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

RAFAELLA LEANDRA SOUZA DO NASCIMENTO

Modelos Lineares Generalizados para Dados Simbólicos do Tipo Poligonal

Recife

2025

RAFAELLA LEANDRA SOUZA DO NASCIMENTO

Modelos Lineares Generalizados para Dados Simbólicos do Tipo Poligonal

Tese apresentada ao Programa de Pós-Graduação em Ciências da Computação da Universidade Federal de Pernambuco, como requisito para a obtenção do título de Doutor em Ciências da Computação.

Área de Concentração: Inteligência Computacional

Orientador (a): Profa. Dra. Renata Maria Cardoso Rodrigues de Souza

Coorientador (a): Prof. Dr. Francisco José de Azevedo Cysneiros

Recife

2025

.Catalogação de Publicação na Fonte. UFPE - Biblioteca Central

Nascimento, Rafaella Leandra Souza do.
Modelos Lineares Generalizados para Dados Simbólicos do tipo poligonal / Rafaella Leandra Souza do Nascimento. - Recife, 2025.
112f.: il.

Tese (Doutorado)- Universidade Federal de Pernambuco, Centro de Informática, Programa de Pós-Graduação em Ciências da Computação, 2025.
Orientação: Renata Maria Cardoso Rodrigues de Souza.
Coorientação: Francisco José de Azevedo Cysneiros.

1. Modelos Lineares Generalizados; 2. Regressão; 3. Análise de Dados Simbólicos; 4. Dados poligonais; 5. Análise residual. I. Souza, Renata Maria Cardoso Rodrigues de. II. Cysneiros, Francisco José de Azevedo. III. Título.

UFPE-Biblioteca Central

Rafaella Leandra Souza do Nascimento

“Modelos Lineares Generalizados para Dados Simbólicos do Tipo Poligonal”

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 Doutora em Ciência da Computação. Área de Concentração: Inteligência Computacional.

Aprovado em 18/08/2025

_____

Orientadora: Profa. Renata Maria Cardoso Rodrigues de Souza

BANCA EXAMINADORA

_____

Prof. Adriano Lorena Inácio de Oliveira
Centro de Informática /UFPE

_____

Prof. Adenilton José da Silva
Centro de Informática / UFPE

_____

Profa. Dra. Roberta Andrade de Araújo Fagundes
Escola Politécnica de Pernambuco/UPE

_____

Prof. Getúlio José Amorim do Amaral
Departamento de Estatística / UFPE

_____

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

Dedico este trabalho à minha família, em especial aos meus pais, Leandro e Lucia, cuja dedicação, amor e apoio sempre foram a base da minha caminhada.

AGRADECIMENTOS

Agradeço, primeiramente, a Deus, pelo dom da vida e pela saúde que me permitiram trilhar esta jornada com perseverança e fé, enfrentando desafios e celebrando conquistas.

À minha família, base de tudo: meus pais, Leandro e Lúcia, e meu irmão Laudson, pelo amor, apoio e incentivo constante. Agradeço também as minhas avós, tios e primos, pela presença carinhosa, pelas palavras de encorajamento e pela torcida em cada etapa desta caminhada.

À minha orientadora, Profa. Renata, minha profunda gratidão pela orientação, dedicação, paciência e apoio ao longo desta pesquisa. Ao meu coorientador, Prof. Francisco, ou Cysneiros, pois seu olhar atento e crítico, suas contribuições e sua disposição para o diálogo foram fundamentais para o amadurecimento deste trabalho. Agradeço aos dois também por todas as oportunidades que fui encorajada a abraçar, que me desafiaram a sair da zona de conforto, fortaleceram minha confiança e contribuíram para o meu crescimento pessoal, acadêmico e profissional.

Aos colegas do grupo de pesquisa do CIn, pelos momentos de colaboração e pelas trocas de ideias, em especial, Wanessa e Wagner.

Aos amigos que consquistei na UFRPE, Geraldo e Mário, pelo apoio mútuo, pelas contribuições e parceria profissional que até hoje temos.

Aos amigos que a vida me deu, e que seguem firmes ao meu lado com apoio, torcida e amizade sincera — ainda que, por vezes, em meio à saudade: Steliane, Izabele, Anninha, Cássio e Fábio.

À toda equipe da Faculdade Senac, pelo apoio constante ao longo desta jornada, em especial à Helô, pelas flexibilidades necessárias que possibilitaram o cumprimento das atividades do doutorado.

Por fim, agradeço à CAPES pelo apoio financeiro desta pesquisa.

"E todo o propósito da educação é transformar espelhos em janelas" (HARRIS, 1978).

RESUMO

A Análise de Dados Simbólicos é uma abordagem que visa desenvolver métodos para dados descritos por variáveis através de diferentes representações, como conjuntos de categorias, lista de valores, intervalos, distribuição de probabilidade, entre outros. Os métodos de regressão são amplamente estudados neste contexto e diferentes modelos têm sido propostos, inclusive pelo tipo de representação que estes dados podem assumir. Os Modelos Lineares Generalizados constituem uma classe de modelos de regressão que permite a modelagem de dados provenientes de diferentes distribuições da família exponencial. Esses modelos utilizam uma função de ligação para relacionar a média da variável resposta a uma combinação linear das variáveis explicativas, ampliando assim a aplicabilidade dos métodos preditivos a diversos cenários. Neste contexto, o objetivo deste trabalho consiste em propor uma extensão de Modelos Lineares Generalizados para dados simbólicos do tipo poligonal. Esse tipo de variável visa conservar a variabilidade original presente em dados agrupados por um caminho de agregação. Foram considerados modelos com as distribuições Gama, Normal Inversa e Binomial. Nos modelos com distribuições contínuas, são propostos resíduos poligonais, avaliados por meio de abordagem gráfica e descritiva, além da análise da função linear predita e definição de uma medida de qualidade. Para o modelo Binomial, baseado na regressão logística, são desenvolvidas regras de classificação para os dados poligonais. Os resultados obtidos demonstram a aplicabilidade e a eficácia dos métodos propostos em cenários com dados simulados e reais. As discussões são fundamentadas em gráficos de diagnóstico, testes estatísticos e ganhos relativos com base no erro de predição, acurácia e precisão. Portanto, esta pesquisa resulta em uma abordagem de predição e diagnóstico de modelos que contribui para o avanço dos estudos em diversos cenários de dados simbólicos.

Palavras-chaves: Modelos Lineares Generalizados. Regressão. Análise de Dados Simbólicos. Dados Poligonais. Análise Residual.

ABSTRACT

Symbolic Data Analysis is an approach aimed at developing methods for data described by variables with different representations, such as sets of categories, lists of values, intervals, probability distributions, among others. Regression methods are widely studied in this context, and various models have been proposed, depending on the type of data representation. Generalized Linear Models (GLMs) constitute a class of regression models that allow modeling data from different distributions belonging to the exponential family. These models use a link function to relate the mean of the response variable to a linear combination of explanatory variables, thus expanding the applicability of predictive methods to various scenarios. In this context, the objective of this work is to propose an extension of Generalized Linear Models for symbolic data of the polygonal type. This type of variable aims to preserve the original variability present in grouped data through an aggregation pathway. Models based on Gamma, Inverse Gaussian, and Binomial distributions were considered. For models with continuous distributions, polygonal residuals are proposed and evaluated using graphical and descriptive approaches, in addition to analyzing the predicted linear function and defining a quality measure. For the Binomial model, based on logistic regression, classification rules are developed for the polygonal data. The results demonstrate the applicability and effectiveness of the proposed methods in both simulated and real data scenarios. The discussions are supported by diagnostic plots, statistical tests, and relative gains based on prediction error, accuracy and precision. Therefore, this research results in a prediction and diagnostic approach for models, contributing to the advancement of studies in various symbolic data scenarios.

Keywords: Generalized Linear Models. Regression. Symbolic Data Analysis. Polygonal-valued Data. Residual Analysis.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ADS	Análise de Dados Simbólicos
AM	Aprendizado de Máquina
BGLM	*Bivariate Generalized Linear Model*
DPR	Desvio Padrão Residual
EMQA	Erro Médio Quadrático da Área
EMQAC	Erro Médio Quadrático da Área e Centro Conjuntamente
EMQCR	Erro Médio Quadrático do Centro e Raio Conjuntamente
EMQDV	Erro Médio Quadrático da Distância dos Vértices
GR	Ganho Relativo
IA	Inteligência Artificial
IDPC-PP	Modelo de Classificação Intervalar baseado em Probabilidade *a Posteriori* Combinada
MC	Monte Carlo
MD	Mineração de Dados
MLG	Modelos Lineares Generalizados
PBIVAR	Modelo de Regressão Linear Bivariado
PMLG	Modelo Linear Generalizado Poligonal
PRL	Modelo de Regressão Linear Poligonal
PSO	*Particle Swarm Optmization*

LISTA DE ALGORITMOS

SUMÁRIO

1 INTRODUÇÃO

Extrair informação, armazenar e encontrar relações em grandes quantidades de dados é um dos tópicos centrais da atualidade. Processar os dados e obter conhecimento possibilita aos diferentes setores da sociedade definir estratégias e intervenções que mitiguem problemas e expliquem cenários relacionando variáveis (OUSSOUS et al., 2018). Nas últimas décadas, a sociedade tem vivenciado um rápido crescimento na geração e no uso de dados, impulsionado por sensores, redes sociais, transações digitais e dispositivos conectados, na qual em torno de 98% dos dados armazenados na web já tinham sido gerados em meados de 2015 (MACHADO, 2018). Esse fenômeno, conhecido como *big data*, envolve grandes volumes de dados produzidos em diferentes contextos (RAO et al., 2019).

Diante desse cenário, surgem novos desafios e oportunidades para o armazenamento, processamento e análise eficiente desses dados, a fim de extrair conhecimento relevante e apoiar a tomada de decisões em tempo real. A complexidade dos dados atuais excede as capacidades dos métodos tradicionais de armazenamento e análise, exigindo novas abordagens para capturar, processar e interpretar informações de forma eficiente. Estima-se que até 2028 a criação global de dados cresça para mais de 380 zetabytes (STATISTA, 2025). O desenvolvimento de novas técnicas é essencial para descobrir novos benefícios para diversas aplicações.

Esse movimento impulsiona a demanda por profissionais capazes de lidar com dados complexos, tanto no setor privado quanto no público e na academia, exigindo competências que vão desde a modelagem de dados e estatística até o desenvolvimento de soluções baseadas em Inteligência Artificial (IA). Além disso, aplicações dessas tecnologias já estão presentes em áreas como saúde, educação, logística, finanças e gestão pública, contribuindo diretamente para a inovação, a eficiência operacional e a tomada de decisões baseada em evidências. Segundo o relatório *Future of Jobs*[1] , publicado em 2020 pelo *World Economic Forum*, o cientista de dados ocupa o primeiro lugar na lista de carreiras promissoras para os próximos anos, seguido pelo especialista em IA e aprendizado de máquina, além do profissional focado em *big data*.

Dessa forma a área de ciência de dados e *big data* tem se consolidado como uma das mais estratégicas e promissoras no cenário contemporâneo. Entretanto, o crescente volume e a heterogeneidade dos dados coletados apresentam desafios que vão além do simples armazenamento e processamento em larga escala. Muitas vezes, os dados consistem em informações agregadas

[1] *Future of Jobs* 2020 - WEF.

ou outras estruturas complexas que não podem ser adequadamente capturadas por técnicas tradicionais de análise baseadas em valores pontuais, chamados de dados clássicos. É nesse contexto que a Análise de Dados Simbólicos (ADS) destaca-se, permitindo a representação e o tratamento de dados com estruturas internas específicas, preservando suas características originais (BILLARD; DIDAY, 2006).

A ADS fornece métodos para lidar com variáveis simbólicas, como intervalos, histogramas, conjuntos multivalorados e distribuições, presentes em diversas áreas. Essa abordagem amplia as possibilidades analíticas, viabilizando a extração de padrões e insights que técnicas tradicionais não capturam (BILLARD; DIDAY, 2006). Assim, a ADS configura-se como uma extensão necessária e complementar para mineração e análise de dados, sobretudo diante da complexidade crescente dos dados contemporâneos. Além disso, oferece ferramentas que permitem o processamento e análise de grandes volumes, possibilitando a descrição de grupos ou classes, a redução da dimensionalidade e a preservação da diversidade e confidencialidade dos dados.

Em diversos contextos reais, os dados são coletados originalmente em formatos simbólicos, como listas, intervalos ou histogramas. Por exemplo, variáveis meteorológicas, como temperatura, umidade, precipitação e velocidade do vento, são frequentemente registradas como intervalos ao longo do tempo. No cenário educacional, o desempenho individual de alunos em exames pode ser agregado para representar o desempenho por escolas ou regiões, considerando a variabilidade interna dessas agregações, o que é fundamental para estudos que envolvem grupos de interesse (NASCIMENTO et al., 2022).

Portanto, a ADS oferece uma estrutura que incorpora a variabilidade observada na representação dos dados, utilizando métodos que a consideram explicitamente. Além disso, constitui um conjunto de ferramentas capaz de lidar com dados massivos e heterogêneos. Essa nova forma de representação implica que as variáveis assumem formatos distintos, o que tem sido amplamente estudado na ADS, gerando técnicas específicas para cada tipo de dado simbólico apresentado. Dessa forma, a análise exploratória e a modelagem estatística clássicas são estendidas para os dados simbólicos (DIDAY, 2016).

Visando contribuir para os avanços práticos e teóricos da modelagem estatística e computacional, este trabalho apresenta uma abordagem preditiva e diagnóstica para dados simbólicos. O presente capítulo fundamenta essa abordagem, expõe seus objetivos e descreve a organização dos capítulos subsequentes.

1.1 MOTIVAÇÃO

Com o rápido avanço da ciência da informação e das tecnologias digitais, novas técnicas de mineração de dados, métodos computacionais e ferramentas de código aberto têm sido amplamente desenvolvidas e utilizadas para viabilizar o uso de *big data* (ABDALLA, 2022). A crescente disponibilidade de grandes volumes de dados tem ampliado as possibilidades de atender a demandas empresariais e sociais, tornando-se um recurso essencial em diversos contextos. Atualmente, o *big data* é aplicado em sistemas de recomendação, análise preditiva, detecção de padrões e elaboração de relatórios estatísticos, com impacto direto em áreas como gestão organizacional, meio ambiente, saúde, educação, redes sociais, cidades inteligentes e transmissão de dados (OUSSOUS et al., 2018). Essas aplicações têm se mostrado fundamentais no suporte a processos de recomendação, previsão e tomada de decisão, fortalecendo práticas baseadas na análise e no uso estratégico de dados.

Diante desse cenário, organizações de diferentes setores da sociedade estão cada vez mais dependentes do conhecimento extraído desses grandes volumes de dados e torna-se necessário utilizar modelos e algoritmos complexos capazes de produzir decisões e resultados confiáveis e repetíveis, além de descobrir *insights* ocultos por meio de análises de dados correlacionados (TIEN, 2017). Nesse contexto, a qualidade das decisões está diretamente vinculada à capacidade de compreender os dados disponíveis, integrar fontes diversas de informação e aplicar modelos analíticos robustos que possibilitem a geração de conhecimento útil, estratégico e aplicável a diferentes realidades.

Os algoritmos e técnicas da Mineração de Dados (MD) fornecem algumas das ilustrações mais claras dos princípios da ciência de dados, a qual é a interseção entre ciência da computação, estatística e domínios de estudo (SKIENA, 2017). Da estatística vêm a análise exploratória de dados, os testes de significância e a visualização de dados. Sobre o domínio do problema, é necessário ter uma sólida compreensão do cenário em que se está trabalhando para entender claramente os problemas do negócio e os padrões para avaliar quando eles forem adequadamente alcançados. E por fim, o conhecimento da ciência da computação permite o desenvolvimento da aprendizagem estatística e do Aprendizado de Máquina (AM) com tecnologias de computação de alto desempenho.

No entanto, quando as entidades em análise da MD não são elementos isolados, mas grupos reunidos com base em alguns critérios determinados devendo-se levar em conta a variabilidade inerente a cada grupo, abordagens específicas são necessárias (BRITO, 2014). Assim, a ADS

possibilita a agregação de dados no grau de granularidade definido, mantendo as informações sobre a variabilidade intrínseca dos dados para então analisar os dados resultantes a partir de análises estatísticas e de MD específicas.

Os dados presentes em bases de dados simbólicas representam uma extensão das informações contidas em bases de dados clássicas, apresentando-as de forma agregada. Esta característica alerta para a necessidade de desenvolver metodologias que considerem a complexidade, imprecisão e variabilidade presentes nas estruturas formadas (BILLARD; DIDAY, 2006). Os dados podem ser representados de diferentes formas, como listas, intervalos, histogramas, distribuições de frequência ou de probabilidade. Neste trabalho, destaca-se também a variável poligonal, que será explorada com maior aprofundamento.

Os dados simbólicos do tipo poligonal constituem uma nova representação de dados em ADS introduzida por Silva, Souza e Cysneiros (2019a). Para este tipo de variável, novas medidas de análise foram propostas como média, variância, histograma entre outros, assim como um modelo de regressão linear para dados do tipo polígonos. Em Silva, Souza e Cysneiros (2020) os experimentos mostraram a aplicabilidade da variável poligonal no cenário educacional para previsão de desempenho escolar.

Posteriormente, considerando a abordagem não supervisionada, Silva et al. 2023 apresentaram o primeiro algoritmo de clusterização dinâmica para dados simbólicos poligonais, com o objetivo de extrair informações de perfis de periódicos científicos. Em Srakar e Vecco 2021, um algoritmo de agrupamento para dados simbólicos poligonais é aplicado à análise de regimes empreendedores, proporcionando insights mais ricos do que os métodos tradicionais baseados em intervalos. Esses trabalhos demonstraram desempenho superior em relação aos métodos desenvolvidos para dados com valores intervalares, destacando o potencial das representações poligonais e abrindo caminho para novas investigações nesta área emergente.

Dito isto, Diday (2016) indica as razões para induzir dados simbólicos: (a) leva em consideração a variabilidade intrínseca a cada unidade; (b) garante a confidencialidade dos indivíduos; (c) agregar dados reduz o número de indivíduos e o número de variáveis definidas pelo valor único de cada categoria e, (d) transforma dados complexos não estruturados em simbólicos estruturados e possibilita a aplicação de ferramentas simbólicas. Além destas razões, Silva, Souza e Cysneiros (2019a) ressaltam que na agregação poligonal mais informações são armazenadas pois considera-se a média e a variância dos dados diferentemente da representação intervalar, a qual considera os limites inferior e superior de cada classe.

Introduzir uma nova variável exige o desenvolvimento de novas ferramentas de análise,

pois a maioria dos conceitos e métodos foram projetados principalmente para observações de valor clássico (BRITO, 2014). O desenvolvimento de novas ferramentas para variável poligonal se faz necessário, as quais sejam capazes de explorar, analisar e modelar variáveis assim como dar suporte à verificação de propriedades estatísticas, diagnóstico de modelos, distribuições teóricas, entre outras. Sabe-se que ADS amplia a análise de dados e diversas técnicas têm sido propostas, em especial as técnicas de regressão.

Na literatura simbólica mantém-se as suposições básicas da literatura clássica para regressão linear. No entanto, vale destacar que em muitos contextos de dados reais algumas destas suposições podem ser violadas, e portanto, não será apropriado utilizar o modelo de mínimos quadrados ordinários.

Neste contexto, os Modelos Lineares Generalizados (MLG) constituem um conjunto de modelos de regressão mais flexíveis às suposições supracitadas. Os dados podem ser oriundos de diferentes distribuições de probabilidade revelando uma relação não linear entre a variável resposta e a explicativa. Assim, os MLG utilizam funções de ligação que possibilitam relacionar a média da variável resposta à combinação linear da variável explicativa, estendendo a aplicabilidade dos métodos preditivos.

Portanto, este trabalho se faz significativo à medida que busca contribuir com a formação de uma abordagem de MLG aplicada a dados simbólicos do tipo poligonal. É realizada uma análise experimental para avaliação das técnicas de predição e diagnósticos propostas, buscando predizer variáveis de bases de dados simulados e dos atuais cenários de dados reais. Além disso, os resultados deste trabalho contribuem para a ampliação do acervo de informações da comunidade científica de ADS.

1.2 OBJETIVOS

O objetivo deste trabalho consiste em desenvolver uma abordagem para análise de Modelos Lineares Generalizados aplicados a dados simbólicos do tipo poligonal. Espera-se que os resultados obtidos e analisados ajudem a ratificar o desenvolvimento desta abordagem na predição de variáveis simbólicas nos diferentes cenários de dados da atualidade. Como objetivos específicos lista-se:

1. Definir modelos lineares generalizados aplicados a dados tipo poligonal.

2. Verificar a adequação de modelos aplicados a dados tipo poligonais através da definição

de resíduo poligonal.

3. Introduzir medidas de avaliação do erro preditivo baseadas em distância de vértices dos polígonos.

4. Criar um ambiente experimental para a avaliação do modelo proposto, utilizando bases de dados reais e simuladas.

5. Avaliar o desempenho da técnica proposta pelo erro de predição através do método de simulação Monte Carlo, comparando com técnicas da literatura de ADS.

6. Contribuir com a área de ADS, introduzindo uma modelagem de análise e predição de dados tipo poligonais e estendendo a aplicabilidade das técnicas de regressão com MLG nesta representação de dados.

1.3 QUESTÕES DE PESQUISA

Neste trabalho são apresentados métodos, experimentos simulados e aplicações em conjuntos de dados simulados e reais que visam responder as seguintes questões:

- Como Modelos Lineares Generalizados podem ser estendidos para variáveis poligonais simbólicas?
- Como resíduos poligonais podem ser definidos e utilizados na avaliação da qualidade do ajuste dos modelos?
- Como aplicar regressão logística em contextos em que os preditores são dados poligonais simbólicos?
- Quais regras de classificação baseadas em probabilidades *a posteriori* são mais eficazes nesse contexto?
- Os modelos desenvolvidos são eficazes na análise de conjuntos de dados reais?
- Como o desempenho dos modelos poligonais se compara a modelos baseados em intervalos em diferentes cenários de variabilidade e sobreposição de classes?

1.4 ORGANIZAÇÃO DA TESE

Os capítulos restantes desta tese encontram-se estruturados da seguinte forma:

2 INTRODUÇÃO: apresenta os principais conceitos relacionados à ADS e diferentes representações de dados, incluindo os dados tipo intervalar e poligonal. Em relação às modelagens aplicadas em ADS, o foco desta pesquisa concentram-se nas técnicas de regressão, portanto, definições que permeiam estes cenários são desenvolvidas. Além disso, apresenta uma visão geral sobre os trabalhos relacionados ao tema desta tese.

3 INTRODUÇÃO: explana sobre os materiais e métodos propostos nesta tese para definição da abordagem de MLG para dados simbólicos tipo poligonais. É descrito uma metodologia para gerar e descrever dados poligonais e construir MLG. Ainda define o diagnóstico de modelos, a partir dos resíduos poligonais e análise preditiva, a qual descreve medidas de erro de predição e de desempenho a partir de regras de classificação.

4 INTRODUÇÃO: descreve e discute os resultados dos experimentos efetuados para análise e avaliação dos métodos desenvolvidos utilizando dados simulados. Considera-se distribuições de dados assimétricos, como a Gama e a Normal Inversa.

5 INTRODUÇÃO: explana sobre os resultados dos métodos propostos nesta tese em cenários de dados reais. O capítulo ilustra a aplicabilidade da metodologia desenvolvida em fazer predições de variável meteorológica.

6 INTRODUÇÃO: apresenta e discute os resultados dos experimentos conduzidos para avaliar os métodos propostos com dados simulados, considerando a distribuição binomial avaliando cenários de classificação binária.

7 INTRODUÇÃO: explora a aplicação dos métodos desenvolvidos em dados reais, demonstrando sua efetividade na predição da variável relacionada à detecção de notícias *fake*.

8 INTRODUÇÃO: apresenta as considerações finais sobre os principais tópicos abordados, como contribuições e direcionamentos para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Este Capítulo explana os conceitos fundamentais utilizados como embasamento teórico e entendimento da abordagem proposta nesta tese. São discutidas definições da literatura de Análise de Dados Simbólicos (ADS) as quais permeiam representação de dados, métodos de regressão e de análise estatísticas para dados simbólicos do tipo intervalar e poligonal. Também, relacionam os principais trabalhos referentes ao tema desta pesquisa.

2.1 CONHECIMENTO A PARTIR DOS DADOS

O termo *big data* define conjuntos de dados grandes, complexos, diversos e heterogêneos que são gerados por diferentes fontes. Devido ao rápido avanço das tecnologias de hardware e das mídias de armazenamento digital, estes dados - provenientes de sensores, fluxos de cliques em sites, transações comerciais e econômicas e redes sociais - podem ser capturados, gerenciados, processados e analisados de forma estratégica (RAO et al., 2019). Os setores da sociedade estão conscientes de que a análise de dados está se tornando cada vez mais um fator vital para ser competitivo, descobrir novos *insights* e personalizar serviços (OUSSOUS et al., 2018). Para isto, técnicas específicas são necessárias para lidar com as particularidades de cada conjunto de dados. Nesse cenário, as ferramentas de Mineração de Dados (MD) e Aprendizado de Máquina (AM) auxiliam na descoberta de padrões e na geração de conhecimento útil para diversas organizações e aplicações.

A MD permite a aquisição de conhecimento que pode ser explorado de ângulos diferentes resultando em tomadas de decisão consistentes, controle de processos, gerenciamento de informação e processamento de consultas (WITTEN et al., 2005). Este conjunto de ferramentas é um campo que abrange diferentes áreas como AM, estatística, tecnologias de banco de dados, visualização e recuperação de informações (WLODARCZAK; ALLY; SOAR, 2015) resultando em extração de padrões desconhecidos, tendências inesperadas ou outras relações presentes (WITTEN et al., 2005). Portanto, é considerada como uma das fronteiras mais importantes em sistemas de banco de dados e um dos mais promissores desenvolvimentos interdisciplinares na indústria da informação (HAN; PEI; TONG, 2011).

Considerando as etapas do processo de descoberta de conhecimento, tem-se: (1) pré-processamento dos dados, constituindo o entendimento do problema e o tratamento dos dados;

(2) construção de padrões e modelos através da execução de algoritmos para extração de padrões; e (3) pós-processamento de dados, o qual refere-se a compreensão das saídas para geração de conhecimento. A escolha da técnica mais adequada depende de aspectos como a área do problema e dos dados disponíveis (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Os dados clássicos são representados por uma estrutura matricial $n \times p$, na qual cada linha representa uma entidade e cada coluna pertence a uma variável que pode ser numérica ou categórica. Outra característica é que um único valor é registrado para cada variável e para cada registro. No entanto, em algumas situações as unidades de interesse estão em um nível superior necessitando agregar os valores observados previamente à análise de dados (BRITO, 2014). Uma abordagem de agregação é calcular indicadores (como médias, medianas e desvios) para que os dados sejam ajustados à matriz $n \times p$ e assim, métodos clássicos de análises possam ser aplicados. Quando o tamanho da amostra é pequeno, esta abordagem extrai com facilidade as informações desejadas, porém esta prática acarreta considerável perda de informação, como a variabilidade intrínseca nos dados.

Dentre as abordagens oriundas da AM e da estatística que dão suporte à MD, a ADS apresenta uma extensão dos dados clássicos que se dá através de representação e análise de dados considerados de nível superior. Portanto, novos tipos de variáveis foram introduzidas as quais não são representados por valores reais ou categorias únicas, mas por conjuntos, intervalos ou distribuições de um determinado domínio (BRITO, 2014). As próximas seções abordam as características dos dados simbólicos e as ferramentas de análise e modelagem desenvolvidas na literatura.

2.2 DADOS SIMBÓLICOS

Os dados presentes em bases de dados simbólicas representam uma extensão das informações contidas em bases de dados clássicas, apresentando-as de forma agregada, nas quais as linhas correspondem aos indivíduos ou classes e as colunas são as variáveis simbólicas que caracterizam os indivíduos. Os objetos podem ser representados por conjuntos de categorias, intervalos, histogramas, distribuições de frequência entre outros.

Considere como exemplo um conjunto de dados com informações sobre pacientes diagnosticados com Covid-19 de diferentes cidades de um país (NASCIMENTO et al., 2022). As variáveis clássicas incluem informações pessoais e demográficas, características clínicas, resultados laboratoriais e opções de tratamento. Assim, as entidades individuais no conjunto de

dados clássico são os pacientes e as cidades podem agregá-los para obter um novo conjunto de dados referente a diferentes variáveis simbólicas, como mostrada na Tabela 1. As cidades são novas unidades, chamadas classes (DIDAY, 2016), e a variabilidade entre os pacientes dentro de suas cidades (classes) é descrita por variáveis simbólicas que expressam a variabilidade dos pacientes dentro de cada cidade.

Tabela 1 – Tabela com dados simbólicos de pacientes com Covid-19.

Cidade	**Sexo**	**Peso**	...	**Internamento**	**Condição Clínica**
C_1	{(0,6)F,(0,4)M}	[25,5; 128,16]		[10; 33]	{Leve, Urgente, Grave}
⋮	⋮	⋮		⋮	⋮
C_{200}	{(0,8)F,(0,2)M}	[19,30; 88,34]		[2; 55]	{Leve, Urgente, Grave}

Seja uma classe, a notação que a define é dada por $w \in S = \{w_1, ..., w_m\}$, onde m representa o número de classes (SILVA; SOUZA; CYSNEIROS, 2019a). Como no exemplo da Tabela 1, o registro C_1, na variável Internamento (em dias), agrupa todos os pacientes que compõe a classe cujo domínio é $D = \{x | x \in [10; 33]\}$. Esse domínio é chamado de descrição.

