativa dos limites da variável resposta. Esta parametrização permitem o ajuste dos pontos nas variáveis regressoras que dão as melhores estimativas para os limites da variável resposta.

No caso do modelo de regressão intervalar que assume distribuições de probabilidade para os erros, Domingues, Souza e Cysneiros (2010) propuseram uma metodologia de análise de dados intervalares utilizando como base o modelo de regressão linear simétrica. Baseados na teoria do modelo linear generalizado, Neto, Cordeiro e Carvalho (2011) introduziram um modelo de regressão bivariada simbólica para dados intervalares. Souza, Queiroz e Cysneiros (2011) propuseram modelos de regressão linear logística para os limites inferiores e superiores dos intervalos, em conjunto e separadamente. Fagundes, Souza e Cysneiros (2013) introduziram um modelo de regressão robusta para a estimativa e a predição de intervalos na presença de $outliers$. Adicionalmente, outra proposta interessante é apresentada por Neto e Anjos (2015), os autores neste trabalho consideram um modelo de regressão para dados de tipo intervalo com estrutura de cópulas obtendo resultados relevantes quando comparado com outros modelos da literatura.

Recentemente, Neto e Carvalho (2017) introduziram um modelo de regressão não linear para dados de tipo intervalo. Hao e Guo (2017) propuseram um novo modelo de regressão para o centro e rango dos intervalos que adicionou várias restrições não negativas para manter a coerência matemática. Reyes, Souza e Cysneiros (2017) propuseram um único modelo de regressão não linear simétrica para ajustar dados de tipo intervalo. Uma característica importante deste novo modelo é que a estimativa e a previsão são menos sensíveis a valores discrepantes. Outro trabalho interessante é o apresentado por Neto e Carvalho (2018), eles propuseram ajustar uma regressão linear robusta usando um $kernel$ exponencial para penalizar os $outliers$ no centro e no range dos intervalos, as estimativas são baseadas no método dos mínimos quadrados. Nos últimos anos, Zhang, Beranger e Sisson (2020) propuseram um modelo de geração de dados para construir a função de verossimilhança com variáveis de tipo intervalo.

No contexto de modelos de regressão para outras variáveis simbólicas, Billard e Diday (2006) propuseram os primeiros modelos de regressão linear para variáveis com valores de histograma. Esses modelos são uma extensão dos modelos para variáveis intervalares propostos pelos mesmos autores. Outro modelo alternativo foi proposto por Verde e Irpino (2010), Verde

e Irpino (2015), os autores tiveram em conta todas as distribuições que são representadas por funções quartis. O modelo proposto se baseia na exploração das propriedades de uma decomposição da distância de Mallows (IRPINO; ROMANO, 2007) (que os autores denominam distância de Wasserstein). Dias e Brito (2015) também propuseram um modelo de regressão linear para variáveis simbólicas de histogramas, que considera dados com variabilidade e permite prever valores de histogramas sem forçar uma relação linear direta. Silva, Souza e Cysneiros (2018) introduziram um novo tipo de dados simbólicos, chamados de dados poligonais simbólicos e um modelo de regressão linear é proposto para este tipo de dados.

Por último, no domínio de séries temporais, Arroyo e Maté (2006) fornecem medidas de precisão para séries temporais intervalares baseadas em distâncias de $Ichino - Yaguchi$ e $Hausdorff$. Maia, Carvalho e Ludermir (2006) apresentaram duas abordagens para a previsão de series temporais considerando variáveis simbólicas intervalares. O primeiro método ajusta dois modelos independentes (*Auto-regressive Moving Average Models* (ARMA)) sobre os centros e as amplitudes dos intervalos. O segundo método baseia-se em uma abordagem híbrida e combina um modelo ARMA com uma rede neural *Multilayer Perceptron* (MLP) (MAIA; CARVALHO; LUDEMIR, 2008). Maté e Arroyo (2009), formularam novos modelos para a previsão de séries temporais simbólicas com dados de tipo histograma. Por fim, García-Ascanio e Maté (2010) promovem uma comparação entre modelos de Vetores Auto-Regressivos (VAR) e MLP para dados de tipo intervalo na previsão da demanda de energia elétrica.

2.1.1 Dados Simbólicos

Os dados simbólicos são extensões de tipos de dados clássicos. Em conjuntos de dados convencionais, os objetos são individualizados, enquanto em dados simbólicos estes são unificados por relacionamentos. Em geral, dados simbólicos são mais complexos do que os dados convencionais nos seguintes aspectos (GOWDA; DIDAY, 1991; GOWDA; RAVI, 1995):

1. Todos os indivíduos de um conjunto de dados simbólicos podem ou não ser definidos pelas mesmas variáveis.

2. Cada variável pode ter mais do que um valor ou mesmo um intervalo de valores.

3. Em dados simbólicos complexos, os valores que as variáveis adquirem podem incluir um ou mais objetos elementares.

4. A descrição de um dado simbólico pode depender das relações existentes entre outros dados.

5. Os valores das variáveis simbólicas podem ser tipicamente frequências de ocorrência relativa ou indicar semelhança e nível de importância de outros valores.

A característica fundamental dos dados simbólicos é que eles permitem a descrição de elementos ou fenômenos onde há uma variabilidade interna. Essa variabilidade surge naturalmente ao agregar observações individuais em classes. Agregação significa a coleção de observações que satisfazem um requisito que permite que elas sejam agrupadas, sendo que é possível distinguir dois tipos de agregação (BRITO, 2014):

- **Agregação Temporal:** aplica-se quando os dados são recolhidos e observados ao longo de um determinado período de tempo para o mesmo indivíduo ou entidade (por exemplo um dia). Um exemplo de uma agregação temporal é o registro horário dos batimentos cardíacos de um paciente, sendo que estes devem ser agregados ao nível do dia, para cada um dos pacientes.

- **Agregação Contemporânea:** aplica-se quando os dados são recolhidos no mesmo tempo e a unidade estatística de interesse encontra-se a um nível superior ao das observações. As novas classes são constituídas a partir de agregações dos valores individuais segundo características específicas. Um exemplo de uma agregação contemporânea é efetuada nos censos, onde os indivíduos são agregados pela respectiva região.

A agregação permite resumir grande bases de dados sem perda significativa de informação e descobrir novos tipos de conhecimento complementar que não estão disponíveis no nível individual. Permite também reduzir o número de indivíduos e de dados faltantes. Resolve as questões de confidencialidade dos dados e facilita a interpretação do resultados. Além disso, permite distinguir entre observações de primeiro nível e segundo nível. O primeiro nível representa os indivíduos e o segundo nível representa os grupos. Também é possível que existam observações de terceiro nível ou níveis ainda mais altos. Esse pode ser o caso dos dados coletados de um país, por exemplo, onde os dados individuais podem ser agregados por cidades e logo por regiões, e depois analisar os dados entre as diferentes regiões.

Dados simbólicos permitem representar observações de segundo nível e nível superior. Dados simbólicos também podem representar dependências lógicas, taxonômicas ou hierárquicas. Uma dependência lógica pode ser, por exemplo, uma regra que é adicionada para manter a

integridade ao agregar dados. As taxonomias permitem representar variáveis na forma de uma árvore invertida onde cada nível representa um nível de generalidade: as folhas representam o nível mais baixo e a raiz representa o mais alto. Por outro lado, as hierarquias permitem estabelecer relações mãe-filha entre variáveis, de modo que uma variável filha é apenas operativa dependendo do resultado da variável mãe no nível superior (DIDAY, 2016).

2.1.2 Tabela de Dados Simbólicos

A premissa é que o processo de obtenção de dados simbólicos deve preservar o máximo de informação possível sobre os dados e ao mesmo tempo diminuir consideravelmente o tamanho inicial da tabela de dados. Como resultado dessa transformação são geradas novas tabelas de dados, chamadas de tabelas de dados simbólicos, onde as classes de indivíduos são descritas por pelo menos uma variável simbólica. Assim, nestas tabelas, as linhas correspondem aos indivíduos ou classes de indivíduos e as colunas são as variáveis simbólicas que descrevem esses indivíduos ou classes (DIDAY, 2016).

Considere o seguinte exemplo extraído de Diday e Noirhomme-Fraiture (2008): em uma ilha vivem 600 pássaros, sendo 400 andorinhas, 100 avestruzes e 100 pinguins. A Tabela (1) consiste de 600 entradas com a informação referente à espécie, capacidade de voo e tamanho para cada um dos pássaros observados na ilha. Entretanto a Tabela (2) mostra, em apenas 3 entradas, os dados simbólicos obtidos pelo processo de ADS agrupando as aves por espécie, na qual também foi adicionada uma nova informação referente à migração dos pássaros em diferentes períodos do ano.

Tabela 1 – Descrição simplificada da base de dados de 600 pássaros com 3 variáveis.

Pássaro	Espécie	Voadora	Tamanho (cm)
1	Pinguim	{NÃO}	80
⋮	⋮	⋮	⋮
599	Andorinha	{SIM}	70
600	Avestruz	{NÃO}	125

Fonte: Elaborada pelo autor (2021)

Na Tabela (2), notamos que as variáveis do conjunto de dados original foram transformadas para variáveis simbólicas. Uma nova variável para representar a migração das aves foi adicionada. Dita variável expressa que 90% das andorinhas migram, que todos os pinguins

migram e que nenhum avestruz migra.

Tabela 2 – Descrição das 3 espécies de aves com conceito de migração.

Espécie	Voadora	Tamanho (cm)	Migração
Pinguim	{NÃO}	[70; 95]	[100% sim, 0% não]
Andorinha	{SIM}	[60; 85]	[90% sim, 10% não]
Avestruz	{NÃO}	[85; 160]	[0% sim, 100% não]

Fonte: Elaborada pelo autor (2021)

Outro exemplo de tabela de dados simbólicos é apresentado na Tabela (3), onde as linhas são indivíduos e as colunas são três variáveis simbólicas: pulso (expresso por um intervalo), marca de automóvel (expresso por um conjunto de categorias) e se faz academia (expresso por uma distribuição de pesos), também obtidas de uma agregação pelo processo ADS.

Tabela 3 – Exemplo de uma tabela de dados simbólicos.

ID	Pulso	Marca Automóvel	Faz academia
1	[58; 90]	{Ford, Fiat}	{(3/4)sim, (1/4)não}
2	[47; 68]	{Ford, Fiat, BMW}	{(1/4)sim, (5/3)não}
3	[32; 114]	{Volkswagen, Chevrolet}	{(4/5)sim, (1/2)não}

Fonte: Elaborada pelo autor (2021)

Por outro lado, a Tabela (4) apresenta dados "naturalmente" simbólicos de tipo intervalo, que são os dados das temperaturas mensais mínimas e máximas registradas em 60 estações meteorológicas na China no ano 1988.

Tabela 4 – Exemplo de uma tabela simbólica com dados "naturalmente" intervalares.

Estações	Janeiro	Fevereirol	$\cdots$	Novembro	Dezembro
AnQing	[1,8; 7,1]	[2,1; 7,2]	$\cdots$	[7,8; 17,9]	[4,3; 11,8]
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
ZhoJing	[2,7; 8,4]	[2,7; 8,7]	$\cdots$	[8,2; 20]	[5,1; 13,3]

Fonte: (BILLARD; DIDAY, 2006, p. 203)

Assim, podemos resumir que cada célula de uma tabela de dados simbólicos pode conter diferentes tipos de dados, em particular (BOCK; DIDAY, 2000):

a) Um único valor quantitativo.

b) Um único valor categórico.

c) Um conjunto de valores ou categorias.

d) Um intervalo.

e) Um conjunto de valores com pesos associados.

2.1.3 Varáveis Simbólicas

Como foi mencionado anteriormente, as variáveis simbólicas podem assumir uma distribuição, um intervalo, um histograma, uma lista de categorias ou números, etc. (BOCK; DIDAY, 2000; BILLARD; DIDAY, 2006; DIDAY; NOIRHOMME-FRAITURE, 2008). A ideia principal é que as variáveis simbólicas são capazes de assumir a variabilidade interna dos indivíduos ou classes de indivíduos, descrevendo unidades de interesse. As classes de indivíduos podem ser obtidas resumindo grandes conjuntos de dados e elas são consideradas como novas unidades de nível mais alto de generalização do que indivíduos. As razões são:

1. Extrair novos e complementares tipos de conhecimento não disponíveis ao nível dos indivíduos;
2. Estudar os dados por unidades dadas no nível necessário de generalização;
3. Resumir os dados sem perda significativa de informação;
4. Resumir dados perdidos e $outliers$; e manter a confidencialidade;
5. Facilitar a interpretação dos resultados.

Os tipos de variáveis simbólicas mais comuns são: variáveis multi-valoradas, variáveis de tipo intervalo e variáveis modais.

- Uma variável simbólica **Multi-valorada** Y expressa a variabilidade interna de uma classe de indivíduos por uma lista de números, categorias ou intervalos. Esses tipos de variáveis podem ser:
  - **Multi-valorada não-ordinal** se seus valores $Y(i)$ correspondem a subconjuntos finitos do domínio $D : |Y(i)| < \infty$ para todos os indivíduos $i \in E$. Exemplo, seja E o conjunto de cidades no Brasil e Y a variável que armazena os bancos das cidades. Então pode-se ter que, $Y(Recife) = \{Bradesco, Caixa, Citibank, BB\}$.

  - **Multi-valorada ordinal** se D suporta uma relação de ordem $\prec$, tal que, para quaisquer pares de elementos $a, b \in D$, temos que $a \prec b$ ou $b \prec a$. Na prática, $a \prec b$ é interpretado como a antecede b ou a é menor que b. Para quaisquer dois indivíduos $i, j \in E$, em que $a = Y(i)$ e $b = Y(j)$ são os valores observados para a variável Y, é possível definir qual deles é estritamente "melhor" do que o outro sem a utilização de qualquer escala numérica: $a \prec b$ ou $b \prec a$. Um exemplo desse tipo de variável é, $Y = \{$*Qualidade do produto*$\}$ e $D = \{$*excelente*, *bom*, *razoável*, *pobre*, *insuficiente*$\}$.

- Uma variável simbólica Y é definida como **intervalar** se ela representa uma realização $\xi = [a; b] \subset R^1$, com $a \leq b$ e $\{a, b\} \in R^1$. Por exemplo, seja E um grupo de homens e $Y =$ o tempo semanal de lazer (em horas), para os indivíduos $i, j \in E$ é possível ter: $Y(i) = [3; 5]$ e $Y(j) = [7; 9]$.

- Todas as variáveis definidas anteriormente são também conhecidas como variáveis simbólicas *booleanas*. Existem também as variáveis modais. Variáveis modais são variáveis de multi-estado com uma frequência, probabilidade ou peso anexado a cada um dos valores específicos nos dados. Ou seja, uma variável simbólica Y é definida como **modal** se para cada indivíduo $i \in E$, essa variável além de apresentar um subconjunto de categorias $Y(i) \subseteq D$, apresenta também uma frequência, um histograma, uma função de distribuição empírica, uma distribuição de probabilidade, um modelo, um peso ou algo assim associado a cada categoria $l \in Y(i)$. Por exemplo, se três dos irmãos de um indivíduo i são diabéticos e um não, a variável que descreve a propensão ao diabetes do individuo i pode ser uma variável modal representada por $Y(i) = 3/4$ diabetes, $1/4$ não diabetes.

As variáveis simbólicas de tipo $boxplots$ foram inicialmente definidas em Arroyo, Maté e Roque (2008) como um novo tipo de variável simbólica modal e aplicadas a problemas de agrupamento hierárquico em relação a diferentes medidas de distância para $boxplots$. Arroyo, Maté e Roque (2008) considera variáveis de $boxplots$ como um caso especial de variáveis modais com valores de intervalo. Drago (2015) assume que as variáveis simbólicas de tipo $boxplots$ são descritas por cinco variáveis quantitativas clássicas $\{m, Q1, Q2, Q3, M\}$ em que, m e M são variáveis clássicas associadas aos valores mais baixo e mais alto, respectivamente, de uma classe.

Uma variável simbólica de tipo $boxplot$ pode ser definida como $X = [m, q_1, q_2, q_3, M] \subset \mathbb{R}$, com $m \leq q_1 \leq q_2 \leq q_3 \leq M$, em que m representa o mínimo, q_1 o primeiro quartil, q_2 a mediana, q_3 o terceiro quartil e M o máximo dos valores de X. No contexto de ADS, X pode ser visto como uma lista de dados multi-valorados de quartis e o analista pode querer estudar um único quartil, dois quartis ou todos os quartis. Os quartis também são frequentemente usados como uma medida de dispersão dos dados que é chamado de intervalo interquartil: $IQR = q_3 - q_1$.

2.2 REGRESSÃO LINEAR PARA INTERVALOS

Regressão linear é um método estatístico que examina as relações entre duas ou mais variáveis. Dois tipos de variáveis estão envolvidos: a variável resposta (ou dependente) e as variáveis regressoras (ou independentes). Cada variável assume um único valor (dados quantitativos)/categorias (dados qualitativos) para cada unidade do conjunto de dados. O objetivo principal é encontrar uma equação linear que represente a variável resposta com base nos regressores. Essa modelagem é simples e não aborda a variabilidade e a incerteza presentes nos dados (RENCHER; SCHAALJE, 2008; MONTGOMERY; PECK; VINING, 2001; DRAPER; SMITH, 1998; SEBER; LEE, 2012). A metodologia de regressão linear é aplicada em diversas áreas de pesquisa, como a financeira, a epidemiológica, a médica, a econômica, etc.

Vários modelos de regressão para dados simbólicos do tipo intervalo têm sido introduzidos na literatura. A maioria desses modelos usam a minimização da soma dos quadrados dos desvios para estimar os parâmetros. Este método dos mínimos quadrados ordinários (*Ordinary Least Squares* (OLS)) tem a vantagem de ser computacionalmente simples e de fornecer os melhores estimadores lineares não-viesados para os parâmetros do modelo (MONTGOMERY; PECK; VINING, 2015).

2.2.1 Regressão Linear

Os três modelos principais de regressão linear simbólica, encontrados na literatura são o Método do Centro (MC), o Método do Mínimo e o Máximo (MinMax) e o Método do Centro e Range (MCR). O processo para a estimativa dos parâmetros da regressão linear nos três métodos é baseado na minimização de critérios predeterminados usando OLS.

O MC consiste em ajustar um MRLC aos pontos médios (centros) dos intervalos e em

seguida aplicar esse modelo aos limites inferior e superior dos intervalos das variáveis preditoras para prever, respectivamente, os limites inferior e superior dos intervalos da variável resposta (BILLARD; DIDAY, 2000). O centro é dado por:

$$\varepsilon_i^c = \frac{(\varepsilon_i^{inf} + \varepsilon_i^{sup})}{2} \tag{2.1}$$

Os limites inferior e superior da variável resposta são preditos através da aplicação do vetor de parâmetros β aos limites inferiores e superiores das variáveis regressoras. O vetor β é o mesmo para os modelos aplicados a ambos os limites inferiores e superiores.

O MinMax proposto por Billard e Diday (2002) ajusta dois MRLC independentes para os limites inferiores e superiores das variáveis simbólicas e essa é a sua grande diferença quando comparado com o MC. Considere o conjunto de variáveis $X_1, X_2, \cdots, X_p$ como variáveis regressoras relacionadas com uma variável resposta Y através do modelo linear:

$$\begin{aligned} y_i^{inf} &= \beta_0^{inf} + \beta_1^{inf} a_{i1} + \cdots + \beta_p^{inf} a_{ip} + \varepsilon_i^{inf} \\ y_i^{sup} &= \beta_0^{sup} + \beta_1^{sup} b_{i1} + \cdots + \beta_p^{sup} b_{ip} + \varepsilon_i^{sup} \end{aligned} \tag{2.2}$$

Os valores preditos para os limites inferior e superior $\hat{y} = [\hat{y}^{inf}; \hat{y}^{sup}]$ da variável resposta Y depois de aplicar o modelo são dados por:

$$\hat{y}^{inf} = (\mathbf{x}^{inf})^T \hat{\beta}^{inf} \quad \text{e} \quad \hat{y}^{sup} = (\mathbf{x}^{sup})^T \hat{\beta}^{sup}$$

com

$$(\mathbf{x}^{inf})^T = (1, a_1, \cdots, a_p) \quad \text{e} \quad (\mathbf{x}^{sup})^T = (1, b_1, \cdots, b_p)$$

$$\hat{\beta}^{inf} = \left(\hat{\beta}_0^{inf}, \hat{\beta}_1^{inf}, \cdots, \hat{\beta}_p^{inf}\right)^T \quad \text{e} \quad \hat{\beta}^{sup} = \left(\hat{\beta}_0^{sup}, \hat{\beta}_1^{sup}, \cdots, \hat{\beta}_p^{sup}\right)^T$$

O MCR foi proposto por Neto e Carvalho (2008), e estabelece as somas dos quadrados dos erros relativos aos centros e amplitudes dos intervalos como critérios de minimização independentes para a estimativa dos parâmetros, obtendo-se um modelo para o centro e outro para a amplitude. A expectativa é que com a inclusão das informações contidas nas amplitudes dos intervalos melhore a predição do modelo. O ajuste dos limites inferior e superior da variável resposta é realizado através da aplicação do vetor de parâmetros β aos centros e amplitudes das variáveis regressoras.

Sejam, y^c e x_j^c (j = 1,2,...,p) os vetores de valores relativos aos centros (c) dos intervalos das variáveis intervalares y e x_j (j = 1,2,...,p). Além disso, considere y^r e x_j^r (j = 1,2,...,p) variáveis quantitativas que assumem como valores a metade das amplitudes (r) dos intervalos

das variáveis intervalares y e x_j (j = 1,2,...,p). Considere y^c e y^r como variáveis resposta e x_j^c e x_j^r (j = 1,2,...,p) um conjunto de variáveis regressoras relacionadas por:

$$\begin{aligned} y_i^c &= \beta_0^c + \beta_1^c x_{i1}^c + \cdots + \beta_p^c x_{ip}^c + \varepsilon_t^c \qquad i = 1, \ldots, n \\ y_i^r &= \beta_0^r + \beta_1^r x_{i1}^r + \cdots + \beta_p^r x_{ip}^r + \varepsilon_i^r \qquad t = 1, \ldots, n \end{aligned} \tag{2.3}$$

em que,

$$x_{ij}^c = \frac{(a_{ij} + b_{ij})}{2} \quad \text{e} \quad x_{ij}^r = \frac{(b_{ij} - a_{ij})}{2}$$
$$y_i^c = \frac{(y_i^{inf} + y_i^{sup})}{2} \quad \text{e} \quad y_i^r = \frac{(y_i^{sup} - y_i^{inf})}{2}$$

O valor $y = [y^{inf}, y^{sup}]$ é predito a partir dos valores $\hat{y}^c$ e $\hat{y}^r$ como mostrado a seguir:

$$\hat{y}^{inf} = \hat{y}^c - \hat{y}^r \quad e \quad \hat{y}^{sup} = \hat{y}^c + \hat{y}^r$$

onde, $\hat{y}^c = (\mathbf{x}^c)^T \hat{\beta}^c$, $\hat{y}^r = (\mathbf{x}^r)^T \hat{\beta}^r$, $(\mathbf{x}^c)^T = (1, x_1^c, \cdots, x_p^c)$, $(\mathbf{x}^r)^T = (1, x_1^r, \cdots, x_p^r)$, $\hat{\beta}^c = (\hat{\beta}_0^c, \hat{\beta}_1^c, \cdots, \hat{\beta}_p^c)^T$ e $\hat{\beta}^r = (\hat{\beta}_0^r, \hat{\beta}_1^r, \cdots, \hat{\beta}_p^r)^T$.

2.2.2 Regressão Linear Paramétrica

Souza et al. (2017) propõem um novo método de regressão linear por meio de uma representação paramétrica para intervalos. Nessa abordagem, os autores propõem a utilização de transformações para intervalos como mecanismo de auxílio para garantir a coerência matemática na predição.

Dado um intervalo $\gamma = [\underline{\gamma}, \overline{\gamma}]$, com $\underline{\gamma} \leq \overline{\gamma}]$, um ponto q que se encontra dentro de γ, pode ser determinado pela equação parametrizada da reta, apresentada na equação 2.4

$$q(\lambda) = \underline{\gamma}(1 - \lambda) + \overline{\gamma}\lambda \tag{2.4}$$

com $0 \leq \lambda \leq 1$ (MCCREA, 2012). Fixando um valor para λ, um intervalo se reduz a um ponto. Os limites do intervalo são obtidos com $\lambda = 0$ e $\lambda = 1$, tais que, $q(0) = \underline{\gamma}$ e $q(1) = \overline{\gamma}$. O centro do intervalo é obtido quando $\lambda = 0,5$

Considere que para cada variável regressora X_j, para $j = 1; \cdots; p$, seja fixado um valor λ_j, para determinar pontos nas amostras desta variável, através da parametrização intervalar. Desta forma, para os intervalos $x_{ji} (i = 1; \cdots; n)$, definem-se os pontos parametrizados q_{ji}, como dado na equação 2.5, baseada na parametrização intervalar definida na equação 2.4,

$$q_{ji} = \underline{x}_{ji}(1 - \lambda_j) + \overline{x}_{ji}\lambda_j \tag{2.5}$$

Propõe-se a modelagem do limite inferior da resposta baseando-se nestes pontos parametrizados, como mostrado na equação 2.6

$$\underline{x}_i = \beta_0^{inf} + \sum_{j=1}^{p} \beta_j^{inf} q_{ij} + \varepsilon_i^{inf}, \tag{2.6}$$

em que $\beta_j^{inf}(j = 0; \cdots; p)$ são os coeficientes desconhecidos do modelo e $\varepsilon_i^{inf}(i = 1; \cdots; n)$ são os erros.

Os valores de λ_j definidos na parametrização não precisam ser fixados previamente. Eles são implicitamente determinados quando o método dos mínimos quadrados é aplicado. Assim, o modelo determina, automaticamente, os pontos nos regressores que garantem o melhor ajuste. Fixando a variável X_j e sabendo que $\alpha_j^{inf} = \beta_j^{inf}(1 - \lambda_j)$ e $\omega_j^{inf} = \beta_j^{inf}\lambda_j$, o valor λ_j pode ser determinado pela equação 2.7

$$\lambda_j = \frac{\omega_j^{inf}}{\alpha_j^{inf} + \omega_j^{inf}}. \tag{2.7}$$

Para o limite superior da variável resposta aplica-se o mesmo modelo baseado na parametrização da reta. Para mais detalhes ver o trabalho de Souza et al. (2017).

2.3 SÉRIES TEMPORAIS SIMBÓLICAS

As origens dos estudos das séries temporais remontam ao início do século XIX, quando Laplace realizou um estudo sobre o efeito das fases da lua nas marés e os movimentos do ar na Terra. Neste período, são essenciais as contribuições de Fourier, que pesquisou as funções periódicas; de Yule, que propôs em 1927 os processos auto-regressivos para explicar as manchas solares (YULE, 1927); e de Slutsky que estudou os processos de médias móveis para representar ciclos econômicos (SLUTZKY, 1937).

Muitos outros pesquisadores, como Kolmogorov, Wiener, Cramer, Bartlett e Tukey, fizeram contribuições notáveis durante a primeira metade do século XX. Mas foi na segunda metade do século XX que se consolidou área de previsão de séries temporais, com três trabalhos fundamentais: o desenvolvimento, por Holt e Winters, de métodos de previsão baseados em suavização exponencial (HOLT, 1957; WINTERS, 1960); A proposta de Box e Jenkins de uma metodologia unificada para prever séries estacionárias e não estacionárias (com e sem sazonalidade) que dá origem aos famosos modelos ARIMA (BOX; JENKINS, 1970); e o desenvolvimento de Kalman de um procedimento para estimar variáveis de estado e antecipar observações futuras em sistemas lineares (KALMAN, 1960).

Essas contribuições fazem a previsão de séries temporais entrar em uma fase de maturidade e surge assim o *International Institute of Forecasters* (IFF), promotor de fóruns e periódicos de previsão, como o *Journal of Forecasting* (em 1982), o *International Journal of Forecasting* (em 1985) e o *International Symposium on Forecasting*, uma conferência anual de especialistas sobre o assunto realizada desde 1982. Outro evento importante ocorreu em 2003 com a atribuição do Prêmio Nobel de Economia para Clive Granger e Robert Engle, que reconheceu o enorme impacto de seu trabalho sobre a co-integração de séries temporais na teoria econômica (ENGLE; GRANGER, 1987). Em 2006, o IFF comemorou seu 25º aniversário e fez uma retrospectiva dos avanços mais significativos na área durante esse período, que estão incluídos na edição especial da revista *International Journal of Forecasting* (GOOIJER; HYNDMAN, 2006).

Uma série temporal é o resultado da observação dos valores de uma variável ao longo do tempo em intervalos regulares (diariamente, mensalmente, etc.) (BOX; JENKINS, 1970). Para formular uma definição formal de séries temporais, é requerido o conceito de processo estocástico. Um processo estocástico é um conjunto de variáveis aleatórias X_t onde o índice t recebe valores em um conjunto T de instantes temporários para os quais o processo está definido. Formalmente, é denotado como

$$\{X_t\}\ tal\ que\ t \subset T\ com\ T \subseteq \mathbb{R}. \tag{2.8}$$

Cada uma das variáveis aleatórias do processo segue sua própria função de distribuição de probabilidade.

Nos processos estocásticos que consideraremos, T é discreto, os valores $t \subset T$ são uniformemente espaçados ao longo do tempo e o valor observado da variável aleatória X_t em t, que é denotado como x_t, não depende em nenhum caso de valores futuros. Uma série temporal x_t é a realização finita de um processo estocástico X_t dessas características.

Em séries temporais, apenas uma modalidade do processo estocástico a ser estudado está usualmente disponível, isto é, a série temporal x_t é geralmente a única observação disponível de X_t. Este fato, complica sua análise e a faz diferente de outras áreas da estatística onde há um conjunto de realizações das variáveis aleatórias de interesse. Ter uma única realização do processo estocástico subjacente torna mais difícil determinar as propriedades desse processo, que é precisamente o objetivo da análise de séries temporais. Para estimar essas propriedades, é necessário que a série seja estacionária.

Estacionariedade: Uma série temporal é estacionária quando suas características estatísticas como média, variância, autocorrelação, são constantes ao longo do tempo. É uma série

que se desenvolve aleatoriamente no tempo, em torno de uma média constante (CHATFIELD, 2001).

Série estritamente estacionária: Uma série temporal estritamente estacionária é aquela para a qual o comportamento probabilístico de cada coleção de valores $x_{t1}, x_{t2}, \cdots, x_{tk}$ é idêntico ao do conjunto de deslocamentos temporais $x_{t1+h}, x_{t2+h}, \cdots, x_{tk+h}$, para quaisquer $t_1, \cdots, t_k, h$ de T (PEÑA, 2005).

Série fracamente estacionária: Uma série temporal fracamente estacionária é um processo de variância finita tal que (PEÑA, 2005):

(i) a função de média μ_t, é constante e não depende do tempo t, e

(ii) a função de autocovariância $\gamma(s,t)$, depende de s e t somente através de sua diferença $|s-t|$.

De acordo com o conceito clássico de séries temporais, as observações de uma série são valores pontuais, isto é, cada instante temporal t é descrito por um único valor x_t da variável X_t. Estes valores pontuais não são capazes de representar a variabilidade da observação no tempo t. Isso não é necessário em muitos casos, mas existem outros, onde é aconselhável ser capaz de capturar a variabilidade da observação de alguma forma. Para os casos em que a variabilidade é importante, esta pesquisa propõe a utilização de séries temporais onde cada observação é descrita por tipo especial de dados simbólicos, chamados dados simbólicos multi-valorados.

As séries temporais simbólicas, ou seja, as séries temporais em que a variável observada ao longo do tempo é uma variável simbólica, permitem representar a variabilidade que ocorre em cada instante temporal. Essa variabilidade pode ser representada, por exemplo, por um $boxplot$, um histograma ou um intervalo. A fim de propor uma definição mais formal de séries temporais simbólicas, o conceito de variável simbólica aleatória e processo estocástico simbólico deve primeiro ser definido.

Variável simbólica aleatória: Uma variável aleatória simbólica é uma função que atribui um valor simbólico a cada elemento do espaço amostral. No caso das variáveis simbólicas multi-valoradas de tipo quartis, cada elemento do espaço amostral será associado a uma lista de quartis e representado por um $boxplots$. O leitor pode ver as definições específicas de cada variável aleatória simbólica nas Subseções (2.1.3) e (**??**).

Processo estocástico simbólico: Um processo estocástico simbólico é definido por um conjunto de variáveis aleatórias simbólicas $\{X_t\}$, onde cada variável X_t é indexada por um

índice t, tal que $t \subset T$ com $T \subseteq \mathbb{R}$, ou seja, T denota os instantes para os quais o processo é definido.

Série temporal simbólica: Uma série temporal simbólica é a realização de um processo estocástico simbólico. Ou seja, é uma amostra do tamanho um do vetor de variáveis aleatórias simbólicas que caracteriza o processo estocástico simbólico.

A maioria das contribuições que apareceram ao longo desses anos dedicadas a séries temporais simbólicas não aborda a questão sob a perspectiva de processos simbólicos estocásticos, mas assume uma perspectiva mais pragmática, concentrando-se unicamente na previsão da série. Esta pesquisa também adotará essa perspectiva.

Teles e Brito (2005) expõem uma primeira aproximação à previsão das séries temporais simbólicas intervalares baseada nos modelos ARMA. Nesta abordagem os autores modelam a séries dos mínimos e a série dos máximos de maneira independente por modelos ARMA com os mesmos parâmetros mas com diferentes variáveis. Pouco depois, Arroyo et al. (2007) adaptam métodos de suavização exponencial para séries temporais de intervalos com a ajuda da aritmética intervalar.

Entretanto, Maia, Carvalho e Ludemir (2006) e Maia, Carvalho e Ludemir (2008) tratam os conceitos de séries temporais intervalares mediante a descomposição da serie temporal intervalar em uma série de centros e outra de ranges e ajustam para cada uma delas um modelo ARMA ou um modelo híbrido que combina ARMA e perceptron multi-capas. Maté e Arroyo (2009) e Arroyo, Gonzáles-Rivera e Maté (2011) também apresentaram contribuições nesta área adaptando algoritmos clássicos, como filtros de suavização e *k-Neighbors Neighbours* (K-NN) não paramétrico às séries temporais de intervalos e histogramas. O K-NN proposto baseia-se na escolha de uma distância que é usada para medir as diferenças entre as sequências de histogramas e calcular as previsões.