Este paradigma apresenta diversos tipos de representações para os dados como variáveis multivaloradas, intervalares, modais, histogramas de variáveis intervalares (DIDAY, 2016) e mais recente a variável poligonal (SILVA; SOUZA; CYSNEIROS, 2019a). Estes novos tipos de variáveis exigiram da comunidade científica novas ferramentas, por exemplo: medidas descritivas usuais como média, variância, correlação, distribuição de probabilidade, histogramas e outras foram recriadas para esta nova estrutura de dados (CARVALHO, 1995; BERTRAND; GOUPIL, 2000; BILLARD; DIDAY, 2003). A seguir, descreve-se a representação e ferramentas de análises para dados intervalar e poligonal.

2.2.1 Dados Simbólicos Intervalares

Dados simbólicos do tipo intervalo são geometricamente representados por meio de uma semi-reta $[a, b]$, com $a \neq b$. A combinação de p variáveis intervalares é geometricamente representada por um hiper-retângulo p-dimensional. Por exemplo, para $p = 2$, obtém-se um retângulo gerado pela combinação de duas variáveis intervalares. Esta representação pode ser vista na Figura 1

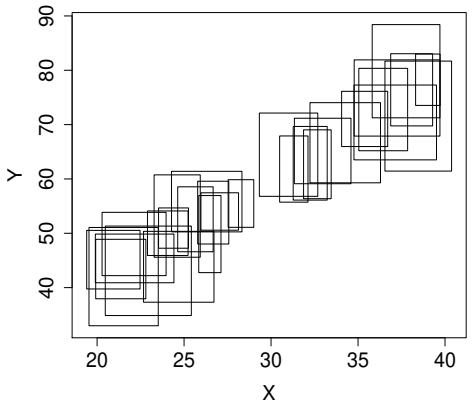


Figura 1 – Gráfico de dispersão com $p = 2$ variáveis do tipo intervalar.

Os dados podem ser naturalmente intervalares, como no caso da medição da temperatura de uma determinada região ao longo de um período, onde se registram valores mínimos e máximos. Outra forma é a transformação de tabelas clássicas em tabelas de dados intervalares, em que os limites inferior e superior do intervalo são definidos por $a_u = \min_{i \in \Omega_u} x_i$ e $b_u = \max_{i \in \Omega_u} x_i$, onde Ω_u é o conjunto de valores x_i pertencentes à categoria w_u. Técnicas para análise de dados simbólicos com valores intervalares possuem uma vasta literatura, com destaque especial para aplicações em modelos de regressão. Essas abordagens têm se mostrado eficazes na modelagem de variabilidade interna dos dados intervalares.

O primeiro trabalho no modelo de regressão para dados simbólicos tipo intervalar pode ser encontrado em Billard e Diday (2000) e Billard e Diday (2002). Lima Neto e De Carvalho (2008) consideraram uma representação para intervalo baseada no centro e na amplitude do intervalo. Além disso, eles desenvolveram um modelo de regressão baseado nesta representação. Lima Neto e De Carvalho (2010) propuseram um modelo de regressão linear restrito na representação do centro e do intervalo para garantir a coerência matemática entre os valores previstos dos limites inferior e superior dos intervalos.

Já em Fagundes, Souza e Cysneiros (2013) foi apresentado um método de previsão robusto para dados simbólicos de valor intervalar baseado na metodologia de regressão linear robusta. Os autores ainda indicam que problemas na escolha do mínimo-máximo podem surgir quando estes valores extremos são, de fato, outliers ou quando o conjunto de indivíduos a generalizar é composto por subconjuntos de diferentes distribuições, definindo o outlier intervalar. Hao e Guo (2017) apresentaram o modelo de regressão restrita para intervalos baseados em mínimos quadrados ordinários.

Souza et al. (2017) introduziu o método parametrizado, um modelo de regressão linear baseado na representação mínimo-máximo. Soares e Fagundes (2018) propuseram uma regressão quantílica intervalar para dados simbólicos intervalados representados por centros e intervalos. Lima Neto e De Carvalho (2018) introduziram um modelo robusto baseado no modelo de mínimos quadrados ponderados. Reyes et al. (2019) propõem um modelo linear para estimar o risco sistemático na precificação de ativos de capital e exemplifica a capacidade do modelo usando os preços diários de alta e baixa na Microsoft.

Embora existam diferentes abordagens de regressão para dados simbólicos tipo intervalar na literatura de ADS, é importante verificar se o modelo funciona bem para os dados coletados. Para isso, podem ser utilizadas medidas de diagnóstico e ferramentas gráficas baseadas em resíduos. Nesse contexto, Lima Neto et al. (2011) propôs o primeiro conceito de resíduos para dados simbólicos tipo intervalar como um valor contínuo único e considerou este conceito para o cálculo de medidas diagnósticas. Este conceito foi utilizado em relação a um modelo que os autores também introduziram. Este modelo assumiu a variável de resposta simbólica com valor de intervalo como um vetor aleatório bivariado com uma distribuição gaussiana bivariada. Os resíduos foram utilizados para fazer inferências sobre a distribuição das respostas, identificar outliers, entre outros aspectos.

Já em Nascimento et al. (2022) um novo conceito de resíduos para dados simbólicos de valor intervalar é introduzido. Esta definição considera os limites inferior e superior dos resíduos conjuntamente, diferentemente das definições encontradas na literatura (NETO; CORDEIRO; CARVALHO, 2011; XU, 2010) as quais consideram o resíduo intervalar baseado em resíduos estatísticos para dados clássicos. Esta abordagem leva em consideração a variabilidade intrínseca a cada classe para definir os resíduos (limites inferior e superior). Além disso, os autores consideram a versão ordinária e padronizada dos resíduos e ferramentas gráficas para investigar a adequação dos modelos de regressão linear.

2.2.1.1 Resíduo Intervalar

Na literatura clássica de modelos de regressão, as premissas básicas de regressão são: i) a relação entre a variável resposta e explicativas ser aproximadamente linear; ii) erro com média zero e variância constante; iii) erros não correlacionados e iv) erros que seguem distribuição aproximadamente normal. No entanto, estas suposições também são verificadas a partir da abordagem intervalar (NASCIMENTO et al., 2022).

Seja $\Omega = 1, \ldots, n$ um conjunto de dados de n objetos, cada um descrito por um vetor de intervalo $(\mathbf{x}_i, y_i)$ onde $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ com $x_{ij} = [a_{ij}, b_{ij}] \in \Im = \{[a, b] : a, b \in \Re, a \leq b\}$ $(j = 1, \ldots, p)$ e $y_i = [\alpha_i, \lambda_i] \in \Im = \{[\alpha, \lambda] : \alpha, \lambda \in \Re, \alpha \leq \lambda\}$. Os objetos são descritos por dados de centro e range dos intervalos. Seja $\mathbf{Y} = (y_1^c, \ldots, y_n^c, y_1^r, \ldots, y_n^r)^T$ a variável de resposta simbólica com valor intervalar com $y_i^c = (\alpha_i + \lambda_i)/2$ e $y_i^r = (\lambda_i - \alpha_i)$.

Considere $\mathbf{X} = (\boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{X_3}, \boldsymbol{X_4})$ a matriz de variáveis explicativas simbólicas com valor intervalar, com $\boldsymbol{X_1} = (\mathbf{1}_n^T, \mathbf{0}_n^T)^T$, $\boldsymbol{X_2} = (\mathbf{0}_n^T, \mathbf{1}_n^T)^T$, $\boldsymbol{X_3} = (\boldsymbol{x}_c^T, \mathbf{0}_n^T)^T$ e $\boldsymbol{X_4} = (\mathbf{0}_n^T, \boldsymbol{x}_r^T)^T$ onde $\boldsymbol{x}_c = (x_{1j}^c, \ldots, x_{nj}^c)^T$ com $x_{ij}^c = (a_{ij} + b_{ij})/2$, $\boldsymbol{x}_r = (x_{1j}^r, \ldots, x_{nj}^r)^T$ com $x_{ij}^r = (b_{ij} - a_{ij})$ $(j = 1, \ldots, p)$ e $\mathbf{0}_n$ e $\mathbf{1}_n$ são vetores zero e um, respectivamente. Em relação ao vetor $\mathbf{Y}$ e à matriz $\mathbf{X}$, a equação de regressão linear pode ser escrita da seguinte forma:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.1}$$

onde $\boldsymbol{\beta} = (\beta_0^c, \beta_1^c, \ldots, \beta_p^c,\ \beta_0^r, \beta_1^r,\ \ldots, \beta_p^r)^T$ é um vetor de parâmetros, $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}^c, \boldsymbol{\epsilon}^r)^{\ T}$ é um vetor erros com $\boldsymbol{\epsilon}^c = (\epsilon_1^c, \ldots, \epsilon_n^c)^T$ e $\boldsymbol{\epsilon}^r = (\epsilon_1^r, \ldots, \epsilon_n^r)^T$. Sejam os resíduos para o centro e o range de dados simbólicos com valor intervalar como $r_i^c = y_i^c - \hat{y}_i^c$ e $r_i^r = y_i^r - \hat{y}_i^r$. Assim, o resíduo ordinário intervalar (Δ) é definido como:

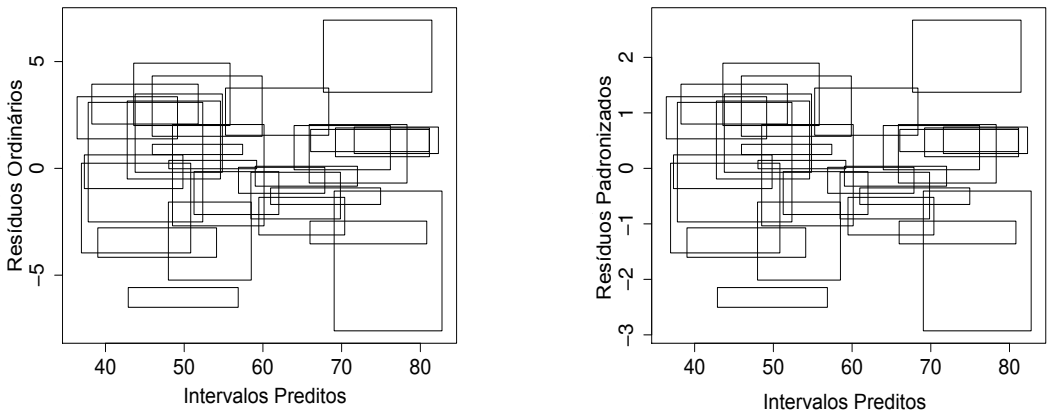
$$\begin{aligned}
\Delta_i &= [r_{il}, r_{iu}] = [(\alpha_i - \hat{\alpha}_i), (\lambda_i - \hat{\lambda}_i)] \\
&= [(y_i^c - y_i^r/2) - (\hat{y}_i^c - \hat{y}_i^r/2), (y_i^c + y_i^r/2) - (\hat{y}_i^c + \hat{y}_i^r/2)] \\
&= [(y_i^c - \hat{y}_i^c) - (y_i^r - \hat{y}_i^r)/2, (y_i^c - \hat{y}_i^c) + (y_i^r - \hat{y}_i^r)/2].
\end{aligned} \tag{2.2}$$

A versão padronizada para o resíduo Δ_i é definida como

$$\Delta_i^S = \left[\frac{r_{il}}{DPR}, \frac{r_{iu}}{DPR}\right]. \tag{2.3}$$

O elemento Desvio Padrão Residual (DPR) é o desvio padrão para o intervalo residual Δ, o qual é mostrado na Equação 2.4, seguindo a definição de desvio padrão para dados simbólicos tipo intervalar apresentados em Bertrand e Goupil (2000). Os exemplos mostrados na Figura 2b (a) e (b) sugerem que os erros são homocedásticos e aleatórios para resíduos ordinários e padronizados, respectivamente, sendo a variância constante e a suposição de linearidade satisfeita.

$$DPR = \sqrt{\frac{1}{3n}\sum_{i\in\Omega}(r_{iu}^2 + r_{iu}r_{il} + r_{il}^2) - \frac{1}{4n^2}\left[\sum_{i\in\Omega}\frac{r_{iu} + r_{il}}{2}\right]^2}. \tag{2.4}$$



(a) Ordinário (b) Padronizado

Figura 2 – Resíduos intervalares quando as suposições do modelo de regressão linear intervalar são satisfeitas.

Além disso, Nascimento et al. (2022 propuseram analisar os resíduos intervalar a partir do histograma intervalar. Portanto, medidas descritivas para dados simbólicos de valor intervalar são apresentadas. O k-ésimo momento e as medidas descritivas para dados simbólicos tipo intervalar são baseados em uma função de densidade empírica para o intervalo como encontrado em Bock e Diday (2000) e Billard e Diday (2000).

Dada uma variável simbólica com valor de intervalo Z, medida por para cada elemento da amostra aleatória $E = \{1, \ldots, n\}$. Para cada $i \in E$ denota-se $[a_i, b_i]$ um intervalo. Uma função de distribuição empírica de Z é uma função de n distribuições uniformes dada por

$$F_Z(\xi) = \frac{1}{n}\left\{\sum_{\xi \in Z(i)}\left(\frac{\xi - a_i}{b_i - a_i}\right) + \frac{\#\{i|\, \xi \geq b_i\}}{n}\right\}. \tag{2.5}$$

De acordo com Bertrand e Goupil (2000) a função densidade empírica de Z baseada na Equação (2.5) é definida como:

$$f(\xi) = \frac{1}{n}\sum_{i:\xi \in Z(i)}\frac{1}{b_i - a_i}. \tag{2.6}$$

O k-ésimo momento para uma variável simbólica intervalar Z é definido na Equação (2.7), onde $k = 0, 1, 2, 3, 4, \ldots$.

$$M_k = \int_{-\infty}^{+\infty} \xi^k \frac{1}{n}\sum_{i:\xi \in Z(i)}\frac{1}{b_i - a_i} d\xi. \tag{2.7}$$

O primeiro e segundo momentos empíricos para dados simbólicos intervalar são definidos

em Bertrand e Goupil (2000), respectivamente, pelas Equações (2.8) e (2.9).

$$M_1 = \frac{1}{n}\sum_{i\in E}\frac{b_i + a_i}{2}. \tag{2.8}$$

$$M_2 = \frac{1}{3n}\sum_{i\in E}[b_i^2 + b_i a_i + a_i^2]. \tag{2.9}$$

O terceiro e o quarto momento empírico foram apresentados em Nascimento et al. (2022 e representados, respectivamente, nas Equações (2.10) e (2.11).

$$M_3 = \frac{1}{4n}\sum_{i\in E}(b_i^3 + a_i^3 + a_i^2 b_i + a_i b_i^2). \tag{2.10}$$

$$M_4 = \frac{1}{5n}\sum_{i\in E}\frac{b_i^5 - a_i^5}{b - a}. \tag{2.11}$$

De acordo com Bertrand e Goupil (2000) e as Equações (2.8) e (2.9), a média e a variância empírica para dados simbólicos intervalares são apresentadas, respectivamente, como:

$$ME = \frac{1}{n}\sum_{i\in E}\frac{b_i + a_i}{2}, \tag{2.12}$$

$$VA = \frac{1}{3n}\sum_{i\in E}(a_i^2 + a_i b_i + b_i^2) - \frac{1}{4n^2}\left[\sum_{i\in E}(a_i + b_i)\right]^2. \tag{2.13}$$

E por fim, a assimetria e a curtose empírica para o dado simbólico intervalar são definidas, respectivamente, como segue as Equações (2.14) e (2.15).

$$SK = SK = M_3 - 3M_1M_2 + 2M_1^3. \tag{2.14}$$

$$KU = M_4 + 6M_1^2M_2 - 3M_1^4. \tag{2.15}$$

2.2.2 Dados Simbólicos Poligonais

Os dados simbólicos do tipo poligonal possuem como descrição um polígono e foram introduzidos por Silva, Souza e Cysneiros 2019a. Dessa forma, Z é uma variável aleatória simbólica

poligonal quando assume valores em um polígono da forma $Z = \xi = (a_1, b_1), ..., (a_l, b_l) \subset \mathbb{R}^2$ em que os segmentos de reta que ligam esses pontos formam uma figura poligonal. Outra forma de representar a variável é $Z = \xi = (\xi_1, \xi_2)$, onde $\xi_1 = a_1, ..., a_l$ e $\xi_2 = b_1, ..., b_l$, ou seja, os valores que a variável pode assumir no eixo das abcissas e no eixo das ordenadas, respectivamente.

Para agregar dados clássicos e transformá-los dados simbólicas poligonais, cada classe é transformada em um polígono com número de lados desejada $l \leq n$, onde n é o número de elementos. O método de representação para dados simbólicos poligonais é baseado em dois valores - $[centro, raio]$ -, sendo apta para representar polígonos regulares (SILVA; SOUZA; CYSNEIROS, 2019a). Esta representação transforma uma variável unidimensional em bidimensional utilizando coordenadas polares. Para exemplificar a transformação de dados em variáveis poligonais, considere a Tabela 2 sendo a descrição de dados clássicos do desempenho de alunos matriculados em uma determinada cidade.

Tabela 2 – Tabela com dados clássicos de alunos de uma cidade.

Aluno	**Cidade**	**Escola**	...	**Nota**$_1$	**Nota**$_2$
A_1	C_1	E_1		7,7	8,2
A_2	C_1	E_1		8,0	6,8
A_3	C_1	E_2		8,8	8,6
A_4	C_1	E_2		8,5	7,5
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
A_{119999}	C_1	E_{135}		9,9	10,0
A_{120000}	C_1	E_{135}		8,9	9,5

Seja n_j o número de indivíduos na classe j. Cada indivíduo é descrito por uma variável contínua X. Um polígono P_j, com L vértices, para $L \leq n_j$, inscrito em uma circunferência, pode ser definido como:

$$P_{j\ell} = (a_{j\ell}, b_{j\ell}) = \left(c_j + r_j \cos\left(\frac{2\pi\ell}{L}\right),\ c_j + r_j \sin\left(\frac{2\pi\ell}{L}\right) \right), \tag{2.16}$$

em que c_j representa o centro do polígono da classe j (isto é, a média de X na classe j) e $r_j = 2 \times \text{dp}(x_j)$ é o raio do polígono (ou da circunferência), sendo $\text{dp}(x_j)$ o desvio padrão de X na classe j, respectivamente. Cada $P_{j\ell}$ representa os pares de pontos que formam vértices do polígono regular P_j, com $\ell = 1, 2, \ldots, L$, onde $L \in \mathbb{N} \geq 3$ é o número de vértices do polígono (SILVA; SOUZA; CYSNEIROS, 2019a).

Com esta definição, a transformação dos dados da Tabela 2 em dados simbólicos baseados em centro e raio é representada na Tabela 3.

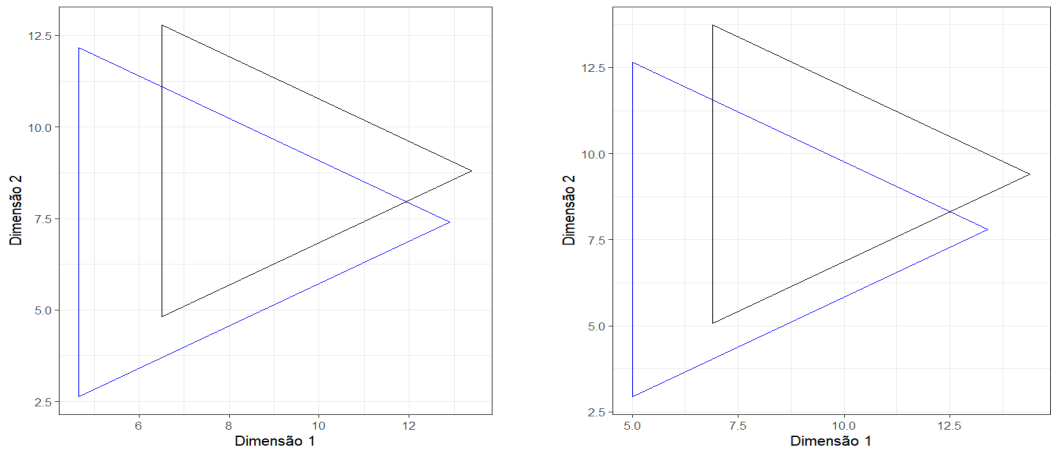
Tabela 3 – Tabela com dados de (centro; raio) para dados simbólicos poligonais.

Escola	...	**Nota**$_1$	**Nota**$_2$
E_1		(7,4; 5,5)	(7,8; 5,6)
E_2		(8,8; 4,6)	(9,4; 5,0)
⋮		⋮	⋮
E_{135}		(9,7; 2,5)	(9,2; 3,3)

Considerando a Tabela 3 pode-se construir os vértices que formam os polígonos definindo o número de dados l e aplicando a Equação (2.16). Agora, a partir dos dados forma-se a Tabela 4 onde ser visto que cada variável a descreve um indivíduo u (escola) por $Z_{\omega u} = \xi_u = (a_{u1}, b_{u1}), ..., (a_{ul}, b_{ul})$, com $l = 3$. A Figura 3 reconstrói os polígonos dos objetos E_1 e E_2, com base nas variáveis Nota$_1$ e Nota$_2$.

Tabela 4 – Tabela com dados simbólicos poligonais.

Escola	...	**Nota**$_1$	**Nota**$_2$
E_1		(7,2; 13,4), (2,2; 3,5), (13,4; 1,3)	(7,5;13,2), (13,2; 4,8), (2,2; 5,3)
E_2		(7,8; 12,6), (12,6; 4,2), (4,2; 7,8)	(10,2; 14,4), (14,4; 2,5), (4,4; 8,8)
⋮		⋮	⋮
E_{135}		(9,3; 12,1), (12,1; 5,5), (6,3; 8,8)	(8,5; 12,5), (12,5; 6,2), (5,9; 6,5)



(a) Polígonos Observados para nota 1 (b) Polígonos Observados para nota 2

Figura 3 – Polígonos obtidos para duas classes, representando a nota 1 (a) e nota 2 (b) das escolas.

O modelo linear para dados simbólicos poligonais baseados é descrito no Algoritmo 1. Considera-se apenas a relação linear entre o centro de y e o centro de x_j $(j = 1, ..., p)$ assim

como entre os raio de y e o raio de x_j (SILVA; SOUZA; CYSNEIROS, 2019a). Neste trabalho o método descrito é referido como Modelo de Regressão Linear Poligonal (PRL).

Algoritmo 1: Método PRL

1: **Entrada:** Conjunto de dados simbólicos poligonais com m observações e L vértices.
2: **Calcule** $(\hat{\beta})^T = (\hat{\beta}^c, \hat{\beta}^r)$.
3: **Calcule** $\hat{\mathbf{y}} = \hat{\beta}\mathbf{X}$
4: **Para todo** $i \leftarrow 1$ **até** *m* **faça:**
5: **Se** $\hat{y}_i^r < 0$ **então:**
6: $\hat{y}_i^r = 0$.
7: **Fim Para**
8: **Calcule** $\hat{e} = \mathbf{y} - \hat{\mathbf{y}}$.
9: **Calcule** as métricas de desempenho
10: **Construa** o polígono predito através da Equação 2.16.

Seja $(\Omega, \mathbb{A}, P)$ um espaço arbitrário de probabilidade e seja $Z = \xi = (\xi_1, \xi_2)$ uma função de valor real em Ω, define-se Z como uma distribuição uniforme no polígono P não auto-intersectável dada por

$$F_z(\xi) = \begin{cases} 0, & \xi_1 < a_1 \; ou \; \xi_2 < b_1, \\ \frac{(\xi_1 - a_1)(\xi_2 - b_1)}{A}, & \text{se } a_1 \leq \xi_1 \leq b_2 \; e \; a_1 \leq \xi_2 \leq b_2, \\ 1, & \text{caso contrario.} \end{cases} \tag{2.17}$$

Sabendo que a distribuição segue a hipótese de equidistribuição, nós definimos a mistura de distribuições uniformes poligonais dada por

$$F_z(\xi) = \begin{cases} 0, & \xi_1 < a_1 \; ou \; y < b_1, \\ \frac{1}{m} \sum_{u \in S} \frac{(\xi_1 - a_{u,1})(\xi_2 - b_{u,1})}{A_u}, & \text{se } a_{u1} \leq x \leq a_{u2} \; e \; b_{u1} \leq y < b_{u2}, \\ 1, & \text{caso contrario.} \end{cases} \tag{2.18}$$

A Função de Distribuição de Probabilidade (FDP) empírica para a mistura de m distribuições uniformes num polígono qualquer não auto-intersectável dada por

$$f_z(\xi) = \begin{cases} \frac{1}{m} \sum_{u \in S} \frac{1}{A_u}, & \text{se } \xi \in P, \\ 0, & \text{caso contrario.} \end{cases} \tag{2.19}$$

Considerando a Função de Distribuição Acumulada (FDA) definida na Equação 2.18 e que o primeiro momento estatístico coincide com o centro de gravidade, Silva, Souza e Cysneiros (2019a) propõem que a média poligonal empírica $(X_g, Y_g) = (\bar{X}, \bar{X})$ seja dada por

$$\begin{aligned}\bar{Z} = (&\frac{1}{6_m}\sum_{u\in S}\sum_{i=1}^{N}\frac{(a_{u,i}a_{u,i+1})\,(a_{u,i}b_{u,i+1} - a_{u,i+1}b_{u,i})}{A_u},\\ &\frac{1}{6_m}\sum_{u\in S}\sum_{i=1}^{N}\frac{(b_{u,i}b_{u,i+1})\,(a_{u,i}b_{u,i+1} - a_{u,i+1}b_{u,i})}{A_u}).\end{aligned} \tag{2.20}$$

Já a variância, considerando a FDA e que o segundo momento de área é igual ao segundo momento estatístico, aplica-se o modelo de mistura de densidades uniformes no polígono e deriva-se o segundo momento empírico para Z $(M_2(Z) = (M_2(\xi_1), M_2(\xi_2)))$ dado por

$$\begin{aligned}M_2(Z) = (&\frac{1}{12m}\sum_{u\in S}\sum_{i=1}^{N}\frac{\left(a_{u,i}^2 + a_{u,i}a_{u,i+1} + a_{u,i+1}^2\right)(a_{u,i}b_{u,i+1} - a_{u,i+1}b_{u,i})}{A_u},\\ &\frac{1}{12m}\sum_{u\in S}\sum_{i=1}^{N}\frac{\left(b_{u,i}^2 + b_{u,i}b_{u,i+1} + b_{u,i+1}^2\right)(a_{u,i}b_{u,i+1} - a_{u,i+1}b_{u,i})}{A_u}).\end{aligned} \tag{2.21}$$

Seja um super retângulo que contem todos os polígonos $R_0 \overset{def}{=} [\alpha_0, \alpha_r]\times[\beta_0, \beta_r]$. A frequência observada para o histograma bivariado no sub-retângulo $R_g = [\alpha_{g-1}, \alpha_g] \times [\beta_{g-1}, \beta_g], g = 1, ...r$, onde r é o número de sub-retângulos que compõem o gride do histograma é dada por

$$f_g = \sum_{u\in S}\frac{\text{area}\,(Z\,(u)\cap R_g)}{\text{area}\,(Z\,(u))}. \tag{2.22}$$

Além disso, a frequência relativa é calculada como

$$p_g = \frac{f_g}{m}, \tag{2.23}$$

onde p_g é probabilidade de um indivíduo em S está no sub-retângulo R_g. O histograma para a variável poligonal Z é a representação gráfica de $\{(R_g, f_g), g = 1, ..., r\}$. Dessa forma, para ilustrar graficamente o histograma com altura f_g sob o sub-retângulo R_g, então o volume é p_g dado pela Equação (2.24). Além destas medidas estatísticas, Silva, Souza e Cysneiros (2019a) também definem a covariância, correlação e coeficiente de variação poligonal.

$$p_g = (\alpha_g - \alpha_{g-1}) \times (\beta_g - \beta_{g-1}) \times f_g. \tag{2.24}$$

Considerando esta nova aplicação em ADS, Silva, Souza e Cysneiros (2020) investigaram a proficiência em português e matemática de estudantes brasileiros no último ano do ensino

fundamental, utilizando o modelo de regressão simbólica poligonal. Ainda, introduziram um conjunto de ferramentas para dados simbólicos poligonais no ambiente R, com a biblioteca *psda* (SILVA; SOUZA; CYSNEIROS, 2019b). Esta biblioteca implementa as medidas descritivas, o modelo de regressão e a representação gráfica da variável poligonal introduzidas por Silva, Souza e Cysneiros (2019a).

A principal vantagem de agregar os dados através desta abordagem é a quantidade de informação armazenada se comparada com o método tradicional de agregação intervalar $[min, max]$ (BILLARD; DIDAY, 2006; SILVA; SOUZA; CYSNEIROS, 2019a).

2.3 MODELOS LINEARES GENERALIZADOS

Os modelos de regressão linear baseados nos mínimos quadrados ordinários possuem suposições, como normalidade dos erros associados ao modelo, variável resposta numérica e variância constante, a qual não é verdadeira para todos os dados (MONTGOMERY; PECK; VINING, 2012). Além disso, pode-se facilmente violar as suposições quando a variável resposta é binária ou relacionada à processos de contagem.