Além disso, encontramos na literatura algumas aplicações interessantes como a proposta de Cheun (2007), que apresenta um modelo empírico para valores máximos e mínimos diários para as ações da bolsa de valores dos EUA usando o *Vector Error Correction Model* (VECMN) e a proposta de Han et al. (2008), que propõe um modelo linear de intervalo para investigar as relações dinâmicas entre processos intervalares.

Em relação às séries temporais simbólicas de $boxplots$, apenas o trabalho de Drago (2015) propõe um método para a previsão dessas séries simbólicas. Na abordagem proposta, o autor considera a representação das classes por $boxplots$, que são descritos por cinco variáveis quantitativas clássicas $\{m, Q^1, Q^2, Q^3, M\}$ em que m e M são variáveis clássicas associadas

aos valores mais baixos e mais altos, respectivamente, de uma classe, Q^1, Q^2, Q^3 são os valores associados aos quartis. Assim, cinco curvas são ajustadas para estimar os $boxplots$.

2.3.1 Séries Temporais de Intervalos

As séries temporais de intervalos permitem representar situações nas quais as observações são afetadas pela variabilidade, o que torna mais apropriado representá-las por meio de um intervalo. Uma série temporal de intervalos, $\{[X]_t\}$, pode ser definida como uma sequência de intervalos de valores que são observados em instantes sucessivos no tempo indicado por $t = 1, \ldots, n$, e em que cada intervalo é representado por

$$\{[X]_t\} = [X_{t,L}, X_{t,U}]$$

em que, $X_{t,L}$ é o limite inferior do intervalo e $X_{t,U}$ o limite superior do intervalo. Alternativamente, o intervalo pode ser representado pelos centros e amplitudes (ARROYO; MATÉ, 2006).

Um intervalo é perfeitamente definido por dois valores: seus limites superior e inferior ou, equivalentemente, seu centro e sua amplitude. De fato, qualquer par retirado de entre estes quatro valores possíveis serve para definir o intervalo, uma vez que permitem obter os outros dois valores, ou seja, a partir do centro e limite inferior, podemos obter a amplitude e limite superior. No entanto, conceitualmente, é mais correto tomar os extremos, porque denotam os limites do intervalo, ou o centro e a amplitude, porque enfatizam em características fundamentais nas estatísticas, como a tendência central (MAIA; CARVALHO; LUDEMIR, 2008).

2.3.2 Séries Temporais de Histogramas

As séries temporais do histograma são uma ferramenta que permite representar séries temporais de distribuições, ou seja, séries nas quais cada instante de tempo é descrito por uma distribuição. Uma série temporal de histogramas $\{h_{X_t}\}$ pode ser definida como uma sequência de distribuições observadas em instantes sucessivos no tempo denotados por $t = 1, \ldots, n$, em que cada distribuição é representada por um histograma h_{X_t} definido como

$$h_{X_t} = \{([I]_{t,1}, \pi_{t,1}), \ldots, ([I]_{t,pt}, \pi_{t,pt})\}, \qquad t = 1, \ldots, n,$$

em que, $\pi_{t,i}$, $i = 1, \ldots, pt$ é uma distribuição de frequência ou de probabilidade no domínio considerado que cumpre que $\pi_{t,i} \geq 0$ e que $\sum_{i=1}^{pt} \pi_{t,i} = 1$. $[I]_{t,1}$ é um intervalo definido como,

$[I]_{t,1} = [\underline{I}_{t,i}, \overline{I}_{t,i})$.

Do ponto de vista de ADS, uma série temporal de histogramas é definida como uma série temporal em que as observações são realizações de variáveis aleatórias simbólicas de histograma. Cada histograma representa a densidade observada em cada instante temporal (MATÉ; ARROYO, 2009).

3 REGRESSÃO LINEAR PARAMETRIZADA PARA DADOS SIMBÓLICOS DE BOXPLOTS

Este capítulo descreve um novo método de regressão linear parametrizada para variáveis simbólicas de tipo $boxplot$. Nesta nova abordagem as variáveis regressoras de tipo $boxplot$ são parametrizadas através da equação da reta. Cinco modelos independentes são ajustados para a estimativa da variável resposta de tipo $boxplot$. Além disso, é proposto um critério para verificar a coerência matemática das previsões do modelo, antes de construir a regressão. Se o critério indicar que a coerência falha, é desenvolvida uma extensão da transformação $Box-Cox$ para dados simbólicos de tipo $boxplots$ que resolve o problema da coerência matemática. Também são apresentados os resultados obtidos através de simulações de Monte Carlo sobre diferentes conjuntos de dados simbólicos multi-valorados reais e sintéticos, considerando dados relevantes e pouco explorados no domínio ADS. Por último, é apresentada uma aplicação do mundo real para previsão de altas temperaturas que dá suporte à identificação de falhas nos equipamentos de uma Usina Termo-eléctrica (UTE) brasileira.

3.1 MÉTODO MRPB

No Método de Regressão Linear Parametrizada para Boxplots (MRPB) introduzido nesta seção, os valores m, q_1, q_2, q_3 e M para representar os $boxplots$, serão, respectivamente, o mínimo, primeiro quartil, mediana, terceiro quartil e o máximo dos $boxplots$

Seja Ω, um conjunto de n elementos e cada elemento é descrito por:

- um conjunto de variáveis regressoras de tipo $boxplot$ p, X, com $\mathbf{x}_i = (x_i^1, x_i^2, ..., x_i^P)$ em que, $x_i^j = \{x_i^{jm}, x_i^{jQ_1}, x_i^{jQ_2}, x_i^{jQ_3}, x_i^{jM}\}$, para $i = 1 \ldots, n$, e
- uma variável resposta de tipo $boxplot$, descrita por Y, como $y_i = \{y_i^m, y_i^{Q_1}, y_i^{Q_2}, y_i^{Q_3}, y_i^M\}$, para $i = 1 \ldots, n$.

De acordo com Souza et al. (2017), um intervalo tem uma representação geométrica como um segmento de reta. Assim, a equação parametrizada da reta pode ser usada para alcançar todos os pontos dentro do intervalo. Essa ideia pode ser estendida para $boxplots$, usando quatro intervalos que estão dentro deles: $[m, q_1]$, $[q_1, q_2]$, $[q_2, q_3]$, $[q_3, M]$. Desta forma, é possível alcançar todos os pontos dentro de um $boxplot$. Dado um intervalo genérico $\gamma = [a, b]$, com $a \leq b$, qualquer ponto p que se encontra dentro de γ, pode ser alcançado usando a Equação

(3.1)

$$p(\lambda) = a(1-\lambda) + b\lambda, \tag{3.1}$$

em que, λ é um parâmetro e $0 \leq \lambda \leq 1$ (SOUZA et al., 2017). Se $\lambda = 0$, temos da equação (3.1) $p = a$. Se $\lambda = 1$, temos $p = b$. E se $\lambda = 0.5$, temos que p é o centro do intervalo. O Método parametrizado (MP) mostra que usando os limites do intervalo como pontos regressores, é possível determinar o melhor valor de λ que será utilizado para produzir o melhor ajuste para a variável resposta.

Considere que para cada variável regressora X_j, para $j = 1, ..., p$, seja fixado um valor λ, para determinar pontos nas amostras desta variável, através da parametrização intervalar em cada um dos intervalos que estão dentro do $boxplots$. Desta forma, para os intervalos γ, definem-se os pontos parametrizados p, como dado na Equação (3.1). Propõe-se a modelagem do limite inferior do intervalo da resposta baseando-se nestes pontos parametrizados, como mostrado na Equação (3.2),

$$y_i^a = \beta_0^a + \sum_{j=1}^{P} \beta_j^a p_{ji} + \varepsilon_i^a \tag{3.2}$$

Então, usando os 4 intervalos que formam cada $boxplot$ do Ω e utilizando as equações doMP para intervalo, propomos modelar a variável resposta com base em:

$$y_i^m = \beta_0^m + \sum_{j=1}^{P} \left(\beta_1^m x_i^{jm} + \beta_1^{Q_1} x_i^{jQ_1} + \beta_1^{Q_2} x_i^{jQ_2} + \beta_1^{Q_3} x_i^{jQ_3} + \beta_1^M x_i^{jM}\right) + \varepsilon_i^m,$$

$$y_i^{Q_1} = \beta_0^{Q_1} + \sum_{j=1}^{P} \left(\beta_2^m x_i^{jm} + \beta_2^{Q_1} x_i^{jQ_1} + \beta_2^{Q_2} x_i^{jQ_2} + \beta_2^{Q_3} x_i^{jQ_3} + \beta_2^M x_i^{jM}\right) + \varepsilon_i^{Q_1},$$

$$y_i^{Q_2} = \beta_0^{Q_2} + \sum_{j=1}^{P} \left(\beta_3^m x_i^{jm} + \beta_3^{Q_1} x_i^{jQ_1} + \beta_3^{Q_2} x_i^{jQ_2} + \beta_3^{Q_3} x_i^{jQ_3} + \beta_3^M x_i^{jM}\right) + \varepsilon_i^{Q_2},$$

$$y_i^{Q_3} = \beta_0^{Q_3} + \sum_{j=1}^{P} \left(\beta_4^m x_i^{jm} + \beta_4^{Q_1} x_i^{jQ_1} + \beta_4^{Q_2} x_i^{jQ_2} + \beta_4^{Q_3} x_i^{jQ_3} + \beta_4^M x_i^{jM}\right) + \varepsilon_i^{Q_3},$$

$$y_i^M = \beta_0^M + \sum_{j=1}^{P} \left(\beta_5^m x_i^{jm} + \beta_5^{Q_1} x_i^{jQ_1} + \beta_5^{Q_2} x_i^{jQ_2} + \beta_5^{Q_3} x_i^{jQ_3} + \beta_5^M x_i^{jM}\right) + \varepsilon_i^M,$$

em que, $\beta_0^m, \beta_0^{Q_1}, \beta_0^{Q_2}, \beta_0^{Q_3}, \beta_0^M, \beta_1^m, \beta_1^{Q_1}, \beta_1^{Q_2}, \beta_1^{Q_3}, \beta_1^M, \beta_2^m, \beta_2^{Q_1}, \beta_2^{Q_2}, \beta_2^{Q_3}, \beta_2^M$,
$\beta_3^m, \beta_3^{Q_1}, \beta_3^{Q_2}, \beta_3^{Q_3}, \beta_3^M, \beta_4^m, \beta_4^{Q_1}, \beta_4^{Q_2}, \beta_4^{Q_3}, \beta_4^M, \beta_5^m$, e $\beta_5^{Q_1}, \beta_5^{Q_2}, \beta_5^{Q_3}, \beta_5^M$
são os coeficientes do modelo e $\varepsilon_i^m, \varepsilon_i^{Q_1}, \varepsilon_i^{Q_2}, \varepsilon_i^{Q_3}, \varepsilon_i^M$, para $i = 1, \ldots, n$, são os erros do modelo.

Uma característica desejável dos modelos de regressão com variáveis simbólicas é manter a coerência matemática dos intervalos previstos. Ou seja, que os limites superiores dos intervalos

sejam sempre maiores ou iguais aos inferiores. Desta forma, definimos a transformação Box-Cox para dados simbólicos de tipo $boxplot$.

O modelo de regressão linear assume que as variáveis regressoras têm uma relação linear com a variável resposta. Infelizmente, essa suposição pode não ser verdadeira. Nestes casos, uma transformação pode ser aplicada para linearizar a dependência entre as variáveis envolvidas (MONTGOMERY; PECK; VINING, 2001; DRAPER; SMITH, 1998). Box e Cox (1964) apresentaram uma família de transformações parametrizadas monotonicamente crescentes. Para um ponto $w \subset \mathbb{R}$, a transformação $Box - Cox$ é mostrada na equação (3.3) a seguir.

$$w^k = \begin{cases} \frac{(w+k_2)^{k^1}-1}{k_1}, & \text{if } k_1 \neq 0, \\ log(w+k_2), & \text{if } k_1 = 0 \end{cases} \tag{3.3}$$

em que, k é um parâmetro de transformação, com $k = (k_1, k_2)$. k_1 pode assumir qualquer valor real, mas k_2 deve satisfazer: $w + k_2 > 1$.

Da mesma forma, podemos estender a transformação $Box-Cox$ para dados de simbólicos de tipo $boxplot$. Dado $y = \{m, Q_1, Q_2, Q_3, M\}$, a transformação $Box - Cox$ é dada pela equação (3.4).

$$y^k = \begin{cases} \left[\frac{(m+k_2)^{k^1}-1}{k_1}, \frac{(Q_1+k_2)^{k^1}-1}{k_1}, \frac{(Q_2+k_2)^{k^1}-1}{k_1}, \frac{(Q_3+k_2)^{k^1}-1}{k_1}, \frac{(M+k_2)^{k^1}-1}{k_1}\right], k_1 \neq 0, \\ [\log(m+k_2), \log(Q_1+k_2), \log(Q_2+k_2), \log(Q_3+k_2), \log(M+k_2)], k_1 = 0 \end{cases} \tag{3.4}$$

em que, $k = (k_1, k_2)$. k_1 pode assumir qualquer valor real, mas k_2 deve satisfazer: $m + k_2 > 1$. Como a expressão $Box - Cox$ é crescente monotonicamente, y^k mantém a coerência matemática. Os valores k_1 e k_2 podem ser obtidos empiricamente ou por métodos de busca computacional.

Usando a Equação (3.4), é possível obter a inversa da transformação $Box - Cox$ para dados de $boxplot$, conforme apresentado na Equação 3.5

$$y = \begin{cases} [\exp\left(\frac{\log(1+k_1(m+k_2)^{k_1})}{k_1}\right), \exp\left(\frac{\log(1+k_1(Q_1+k_2)^{k_1})}{k_1}\right), \exp\left(\frac{\log(1+k_1(Q_2+k_2)^{k_1})}{k_1}\right), \\ \exp\left(\frac{\log(1+k_1(Q_3+k_2)^{k_1})}{k_1}\right), \exp\left(\frac{\log(1+k_1(M+k_2)^{k_1})}{k_1}\right)], k_1 \neq 0, \\ [\exp((m+k_2)^{k_1}), \exp((Q_1+k_2)^{k_1}), \exp((Q_2+k_2)^{k_1}), \exp((Q_3+k_2)^{k_1}), \\ \exp((M+k_2)^{k_1})], k_1 = 0 \end{cases} \tag{3.5}$$

O teste deve ser aplicado para todos os quatro intervalos que compõem os $boxplots$ do conjunto utilizado. Se todos os valores obtidos forem positivos, a coerência matemática é garantida. No entanto, se houver um único valor negativo, a transformação $Box - Cox$ para $boxplots$ deve ser aplicada à variável resposta.

Para um intervalo $[a, b]$, verifica-se a coerência matemática da regressão verificando se $H(b - a) \geq 0$, em que, H é a matriz de projeção de regressão. Assim, para cada $boxplots$, quatro intervalos, que definem um $boxplots$ do conjunto de dados de entrada, são avaliados: $[y^m, \hat{y}^{Q_1}], [y^{Q_1}, y^{Q_2}], [y^{Q_2}, y^{Q_3}]$ e $[y^{Q_3}, y^M]$. Como mostrado por (SOUZA et al., 2017), se todos os valores do cálculo $H(b - a)$ forem positivos, a coerência matemática é garantida para a previsão de valores dentro do casco convexo gerado pelo conjunto de treinamento. No entanto, se houver um único valor negativo, a transformação $Box-Cox$ para $boxplots$ deve ser aplicada à variável de resposta. Os parâmetros k_1 e k_2 devem ser selecionados como $H(b^k - a^k) \geq 0$, para todos os intervalos na amostra de treinamento.

Regra de predição: Dado um novo elemento s descrito por p variáveis regressoras $\mathbf{x}_s = (x_s^1, \ldots, x_s^p)$ com $x_s^j = \{x_s^{jm}, x_s^{jq_1}, x_s^{jq_2}, x_s^{jq_3}, x_s^{jM}\}$ $(j = 1, \ldots, p)$, o $boxplot$ previsto para a variável resposta é construído por $\hat{y}_s = \{\hat{y}_s^m, \hat{y}_s^{q_1}, \hat{y}_s^{q_2}, y_s^{q_3}, y_s^M\}$. Quando a transformação $Box - Cox$ é aplicada, as previsões são computadas em um espaço transformado. Para obter uma previsão associada aos dados originais, a transformação $Box - Cox$, definida na Equação (3.5), deve ser utilizada nas estimativas.

3.2 AVALIAÇÃO EXPERIMENTAL

Esta seção compara o ajuste do MRPB e dos métodos propostos na literatura ADS para dados intervalares: MC (BILLARD; DIDAY, 2000), MinMax (BILLARD; DIDAY, 2002) e MCR (NETO; CARVALHO, 2008). Dados sintéticos são gerados para analisar o ajuste desses métodos sob diferentes configurações. Alguns conjuntos de dados reais são, também, apresentados e confirmam a adaptabilidade do MRPB. A precisão dos modelos é estimada usando o Magnitude Média dos Erros Relativos (MMRE) (KITCHENHAM et al., 2001; FOSS et al., 2003; FAGUNDES; SOUZA; CYSNEIROS, 2014) através de simulações de Monte Carlo. Essa medida foi escolhida de acordo com (FOSS et al., 2003), que afirma que o MMRE é uma medida forte para verificar a adequação das previsões de um modelo linear. Quanto mais próximo seu valor estiver de 0, melhor será a adequação do modelo ajustado aos dados.

Nos experimentos, usamos o método de validação *Hold Out* para os dados simbólicos sintéticos, e o método de validação *Leave One Out* para os dados simbólicos reais. Isso se deve aos diferentes tamanhos dos conjuntos de dados, já que o método de validação *Hold Out* é eficaz e computacionalmente barato em conjuntos de dados muito grandes. Além disso, a implementação é simples e rápida. No entanto, para conjuntos de dados pequenos, o método

de validação *Leave One Out* é mais útil, pois permite que a menor quantidade de dados seja removida dos dados de treinamento em cada iteração.

O MMRE é apresentado na equação (3.6).

$$\frac{1}{5n}\sum_{i=1}^{n}\left\{\left|\frac{y_i^m-\hat{y}_i^m}{\hat{y}_i^m}\right|+\left|\frac{y_i^{Q_1}-\hat{y}_i^{Q_1}}{\hat{y}_i^{Q_1}}\right|+\left|\frac{y_i^{Q_2}-\hat{y}_i^{Q_2}}{\hat{y}_i^{Q_2}}\right|+\left|\frac{y_i^{Q_3}-\hat{y}_i^{Q_3}}{\hat{y}_i^{Q_3}}\right|+\left|\frac{y_i^{M}-\hat{y}_i^{M}}{\hat{y}_i^{M}}\right|\right\} \tag{3.6}$$

O MMRE para boxplots é usado para verificar a adequação das previsões do modelo linear. Modelos melhores são ajustados quando o valor do MMRE está próximo de 0. Além disso, são considerados o valor médio e o desvio padrão do MMRE. Para determinar se alguma das diferenças entre os modelos testados é estatisticamente significativa, usamos o teste de Friedman. A hipótese nula afirma que o valor médio para cada uma das populações é igual. Um nível de significância de 5% foi utilizado para comparar os resultados entre as abordagens..

3.2.1 Simulação de Monte Carlo

Os experimentos consistem em uma sequência de algoritmos organizados no $framework$ de uma simulação de Monte Carlos com 1000 iterações. Cinco configurações diferentes são usadas para gerar esses conjuntos de dados e 2 cenários para cada (dados sem ruído e dados com ruído). Assim, 400 $boxplots$ de tamanho 50 são gerados para a variável regressora e 400 $boxplots$ de tamanho 50 para variável resposta usando a equação (3.7),

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{3.7}$$

em que, x_i segue uma distribuição uniforme no intervalo $[-10; 10]$ (U(-10,10)) e ε_1 segue uma distribuição normal com média 0 e variância 1 (N(0,1)).

Foi utilizado o método de validação *Hold Out*. Cada conjunto simulado foi dividido em 80% para treinamento e 20% para teste. O conjunto de dados de treinamento possui 300 $boxplots$ que são escolhidos aleatoriamente. Cinco configurações são definidas para avaliar o desempenho do MRPB utilizando diferentes dados entre as cinco curvas m, Q_1, Q_2, Q_3 e M. As configurações são definidas de acordo com os parâmetros de distribuição uniforme conforme descrito na Tabela 5. Na configuração 1, todas as curvas têm valores iguais para os parâmetros β_0 e β_1 e na configuração 5 todas as curvas têm valores diferentes para o parâmetro β_0.

Vale ressaltar que, as curvas não possuem a mesma angulação e em todos os casos o parâmetro é gerado aleatoriamente. Para obter o cenário II, um ruído de distribuição gaussiana com média 0 e variância 4 foi adicionado aos dados simulados.

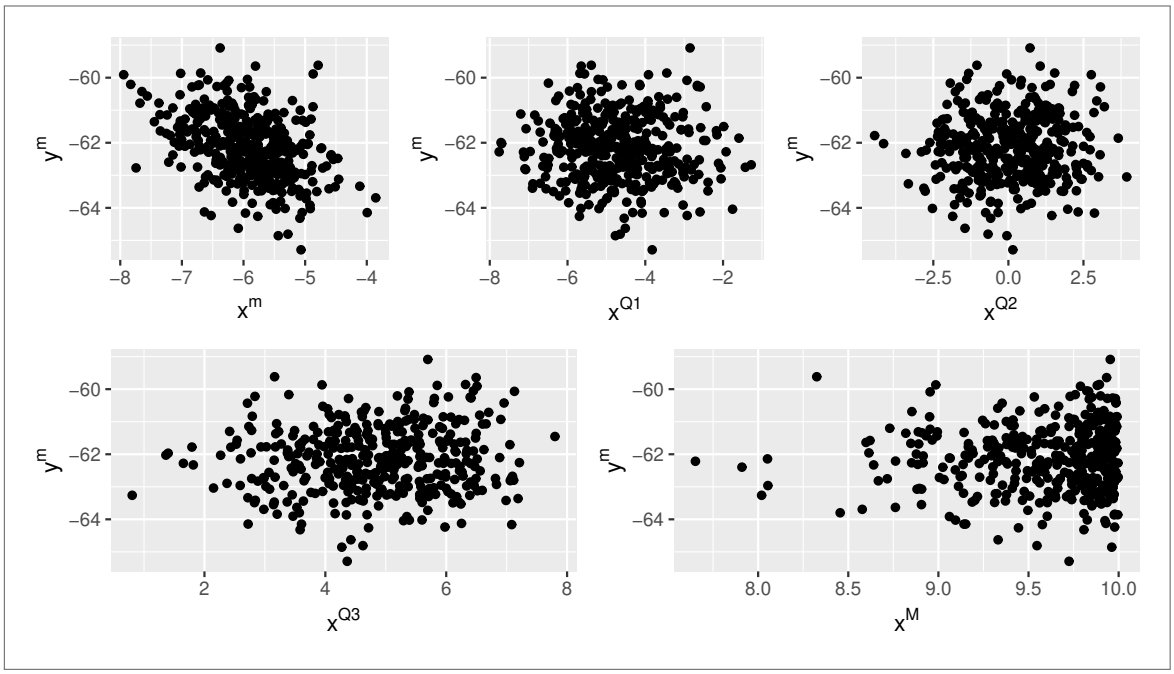
Tabela 5 – Configurações e parâmetros usadas para simular as curvas dos boxplots.

Config./Par.	y^m	y^{Q_1}	y^{Q_2}	y^{Q_3}	y^M
1 $/\beta_0$	U$(-5;5)$	U$(-5;5)$	U$(-5;5)$	U$(-5;5)$	U$(-5;5)$
1 $/\beta_1$	U$(-5;5)$	U$(-5;5)$	U$(-5;5)$	U$(-5;5)$	U$(-5;5)$
2 $/\beta_0$	U$(-67.5;-62.5)$	U$(-5;5)$	U$(-5;5)$	U$(-5;5)$	U$(-5;5)$
2 $/\beta_1$	U$(-5;5)$	U$(-5;5)$	U$(-5;5)$	U$(-5;5)$	U$(-5;5)$
3 $/\beta_0$	U$(-67.5;-62.5)$	U$(-5;5)$	U$(-5;5)$	U$(-5;5)$	U$(62.5;67.5)$
3 $/\beta_1$	U$(-5;5)$	U$(-5;5)$	U$(-5;5)$	U$(-5;5)$	U$(-5;5)$
4 $/\beta_0$	U$(-67.5;-62.5)$	U$(-50.0;-40.5)$	U$(-5;5)$	U$(-5;5)$	U$(62.5;67.5)$
4 $/\beta_1$	U$(-5;5)$	U$(-5;5)$	U$(-5;5)$	U$(-5;5)$	U$(-5;5)$
5 $/\beta_0$	U$(-67.5;-62.5)$	U$(-50.0;-40.5)$	U$(-5;5)$	U$(40.5;50.0)$	U$(62.5;67.5)$
5 $/\beta_1$	U$(-5;5)$	U$(-5;5)$	U$(-5;5)$	U$(-5;5)$	U$(-5;5)$

Fonte: Elaborada pelo autor (2021)

As figuras 1 a 5 mostram os gráficos de dispersão para os dados sintéticos referentes à configuração 5 que envolve diferentes valores de parâmetros para as curvas. Cada figura apresenta a relação linear entre a variável resposta e uma variável regressora associada a cinco curvas.

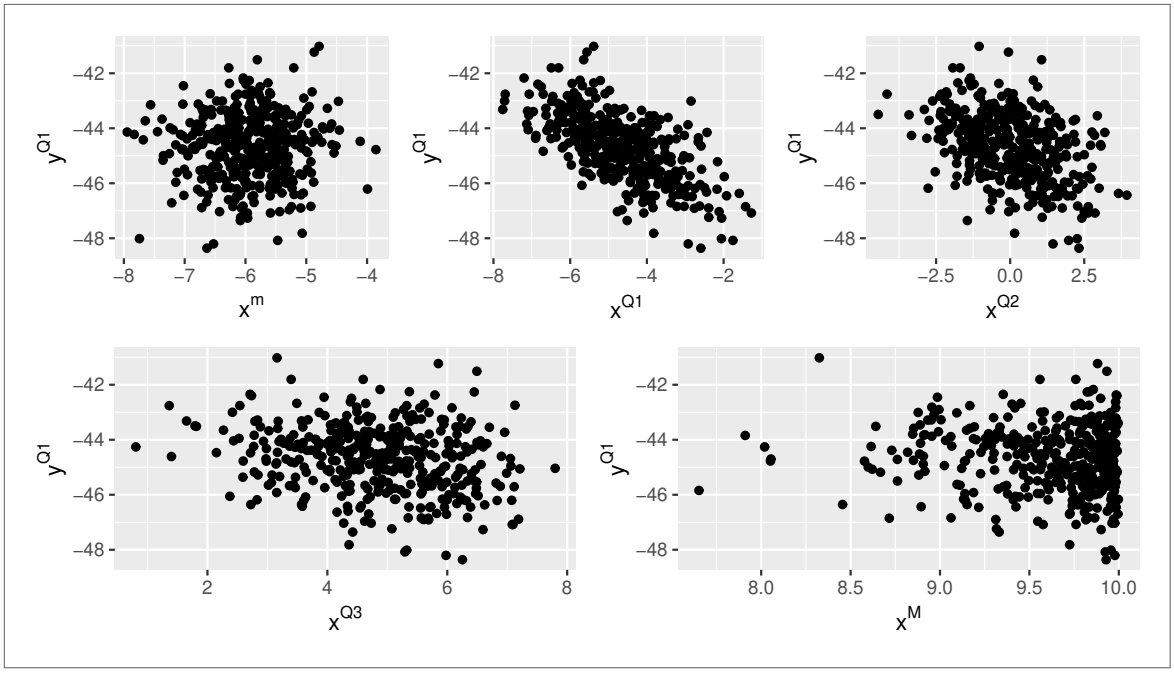
Figura 1 – Relação linear entre a variável resposta y^m e uma regressora para a config. 5



Fonte: Elaborada pelo autor (2021)

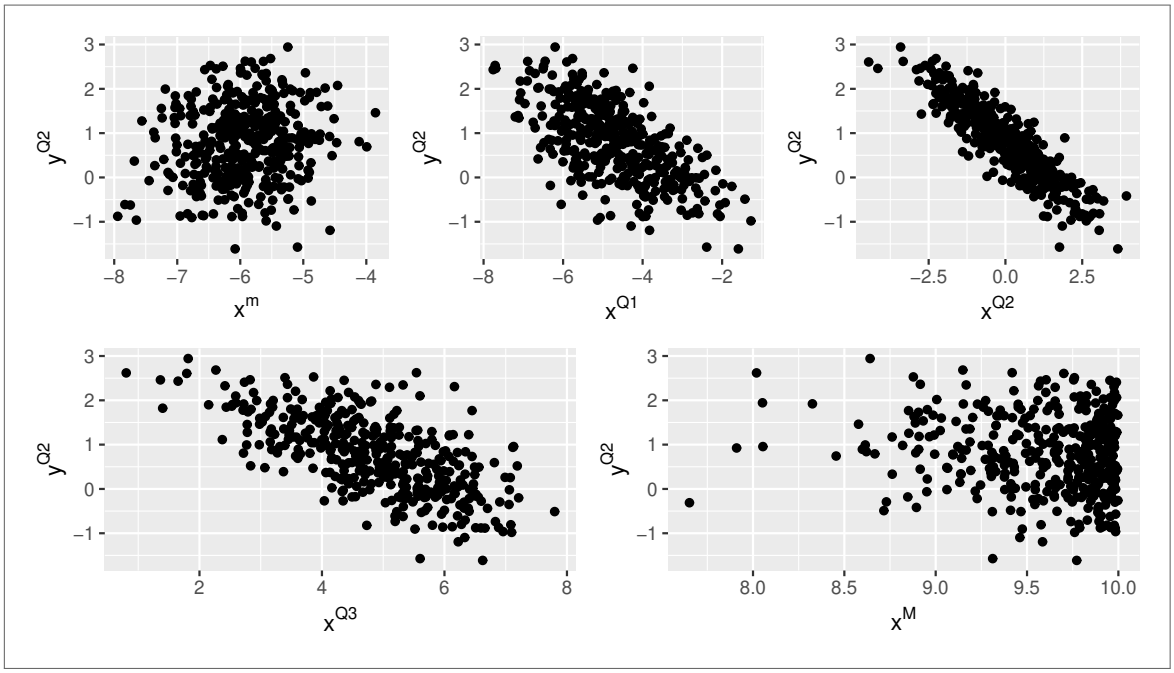
A Tabela 7 mostra os resultados do MMRE para as cinco configurações e os cenários I (sem dados ruidosos) e II (com dados ruidosos). A partir desses resultados, podemos observar

Figura 2 – Relação linear entre a variável resposta y^{Q_1} e uma regressora para a config. 5



Fonte: Elaborada pelo autor (2021)

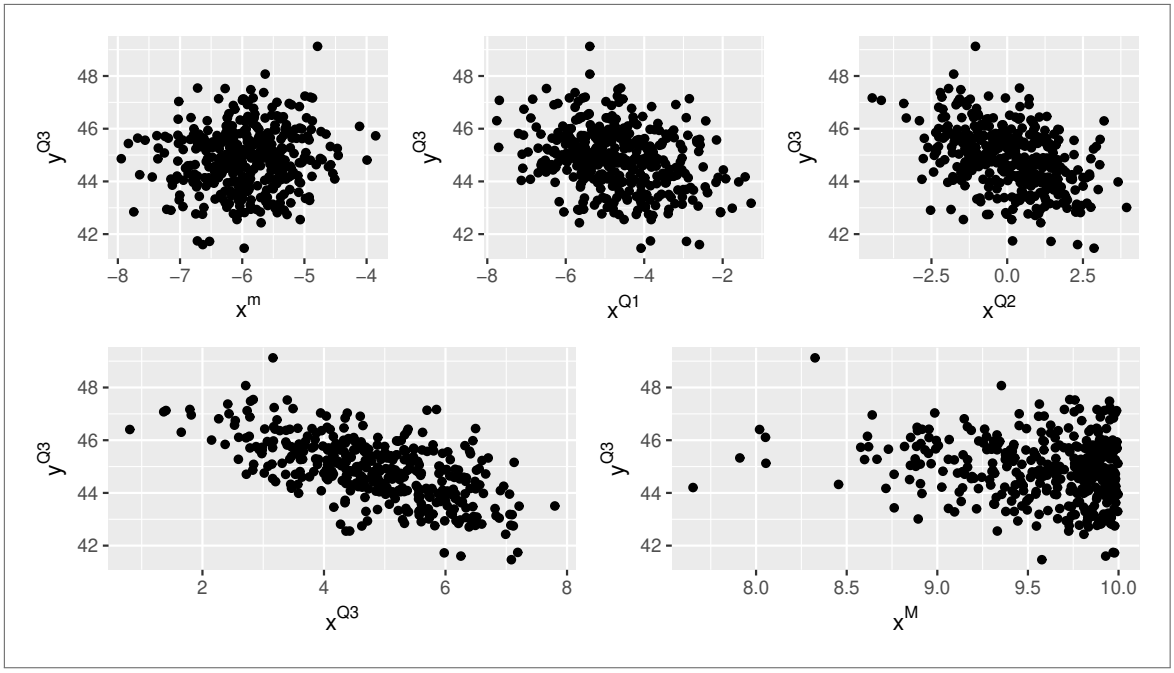
Figura 3 – Relação linear entre a variável resposta y^{Q_2} e uma regressora para a config. 5



Fonte: Elaborada pelo autor (2021)

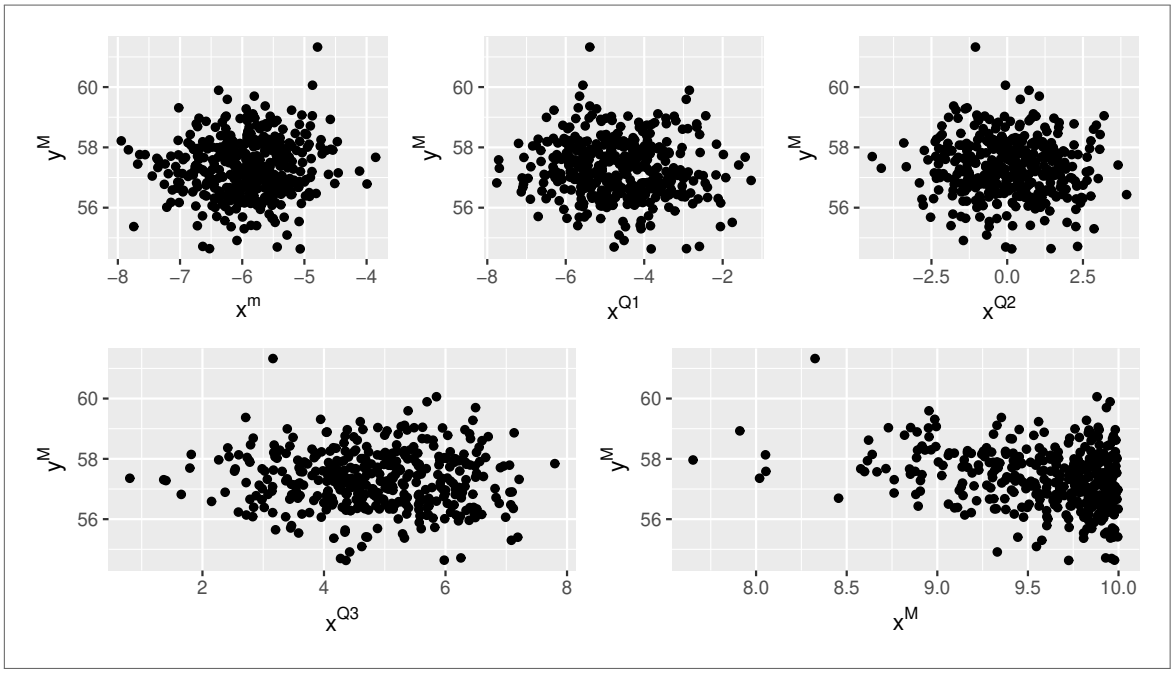
que o MRPB atinge o melhor $MMRE$ para conjunto de dados sem e com dados ruidosos quando as cinco curvas de dados são geradas em relação a diferentes valores de parâmetros da distribuição uniforme. Podemos dizer que a configuração 5 é o caso mais geral. Como esperado, o MMRE para o conjunto de dados ruidosos é maior do que o conjunto de dados

Figura 4 – Relação linear entre a variável resposta y^{Q_3} e uma regressora para a config. 5



Fonte: Elaborada pelo autor (2021)

Figura 5 – Relação linear entre a variável resposta y^{M} e uma regressora para a config. 5



Fonte: Elaborada pelo autor (2021)

sem ruído. Os menores valores de MMRE para ambos os cenários são obtidos da configuração 5. Esse resultado é importante, pois a configuração 5 envolve valores de parâmetros diferentes para as curvas. É a configuração mais complexa neste estudo.

A Tabela 12 mostra os resultados obtidos da comparação do desempenho do MRPB pro-

Tabela 6 – Média e desvio padrão entre parêntesis do MMRE calculado a partir de 1000 repetições de Monte Carlo.

Configuração	Cenário I	Cenário II
1	0.3548 (1.2263)	0.7661 (1.6874)
2	0.2995 (0.6032)	0.6698 (1.6223)
3	0.2850 (0.4086)	0.5416 (0.6664)
4	0.2410 (0.5722)	0.4738 (0.4554)
5	0.1746 (0.2193)	0.3413 (0.2396)

Fonte: Elaborada pelo autor (2021)

posto com os modelos da literatura de ADS para dados intervalares (MC (BILLARD; DIDAY, 2000), MinMax (BILLARD; DIDAY, 2002) e MCR (NETO; CARVALHO, 2008)). A Tabela 12 mostra a média e o desvio padrão (entre parêntesis) do MMRE para estes modelos nos cenários I e II baseados na configuração 5. Além disso, consideramos três versões do nosso modelo MRPB usando três curvas para representar os $boxplots$, ou seja, três curvas são ajustadas. Eles são:

- MRPB1 que considera as curvas Q_1, Q_2 e Q_3 para cada variável. Assim, temos como saída do modelo $\hat{Q}_1, \hat{Q}_2$ e $\hat{Q}_3$. Os $boxplots$ previstos são construídos assumindo que $\hat{m} = \hat{Q}_1 - 1.5(\hat{Q}_3 - \hat{Q}_1)$ e $\hat{M} = \hat{Q}_3 + 1,5(\hat{Q}_3 - \hat{Q}_1)$;

- MRPB2 que considera as curvas Q_1, Q_2 e $I = Q_3 - Q_1$ para cada variável. Assim, temos como saída do modelo $\hat{Q}_1, \hat{Q}_2$ e $\hat{I}$. Os $boxplots$ previstos são construídos assumindo que $\hat{m} = \hat{Q}_1 - 1.5\hat{I}$, $\hat{Q}_3 = \hat{I} - \hat{Q}_1$ e $\hat{M} = \hat{Q}_3 + 1,5\hat{I}$ e

- MRPB3 que considera as curvas Q_2, Q_3 e I para cada variável. Assim, temos como saída do modelo $\hat{Q}_2, \hat{Q}_3$ e $\hat{I}$. Os $boxplots$ previstos são construídos assumindo que $\hat{Q}_1 = \hat{Q}_3 - \hat{I}$ $\hat{m} = \hat{Q}_1 - 1.5\hat{I}$ e $\hat{M} = \hat{Q}_3 + 1,5\hat{I}$

Os resultados da Tabela 12 destacam que o MRPB proposto (usando as curvas m, Q_1, Q_2, Q_3 e M) para prever variáveis de tipo $boxplot$ supera todos os modelos testados. Além disso, podemos dizer que o MRPB1 (usando as curvas Q_1, Q_2 e Q_3) é a segunda melhor opção. Entre todas as variantes apresentadas, MRPB2 e MRPB3 usando a curva I, apresentam os piores resultados em termos de MMRE. Isso porque o uso do intervalo interquartil aumenta o erro. Como esperado, MC para dados de intervalo mostra os valores mais altos de MMRE.

Além disso, o valor p da comparação dos resultados de todos os modelos por meio do teste de Friedman foi de 0,015. Como esse valor p é menor que 0,05, podemos rejeitar a hipótese nula $H0$ a um nível de significância de 5%. Em palavras simples, temos provas suficientes

Tabela 7 – Média e desvio padrão entre parêntesis do MMRE calculado a partir de 1000 repetições de Monte Carlo para os diferentes modelos.

Modelos	Cenário I	Cenário II
MC	2.0965 (12.109)	3.6214 (10.302)
MinMax	0.4894 (0.2097)	0.5716 (0.1909)
CRM	0.3753 (0.2448)	0.5552 (0.1971)
MRPB	**0.1746** (0.2193)	**0.3413** (0.2396)
MRPB1	0.2774 (0.7858)	0.4563 (0.7903)
MRPB2	0.9285 (0.5278)	1.0573 (0.6166)
MRPB3	0.9050 (0.6086)	1.0221 (0.6396)

Fonte: Elaborada pelo autor (2021)

para dizer que houve uma diferença estatisticamente significativa entre os desempenhos dos modelos comparados. Assim, o teste Nemenyi foi realizado para encontrar exatamente quais modelos possuem médias diferentes. O teste $post - hoc$ de Nemenyi produz os valores de p para cada comparação de médias em pares. Esses valores mostraram que não houve diferenças significativas entre MRPB2 e MRPB3 $(Z = -0,061, p = 0,952)$ ou entre MinMax, MCR e MRPB1 $(Z = -1,811, p = 0,070)$, apesar da média de MMRE para MRPB1 ser menor. No entanto, houve redução estatisticamente significativa no MMRE para MRPB vs outros modelos, quando comparados ao MC $(Z = -2,636, p = 0,008)$.

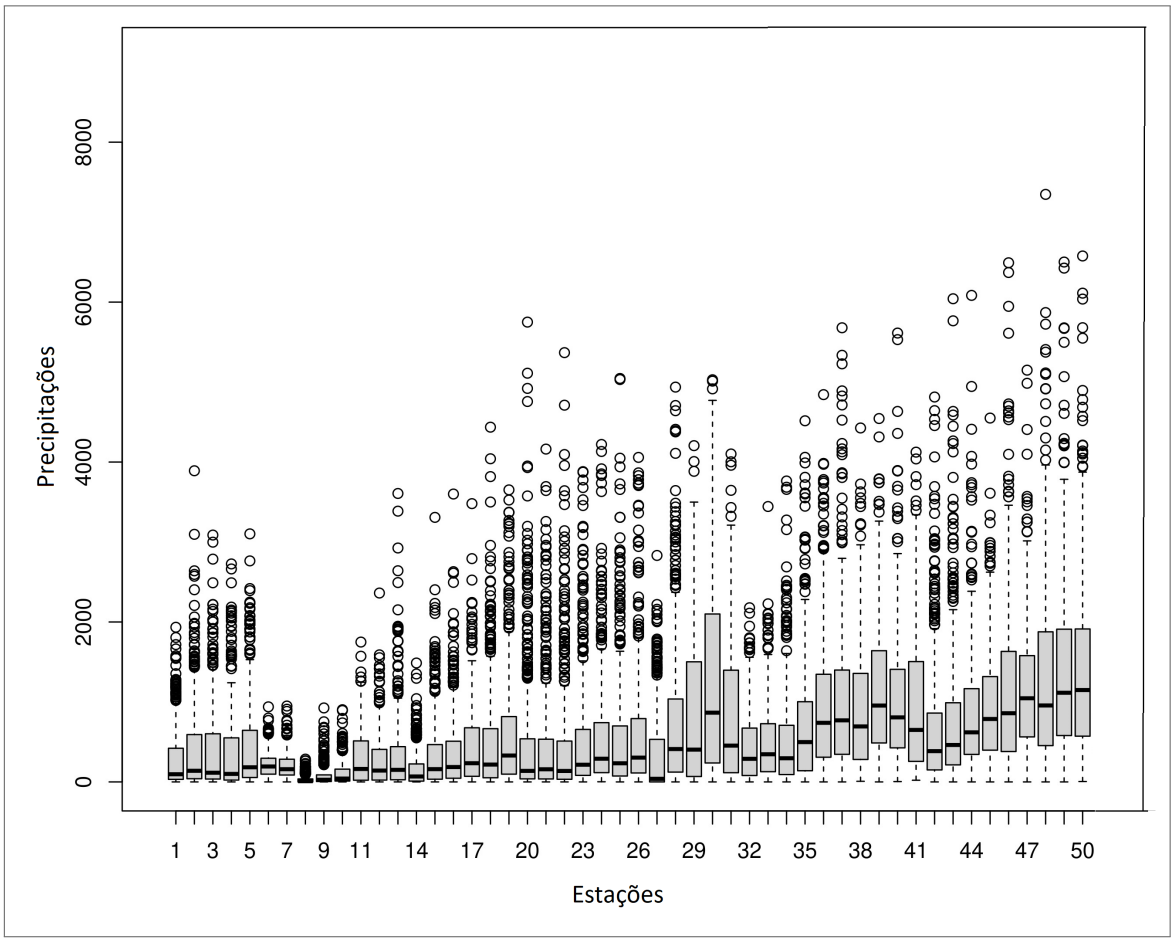
3.2.2 Base de Dados Reais

O conjunto de dados foi extraído do Banco de Dados Climático Instrumental de Longo Prazo da República Popular da China. Esta base de dados contém as precipitações mensais na China registradas por 60 estações meteorológicas entre 1951 e 1988. Cada estação representa uma classe em que o registro das precipitações é a variável regressora (X) e a temperatura máxima é a variável resposta (Y). Assim, são 60 classes de 437 valores cada, representadas por variáveis simbólicas do tipo $boxplot$ (ver Figuras 6 e 7).

O método de validação *Leave One Out* foi utilizado para realizar a estimação do modelo. Para um conjunto de dados com amostras *n*, são construídos modelos *n*, eliminando, a cada vez, uma amostra, que é utilizada para previsão e cálculo do MMRE. Em outras palavras, o conjunto de treinamento terá o comprimento $k - 1$ e o conjunto de teste será uma única amostra dos dados.

A análise do intervalo previsto, conforme apresentado anteriormente, mostra que nenhuma

Figura 6 – Boxplots das precipitações (variável regressora) observadas nas 60 estações da China.



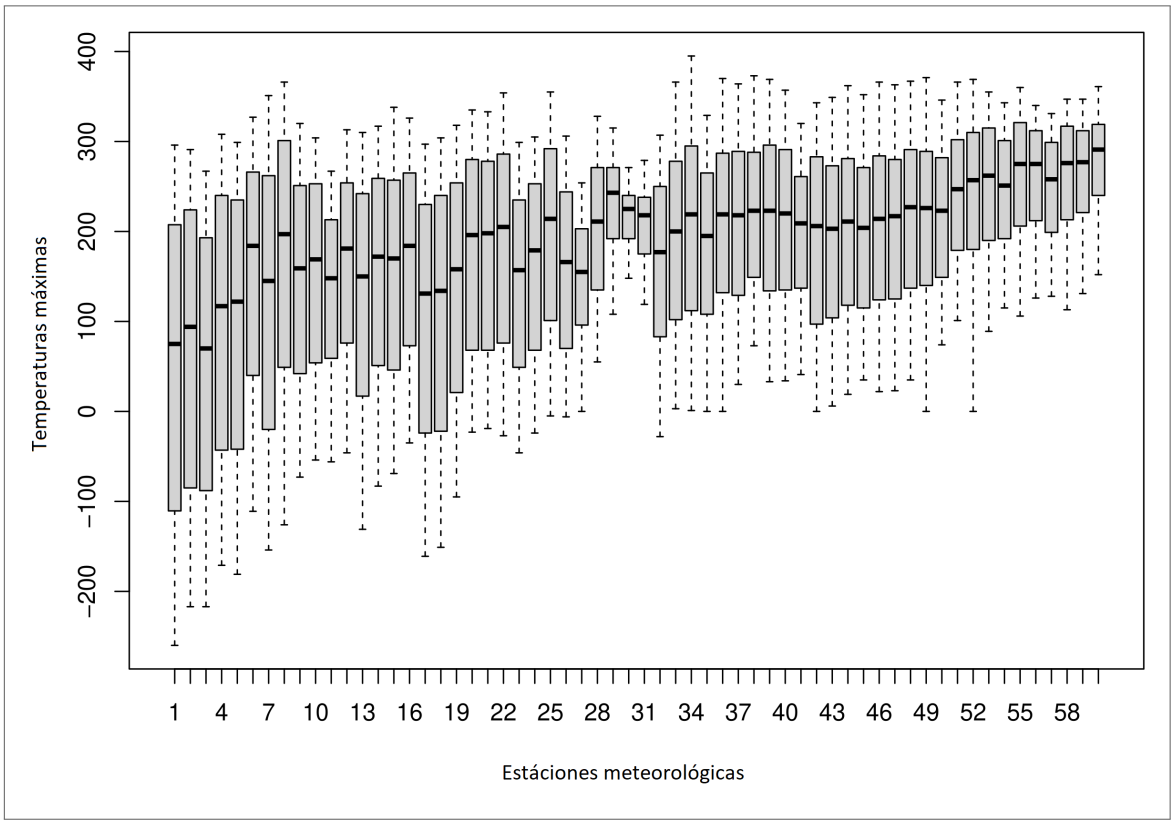
Fonte: Elaborada pelo autor (2021)

transformação precisa ser aplicada à resposta, portanto a coerência matemática já está garantida.

A Tabela 8 apresenta os valores do MMRE para MC,MinMax, MCR, MRPB, MRPB1, MRPB2 e MRPB3. Novamente, MRPB teve o melhor desempenho para ajustar os dados reais. O valor p da comparação desses resultados usando o teste de Friedman foi 0,003. Como esse valor p é menor que 0,05, podemos rejeitar a hipótese nula H_0 a um nível de significância de 5%. Isso indica que houve diferença estatisticamente significativa entre os desempenhos dos modelos. O teste Nemenyi mostrou que não houve diferenças significativas entre o MRPB e os demais modelos.

Este resultado revela a aplicabilidade do MRPB para ajustar o modelo de regressão linear para $boxplot$ com dados reais. MC teve o pior desempenho.

Figura 7 – Boxplots da temperatura máxima (variável resposta) observada nas 60 estações na China.



Fonte: Elaborada pelo autor (2021)

Tabela 8 – MMRE para o conjunto de dados reais.

Modelo	MMRE
MC	0.5083
MinMax	0.1857
CRM	0.1954
MRPB	**0.1436**
MRPB1	0.2070
MRPB2	0.1806
MRPB3	0.1745

Fonte: Elaborada pelo autor (2021)

3.3 APLICAÇÃO NO SETOR ELÉTRICO BRASILEIRO

No Brasil, a geração termelétrica vem sofrendo um aumento considerável no número de despachos para atendimento da demanda do Sistema Interligado Nacional (SIN). Dentre os inúmeros motivos, é possível citar a possibilidade de produção em uma quantidade constante durante todo o ano, restrições hidrológicas, atrasos para entrada em operação de novos em-

preendimentos de geração e transmissão e o baixo volume armazenado de energia. Situações como essas, demandam ajustes significativos nos planos de manutenção das UTE, para evitar os riscos de indisponibilidade momentânea de um ou mais ativos, e garantindo a disponibilidade e a confiabilidade energética (MORADI; CHAIBAKHSH; RAMEZANI, 2018). Em setores regulados, o custo da falha tende a ser alto pois está relacionado à indisponibilidade de serviços essenciais e sem reposição, dessa forma, as falhas podem gerar perda de faturamento e penalidades contratuais ou regulatórias.

TermoCabo, é uma UTE que está localizada no Cabo de Santo Agostinho no estado de Pernambuco, nordeste do Brasil. Termocabo foi constituída em agosto de 2001 objetivando a produção e comercialização de 48 MW de energia elétrica. Em janeiro de 2002 foi assinado um Plano Plurianual com a Comercializadora Brasileira de Energia Emergencial (CBEE) para operação comercial de 40 meses com duração até dezembro de 2005. A usina, considerada emergencial, entrou em operação comercial em setembro de 2002, sua obra durou 7 meses (TERMOCABO, 2019).

As UTEs produzem eletricidade queimando combustíveis como carvão, gás natural e óleo combustível. O vapor produzido pela queima aciona turbinas conectadas a geradores que, por sua vez, produzem eletricidade. No Brasil, as UTEs são uma estratégia do setor elétrico para garantir o fornecimento de energia e reduzir o risco de deficit no sistema em períodos de condições hidrológicas críticas ou indisponibilidade de geração de energia eólica e solar. A ociosidade por longos períodos está associada a contratempos relacionados à operação e manutenção das UTEs brasileiras. Além disso, a interrupção não programada do fornecimento de energia elétrica decorrente de falhas nestes equipamentos resulta em multas da ANEEL. Além dos custos de manutenção corretiva ou substituição de equipamentos, quando se verifica que a indisponibilidade da usina foi decorrente de mau planejamento, manutenção ou operação, a concessionária poderá sofrer advertência com multa de até 1% do valor das vendas anuais. Nesse sentido, é um desafio para as UTEs brasileiras garantir sua disponibilidade por meio de métodos tradicionais de manutenção para evitar penalidades.

Nesta seção, avaliamos uma aplicação da abordagem proposta usando dados fornecidos pela UTE TermoCabo. O estudo apresentado faz parte do projeto de P&D "PD-02901-0003/2019: Mapa de risco em tempo real baseado em aprendizado de máquina aplicado em manutenção preditiva", regulamentado pela ANEEL, Brasil. Este é um projeto financiado com recursos P&D ANEEL e encontra-se em andamento dentro da Termocabo, em parceria com a In Forma Software S.A. O produto principal deste projeto será um sistema inteligente, baseado

em Aprendizagem de Máquina, que permita predizer o nível de risco de falha em tempo real de componentes da instalação a partir de dados cadastrais e comportamentais, apresentando os resultados em um $Dashboard$, onde será possível monitorar os equipamentos com informações visuais e utilizar como apoio à tomada de decisão, possibilitando a geração de indicadores e posterior avaliação do impacto sobre os KPIs (do inglês, *Key Performance Indicator*) do negócio. A conclusão do projeto, está prevista para março de 2022.

3.3.1 Dados da Usina Termo-elétrica da TermoCabo

A TermoCabo possui três motores Wärtsilä 18V46 com uma potência elétrica total de 48 MW. Cada motor possui quatro sistemas: Temperatura, Água de Resfriamento, Lubrificação e Ar de Partida. Todos esses sistemas são monitorados por sensores que acionam alarmes quando as medidas de interesse atingem determinados limites. Assim, a detecção de anomalias nos motores é muito simples, pois se baseia apenas no monitoramento das variáveis individuais com limites definidos por especialistas. Anomalias que não são detectadas em estágios iniciais (ou não detectadas) pode resultar em quebra de equipamentos ou multas devido ao fornecimento reduzido de energia. No presente trabalho, o sistema alvo é o sistema de água de resfriamento dos motores.

O sistema de água de resfriamento utiliza água doce tratada quimicamente para controlar a temperatura dos motores. Ele inclui sensores para monitorar a pressão e a temperatura do sistema e é dividido em um circuito de água de resfriamento de Baixa Temperatura (LT) e de Alta Temperatura (HT).

Cada motor tem dezoito cilindros, nove no lado A e nove no lado B, e cada cilindro tem três camisas. O circuito HT é usado para controlar as camisas dos cilindros, cabeçotes e o primeiro estágio do resfriador de ar de admissão. Devido às altas temperaturas de combustão alcançadas pelos gases de combustão nos cilindros do motor, as falhas do sistema HT podem levar a problemas como falha no cilindro, redução da eficiência volumétrica do motor e distorção dos componentes do motor.

Um conjunto de dados, com 597 variáveis, foi produzido a partir dos valores registrados pelos sensores para monitorar a operação do circuito HT do motor 1. Vale ressaltar que, esses dados foram coletados nos dias em que o motor deu partida. Isso em razão de, por se tratar de uma usina de reserva, a Termocabo não está sempre em operação e é despachada conforme solicitações do Operador Nacional do Sistema Elétrico (ONS) ou por necessidade do agente

(geração por inflexibilidade).

O próximo passo foi definir as variáveis de interesse. Algumas técnicas de seleção e redução de variáveis foram aplicadas nos dados (por exemplo: Análise de Componentes Principais, correlação e técnicas de Aprendizagem de Máquina) que mostraram que as variáveis possuem forte relação. Ou seja, muitas destas variáveis podem ser consideradas redundantes uma vez que carregam as informações umas das outras. No contexto de medicina, por exemplo, uma situação como essa seria as variáveis Peso e o Índice de Massa Corpórea (Peso dividido pela Altura). O IMC carrega a informação de peso, sendo muitas vezes redundante usar as duas variáveis juntas.

Assim, com os resultados obtidos das técnicas aplicadas e com ajuda do conhecimento de um especialista em engenharia mecânica foram selecionadas as variáveis mostradas na Tabela 9 como as mais interessantes para predizer dados de $boxplots$ usando o modelo MRPB proposto neste capítulo.

Tabela 9 – Variáveis selecionadas do sistema de resfriamento.

Variáveis	Descrição	Unidade	Valor Min.	Valor Máx.
Y_1	Temp. A da saída de água HT	°C	0	120
Y_2	Temp. B da saída de água HT	°C	0	120
X_1	Pressão de entrada de água HT	bar	0	6
X_2	Freq. do Motor 1	Hz	0	30

Fonte: Elaborada pelo autor (2021)

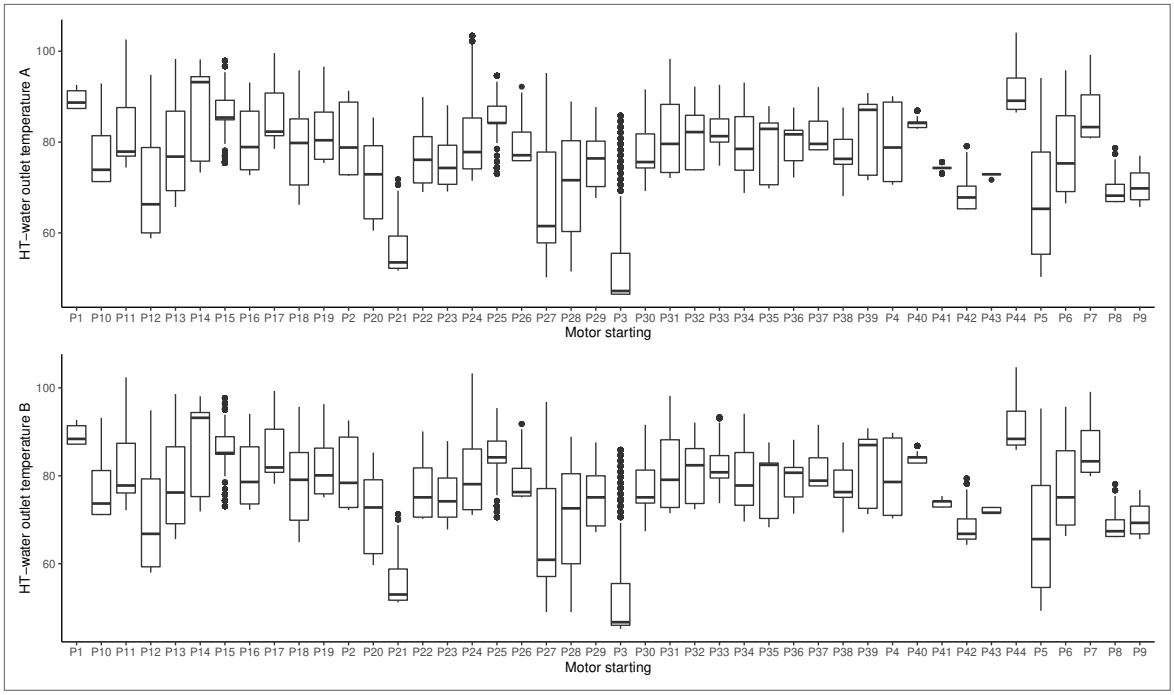
O conjunto de dados selecionado possui $N = 44712$ elementos que são agregados e representados por 44 $boxplots$ representando cada partida do motor 1 registrada. O procedimento de agregação permite reter as informações presentes nos dados e adicionar informações sobre a variação temporal. Figura 8 exibe 2 variáveis de tipo $boxplot$ construídas para as temperaturas A e B de saída de água HT, respectivamente. Enquanto a Figura 9 exibe as variáveis regressoras (pressão de entrada de água HT e frequência do motor 1) do tipo $boxplot$.

3.3.2 Ajuste do Modelo de Regressão nos Dados da TermoCabo

Dois modelos baseados em MRPB são ajustados para as temperaturas de saída de água HT, um para a temperatura A (Y_1) e outro para a temperatura B (Y_2) conforme descrito nas equações (3.8) e (3.9).

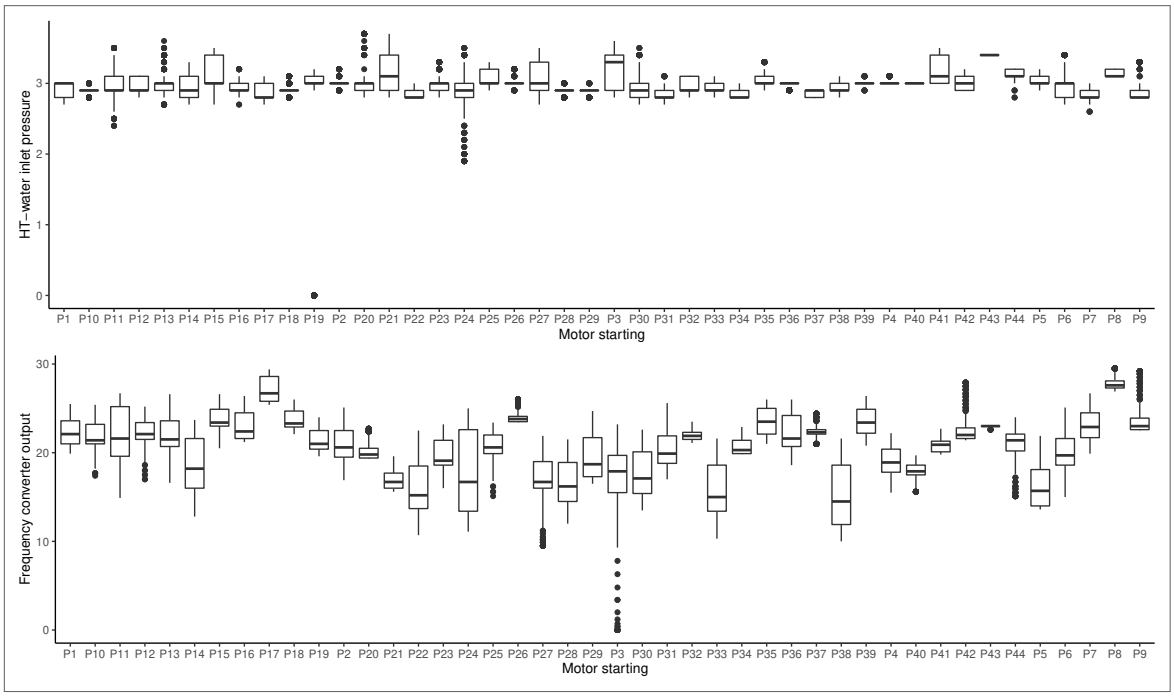
$$Y_1 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \tag{3.8}$$

Figura 8 – Boxplots das temperaturas A e B de saída de água HT para cada partida de motor registrada.



Fonte: Elaborada pelo autor (2021)

Figura 9 – Boxplots da pressão de entrada de água HT e frequência do motor 1 para cada partida registrada.



Fonte: Elaborada pelo autor (2021)

$$Y_2 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \tag{3.9}$$

As variáveis foram normalizadas para evitar que diferenças de escala afetassem o ajuste

dos modelos. Cada variável foi transformada em uma escala comum (ou seja, [0,1]), com base na equação (3.10):

$$x_i^{j*} = \frac{x_i^j - x_{Min}^j}{x_i^j - x_{Max}^j} \tag{3.10}$$

em que, x_i^j é o valor original, x_i^{j*} é o valor normalizado, x_{Max}^j e x_{Min}^j são respectivamente os valores máximo e mínimo observados para o atributo j-ésimo nos dados.

A seguir, apresentamos as equações de regressão dos modelos MRPB ajustados ao conjunto de dados de tipo $boxplot$ da TermoCabo para as temperaturas A e B de saída de água HT: Para a temperatura A (Y_1):

$$\begin{aligned}\hat{Y}_1^m = 114.5 \;&-\; 2.6X_1^m - 5.2X_1^{Q_1} - 4.1X_1^{Q_2} + 0.5X_1^{Q_3} - 11.5X_1^M \\ &+\; 0.6X_2^m - 3X_2^{Q_1} + 3.4X_2^{Q_2} + 2.2X_2^{Q_3} - 1.8X_2^M,\end{aligned}$$

$$\begin{aligned}\hat{Y}_1^{Q_1} = 110.1 \;&-\; 1.5X_1^m + 6.9X_1^{Q_1} - 19.7X_1^{Q_2} + 4.9X_1^{Q_3} - 8.9X_1^M \\ &+\; 0.3X_2^m - 0.3X_2^{Q_1} + 0.9X_2^{Q_2} + 4.1X_2^{Q_3} - 2.2X_2^M,\end{aligned}$$

$$\begin{aligned}\hat{Y}_1^{Q_2} = 120.2 \;&-\; 0.9X_1^m + 26.9X_1^{Q_1} - 22.6X_1^{Q_2} - 14.9X_1^{Q_3} - 4.5X_1^M \\ &+\; 0.4X_2^m - 2.4X_2^{Q_1} + 2X_2^{Q_2} + 3.2X_2^{Q_3} - 2.7X_2^M,\end{aligned}$$

$$\begin{aligned}\hat{Y}_1^{Q_3} = 157.9 \;&-\; 1.6X_1^m + 30.1X_1^{Q_1} - 35.7X_1^{Q_2} - 18.4X_1^{Q_3} + 0.4X_1^M \\ &-\; 0.3X_2^m - 0.1X_2^{Q_1} - 0.3X_2^{Q_2} + 3.6X_2^{Q_3} - 2.8X_2^M,\end{aligned}$$

$$\begin{aligned}\hat{Y}_1^M = 164.2 \;&-\; 3.3X_1^m + 23.1X_1^{Q_1} - 36.4X_1^{Q_2} - 12.7X_1^{Q_3} + 2.8X_1^M \\ &-\; 1.4X_2^m + 0.9X_2^{Q_1} - 0.6X_2^{Q_2} + 3.7X_2^{Q_3} - 2.3X_2^M.\end{aligned}$$

Para a temperatura B (Y_2):

$$\begin{aligned}\hat{Y}_1^m = 115.3 \;&-\; 2.8X_1^m - 8.4X_1^{Q_1} + 0.1X_1^{Q_2} - 0.4X_1^{Q_3} - 12.3X_1^M \\ &+\; 0.6X_2^m - 2.9X_2^{Q_1} + 3.1X_2^{Q_2} + 2.1X_2^{Q_3} - 1.6X_2^M,\end{aligned}$$

$$\begin{aligned}\hat{Y}_1^{Q_1} = 109.5 \;&-\; 1.5X_1^m + 7.3X_1^{Q_1} - 19.1X_1^{Q_2} + 4.5X_1^{Q_3} - 9.4X_1^M \\ &+\; 0.3X_2^m - 0.1X_2^{Q_1} - 1.1X_2^{Q_2} + 3.9X_2^{Q_3} - 2.1X_2^M,\end{aligned}$$

$$\begin{aligned}\hat{Y}_1^{Q_2} = 126.1 \;&-\; 1X_1^m + 27.9X_1^{Q_1} - 25.5X_1^{Q_2} - 14.7X_1^{Q_3} - 4.3X_1^M \\ &+\; 0.4X_2^m - 2.9X_2^{Q_1} + 2.8X_2^{Q_2} + 3.1X_2^{Q_3} - 2.9X_2^M,\end{aligned}$$

$$Y_1^{Q_3} = 164.3 - 1.7X_1^m + 30.3X_1^{Q_1} - 37.9X_1^{Q_2} - 18.3X_1^{Q_3} + 0.2X_1^M - 0.3X_2^m - 0.4X_2^{Q_1} + 0.4X_2^{Q_2} + 3.4X_2^{Q_3} - 2.7X_2^M,$$

$$\hat{Y}_1^M = 158.4 - 3.1X_1^m + 26.1X_1^{Q_1} - 36.8X_1^{Q_2} - 12.8X_1^{Q_3} + 2.4X_1^M - 1.5X_2^m + 1.3X_2^{Q_1} - 1.3X_2^{Q_2} + 3.8X_2^{Q_3} - 2.1X_2^M.$$

3.3.3 Avaliação do Modelo de Regressão Ajustado

A partir dos modelos MRPB ajustados, podemos prever as variáveis temperatura A (Y_1) e temperatura B (Y_2) e avaliar os dois modelos calculando a medida de MMRE. Além disso, as versões do modelo MRPB baseado em três curvas (MRPB1, MRPB2 e MRPB3) são construídas e avaliadas nesta subseção.