Os Modelos Lineares Generalizados (MLG) ampliam as possibilidades de modelagem da variável resposta ao contemplar distribuições pertencentes à família exponencial, flexibilizando a relação funcional entre a variável resposta e as variáveis explicativas (PAULA, 2013). Assumindo que as respostas seguem uma distribuição pertencente à família exponencial, os MLG permitem componentes sistemáticos mais gerais para o modelo (DUNN; SMYTH, 2018).

A função densidade de uma variável aleatória Y pertencente à família exponencial pode ser expressa como:

$$f\left(y;\theta,\phi\right)=\exp\left[\phi\left\{y\theta-b\left(\theta\right)\right\}+c\left(y,\phi\right)\right]\ . \tag{2.25}$$

De acordo com Paula (2013), $E(Y)=\mu=b'(\theta)$, $Var(Y)=\theta^{-1}b''(\theta)=\phi^{-1}V$, em que $V=V(\mu)=d\mu/d\theta$ é a função de variância e $\phi^{-1}>0$ é o parâmetro de dispersão ou precisão, têm-se ainda ϕ que será o parâmetro de localização. A função de variância desempenha um papel importante na família exponencial, uma vez que a mesma caracteriza a distribuição e para algumas distribuições a variância muda conforme a sua média (PAULA, 2013).

Os componentes aleatórios e sistemáticos especificam formas para os MLG, e fazem parte dos seguintes elementos que os definem:

1. A distribuição de probabilidade da variável resposta Y_i, com $i=1,...,n$, pertencendo

à família exponencial dada pela Equação 2.25 determina o componente do aleatório do modelo. Esta distribuição pode ser sugerida pela variável resposta (como exemplo, proporções sugerem uma distribuição Binomial) ou por conhecer como a variância muda com a média.

2. O componente sistemático, sendo $g(\mu_i) = \eta_i$ sendo,

$$\eta_i = \beta_0 + \beta_i X_{i1} + \beta_i X_{i2} + ... + \beta_k X_{ik}, \tag{2.26}$$

onde η_i é o preditor linear que pode ser utilizado para fazer predições e se relacionam com a média da variável resposta μ_i.

3. Os MLG assumem uma função monótona e diferenciável que liga o preditor linear η_i à média μ_i, cuja função é adequada para relacionar os componentes aleatórios e sistemáticos do modelo, denominada função de ligação $g(.)$. A função de ligação por ser invertível, transforma a esperança da variável resposta no preditor linear, como mostra a Equação (2.27). A função de ligação inversa $g^{-1}(.)$ também é chamada de função média.

$$E(Y_i) = \mu_i = g^{-1}(\eta_i) = g^{-1}(\beta_0 + \beta_i X_{i1} + \beta_i X_{i2} + ... + \beta_k X_{ik}). \tag{2.27}$$

Os casos particulares e mais conhecidos da distribuição exponencial são os modelos contínuos os quais incluem distribuições Normal, Gama e Normal Inversa. Já os modelos discretos incluem as distribuições de Poisson e Binomial. Portanto, a família de distribuições exponencial permite que os MLG sejam ajustados a vários de tipos de dados, incluindo dados binários, proporções, contagens e dados contínuos assimétricos e positivos (DUNN; SMYTH, 2018), como mostrado na Tabela 5.

Em relação as ligações canônicas, sua utilização implica em algumas interessantes propriedades pois simplifica as estimativas de máxima verossimilhança dos parâmetros do modelo, mas também o cálculo do intervalo de confiança para a média da resposta (MYERS et al., 2002). Contudo, isto não implica em qualidade do ajuste de modelo, sendo apropriadas diferentes funções de ligação diferentes das canônicas.

Tabela 5 – Distribuições para a variável resposta Y e a natureza dos dados.

Distribuição	**Tipo de Dados**
Binomial	Proporção
Poisson	Contagem
Normal	Contínuos
Normal Inversa	Contínuos Assimétricos
Gama	Contínuos Assimétricos

A Tabela 6 apresenta as funções θ, ϕ, $b(\theta)$ e $c(y, \phi)$ específicas para cada uma destas distribuições, assim como suas respectivas ligações canônicas.

Tabela 6 – Funções da família exponencial.

Distribuição	$b(\theta)$	θ	ϕ	$V(\mu_i)$	$g(\theta)$
Binomial	$n\log(1-\mu)$	$\log\{\mu/(1-\mu)\}$	1	$\mu(1-\mu)$	$\log\{\mu/(1-\mu)\}$
Poisson	ϵ^{θ}	$\log\mu$	1	μ	$\log\mu$
Normal	$\theta^2/2$	μ	σ^{-2}	1	μ
Normal Inversa	$1/\mu$	$-1/2\mu^2$	θ^2	μ^3	$1/\mu^2$
Gama	-log$(-\theta)$	$-1/\mu$	$1/\alpha$	μ^2	$1/\mu$

Em modelos de regressão é importante verificar possíveis afastamentos de pontos observados com os pontos do modelo estimado, levando em consideração a parte aleatória e a parte sistemática do modelo. Os resíduos no contexto dos MLG são utilizados para explorar a adequação do modelo ajusta no que diz respeito a escolha da distribuição proposta para a variável resposta. A importância é verificar desvios sistemáticos, ocasionado pela escolha inadequada da função de ligação e da função de variância.

2.3.1 Modelos Lineares Generalizados na Análise de Dados Simbólicos

Assim como nos MLG clássicos (DUNN; SMYTH, 2018), o modelo BGLM estudado por Neto et al. (2009) para dados intervalar também é formado por um componente aleatório e um componente sistemático. A abordagem foi construída a partir do modelo clássico *Bivariate Generalized Linear Model* (BGLM) proposto por Iwasaki e Tsubak 2005. No componente aleatório, considera-se o vetor bivariado

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix},$$

pertencente à família exponencial bivariada. No contexto de ADS com variáveis tipo intervalar, pode-se considerar as variáveis aleatórias Y_1 e Y_2 como, por exemplo, limites inferior e superior ou no centro e range dos intervalos, respectivamente. Lima Neto et al. (2009) indicam que o componente sistemático, formado pelas variáveis explicativas responsáveis pela variabilidade de Y_1 e Y_2 , é definido por

$$\eta_1 = g_1(\mu_1) = \beta_1 \mathbf{X_1} \text{ e } \eta_2 = g_2(\mu_2) = \beta_2 \mathbf{X_2}, \tag{2.28}$$

em que $\mathbf{X_1}$ e $\mathbf{X_2}$ são são matrizes formadas por variáveis explicativas, β_1 e β_2 são os vetores de parâmetros e $g_1(\mu_1)$ e $g_2(\mu_2)$ são as funções de ligação. Os experimentos desenvolvidos por Lima Neto et al. (2009) consideram que o vetor aleatório $\mathbf{Y}$ para os limites inferior e superior segue a distribuição normal e as funções de ligação $g_1(\mu_1)$ e $g_2(\mu_2)$ são identidade ($\eta = \mu$). Os resultados foram comparados com os métodos introduzidos por Bilard e Diday (2000) e Lima Neto e De Carvalho (2008).

No contexto da representação centro-range, Lima Neto, Cordeiro e Carvalho (2011) propuseram modelos de regressão simbólica bivariada baseadas em MLG. Com este trabalho os autores ampliaram as possibilidades de lidar com a variável resposta com dados simbólicos tipo intervalar, que agora podem constituir diferente distribuições. Aplicações em dados simulados ilustraram a usabilidade da abordagem proposta.

Como parte das abordagens baseadas nos MLG voltadas à regressão logística, classificadores típicos para dados intervalares realizam predições, em geral, a partir das estimativas dos limites inferior e superior (SOUZA; QUEIROZ; CYSNEIROS, 2011; BARROS; CARVALHO; NETO, 2012). Em (SOUZA; QUEIROZ; CYSNEIROS, 2011), foi proposta quatro regras de classificação que combina as previsões derivadas desses limites da representação intervalar:

- IDPC-CSP: O classificador IDPC-CSP estima a probabilidade de um padrão pertencer à classe k utilizando como covariáveis os centros dos intervalos, calculados pela média dos limites inferior e superior. Um modelo de regressão logística multinomial é ajustado com base nesse vetor de centros, e os parâmetros são estimados por máxima verossimilhança.

- IDPC-SP: O classificador IDPC-SP estima a probabilidade de classe considerando conjuntamente os limites inferior e superior de cada intervalo como covariáveis. O vetor de

entrada possui $2p$ componentes, e um modelo logístico multinomial é ajustado para cada classe com base nesse vetor. Os parâmetros são estimados por máxima verossimilhança, utilizando K transformações logísticas no caso multiclasse.

- IDPC-VSP: O classificador IDPC-VSP estima a probabilidade de classe utilizando os vértices dos hipercubos definidos pelos limites inferior e superior dos intervalos. Cada padrão intervalar é representado por 2^p vértices, e um modelo de regressão logística multinomial é ajustado com base nesses vértices. Os parâmetros do modelo são estimados por máxima verossimilhança.

- IDPC-PP: O último classificador proposto, Modelo de Classificação Intervalar baseado em Probabilidade *a Posteriori* Combinada (IDPC-PP), estima a probabilidade de classe combinando duas regressões logísticas multinomiais ajustadas separadamente aos limites inferior e superior dos intervalos. A probabilidade a posteriori final é obtida pela média das probabilidades estimadas por cada modelo. Os parâmetros são estimados por máxima verossimilhança. Os autores demonstraram, com experimentos em bases reais e sintéticas, que o IDPC-PP apresentou menores erros de classificação em relação a outros classificadores intervalares propostos.

Com base nos estudos apresentados, observa-se que os classificadores intervalares, em especial o IDPC-PP, têm se mostrado eficazes na modelagem de dados simbólicos intervalares. Tal abordagem fundamenta a proposta deste trabalho, que visa estender esses conceitos para dados poligonais por meio de regras de classificação baseadas em regressão logística. A próxima seção apresenta em detalhes a metodologia adotada para essa extensão.

3 MODELOS LINEARES GENERALIZADOS PARA DADOS SIMBÓLICOS DO TIPO POLIGONAL

Este capítulo introduz a abordagem de Modelos Lineares Generalizados (MLG) para dados simbólicos tipo poligonal, referenciada nesta tese como Modelo Linear Generalizado Poligonal (PMLG). Conforme discutido no Capítulo 2, esse conjunto de modelos foi previamente explorado na área da Análise de Dados Simbólicos (ADS) apenas para dados simbólicos do tipo intervalar (NETO et al., 2009; NETO; CORDEIRO; CARVALHO, 2011). Para dados poligonais, os estudos anteriores se restringiram à aplicação do método dos mínimos quadrados ordinários (SILVA; SOUZA; CYSNEIROS, 2019a; SILVA; SOUZA; CYSNEIROS, 2019b).

Este capítulo descreve o modelo proposto para dados oriundos de diferentes distribuições. Define-se o conceito de resíduo poligonal, uma vez que a literatura atual ainda se apoia em estatísticas clássicas para a análise de resíduos em modelagens poligonais. Também são apresentadas métricas de desempenho baseadas nos erros de predição, as quais ampliam a aplicabilidade dos modelos lineares no contexto da ADS. No caso da distribuição Binomial, são estabelecidas regras de classificação fundamentadas na regressão logística.

3.1 MODELO LINEAR GENERALIZADO POLIGONAL

O PMLG possui um componente aleatório $\mathbf{Y}$ pertencente à família exponencial, sendo um vetor de variáveis Y^c e Y^r para valores de centros e de raios com médias μ^c e μ^r, respectivamente. O componente sistemático é definido por um preditor linear η, onde

$$\eta^c = \mathbf{x^{cT}}\beta^{\mathbf{c}} \text{ e} \tag{3.1}$$

$$\eta^r = \mathbf{x^{rT}}\beta^{\mathbf{r}}, \tag{3.2}$$

sendo $\beta^{\mathbf{c}}$ e $\beta^{\mathbf{r}}$ vetores de parâmetros com $p < m$, $\mathbf{x^{cT}} = (x^c_{j1}, ..., x^c_{jp})$ e $\mathbf{x^{rT}} = (x^r_{j1}, ..., x^r_{jp})$ sendo matrizes formadas por variáveis explicativas com $j = 1, ..., m$. Ainda, funções de ligação $g^c(\mu^c) = \eta^c$ e $g^r(\mu^r) = \eta^r$. Em MLG, a solução para o vetor de parâmetros desconhecidos $\hat{\beta} = (\beta^c, \beta^r) = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X^T W X})^{-1}\mathbf{X^T W}^{1/2}$, sendo $\mathbf{W}$ o elemento de reponderação da matriz.

A predição de um novo exemplo é calculada a partir dos preditores lineares para o centro e o raio, dados por $\hat{\eta}^c = \mathbf{X}^c\hat{\boldsymbol{\beta}}^c$ e $\hat{\eta}^r = \mathbf{X}^r\hat{\boldsymbol{\beta}}^r$, respectivamente, e as médias correspondentes são

obtidas pela função inversa da ligação: $\hat{\mu}^c = (g^c)^{-1}(\hat{\eta}^c)$ e $\hat{\mu}^r = (g^r)^{-1}(\hat{\eta}^r)$. O Algoritmo 2 descreve os passos do PMLG.

Algoritmo 2: PMLG: Modelos Lineares Generalizados para Dados Poligonais

1: **Entrada:** Conjunto de dados simbólicos poligonais com m observações e L vértices.
2: **Saída:** Preditores de centros e raios $\hat{\eta}_i^c$ e $\hat{\eta}_i^r$, respectivamente.
3: **Início**
4: **Defina** as funções de ligação $g^c(\cdot)$ e $g^r(\cdot)$ conforme a distribuição da variável resposta.
5: **Estime** os parâmetros dos modelos para centro e raio:
6: $\hat{\beta}^c \leftarrow$ estimação via máxima verossimilhança para o centro;
7: $\hat{\beta}^r \leftarrow$ estimação via máxima verossimilhança para o raio;
8: **Para** $i = 1$ até m **faça**:
9: **Calcule** os preditores lineares: $\hat{\eta}_i^c = \mathbf{x}_i^{cT}\hat{\beta}^c$ e $\hat{\eta}_i^r = \mathbf{x}_i^{rT}\hat{\beta}^r$;
10: **Aplique** as funções de ligação inversas:
11: $\hat{\mu}_i^c = (g^c)^{-1}(\hat{\eta}_i^c), \quad \hat{\mu}_i^r = (g^r)^{-1}(\hat{\eta}_i^r)$;
12: **Fim Para**
13: **Aplique** a métrica de avaliação do modelo.
14: **Fim**

Dependendo do tipo da variável resposta, diferentes distribuições pertencentes à família exponencial podem ser consideradas no escopo do PMLG. Dessa forma, o PMLG generaliza a abordagem tradicional dos MLG ao permitir a modelagem de variáveis simbólicas poligonais em diferentes contextos:

- Distribuições contínuas: modelagem de medidas reais de centro e raio, adequada para problemas de regressão com distribuições contínuas. O polígono $\hat{P}_i$ $(i = 1, ..., m)$ com L vértices é obtido a partir da Equação (3.3):

$$\hat{P}_{il} = \left(\hat{\mu}_i^c + \hat{\mu}_i^r cos(\frac{2\pi r}{l}),\ \hat{\mu}_i^c + \hat{\mu}_i^r sin(\frac{2\pi r}{l})\right), \text{ onde } l = 1, ..., L. \tag{3.3}$$

  Desta forma, obtêm-se os pares de vértices preditos da i-ésima observação que reconstrói o polígono. Se $\hat{\mu}_i^r < 0$, então $\hat{\mu}_j^r = 0$, configurando um polígono degenerado.

- Distribuições discretas (Binomial): modelagem de dados categóricos, adequada para problemas de classificação, em que se busca prever a categoria ou rótulo a que pertence

cada observação. Podem ser aplicadas regras de classificação para uma resposta discreta com base em $\hat{\mu}_i^c$ e $\hat{\mu}_i^r$ da i-ésima observação.

3.2 PMLG PARA DISTRIBUIÇÕES CONTÍNUAS

Nesta seção, apresenta-se a formalização do PMLG para distribuições contínuas, com foco na definição de resíduos e nas métricas utilizadas para avaliação preditiva. Introduz-se um resíduo poligonal baseado na diferença entre os vértices dos polígonos observados e preditos. Também são discutidas diferentes métricas de desempenho que permitem comparar a qualidade preditiva de modelos e estratégias de regressão aplicados a dados poligonais.

3.2.1 Resíduo Poligonal baseado nos Vértices

Em regressão linear clássica, um resíduo é definido como a diferença entre o valor observado e o valor predito baseado na equação de regressão. A análise de resíduos é um passo essencial para identificar os efeitos de desvios de suposições de um modelo de regressão. Uma análise residual comum para dados simbólicos com valor de intervalo é construída a partir da representação centro e intervalo, ou seja, é baseada na análise de resíduos para dados clássicos.

Em Nascimento et al. (2022), são investigadas as premissas do modelo de regressão linear a partir dos resíduos intervalares. Esta abordagem considera os centros e os intervalos dos resíduos resultando em uma medida única. Além disso, um estudo foi realizado a partir da definição de resíduo padronizado. Na abordagem poligonal introduzida por Silva, Souza e Cysneiros (2019a) a diferença entre observados e preditos é calculada a partir das áreas poligonais. No entanto, ainda não há um estudo detalhado acerca dos resíduos.

Assim, este trabalho propõe o resíduo poligonal ordinário, ou seja, a forma poligonal da diferença entre os polígonos observados e preditos, tendo como base a Equação 2.16. Seja Ω um espaço de polígonos e Z uma variável aleatória $Z : \Omega \to \mathbb{R}^2$ que assume valores no polígono P com L vértices. Então $Z = \xi = \{(a_1, b_1), ..., (a_L, b_L)\} \subset \mathbb{R}^2$. Ele pode ser reescrito como $Z = \xi = (\xi_1, \xi_2)$, onde $\xi_1 = \{a_1, ..., a_L\}$ e $\xi_2 = \{b_1, ..., b_L\}$. Define-se a diferença de resíduos pela Equação (3.4, onde c é o centro observado e $\hat{c}$ é o predito, r é o raio observado e $\hat{r}$ é o predito. Essa métrica quantifica a discrepância entre o polígono observado e o predito em termos das componentes estruturais que o definem, centro e raio. Valores maiores indicam maior divergência entre o polígono gerado pelo modelo e aquele observado nos dados.

$$
\begin{aligned}
\Delta_i &= \left[\left(a_{il} - \hat{a}_{il}\right), \left(b_{il} - \hat{b}_{il}\right)\right] \\
&= \left[\left(c_i + r_i \cos\left(\frac{2\pi l}{L}\right)\right) - \left(\hat{c}_i + \hat{r}_i \cos\left(\frac{2\pi l}{L}\right)\right), \left(c_i + r_i \sin\left(\frac{2\pi l}{L}\right)\right) - \left(\hat{c}_i + \hat{r}_i \sin\left(\frac{2\pi l}{L}\right)\right)\right] \\
&= \left[\left((c_i - \hat{c}_i) + r_i \cos\left(\frac{2\pi l}{L}\right)\right) - \left(\hat{c}_i + \hat{r}_i \cos\left(\frac{2\pi l}{L}\right)\right), \left(c_i + r_i \sin\left(\frac{2\pi l}{L}\right)\right) - \left(\hat{c}_i + \hat{r}_i \sin\left(\frac{2\pi l}{L}\right)\right)\right] \qquad (3.4)
\end{aligned}
$$

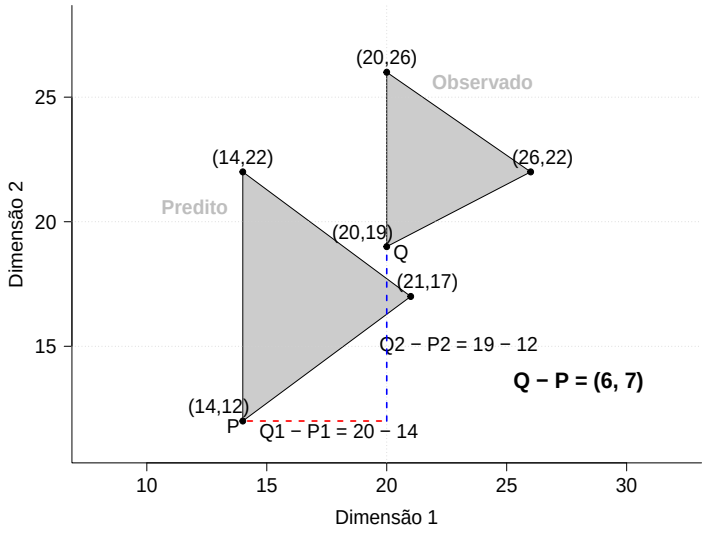
3.2.2 Métricas de Desempenho

A avaliação da qualidade preditiva de modelos é uma etapa essencial em qualquer abordagem estatística ou de aprendizado de máquina, especialmente em contextos que envolvem representações simbólicas, como os dados poligonais. A escolha das métricas de avaliação influencia diretamente a interpretação dos resultados, podendo ressaltar ou ocultar características relevantes do modelo. Nesse contexto, o trabalho de Silva, Souza e Cysneiros (2019a) desenvolveu um método de avaliação de performance de modelos denominado Erro Médio Quadrático da Área (EMQA), o qual é dado por:

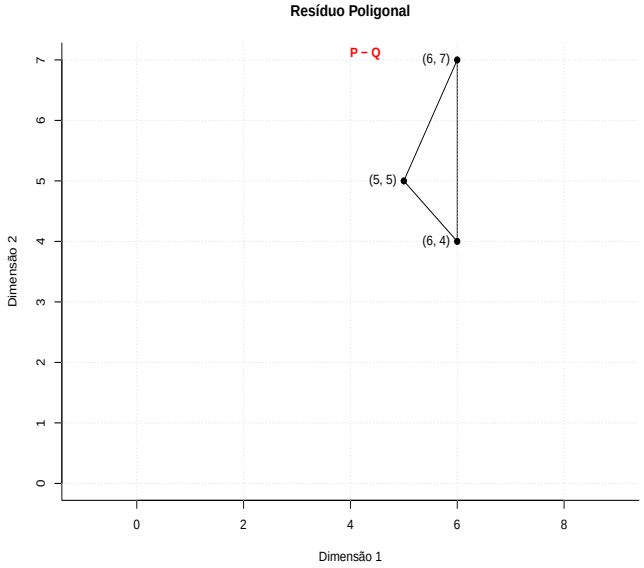
$$
EQMA = \sqrt{\frac{1}{r}\sum_{u=1}^{r}\left[(area(P_u) - area\hat{P}_u)\right]^2}, \qquad (3.5)
$$

onde P_u é o polígono observado e $\hat{P}_u$ o polígono predito. Observa-se que essa medida considera apenas a área dos polígonos, desconsiderando a posição que eles ocupam no espaço $\mathbb{R}^2$. A implicação disso é que os polígonos podem ter valores de raio semelhantes, resultando em áreas próximas ou equivalentes, mas apresentar grande dispersão em relação aos centros e, consequentemente, em suas posições, tornando essa medida de qualidade incompleta.

A Figura 4 ilustra essa discussão. Em (a), apresenta-se a representação de dois polígonos que possuem formas semelhantes e áreas próximas. Com base na diferença de áreas, o resíduo calculado é pequeno. Contudo, ao considerar a posição dos polígonos, observa-se que a diferença entre seus centros é mais significativa. Já em (b), está apresentada a proposta de resíduo, que representa a diferença entre os vértices dos polígonos.



(a) Polígono Observado e Predito



(b) Polígono Residual

Figura 4 – Representação da diferença entre vértices dos Polígonos.

Dito isto, este trabalho propõe uma medida de performance baseada nos vértices dos polígonos. Com base no resíduo poligonal proposto na Equação 3.4, define-se o Erro Médio Quadrático da Distância dos Vértices (EMQDV).

$$EMQDV = \sqrt{\frac{1}{2r}\sum_{u=1}^{r}\left[(a_{ul} - \hat{a}_{ul}) + (b_{ul} - \hat{b}_{ul})\right]^2}, \tag{3.6}$$

onde a e $\hat{a}$ constituem os valores do eixo das abcissas,b e $\hat{b}$ os valores do eixo das coordenadas e $l = 1, ..., L$. A distância Euclideana é considerada, na qual obtém-se o somatório das diferenças de cada par de vértice.

Neste trabalho, ainda se considera a avaliação dos erros de predição sob dois métodos de desempenho. O primeiro definido como Erro Médio Quadrático da Área e Centro Conjuntamente (EMQAC), é uma adaptação do modelo proposto por Silva, Souza e Cysneiros (2019a) o qual é acrescido do valor de centro, ou posição como definido na Equação 3.7.

$$EMQAC = \sqrt{\frac{1}{2r}\sum_{u=1}^{r}\left[(area(P_u) - area\hat{P}_u) + (centro_u - c\hat{entro}_u)\right]^2}. \tag{3.7}$$

O segundo é baseado apenas em valores de centro e raio, definido na Equação 3.8 e referenciado por Erro Médio Quadrático do Centro e Raio Conjuntamente (EMQCR). Portanto, avaliam-se os modelos de regressão e o resultado das quatro medidas métricas do erro definidas. Os cenários de avaliação consideram bases de dados simuladas e reais.

$$EMQCR = \sqrt{\frac{1}{2r}\sum_{u=1}^{r}\left[(centro_u - c\hat{entro}_u) + (raio_u - r\hat{aio}_u)\right]^2}. \tag{3.8}$$

3.3 PMLG PARA DISTRIBUIÇÕES DISCRETAS

A modelagem de variáveis categóricas é uma etapa central em diversos problemas de classificação. Nesse contexto, distribuições da família exponencial, como a Binomial, fornecem uma base probabilística para a construção de modelos preditivos. Quando a variável resposta segue uma distribuição Binomial, uma abordagem comum é utilizar a regressão logística. Por exemplo, um MLG com função de ligação *logit* para uma variável Binomial é dado por:

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \mathbf{X}\boldsymbol{\beta},$$

onde p é a probabilidade de sucesso, $\mathbf{X}$ são as variáveis explicativas, e $\boldsymbol{\beta}$ são os coeficientes do modelo. Esse modelo é amplamente utilizado para prever probabilidades de eventos binários. A função inversa do *logit* transforma o preditor linear $\mathbf{X}\boldsymbol{\beta}$ no intervalo $(0, 1)$, sendo expressa por:

$$p = \frac{e^{X\beta}}{1 + e^{X\beta}}. \tag{3.9}$$

No contexto do PMLG, essa abordagem é estendida para representar uma variável resposta simbólica poligonal, modelada por meio de dois componentes aleatórios Y^c e Y^r, que representam, respectivamente, o centro e o raio do polígono. Ambos os componentes são considerados pertencentes à família exponencial. Sejam $p^c = P(Y = 1 \mid x^c)$ e $p^r = P(Y = 1 \mid x^r)$ as

probabilidades de sucesso associadas às componentes de centro e raio das variáveis preditoras, respectivamente. Os modelos logísticos são definidos por:

$$\eta^c = \text{logit}(p^c) = \ln\left(\frac{p^c}{1-p^c}\right) = \mathbf{X}^c\boldsymbol{\beta}^c,$$

$$\eta^r = \text{logit}(p^r) = \ln\left(\frac{p^r}{1-p^r}\right) = \mathbf{X}^r\boldsymbol{\beta}^r,$$

onde $\mathbf{X}^c$ e $\mathbf{X}^r$ são as variáveis explicativas para o centro e o raio, $\boldsymbol{\beta}^c$ e $\boldsymbol{\beta}^r$ os são os respectivos coeficientes, e $p^c = \frac{1}{1+\exp(\eta^c)}$ e $p^r = \frac{1}{1+\exp(\eta^r)}$.

Dessa forma, nesta seção apresentam-se três regras de classificação baseadas no PMLG com distribuição Binomial: a primeira é baseada na média aritmética das predições; a segunda utiliza uma média ponderada otimizada por meio de um algoritmo de otimização; e a terceira baseia-se em uma representação por protótipos e probabilidades. Essas regras são aplicadas considerando que a predição de um novo exemplo é calculada a partir das predições do centro e do raio, dadas por $\hat{\eta}_i^c = \left(X_i^c\hat{\beta}_p^c\right)$ e $\hat{\eta}_i^r = \left(X_i^r\hat{\beta}_p^r\right)$ resultando nas probabilidades *a posteriori* $\hat{p}_i^c = \hat{\mu}_i^c = g^{c^{-1}}\left(\hat{\eta}_p^c\right)$ e $\hat{p}_i^r = \hat{\mu}_i^r = g^{r^{-1}}\left(\hat{\eta}_p^r\right)$. Além disso, essa abordagem pode ser estendida para problemas com mais de duas classes por meio da técnica "um contra todos" (*one-vs-all*).