A coerência matemática já é garantida pelo método proposto MRPB, consequentemente nenhuma transformação precisa ser aplicada à variável resposta.

A Tabela 10 exibe os MMRE para MRPB, MRPB1, MRPB2, MRPB3, MC, MinMax e MCR. A partir desta tabela, concluímos que o MRPB supera todos os modelos para esta aplicação. Por outro lado, os métodos MRPB1, MRPB2 e MRPB3 apresentaram desempenho muito semelhante entre si. MC e MCR apresentaram os maiores valores de MMRE, apresentando os piores desempenhos. Esse resultado indica que o MRPB teve um ajuste melhor do que os demais modelos propostos na literatura ADS. O MinMax (BILLARD; DIDAY, 2002) é a segunda opção de modelo neste estudo e a melhor opção se os dados foram intervalares.

Tabela 10 – MMRE calculado para o ajuste dos modelos para as temperaturas A e B de saída de água HT.

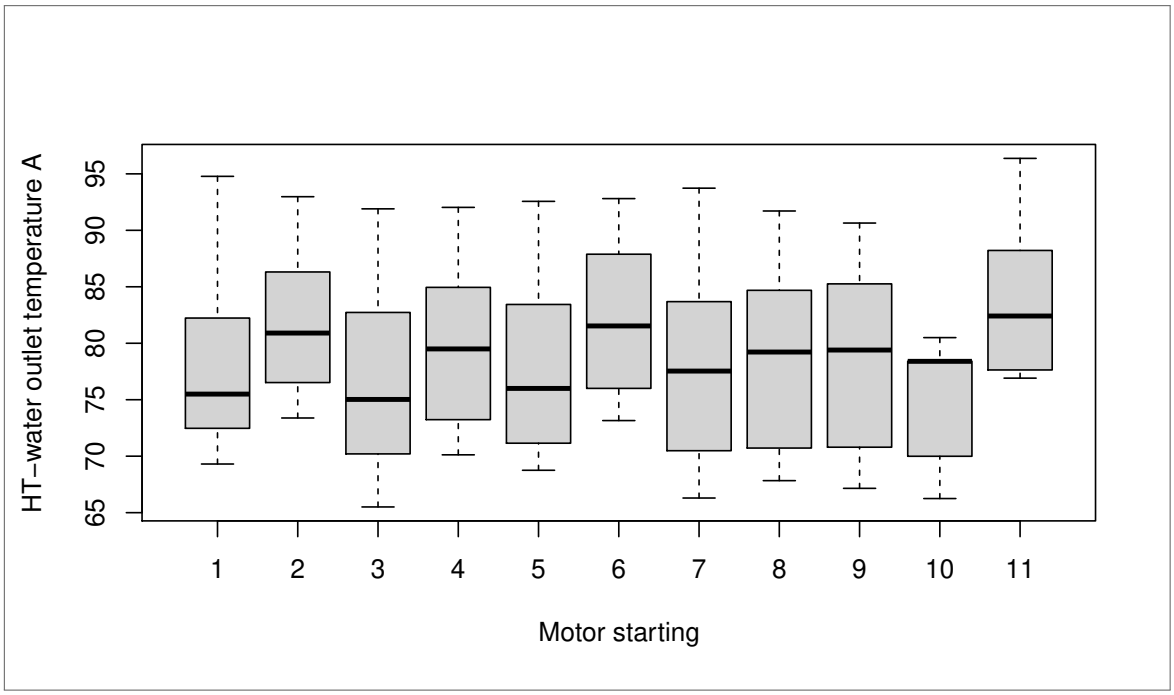
Modelo	**MMRE (Y_1)**	**MMRE (Y_1)**
MC	0.2044	0.2020
MinMax	0.0072	0.0076
MCR	0.2077	0.2051
MRPB	**0.0055**	**0.0056**
MRPB1	0.0108	0.0104
MRPB2	0.0187	0.0193
MRPB3	0.0105	0.0107

Fonte: Elaborada pelo autor (2021)

As Figuras 10 e 11 mostram os $boxplots$ estimados das temperaturas A e B de saída de água HT, respectivamente, usando o MRPB para as 11 próximas partidas do motor. Ambas

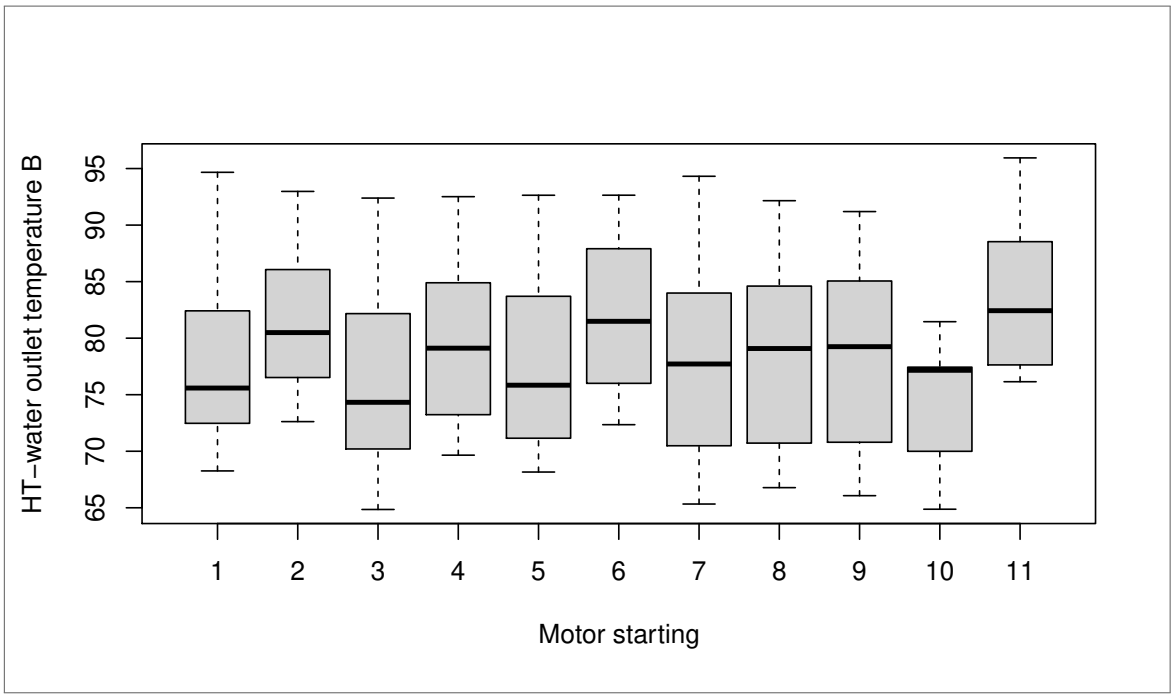
as previsões tiveram um desempenho semelhante variando entre 65 e 95A °C. Com esses resultados é esperado que as próximas partidas do motor 1 sejam bem-sucedidas, porque as temperaturas A e B não excedem os limites permitidos (consulte a Tabela 9), o que pode levar a uma falha.

Figura 10 – Boxplots estimados da temperatura A de saída de água HT para as próximas 11 partidas do motor 1.



Fonte: Elaborada pelo autor (2021)

Figura 11 – Boxplots estimados da temperatura B de saída de água HT para as próximas 11 partidas do motor 1.



Fonte: Elaborada pelo autor (2021)

3.4 CONCLUSÕES DO CAPÍTULO

Este capítulo propôs o novo método de regressão linear parametrizada para variáveis simbólicas de tipo $boxplot$ (MRPB). Este modelo é construído com base em cinco valores para as variáveis de resposta e regressoras: mínimo, primeiro quartil, segundo quartil (ou mediana), terceiro quartil e máximo. Uma vantagem da abordagem proposta é a utilização do método dos mínimos quadrados, sem nenhuma suposição para a distribuição de probabilidade dos erros. Além disso, a coerência matemática de qualquer previsão é controlada. Para isso, este capítulo também propôs uma extensão da transformação $Box - Cox$ para dados de $boxplots$ a fim de resolver o problema de coerência matemática.

Experimentos no âmbito de simulações de Monte Carlo sobre diferentes conjuntos de dados reais e sintéticos, bem como uma aplicação para dados de $boxplot$ da vida real, demonstram a robustez e adaptabilidade do modelo introduzido. Nesse contexto, torna-se uma boa opção analisar a dependência linear entre variáveis de $boxplot$ e prever valores desconhecidos de resposta com base nos valores das variáveis regressoras.

4 PREVISÃO DE SÉRIES TEMPORAIS SIMBÓLICAS DE BOXPLOTS

A previsão de séries temporais pode ser considerada como um problema de modelagem (SORJAMAA; HAO; LENDASSE, 2005): um modelo é construído entre as entradas e as saídas e, em seguida, é usado para prever os valores futuros com base nos valores anteriores.

No Capítulo 2, as séries temporais simbólicas de intervalos e histogramas foram abordadas. O intervalo representa a variabilidade de uma observação em um determinado momento através de um intervalo de valores. No entanto, o intervalo não informa o que acontece entre os extremos considerados, isto é, não indica como as observações são distribuídas dentro do intervalo. Por outro lado, o histograma permite cobrir esta falta, mas não permite ao analista focar nas características que mais lhe interessam, como: um conjunto de quantis, uma parte do intervalo da variável ou os valores dos $outliers$. Para isso, é necessário um dado simbólico que represente uma distribuição completa dos dados quantitativos e que identifique se existem $outliers$ e quais são seus valores.

Neste trabalho, utilizamos a previsão direta para realizar a previsões de longo prazo para séries temporais simbólicas multi-valoradas do tipo quartis, como caso especial séries representadas por $boxplots$. Este capítulo abordará diferentes aspectos deste novo tipo de séries temporais, como as diferentes agregações para sua obtenção, o uso de quartis e os $boxplots$. Além, disso, é descrita a abordagem proposta e o modelo desenvolvido.

4.1 PRECEDENTES DAS SÉRIES TEMPORAIS DE BOXPLOTS

Em publicações e sites dedicados a finanças geralmente mostram-se os preços de abertura, fechamento, mínimos e máximos atingidos por diferentes ativos financeiros durante o dia ou durante a semana. Esses quatro valores dão origem a uma representação gráfica chamada gráficos de velas ou $candlesticks$. Um exemplo dessa representação pode ser visto na Figura (12). Os gráficos de velas foram criados no século XVIII por Munehisa Homna, um comerciante de arroz japonês, para representar as variações diárias que o preço do arroz atingia no mercado. Na área de finanças, há toda uma teoria para interpretar esses gráficos em busca de sinais de compra ou de venda para apoiar a toma de decisões, ver, por exemplo, Morris (2006). No entanto, essa teoria está longe da análise estatística ou a predição de séries temporais. A existência dos $candlesticks$ mostra que nessa área é útil obter representações e previsões

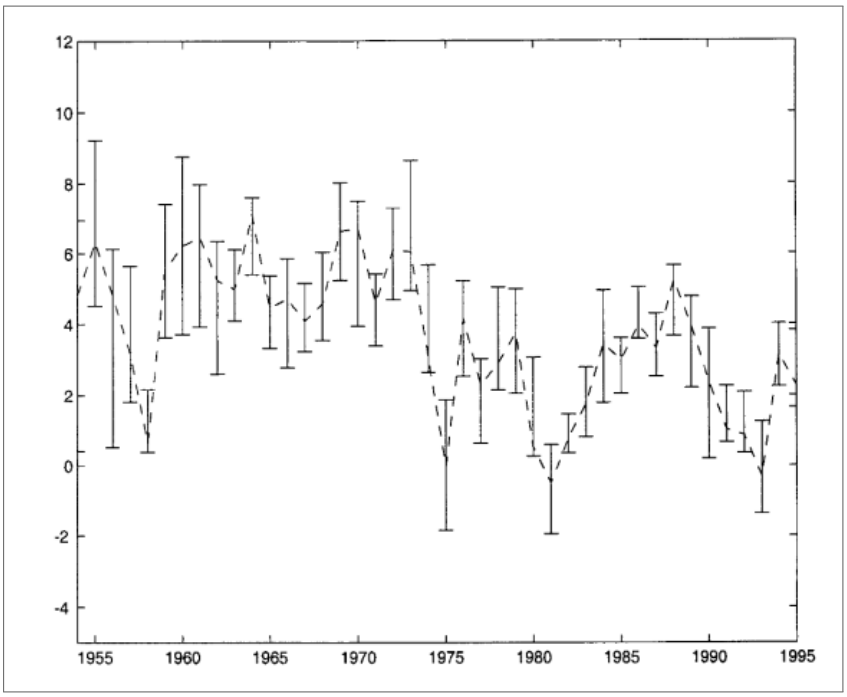
Figura 12 – Gráfico de velas das cotações diárias do USD entre 23 de junho e 16 de agosto de 2019

Fonte: (YAHOO, 2019)

temporais que transcendem as séries temporais pontuais e que representem a distribuição que os valores da ação seguem ao longo do dia.

Outro precedente interessante para a representação de séries temporais de $boxplots$ é mostrado em Zellner e Tobias (2000). Este artigo analisa as taxas de crescimento anual de 18 países industrializados, a fim de prever a mediana dessas taxas de duas maneiras, agregadas e desagregadas. Curiosamente, quando se trata de representar graficamente as séries temporais de medianas, os autores escolhem desenhar a mediana dentro de intervalos que representam o intervalo interquartil das 18 taxas de crescimento anual (Figura (13)). Se o mínimo e o máximo aparecessem a cada vez, a representação resultante seria perfeitamente um $boxplot$.

Figura 13 – Medianas e intervalos interquartil das taxas de crescimento anual (em %) de 18 países industrializados



Fonte: (ZELLNER; TOBIAS, 2000, p. 460)

4.2 SÉRIES TEMPORAIS SIMBÓLICAS MULTI-VALORADAS DE QUARTIS

No contexto de ADS, dados simbólicos multi-valorados são unidades descritas por subconjuntos finitos de entidades individuais com uma propriedade comum. A variabilidade entre diferentes entidades é levada em conta (DIDAY, 2016). Seja Y_s uma série temporal de s instantes de tempo ($s = 1, \ldots, N$). Considere a escolha de um intervalo temporal s (por exemplo: dia, semana, mês, etc.) para obter uma sequência de subconjuntos ou classes de números reais que pertencem a Y_s. Essa escolha depende das características específicas dos dados que o analista deseja estudar. Considere uma classe definida por uma sequência ordenada de v instantes sucessivos no tempo $\{y_1, \ldots, y_v\} \subset Y_s$.

Seja Q_t, $t = 1, \ldots, n$, uma série de dados numéricos com múltiplos valores indexados (listados ou representados graficamente) ordenados no tempo, em que $Q_t = \{Q1_t, Q2_t, Q3_t\}$, $Q1_t$ o quartil inferior, $Q2_t$ a mediana e $Q3_t$ o quartil superior da t-ésima classe de tamanho v descrita por $\{y_{t1}, \ldots, y_{tv}\}$ no domínio considerado que satisfaz: $-\infty < Q1_t \leq Q2_t \leq Q3_t < \infty$ $(t = 1, \ldots, n)$. Aqui, Q_t, $t = 1, \ldots, n$, pode ser visto como uma série de dados com múltiplos valores de quartis e o analista pode querer estudar um único quartil, dois quartis ou todos os quartis. Os quartis também são frequentemente usados como uma medida de dispersão dos

dados denominada intervalo interquartil em que: $I = Q3 - Q1$. A notação usada neste capítulo para representar as séries temporais e os dados simbólicos não coincide com a apresentada no Capítulo 2, mas é a que foi estimada como mais apropriada e menos confusa para trabalhar com quartis no contexto de séries temporais.

De acordo com a definição de variáveis simbólicas multi-valoradas apresentada por Diday (2016), os dados de quartis podem ser considerados como um caso particular de dados simbólicos multi-valorados na estrutura ADS. As razões são:

- São flexíveis para descrever classes, pois permitem ao analista selecionar um ou mais segmentos específicos de dados para estudar. Cada unidade é descrita por uma lista de vetores de valores.

- Ocupam menos espaço e tempo de execução quando comparados aos dados de histograma, o que é muito útil para comparar distribuições em grandes conjuntos de dados simbólicos.

- Permitem representar a distribuição com base nas características que leva em conta a variabilidade como o intervalo interquartil.

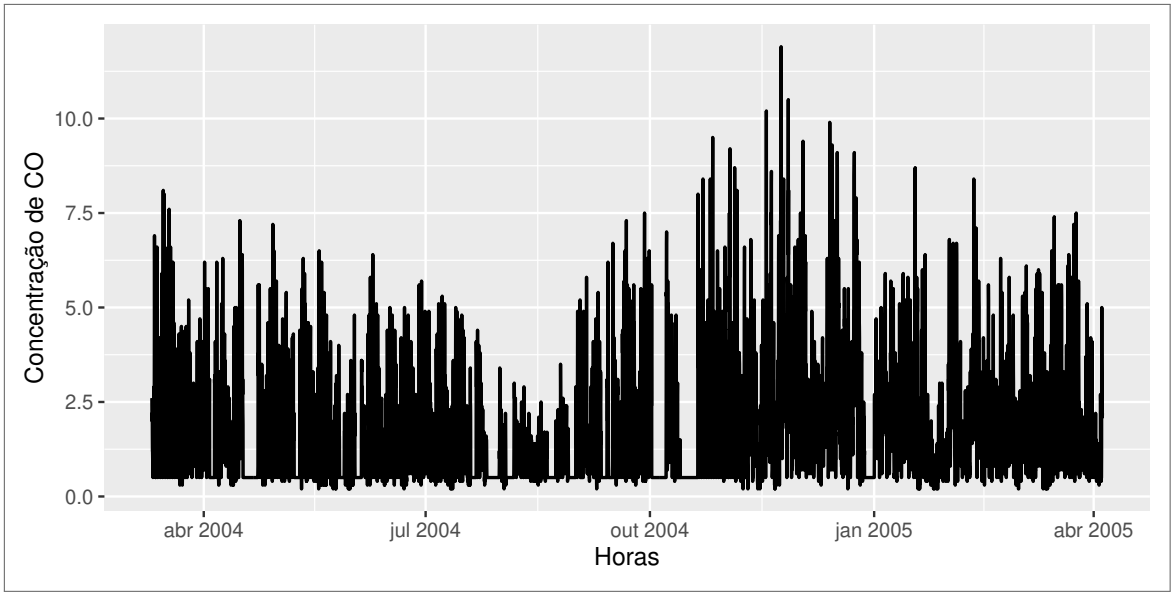
O exemplo a seguir ilustra o conceito de série temporal simbólica multi-valorada de tipo quartis. Considere uma série temporal dos registros da concentração de CO no ar por hora em uma cidade italiana (veja a Figura 14). O conjunto de dados contém 9357 instâncias de respostas por hora de uma matriz de 5 sensores químicos de óxido de metal integrados em um dispositivo multissensor de qualidade de ar. Os dados foram registrados entre março de 2004 e fevereiro de 2005 (um ano) e foram obtidos do repositório UCI [1].

Para ilustrar o processo de agregação em diferentes níveis, inicialmente, a série temporal foi agregada por dia produzindo 391 classes e cada uma delas é descrita por quartis simbólicos. A Figura (15) mostra as 3 curvas que se referem ao quartil inferior, à mediana e o quartil superior obtidos desta agregação. Pode se observar que as curvas obtidas são muito semelhantes à série temporal original e a tarefa de extração de conhecimento não é fácil.

No caso da agregação por semana, foram obtidas 57 semanas e a Figura (16) mostra as 3 curvas que se referem aos quartis simbólicos. Neste caso, as curvas são mais suaves que as curvas na Figura (15), especialmente a curva do quartil inferior ($Q1$), que apresenta um

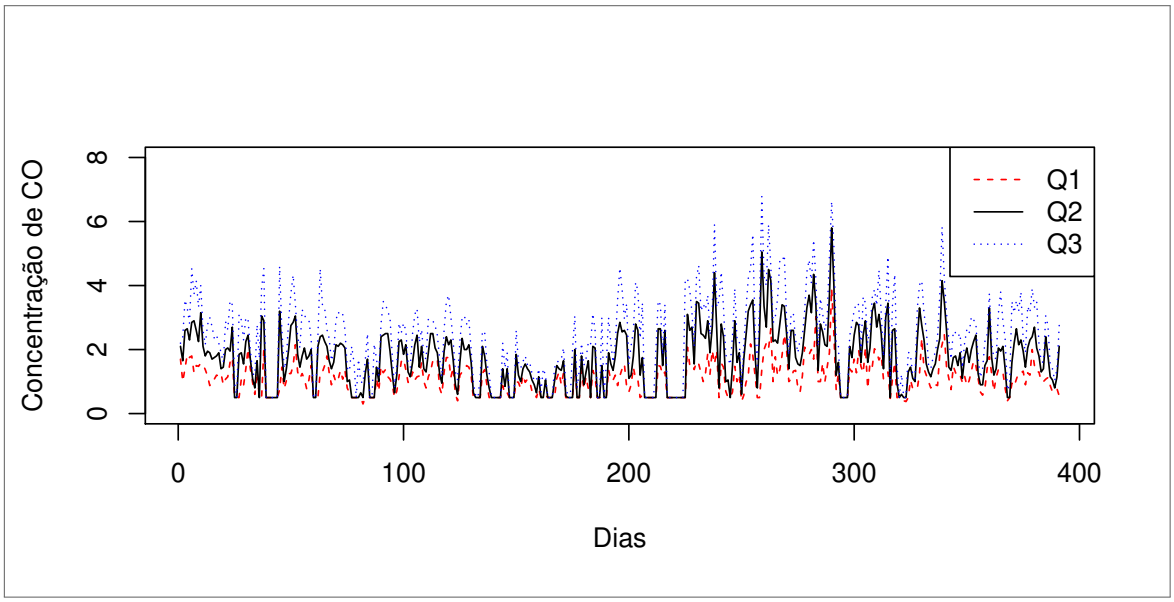
[1] Repositório de bancos de dados, teorias de domínio e gerador de dados que são usados pela comunidade de aprendizado de máquina para a análise empírica de algoritmos de aprendizado de máquina. Saiba mais em:(FAN; POH, 2009)

Figura 14 – Série temporal dos registros da concentração de CO no ar por hora em uma cidade italiana.



Fonte: Elaborada pelo autor (2021)

Figura 15 – Curvas do quartil inferior, mediana e quartil superior após agregação por dias.

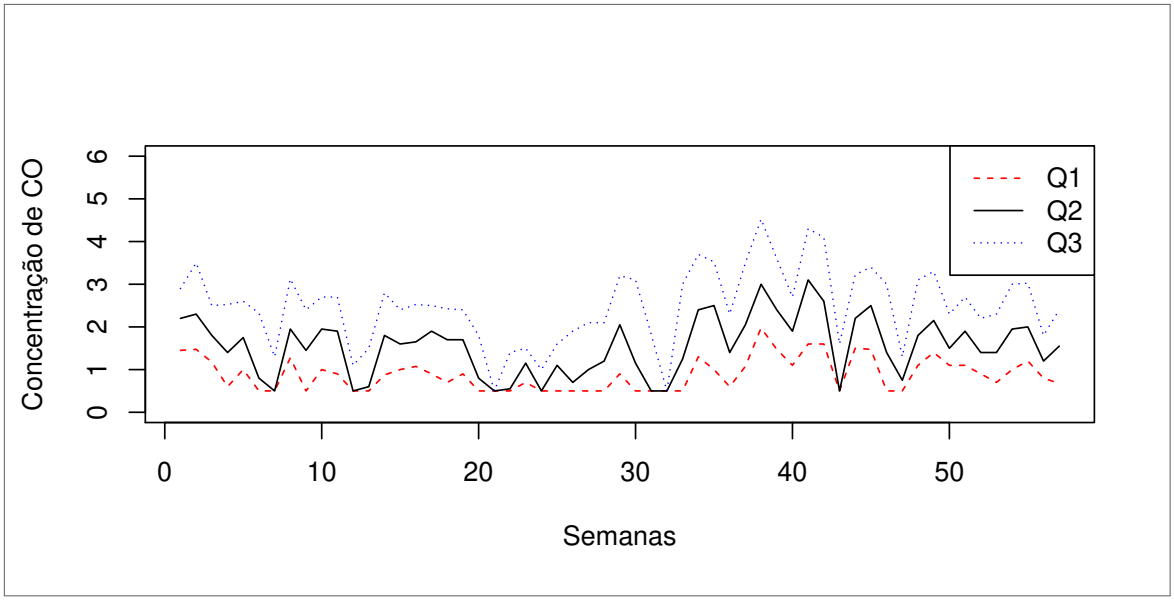


Fonte: Elaborada pelo autor (2021)

comportamento diferente das curvas que representam a mediana e o quartil superior ($Q2$ e $Q3$, respectivamente). O comportamento das curvas após a agregação por semanas facilita a extração de conhecimento quando comparado com a série original ou a agregação por dias.

Finalmente, a série temporal também foi agregada por mês e 14 classes foram obtidas. A Figura (17) mostra as 3 curvas que representam os quartis simbólicos. Nesta figura, pode-se observar que as curvas são muito mais suaves que os casos anteriores e o quartil inferior ($Q1$)

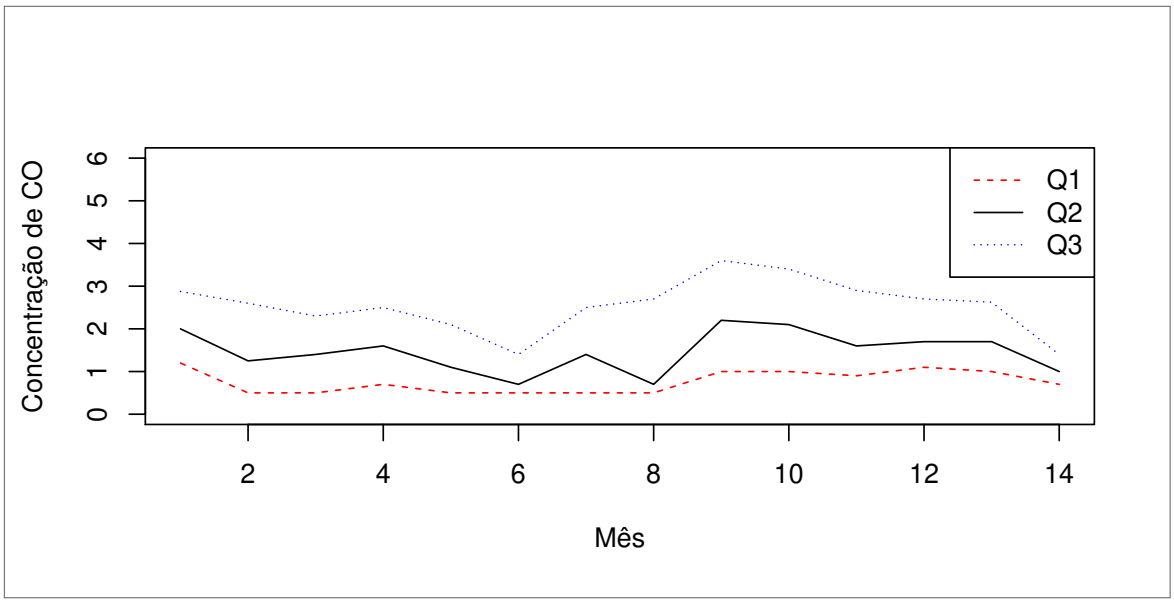
Figura 16 – Curvas do quartil inferior, mediana e quartil superior após agregação por semanas.



Fonte: Elaborada pelo autor (2021)

está muito próximo de ser uma reta e mostra comportamento bem diferente dos comportamentos das curvas da mediana ($Q2$) e o quartil superior ($Q3$).

Figura 17 – Curvas do quartil inferior, mediana e quartil superior após agregação por mês.



Fonte: Elaborada pelo autor (2021)

Analisando os 3 processos de agregação ilustrados, podemos recomendar a agregação por semanas, porque mantém o comportamento da série original e permite reduzir o conjunto de dados sem perda significativa de informação. No entanto, a escolha do período temporal para a agregação da série depende dos interesses do analista.

Os dados de quartis representam a variabilidade entre indivíduos nas classes como um todo e são menos sensíveis a $outliers$ do que os dados clássicos. Além disso, o conceito acima pode ser facilmente estendido para decis ou percentis. Estas são estatísticas que geralmente se aplicam a grandes conjuntos de dados clássicos, e sua escolha depende de um analista querer examiná-las. Neste contexto, a abordagem proposta também permite estimar $boxplots$, que são gráficos que mostram a distribuição de dados com base em um resumo de cinco valores: mínimo (m), quartil inferior ($Q1$), mediana ($Q2$), quartil superior ($Q3$) e máximo (M). Os $boxplots$ podem indicar se a distribuição de dados é simétrica, quão rigorosamente os dados são agrupados, se existem $outliers$ e quais são seus valores. Aqui, os valores mínimo e máximo dos $boxplots$ são obtidos por:

$$m = Q1 - 1.5(Q3 - Q1) \text{ and } M = Q3 + 1.5(Q3 - Q1).$$

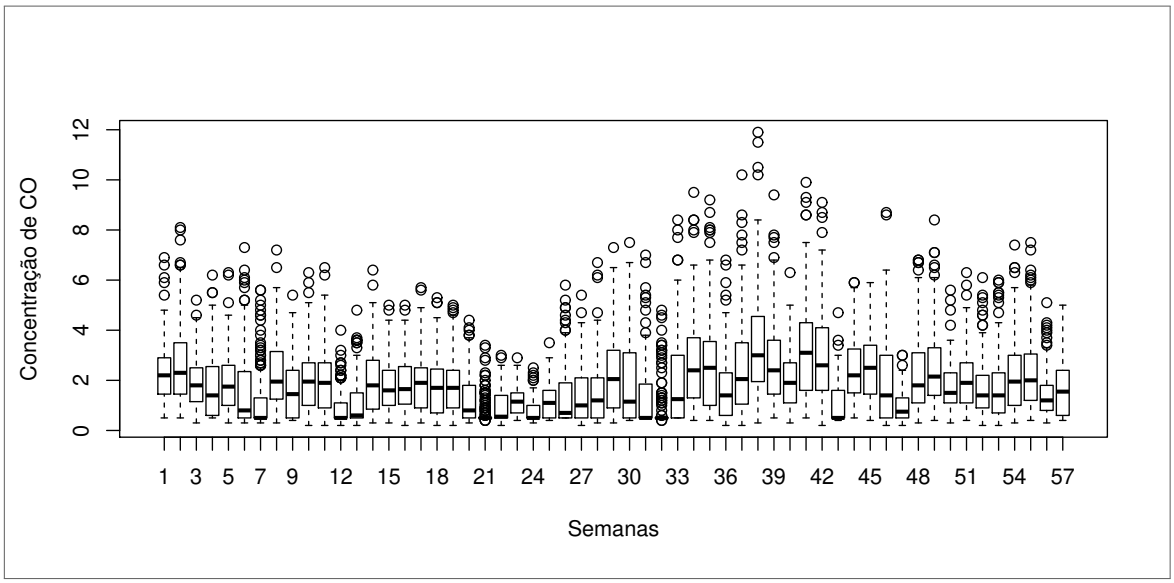
4.2.1 Séries Temporais Simbólicas de Boxplots

As séries temporais simbólicas de tipo $boxplots$ mostram conjuntamente, a posição dos dados (a mediana), a dispersão (o intervalo interquartil, que é a diferença entre o quartil superior e o quartil inferior), as caudas simétricas e de distribuição ao longo do tempo. A Figura (18) mostra uma série temporal simbólica de tipo $boxplots$ obtida a partir da agregação por semanas apresentadas anteriormente para a série temporal dos registros da concentração de CO no ar por hora em uma cidade italiana. Em geral, os $boxplots$ parecem preservar a estrutura das séries temporais, mas mostram padrões relevantes adicionais nos dados, por exemplo, $outliers$.

4.2.2 Seleção de Protótipos

Outra tarefa importante na previsão de séries temporais é selecionar um subconjunto representativo de entrada, de modo que, o armazenamento e o custo computacional diminuam, sem influenciar a precisão dos resultados. Nesse contexto cabe destacar a seleção de protótipos, que é uma técnica que visa a obtenção de instâncias representativas dos dados do problema em mãos. Tais instâncias são chamadas de protótipos. A obtenção destes protótipos, além de promover a redução dos dados, também tem por objetivo filtrar ruídos e compactar dados redundantes.

Figura 18 – Série temporal simbólica de tipo $boxplots$ dos registros da concentração de CO no ar por hora em uma cidade italiana.



Fonte: Elaborada pelo autor (2021)

Dentre as principais abordagens para a seleção de protótipos destacam-se: a abordagem incremental (AHA, 1992), a abordagem decremental (WILSON; MARTINEZ, 2000) e o método $batch$ (TOMEK, 1976a). Recentemente, também foram propostas outras abordagens, como algoritmos evolutivos (ISHIBUCHI; NAKASHIMA; NII, 2001), algoritmos baseados em $boosting$ (SEBBAN; NOCK; LALLICH, 2002), e técnicas de poda (ZUBEK; DIETTERICH, 2002).

A maioria dos métodos encontrados na literatura para a seleção de protótipos foi desenvolvido para problemas de classificação (PEKALSKA; DUIN; PACLíK, 2006), enquanto apenas alguns artigos focam na seleção de protótipos para regressão. Isso é devido ao fato de que o problema de seleção de protótipos para regressão é muito mais complexo. A razão é que nos problemas de classificação apenas os limites entre as classes devem ser determinados com precisão, enquanto nos problemas de regressão o valor de saída deve ser calculado adequadamente em cada ponto do espaço de entrada. Além disso, a decisão em problemas de classificação é frequentemente bidirecional ou há no máximo várias classes diferentes, enquanto que em problemas de regressão, a saída do sistema é contínua, portanto há um número ilimitado de valores possíveis previstos pelo sistema. Isso faz com que a compactação do conjunto de dados obtida pela seleção de protótipos possa ser muito maior em problemas de classificação do que em problemas de regressão. Além disso, a decisão sobre a rejeição de um determinado vetor nos problemas de classificação pode ser tomada com base na classificação correta ou incorreta do vetor por algum algoritmo.