3.3.1 Regra de Classificação Baseada na Média Aritmética das Predições

Nesta regra a ideia é assumir que os dados de centro e raio das variáveis preditoras possuem o mesmo peso na obtenção da probabilidade *a posteriori* associada a $\mathbf{x}$. Seja $P(Y = 1 \mid \mathbf{x})$ a probabilidade *a posteriori* associada a $\mathbf{x}$. O modelo logístico para dados poligonais combina as predições para o centro e o raio tomando a média de suas probabilidades *a posteriori*. A probabilidade *a posteriori* combinada para $Y = 1$ é dada por:

$$\hat{P}(Y = 1 \mid \mathbf{x}) = \frac{\hat{p}^c + \hat{p}^r}{2},$$

onde $\hat{p}^c$ e $\hat{p}^r$ representam as probabilidades a posteriori estimadas a partir dos modelos logísticos ajustados para o centro e o raio, respectivamente. A decisão final de classificação é obtida comparando $\hat{P}(Y = 1 \mid \mathbf{x})$ com um limiar τ, fixado em 0,5. Assim, a classe predita $\hat{y}$ é definida por:

$$\hat{y} = \begin{cases} 1, & \text{se } \hat{P}(Y = 1 \mid \mathbf{x}) \geq \tau, \\ 0, & \text{caso contrário.} \end{cases}$$

Essa estratégia de combinação é comumente utilizada em problemas de classificação que envolvem dados simbólicos ou intervalares, nos quais modelos separados são construídos para diferentes componentes do intervalo, tipicamente os limites inferior e superior (SOUZA; QUEIROZ; CYSNEIROS, 2011). O Algoritmo 3 apresenta os passos da regra de classificação.

Algoritmo 3: Regra de classificação para o PMLG baseado na média aritmética

Entrada: Conjunto de dados poligonais com m observações e L vértices.
Saída: Probabilidade *a posteriori* $\hat{P}(Y = 1 \mid \mathbf{x})$.
Início
    Defina as funções de ligação $g^c(\cdot)$ e $g^r(\cdot)$ como *logit*.
    Estime os parâmetros dos modelos para centro e raio:
        $\hat{\beta}^c \leftarrow$ estimação via máxima verossimilhança para o centro;
        $\hat{\beta}^r \leftarrow$ estimação via máxima verossimilhança para o raio;
    Para $i = 1$ até m **faça**:
        Calcule os preditores lineares: $\hat{\eta}_i^c = \mathbf{x}_i^{cT}\hat{\beta}^c$ e $\hat{\eta}_i^r = \mathbf{x}_i^{rT}\hat{\beta}^r$;
        Aplique as funções de ligação inversas:
            $\hat{p}_i^c = \hat{\mu}_i^c = (g^c)^{-1}(\eta_i^c)$ e $\hat{p}_i^r = \hat{\mu}_i^r = (g^r)^{-1}(\eta_i^r)$;
    Fim Para
    Para cada nova observação $\mathbf{x}$ **faça:**
        Compute $\hat{P}(Y = 1 \mid \mathbf{x}) = \frac{\hat{p}^c + \hat{p}^r}{2}$.
    Fim Para
Fim

3.3.2 Regra de Classificação Baseada na Média Otimizada das Predições

O processo de ponderação na representação de dados em ADS tem sido abordado em alguns trabalhos. Em (ARAÚJO et al., 2017), o objetivo foi ajustar a influência relativa dos limites inferior e superior dos intervalos na medida de distância utilizada para a classificação. Para isso, foi considerado um parâmetro de controle $\tau \in [0, 1]$, avaliado por meio da variação de seus valores a fim de analisar o impacto dos limites na performance do classificador. Diferentemente

da média aritmética simples, que atribui o mesmo peso fixo aos extremos do intervalo (ou seja, trata os limites inferior e superior como igualmente relevantes), o processo de ponderação possibilita controlar a influência de cada limite conforme as características do problema.

Por sua vez, a regra de classificação baseada na média ponderada das predições proposta utiliza o algoritmo *Particle Swarm Optmization* (PSO) (KENNEDY; EBERHART, 1995), no português Otimização por Enxame de Partículas, uma metaheurística populacional inspirada na inteligência coletiva de enxames, para determinar o valor ótimo do parâmetro λ, que atua como um fator de ponderação responsável por equilibrar as contribuições das predições associadas ao centro e ao raio. O PSO tem recebido grande atenção na comunidade científica devido ao seu desempenho em resolver problemas complexos de otimização sem a necessidade de suposições sobre a função objetivo (GAD, 2022). Um enxame de partículas atualiza suas posições de uma iteração para a próxima, permitindo que o algoritmo PSO realize efetivamente o processo de busca. Para encontrar a solução ótima, cada partícula se move em direção à sua melhor posição anterior e à melhor posição global identificada dentro do enxame (KENNEDY; EBERHART, 1995; GAD, 2022).

Neste contexto, a otimização de λ é realizada para maximizar a acurácia da classificação, definida como:

$$\text{Acurácia} = \frac{\sum_{i=1}^{n} I(y_i = \hat{y}_i)}{n},$$

onde $I(y_i = \hat{y}_i)$ é a função indicadora que vale 1 se y_i for igual à $\hat{y}_i$, e 0 caso contrário. Após a otimização, o valor de λ é usado na regra final de classificação. A probabilidade *a posteriori* combinada para a classe é dada por:

$$\hat{P}(Y = 1 \mid \mathbf{x}) = \lambda \times \hat{p}^c + (1 - \lambda) \times \hat{p}^r,$$

onde $\hat{p}^c$ e $\hat{p}^r$ são as probabilidades a posteriori estimadas pelos modelos logísticos para o centro e o raio, respectivamente, e λ é um peso obtido por meio de um processo de otimização que equilibra as contribuições dessas duas representações. Esse elemento permite ajustar dinamicamente a influência relativa de centro e raio no modelo final. A otimização de λ visa maximizar a acurácia, tornando o modelo mais flexível e mais robusto, principalmente em casos onde uma das representações pode ser mais informativa que a outra. O Algoritmo 4 apresenta os passos da regra de classificação.

Algoritmo 4: Regra de classificação do PMLG baseada em Otimização da Média

Entrada: Conjunto de dados poligonais com m observações e L vértices.
Saída: Probabilidade *a posteriori* $\hat{P}(Y = 1 \mid \mathbf{x})$.
Início
    Defina as funções de ligação $g^c(\cdot)$ e $g^r(\cdot)$ como *logit*.
    Estime os parâmetros dos modelos para centro e raio:
        $\hat{\beta}^c \leftarrow$ estimação via máxima verossimilhança para o centro;
        $\hat{\beta}^r \leftarrow$ estimação via máxima verossimilhança para o raio;
    Para $i = 1$ até m **faça**:
        Calcule os preditores lineares: $\hat{\eta}_i^c = \mathbf{x}_i^{cT}\hat{\beta}^c$ e $\hat{\eta}_i^r = \mathbf{x}_i^{rT}\hat{\beta}^r$;
        Aplique as funções de ligação inversas:
            $\hat{p}_i^c = \hat{\mu}_i^c = (g^c)^{-1}(\hat{\eta}_i^c)$ e $\hat{p}_i^r = \hat{\mu}_i^r = (g^r)^{-1}(\hat{\eta}_i^r)$;
    Fim Para
    Aplique PSO para encontrar o valor ótimo λ^* que maximiza a acurácia;
    Para cada nova observação $\mathbf{x}$ **faça:**
        Calcule $\hat{P}(Y = 1 \mid \mathbf{x}) = \lambda \times \hat{p}^c + (1 - \lambda) \times \hat{p}^r$.
    Fim Para
Fim

Nesta tese, a otimização do parâmetro λ foi realizada utilizando o algoritmo PSO, implementado na função `psoptim` do pacote `pso` em R. Com o objetivo de garantir a reprodutibilidade, os principais parâmetros do algoritmo foram explicitamente documentados na Tabela 7. Apenas o número máximo de iterações (`maxit`) foi modificado, sendo aumentado do valor padrão de 100 para 500, a fim de assegurar a convergência.

Tabela 7 – Configurações dos parâmetros do algoritmo PSO.

Parâmetro	**Valor padrão**	**Valor utilizado**
Tamanho do enxame (s)	40	40
Número máximo de iterações ($maxit$)	100	**500**
Número máximo de avaliações ($maxf$)	∞	∞
Peso de inércia (w)	0.721	0.721
Coeficiente cognitivo ($c.p$)	1.193	1.193
Coeficiente social ($c.g$)	1.193	1.193
Tolerância absoluta ($abstol$)	$-\infty$	$-\infty$
Nível de rastreamento ($trace$)	0	0
Frequência de relatório ($report$)	10	10

3.3.3 Regra de Classificação Baseada em Protótipos Poligonais

Em aprendizagem de máquina, um protótipo refere-se a uma representação simplificada das características centrais ou típicas de uma classe ou grupo de dados. Ele serve como referência para a comparação e classificação de novas observações, geralmente com base em uma medida de distância. Com base nessa abordagem, propõe-se uma regra de classificação baseada em protótipos para dados poligonais, na qual uma nova observação é atribuída à classe do protótipo poligonal mais próximo.

O procedimento de construção dos protótipos é descrito a seguir. Dado um conjunto de dados rotulado com duas classes distintas, $k \in \{0, 1\}$, define-se um total de z protótipos para cada classe com base nos valores ajustados de centro e raio observados no conjunto de treinamento. Esses protótipos são determinados a partir dos quantis empíricos dos valores ordenados de centro e raio da respectiva classe.

A partir dos z pares representativos de centro e raio, é possível reconstruir os polígonos correspondentes utilizando a Equação 2.16. Especificamente, para o j-ésimo protótipo da classe k, os valores de centro e raio são definidos como:

$$\mathbf{C}_j^{(k)} = q_{\frac{j}{z}}(\mathbf{C}^{(k)}), \quad \mathbf{R}_j^{(k)} = q_{\frac{j}{z}}(\mathbf{R}^{(k)}), \tag{3.10}$$

onde $q_{\frac{j}{z}}(\cdot)$ denota o quantil $\frac{j}{z}$-ésimo dos valores ajustados na classe k, com $j = 1, 2, \ldots, z$ e $k \in \{0, 1\}$. O Algoritmo 5 descreve as etapas desse processo.

Algoritmo 5: Construção de Protótipos Poligonais por Classe

Entrada: Conjunto de valores ajustados de centro C_i, raio R_i e rótulo $y_i \in \{0, 1\}$.
Número de protótipos por classe: z.

Saída: Protótipos $(\mathbf{C}_j^{(k)}, \mathbf{R}_j^{(k)})$ para cada classe $k \in \{0, 1\}$, $j = 1, \ldots, z$.

1: **Início**
2: **Para** cada classe $k \in \{0, 1\}$ **faça**:
3: Ordenar os vetores $\mathbf{C}^{(k)}$ e $\mathbf{R}^{(k)}$.
4: **Para** cada $j = 1$ até z **faça**:
5: Calcular o quantil $\frac{j}{z}$ para o centro: $\mathbf{C}_j^{(k)} = q_{\frac{j}{z}}(\mathbf{C}^{(k)})$.
6: Calcular o quantil $\frac{j}{z}$ para o raio: $\mathbf{R}_j^{(k)} = q_{\frac{j}{z}}(\mathbf{R}^{(k)})$.
7: Reconstruir o polígono correspondente utilizando a Equação 2.16.
8: **Fim Para**
9: **Fim Para**
10: **Retorne** os z protótipos e seus respectivos polígonos para cada classe k.
11: **Fim**

A regra de classificação é baseada na proximidade entre uma nova observação predita e os protótipos das classes 0 e 1, considerando a distância euclidiana entre os centros e os raios.

Seja $\mathbf{x} = (x_C, x_R)$ a nova observação, onde $\hat{p}^c$ e $\hat{p}^r$ são as probabilidades a posteriori estimadas pelos modelos logísticos para o centro e o raio, respectivamente. Esta entrada será atribuída à classe $k \in \{0, 1\}$ cujo protótipo for mais próximo, segundo a seguinte regra:

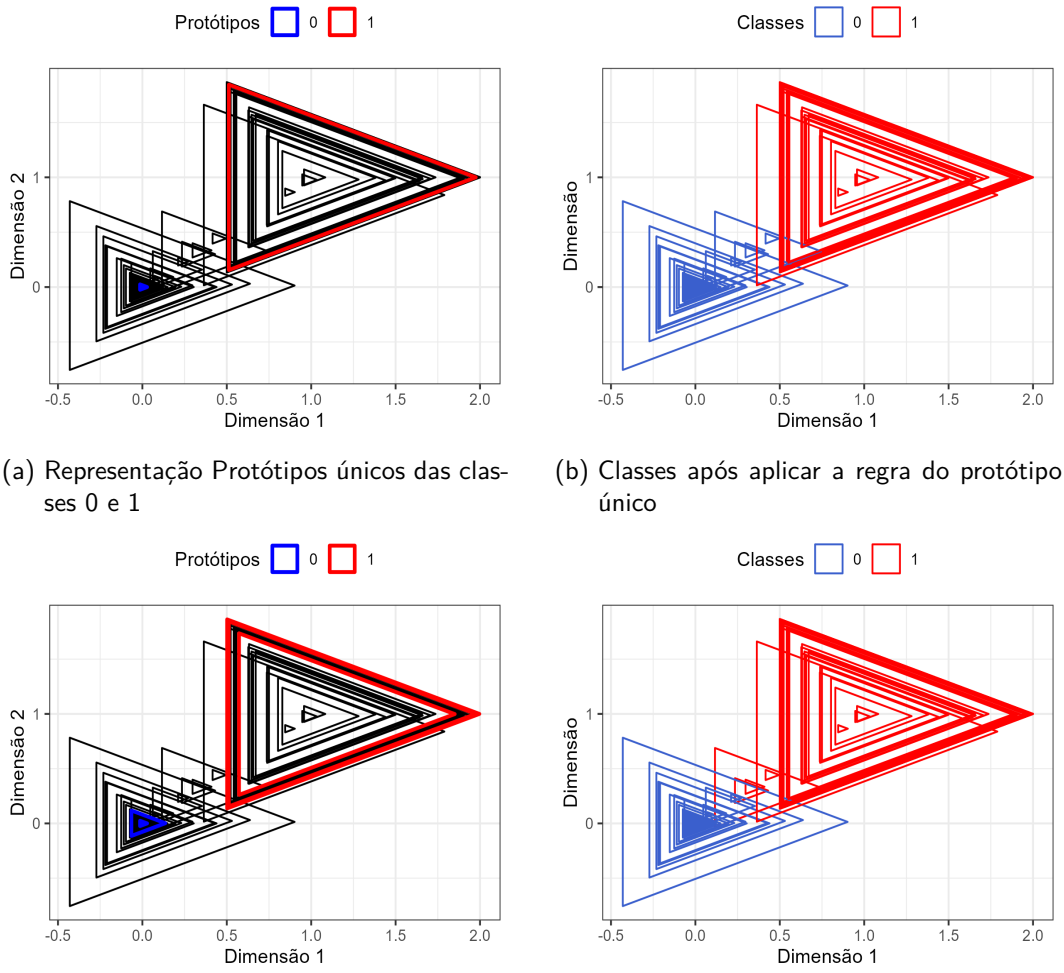
$$k = \arg \min_{k \in \{0,1\},\, j \in \{1,\ldots,z\}} \left(\sqrt{(\hat{p}^c - \mathbf{C}_j^{(k)})^2 + (\hat{p}^r - \mathbf{R}_j^{(k)})^2} \right). \tag{3.11}$$

O Algoritmo 6 apresenta os passos da regra de classificação baseada em protótipos. Essa abordagem considera a proximidade de uma nova observação predita aos protótipos, os quais são representados por polígonos específicos de cada classe. Considerando as classes $k \in \{0, 1\}$, espera-se que os protótipos polígonais associados à classe 0 possuam centros e raios próximos de 0, enquanto aqueles da classe 1 apresentem centros e raios próximos de 1.

Algoritmo 6: Regra de classificação do PMLG baseado em Protótipos Poligonais

Entrada: Conjunto de dados poligonais com m observações e L vértices.
Saída: Probabilidade *a posteriori* para a classe k.
Início
    Defina as funções de ligação $g^c(\cdot)$ e $g^r(\cdot)$ como *logit*.
    Estime os parâmetros dos modelos para centro e raio:
        $\hat{\beta}^c \leftarrow$ estimação via máxima verossimilhança para o centro;
        $\hat{\beta}^r \leftarrow$ estimação via máxima verossimilhança para o raio;
    Para $i = 1$ até m **faça**:
        Calcule os preditores lineares: $\hat{\eta}_i^c = \mathbf{x}_i^{cT}\hat{\beta}^c$ e $\hat{\eta}_i^r = \mathbf{x}_i^{rT}\hat{\beta}^r$;
        Aplique as funções inversas da ligação:
            $\hat{p}_i^c = \hat{\mu}_i^c = (g^c)^{-1}(\hat{\eta}_i^c)$ e $\hat{p}_i^r = \hat{\eta}_i^r = (g^r)^{-1}(\hat{\eta}_i^r)$;
    Fim Para
    Para cada classe k **faça:**
        Calcule z protótipos $(\mathbf{C}_j^{(k)}, \mathbf{R}_j^{(k)})$ usando os quantis de $\hat{p}^c$ e $\hat{p}^r$ na classe k, para $j = 1, \ldots, z$;
    Fim Para
    Para cada observação $\mathbf{x} = (x_C, x_R)$ **faça:**
        Atribua à classe
            $$k = \arg\min_{k\in\{0,1\},\, j\in\{1,\ldots,z\}} \left(\sqrt{(\hat{p}^c - \mathbf{C}_j^{(k)})^2 + (\hat{p}^r - \mathbf{R}_j^{(k)})^2} \right).$$
    Fim Para
Fim

Para representar esta última regra de classificação, a Figura 5 mostra dois exemplos da atuação da regra, com um protótipo e três protótipos, para um mesmo cenário de dados. Pode ser observado que aumentar o número de protótipos ocasiona mudança na classificação de algumas entrada de dados.



(a) Representação Protótipos únicos das classes 0 e 1

(b) Classes após aplicar a regra do protótipo único

(c) Representação de três protótipos das classes 0 e 1

(d) Classes após aplicar a regra dos três protótipos

Figura 5 – Representação da classificação usando a regra baseada em protótipos.

3.4 ABORDAGEM DE MODELAGEM POLIGONAL

Este capítulo apresentou uma estrutura unificada para modelagem estatística de dados poligonais, abordando tanto variáveis respostas contínuas quanto discretas. Partindo de uma base de dados poligonal representada por centro e raio, o capítulo desenvolve abordagens distintas conforme a natureza da variável de interesse.

Para variáveis resposta com distribuição contínua e assimétrica (como distribuições Gama e Normal Inversa), é proposto o uso do Modelo Linear Generalizado Poligonal (PMLG) com a análise de resíduos poligonais ordinários. A avaliação dos modelos é realizada através de quatro métricas de erro médio quadrático. Para variáveis discretas com distribuição Binomial, o mesmo framework PMLG é adaptado com três regras de predição: média aritmética das predições, média otimizada via Otimização por Enxame de Partículas e método baseado em

protótipos poligonais. O desempenho preditivo é avaliado através de acurácia e precisão.

O fluxo metodológico resultante é ilustrado na Figura 6. A figura fornece um guia para análise preditiva de dados poligonais, estabelecendo padrões de avaliação para diferentes contextos estatísticos.

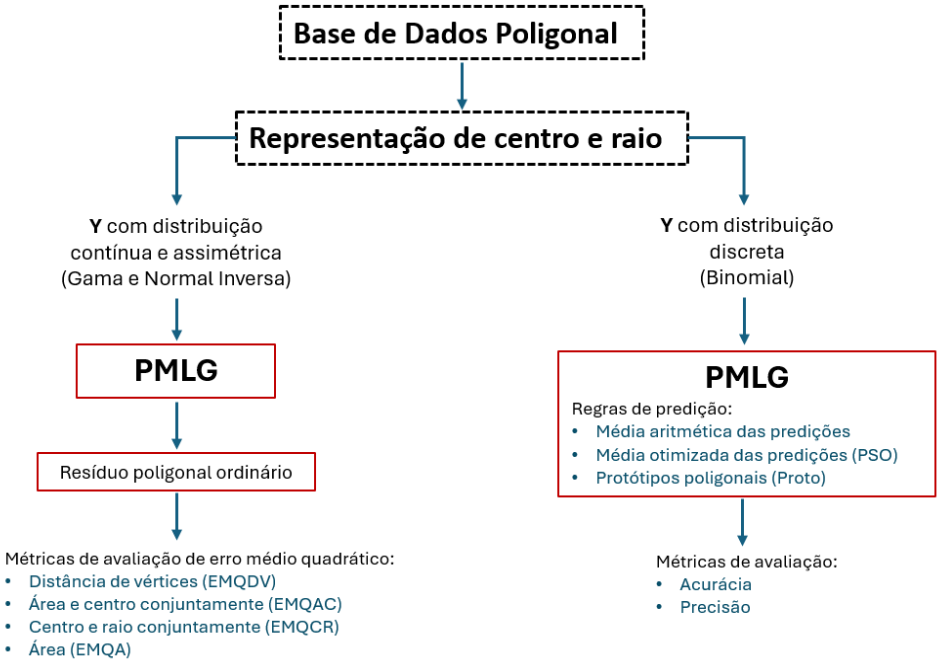


Figura 6 – Fluxo metodológico para modelagem poligonal com diferentes distribuições de Y, mostrando técnicas de predição e métricas de avaliação propostas.

4 AVALIAÇÃO EXPERIMENTAL COM DADOS POLIGONAIS GERADOS A PARTIR DE DISTRIBUIÇÕES CONTÍNUAS ASSIMÉTRICAS

A análise experimental inicia-se com a geração dos conjuntos de dados poligonais e técnicas de visualização para estes conjuntos. A metodologia segue uma sequência de algoritmos organizados em simulações Monte Carlo (MC) para avaliação da proposta em dois contextos: Diagnóstico e Preditivo. A abordagem proposta neste trabalho, denominada Modelo Linear Generalizado Poligonal (PMLG), é comparada aos métodos introduzidos em Silva, Souza e Cysneiros (2019a) e Neto, Cordeiro e Carvalho (2011), referenciados como Modelo de Regressão Linear Poligonal (PRL) e Modelo de Regressão Linear Bivariado (PBIVAR), respectivamente. O modelo PBIVAR foi adaptado para operar com os dados em termos de centro e raio, possibilitando a comparação com os modelos baseados em representações poligonais. Os experimentos desta pesquisa foram realizados na linguagem R (R Core Team, 2020).

4.1 CONFIGURAÇÕES DOS DADOS SIMULADOS

4.1.1 Cenário 1: Distribuição Gama

O primeiro cenário de dados simulados considera a distribuição Gama. O Algoritmo 7 descreve o processo de geração de dados, considerando a função de ligação canônica da distribuição Gama, a recíproca. A representação dos dados de centro e raio é mostrado na Figura 7 para uma amostra de dados. Os histogramas da variável resposta revelam assimetria à direita, e os gráficos de dispersão da variável resposta em função da variável explicativa revela que a variância não é constante.

Algoritmo 7: Geração de conjuntos simulados com distribuição Gama

1: **Requerer** $n = 100$.
2: **Defina** a função de ligação recíproca para a distribuição Gama.
3: **Defina** x_i^c obtido de uma distribuição $U(a = 0; b = 10)$.
4: **Defina** x_i^r obtido de uma distribuição $U(a = 0; b = 5)$.
5: **Calcule** $y_i^c = 1,0 + 0,5x_i^c$ obtido de uma distribuição $G(\mu = 0{,}3;\ \phi = 3)$.
6: **Calcule** $y_i^r = 1,0 + 0,5x_i^r$ obtido de uma distribuição $G(\mu = 0{,}5;\ \phi = 8)$.
7: **Compute** os vértices dos polígonos com a Equação (2.16).

A Figura 8 descreve a variável resposta poligonal (centro, raio) dos dados gerados, aplicando a Equação (2.16) implementada na biblioteca *psda* (SILVA; SOUZA; CYSNEIROS, 2020). Para o cenário de dados com distribuição Gama, gera-se duas bases dados poligonais, com 5 e 10 vértices. Percebem-se os centros dos polígonos destacados em azul, os quais possuem maior concentração no início da distribuição, revelando maior proximidade entre os polígonos. Em relação aos raios, há uma variação quanto à área formada pelos polígonos, no entanto, este aspecto foi controlado na geração dos valores de raio oriundos da distribuição Gama para que não houvesse valores extremos.

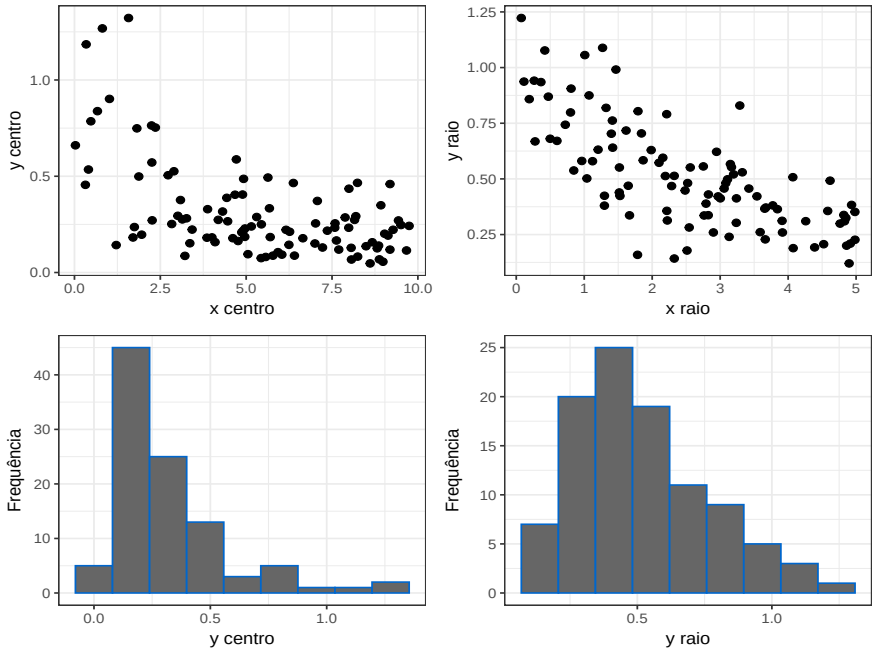
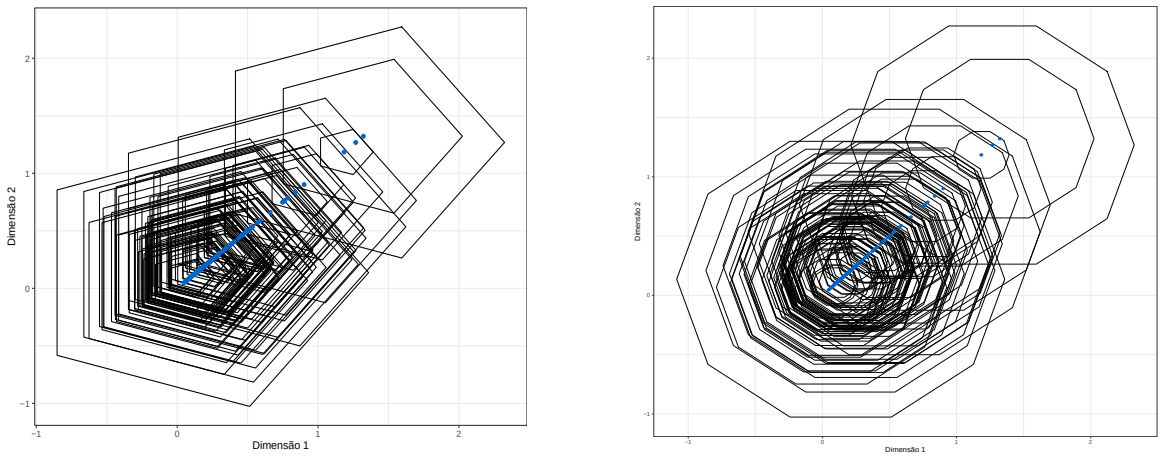


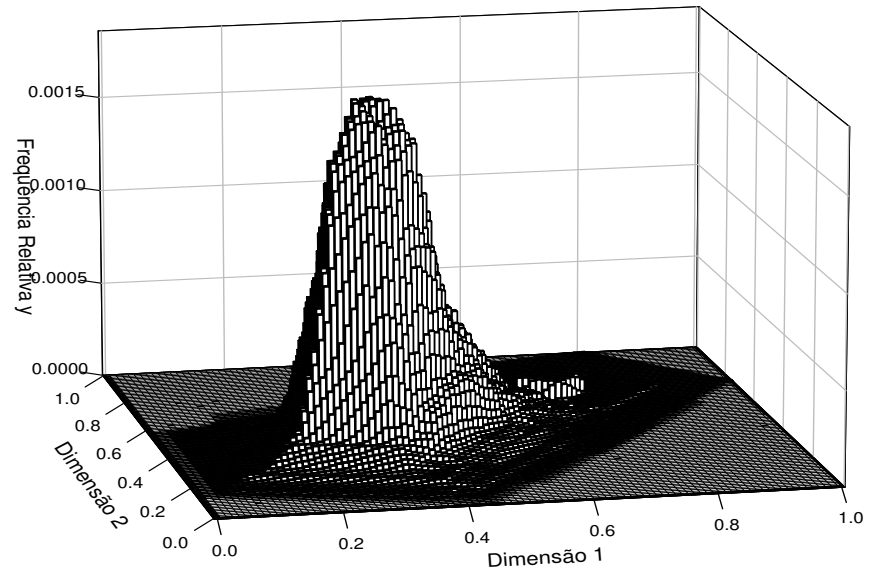
Figura 7 – Representação do centro e raio da variável resposta com distribuição Gama.