Em problemas de regressão, deve-se estabelecer um limiar que defina a diferença entre o valor previsto e o valor real da saída do vetor. Determinar esse limite é um problema que precisa atenção. Outro problema é a medida de erro, que nos problemas de classificação é muito simples, enquanto na regressão, a medida de erro pode ser definida de várias maneiras e em soluções práticas nem sempre as simples definições de erros como o *Mean Square Error* (MSE) funcionam bem (KORDOS; BLACHNIK; WIECZOREK, 2011).

A continuação são relacionados alguns dos trabalhos mais relevantes encontrados na literatura. Wilson e Martinez (2000) por exemplo, apresentam uma análise de várias técnicas de redução comparando suas performances em 31 diferentes tarefas de classificação e propõem também um conjunto de seis algoritmos mais robustos quanto à presença de ruídos nos dados. Batista, Prati e Monard (2004) aplicaram técnicas de redução de instâncias: *Tomek Links* (TOMEK, 1976b), *Condensed Nearest Neighbor Rule* (CNN) (HART, 1968), *One-Sided Selection* (OSS) (KUBAT; MATWIN, 1997) e *Neighborhood Cleaning Rule* (NCR) (LAURIKKALA, 2001), como uma alternativa de solução para problema de classes desbalanceadas. Nesta tarefa de classificação, um dos métodos mais populares que usa a seleção de protótipos é o *Learning Vector Quantization* (LVQ) introduzido por Kohonen (1986). A partir de um conjunto inicial de protótipos aplica-se um algoritmo supervisionado que realiza o ajuste destes protótipos com a finalidade de aproximar as fronteiras de classificação. Em Salvador, Derrac e Ramon (2012) pode ser encontrada uma grande pesquisa que inclui quase 70 diferentes algoritmos de seleção de protótipos para problemas de classificação. Todos essas abordagens mencionadas conseguiram melhorar a qualidade e a velocidade dos algoritmos de classificação, filtrando ruídos e compactando exemplos redundantes.

Em problemas de regressão, Zhang, Yim e Yang (1997) apresenta um método para selecionar os vetores de entrada ao calcular a saída usando o algoritmo k-NN. Tolvi (2004) usa um algoritmo genético para selecionar os subconjuntos de entrada em modelos de regressão linear. Em seus trabalhos, Guillen, Herrera e Rubio (2008), Guillen et al. (2009), Guillen et al. (2010) discutiram o conceito de informação mútua usada para seleção de protótipos em problemas de regressão. Mais recentemente, Kordos, Białka e Blachnik (2013) apresentam três algoritmos de seleção de protótipos para problemas de regressão, que ampliam as capacidades dos algoritmos CNN, *Edited Nearest Neighbor Rule* (ENN) e *Cellular Automata* (CA) usados para problemas de classificação. A pesquisa apresentada aqui é motivada pela abordagem de seleção de protótipos proposta por Guillen et al. (2010) baseada na informação mútua. Os autores propõem um estimador de Informação Mútua (IM) baseado em vizinhos mais próximos,

o que permite estimar IM diretamente do conjunto de dados. O objetivo é escolher exemplos relevantes de um conjunto de dados, sem a necessidade de gerar novos dados, permitindo remover os valores discrepantes e o ruído de conjuntos de dados altamente distorcidos. O algoritmo aplicado determina a perda de IM em relação aos seus vizinhos de tal forma que se a perda de IM é semelhante à instância próxima estudada, então esta instância deve ser incluída no conjunto de dados de treinamento. Essa abordagem se mostrou bem sucedida em situações em que foi aplicada a conjuntos de treinamento artificialmente distorcidos, adicionando ruído ou $outliers$.

4.2.2.1 Informação Mútua

O conceito clássico de informação mútua (também chamado entropia cruzada) torna a teoria da informação uma estrutura interessante para abordagens de filtragem, já que é considerada como um bom indicador de relevância entre duas variáveis aleatórias (COVER; THOMAS, 1991).

Sejam Z e W duas variáveis aleatórias contínuas, com uma função de densidade de probabilidade conjunta $p(z, w)$ e probabilidades marginais $p(z)$ e $p(w)$, respectivamente. A informação mútua entre Z e W pode ser calculada como (COVER; THOMAS, 1991)

$$IM(Z;W) = \int\int p(z,w)\ log\frac{p(z,w)}{p(z)p(w)}dzdw. \tag{4.1}$$

A equação 4.1 pode ser reescrita em termos de entropia e entropia condicional, respectivamente definidas como (SHANNON, 1948)

$$H(Z) = -\int p(z)\ log\ p(z)dz \tag{4.2}$$

e

$$H(W|Z) = -\int\int p(z,w)\ log\frac{p(z)}{p(z,w)}dzdw. \tag{4.3}$$

Usando a equações 4.1, 4.2 e 4.3, é possível escrever a informação mútua como

$$IM(Z;W) = H(W) - H(W|Z). \tag{4.4}$$

$I(Z;W)$ tem duas propriedades principais que o distinguem de outras medidas de dependência: primeiro, a capacidade de medir qualquer tipo de relação entre variáveis. Esta propriedade tem sua raiz na medida em que a informação mútua é construída a partir de probabilidades conjuntas e marginais das variáveis e não utiliza estatísticas de qualquer grau

ou ordem. A segunda propriedade é sua variância em transformações espaciais. Isto é baseado no fato de que o argumento do logaritmo na equação 4.1 é não-dimensional, assim o valor integral não depende das coordenadas escolhidas (KULLBACK, 1997).

4.2.3 Algoritmo de Seleção de Protótipos usando Informação Mútua

Quando os conjuntos de dados do mundo real são examinados, a necessidade imperativa de métodos de seleção de protótipos se torna cada vez mais clara. Por um lado, o tamanho médio do conjunto de dados está se tornando cada vez maior. Por outro lado, conjuntos de dados reais geralmente contêm instâncias ruidosas, $outliers$ e anomalias. A seleção de um subconjunto adequado de instâncias é, portanto, uma boa opção para diminuir o tamanho da amostra, possibilitando seu tratamento posterior.

Existem duas abordagens para a seleção de protótipos: $Wrapper$ (tenta projetar o modelo ao mesmo tempo em que realiza a seleção dos protótipos) e $Filter$ (consistem em um pré-processamento dos dados de entrada para que o modelo seja construído posteriormente). O conceito clássico de informação mútua torna a teoria da informação uma estrutura interessante para abordagens de $Filter$, pois é considerada um bom indicador de relevância entre duas variáveis aleatórias (COVER; THOMAS, 1991).

A abordagem de previsão apresentada neste documento aplica a cada classe de uma série temporal simbólica Q_t, o método de seleção de protótipos baseado na informação mútua que foi proposto por Guillen et al. (2010).

Para fazer isso, a informação mútua de cada classe descrita pelo conjunto de dados univariado $W = \{y_1, \ldots, y_v\}$ é computada como a quantidade de conhecimento obtido de $\{y_{s+h}\}$ quando se observa $\{y_s\}$. Após a equação (4.4) considere $Z = \{y_s\}$ com $(s = 1, \ldots, (v-1))$ como um subconjunto de dados univariado de tamanho $v-1$. Aqui, a informação mútua pode ser entendida como o grau redução de incerteza (medido pela entropia) nos valores de W quando Z é conhecido. Se W denota uma saída de destino para prever e Z um subconjunto de dados de entrada, a informação mútua tem uma interpretação bastante natural como um critério de seleção.

Inicialmente, todas as unidades de uma classe são protótipos. Portanto, se for calculada a informação mútua entre os vetores de entrada e de saída, é possível dizer que, se um protótipo significativo for removido dos dados de entrada, a informação mútua relativa aos seus vizinhos diminuirá e vice-versa. Assim, para avaliar a relevância de um protótipo, calcula-

se a informação mútua entre Z e W, eliminando este protótipo de cada vez (com reposição) do conjunto W. Além disso, os K vizinhos mais próximos deste protótipo são considerados baseados na distância euclidiana e suas informações mútuas são calculadas. Um parâmetro (α), chamado "limiar de informação mútua" determina quão diferente a informação mútua do protótipo deve ser em comparação à informação mútua de cada um dos seus vizinhos. O algoritmo (1) apresentado a seguir foi desenvolvido para o método de seleção de protótipos baseado no critério de informação mútua.

Data: $W = \{y_1, \ldots, y_v\}$; Z; α; K; v

Result: Um subconjunto de protótipos selecionados δ

begin

Definir M como uma matriz $v \times K$;

for $i = 1 : v$ **do**

Computar os K vizinhos mais próximos (K-NN) de y_i de W;

Salvar os índices i das linhas da matriz M.

end

Definir $\Psi = \{\Psi_1, \ldots \Psi_v\}$;

for $i = 1 : v$ **do**

$Z_i = Y$ sem i-ésimo elemento de W;

Calcular $\Psi_i = IM(Z_i; W)$ usando a equação (4.4);

end

Normalizar Ψ no range [0, 1] ;

`/* MODULO FILTRAR */`

for $i = 1 : v$ **do**

$cont = 0$;

for $j = 1 : K$ **do**

$D = \Psi_i - \Psi_{M(i,j)}$;

if $D > \alpha$ **then**

$cont = cont + 1$;

end

end

if $cont < K$ **then**

$\delta = \delta \cup W_i$ `/* Seleciona o protótipo */`

end

end

retornar δ como o conjunto de protótipos selecionados;

end

Algoritmo 1: Algoritmo para a seleção de protótipos usando informação mútua

Para explicar como o algoritmo funciona, um exemplo Toy é descrito a continuação. Considere como dados de entrada: uma classe de séries temporais com cinco elementos representados pelo vetor $W =$ (0,1064, 0,3803, 0,9427, 0,2161, 0,6775); $\alpha =$ 0,01; $K =$ 3 e $v =$ 5.

Inicialmente, são calculados os $3 - NN$ para cada elemento de W. Cada linha da matriz M é preenchida com índices de 3 vizinhos mais próximos de um elemento de W. Assim, a matriz M é dada como:

$$\begin{bmatrix} 4 & 2 & 5 \\ 4 & 1 & 5 \\ 5 & 2 & 4 \\ 1 & 2 & 5 \\ 3 & 2 & 4 \end{bmatrix}$$

A seguir, a informação mútua entre Z e W, $IM(Z, W)$, é calculada da seguinte forma: Para $i = 1$; $Z = \{$0,3803, 0,9427, 0,2161, 0,6775$\}$ e temos $\Psi_1 = IM(Z, W) = 0,445$. Para $i = 2$; $Z =$ (0,1064, 0,9427, 0,2161, 0,6775) e $W =$ (0,1064, 0,3803, 0,9427, 0,2161, 0,6775), temos $\Psi_2 = IM(Z, W) = 0,293$. Após calcular $\{\Psi_1, \ldots, \Psi_v\}$ e normalizar esses valores no intervalo [0, 1], temos

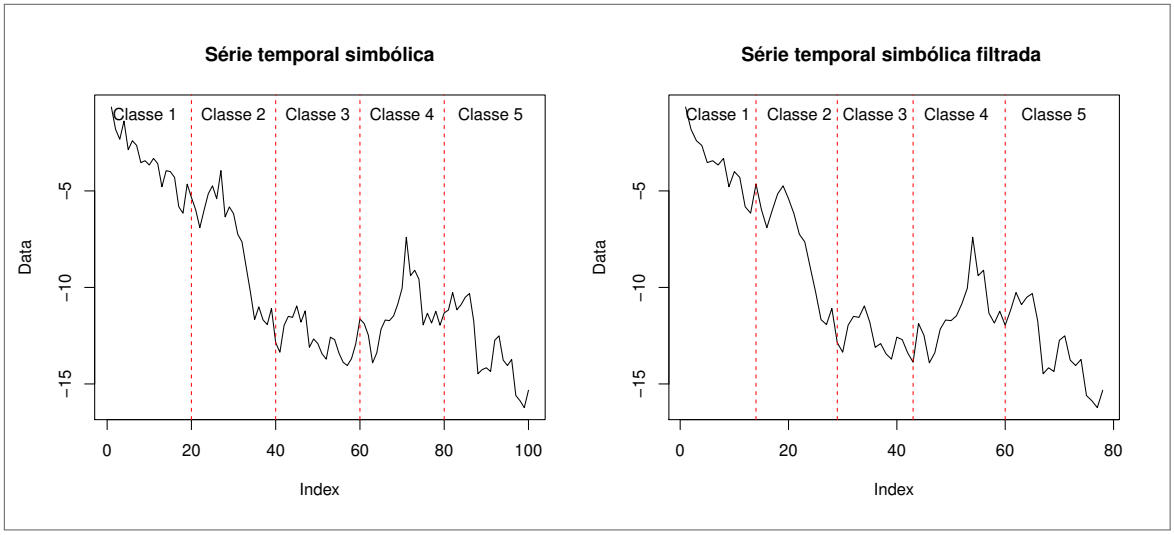
$$\Psi = \{0,445; 0,293; 0,165; 0,064; 0,827\}$$

Na próxima etapa, os valores de Ψ são filtrados de acordo com a diferença entre Ψ_i $(i = 1 \ldots, 5)$ e Ψ_j $(j = 1, 2, 3)$ onde j está associado ao vizinho mais próximo do elemento Ψ_i. Assim, para $i = 1$; temos $cont = 0$ e $\Psi_1 = 0,445$: para $j = 1$; $D = 0,445 - 0,064$ como $D > 0,01$ então $cont = 0 + 1$; para $j = 2$; $D = 0,445 - 0,293$ como $D > 0,01$ então $cont = 1 + 1 = 2$; para $j = 3$; $D = 0,445 - 0,827$ como $D < 0,01$ então $cont = 2$. Como $cont = 2 < K = 3$, o primeiro elemento de W é selecionado como protótipo, ou seja, $\delta_1 = W_1 =$ 0,1064. Após verificar se cada elemento de W é um protótipo, temos como saída do algoritmo para este exemplo Toy: $\delta = \{$0.1064, 0.3803, 0.9427, 0.2161$\}$.

A Figura (19) apresenta parte de uma série temporal simbólica antes e depois da seleção do protótipo. O exemplo é uma caminhada aleatória distorcida com valores aleatórios de tamanho $n = 1000$. A série temporal simbólica foi obtida pela agregação de dados em 50 classes de tamanho 20 cada. Após a aplicação do Algoritmo (1) para cada classe, a série temporal original foi reduzida para 26,9% e a nova série obtida tem 739 protótipos. Podemos observar a partir desta figura que ambas as séries temporais são muito semelhantes, embora tenham tamanhos

diferentes. Ou seja, o comportamento da série temporal filtrada não mudou, mostrando que a seleção do protótipo parece efetivamente reduzir o conjunto de dados sem perder informações significativas.

Figura 19 – Parte de uma série temporal simbólica antes e depois da seleção do protótipo.



Fonte: Elaborada pelo autor (2021)

4.2.4 Modelo Autoregressivo Vetorial

O VAR é um dos modelos mais bem sucedidos, flexíveis e fáceis de usar para a análise de séries temporais multivariadas. O modelo VAR provou ser especialmente útil para descrever o comportamento dinâmico de séries temporais econômico-financeiras e para previsão. Muitas vezes, fornece previsões superiores às de modelos de séries temporais univariadas e modelos elaborados de equações simultâneas baseados em teoria (HAMILTON, 1994).

O objetivo de desenvolver um VAR é identificar as relações entre séries temporais lineares para obter previsões acuradas e precisas. Em um modelo VAR estrutural, a trajetória temporal de cada variável é influenciada pelas defasagens de todas as variáveis incluídas. Seja, $y_t = (y_{1t}, y_{2t}, ..., y_{nt})'$ um vetor de variáveis de séries temporais $(n \times 1)$. O modelo básico do VAR(p) tem a forma:

$$y_t = \theta + \Phi_1 y_{t-1} + \Phi_2 y_{t-2} + ... + \Phi_p y_{t-p} + \varepsilon_t, \tag{4.5}$$

em que, Φ_i são $(n \times n)$ matrizes de coeficientes invariantes no tempo e ε_t é um vetor de ruído branco com média zero não observável $(n \times 1)$ e com matriz de covariância invariável no tempo $\sum$.

Considere três variáveis de séries temporais diferentes, denotadas por: $(y_{t1}, y_{t2}$ e $y_{t3})$. A equação do VAR para cada uma é dada por:

$$y_{1t} = \theta_1 + \phi_{11} y_{1t-1} + \phi_{12} y_{2t-1} + \phi_{13} y_{3t-1} + \varepsilon_{1t},$$

$$y_{2t} = \theta_2 + \phi_{21} y_{1t-1} + \phi_{22} y_{2t-1} + \phi_{23} y_{3t-1} + \varepsilon_{2t},$$

$$y_{3t} = \theta_3 + \phi_{31} y_{1t-1} + \phi_{32} y_{2t-1} + \phi_{33} y_{3t-1} + \varepsilon_{3t}.$$

Observe que noVAR, y_{1t}, y_{2t} e y_{3t} são relacionadas por meio de sua covariância. Além disso, cada equação tem os mesmos regressores — valores defasados. Portanto, o modelo VAR(p) é apenas um modelo de regressão aparentemente não relacionado com variáveis defasadas e termos determinísticos como regressores comuns.

A estimação dos parâmetros e da matriz de covariância do VAR é simples. Para $Y_t = (y_1, y_2, ..., y_T)$ e $Z = (z_1, z_2, ..., z_T)$ com z como um vetor de valores defasados de y e possíveis termos determinísticos, o estimador de mínimos quadrados dos parâmetros é $\hat{A} = YZ(ZZ')^{-1}$. A matriz de covariância é então obtida de $\frac{1}{T-K}(Y - \hat{A}Z)(Y - \hat{A}Z)'$, em que K é o número de parâmetros estimados.

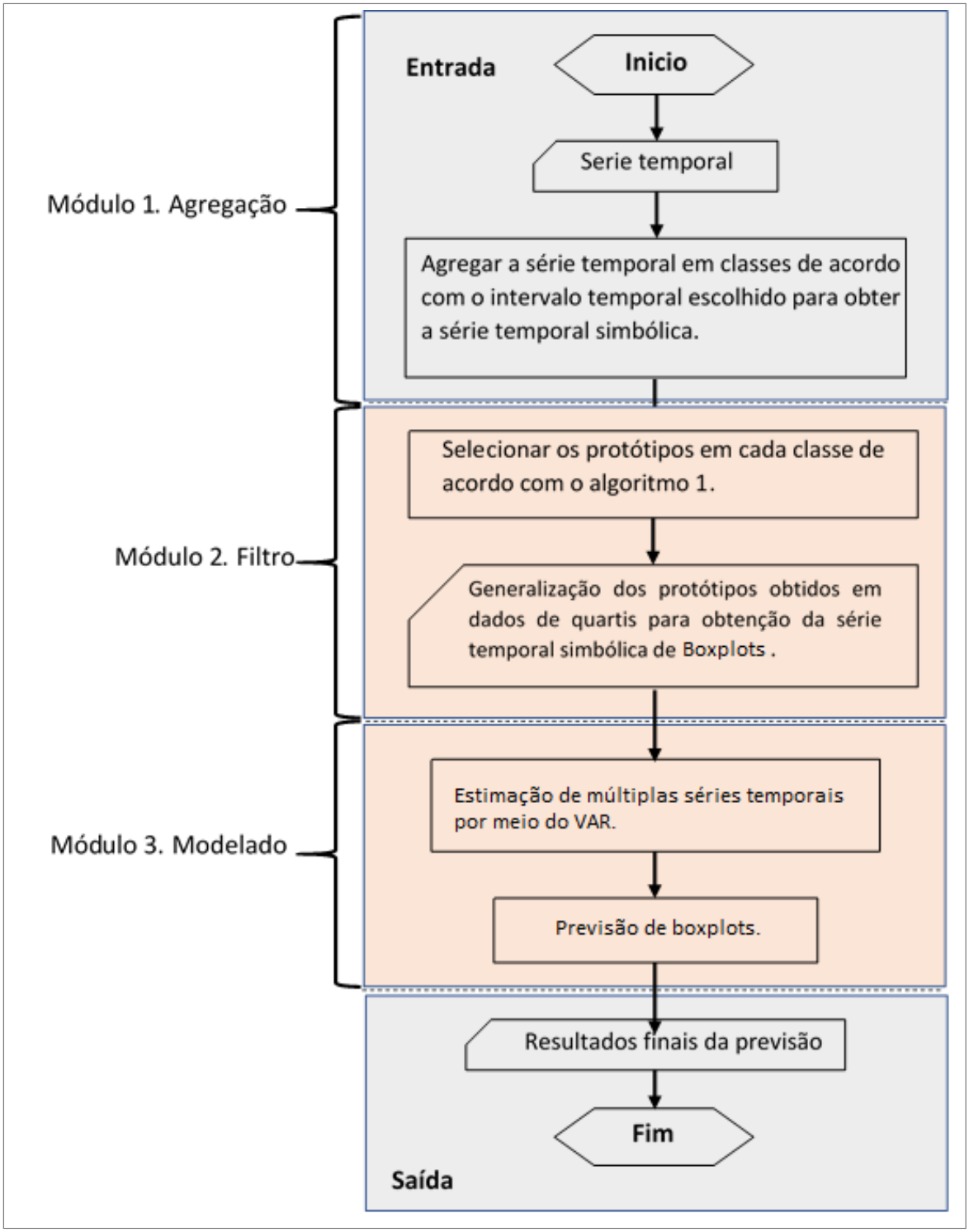
4.3 ABORDAGEM PROPOSTA

Nesta seção é apresentada uma abordagem em três etapas para previsão de séries temporais que combina o uso de dados simbólicos multi-valorados e seleção de protótipos para reduzir conjuntos de dados de alta dimensão, eliminar ruídos e extrair novos conhecimentos. A Figura (20) fornece uma visão geral da abordagem proposta.

Inicialmente, (i) os dados da série temporal são agregados em classes relacionadas ao intervalo temporal escolhido (dia, semana e mês). Essa escolha dependerá das características específicas dos dados que o analista deseja estudar. (ii) o algoritmo (1) é aplicado para encontrar classes mais representativas e são construídos $boxplots$. (iii) é feita a estimação de múltiplas séries temporais aplicada a dados de quartis simbólicos por meio do VAR.

Em comparação com o trabalho introduzido por (DRAGO, 2015), a nossa proposta tem três vantagens: i) a representação por uma lista de três valores numéricos reduz a quantidade de curvas ajustadas e permite gerar um modelo parcimonioso sem perda de precisão, ii) uso de um modelo que ajusta três curvas conjuntamente enquanto que o trabalho apresentado em (DRAGO, 2015) ajusta 5 curvas independentemente, e iii) devido ao uso da seleção de

Figura 20 – Abordagem proposta em três etapas.



Fonte: Elaborada pelo autor (2021)

protótipos, a abordagem resulta em uma série temporal robusta contra os efeitos do ruído nos $boxplots$ estimados. A abordagem discutida em(DRAGO, 2015) não usa filtro. Além disso, nossa abordagem é o primeiro método de séries temporais múltiplas para dados de quartis simbólicos.

Em comparação com a abordagem clássica para dados pontuais baseada no modelo *Autoregressive Integrated Moving Average* (ARIMA), a nossa abordagem permite levar em conta a variabilidade quando o analista tem interesse nas unidades de tempo de um nível superior da população que está sendo estudada, como por exemplo, nível de dia quando os dados são descritos por nível de hora ou minuto. Nesta situação, é necessário agregar dados e representá-los sem perda de muitas informações. Aqui o ADS estende a entrada padrão a um conjunto de classes de entidades individuais que associam a cada classe um valor simbólico. O modelo de dados clássico associa a cada classe um único valor, por exemplo, a média.

4.3.1 Modelo para Previsão de Séries Temporais de Boxplots

Conforme discutido anteriormente, a abordagem proposta neste capítulo considera séries temporais simbólicas representadas por $boxplots$ como um caso especial de séries temporais simbólicas multi-valoradas do tipo quartis. Cada classe desta série de tamanho n é descrita por uma variável simbólica multi-valorada do tipo quartis Q_t como:

$$Q_t = \{q1_t, q2_t, q3_t\}$$

Em forma matricial (notação mais compacta) como:

$$\begin{bmatrix} q1_t \\ q2_t \\ q3_t \end{bmatrix} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix} + \begin{bmatrix} \phi_{11} & \phi_{12} & \phi_{13} \\ \phi_{21} & \phi_{22} & \phi_{23} \\ \phi_{31} & \phi_{32} & \phi_{33} \end{bmatrix} \times \begin{bmatrix} q_{1t-1} \\ q_{2t-1} \\ q_{3t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \\ \varepsilon_{3t} \end{bmatrix}$$

Nestes modelos anteriores: θ e ϕ são os parâmetros a serem estimados pelo método dos OLS.

Com base em $\hat{Q} = \{\hat{q1}_t, \hat{q2}_t, \hat{q3}_t\}$, a regra de previsão para construir os $boxplots$ estimados é dada por:

$$\hat{m}_t = \hat{q1}_t - 1.5\hat{I}_t; \quad \hat{M}_t = \hat{q3}_t + 1.5\hat{I}_t; \tag{4.6}$$

onde, $\hat{I} = \hat{q3}_t - \hat{q1}_t$ é o intervalo interquartil previsto.

O procedimento completo para a modelagem das séries temporais simbólicas que é realizado até aqui é descrito a seguir:

(1. Entrada dos dados)

- Seja $\{Y_s\}$, $s = 1, \ldots, N$ uma série temporal com N valores numéricos.
- Agregar a série em classes: Analise a frequência da série temporal e defina o interesse da pesquisa. Gere amostras menores da série, para que cada partição possa representar um intervalo de tempo específico s.

(2. Filtragem e generalização dados)

- Utilize o algoritmo (1) para selecionar os protótipos em cada classe agregada $\{y_1, \ldots, y_v\}$.
- Generalize os dados para construir $boxplots$ com os protótipos obtidos em cada classe.
- Obtenha a série temporal simbólica de $boxplots$ Q_t.

(3. Previsão de boxplots)

- Ajuste o modelo VAR.
- Compute el valor mínimo $\{\hat{m}_t\}$, o máximo $\{\hat{M}_t\}$ do $boxplot$ baseado nas estimativas obtidas dos modelos VAR.
- Construa os novos $boxplots$.
- Compute o MMRE.

4.4 ANÁLISE EXPERIMENTAL

Esta seção mostra a validação da abordagem proposta através de experimentos com séries temporais simuladas de diferentes graus de dificuldade. Está organizado com base em quatro análises diferentes:

1. O primeiro apresenta um estudo avaliando a importância de filtrar uma série temporal.
2. O segundo compara o modelo VAR multivariado com o ajuste de três modelos ARIMA independentes.

3. O terceiro avalia a importância de considerar três curvas ao invés de cinco e compara nossa abordagem com a proposta introduzida em (DRAGO, 2015).

4. O quarto compara nossa abordagem com uma abordagem clássica usando o modelo ARIMA para séries temporais descritas por uma unidade agregada (média).

Inicialmente, a precisão dos modelos será avaliada em termos do MMRE (FAGUNDES; SOUZA; CYSNEIROS, 2014) que é estimado através de simulações de Monte Carlo, usando o método $Hold\,Out$. É importante ressaltar que para calcular o MMRE usando uma lista de três valores contínuos, os valores $\hat{m}_t$ e $\hat{M}_t$ são calculados pela equação (4.6) a partir dos valores previstos do primeiro, segundo e terceiro quartil.

O critério proposto de MMRE para dados simbólicos de $boxplots$ é dados por:

$$MMRE = \frac{1}{5t}\sum_{i=1}^{t}\left\{\left|\frac{m_t - \hat{m}_t}{m_t}\right| + \left|\frac{q1_t - \hat{q1}_t}{q1_t}\right| + \left|\frac{q2_t - \hat{q2}_t}{q2_t}\right| + \left|\frac{q3_t - \hat{q3}_t}{q3_t}\right| + \left|\frac{M_t - \hat{M}_t}{M_t}\right|\right\}$$

Além disso, foram considerados o valor médio e o desvio padrão do MMRE, bem como o teste t-Student para amostras pareadas ao nível de significância de 5%, o qual foi utilizado para comparar os resultados entre as abordagens. As estimações e cálculos foram implementados na linguagem R(TEAM, 2006). O R é um ambiente de programação para realização de análises estatísticas de dados e de análises gráficas. Trata-se de uma linguagem orientada a objetos que corresponde a uma versão ampliada e aprimorada da linguagem S. O R é um programa bastante flexível, gratuito e de código livre e encontra-se disponível em https://www.r-project.org. Uma revisão detalhada sobre este ambiente de programação pode ser vista no livro de Venables e Ripley (2002). Estes códigos encontram-se disponível em: https://github.com/Maite-analista/codes_thesis.

4.4.1 Séries Temporais Simuladas

A escolha de séries artificiais se deve ao fato da manipulação da série possibilitando que se agreguem características distintas, como por exemplo: tendências, sazonalidade e ruído. Com essas modificações é possível avaliar a efetividade dos modelos propostos em situações diversas e, se possível, identificar qual representação é recomendada em cada situação. Os experimentos descritos nessa seção consistiram em uma sequência de algoritmos organizados no $framework$ de uma simulação de Monte Carlo com 1000 iterações. Foram escolhidas 3 configurações com características diferentes para a geração das séries temporais artificiais.

A Tabela (11) exibe as configurações que foram usadas, o tamanho de cada série temporal gerada e o número de classes para à agregação.

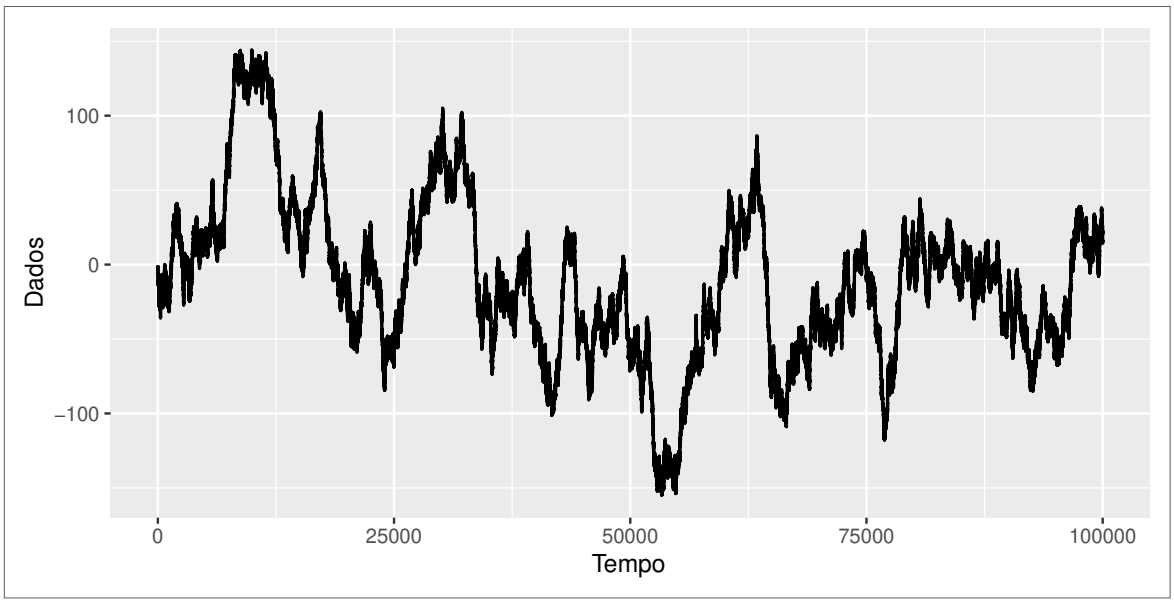
Tabela 11 – Configurações usadas para simular as séries temporais.

Configurações	Equações	N	Quantidade de classes
C_1	$y_t = y_{t-1} + \varepsilon_t(i)$	100000	100
C_2	$x_{n+1} = 0.7x_n(1 - x_n)(ii)$	100000	250
C_3	$x_{n+1} = 1 - 1.4x_n^2 + 0.3x_{n-1}(iii)$	100000	400

Fonte: Elaborada pelo autor (2021)

A primeira configuração, C_1, é o chamado processo de passeio aleatório ($random\ walk$). Um passeio aleatório é definido como um processo onde o valor atual de uma variável é composto pelo valor passado mais um termo de erro definido como um ruído branco (uma variável normal com média zero e variância um). A figura (21) mostra a série temporal gerada com 100000 dados para essa primeira configuração. Para agregação da série temporal foram obtidas 100 classes com 1000 valores cada.