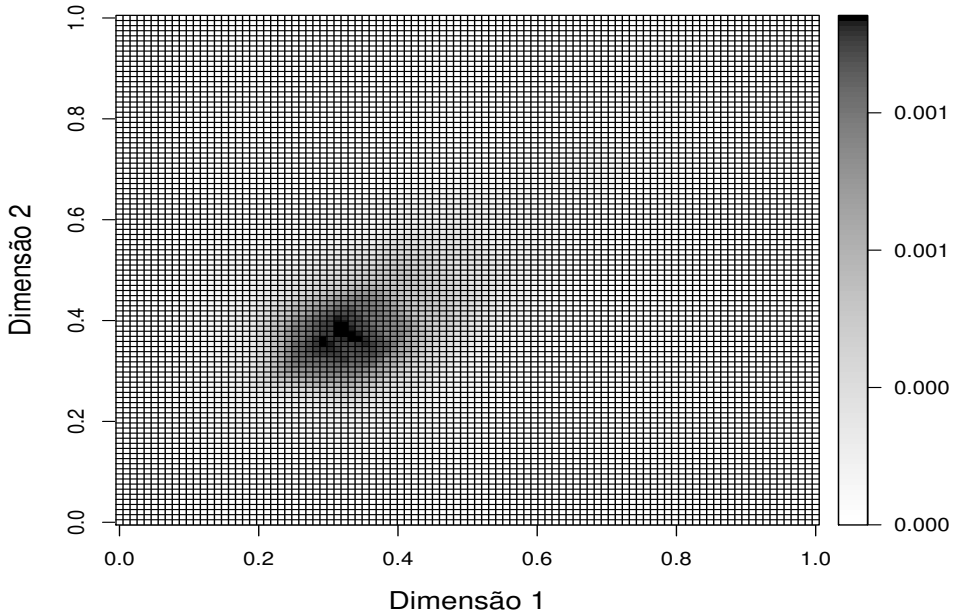
(a) 5 vértices. (b) 10 vértices.

Figura 8 – Representação da variável resposta poligonal com (a) 5 e (b) 10 vértices no cenário de distribuição Gama.

Na Figura 9 observa-se a dispersão da variável resposta poligonal com 5 vértices. Pelo histograma (a), percebe-se distribuição unimodal e assimetria, com concentração de frequência entre 0,2 e 0,5 (b). As seguintes medidas poligonais foram calculadas a partir da biblioteca *psda* (SILVA; SOUZA; CYSNEIROS, 2020) definidas nas Equações (2.20) e (2.21): média poligonal empírica $(0{,}31;\ 0{,}31)^{\mathsf{T}}$ e desvio padrão poligonal empírico $(0{,}35;\ 0{,}35)^{\mathsf{T}}$.



(a) Histograma de frequência relativa da variável resposta.

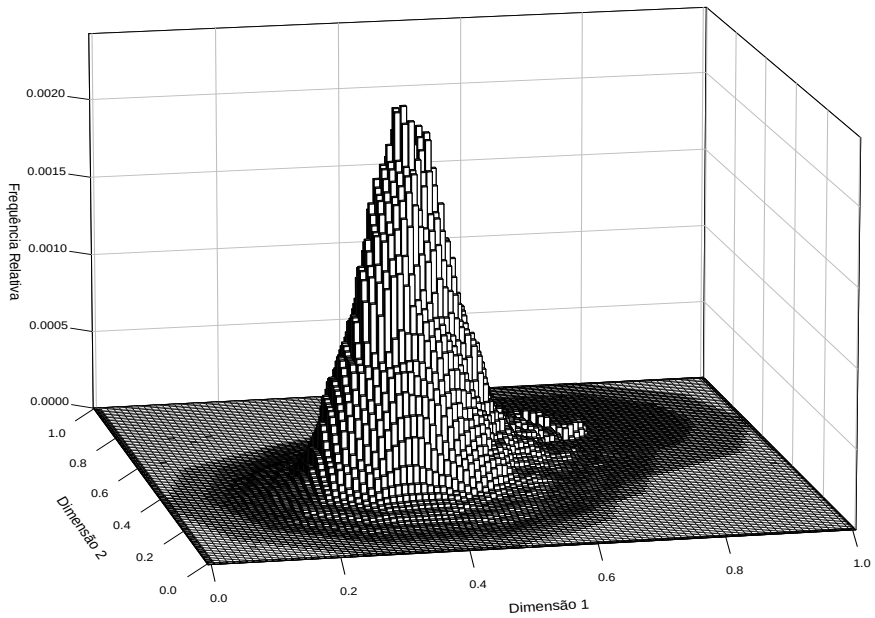


(b) Concentração de frequência da variável resposta poligonal.

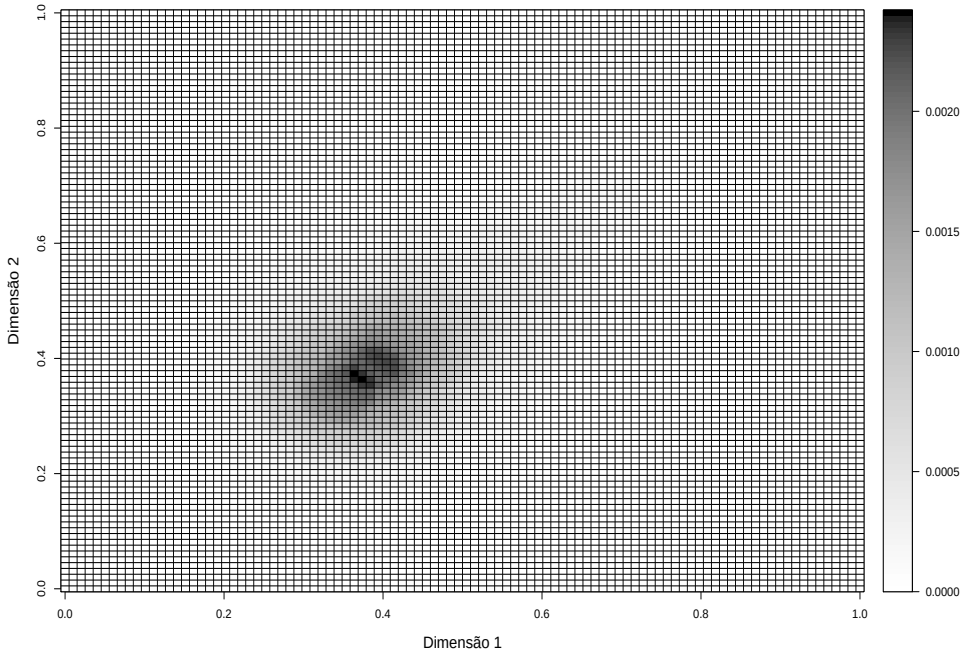
Figura 9 – Variável resposta poligonal com 5 vértices e distribuição Gama.

Já na Figura 10 observa-se o comportamento da variável resposta poligonal com 10 vértices. Lembrando que a configuração que origina os dados clássicos de centro e raio é o mesma,

independente da quantidade de vértices definida. Pelo histograma (a) também percebe-se distribuição com assimetria a direita e concentração de frequência entre 0,2 e 0,5 (b). As seguintes medidas poligonais foram calculadas: média poligonal empírica $(0{,}31;\ 0{,}31)^{\mathsf{T}}$; desvio padrão poligonal empírico $(0{,}37;\ 0{,}37)^{\mathsf{T}}$.



(a) Histograma de frequência relativa da variável resposta.



(b) Concentração de frequência da variável resposta poligonal.

Figura 10 – Variável resposta poligonal com 10 vértices e distribuição Gama.

4.1.2 Cenário 2: Distribuição Normal Inversa

O segundo cenário de dados simulados considera a distribuição Normal Inversa. O Algoritmo 8 descreve o processo de geração, considerando a função de ligação canônica recíproca quadrática. A representação dos dados de centro e raio é mostrada na Figura 11 para uma amostra que representa o cenário gerado. Os histogramas revelam assimetria à direita, e os gráficos de dispersão da variável resposta em função da variável explicativa revelam que a variância não é constante.

Algoritmo 8: Geração de conjuntos simulados com distribuição Normal Inversa

1: **Requerer** $n = 100$.
2: **Defina** a função de ligação como recíproca quadrática.
3: **Defina** x_i^c obtido de uma distribuição $U(a = 0; b = 4)$.
4: **Defina** x_i^r obtido de uma distribuição $U(a = 0; b = 3)$.
5: **Calcule** $y_i^c = 0,5 + 2,5x_i^c$ obtido de uma distribuição $NI(\mu = 0{,}5;\ \phi = 7)$.
6: **Calcule** $y_i^r = 2,5 + 1,5x_i^r$ obtido de uma distribuição $NI(\mu = 0{,}5;\ \phi = 23)$.
7: **Compute** os vértices usando a Equação (2.16).

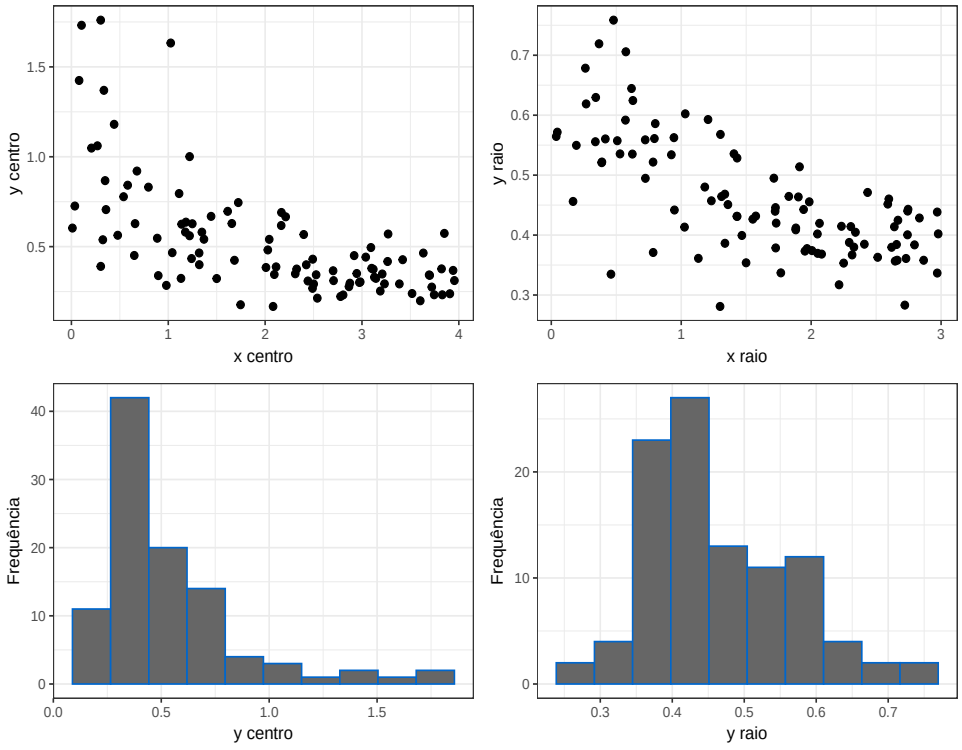
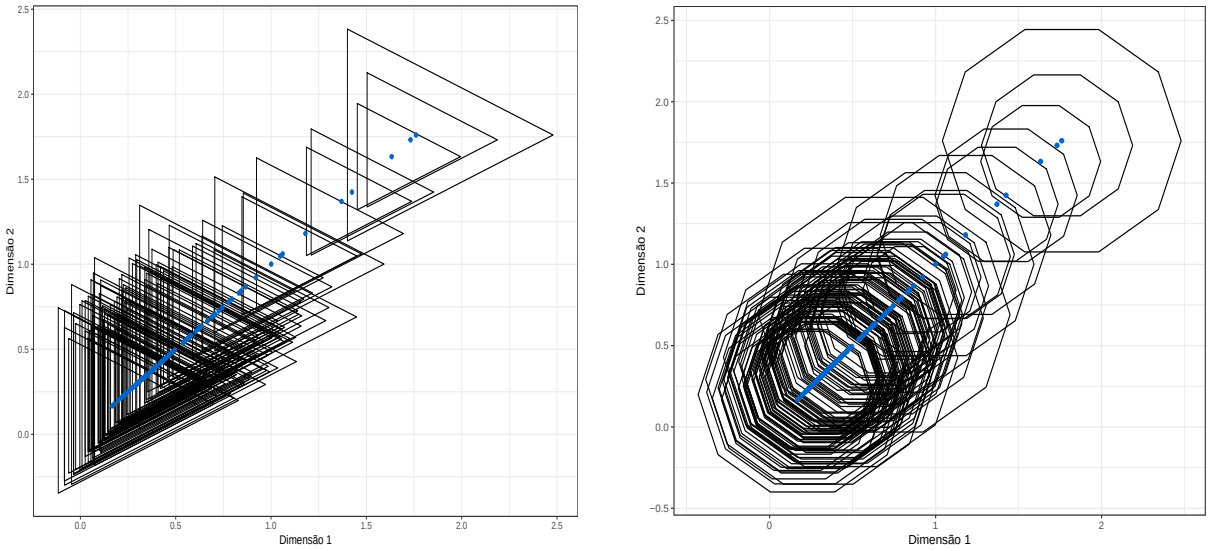


Figura 11 – Representação do centro e raio da variável resposta com distribuição Normal Inversa.

A Figura 12 representa a variável poligonal construída a partir dos dados gerados para o centro e para o raio. Observa-se que os centros dos polígonos, destacados em azul, apresentam maior proximidade no início da distribuição, afastando-se progressivamente ao longo da amostra. Dessa forma, a maior concentração dos dados poligonais encontra-se no intervalo entre 0 e 5, refletindo uma distribuição mais densa nas primeiras observações da série..

Com relação aos raios, nota-se a presença de observações com maior variação de área. Diferentemente do cenário com distribuição Gama, o intervalo de valores adotado neste caso é menor, o que torna os raios mais sensíveis a variações nos dados.



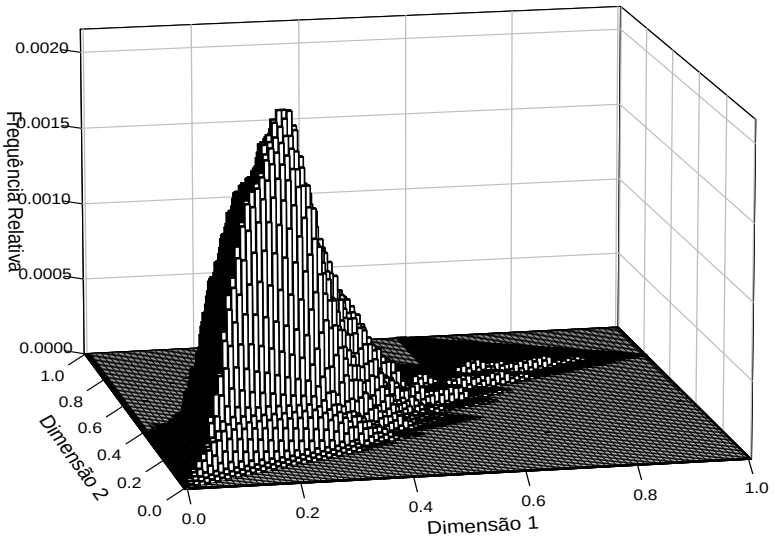
(a) 3 vértices. (b) 10 vértices.

Figura 12 – Representação da variável resposta poligonal com (a) 5 e (b) 10 vértices no cenário de distribuição Normal Inversa.

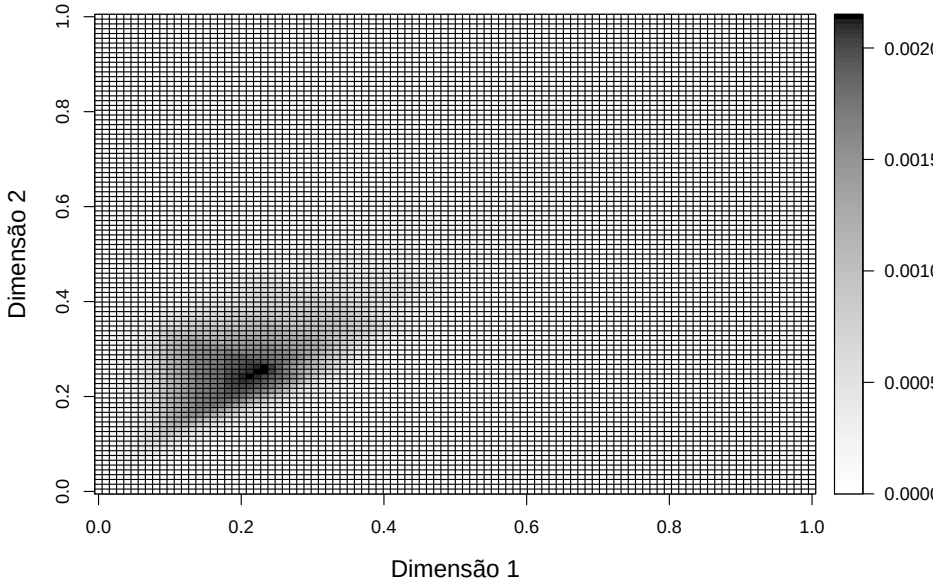
As medidas poligonais da variável resposta com 3 vértices foram obtidas a partir dos valores dos centros e raios estimados, resultando em uma média poligonal empírica igual a $(0{,}52;\ 0{,}52)^{\mathsf{T}}$ e um desvio padrão poligonal empírico de $(0{,}13;\ 0{,}13)^{\mathsf{T}}$. Esses valores refletem uma distribuição centrada em torno do ponto médio da escala considerada, com moderada dispersão. A Figura 13 apresenta a distribuição da variável resposta poligonal simulada segundo uma distribuição Normal Inversa.

No histograma da Figura 13(a), observa-se que a distribuição é unimodal, com o pico de frequência bem definido. Já na subfigura (b), nota-se uma assimetria na distribuição, com concentração mais acentuada de frequências no intervalo entre 0,1 e 0,3. Esses padrões reforçam o comportamento assimétrico e concentrado da variável resposta poligonal nesse cenário de distribuição Normal Inversa com 3 vértices.

Na Figura 14 observa-se a dispersão da variável resposta poligonal com 10 vértices. Pelo histograma (a) percebe-se distribuição unimodal e assimetria e concentração de frequência entre 0,2 e 0,5 (b). As seguintes medidas poligonais foram calculadas: média poligonal empírica $(0{,}31;\ 0{,}31)^{\mathsf{T}}$ e desvio padrão poligonal empírico $(0{,}35;\ 0{,}35)^{\mathsf{T}}$. Ao comparar os dois cenários, nota-se que o desvio padrão aumentou com a variável poligonal com 10 vértices.

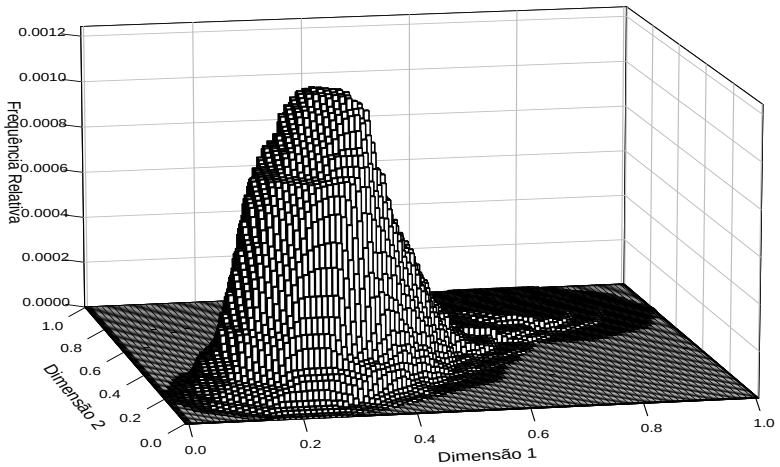


(a) Histograma de frequência relativa da variável resposta.

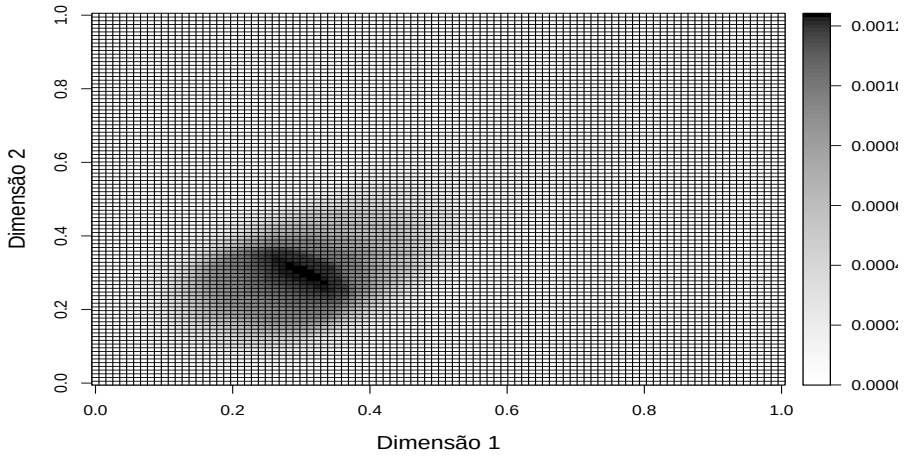


(b) Concentração de frequência da variável resposta poligonal.

Figura 13 – Variável resposta poligonal com 3 vértices e distribuição Normal Inversa.



(a) Histograma de frequência relativa da variável resposta.



(b) Concentração de frequência da variável resposta poligonal.

Figura 14 – Variável resposta poligonal com 10 vértices e distribuição Normal Inversa.

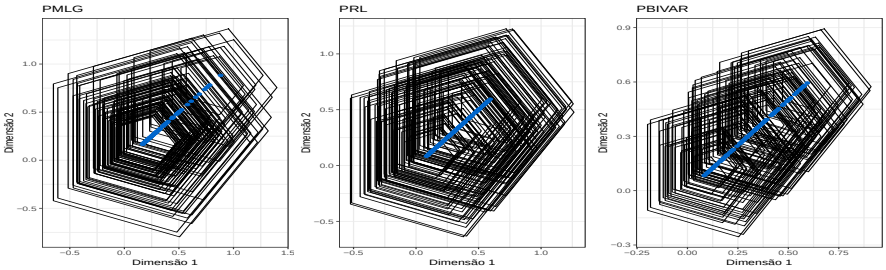
4.2 DIAGNÓSTICO DO MODELO: ANÁLISE DE RESÍDUOS

Nesta seção, são apresentadas as análises de resíduos dos modelos avaliados nos cenários com dados simulados. O resíduo poligonal está definido na Equação 3.4. Medidas descritivas e representações gráficas são fornecidas para auxiliar na interpretação dos resultados.

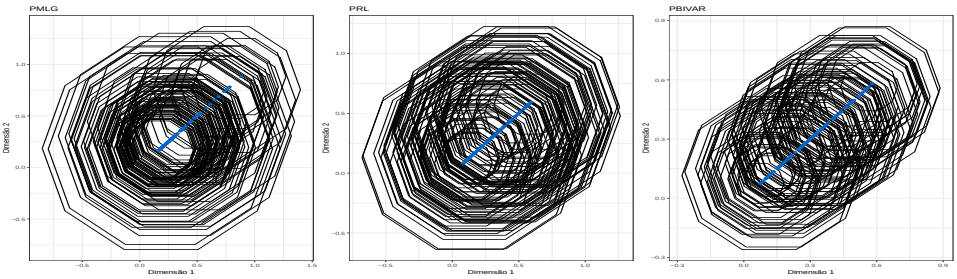
4.2.1 Cenário 1: Distribuição Gama

Antes de fazer a reapresentação gráfica e análise dos resíduos, é necessário observar a variável poligonal predita pelos modelos PMLG, PRL e PBIVAR, revelando distribuições com algumas particularidades. A Figura 15 mostra que o modelo PMLG apresenta polígonos com

centros acumulados no início da distribuições, e mais dispersos ao final, diferentemente do PRL, o qual percebe-se uma distribuição de objetos mais próxima. O modelo PBIVAR, além de exibir uma distribuição igualitária entre os objetos, obtém menor valor de raios.



(a) Variável resposta predita com 5 vértices.



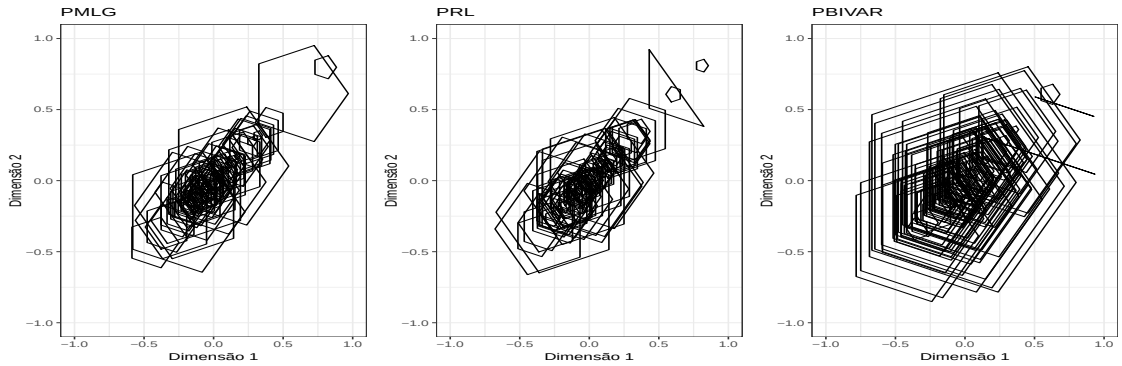
(b) Variável resposta predita com 10 vértices.

Figura 15 – Representação da variável predita poligonal no cenário de distribuição Gama.

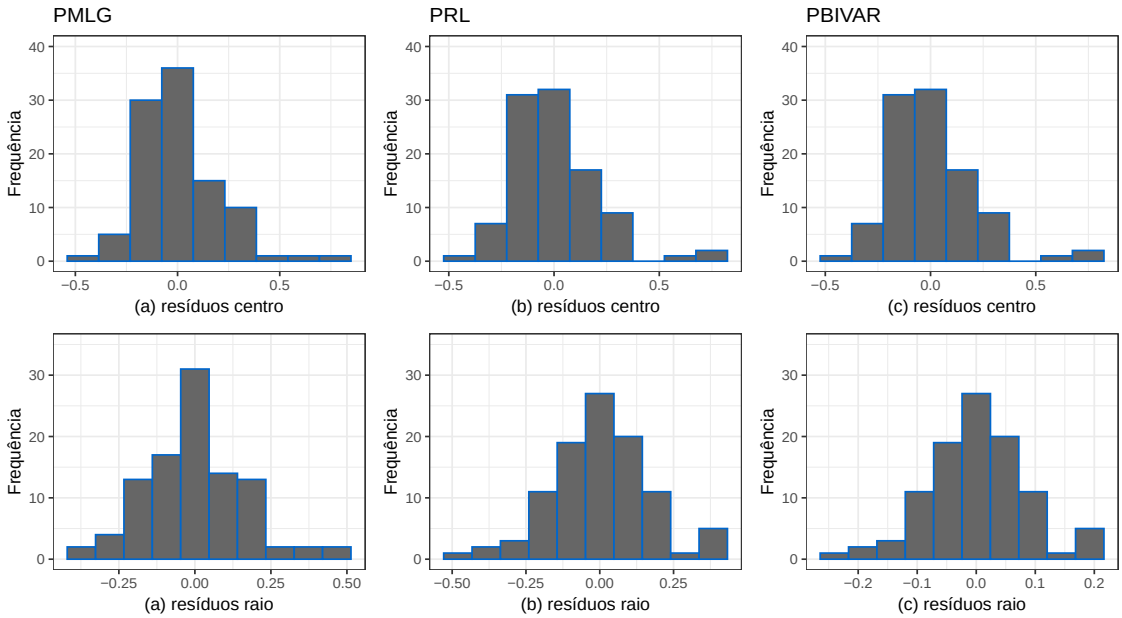
Como definido no Capítulo 3, o resíduo poligonal é calculado a partir das diferenças entre os vértices observados e os vértices preditos. Espera-se que os polígonos residuais sejam, em sua maioria, degenerados e próximos à coordenada (0, 0) do espaço representado. Dito isto, analisa-se os resíduos do cenário de dados Gama com polígonos de 5 vértices.

Percebe-se na Figura 16(a) que os modelos PMLG e PRL apresentam menor área residual e maior proximidade ao ponto (0, 0), indicando menores resíduos tanto para a posição (centro) quanto para a área (raio). A Figura 16(b) detalha os resíduos de centro e raio, evidenciando que o modelo PMLG concentra maiores frequências no valor zero e exibe menor assimetria. A dispersão dos resíduos é apresentada na Figura 16(c). Já na Figura 17, os histogramas permitem analisar a distribuição de frequências dos resíduos, destacando que o modelo PBIVAR possui maior ocupação da área, o que indica resíduos mais dispersos.

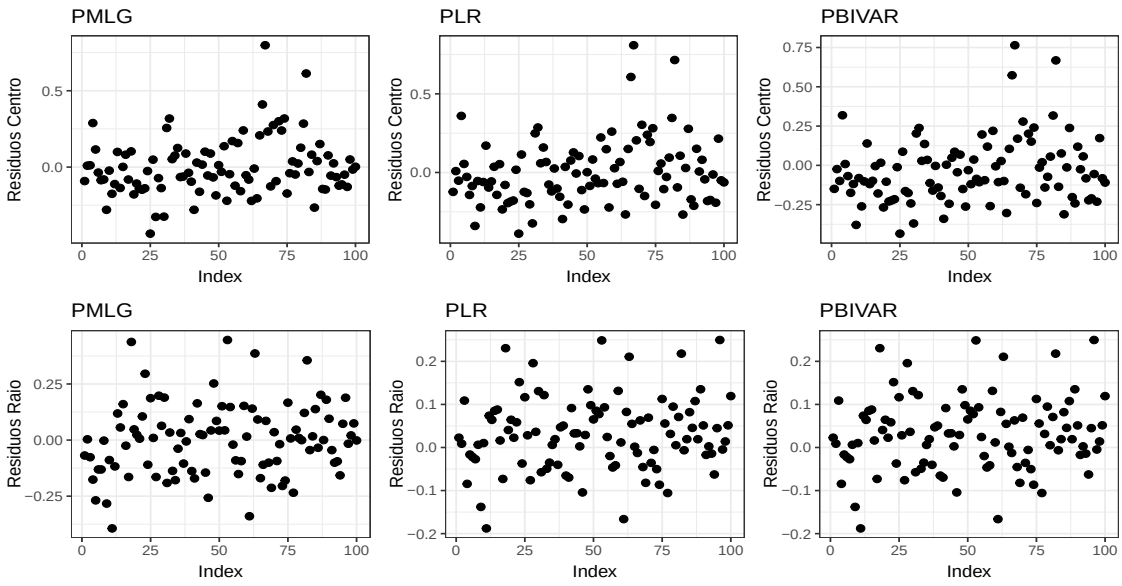
Em relação ao cenário Gama com polígonos de 10 vértices, a Figura 18(a) apresenta menor área residual dos modelos PMLG e PRL, que estão mais próximos das coordenadas (0, 0), indicando pequenos resíduos para centro e raio. As Figuras 16(b) e (c) exibem os resíduos de centro e raio, evidenciando assimetria nos histogramas devido a valores mais altos de resíduos. Já a Figura 19 mostra a concentração de frequências desses resíduos por meio dos histogramas.



(a) Resíduo Poligonal.

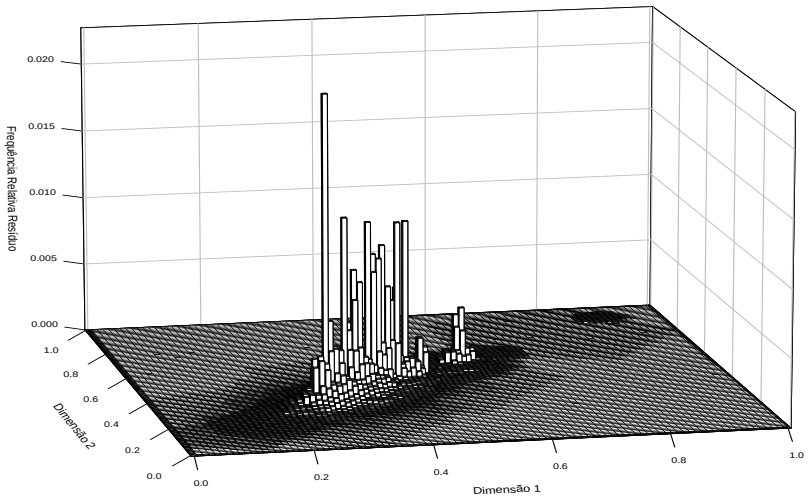


(b) Histograma dos resíduos de centro e raio.

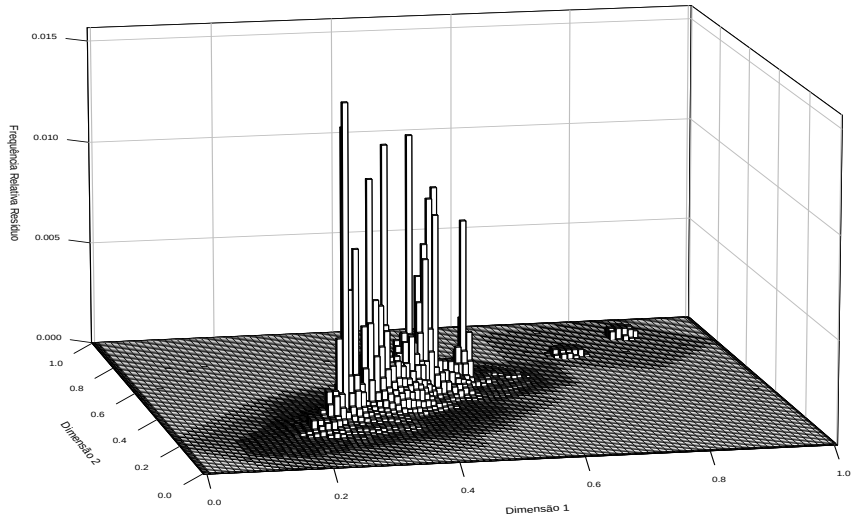


(c) Dispersão Centro e Raio.

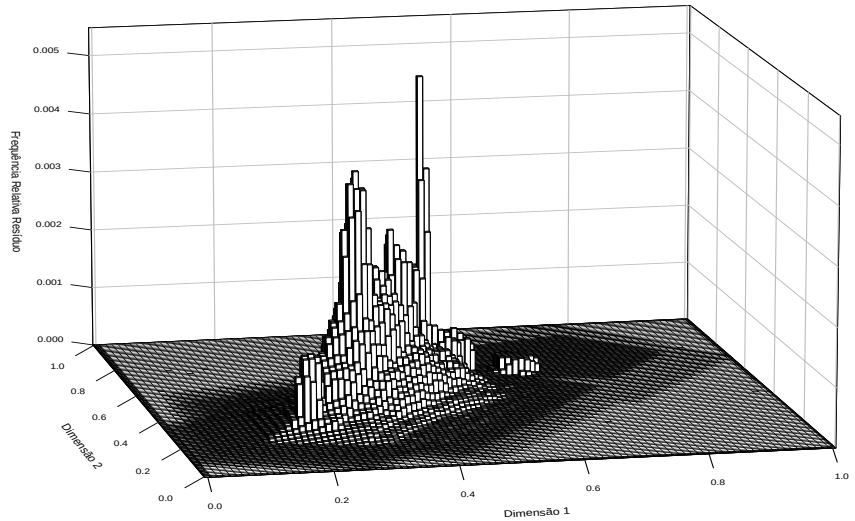
Figura 16 – Representação dos resíduos para polígonos com 5 vértices no cenário de distribuição Gama.



(a) PMLG

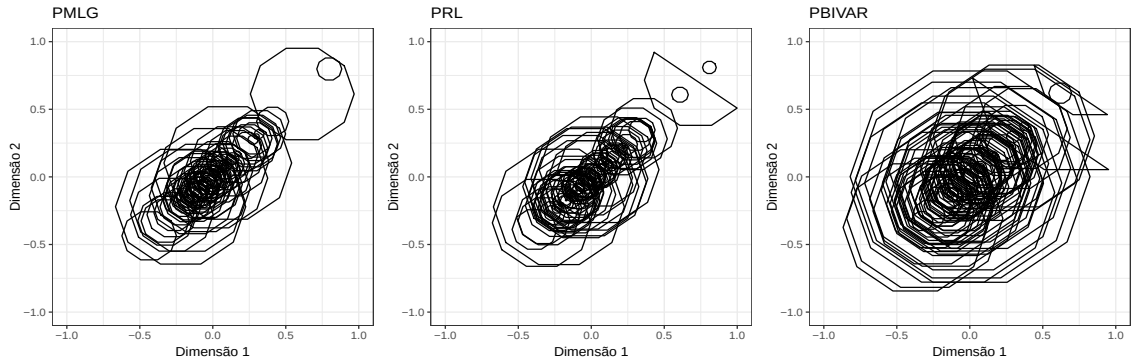


(b) PRL

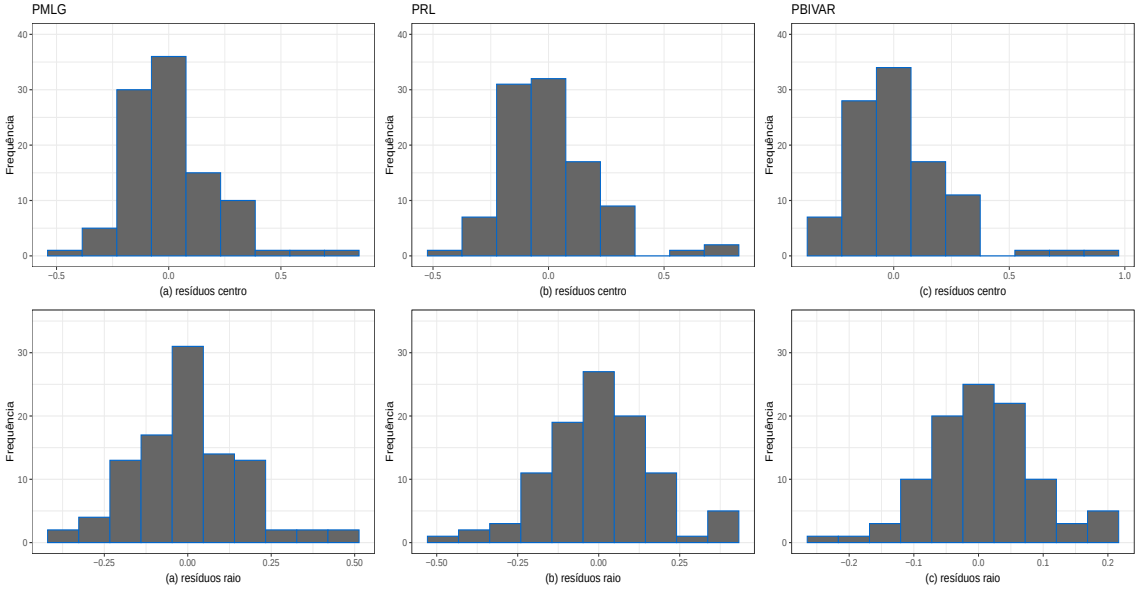


(c) PBIVAR

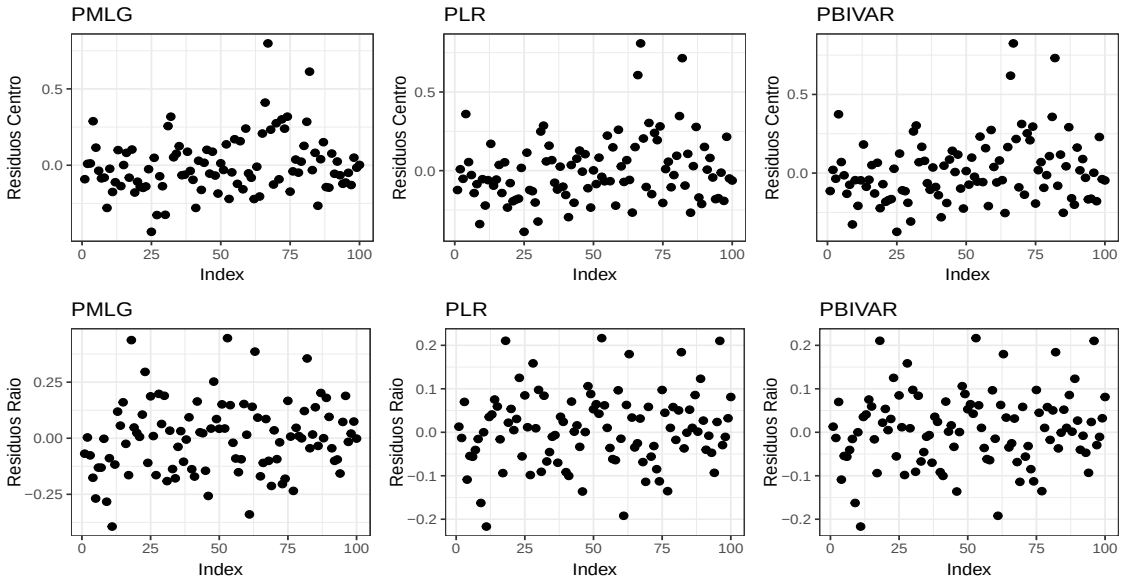
Figura 17 – Histogramas dos resíduos poligonais com 5 vértices no cenário de distribuição Gama.



(a) Resíduo Poligonal.

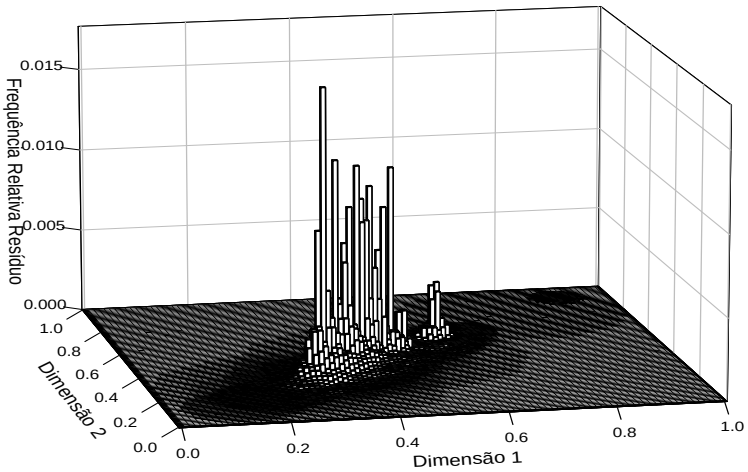


(b) Histograma dos resíduos de centro e raio.

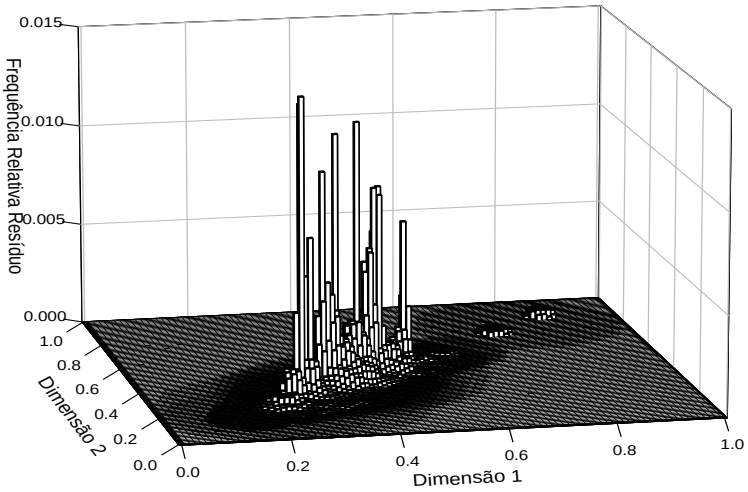


(c) Dispersão Centro e Raio.

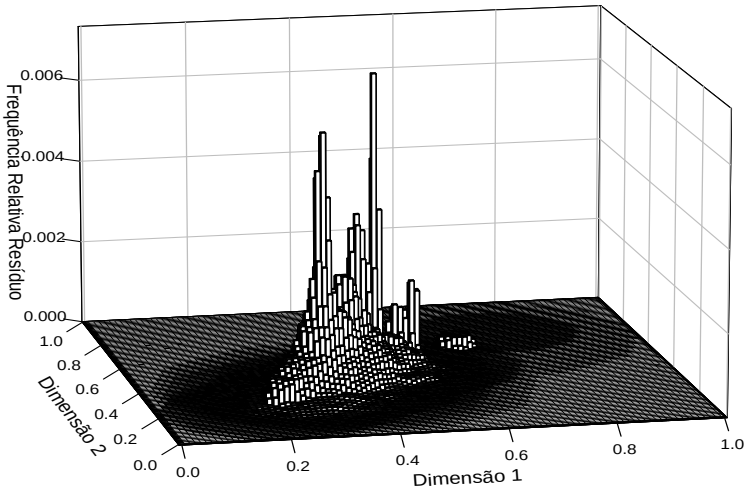
Figura 18 – Representação dos resíduos para polígonos com 10 vértices e distribuição Gama.



(a) PMLG



(b) PRL

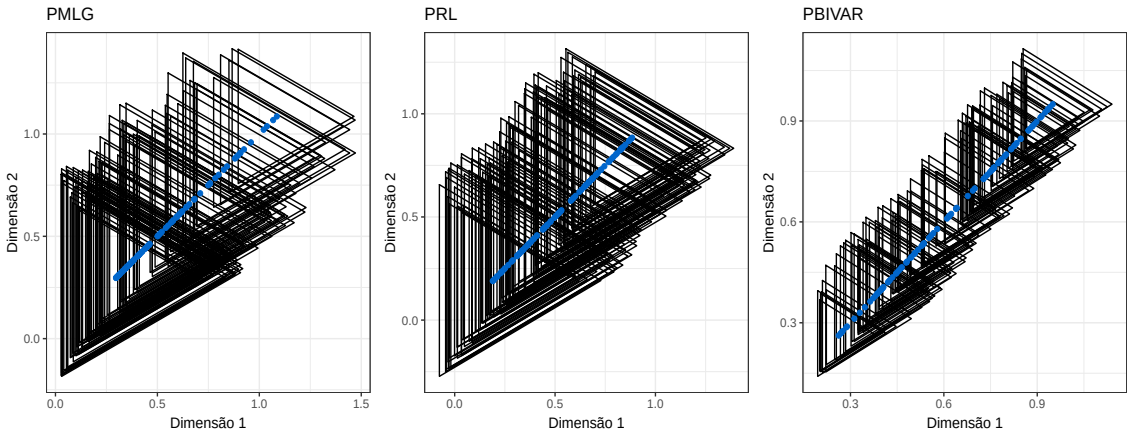


(c) PBIVAR

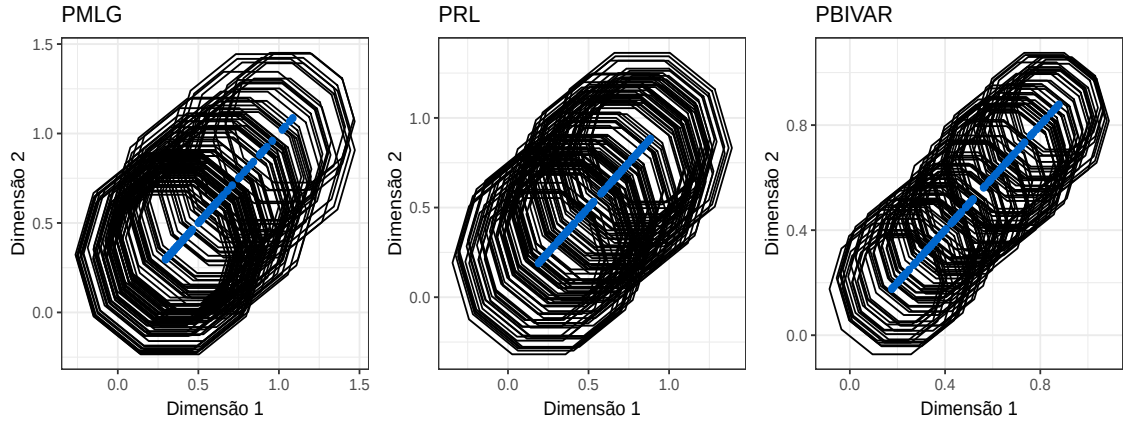
Figura 19 – Histogramas dos resíduos poligonais com 10 vértices no cenário de distribuição Gama.

4.2.2 Cenário 2: Distribuição Normal Inversa

A Figura 20 mostra os polígonos preditos pelos três modelos implementados. Percebe-se que, com (a) 3 ou (b) 10 vértices, o PMLG obteve mais dispersão em relação aos valores de centro do que as demais técnicas. Em relação aos raios, o modelo PBIVAR obteve valores mais baixos mostrado pela distribuição dos polígonos em uma distribuição menor. Comparado esses gráficos com a variável resposta poligonal observada, percebe-se, geometricamente, maior proximidade do modelo PMLG.



(a) Variável resposta predita com 3 vértices.



(b) Variável resposta predita com 10 vértices.

Figura 20 – Representação da variável predita poligonal no cenário de distribuição Normal Inversa.

Em relação aos resíduos poligonais, a Figura 21 exibe a variável poligonal resultante de cada modelo. Os modelos PMLG e PRL possuem menor área, sendo polígonos distribuídos entre (0,5; -0,5) em maioria, e com valores pequenos de raio formando pequenos polígonos. Pode-se notar alguns polígonos residuais degenerados formando pontos e retas nos polígonos dos três modelos.

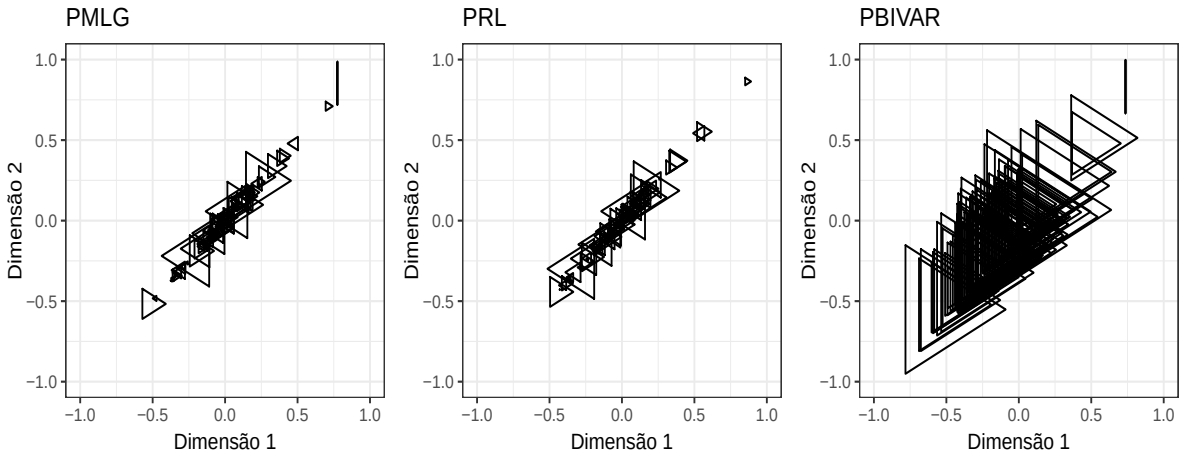


Figura 21 – Representação dos resíduos poligonais com 3 vértices no cenário de distribuição Normal Inversa.

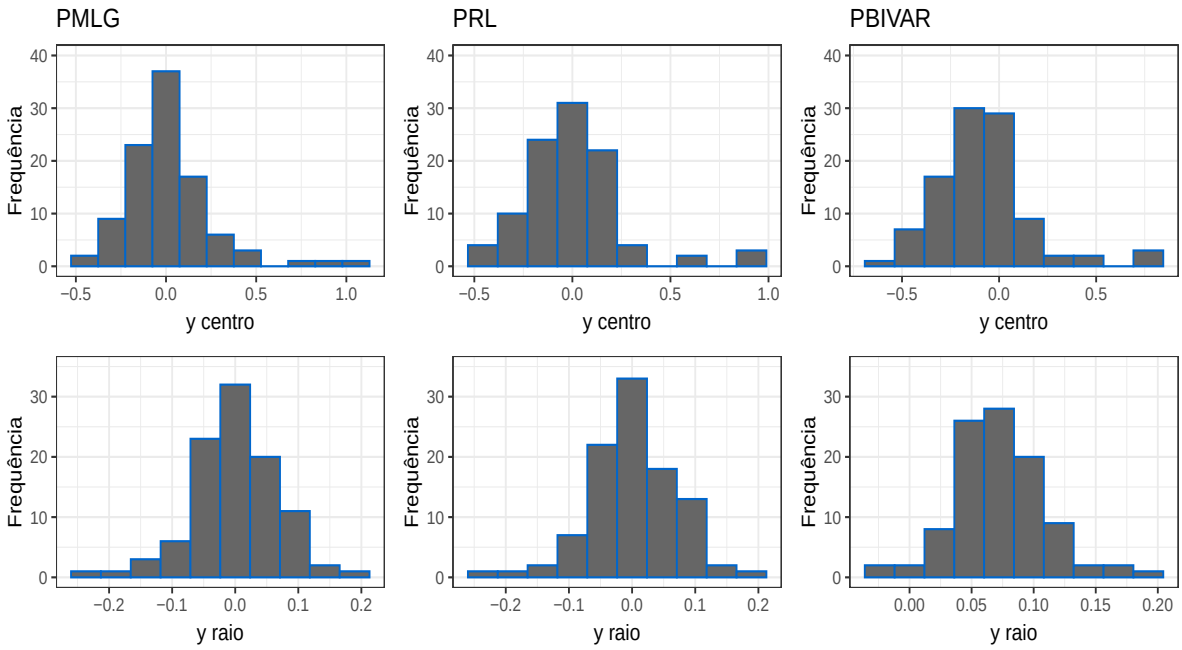
Na Figura 22(a), observa-se que a representação dos resíduos poligonais associados ao centro apresenta um histograma assimetrico. Além disso, destaca-se que o modelo PMLG concentra maior frequência de resíduos em torno do zero, indicando melhor ajuste central em relação aos demais modelos. A distribuição dos resíduos associados ao raio é assimétrica nos modelos PMLG e PRL, enquanto o modelo PBIVAR apresenta um comportamento mais simétrico, com predominância de valores positivos.

Essas características também são evidenciadas na Figura 22(b), a qual representa a dispersão dos resíduos nas duas componentes (centro e raio). Já na Figura 23, verifica-se a concentração de frequências dos resíduos, sendo notável que a região ocupada pelos resíduos do modelo PBIVAR é mais extensa, indicando maior variabilidade residual.

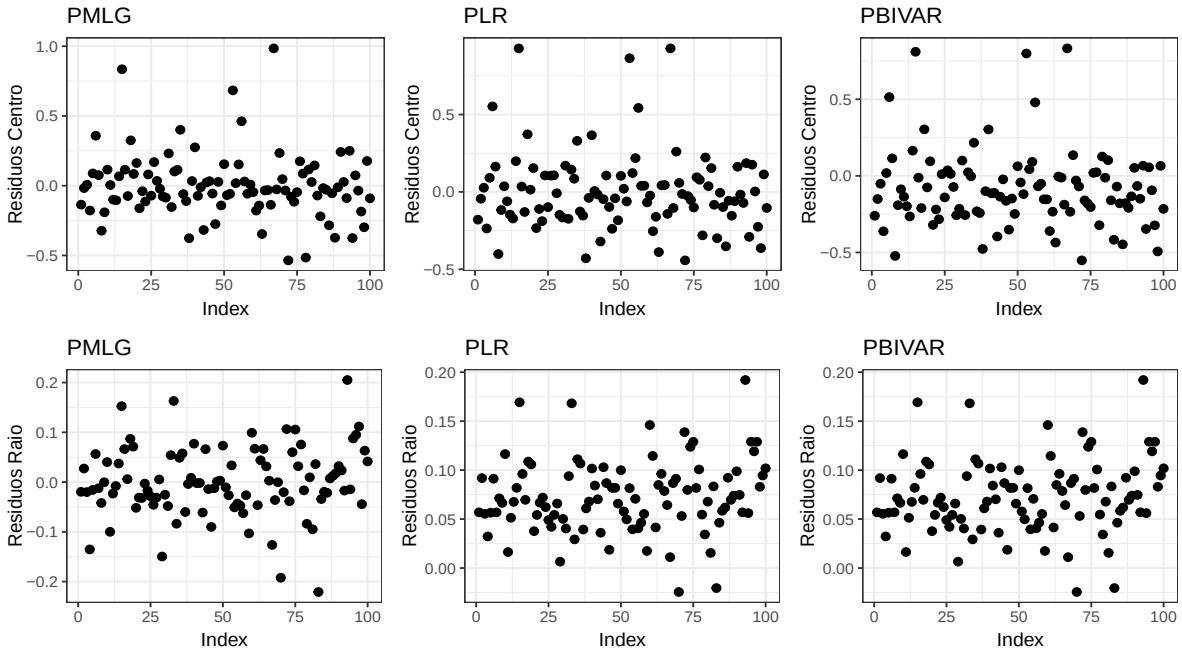
A Figura 24(a) apresenta a representação dos resíduos poligonais considerando 10 vértices. Assim como na análise com 3 vértices, os modelos PMLG e PRL continuam exibindo menor área residual, com os polígonos concentrando-se majoritariamente entre as coordenadas (0,5; -0,5). Nota-se ainda a ocorrência de polígonos degenerados, com formações que se aproximam de pontos, retas ou outras representações poligonais não regulares, indicando possíveis resíduos nulos ou extremos.

Na Figura 24(b), observa-se que a representação dos resíduos do centro mantém o padrão assimétrico, com destaque para a presença de resíduos discrepantes positivos. O modelo PMLG novamente concentra a maior frequência de resíduos no valor zero. A distribuição dos resíduos de raio apresenta assimetria nos três modelos, embora o PBIVAR evidencie maior concentração de valores nulos. A dispersão dos resíduos pode ser analisada com maior clareza na Figura 24(c), enquanto a Figura 25 revela a concentração de frequências residuais, com destaque para o

PBIVAR, cuja área ocupada é mais ampla, sugerindo maior variabilidade em comparação aos demais modelos.