Figura 21 – Série temporal clássica simulada para a configuração C_1

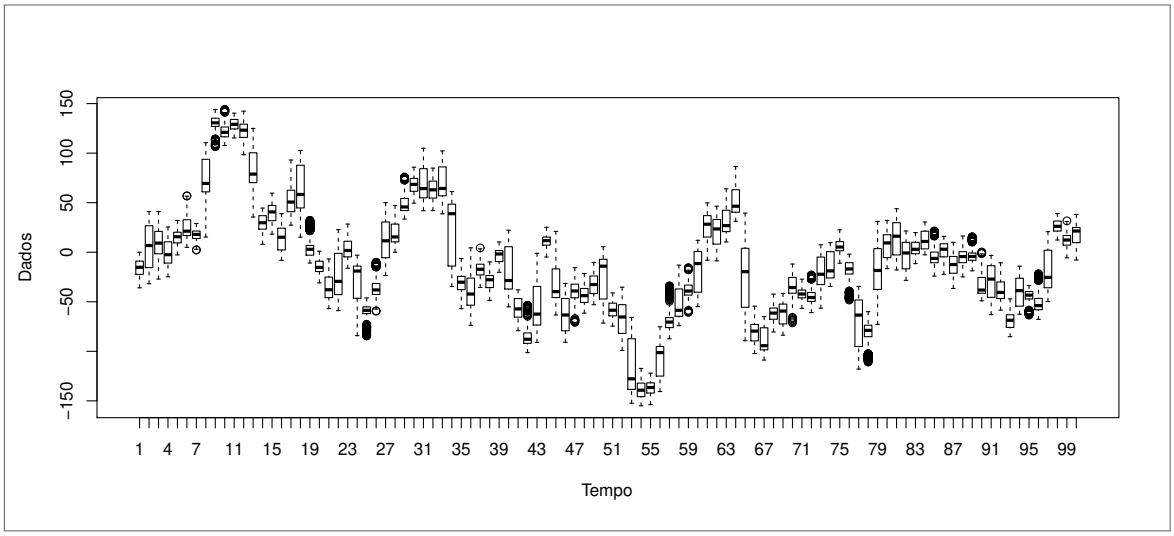


Fonte: Elaborada pelo autor (2021)

A Figura (22) mostra a série temporal simbólica de $boxplots$, onde pode-se observar o mesmo comportamento aleatório da série gerada originalmente. Além disso, os $boxplots$ mostram a variabilidade dos dados e os $outlier$ em cada classe agregada. A Figura (23) contém as 3 curvas dos quartis que descrevem esses $boxplots$. Note que as curvas obtidas são muito seme-

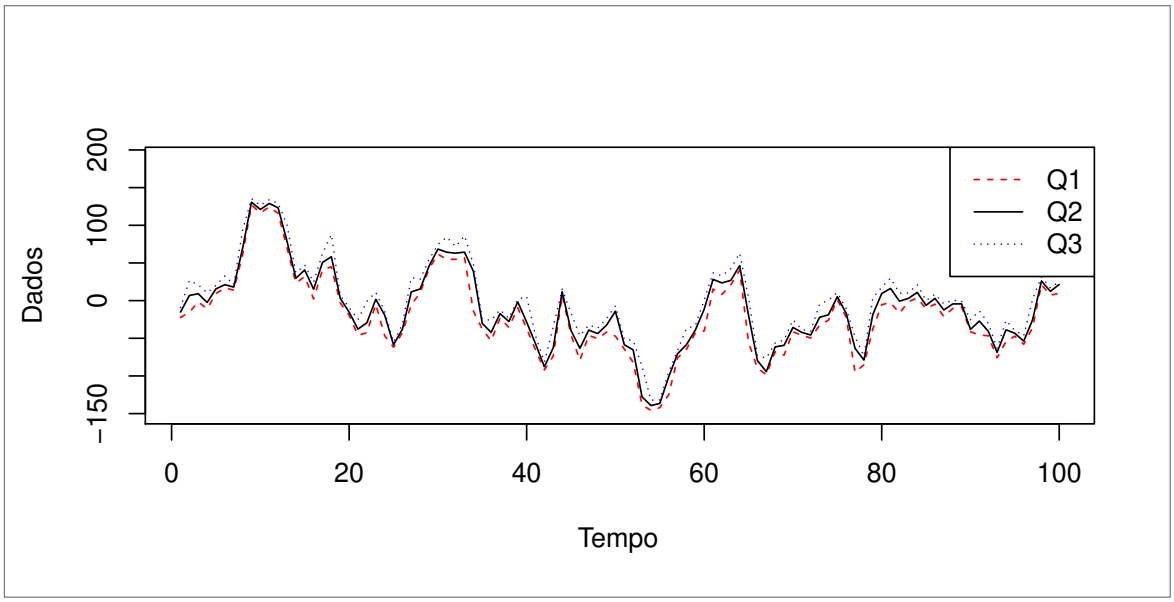
lhantes à série temporal originalmente gerada, corroborando que não houve perda significativa de informação.

Figura 22 – Série temporal de $boxplots$ para a configuração C_1



Fonte: Elaborada pelo autor (2021)

Figura 23 – Curvas da série temporal simbólica Q_t para configuração C_1

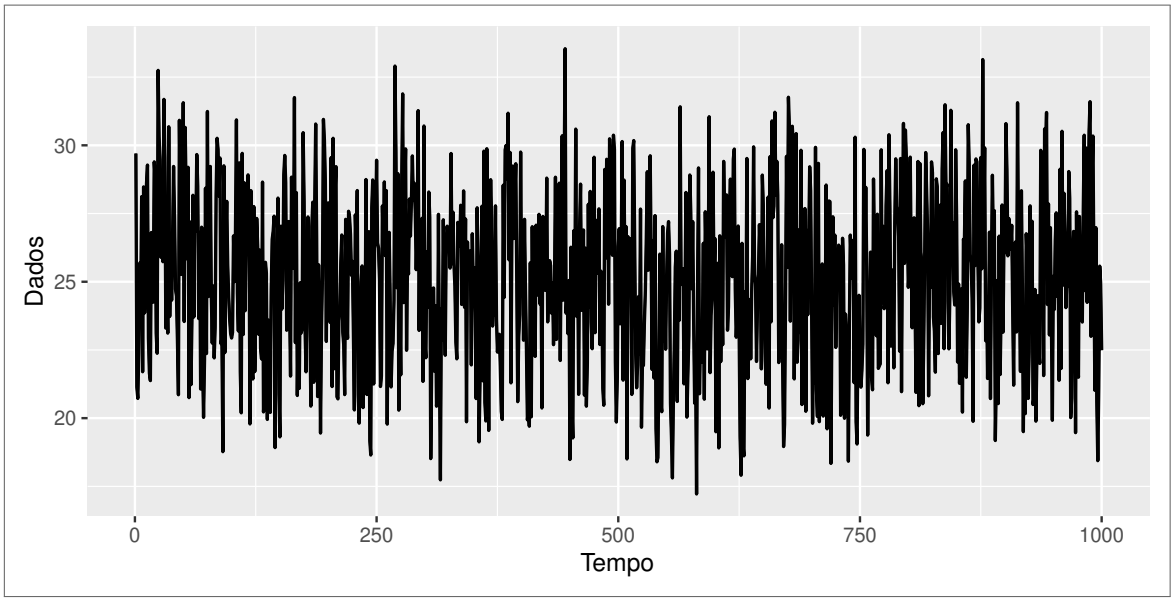


Fonte: Elaborada pelo autor (2021)

A configuração C_2 apresenta o mapa logístico que foi popularizado em um artigo de 1976 do biólogo Robert May (MAY, 1976). Este modelo é baseado na função logística comum da curva-s que mostra como uma população cresce lentamente, depois rapidamente, antes de diminuir à medida que atinge sua capacidade de carga. A função logística usa uma equação diferencial que trata o tempo como contínuo. Em vez disso, o mapa logístico usa uma equação

diferencial não linear para observar etapas de tempo discretas. A série temporal foi gerada com 500000 dados e agregada em 100 classes como 5000 dados cada. A Figura (24) mostra uma parte dessa série temporal para a configuração C_2 referente aos primeiros 1000 dados onde pode-se observar 0 comportamento linear.

Figura 24 – Parte da série temporal clássica simulada para a configuração C_2



Fonte: Elaborada pelo autor (2021)

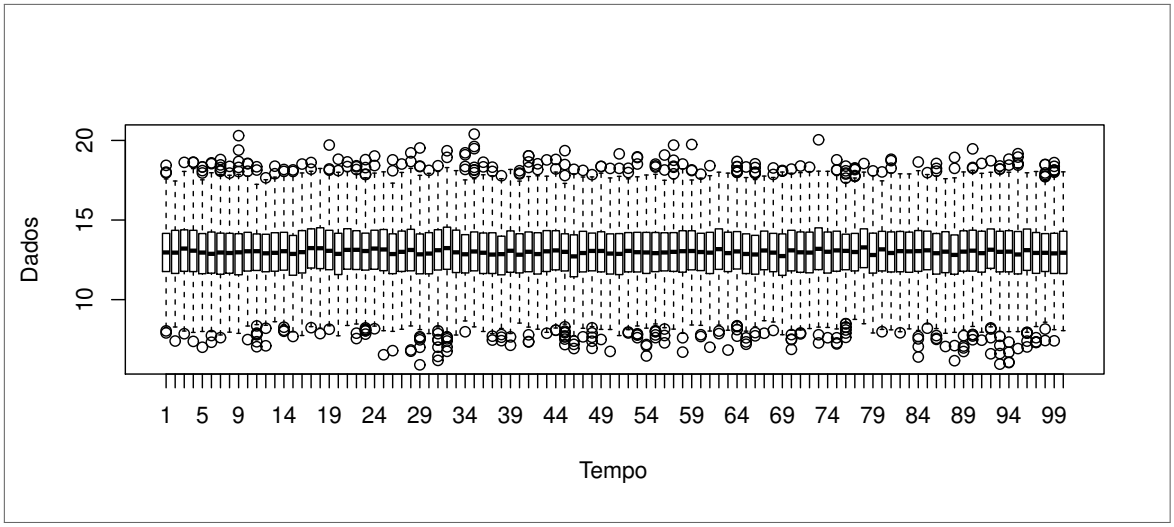
Nas Figuras (25) e (26) são apresentadas a série temporal simbólica de $boxplots$ obtida após a agregação e as 3 curvas dos quartis que descrevem esses $boxplots$. Note que a série de $boxplots$, à diferença da configuração C_1, tem comportamento linear e as curvas obtidas mantêm o comportamento da série temporal originalmente gerada.

Além disso, simulamos uma série temporal não linear: o mapa Hénon (veja Figura (27)). Essa série apresenta comportamento complexo e caótico em duas dimensões e foi uma proposta do astrônomo francês Michel Hénon em 1976 (HÉNON, 1976). Tipicamente, é estudado em áreas de conhecimento como engenharias, matemáticas e física.

O mapa Hénon também pode ser desconstruído em um mapa unidimensional, definido de forma semelhante à sequência de $Fibonacci$ pela equação mostrada na Tabela (11) para a configuração C_3. Geramos um mapa de Hénon de uma só dimensão constituído por 100000 pontos sem adicionar ruído, agregado em 400 classes como 250 dados cada.

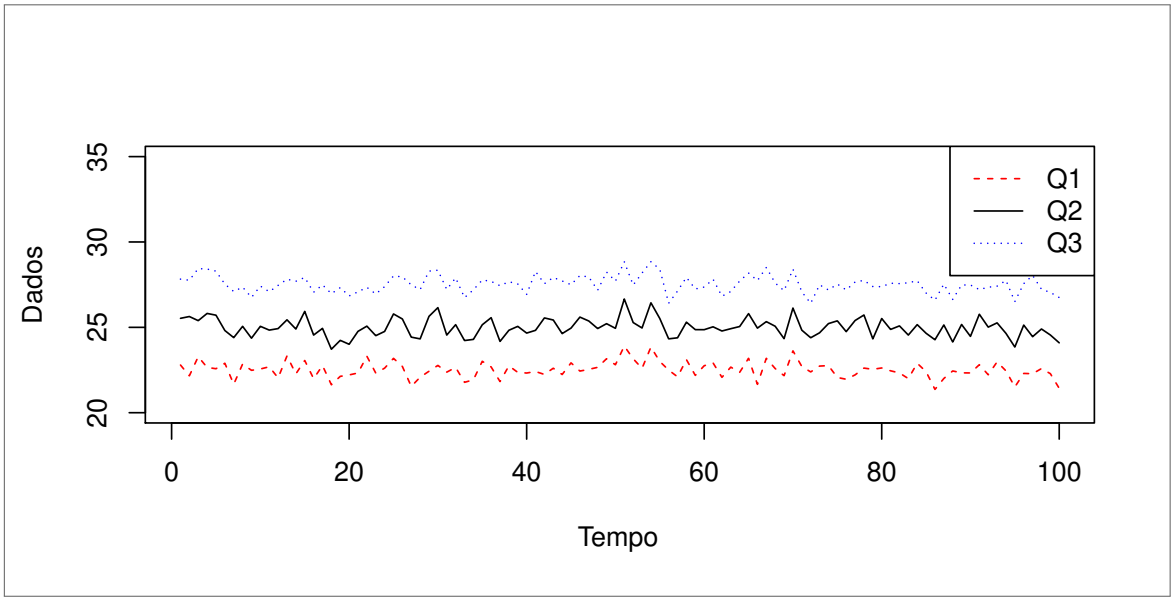
A Figura (28) mostra os primeiros 1000 dados do mapa de Hénon simulado e nas Figuras (29) e (30) parte da série de $boxplots$ e as 3 curvas para a representação I que descrevem esses $boxplots$. Nessas figuras pode-se observar que a série gerada não tem $outliers$ e que as

Figura 25 – Parte da série temporal de $boxplots$ para a configuração C_2



Fonte: Elaborada pelo autor (2021)

Figura 26 – Curvas da parte apresentada da série temporal simbólica Q_t para configuração C_2

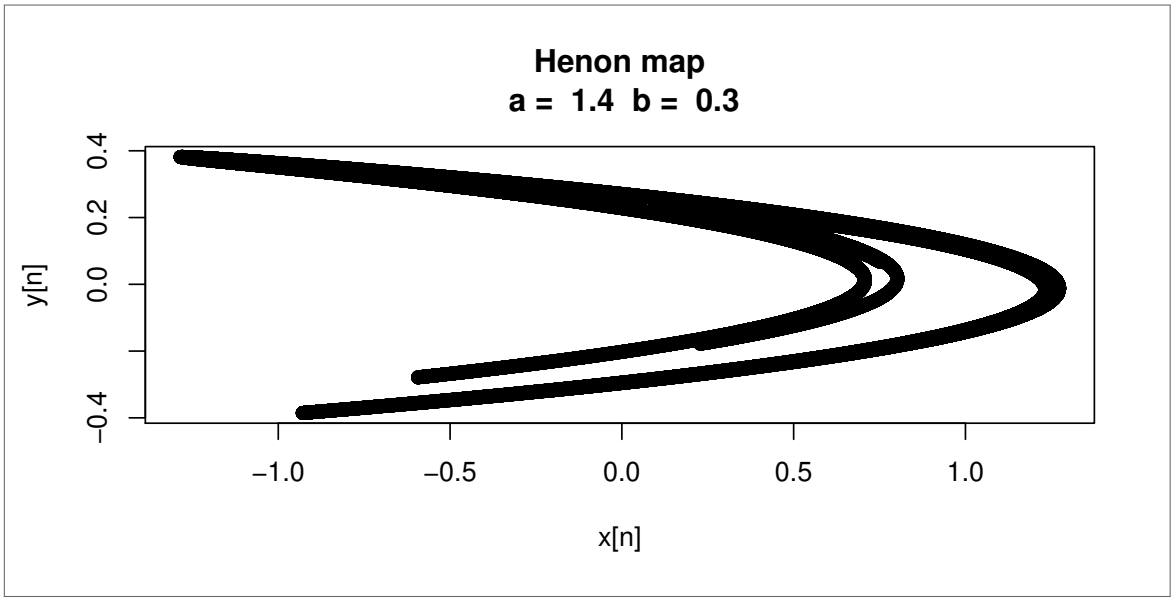


Fonte: Elaborada pelo autor (2021)

instâncias que conformam as classes não têm distribuição simétrica.

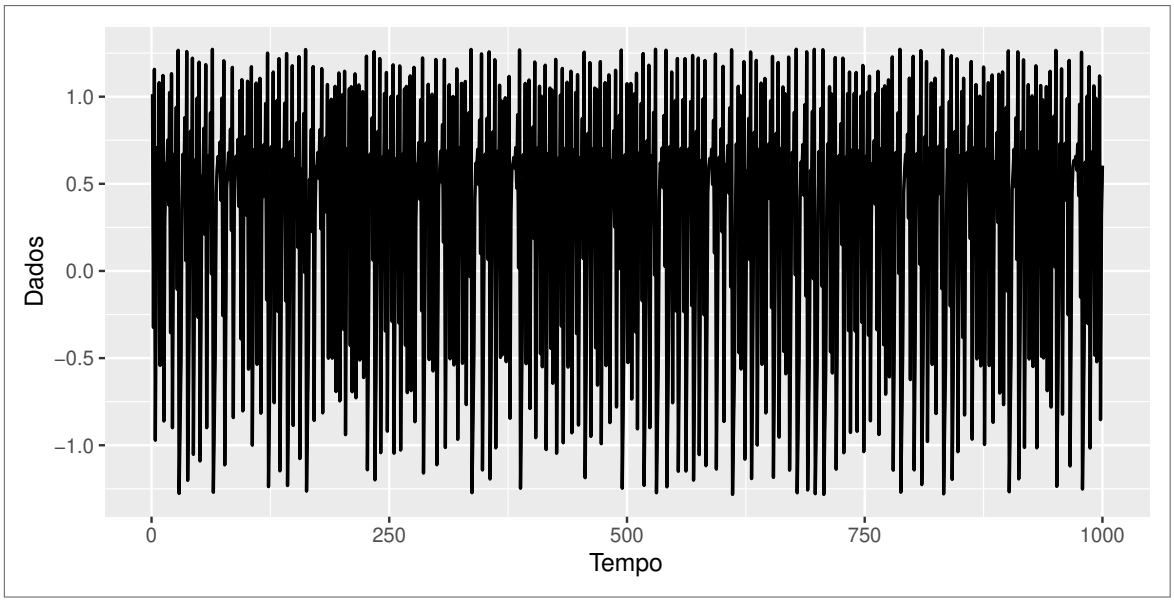
Os testes descritos nesta seção foram realizados utilizando as séries temporais simuladas descritas anteriormente e aplicando-se o método de seleção de protótipos descrito na Subseção 4.2.3. O algoritmo 2, mostra como as séries temporais de $boxplots$ são geradas nesta simulação de Monte Carlo.

Figura 27 – Mapa de Hénon



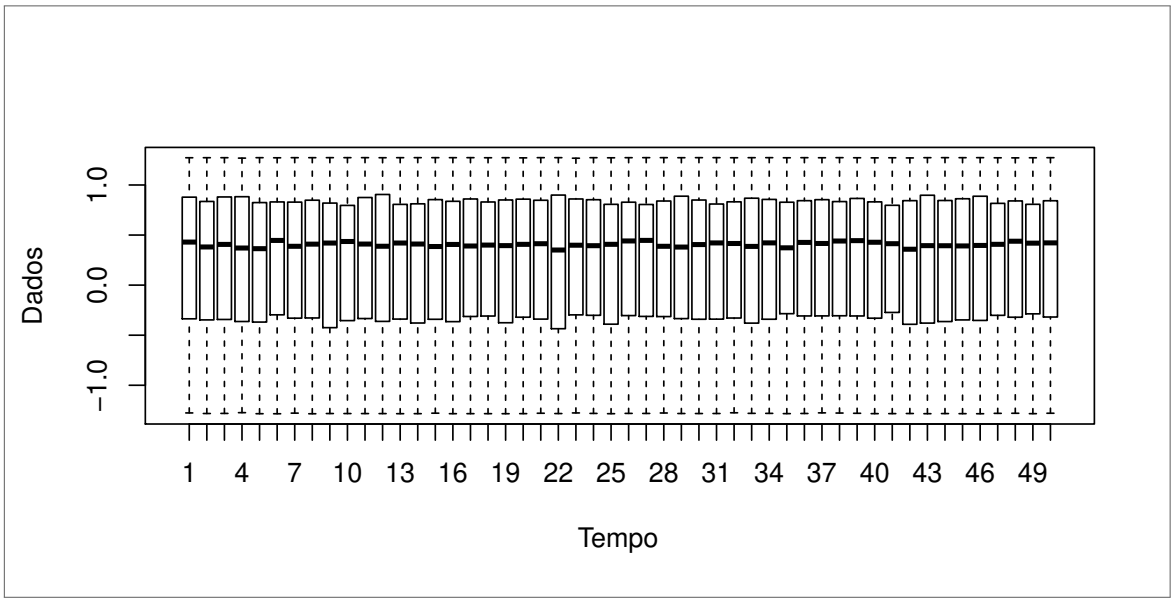
Fonte: (HÉNON, 1976, p. 73)

Figura 28 – Parte da série temporal clássica simulada para a configuração C_3



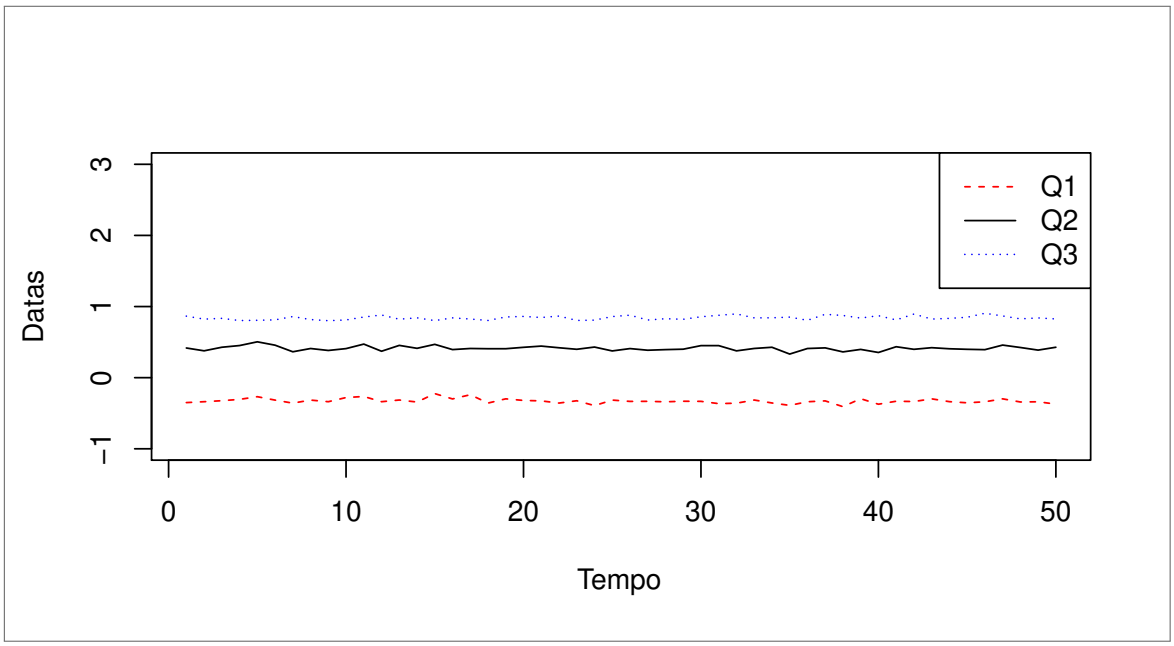
Fonte: Elaborada pelo autor (2021)

Figura 29 – Parte da série temporal de $boxplots$ para a configuração C_3



Fonte: Elaborada pelo autor (2021)

Figura 30 – Curvas da parte apresentada da série temporal simbólica Q_t para configuração C_3



Fonte: Elaborada pelo autor (2021)

begin

Inicializar *MonteCarlo* $= 1000$, N $= 100000$;

for $i = 1$ *: MonteCarlo* **do**

Se a configuração 1 (equação i) é usada, **gerar** uma série temporal de tamanho N de acordo com uma distribuição normal com média 0, desvio padrão 1 e erro ε_t ;

Se a configuração 2 (equação ii) é usada, **gerar** uma série temporal de tamanho N de acordo com uma distribuição normal com média 0, desvio padrão 1 e erro ε_t;

Se a configuração 3 (equação iii) é usada, **gerar** uma série temporal de tamanho N de acordo com uma distribuição normal com média 0, desvio padrão 1 e erro ε_t;

Particionar a série temporal gerada em n classes de tamanho v;

for $c = 1 : n$ **do**

Selecione os protótipos para c-ésima classe de acordo com o algoritmo 1;

end

Obter a série simbólica multi-valorada Q_t, $(t = 1, \ldots, n)$;

Aplicar o modelo VAR para a série multi-valorada Q_t de acordo com as Subseções 4.3.1 e **obter** $\hat{Q}_t = \{\hat{q_1}t, \hat{q_2}t, \hat{q_3}t\}$;

Construir os $boxplots$ estimados de acordo com a regra descrita na Subseção 4.3.1;

Computar o $MMRE$ de acordo com a equação apresentada anteriormente;

end

Computar a média e o desvio padrão de todos os valores $MMRE$;

end

Algoritmo 2: Simulação Monte Carlo.

4.4.1.1 *Análises de Desempenho da Seleção de Protótipos*

Esta subseção avalia a relevância do algoritmo de seleção de protótipos para filtrar as classes, reduzindo o tamanho sem afetar a qualidade da previsão. Todas as séries temporais simbólicas foram particionadas em conjuntos de treinamento com $n-20$ $boxplots$ e teste com 20 $boxplots$. Em relação aos conjuntos de teste, são considerados dois horizontes de previsão: um de 5 passos-à-frente e um de 20 passos-à-frente. Assim, quando nos referimos ao horizonte de previsão de h passos-à-frente, nós realizamos uma sequência de h previsões: de 1 passo-à-frente até h passos-à-frente. Por exemplo, quando consideramos o horizonte de previsão de 5 passos-à-frente, nós prevemos os próximos 5 $boxplots$ futuros. Da mesma maneira, quando consideramos o horizonte de previsão de 20 passos-à-frente, nós prevemos os próximos 20 $boxplots$ futuros.

Em cada réplica de Monte Carlo, computamos o MMRE usando a seleção de protótipos e sem a seleção de protótipos dos quartis para descrever os $boxplots$ e, em seguida, calculamos o valor médio e o desvio padrão das 1000 réplicas. De aqui em diante, chamaremos a abordagem que usa a seleção de protótipos como "CP" e a abordagem que não faz uso da seleção de protótipos como "SP" para facilitar a escrita e a compreensão dos resultados.

Neste contexto foi adotado o teste t-Student bilateral para amostras pareadas a um nível de significância de 5%. Para comparar esses resultados formulamos as seguintes hipótese:
H_0: μMMRE CP $=$ μMMRE SP
H_1: μMMRE CP $\neq$ μMMRE SP

As Tabelas (12) mostra os valores da média e desvio padrão (entre parêntesis) dos MMRE calculados para as configurações C_1, C_2 e C_3 respectivamente, e o p-valor do teste t-Student para amostras pareadas a um nível de significância de 5%.

A partir dos valores da Tabela 12 pode-se observar que os valores p para as configurações C_2 e C_3 foram superior a 0,05, o que sugere a não rejeição da hipótese nula a um nível de significância de 5% (ou seja, a diferença entre as médias das medidas de erro relativas aos modelos CP e SP nessas configurações não é estatisticamente significativas). Esses resultados indicam que os respectivos modelos apresentam resultados similares em termos de precisão na previsão das séries temporais simbólicas de $boxplots$ consideradas. No caso das configurações C_1 os valores p foram inferiores a 0,05, o que sugere a rejeição da hipótese nula a um nível de significância de 5%, ou seja, a média das medidas de erro relativas ao modelo que não usa a seleção de protótipos (SP) foi maior estatisticamente que a média das medidas de erro

Tabela 12 – Média e desvio padrão (entre parêntesis) do MMRE calculado a partir de 1000 repetições de Monte Carlo e o p-valor do teste t-Student para amostras pareadas a um nível de significância de 5%.

Horiz. Previsão	**Configurações**	**MMRE**		p-**valor**
		CP	SP	
Treinamento	C_1	0.4490 (2.5723)	0.9400 (1.1497)	< 0.0001
	C_2	0.2078 (1.0818)	0.2110 (1.0002)	0.1650
	C_3	0.0190 (0.0723)	0.0202 (0.0252)	0.4501
5-passos à frente	C_1	1.2580 (2.0818)	1.6855 (1.0521)	< 0.0001
	C_2	0.5175 (1.9124)	0.5200 (1.0012)	0.1091
	C_3	0.0580 (0.0818)	0.0206 (0.0188)	0.1517
20-passos à frente	C_1	0.6911 (5.7678)	1.0771 (4.0766)	< 0.0001
	C_2	0.2951 (2.7678)	0.2754 (2.0008)	0.2149
	C_3	0.0618 (0.7678)	0.0599 (0.1095)	0.1496

Fonte: Elaborada pelo autor (2021)

relativas ao modelo que usa a seleção de protótipos (CP).

No quesito redução, após a aplicação da seleção de protótipos, as médias de porcentagem de redução de tamanho do conjunto para C_1, C_2 e C_3 são, respectivamente: 45.7%, 23.5% e 26,3%. Note que a configuração C_1 teve o maior percentual de redução. Essa ocorrência se deve ao grande número de anomalias e instâncias ruidosas presentes nesta série temporal (veja a Figura 22). Esses resultados eram esperados, pois na configuração C_1 há um grande número de instâncias ruidosas e a sua compactação foi bem maior quando comparada com as configurações C_2 e C_3. Isto corresponde a dizer que o método de seleção de protótipos apresentado permite reduzir o conjunto de dados a uma amostra representativa do problema em questão, o que minimiza a presença de anomalias, instâncias ruidosas e erros de previsão.

4.4.1.2 Análise de Desempenho dos Modelos VAR e ARIMA para Previsão de Séries Temporais de Boxplots

Nesta subseção, comparamos o desempenho dos modelos VAR e ARIMA na previsão de séries temporais de $boxplots$. Os modelos de ARIMA (BOX; JENKINS, 1970) são os modelos de séries temporais mais importantes e amplamente utilizados, devido à sua relativa simplicidade na compreensão e implementação. ARIMA é um modelo linear criado a partir de uma combinação finita e linear de valores passados da série e de uma combinação linear finita de erros passados. Esses modelos são bastante flexíveis, pois podem representar diferentes tipos

de séries temporais como, *Auto-regressive Model* (AR), *Moving Average Model* (MA) e séries combinadas de AR e MA ARMA. Um processo ARIMA(p, d, q) modela as diferenças estacionárias da série temporal Y_s usando o processo ARMA (p, q). Os valores (p, d, q) são números inteiros, não negativos que indicam rapidamente o modelo ARIMA específico que é usado e são definidos da seguinte forma:

- p: o número de defasagens ($lags$) do modelo auto-regressivo;
- d: o grau de diferenciação (o número de vezes que as observações são diferenciadas);
- q: ordem do modelo de média móvel.

Três modelos independentes de séries temporais (para $Q1_t, Q2_t$ e $Q3_t$) são ajustados com base no ARIMA e a média MMRE é computada. A seleção do melhor modelo ARIMA para ajuste da série temporal foi realizada através da minimização do *Akaike Information Criterion* (AIC) (AKAIKE, 1974). O teste t-Student para amostras pareadas ao nível de significância de 5%, foi utilizado para comparar os resultados entre as abordagens. As hipóteses são:

H_0: μMMRE VAR $\geq$ μMMRE ARIMA.

H_1: μMMRE VAR $<$ μMMRE ARIMA.

A Tabela 13 mostra os valores p do teste Student-t. Podemos observar que como esperado, os valores de p são $<$ 0,05. Ao nível de significância de 5% a hipótese nula é rejeitada para todos os casos. Isso significa que o modelo VAR supera o modelo ARIMA para descrever dados de $boxplots$.

4.4.1.3 Comparando com as Abordagens da Literatura de SDA

De acordo com Drago (2015), dada uma série temporal, cada classe desta série é descrita por um resumo de cinco valores: mínimo (m), quartil inferior ($Q1$), mediana ($Q2$), quartil ($Q3$) e máximo (M) e cinco curvas são ajustadas. Além disso, as estimativas são obtidas usando o modelo clássico ARIMA para cada curva de forma independente. Para avaliar nosso trabalho desde uma perspectiva mais ampla, comparamos com a abordagem proposta em Drago (2015), esta subseção apresenta duas tabelas com os resultados do MMRE obtidos:

1. Os primeiros resultados (Tabela 14) foram obtidos considerando o uso da seleção de protótipos e do modelo VAR para ajustar cinco curvas. A abordagem proposta neste

Tabela 13 – Média e desvio padrão do MMRE para 1000 réplicas de Monte Carlo e o p-valor do teste t-Student para amostras pareadas a um nível de significância de 5% para o modelo VAR e ARIMA.

Horiz. Previsão	**Configurações**	**MMRE**		**p-valor**
		VAR	ARIMA	
Treinamento	C_1	0.4490 (2.5723)	1.0991 (2.1497)	< 0.0001
	C_2	0.2078 (1.0818)	0.5599 (1.0026)	< 0.0001
	C_3	0.0190 (0.0723)	0.0598 (0.0031)	< 0.0001
5-passos à frente	C_1	1.2580 (2.0818)	1.9083 (1.6523)	< 0.0001
	C_2	0.5175 (1.9124)	0.9052 (1.0012)	< 0.0001
	C_3	0.0580 (0.0818)	0.0804 (0.0185)	< 0.0001
20-passos à frente	C_1	0.6911 (5.7678)	1.7571 (3.0766)	< 0.0001
	C_2	0.2951 (2.7678)	0.5400 (1.0008)	< 0.0001
	C_3	0.0618 (0.7678)	0.0999 (0.0095)	< 0.0001

Fonte: Elaborada pelo autor (2021)

trabalho considera apenas três curvas. A ideia é mostrar a importância do uso de três curvas como proposto na nossa abordagem;

2. Os segundos resultados (Tabela 15) foram obtidos assumindo o modelo proposto em Drago (2015) que não utiliza seleção de protótipos e ajusta cinco modelos ARIMA independentes. Vale salientar que é o mesmo modelo originalmente apresentado por Drago (2015) sem fazer nenhuma modificação.

A Tabela 14 mostra a média e o desvio padrão do MMRE para 1000 réplicas de Monte Carlo para o conjunto de treinamento e horizontes de previsão de 5 e 20 passos à frente em relação às configurações C_1, C_2 e C_3. Além disso, esta tabela mostra os valores p do teste t-Student para amostras pareadas a um nível de significância de 5%. Como podemos observar, os valores p são $<$ 0,05 para todos horizontes de previsão. Esses resultados indicam que o uso de três curvas para representar as séries temporais de $boxplots$ é mais adequado que o uso de cinco curvas.

A Tabela 15 apresenta a média e o desvio padrão do MMRE calculado a partir de 1000 réplicas de Monte Carlo para o conjunto de treinamento e horizontes de previsão de 5 e 20 passos à frente em relação às configurações C_1, C_2 e C_3 para abordagem introduzida em (DRAGO, 2015). Pode-se observar claramente que nossa abordagem usando o modelo VAR com três ou com cinco curvas supera a abordagem proposta por Drago (2015).

Tabela 14 – Média e desvio padrão do MMRE para o uso de cinco curvas e os valores p do teste de t-Student na comparação com o modelo proposto neste artigo que usa três curvas.

Configurações	Horiz. Previsão	MMRE	p-valor
C_1	Treinamento	1.0194 (0.0012)	< 0.0001
	5-passos à frente	2.2574 (2.8066)	< 0.0001
	20-passos à frente	2.3443 (5.0113)	< 0.0001
C_2	Treinamento	0.8194 (2.2476)	< 0.0001
	5-passos à frente	1.0449 (2.7890)	< 0.0001
	20-passos à frente	1.3035 (3.6032)	< 0.0001
C_3	Treinamento	0.2284(1.0013)	< 0.0001
	5-passos à frente	0.2557 (2.0076)	< 0.0001
	20-passos à frente	0.3930 (2.6032)	< 0.0001

Fonte: Elaborada pelo autor (2021)

Tabela 15 – Média e desvio padrão do MMRE para a abordagem usando ARIMA, como apresentado em Drago (2015) e os valores p do teste de t-Student na comparação com o modelo proposto neste artigo que usa o modelo VAR.