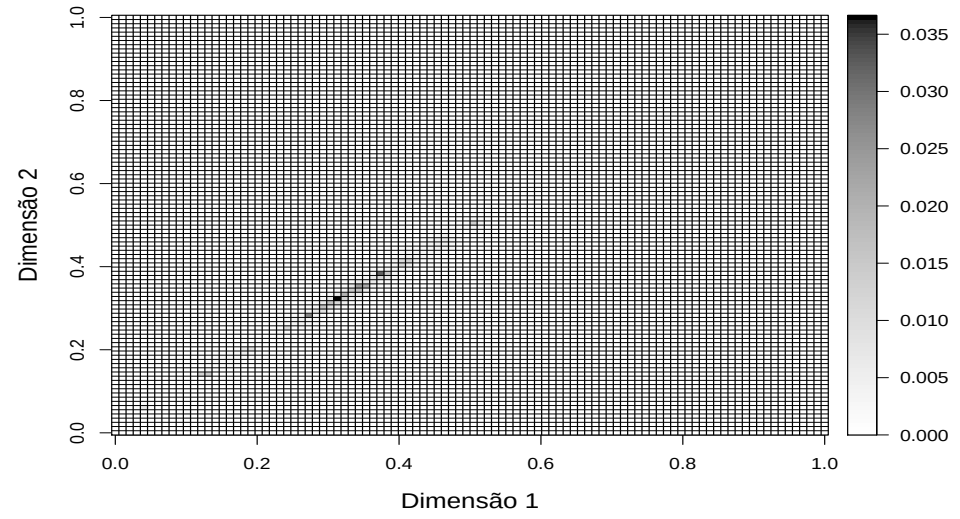


(a) Representação dos Resíduos dos centros e raios.

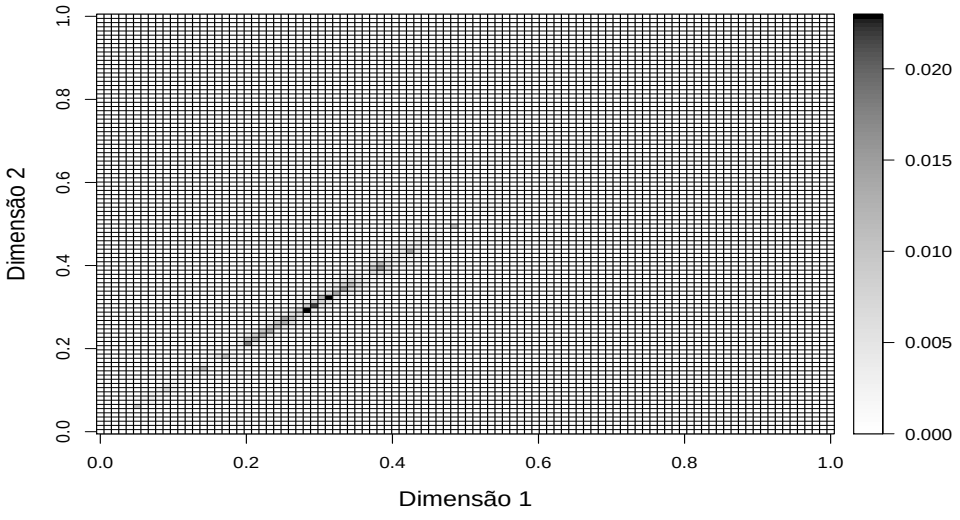


(b) Dispersão dos Resíduos.

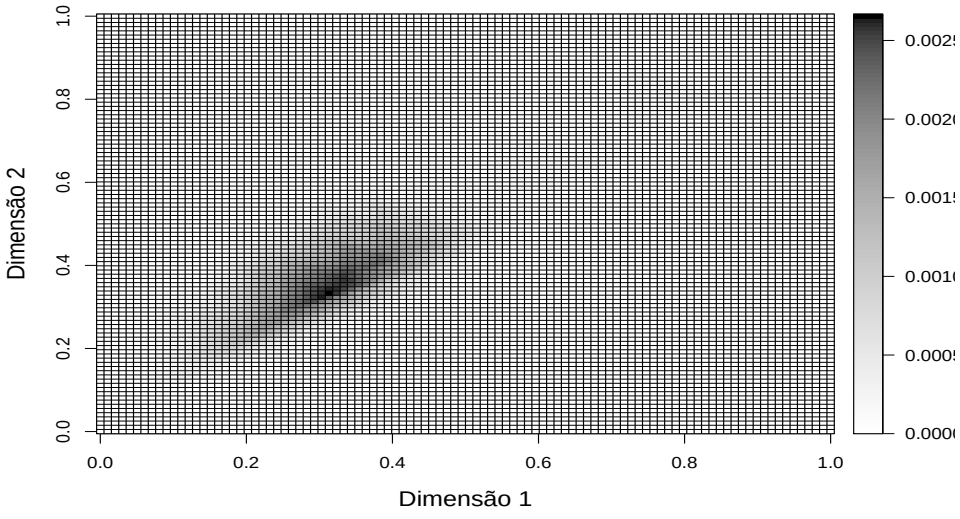
Figura 22 – Representação dos resíduos para polígonos com 3 vértices e distribuição Normal Inversa.



(a) PMLG

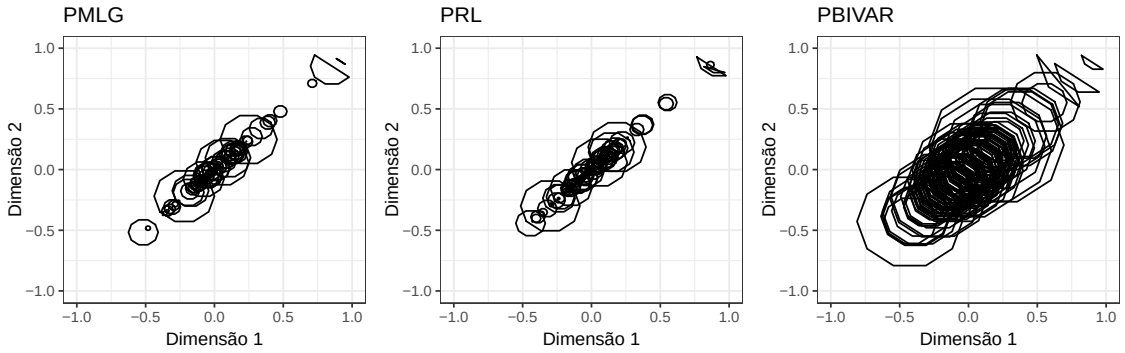


(b) PRL

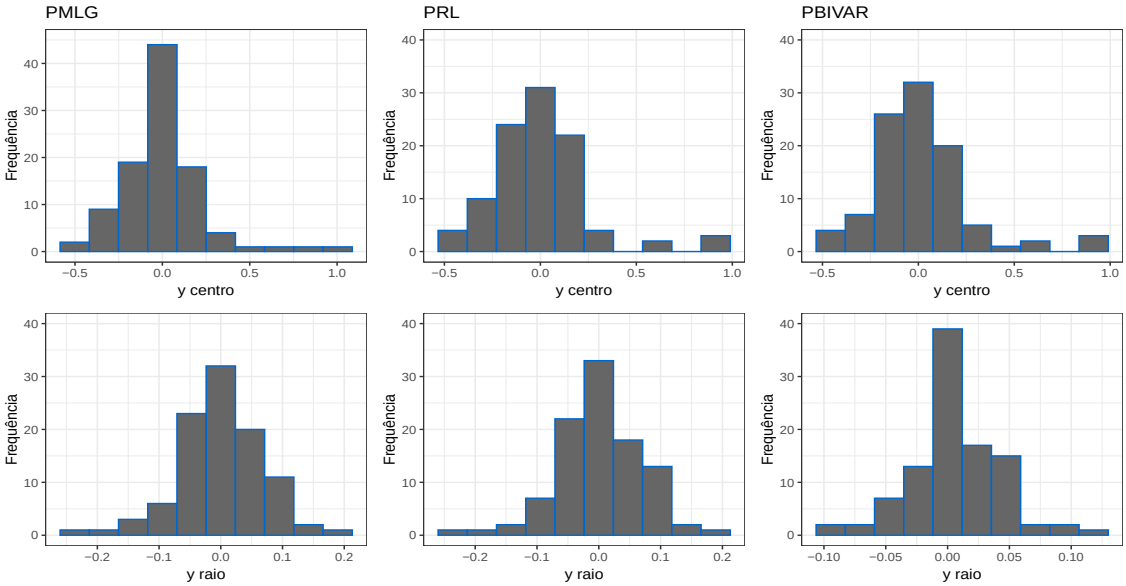


(c) PBIVAR

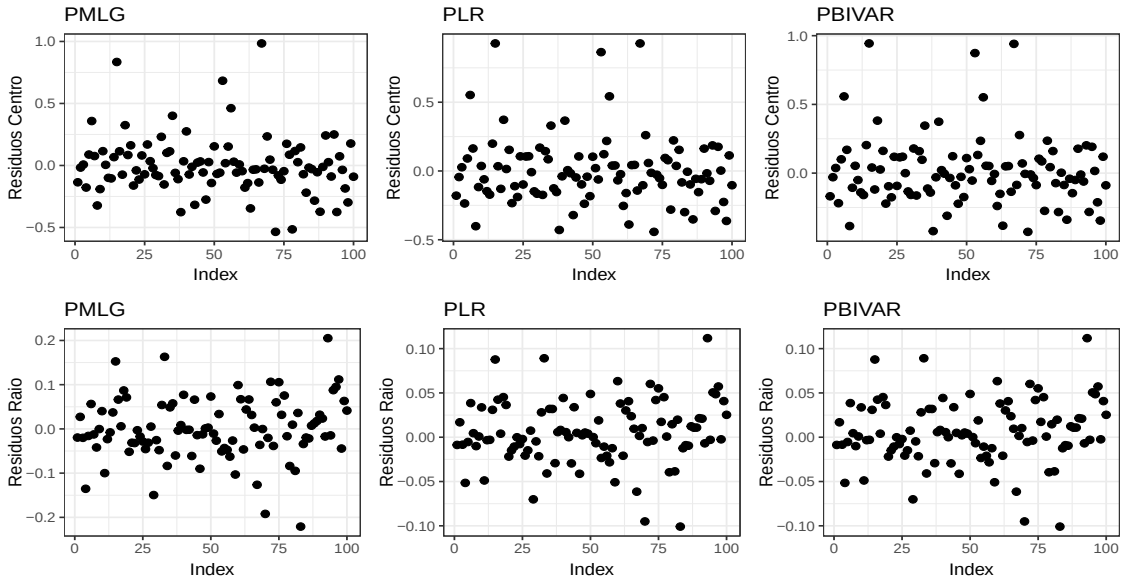
Figura 23 – Concentração de frequência dos resíduos poligonais com 3 vértices para dados com distribuição Normal Inversa.



(a) Representação dos resíduos poligonais.

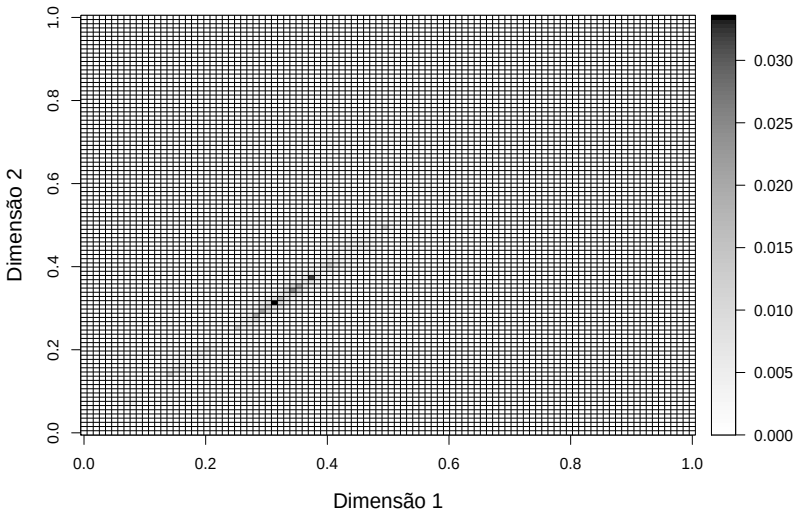


(b) Histograma dos Resíduos de centro e raio.

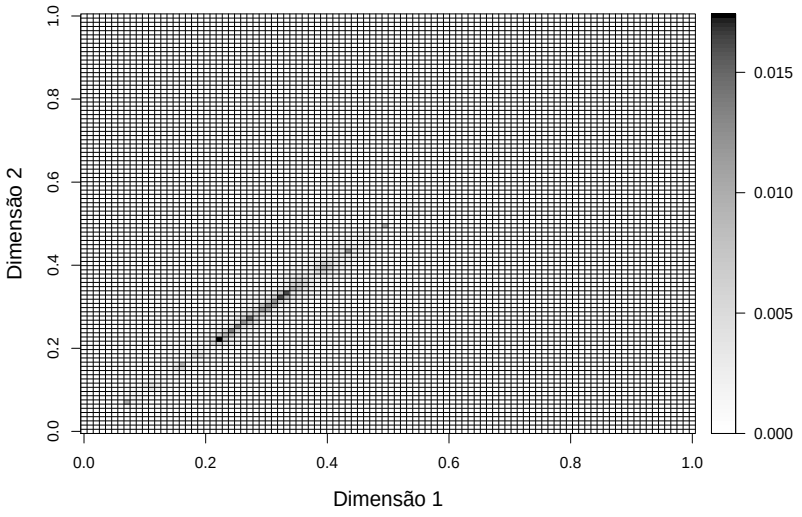


(c) Dispersão dos resíduos.

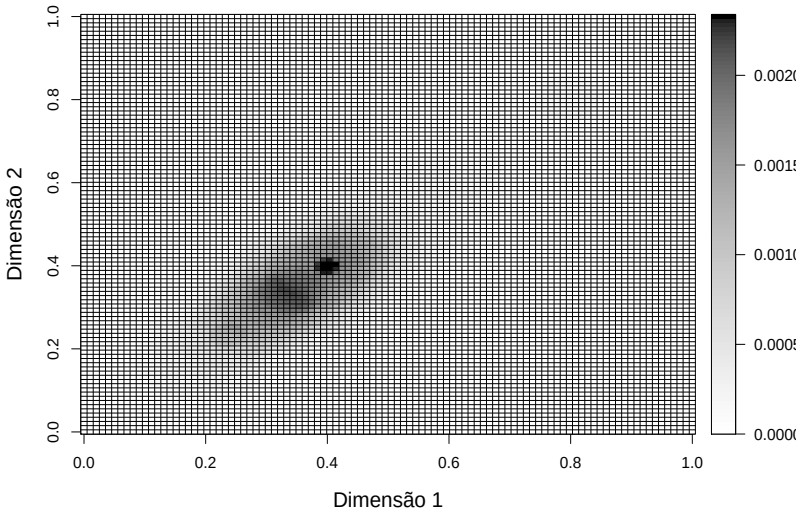
Figura 24 – Representação dos resíduos para polígonos com 10 vértices e distribuição Normal Inversa.



(a) PMLG



(b) PRL



(c) PBIVAR

Figura 25 – Concentração de frequência dos resíduos poligonais com 10 vértices para dados com distribuição Normal Inversa.

4.3 ANÁLISE PREDITIVA

O desempenho dos métodos de predição será mensurado por meio de quatro métricas: Erro Médio Quadrático da Área (EMQA), Erro Médio Quadrático da Distância dos Vértices (EMQDV), Erro Médio Quadrático da Área e Centro Conjuntamente (EMQAC) e Erro Médio Quadrático do Centro e Raio Conjuntamente (EMQCR), definidas na subseção 3.2.2. Para avaliar a abordagem proposta, utiliza-se o método de Monte Carlo (MC) com 100 iterações, em que, a cada repetição, os dados são particionados aleatoriamente em 75% para o treinamento dos modelos e 25% para o teste, conforme descrito no Algoritmo 9.

Algoritmo 9: Método Monte Carlo Para Dados Simulados

1: **Requerer** MC $= 100$.
2: **Requerer** tamanho da base de treino $n1 = 150$.
3: **Requerer** tamanho da base de teste $n2 = 50$.
4: **Se** cenário de dados com distribuição Gama **Então:**
5: **Requerer** número de vértices $L = 5$ ou $L = 10$.
6: **Senão**:
7: **Requerer** número de vértices $L = 3$ ou $L = 10$.
8: **Para todo** $i \leftarrow 1$ **até** MC **faça:**
9: **Gere** uma base de treino de tamanho $n1$.
10: **Gere** uma base de teste de tamanho $n2$.
11: **Aplique** os métodos PMLG, PRL e PBIVAR nos dados de treino.
12: **Aplique** a regra de predição nos dados de teste.
13: **Calcule** as medidas de desempenho usando as Equações (3.5, 3.6, 3.7 e 3.8).
14: **Fim Para**
15: **Calcule** a média e desvio padrão das medidas de desempenho.

Outra forma de medição de desempenho é através do Ganho Relativo (GR). O GR é aplicado para mensurar o ganho em relação a minimização do erro de predição, dado em porcentagem. O cálculo é mostrado na Equação 4.1, onde Medida_a é o resultado de maior valor e Medida_b o valor de interesse. Ainda são abordadas avaliações por meio de testes estatísticos.

$$GR = 100\left(\frac{Medida_a - Medida_b}{Medida_a}\right) \tag{4.1}$$

4.3.1 Distribuição Gama

Considerando os dados simulados a partir da distribuição Gama, a Tabela 8 concentra-se nos resultados obtidos. Os valores destacados (em negrito) enfatizam que o método PMLG apresentou a menor média e erro padrão em todos os casos. Os testes de Wilcoxon realizados para as amostras de erro indicaram, por meio do p-valor, a rejeição da hipótese nula. Estatisticamente, os erros médios do modelo PMLG são inferiores aos dos demais. A tabela também informa o GR do PMLG em relação aos outros modelos.

Tabela 8 – Resultados para o cenário de dados Gama: dados simulados com 5 vértices.

Medida	PMLG	PRL	PBIVAR
EMQDV	**0,672**	0,704	0,742
	(0,065)	(0,071)	(0,066)
	p-valor:	$5{,}2\text{x}10^{-14}$	$2{,}2\text{x}10^{-16}$
	GR:	4,5%	9,4%
EMQA	**0,595**	0,613	0,664
	(0,180)	(0,199)	(0,217)
	p-valor:	$1{,}65\text{x}10^{-3}$	$4{,}78\text{x}10^{-12}$
	GR:	2,9%	10,4%
EMQAC	**0,642**	0,662	0,712
	(0,171)	(0,187)	(0,204)
	p-valor:	$1{,}64\text{x}10^{-3}$	$2{,}49\text{x}10^{-13}$
	GR:	3,0%	9,8%
EMQCR	**0,293**	0,305	0,319
	(0,035)	(0,038)	(0,034)
	p-valor:	$1{,}35\text{x}10^{-11}$	$2{,}2\text{x}10^{-16}$
	GR:	3,8%	8,1%

A Tabela 9 apresenta os resultados dos experimentos considerando os dados da distribuição Gama com 10 vértices. Os valores em negrito indicam que o modelo PMLG obteve os menores erros médios e desvios padrão em todas as medidas avaliadas: EMQDV, EMQA, EMQAC e EMQCR. Os testes estatísticos de Wilcoxon, aplicados às amostras de erro, confirmam a superioridade do modelo PMLG ao apresentarem p-valores significativamente baixos em todas as comparações, indicando rejeição da hipótese nula de igualdade entre os métodos.

Além disso, a Tabela 9 também apresenta o GR do PMLG em relação aos outros métodos. Observa-se que o PMLG alcançou reduções de erro variando entre 2,9% e 10,4%, a depender da medida analisada, reforçando a efetividade do modelo na tarefa de predição poligonal no cenário de dados Gama com maior número de vértices.

Tabela 9 – Resultados para o cenário de dados Gama: dados simulados com 10 vértices.

Medida	PMLG	PRL	PBIVAR
EMQDV	**0,951**	0,996	0,988
	(0,092)	(0,100)	(0,102)
	p-valor:	$5{,}21\text{x}10^{-14}$	$1{,}09\text{x}10^{-11}$
	GR:	4,5%	9,4%
EMQA	**0,736**	0,757	0,762
	(0,223)	(0,246)	(0,252)
	p-valor:	$1{,}65\text{x}10^{-3}$	$4{,}78\text{x}10^{-12}$
	GR:	2,9%	10,4%
EMQAC	**0,774**	0,798	0,804
	(0,215)	(0,242)	(0,236)
	p-valor:	$4{,}28\text{x}10^{-4}$	$8{,}7\text{x}10^{-5}$
	GR:	3,0%	9,8%
EMQCR	**0,293**	0,305	0,319
	(0,035)	(0,038)	(0,039)
	p-valor:	$1{,}35\text{x}10^{-11}$	$6{,}46\text{x}10^{-11}$
	GR:	3,8%	8,1%

4.3.2 Distribuição Normal Inversa

As Tabelas 10 e 11 apresentam os valores médios das métricas de desempenho obtidas para os dados poligonais segundo a distribuição Normal Inversa, com 5 e 10 vértices, respectivamente. Os valores destacados evidenciam que o método PMLG obteve os menores valores médios e desvios padrão de erro em todas as métricas avaliadas.

A análise estatística realizada por meio do teste de Wilcoxon revelou, com base nos p-valores extremamente baixos (inferiores a 10^{-12} em todos os casos), a rejeição da hipótese nula de igualdade de distribuições de erro. Isso reforça que, estatisticamente, o PMLG apresenta desempenho superior em relação aos demais métodos avaliados. Além disso, oGR mostra que, em comparação aos métodos PRL e PBIVAR, o PMLG reduziu significativamente os erros, em especial na métrica EMQA, com ganhos de até 56,3% no cenário com 5 vértices.

Ao comparar os cenários com 5 e 10 vértices, observa-se que o aumento na complexidade geométrica dos dados resultou, de modo geral, a maiores valores médios e desvios padrão nas métricas EMQDV e EMQA, indicando menor estabilidade e precisão dos modelos. Essa maior variabilidade também reduziu os ganhos relativos obtidos, especialmente nos cenários mais complexos. Ainda assim, os resultados confirmam a robustez do método PMLG, que manteve

desempenho superior aos demais, mesmo diante do aumento do número de vértices dos dados poligonais.

Tabela 10 – Resultados para o cenário de dados Normal Inversa: dados simulados com 5 vértices.

Medida	**PMLG**	**PRL**	**PBIVAR**
EMQDV	**0,454**	0,503	0,739
	(0,065)	(0,079)	(0,071)
	p-valor:	$2{,}2\text{x}10^{-16}$	$2{,}2\text{x}10^{-16}$
	GR:	9,7%	38,5%
EMQA	**0,093**	0,095	0,213
	(0,017)	(0,017)	(0,018)
	p-valor:	$4{,}56\text{x}10^{-13}$	$2{,}2\text{x}10^{-16}$
	GR:	2,1%	56,3%
EMQAC	**0,297**	0,318	0,400
	(0,090)	(0,093)	(0,073)
	p-valor:	$2{,}2\text{x}10^{-16}$	$2{,}2\text{x}10^{-16}$
	GR:	6,6%	25,7%
EMQCR	**0,289**	0,310	0,395
	(0,092)	(0,095)	(0,075)
	p-valor:	$2{,}2\text{x}10^{-16}$	$2{,}2\text{x}10^{-16}$
	GR:	6,7%	26,8%

Tabela 11 – Resultados para o cenário de dados Normal Inversa: dados simulados com 10 vértices.

Medida	**PMLG**	**PRL**	**PBIVAR**
EMQDV	**0,728**	0,781	0,785
	(0,138)	(0,151)	(0,154)
	p-valor:	$2{,}2\text{x}10^{-16}$	$2{,}2\text{x}10^{-16}$
	GR:	6,7%	7,2%
EMQA	**0,196**	0,200	0,206
	(0,032)	(0,033)	(0,035)
	p-valor:	$2{,}37\text{x}10^{-12}$	$2{,}77\text{x}10^{-16}$
	GR:	2,0%	4,8%
EMQAC	**0,328**	0,343	0,349
	(0,115)	(0,121)	(0,122)
	p-valor:	$2{,}2\text{x}10^{-16}$	$2{,}2\text{x}10^{-16}$
	GR:	4,4%	6,0%
EMQCR	**0,264**	0,280	0,283
	(0,126)	(0,132)	(0,133)
	p-valor:	$2{,}2\text{x}10^{-16}$	$2{,}2\text{x}10^{-16}$
	GR:	5,7%	6,7%

4.4 CONSIDERAÇÕES SOBRE O CAPÍTULO

Este capítulo apresentou uma avaliação experimental de conjuntos de dados poligonais gerados a partir de distribuições contínuas assimétricas, especificamente as distribuições Gama e Normal Inversa. A análise focou na aplicação do PMLG, comparando-o com os métodos PRL e PBIVAR, tanto no contexto de diagnóstico quanto no preditivo.

No cenário da distribuição Gama, observou-se que os dados poligonais exibiram assimetria à direita e variância não constante, características consistentes com a natureza da distribuição. O PMLG demonstrou maior eficácia na predição dos centros e raios, com resíduos menores e mais próximos de zero, especialmente em comparação ao PBIVAR, que apresentou resíduos mais dispersos. A representação poligonal com 5 e 10 vértices revelou que o aumento no número de vértices pode influenciar a dispersão dos dados, ainda que a estrutura dos centros e raios tenha sido mantida.

No cenário da distribuição Normal Inversa, os resultados corroboraram a presença de assimetria com os centros dos polígonos concentrados nas primeiras observações. O PMLG obteve destaque, mostrando menor área residual e maior precisão na predição. Notou-se também que o desvio padrão poligonal aumentou com a representação de 10 vértices, sugerindo que a complexidade da forma poligonal pode introduzir maior variabilidade nos dados.

A análise de resíduos foi fundamental para avaliar a qualidade dos modelos. Em ambos os cenários, o PMLG e o PRL apresentaram resíduos mais concentrados próximos a zero, indicando um ajuste mais adequado aos dados. Por outro lado, o PBIVAR exibiu resíduos com maior dispersão, o que pode limitar sua aplicação em contextos demaior complexidade.

No próximo capítulo, os modelos são aplicados a um conjunto de dados reais, com o objetivo de avaliar sua eficácia prática e demonstrar sua aplicabilidade em situações concretas.

5 APLICAÇÃO EM DADOS REAIS DE DISTRIBUIÇÃO CONTÍNUA ASSIMÉTRICA

Os resultados dos experimentos apresentados no Capítulo 4 ratificam a aplicabilidade da abordagem proposta em conjuntos de dados simulados provenientes de distribuições contínuas assimétricas, como Gama e Normal Inversa. Neste capítulo, amplia-se a análise com o objeito de verificar sua aplicabilidade em cenário de dados reais

A aplicação do modelo em contextos reais permite avaliar seu desempenho em situações complexas e menos controladas, oferecendo evidências sobre sua robustez e potencial de uso prático. Sabe-se que diversos problemas da sociedade originam variáveis de distribuições positivas, contínuas e assimétricas, como exemplo:

- Saúde: Estudo do tempo de sobrevivência de pacientes em função de idade, estágio da doença, tipo de tratamento, entre outros fatores.
- Meteorologia: Predição de variáveis climáticas como precipitação, velocidade do vento, temperatura e umidade relativa do ar.
- Social: Análise da renda populacional com base em características como escolaridade, ocupação, localização e idade.
- Indústria e Produção: Avaliação da resistência de materiais (exemplo: peças de alumínio) conforme a força aplicada ou composição; ou ainda, tempo até a falha de componentes eletrônicos em função do uso ou da temperatura ambiente.
- Mercado Imobiliário: Estimativa do valor de aluguel ou venda de imóveis com base no ano de construção, número de cômodos, localização, entre outras características.

Portanto, pode-se considerar os Modelos Lineares Generalizados (MLG) em função da característica da variável de interesse possuírem distribuição contínua assimétrica, ou uma variância em função média, exemplificando, tem-se a distribuição gama e normal inversa. A metodologia deste capitulo cumpre a descrição e análise dos dados, além uma sequência de simulações Monte Carlo (MC) para avaliação diagnóstica e preditiva dos modelos.

5.1 CENÁRIO DE APLICAÇÃO: DADOS DA METEOROLOGIA

O conjunto de dados contém características de três importantes cidades do Panamá - Tocumen, San Miguelito e David - , as quais incluem eletricidade utilizada, variáveis meteorológicas, além de informações do calendário escolar, como ser dia útil ou feriado (MADRID; ANTONIO, 2021). Os valores são de 2021, coletados diariamente e a cada hora, formando uma base de dados com 8,760 registros. Para este estudo, considera-se as variáveis meteorológicas da cidade Tocumen as quais podem ser definidas na Tabela 12.

Tabela 12 – Variáveis meteorológicas presentes na análise.

Data	**Precipitação**	**Temperatura**	**Umidade**	**Velocidade do vento**
01/01	0,007	24,9	0,017	22,6
01/01	0,009	24,8	0,017	23,2
01/01	0,011	24,9	0,017	23,2
⋮	⋮	⋮	⋮	⋮
31/12	0.007	29.09	0.017	20.7
31/12	0.004	28.11	0.017	17.6
31/12	0.005	26.99	0.018	13.7

O pré-processamento realizado consiste em transformar as variáveis data e hora, tornando-as características de agregação dos dados no contexto poligonal. Portanto, esta atividade resulta em um conjunto de dados agregados por dia, com 365 observações.

A Tabela 13 exibe os valores de centro e raio das quatro variáveis e a Figura 26 mostra a distribuição de centro e range, assim como a representação gráfica da variável poligonal *precipitação* de 10 vértices, sendo a variável resposta deste estudo. Percebe-se assimetria nos histogramas de centro e raio, e em relação a variável poligonal, alguns polígonos com de valores extremos para o raio. As medidas descritivas da variável resposta poligonal: média empírica poligonal $(0{,}07;\ 0{,}07)^T$ e desvio padrão empírico poligonal $(0{,}06;\ 0{,}06)^T$.

Tabela 13 – Tabela com valores de centro e raio da base de dados de meteorologia.