Configurações	Horiz. Previsão	MMRE	p-valor
C_1	Treinamento	1.9503 (20.7820)	< 0.0001
	5-passos à frente	2.4847 (10.9458)	< 0.0001
	20-passos à frente	2.9877 (26.2648)	< 0.0001
C_2	Treinamento	0.9409 (1.6456)	< 0.0001
	5-passos à frente	1.5133 (3.7600)	< 0.0001
	20-passos à frente	4.3096 (18.5249)	< 0.0001
C_3	Treinamento	0.3657(0.2089)	< 0.0001
	5-passos à frente	0.7160(0.8572)	< 0.0001
	20-passos à frente	1.0572 (2.4352)	< 0.0001

Fonte: Elaborada pelo autor (2021)

Para finalizar este estudo com séries temporais sintéticas, algumas observações podem ser extraídas:

- O algoritmo de seleção de protótipos utilizado neste artigo permitiu reduzir as séries temporais sem degradar a precisão dos resultados. Ele foi capaz de detectar e remover alguns ruídos e/ou instâncias redundantes presentes na série temporal sintética.
- O poder preditivo da abordagem proposta foi superior ao introduzido por Drago (2015) em termos de MMRE. De fato, o uso de três curvas é a melhor opção.

4.4.1.4 Comparando com as Abordagens da Literatura Clássica

Nesta subseção, o desempenho dos modelos ARIMA e *Support Vector Regression* (SVR) para séries temporais clássicas são comparados com a abordagem proposta neste artigo para séries temporais simbólicas de $boxplots$. A medida de desempenho é o MMRE e o objetivo deste estudo é mostrar que o uso de médias não é mais adequado para a previsão de séries temporais que necessitam de agregação porque o interesse está nas classes e não nos indivíduos.

Nesse contexto, as classes são descritas por valores únicos (médias). Deve-se especificar também que as classes não foram filtradas, ou seja, não foi utilizado o algoritmo de seleção de protótipos. A seleção do melhor modelo ARIMA para ajuste da série de médias foi realizada por meio da minimização do AIC.

A média e o desvio padrão do MMRE calculado a partir de 1000 réplicas de Monte Carlo para o conjunto de treinamento e horizontes de previsão de 5 e 20 passos à frente para modelos clássicos ARIMA e SVR são mostrados nas Tabelas 16 e 17. Além disso, são mostrados em cada tabela os valores p do teste t-Student quando comparados os resultados obtidos com nossa abordagem que considera seleção de protótipos, dados quartis e modelo VAR.

Esses resultados reforçam o uso de dados de quartis simbólicos como ferramenta para resumir as classes das séries temporais. Assim, dependendo do conjunto de dados e das necessidades da agregação, o uso de nossa abordagem é adequado e seu poder preditivo é superior ao do ARIMA e SVR clássicos.

Tabela 16 – Média e desvio padrão do MMRE calculados a partir de 1000 réplicas de Monte Carlo (para o conjunto de treinamento, previsão de 5 e 20 passos à frente) para o modelo ARIMA clássico e os valores p de do teste t-Student na comparação com os resultados da abordagem proposta neste trabalho.

Configurações	**Horiz. Previsão**	**MMRE**	p**-valor**
C_1	Treinamento	1.9076 (2.0093)	< 0.0001
	5-passos à frente	1.9314 (5.0716)	< 0.0001
	20-passos à frente	2.6333 (4.4276)	< 0.0001
C_2	Treinamento	1.0331 (0.0014)	< 0.0001
	5-passos à frente	1.0681 (0.0098)	< 0.0001
	20-passos à frente	1.0444 (0.0048)	< 0.0001
C_3	Treinamento	0.1625(1.0103)	< 0.0001
	5-passos à frente	0.1404 (1.0706)	< 0.0001
	20-passos à frente	0.1712 (1.0362)	< 0.0001

Fonte: Elaborada pelo autor (2021)

Tabela 17 – Média e desvio padrão do MMRE calculados a partir de 1000 réplicas de Monte Carlo (para o conjunto de treinamento, previsão de 5 e 20 passos à frente) para o modelo SVR clássico e os valores p de do teste t-Student na comparação com os resultados da abordagem proposta neste trabalho.

Configurações	**Horiz. Previsão**	**MMRE**	p-**valor**
C_1	Treinamento	1.5905 (4.8730)	< 0.0001
	5-passos à frente	1.8135 (8.9009)	< 0.0001
	20-passos à frente	2.7517 (1.4585)	< 0.0001
C_2	Treinamento	1.9681 (2.3001)	< 0.0001
	5-passos à frente	2.6980 (1.0798)	< 0.0001
	20-passos à frente	1.1687 (2.3569)	< 0.0001
C_3	Treinamento	1.1102 (0.0125)	< 0.0001
	5-passos à frente	1.3940 (0.0010)	< 0.0001
	20-passos à frente	1.0796 (0.0041)	< 0.0001

Fonte: Elaborada pelo autor (2021)

4.4.2 Resultados Empíricos com Séries Temporais Reais

Esta subseção demostrará a utilidade da abordagem apresentada nesta tese em aplicações a dados de séries temporais simbólicas de $boxplots$ reais. Nós consideramos 3 séries temporais reais que agregamos em séries temporais simbólicas, aplicamos a seleção de protótipos em cada classe e usamos o modelo descrito anteriormente para a previsão das mesmas. A primeira série temporal analisada foi descrita na Subseção 4.2 (os registros da concentração de CO no ar por hora em uma cidade italiana). Foi usada a agregação por semanas para obter a série temporal simbólica de $boxplots$ (veja a Figura 16).

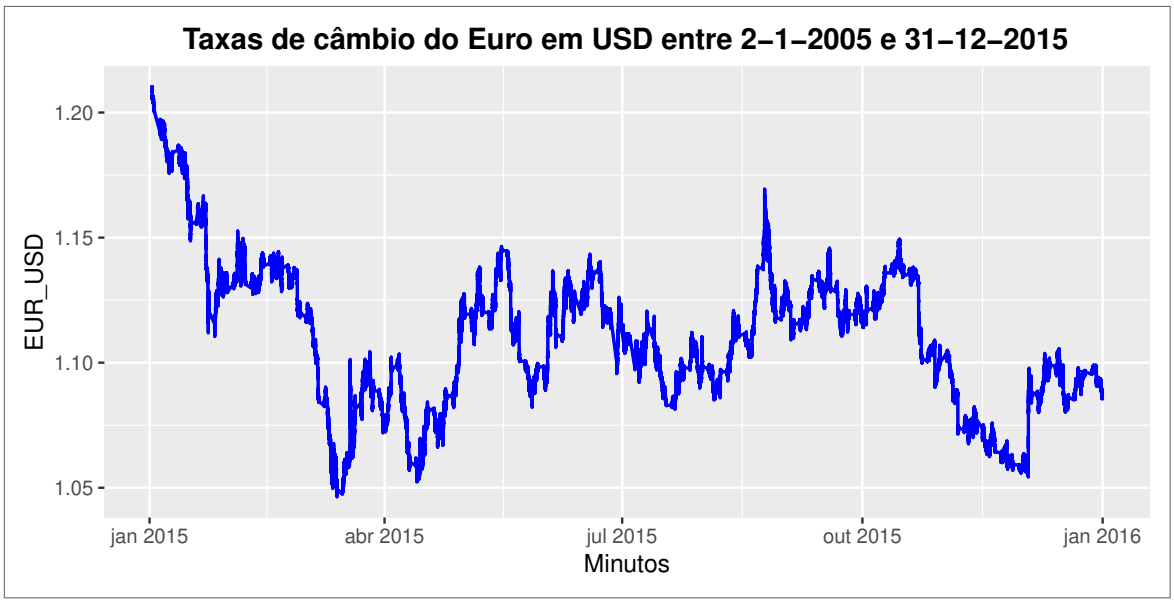
A segunda série utilizada é apresentada na Figura (31) e corresponde à série temporal das taxas de câmbio do euro em USD extraída do site $Yahoo-Finances$. Esse conjunto de dados contém 824706 dados correspondentes às taxas de câmbio do Euro em USD registradas com uma frequência de 5 minutos, no período de 2-1-2005 a 31-12-2015.

Para analisarmos esta série econômica foi preciso calcular os retornos usando a equação (4.7). A razão de utilizarmos série de retornos é que possui propriedades estatísticas mais interessantes do que a série dos preços.

$$R_t = \left(\frac{P_t - P_{t-i}}{P_{t-i}} \right) \times 100 \tag{4.7}$$

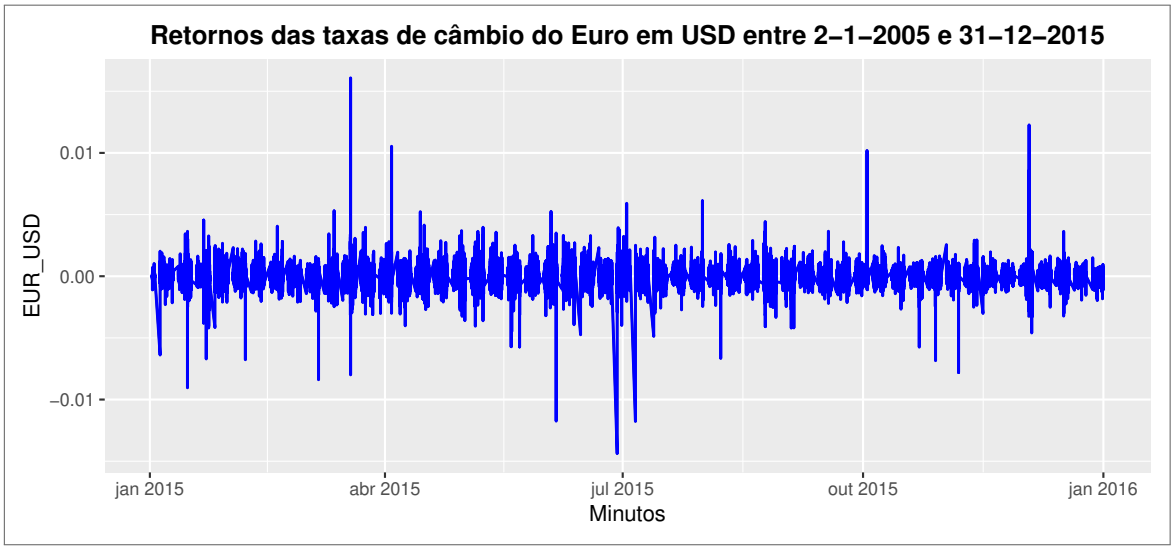
Essa série temporal de retornos refere a agregação temporal explicitada na Subseção (2.1.1). Assim, para obter as variáveis simbólicas, a série foi agrupada por dias, sendo 3443 dias com

Figura 31 – Série temporal das taxas de câmbio do euro em USD.



Fonte: Elaborada pelo autor (2021)

Figura 32 – Série temporal dos retornos das taxas de câmbio do euro em USD.

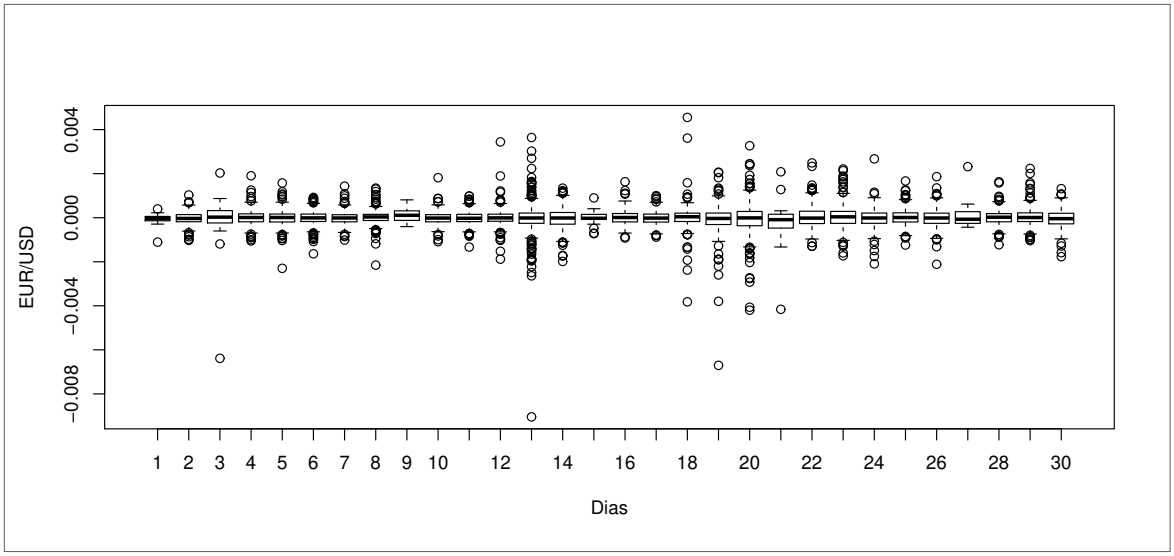


Fonte: Elaborada pelo autor (2021)

288 valores de preço cada. A Figura (33) mostra uma parte dos $boxplots$ de retornos das taxas de câmbio do Euro em USD referente aos primeiros 30 dias, pode-se observar que a série tem vários $outliers$. A Figura (34) mostra as 3 curvas que descrevem esses $boxplots$.

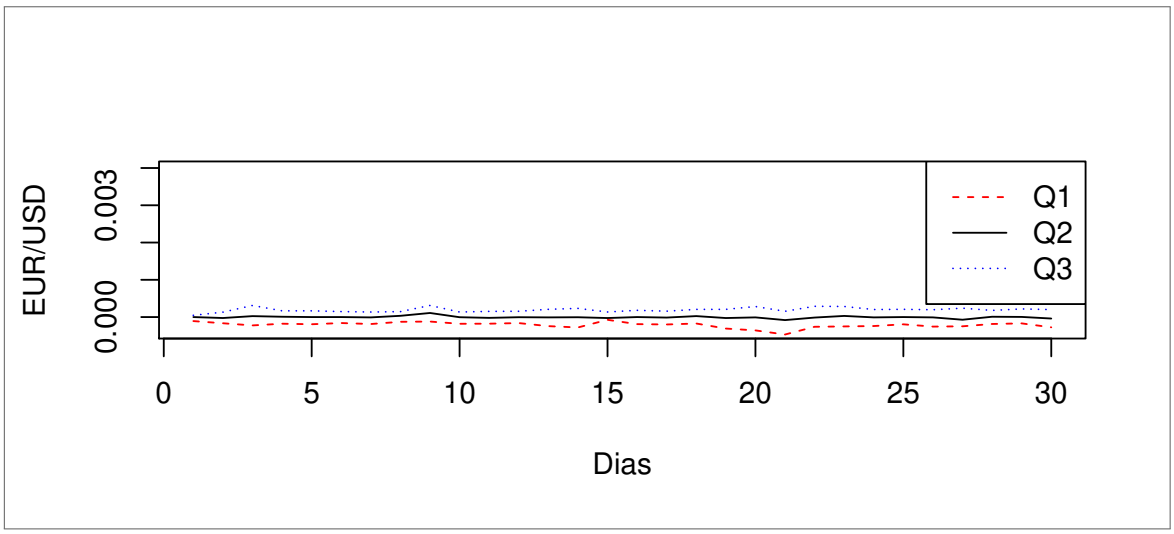
A terceira série temporal foi extraída do Banco de Dados Climáticos Instrumentais de Longo Prazo da República Popular da China. Entre outras variáveis, este banco de dados contém as precipitações mensais na China registradas por 60 estações meteorológicas entre 1951 e 1988. Assim, são 60 séries de 437 valores cada, a Figura (35) mostra 8 de essas séries

Figura 33 – Parte da série temporal simbólica de $boxplots$ dos retornos das taxas de câmbio do EUR em USD.



Fonte: Elaborada pelo autor (2021)

Figura 34 – Curvas da parte apresentada da série temporal simbólica de $boxplots$ dos retornos das taxas de câmbio do EUR em USD, representação II.

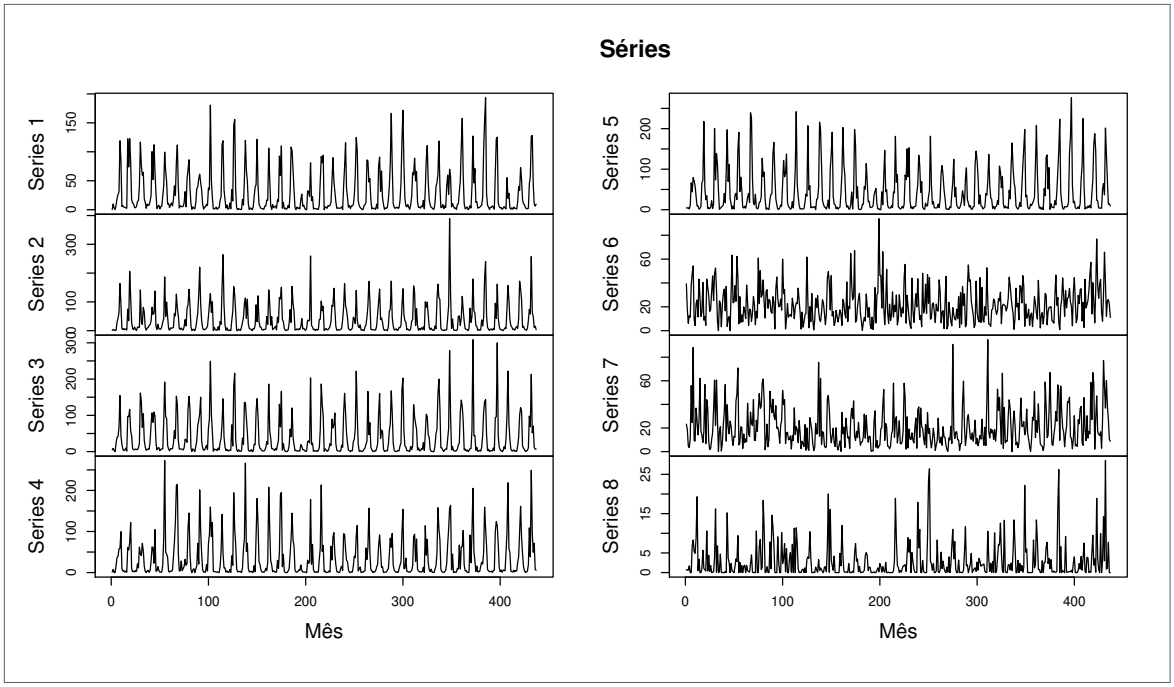


Fonte: Elaborada pelo autor (2021)

temporais.

Esse conjunto de séries temporais das precipitações da China se refere à agregação contemporânea explicitada na Subseção (2.1.1) em que a variável é medida ao longo do tempo nos elementos de um conjunto, mas o interesse não está em saber a evolução de cada um dos elementos e sim do conjunto como um todo. Então as séries foram agregadas por mês, obtendo-se 437 classes que contêm os registros das 60 estações do mês em questão. A Figura (36) mostra uma parte da série temporal simbólica dos $boxplots$ obtida para as precipitações observadas nas 60 estações da República Popular da China de janeiro de 1951 a dezembro de

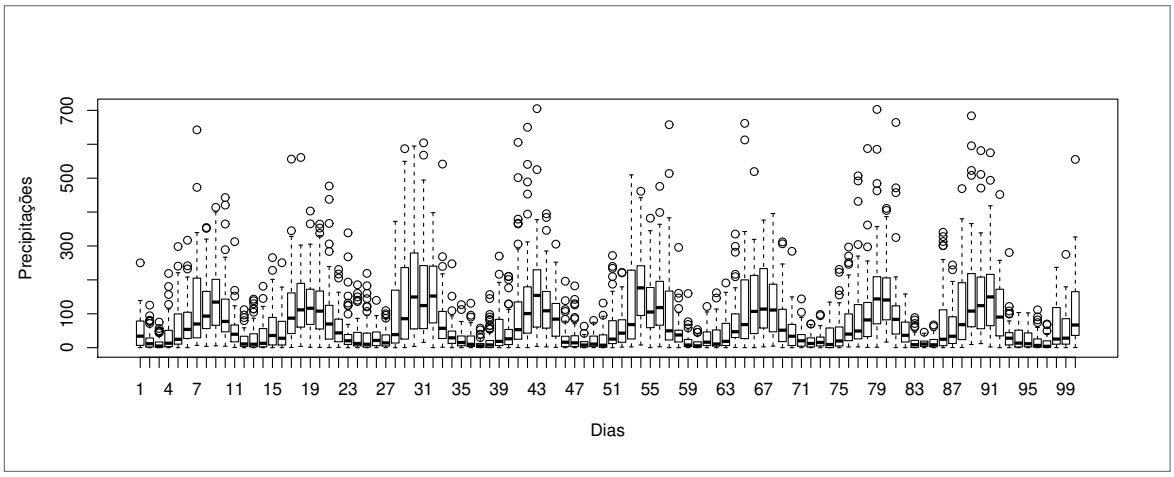
Figura 35 – Séries temporais das precipitações observadas em 8 estações da República Popular da China.



Fonte: Elaborada pelo autor (2021)

1988 e a Figura (37) as 3 curvas que descrevem esses $boxplots$.

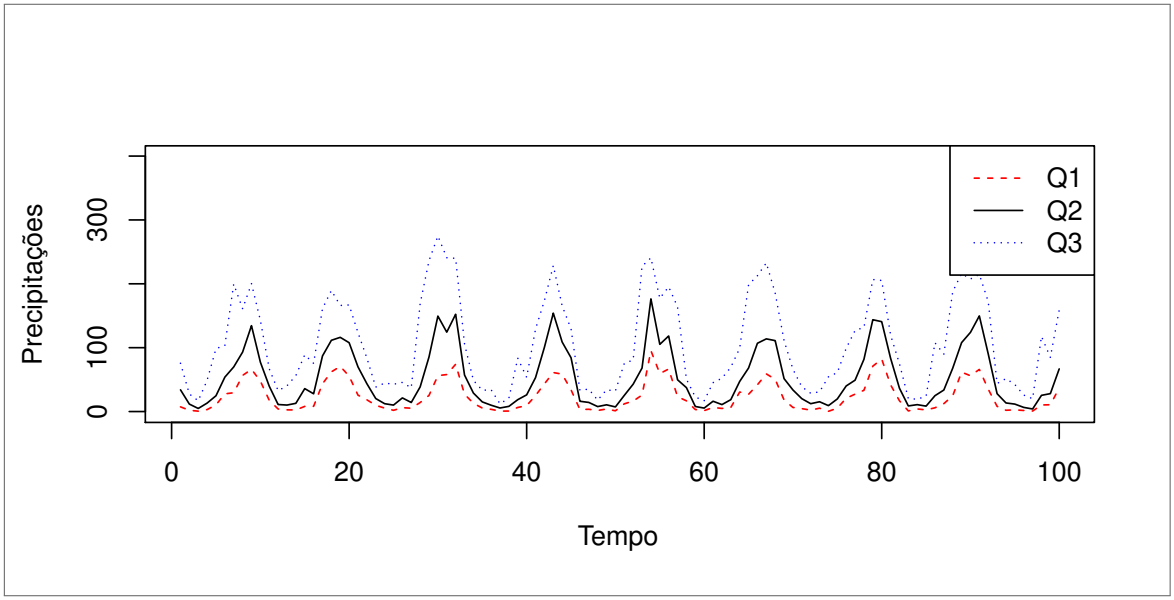
Figura 36 – Parte da série temporal simbólica de $boxplots$ das precipitações na República Popular da China.



Fonte: Elaborada pelo autor (2021)

Um dos procedimentos padrão mais amplamente utilizados para avaliação de modelos em classificação e regressão é a validação cruzada K-fold. No entanto, quando se trata de previsão de séries temporais, devido à inerente correlação serial e potencial não estacionaridade dos dados, sua aplicação não é adequada (BERGMEIR; BENíTEZ, 2012). Racine (2000) propõe validação cruzada "hv-block" que é assintoticamente ótima. É consistente para observações temporalmente dependentes no sentido de que a probabilidade de selecionar o modelo com a

Figura 37 – Curvas da parte apresentada da série temporal simbólica de $boxplots$ das precipitações na República Popular da China, representação I.



Fonte: Elaborada pelo autor (2021)

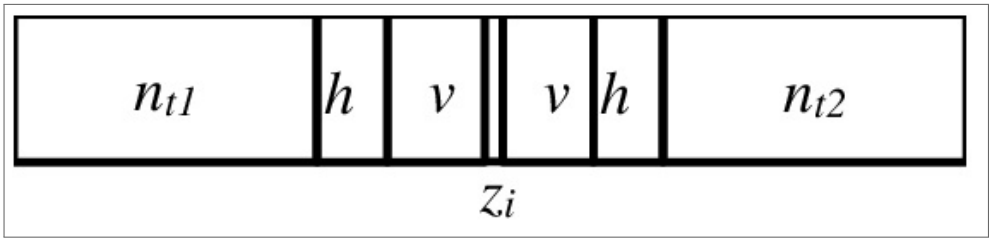
melhor capacidade preditiva converge para 1 a medida que o número total de observações se aproxima do infinito.

A ideia básica é colocar restrições nas relações entre o conjunto de treinamento, o conjunto de validação, o tamanho de um bloco h e o tamanho da amostra. Podemos assim obter um procedimento consistente de validação cruzada.

Dada uma observação Zi (veja a Figura 38), inicialmente removemos v observações em cada lado dela para obter um conjunto de validação de tamanho $2v+1$. Em seguida, removemos h observações em ambos os lados desse conjunto de validação com as observações restantes $n-2v-2h-1$ formando o conjunto de treinamento. O valor de v controla o tamanho do conjunto de validação com $n_v = 2v+1$. O valor de h controla a dependência do conjunto de treinamento de tamanho $n_t = n - 2h - n_v$ e o conjunto de validação de tamanho n_v. Consulte (RACINE, 2000) para uma discussão sobre treinamento, validação e seleção do tamanho da amostra de teste para previsão de séries temporais usando validação cruzada "hv-block".

A Figura 39 a continuação ilustra um exemplo da validação cruzada em um $boxplot$ de teste que representa os valores para um dia. Suponha que o $boxplot$ se refira aos valores da sexta-feira, por exemplo, definido como B_0, e a unidade de defasagem é um dia. Portanto, os dados de treinamento consistem em seis $boxplots$ (quinta, quarta, terça, segunda e domingo na mesma semana dos dados de teste e sábado da semana anterior) e três $boxplots$ para validação (três sextas-feiras das últimas três semanas, B_1, B_2, B_3). Com base na abordagem de validação

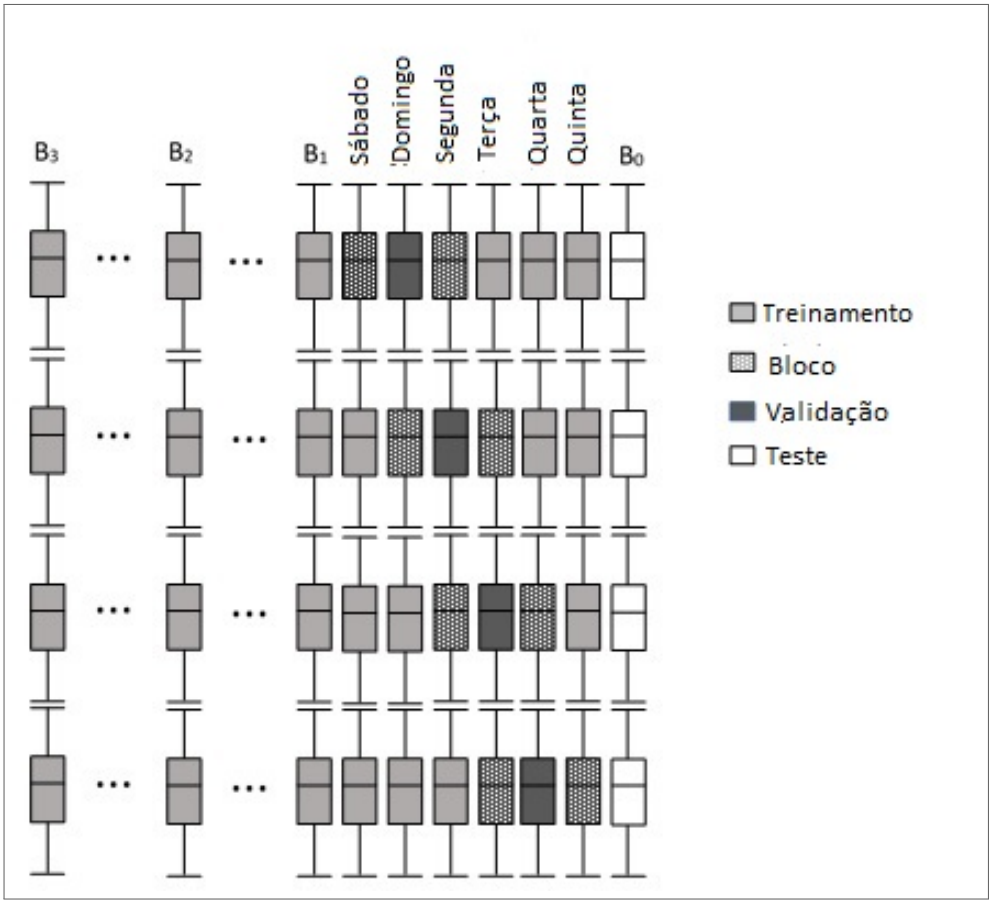
Figura 38 – Ilustração da validação cruzada "hv-block"



Fonte: (RACINE, 2000, p. 54)

cruzada "hv-block", os dados de treinamento são divididos em 4 $folds$ de treinamento e validação. Em cada $fold$, o tamanho dos dados de validação nv e o bloco h são definidos como um $boxplot$ e o restante dos dados é mantido como dados de treinamento.

Figura 39 – Exemplo de validação cruzada "hv-block".



Fonte: Elaborada pelo autor (2021)

A Tabela 18 apresenta a média e o desvio padrão do MMRE para o modelo proposto nesta tese e o modelo de Drago (2015). Além disso, esta tabela mostra os valores p do teste t-Student para amostras pareadas a um nível de significância de 5% da comparação entre a nossa abordagem com a abordagem de Drago (2015). A série temporal das taxas de câmbio

do euro em USD teve uma porcentagem de redução de 20.10%, a série temporal dos registros de concentração de CO no ar 12,20 % e a série temporal de precipitações na República da China 15,81 % após aplicar a seleção de protótipos.

Tabela 18 – Média e desvio padrão (entre parênteses) do MMRE para as séries temporais reais e o p-valor do teste t-Student da comparação entre a nossa abordagem com a abordagem de Drago (2015).

	MMRE		
Série temporal	Abordagem de Drago	Modelo proposto	p-valor
Série temporal da Itália	1.3717 (0.6160)	0.8922 (0.2510)	< 0.0001
Série temporal do USD	1.1844 (0.0877)	0.4369 (0.0633)	< 0.0001
Série temporal da China	1.0573 (0.8026)	0.6692 (0.4520)	< 0.0001

Fonte: Elaborada pelo autor (2021)

Os valores da Tabela 18, mostram claramente que a abordagem proposta neste capítulo, para as séries temporais analisadas, supera a abordagem de Drago (2015).

Por fim, o desempenho do modelo VAR para $boxplots$ foi comparado com os resultados do modelo ARIMA tradicional utilizando as médias para representar as classes das séries temporais simbólicas, a média e o desvio padrão do MMRE para os resultados obtidos são apresentados na Tabela 19. A tabela também mostra os valores p do teste t-Student para amostras pareadas a um nível de significância de 5% da comparação realizada. Novamente, a hipótese nula é rejeitada com $p < 0,05$ mostrando-se que o modelo VAR melhora as previsões para séries temporais de $boxplots$. De acordo com os experimentos realizados, a abordagem apresentada neste capítulo mostrou ser a melhor opção em termos de acurácia na previsão de séries temporais de $boxplots$.

Tabela 19 – Média e desvio padrão (entre parênteses) do MMRE para as séries temporais reais e o p-valor do teste t-Student da comparação entre a nossa abordagem com a abordagem de Drago (2015).

	MMRE		
Série temporal	ARIMA	VAR	p-valor
Série temporal da Itália	0.9421 (0.3975)	0.8922 (0.2510)	< 0.0001
Série temporal do USD	0.5142 (0.0974)	0.4369 (0.0633)	< 0.0001
Série temporal da China	0.8244 (0.6461)	0.6692 (0.4520)	< 0.0001

Fonte: Elaborada pelo autor (2021)

5 CONCLUSÕES

Este capítulo apresenta as principais contribuições produzidas durante os quatro anos de pesquisa para o desenvolvimento desta tese de doutorado. Um artigo referente aos resultados apresentados no Capítulo 3 foi submetido para publicação na *Information Sciences*. Os resultados do Capítulo 4 foram publicados no periódico internacional *Expert Systems with Applications*. Por fim, serão enunciadas possíveis linhas de trabalho futuro, não apenas no nível teórico, mas também no nível prático.

5.1 CONSIDERAÇÕES FINAIS

Esta tese de doutorado apresentou duas novas abordagens para modelagem e predição de dados simbólicos de tipo quartis, como caso especial dados representados por $boxplots$. As duas abordagens propostas foram: o método de regressão linear parametrizada e a previsão de séries temporais. A avaliação foi realizada por meio do comportamento médio da magnitude média dos erros relativos das predições no contexto de experimentos de Monte Carlo, bem como da aplicação em dados reais.

A primeira abordagem é um método em que as variáveis regressoras de tipo $boxplot$ são parametrizadas através da equação da reta. Além disso, uma extensão da transformação $Box - Cox$ resolve o problema da coerência matemática nos $boxplots$ estimados. O método proposto foi comparado com os principais modelos da literatura de ADS para dados intervalares (MC (BILLARD; DIDAY, 2000), MinMax (BILLARD; DIDAY, 2002) e MCR (NETO; CARVALHO, 2008)). Todos os resultados obtidos mostraram que o nosso método supera todos os testados.

A segunda abordagem é um método que combina a seleção de protótipos e o modelo VAR para estimar $boxplots$. Os resultados demonstraram que a seleção dos protótipos usando a informação mútua reduziu o tamanho das séries temporais sem afetar a precisão do modelo. Além disso, observou-se que houve uma melhoria na previsão de séries temporais de $boxplots$ com comportamento linear ou não linear em relação ao modelo posposto por Drago (2015).

Cabe ressaltar que esta tese tem sido um esforço multidisciplinar, não apenas para combinar duas áreas, como a regressão linear e a análise de dados simbólicos, mas também porque, em seu desenvolvimento foram usados tópicos relacionados a áreas tão variadas quanto são a teoria de informação (quando usamos informação mútua para selecionar protótipos), econometria (ao

trabalhar com séries financeiras), meteorologia (como campos de aplicação), etc. A seguir, serão resumidas as contribuições mais importantes desta tese.

5.2 PRINCIPAIS CONTRIBUIÇÕES

A maior parte da análise atual de dados simbólicos se concentra em conjuntos de dados de intervalo. Enquanto que os estudos envolvendo outros dados simbólicos têm sido pouco explorados. Isso se deve ao fato de que a formulação de modelos de regressão não é trivial para variáveis simbólicas de natureza complexa. No entanto, a presente teses explorou as variáveis simbólicas multi-valoradas de tipo quartis: como caso especial os $boxplot$. Esse tipo de variável não é tão simples quanto intervalo e nem tão complexa quanto histograma, mas representa uma distribuição completa dos dados quantitativos e identifica se existem $outliers$ e quais são seus valores. De forma explícita, as principais contribuições deste estudo foram:

- Os dados são agregados em classes relacionadas ao intervalo temporal escolhido (diário, semana e mês). Essa escolha dependerá dos recursos de dados específicos que o analista deseja estudar;
- As classes são descritas por dados simbólicos multi-valorados de tipo quartis e $boxplots$ são construídos;
- Um novo método de regressão linear parametrizada para variáveis simbólicas de tipo $boxplot$;
- Uma extensão da transformação Box-Cox para dados de $boxplots$ para resolver o problema da coerência matemática;
- Uma aplicação do mundo real para previsão de altas temperaturas que dá suporte para identificação de falhas nos equipamentos das UTE brasileiras;
- Aplicação do algoritmo de seleção de protótipos usando informação mútua nas classes da série temporal simbólica;
- Primeiro método de séries temporais múltiplas para dados de quartis simbólicos;
- Proposta de uma nova representação para descrever os $boxplots$ que representam as classes da série temporal simbólica;

- Estimação de múltiplas séries temporais aplicada a dados simbólicos de quartis por meio do modelo vetorial autorregressivo (VAR);
- Experimentos no $framework$ de simulações de Monte Carlo sobre diferentes conjuntos de dados simbólicos multi-valorados reais e sintéticos, que consideram dados relevantes e pouco explorados no domínio de ADS.

As principais conclusões que foram obtidas no desenvolvimento da tese são as seguintes:

- Os $boxplots$, são ferramentas muito versáteis para representar variabilidade, pois são capazes de se adaptar às diferentes necessidades do analista;
- A complexidade envolvida no trabalho com variáveis de tipo $boxplots$ é maior que a envolvida no trabalho com intervalos, mas é muito menor do que a envolvida no trabalho com outras variáveis simbólicas em SDA, como o caso de histogramas;
- O MRPB tem a vantagem de usar método dos mínimos quadrados, sem suposição para a distribuição de probabilidade de erros;
- O MRPB usando as 5 curvas (m, Q_1, Q_2, Q_3 e M) para prever variáveis de tipo $boxplot$ supera todos os modelos testados para conjunto de dados sem e com dados ruidosos;
- A coerência matemática de qualquer previsão é controlada através da transformação $Box - Cox$ para dados de $boxplots$;
- A seleção de protótipos usando informação mútua se mostrou útil na redução do conjunto de dados nas classes da série simbólica sem afetar a qualidade da previsão; além do mais, no caso de séries com ruídos, a seleção de protótipos filtra a série temporal, o que leva a uma diminuição dos erros de previsão;
- A seleção de protótipos usando informação mútua resulta em uma previsão de séries temporais simbólicas robusta contra os efeitos do ruído;
- A capacidade preditiva da abordagem apresentada para séries temporais simbólicas é notável. O modelo multivariado gera previsões mais precisas do que vários modelos univariados, uma vez que as curvas são ajustadas conjuntamente. Em todos os exemplos analisados se obteve melhores resultados que o modelo de referência (DRAGO, 2015);

- A representação por uma lista de três valores numéricos para as séries temporais reduz a quantidade de curvas ajustadas e permite gerar um modelo parcimonioso sem perda de precisão;
- A abordagem proposta para séries temporais simbólicas de $boxplots$ permite levar em conta a variabilidade contida nos dados e supera o modelo clássico ARIMA quando o analista tem interesse em estudar os dados temporais de nível superior aos registrados, como por exemplo, dia quando os dados são descritos por hora ou minuto.

5.3 TRABALHOS FUTUROS

Apesar de considerar-se que o objetivo da tese foi cumprido, há muitas melhorias que podem ser feitas para alcançar melhores resultados. Portanto, a seguir são apresentadas algumas das questões em aberto que merecem mais pesquisas.

1. A primeira sugestão refere-se às as transformações para $boxplots$ a fim de se realizar regressão não-linear e ainda garantir a coerência matemática na predição.
2. A segunda sugestão refere-se à implementação do método de seleção de protótipos. Como apresentado, o método já tinha sido desenvolvido e não foi modificado. Assim, como trabalho futuro poderia ser feita uma otimização do parâmetro (α) chamado "limiar de informação mútua" que o algoritmo possui para melhorar a qualidade da seleção.
3. No contexto de previsão de séries temporais, explorar outros modelos de previsão, tais como o ARMAX e propor uma extensão para previsão de séries temporais de quartis.
4. Por outra parte, também se propõe aplicar as abordagens proposta na predição de outros dados reais relevantes para a sociedade.
5. Propor novos cenários para a validação experimental das soluções propostas, inserindo novas medidas de diagnóstico.

5.4 PUBLICAÇÕES

1. Dailys M.A. Reyes, Renata M.C.R. de Souza, Adriano L.I. de Oliveira, A three-stage approach for modeling multiple time series applied to symbolic quartile data, Expert Systems with Applications, Volume 187, 2022, https://doi.org/10.1016/j.eswa.2021.115884.

2. **(Submetido para publicação)** Dailys M.A. Reyes, Leandro C. Souza, Renata M.C.R. de Souza, Adriano L.I. de Oliveira, Parametrized Linear Regression for Boxplot-Multivalued Data Applied to Brazilian Electric Sector, Information Sciences.

REFERÊNCIAS

AHA, D. W. Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms. *International Journal of Man-Machine Studies*, v. 36, n. 2, p. 267–287, 1992.

AKAIKE, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control*, v. 19, p. 716–723, 1974.

ANGADIA, S. A.; KAGAWADEB, V. C. A robust face recognition approach through symbolic modeling of polar fft features. *Pattern Recognition*, v. 71, p. 235–248, 2017.

ARROYO, J.; GONZÁLES-RIVERA, G.; MATÉ, C. Forecasting with interval and histogram data. In: *Handbook of empirical economics and finance*. [S.l.]: Chapman & Hall, 2011. p. 37–48.

ARROYO, J.; MATÉ, C. Introducing interval time series: Accuracy measures. *COMPSTAT 2006, proceedings in computational statistics*, p. 1139–1146, 2006.

ARROYO, J.; MATÉ, C.; ROQUE, A. M. S. Hierarchical clustering for boxplot variables. In: *Data Science and Classification, Berlin, Heidelberg*. [S.l.: s.n.], 2008. p. 558–573.

ARROYO, J.; ROQUE, A. M. S.; MATÉ, C.; SARABIA, A. Exponential smoothing methods for interval time series. In: *STSP07: First European Symposium on Time Series Prediction (TSP)*. [S.l.]: Proceedings of Symposium, Espoo, Finland, 2007.

BATISTA, G. E.; PRATI, R. C.; MONARD, M. C. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, v. 6, n. 1, p. 20–29, 2004.

BEATON, A. E.; TUKEY, J. W. The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. *Technometrics*, Taylor & Francis Group, v. 16, n. 2, p. 147–185, 1974.

BELSON, W. Matching and prediction on the principle of biological classification. *Applied statistics*, JSTOR, p. 65–75, 1959.

BENJAMINI, Y. Opening the box of a boxplot. *The American Statistician*, v. 42, n. 4, p. 257–262, 1988.

BERGMEIR, C.; BENíTEZ, J. M. On the use of cross-validation for time series predictor evaluation. *Information Sciences*, v. 191, p. 192–213, 2012. ISSN 0020-0255. Data Mining for Software Trustworthiness.

BERTRAND, P.; GOUPIL, F. Descriptive statistics for symbolic data. In: *Analysis of symbolic data*. [S.l.]: Springer, 2000. p. 106–124.

BEZERRA, B. L. D.; CARVALHO, F. D. A. D. Symbolic data analysis tools for recommendation systems. *Knowledge and Information Systems*, v. 26, p. 385–418, 2011.

BILLARD, L. Dependencies in bivariate interval-valued symbolic data. In: *Classification, Clustering, and Data Mining Applications*. [S.l.]: Springer, 2004. p. 319–324.

BILLARD, L.; DIDAY, E. Regression analysis for interval-valued data. In: *Data Analysis, Classification, and Related Methods*. [S.l.]: Springer, 2000. p. 369–374.

BILLARD, L.; DIDAY, E. Symbolic regression analysis. In: *Classification, Clustering, and Data Analysis*. [S.l.]: Springer, 2002. p. 281–288.

BILLARD, L.; DIDAY, E. From the statistics of data to the statistics of knowledge: Symbolic data analysis. *Journal of the American Statistical Association*, v. 98, n. 462, p. 470–487, 2003.

BILLARD, L.; DIDAY, E. Symbolic data analysis: Conceptual statistics and data mining. *England: Wiley & Sons Ltd*, 2006.

BILLARD, L.; DIDAY, E. *Clustering methodology for symbolic data*. [S.l.]: John Wiley & Sons, 2019.

BOCK, H. Automatische klassifikation. *Studia Mathematica/Mathematische Lehrbücher*, v. 24, 1974.

BOCK, H.-H. Clustering algorithms and kohonen maps for symbolic data. *Journal of the Japanese Society of Computational Statistics*, v. 15, n. 2, p. 217–229, 2002.

BOCK, H. H.; DIDAY, E. *Analysis of symbolic data: exploratory methods for extracting statistical information from complex data*. [S.l.]: Springer Science & Business Media, 2000.

BOX, G. E.; JENKINS, G. M. *Time Series Analysis: Forecasting and Control*. [S.l.]: Holden Day, San Fransisco, 1970.

BOX, G. E. P.; COX, D. R. An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, v. 26, n. 2, p. 211–243, 1964.

BRITO, P. Symbolic data analysis: another look at the interaction of data mining and statistics. *WIREs Computational Statistic*, v. 4, n. 4, p. 281–295, 2014.

CARVALHO, F. A. T. D. Histograms in symbolic data analysis. *Annals of Operations Research*, Springer, v. 55, n. 2, p. 299–322, 1995.

CARVALHO, F. d. A. T. D.; PIMENTEL, J. T.; BEZERRA, L. X. T.; SOUZA, R. M. C. R. D. Clustering symbolic interval data based on a single adaptive hausdorff distance. In: IEEE. *2007 IEEE International Conference on Systems, Man and Cybernetics*. [S.l.], 2007. p. 451–455.

CARVALHO, F. d. A. T. D.; SOUZA, R. M. C. R. D.; CHAVENT, M.; LECHEVALLIER, Y. Adaptive hausdorff distances and dynamic clustering of symbolic interval data. *Pattern Recognition Letters*, Elsevier, v. 27, n. 3, p. 167–179, 2006.

CARVALHO, F. d. A. T. de; SOUZA, R. M. C. R. D. Unsupervised pattern recognition methods for interval data using non-quadratic distances. *Electronics Letters*, The Institution of Engineering & Technology, v. 39, n. 5, p. 433–434, 2003.

CAZES, P.; CHOUAKRIA, A.; DIDAY, E.; SCHEKTMAN, Y. Extension de l'analyse en composantes principales à des données de type intervalle. *Revue de Statistique appliquée*, v. 45, n. 3, p. 5–24, 1997.

CHACÓN, J. E.; RODRÍGUEZ, O. Regression models for symbolic interval-valued variables. *Entropy*, Multidisciplinary Digital Publishing Institute, v. 23, n. 4, p. 429, 2021.

CHATFIELD, C. *Time-Series Forecasting*. [S.l.]: Chapman Hall/CRC, London, 2001.

CHEUN, Y. An empirical model of daily highs and lows. *International Journal of Finance and Economics*, v. 12, p. 1–20, 2007.

CHOUAKRIA, A.; DIDAY, E.; CAZES, P. An improved factorial representation of symbolic objects. *Knowledge Extraction from Statistical Data*, v. 301, p. 305, 1998.

CIAMPI, A.; DIDAY, E.; LEBBE, J.; PÉRINEL, E.; VIGNES, R. Growing a tree classifier with imprecise data. *Pattern Recognition Letters*, Elsevier, v. 21, n. 9, p. 787–803, 2000.

CORDEIRO, I. Z. *Ser & Vencer*. [S.l.]: Editora UNOESC, 2017. 214 p.

COSTA, A. F.; PIMENTEL, B. A.; SOUZA, R. M. Clustering interval data through kernel-induced feature space. *Journal of Intelligent Information Systems*, v. 40, n. 1, p. 109–140, 2013.

COVER, T.; THOMAS, J. *Elements of Information Theory*. [S.l.]: Wiley-Interscience, 1991.

DIAS, S.; BRITO, P. Linear regression model with histogram-valued variables. *Statistical Analysis and Data Mining*, v. 8, n. 2, p. 75–113, 2015.

DIDAY, E. The symbolic approach in clustering and relating methods of data analysis: The basic choices. In: *Conference of the International Federation of Classification Societies*. [S.l.: s.n.], 1987. p. 673–684.

DIDAY, E. Introduction à l'approche symbolique en analyse des données. *Revue française d'automatique, d'informatique et de recherche opérationnelle. Recherche opérationnelle*, v. 23, n. 2, p. 193–236, 1989.

DIDAY, E. Des objets de l'analyse des données à ceux de l'analyse des connaissances. *Induction Symbolique et Numérique à partir de données, Kodratoff Y. et Diday E. Eds., CEPADUES*, 1991.

DIDAY, E. Thinking by classes in data science: the symbolic data analysis paradigm. *WIREs Computational Statistic*, v. 8, n. 5, p. 172–205, 2016.

DIDAY, E.; NOIRHOMME-FRAITURE, M. *Symbolic data analysis and the SODAS software*. [S.l.]: Wiley Online Library, 2008.

DIDAY, E.; SIMON, J. Clustering analysis. In: *Digital pattern recognition*. [S.l.]: Springer, 1980. p. 47–94.

DOMINGUES, M. A. O.; SOUZA, R. M. C. R. D.; CYSNEIROS, F. J. A. A robust method for linear regression of symbolic interval data. *Pattern Recognition Letters*, Elsevier, v. 31, n. 13, p. 1991–1996, 2010.

DOUZAL-CHOUAKRIA, A. *Extension des méthodes d'analyse factorielles à des données de type intervalle*. Tese (Doutorado) — Paris IX Dauphine, 1998.

DRAGO, C. Forecasting boxplot time series. In: *International Association for Statistical Computing - Joint Meeting of IASC-ABE Satellite Conference for the 60th ISI WSC, Buzios*. [S.l.: s.n.], 2015.

DRAPER, N. R.; SMITH, H. *Applied regression analysis*. [S.l.]: John Wiley and Sons, 1998. v. 326.

ENGLE, R. F.; GRANGER, C. W. J. Co-integration and error correction: Representation, estimation, and testing. *Econometrica*, v. 55, n. 2, p. 251–276, 1987.

FAGUNDES, R. A.; SOUZA, R. D.; CYSNEIROS, F. Interval kernel regression. *Neurocomputing*, v. 128, p. 371–388, 2014.

FAGUNDES, R. A. A.; SOUZA, R. M. C. R. D.; CYSNEIROS, F. J. A. Robust regression with application to symbolic interval data. *Engineering Applications of Artificial Intelligence*, Elsevier, v. 26, n. 1, p. 564–573, 2013.

FAN, L.; POH, K. L. Improving the naïve bayes classifier. In: *Encyclopedia of Artificial Intelligence*. [S.l.]: Hershey, PA: IGI Global, 2009. p. 879–883.

FOSS, T.; STENSRUD, E.; KITCHENHAM, B.; MYRTVEIT, I. A simulation study of the model evaluation criterion mmre. *IEEE Transactions on Software Engineering*, v. 29, n. 11, p. 985–995, 2003.

GARCÍA-ASCANIO, C.; MATÉ, C. Electric power demand forecasting using interval time series: A comparison between var and imlp. *Energy Policy*, Elsevier, v. 38, n. 2, p. 715–725, 2010.

GETTLER-SUMMA, M.; PARDOUX, C. Symbolic approaches for three-way data. In: *Analysis of Symbolic Data*. [S.l.]: Springer, 2000. p. 342–354.

GOOIJER, J. G. D.; HYNDMAN, R. J. 25 years of time series forecasting. *International Journal of Forecasting*, v. 22, n. 3, p. 443–473, 2006.

GOWDA, K. C.; DIDAY, E. Symbolic clustering using a new dissimilarity measure. *Pattern Recognition*, Elsevier, v. 24, n. 6, p. 567–578, 1991.

GOWDA, K. C.; DIDAY, E. Symbolic clustering using a new similarity measure. *IEEE Transactions on Systems, Man, and Cybernetics*, IEEE, v. 22, n. 2, p. 368–378, 1992.

GOWDA, K. C.; RAVI, T. Divisive clustering of symbolic objects using the concepts of both similarity and dissimilarity. *Pattern recognition*, Elsevier, v. 28, n. 8, p. 1277–1282, 1995.

GUILLEN, A.; HERRERA, L. J.; RUBIO, G. Instance or prototype selection for function approximation using mutual information. In: *European Symposium on Time Series Prediction—ESTSP*. [S.l.: s.n.], 2008. p. 67–75.

GUILLEN, A.; HERRERA, L. J.; RUBIO, G.; POMARES, H.; LENDASSE, A.; ROJAS, I. Applying mutual information for prototype or instance selection in regression problems. In: *ESANN 2009, 17th European Symposium on Artificial Neural Networks, Bruges, Belgium*. [S.l.: s.n.], 2009.

GUILLEN, A.; HERRERA, L. J.; RUBIO, G.; POMARES, H.; LENDASSE, A.; ROJAS, I. New method for instance or prototype selection using mutual information in time series prediction. *Neurocomputing*, v. 73, p. 2030–2038, 2010.

GURU, D. S.; KIRANAGI, B. B.; NAGABHUSHAN, P. Multivalued type proximity measure and concept of mutual similarity value useful for clustering symbolic patterns. *Pattern Recognition Letters*, Elsevier, v. 25, n. 10, p. 1203–1213, 2004.

HAMILTON, J. D. *Time Series Analysis*. 1ª edition. ed. [S.l.]: Princeton University Press, 1994.

HAN, A.; HONG, Y.; LAI, k.; WANG, S. Interval time series analysis with an application to the sterling-dollar exchange rate. *Journal of Systems Science and Complexity*, v. 21, n. 4, p. 558–573, 2008.

HAO, P.; GUO, J. Constrained center and range joint model for interval-valued symbolic data regression. *Computational Statistics & Data Analysis*, v. 16, p. 106–138, 2017.

HART, P. E. The condensed nearest neighbor rule. *IEEE Transactions on Information Theory*, v. 14, n. 3, p. 515–516, 1968.

HÉNON, M. A two-dimensional mapping with a strange attractor. *Communications in Mathematical Physics*, v. 50, n. 1, p. 459–467, 1976.

HOLT, C. C. *Forecasting Seasonals and Trends by Exponentially Weighted Moving Averages*. Tese (Doutorado) — Carnegie Institute of Technology, 1957.

ICHINO, M.; YAGUCHI, H. Generalized minkowski metrics for mixed feature-type data analysis. *IEEE Transactions on Systems, Man, and Cybernetics*, IEEE, v. 24, n. 4, p. 698–708, 1994.

IRPINO, A. "spaghetti" pca analysis: An extension of principal components analysis to time dependent interval data. *Pattern recognition letters*, Elsevier, v. 27, n. 5, p. 504–513, 2006.

IRPINO, A.; ROMANO, E. Optimal histogram representation of large data sets: Fisher vs piecewise linear approximation. In: *EGC*. [S.l.: s.n.], 2007. p. 99–110.

ISHIBUCHI, H.; NAKASHIMA, T.; NII, M. Learning of neural networks with ga-based instance selection. In: *Proceedings Joint 9th IFSA World Congress and 20th NAFIPS International Conference (Cat. No. 01TH8569)*. [S.l.: s.n.], 2001. v. 4, p. 2102–2107.

KALMAN, R. E. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, v. 82, n. 1, p. 35–45, 1960.

KITCHENHAM, B.; PICKARD, L.; MACDONELL, S. G.; SHEPPERD, M. What accuracy statistics really measure. *IEE Proceedings-Software*, v. 148, p. 81–85, 2001.

KOHONEN, T. *Learning vector quantization for pattern recognition*. Tese (Doutorado) — Helsinki University of Technology, 1986.

KORDOS, M.; BIAŁKA, S.; BLACHNIK, M. Instance selection in logical rule extraction for regression problems. In: *ICAISC 2013: Artificial Intelligence and Soft Computing*. [S.l.: s.n.], 2013. v. 7895, p. 167–175.

KORDOS, M.; BLACHNIK, M.; WIECZOREK, T. Temperature prediction in electric arc furnace with neural network tree. In: *ICANN 2011: Artificial Neural Networks and Machine Learning*. [S.l.: s.n.], 2011. p. 71–78.

KUBAT, M.; MATWIN, S. Addressing the curse of imbalanced training sets: One-sided selection. In: *International Conference in Machine Learning*. [S.l.: s.n.], 1997. p. 179–186.

KULLBACK, S. *Information Theory and Statistics*. [S.l.]: New York: Dover, 1997.

LAURIKKALA, J. Improving identification of difficult small classes by balancing class distribution. In: *AIME 2001: Artificial Intelligence in Medicine*. [S.l.: s.n.], 2001. p. 63–66.

LAURO, C. N.; PALUMBO, F. Principal component analysis of interval data: a symbolic data analysis approach. *Computational statistics*, Citeseer, v. 15, n. 1, p. 73–87, 2000.

LAURO, N. C.; VERDE, R.; PALUMBO, F. Factorial discriminant analysis on symbolic objects. *Analysis of symbolic data Exploratory methods for extracting statistical information from complex data. Springer, Berlin Heidelberg New York*, p. 212–233, 2000.

LLATAS, M. B.; M., G.-S. J. Segmentation trees for stratified data. In: *Analysis of Symbolic Data*. [S.l.]: Springer, 2000. p. 266–293.

LUGER, G. *Artificial intelligence: structures and strategies for complex problem solving*. [S.l.]: Pearson education, 2005.

MAIA, A.; CARVALHO, F. A. D.; LUDEMIR, T. B. Symbolic interval time series forecasting using a hybrid model. In: *STSP07: First European Symposium on Time Series Prediction (TSP)*. [S.l.: s.n.], 2006.

MAIA, A. L. S.; CARVALHO, F. d. A. T. D. Fitting a least absolute deviation regression model on interval-valued data. In: SPRINGER. *Brazilian Symposium on Artificial Intelligence*. [S.l.], 2008. p. 207–216.

MAIA, A. L. S.; CARVALHO, F. d. A. T. D.; LUDERMIR, T. B. A hybrid model for symbolic interval time series forecasting. In: SPRINGER. *International Conference on Neural Information Processing*. [S.l.], 2006. p. 934–941.

MAIA, A. L. S.; CARVALHO, F. d. T. D.; LUDEMIR, T. B. Iforecasting models for interval-valued time series. *Neurocomputing*, v. 71, p. 3344–3352, 2008.

MATÉ, C.; ARROYO, J. Forecasting histogram time series with k-nearest neighbours methods. *International Journal of Forecasting*, v. 25, n. 1, p. 192–207, 2009.

MAY, R. M. Simple mathematical models with very complicated dynamics. *Nature*, v. 261, p. 459–467, 1976.

MBALLO, C.; DIDAY, E. Decision trees on interval valued variables. *The electronic journal of symbolic data analysis*, v. 3, n. 1, p. 8–18, 2005.

MCCREA, W. H. *Analytical Geometry of Three Dimensions*. [S.l.]: Dover Publications, 2012.

MICHALSKI, R.; STEPP, R.; DIDAY, E. *A recent advance in data analysis: Clustering objects into classes characterized by conjunctive concepts*. [S.l.: s.n.], 1981.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. *Introduction to linear regression analysis*. [S.l.]: John Wiley and Sons, 2001.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. *Introduction to linear regression analysis*. [S.l.]: John Wiley & Sons, 2015.

MORADI, M.; CHAIBAKHSH, A.; RAMEZANI, A. An intelligent hybrid technique for fault detection and condition monitoring of a thermal power plant. *Applied Mathematical Modelling*, v. 60, p. 34–47, 2018.

MORGAN, J.; SONQUIST, J. Problems in the analysis of survey data, and a proposal. *Journal of the American statistical association*, Taylor & Francis Group, v. 58, n. 302, p. 415–434, 1963.

MORINEAU, A.; SAMMARTINO, A.-E.; GETTLER-SUMMA, M.; PARDOUX, C. Analyses des données et modélisation des séries temporelles. application à la prévision des ventes de périodiques. *Revue de statistique appliquée*, v. 42, n. 4, p. 61–81, 1994.

MORRIS, G. *Candlestick Charting Explained: Timeless Techniques for Trading stocks and Futures*. [S.l.]: McGraw-Hill Education, 2006.

NAGABHUSHAN, P.; GOWDA, K. C.; DIDAY, E. Dimensionality reduction of symbolic data. *Pattern recognition letters*, Elsevier, v. 16, n. 2, p. 219–223, 1995.

NETO, E. d. A. L.; ANJOS, U. U. dos. Regression model for interval-valued variables based on copulas. *Journal of Applied Statistics*, Taylor & Francis, v. 42, n. 9, p. 2010–2029, 2015.

NETO, E. d. A. L.; CARVALHO, F. d. A. T. D. Centre and range method for fitting a linear regression model to symbolic interval data. *Computational Statistics & Data Analysis*, Elsevier, v. 52, n. 3, p. 1500–1515, 2008.

NETO, E. d. A. L.; CARVALHO, F. d. A. T. D. Constrained linear regression models for symbolic interval-valued variables. *Computational Statistics & Data Analysis*, Elsevier, v. 54, n. 2, p. 333–347, 2010.

NETO, E. d. A. L.; CORDEIRO, G. M.; CARVALHO, F. d. A. T. de. Bivariate symbolic regression models for interval-valued variables. *Journal of Statistical Computation and Simulation*, Taylor & Francis, v. 81, n. 11, p. 1727–1744, 2011.

NETO, E. de A. L.; CARVALHO, F. de A.T. de. An exponential-type kernel robust regression model for interval-valued variables. *Information Sciences*, v. 454-455, p. 419–442, 2018.

NETO, E. L.; CARVALHO, F. Nonlinear regression applied to interval-valued data. *Formal Pattern Analysis & Applications*, v. 20, p. 1–16, 08 2017.

NOIRHOMME-FRAITURE, M.; BRITO, P. Far beyond the classical data models: symbolic data analysis. *Statistical Analysis and Data Mining*, Wiley Online Library, v. 4, n. 2, p. 157–170, 2011.

NOIRHOMME-FRAITURE, M.; ROUARD, M. Zoom star: a solution to complex statistical object representation. In: SPRINGER. *Human-Computer Interaction INTERACT'97*. [S.l.], 1997. p. 100–101.

PEKALSKA, E.; DUIN, R.; PACLíK, P. Prototype selection for dissimilarity-based classifiers. *Pattern Recognition*, v. 39, n. 2, p. 189–208, 2006.

PEÑA, D. *Análisis De Series Temporales*. [S.l.]: ALIANZA Editorial, Madrid, 2005.

RACINE, J. Consistent cross-validatory model-selection for dependent data: hv-block cross-validation. *Journal of Econometrics*, v. 99, n. 1, p. 39–61, 2000. ISSN 0304-4076.

RENCHER, A. C.; SCHAALJE, G. B. *Linear models in statistics*. [S.l.]: John Wiley and Sons, 2008.

REYES, D. M.; SOUZA, R.; CYSNEIROS, F. Predicting symbolic interval-valued data through symmetrical nonlinear regression. *International Journal of Business Intelligence and Data Mining*, v. 12, p. 175, 01 2017.

RUSSELL, S.; NORVIG, P.; CANNY, J.; MALIK, J.; EDWARDS, D. *Artificial intelligence: a modern approach*. [S.l.]: Prentice hall Upper Saddle River, 2003. v. 2.

SALVADOR, G.; DERRAC, J.; RAMON, C. Prototype selection for nearest neighbor classification: Taxonomy and empirical study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 34, p. 417–435, 2012.

SAPORTA, G. Probabilités, analyse des données et statistique. *Paris, Éditions Technip*, 1990.

SEBBAN, M.; NOCK, R.; LALLICH, S. Stopping criterion for boosting-based data reduction techniques: from binary to multiclass problems. *Journal of Machine Learning Research*, v. 3, n. 4, p. 863–885, 2002.

SEBER, G. A.; LEE, A. J. *Linear regression analysis*. [S.l.]: John Wiley and Sons, 2012. v. 329.

SHANNON, C. A mathematical theory of communication. *The Bell System Technical Journal*, v. 27, n. 1, p. 379–423, 1948.

SILVA, A. C. G. D. *Dissimilarity fuctions analysis based on dynamic clustering for symbolic data*. 2005.

SILVA, W.; SOUZA, R.; CYSNEIROS, F. Polygonal data analysis: A new framework in symbolic data analysis. *Knowledge-Based Systems*, v. 163, 09 2018.

SILVA, W. J.; SOUZA, R. M.; CYSNEIROS, F. J. A. Bivariate elliptical regression for modeling interval-valued data. *Computational Statistics*, Springer, p. 1–26, 2022.

SILVA, W. J. F.; SOUZA, R. M. C. R.; CYSNEIROS, F. J. A. psda: A tool for extracting knowledge from symbolic data with an application in brazilian educational data. *Soft Computing*, v. 25, p. 1803–1819, 2021.

SLUTZKY, E. The summation of random causes as the source of cyclic processes. *Econometrica*, v. 5, n. 2, p. 105–146, 1937.

SNEATH, P.; SOKAL, R. Numerical taxonomy. *Nature*, Nature Publishing Group, v. 193, n. 4818, p. 855–860, 1962.

SORJAMAA, A.; HAO, J.; LENDASSE, A. Mutual information and k-nearest neighbors approximator for time series prediction. In: *Proceedings of the 15th international conference on Artificial neural networks: formal models and their applications*. [S.l.]: Springer, Berlin, 2005. Part III, p. 553–558.

SOUZA, L. C.; SOUZA, R. M. C. R. D.; AMARAL, G. J. A.; Silva Filho, T. M. A parametrized approach for linear regression of interval data. *Knowledge-Based Systems*, v. 131, p. 149–159, 2017. ISSN 0950-7051.

SOUZA, R. M. C. R. D.; CARVALHO, F. d. A. T. D. Clustering of interval data based on city–block distances. *Pattern Recognition Letters*, Elsevier, v. 25, n. 3, p. 353–365, 2004.

SOUZA, R. M. C. R. D.; QUEIROZ, D. C. F.; CYSNEIROS, F. J. A. Logistic regression-based pattern classifiers for symbolic interval data. *Pattern Analysis and Applications*, Springer, v. 14, n. 3, p. 273–282, 2011.

TEAM, R. D. C. *R: A language and environment for statistical computing*. 2006. [Acessado 16 Dezembro 2018]. Disponível em: <http://www.R-project.org>.

TELES, P.; BRITO, P. Modelling interval time series. In: *3rd IASC World Conference on Computational Statistics Data Analysis, Limassol, Cyprus*. [S.l.: s.n.], 2005.

TERMOCABO. 2019. Disponível em: <http://termocabo.com.br>.

TOLVI, J. Genetic algorithms for outlier detection and variable selection in linear regression models. *Soft Computing*, v. 8, n. 8, p. 527–533, 2004.

TOMEK. An experiment with the edited nearest-neighbor rule. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-6, n. 6, p. 448–452, 1976.

TOMEK, I. Two modifications of cnn. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-6, n. 11, p. 769–772, 1976.

TUKEY, J. W. *Exploratory Data Analysis*. [S.l.]: Addison-Wesley, Reading, MA, 1977.

VENABLES, W.; RIPLEY, B. *Modern applied statistics whit S (4a edição)*. [S.l.]: New York: Springer-Verlag, 2002.

VERDE, R.; IRPINO, A. Ordinary least squares for histogram data based on wasserstein distance. In: *Proceedings of COMPSTAT'2010*. [S.l.: s.n.], 2010. p. 581–588.

VERDE, R.; IRPINO, A. Linear regression for numeric symbolic variables: an ordinary least squares approach based on wasserstein distance. *Advances in Data Analysis and Classification*, v. 9, n. 1, p. 81–106, 2015.

WICKHAM, H.; STRYJEWSKI, L. 40 years of boxplots. *The American Statistician*, 2011.

WILSON, D. R.; MARTINEZ, T. R. Reduction techniques for instance-based learning algorithms. *Machine Learning*, v. 38, n. 3, p. 257–286, 2000.

WINTERS, P. R. Forecasting sales by exponentially weighted moving averages. *Management Science*, v. 6, n. 3, p. 231–362, 1960.

YULE, G. U. On a method of investigating periodicities in disturbed series, with special reference to wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society of London A.*, v. 226, p. 267–298, 1927.

ZELLNER, A.; TOBIAS, J. A note on aggregation, disaggregation and forecasting performance. *Journal of Forecasting*, v. 19, n. 5, p. 457–465, 2000.

ZHANG, J.; YIM, Y.; YANG, J. Intelligent selection of instances for prediction functions in lazy learning algorithms. *Artificial Intelligence Review*, v. 11, p. 175–191, 1997.

ZHANG, X.; BERANGER, B.; SISSON, S. A. Constructing likelihood functions for interval-valued random variables. *Scandinavian Journal of Statistics*, v. 47, n. 1, p. 1–35, March 2020.

ZUBEK, V. B.; DIETTERICH, T. G. Pruning improves heuristic search for cost-sensitive learning. *Proceedings of the International Conference on Machine Learning*, 2002.