Dia	**Precipitação**	**Temperatura**	**Umidade**	**Velocidade do Vento**
D_1	(0,008; 0,010)	(26,845; 4,213)	(0,017; 0,001)	(22,787; 3,998)
D_2	(0,032; 0,038)	(26,368; 3,980)	(0,015; 0,001)	(18,128; 4,137)
⋮	⋮	⋮	⋮	⋮
D_{365}	(0,051; 0,063)	(27,439; 3,278)	(0,018; 0,001)	(17,291; 7,376)

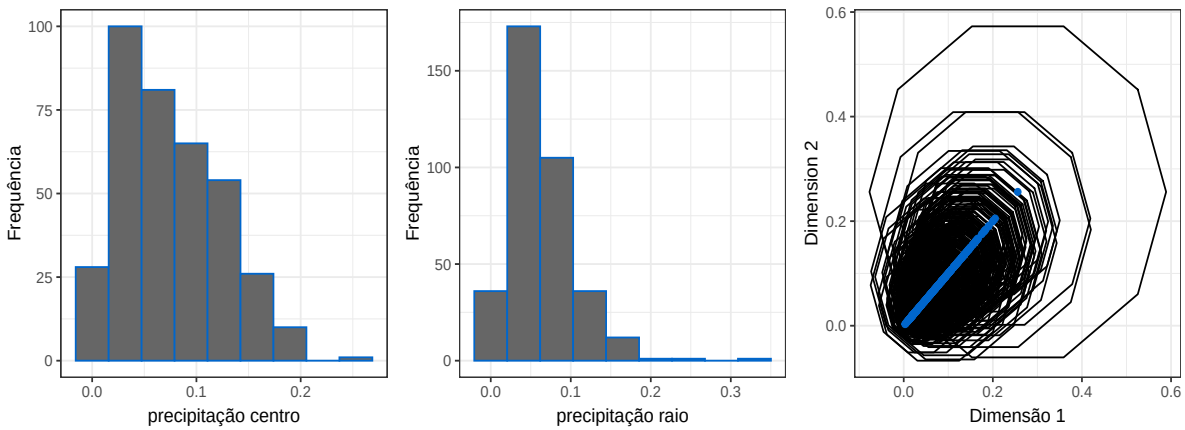


Figura 26 – Variável resposta poligonal com 10 vértices e distribuição gama.

5.2 ANÁLISE DE RESÍDUOS

Na análise residual, executam-se os três modelos para obtenção dos valores preditos e respectivos resíduos. A Figura 27 apresenta a variável predita, destacando que o modelo PMLG demonstra maior proximidade geométrica à variável observada.

Quanto aos resíduos, a Figura 28(a) exibe sua representação poligonal. Observa-se que os modelos apresentaram algumas predições insatisfatórias para o raio, resultando em polígonos com áreas mais amplas. Os histogramas da Figura 28(b) revelam maior simetria dos resíduos no modelo PMLG, com alta concentração de valores próximos de zero e ocorrência reduzida de pontos extremos, tanto para o centro quanto para o raio. A Figura 28(c) apresenta os resíduos dos centros e dos raios, evidenciando uma maior dispersão em torno do zero para o modelo PMLG.

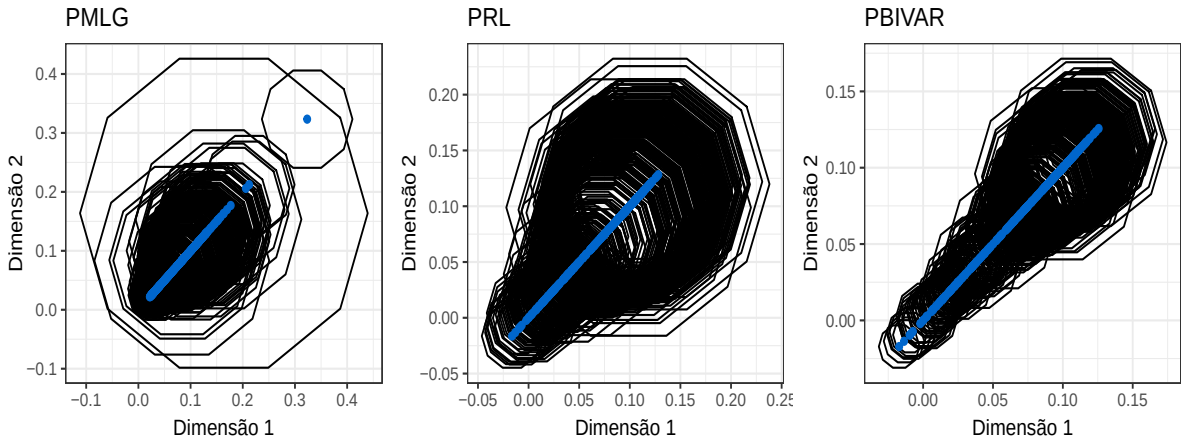
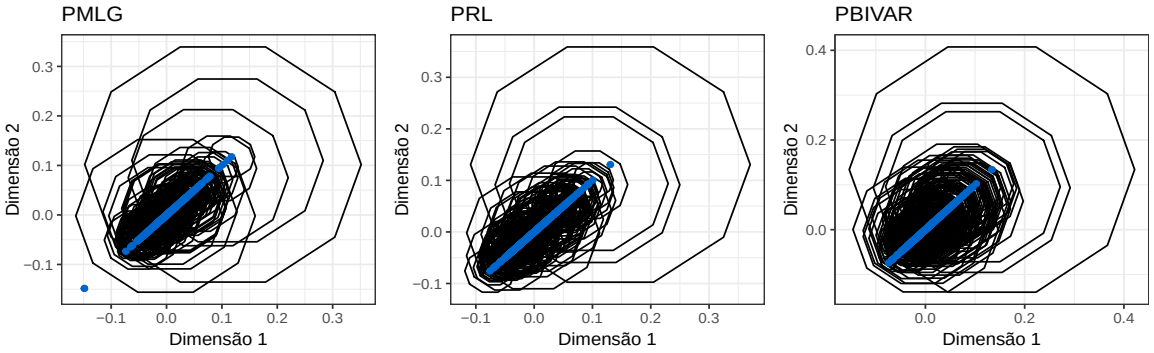
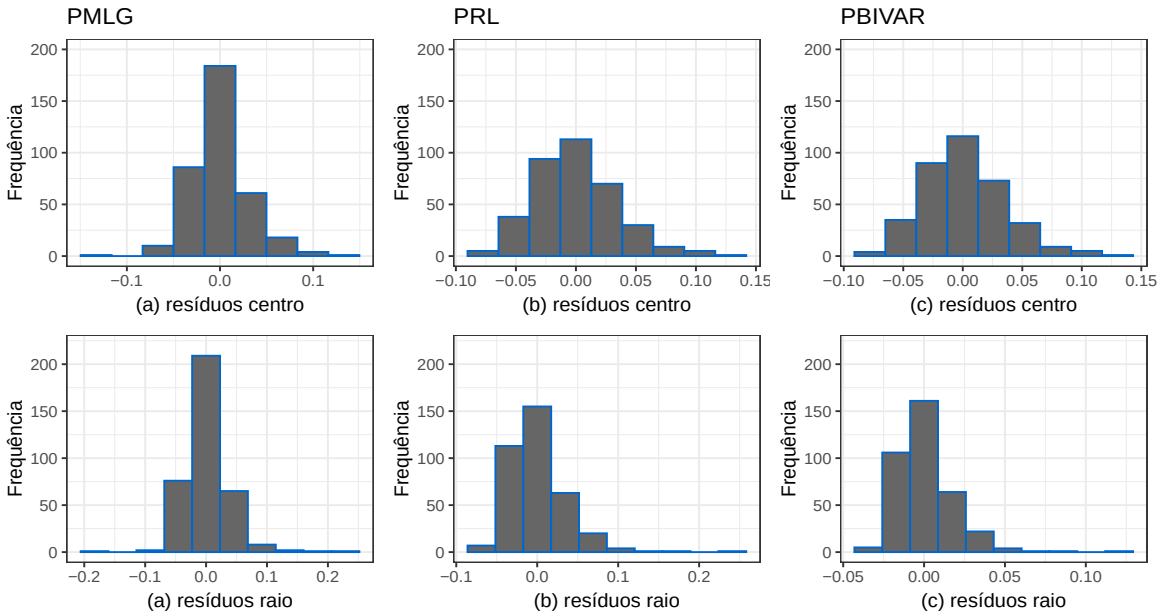


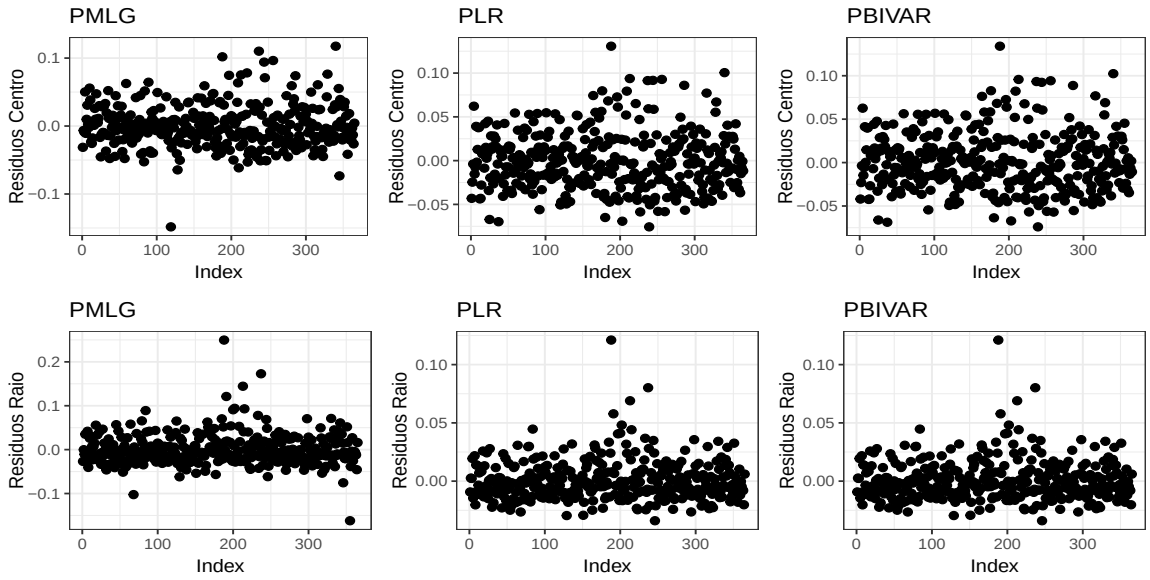
Figura 27 – Variável resposta poligonal com 10 vértices e distribuição gama.



(a) Resíduo Poligonal.



(b) Histograma dos resíduos de centro e raio.



(c) Dispersão Centro e Raio.

Figura 28 – Representação dos resíduos no cenário de dados reais e distribuição Gama.

5.3 ANÁLISE PREDITIVA

A configuração da análise preditiva é apresentada no Algoritmo 10. Aplica-se partições aleatórias *hold-out* no conjunto de dados para mensurar o desempenho dos modelos segundo as métricas de erro EMQA e EMQDV. Neste cenário, forma-se uma modelo gama com função de ligação inversa. Os resultados são apresentados na Tabela 14.

Algoritmo 10: Método Monte Carlo para Dados Reais

1: **Requerer** MC $= 100$.
2: **Requerer** número de vértices $L = 10$.
3: **Para todo** $i \leftarrow 1$ **até** MC **faça:**
4: **Particione** aleatoriamente a base de dados em conjunto de treinamento (75% dos dados) e de teste (25% dos dados).
5: **Aplique** os métodos de regressão (PMLG,PRL e PBIVAR) nos dados de treino.
6: **Aplique** a regra de predição nos dados de teste.
7: **Calcule** as medidas de desempenho usando as Equações 3.5 e 3.6.
8: **Fim Para**
9: **Calcule** a média e desvio padrão das medidas de desempenho.

Como pode ser visto, as duas métricas analisadas diferem quanto o desempenho dos modelos. A métrica EMQDV indica que o modelo PMLG, o qual é baseado em diferença dos vértices, possui menor diferença entre os polígonos observados e preditos com ganho de 5,2% e 8,4% em relação ao PBIVAR e PRL, respectivamente.

Por outro lado, a métrica EMQA baseia-se exclusivamente na diferença de área entre os polígonos, desconsiderando sua posição espacial. Como ilustrado na Figura 27, do ponto de vista geométrico, os modelos PRL e PBIVAR apresentaram maior discrepância, enquanto o modelo PMLG manteve maior aderência à forma observada.

Tabela 14 – Desempenho dos modelos de predição no cenário de dados de meteorologia.

Medida	**PMLG**	**PRL**	**PBIVAR**
EMQDV	**0,109**	0,119	0,115
	(0,066)	(0,069)	(0,069)
EMQA	0,024	0,019	**0,019**
	(0,019)	(0,014)	**(0,014)**

5.4 CONSIDERAÇÕES SOBRE O CAPÍTULO

Este capítulo explorou a aplicação do PMLG em dados reais de meteorologia, consolidando a análise iniciada com conjuntos simulados no capítulo anterior. A utilização de variáveis como precipitação, temperatura, umidade e velocidade do vento permitiu avaliar o desempenho do modelo em um contexto prático.

Os resultados demonstraram que o PMLG manteve sua eficácia mesmo em cenários menos controlados, destacando-se na predição de variáveis com distribuição assimétrica, como a precipitação. A análise de resíduos revelou que o modelo apresentou maior proximidade geométrica em relação aos dados observados, com resíduos mais simétricos e concentrados próximos de zero. Isso reforça sua robustez na predição de centros e raios.

Na avaliação preditiva, o PMLG superou os modelos PRL e PBIVAR na métrica EMQDV, que considera a diferença entre vértices. Embora a métrica EMQA, baseada em áreas, tenha mostrado desempenho semelhante entre os modelos, a análise visual confirmou que o PMLG preservou melhor a forma e a posição dos polígonos preditos.

A transformação desses dados em representações poligonais mostrou-se eficiente para capturar tendências e variações, mesmo na presença de valores extremos. Em síntese, os resultados deste capítulo validam a aplicabilidade do PMLG em problemas reais, por exemplo na meteorologia, onde variáveis assimétricas são comuns.

6 AVALIAÇÃO EXPERIMENTAL COM DADOS POLIGONAIS GERADOS A PARTIR DE DISTRIBUIÇÃO BINOMIAL

Neste Capítulo diferentes cenários de conjuntos de dados poligonais são considerados com o objetivo de avaliar o desempenho do Modelo Linear Generalizado Poligonal (PMLG) ao estimar variáveis com distribuição Binomial. A comparação envolve três regras de classificação: a primeira baseia-se na média aritmética das predições, denominada PMLG; a segunda utiliza a média ponderada otimizada das predições, denotada por PMLG_{PSO}; e a terceira emprega protótipos poligonais, referida como $\text{PMLG}_{\text{Proto}}$, utilizando três protótipos nesses experimentos. Foi conduzido um experimento de Monte Carlo (MC) com 1000 iterações para gerar conjuntos de dados com valores poligonais e avaliar o desempenho dos modelos.

O Modelo de Classificação Intervalar baseado em Probabilidade *a Posteriori* Combinada (IDPC-PP) (SOUZA; QUEIROZ; CYSNEIROS, 2011) foi comparado com as propostas poligonais. Para avaliar o desempenho dos modelos, foram utilizadas métricas de classificação, como a acurácia e a precisão por classe. A acurácia é calculada pela razão entre o número de classificações corretas e o total de observações, dada por:

$$\text{Acurácia} = \frac{TP + TN}{TP + TN + FP + FN},$$

onde TP, TN, FP e FN são, respectivamente, os verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos. Já a precisão para a classe positiva é:

$$\text{Precisão} = \frac{TP}{TP + FP}.$$

Além disso, é calculado o percentual de vitórias de cada modelo, definido como a proporção de execuções nas quais um modelo obteve o melhor desempenho em relação aos demais, com base nas métricas avaliadas. A comparação entre os algoritmos foi realizada por meio do teste de Friedman, utilizado para detectar diferenças significativas no desempenho entre múltiplos métodos, e do teste de Wilcoxon para amostras pareadas, empregado para comparações estatísticas entre pares de modelos.

6.1 CONFIGURAÇÕES DOS DADOS SIMULADOS

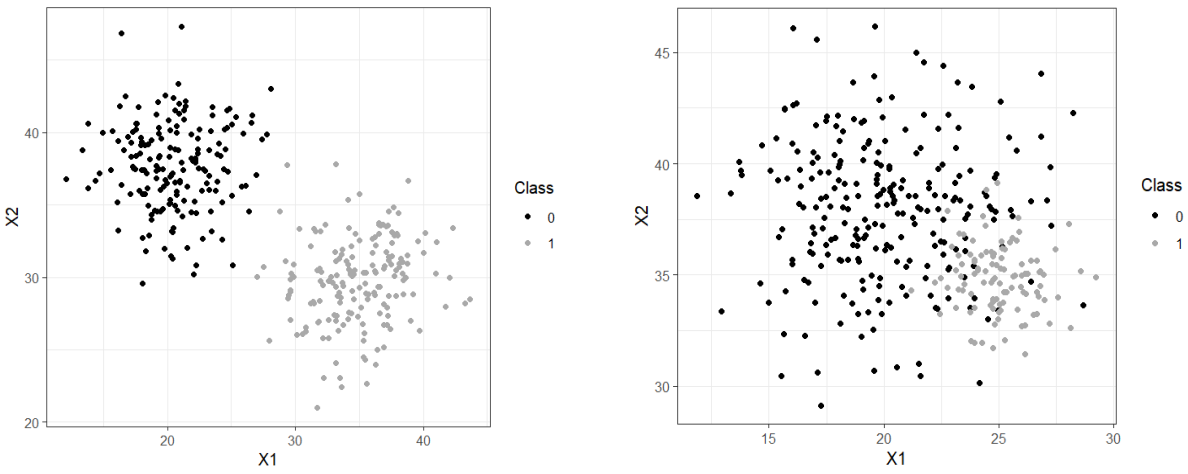
Para demonstrar a aplicabilidade da abordagem proposta, foram inicialmente construídos dois conjuntos de dados semente distintos, cada um contendo 500 observações distribuídas

em duas classes. Esses conjuntos foram gerados a partir de distribuições normais bivariadas com características específicas. Para cada conjunto de dados semente, foram aplicadas quatro configurações distintas de parâmetros, com o intuito de gerar conjuntos de dados simbólicos com diferentes níveis de complexidade e variabilidade. A Figura 29 apresenta os padrões de dispersão dos conjuntos semente, cujas configurações estão detalhadas a seguir:

1. Conjunto Semente I: Classes Balanceadas e Bem Separadas:
   - Classe 1 $(n = 250)$: $\boldsymbol{\mu} = (15, 5)^\top$, $\sigma_1^2 = 64$, $\sigma_2^2 = 9$ e $\sigma_{12} = 0$.
   - Classe 2 $(n = 250)$: $\boldsymbol{\mu} = (30, 10)^\top$, $\sigma_1^2 = 25$, $\sigma_2^2 = 36$ e $\sigma_{12} = 0$.
2. Conjunto Semente II: Classes Desbalanceadas e Sobrepostas:
   - Classe 1 $(n = 350)$: $\boldsymbol{\mu} = (20, 38)^\top$, $\sigma_1^2 = 9$, $\sigma_2^2 = 9$ e $\sigma_{12} = 0$.
   - Classe 2 $(n = 150)$: $\boldsymbol{\mu} = (25, 35)^\top$, $\sigma_1^2 = 2$, $\sigma_2^2 = 2$ e $\sigma_{12} = 0$.

A partir de cada semente bivariada $(s_1, s_2)^T$, é gerada uma classe de dados bivariados. O tamanho n de cada classe é definido segundo uma distribuição uniforme $U[15, 20]$. As unidades de cada classe $\{u_1, \ldots, u_n\}$ são geradas a partir de uma distribuição de probabilidade bivariada com componentes independentes. Dado um $n \sim U[15, 20]$, um vetor bivariado (u_1, u_2) pode ser definido da seguinte forma:

- Normal: os componentes u_1 e u_2 seguem, respectivamente, $N(s_1, \delta)$ e $N(s_2, \delta)$.
- Gama: ambos os componentes u_1 e u_2 seguem $\Gamma(\delta_1, \delta_2)$.



(a) Semente 1 (b) Semente 2

Figura 29 – Cenários de dados semente: (a) classes balanceadas e bem separadas; (b) classes desbalanceadas e sobrepostas.

A Tabela 15 apresenta os parâmetros utilizados para a geração das classes de acordo com cada distribuição de probabilidade. Esses valores de parâmetros são válidos tanto para os conjuntos sintéticos de sementes 1 quanto 2. Cada classe corresponde a um subconjunto de unidades agregadas, que pode ser descrito por dados poligonais, cujos centro e raio são utilizados na representação simbólica.

Assim, essa abordagem permite a geração de conjuntos de dados simbólicos com variabilidade controlada, possibilitando a avaliação comparativa de modelos de classificação sob diferentes condições, como classes bem separadas e classes com sobreposição. Nesse experimento cada polígono foi representado com vértices $\ell = 5$.

Tabela 15 – Parâmetros das Distribuições que geram as classes.

Normal (σ)	**Gama (k, θ)**
$\delta = 1$	$[\delta_1, \delta_2] = [1, 1]$
$\delta = 2$	$[\delta_1, \delta_2] = [1, 2]$
$\delta = 3$	$[\delta_1, \delta_2] = [1, 3]$
$\delta = 4$	$[\delta_1, \delta_2] = [4, 2]$
$\delta = 6$	$[\delta_1, \delta_2] = [9, 2]$
$\delta = 8$	$[\delta_1, \delta_2] = [16, 2]$
$\delta = 9$	$[\delta_1, \delta_2] = [81, 1]$
$\delta = 10$	$[\delta_1, \delta_2] = [25, 2]$

6.2 ANÁLISE PREDITIVA

A Tabela 16 resume o desempenho dos modelos avaliados sob diferentes níveis crescentes de variabilidade ($\sigma = 1$ a 10) em cenário de geração de dados com distribuição normal e classes balanceadas. Os resultados apresentados incluem a média e o desvio padrão da acurácia, bem como as médias de precisão por classe, calculadas a partir das réplicas de MC.

O modelo PMLG$_{\text{PSO}}$ obteve as maiores médias de acurácia, variando de $0{,}985$ $(0{,}010)$ para $\sigma = 1$ até $0{,}962$ (0,016) para $\sigma = 10$, superando tanto o modelo PMLG original quanto a variante baseada em protótipos, PMLG$_{\text{Proto3}}$. Essa superioridade também se reflete nas taxas de vitória, especialmente em cenários com maior variabilidade, nas quais o modelo PMLG$_{\text{PSO}}$ ultrapassa 70% de vitórias para $\sigma \geq 8$. Embora o modelo PMLG padrão apresente desempenho estável com acurácia inferior, o modelo PMLG$_{\text{Proto3}}$ destaca-se pela elevada precisão na classe 0. O modelo intervalar IDPC-PP, obteve os menores valores de desempenho, com acurácia

inferior a 0,91 e taxas de vitória abaixo de 2% nos cenários com $\sigma \geq 9$. Pode-se observar que, à medida que a variabilidade dos dados aumenta, o modelo intervalar reduz sua proporção de vitórias, enquanto o modelo PMLG_{PSO} apresenta um aumento correspondente.

A análise estatística realizada por meio do teste de Friedman confirma diferenças significativas entre os modelos avaliados (p-valor $< 10^{-150}$). Os testes *pos-hoc* de Wilcoxon reforçam essas evidências, indicando significância estatística em praticamente todas as comparações pareadas. Em poucos casos a diferença entre os modelos $\text{PMLG}_{\text{Proto3}}$ e IDPC-PP foi pequena, ficando próxima ao limite de significância estatística. De modo geral, os resultados evidenciam que o modelo PMLG_{PSO} apresenta maior robustez e capacidade de generalização diante de diferentes níveis de variabilidade nos dados, consolidando-se como a abordagem mais confiável entre os modelos considerados.

Tabela 16 – Normal: Média e Desvio Padrão da Acurácia e da Precisão para Cenário de Dados 1.

σ	**Modelo**	**Acurácia**	**Precisão 0**	**Precisão 1**	**Vitórias (%)**
1	PMLG	0,9774 (0,0130)	0,9776 (0,0176)	0,9778 (0,0179)	29,4%
	PMLG_{PSO}	**0,9851** (0,0101)	0,9851 (0,0142)	0,9854 (0,0142)	25,0%
	$\text{PMLG}_{\text{Proto3}}$	0,9796 (0,0122)	0,9894 (0,0127)	0,9707 (0,0203)	10,5%
	IDPC-PP	0,9815 (0,0113)	0,9815 (0,0164)	0,9822 (0,0154)	35,1%
	Teste de Friedman: $p = 3,37 \times 10^{-154}$		*Post-hoc* Wilcoxon: todas comparações p-valor $<$ 0,001		
2	PMLG	0,9763 (0,0132)	0,9764 (0,0186)	0,9769 (0,0180)	25,2%
	PMLG_{PSO}	**0,9843** (0,0104)	0,9843 (0,0151)	0,9847 (0,0143)	35,1%
	$\text{PMLG}_{\text{Proto3}}$	0,9787 (0,0120)	0,9888 (0,0128)	0,9696 (0,0201)	9,5%
	IDPC-PP	0,9794 (0,0115)	0,9792 (0,0165)	0,9802 (0,0162)	30,2%
	Teste de Friedman: p-valor $= 8,99 \times 10^{-154}$				
	Post-hoc Wilcoxon: todas comparações p-valor $< 0,001$ exceto IDPC-PP *vs* $\text{PMLG}_{\text{Proto3}}$ (p-valor$= 0,0353$)				
3	PMLG	0,9746 (0,0134)	0,9748 (0,0189)	0,9751 (0,0185)	23,1%
	PMLG_{PSO}	**0,9832** (0,0104)	0,9834 (0,0151)	0,9835 (0,0149)	43,1%
	$\text{PMLG}_{\text{Proto3}}$	0,9770 (0,0126)	0,9882 (0,0133)	0,9671 (0,0218)	9,9%
	IDPC-PP	0,9760 (0,0130)	0,9761 (0,0186)	0,9766 (0,0177)	23,9%
	Teste de Friedman: $p = 2,89 \times 10^{-167}$				
	Post-hoc Wilcoxon: todas comparações p-valor$< 0,001$ exceto IDPC-PP *vs* $\text{PMLG}_{\text{Proto3}}$ (p-valor$= 0,0072$)				
4	PMLG	0,9735 (0,0136)	0,9744 (0,0188)	0,9735 (0,0194)	22,4%
	PMLG_{PSO}	**0,9822** (0,0109)	0,9825 (0,0155)	0,9823 (0,0154)	49,7%
	$\text{PMLG}_{\text{Proto3}}$	0,9756 (0,0130)	0,9880 (0,0132)	0,9645 (0,0225)	10,7%
	IDPC-PP	0,9715 (0,0133)	0,9717 (0,0189)	0,9721 (0,0188)	17,2%
	Teste de Friedman: $p = 3,13 \times 10^{-194}$		*Post-hoc* Wilcoxon: todas comparações p-valor$< 0,001$		
6	PMLG	0,9677 (0,0148)	0,9676 (0,0208)	0,9686 (0,0198)	18,8%
	PMLG_{PSO}	**0,9780** (0,0117)	0,9781 (0,0171)	0,9784 (0,0158)	62,8%
	$\text{PMLG}_{\text{Proto3}}$	0,9699 (0,0143)	0,9860 (0,0141)	0,9557 (0,0236)	12,0%
	IDPC-PP	0,9555 (0,0174)	0,9562 (0,0238)	0,9560 (0,0230)	6,4%
	Teste de Friedman: p-valor$< 1 \times 10^{-300}$		*Post-hoc* Wilcoxon: todas comparações p-valor$< 0,001$		
8	PMLG	0,9591 (0,0167)	0,9602 (0,0228)	0,9592 (0,0241)	15,4%
	PMLG_{PSO}	**0,9714** (0,0136)	0,9722 (0,0191)	0,9714 (0,0200)	71,5%
	$\text{PMLG}_{\text{Proto3}}$	0,9614 (0,0158)	0,9824 (0,0160)	0,9433 (0,0271)	11,3%
	IDPC-PP	0,9337 (0,0201)	0,9347 (0,0268)	0,9343 (0,0281)	1,8%
	Teste de Friedman: p-valor$< 1 \times 10^{-300}$		*Post-hoc* Wilcoxon: todas comparações p-valor$< 0,001$		
9	PMLG	0,9541 (0,0179)	0,9549 (0,0248)	0,9545 (0,0246)	14,9%
	PMLG_{PSO}	**0,9670** (0,0148)	0,9678 (0,0209)	0,9671 (0,0212)	71,8%
	$\text{PMLG}_{\text{Proto3}}$	0,9556 (0,0175)	0,9800 (0,0173)	0,9348 (0,0287)	12,3%
	IDPC-PP	0,9204 (0,0225)	0,9219 (0,0300)	0,9209 (0,0298)	1,0%
	Teste de Friedman: p-valor$< 1 \times 10^{-300}$				
	Post-hoc Wilcoxon: todas comparações $p < 0,001$ exceto PMLG *vs* $\text{PMLG}_{\text{Proto3}}$ (p-valor$= 0,00248$)				
10	PMLG	0,9477 (0,0196)	0,9480 (0,0264)	0,9486 (0,0251)	13,2%
	PMLG_{PSO}	**0,9621** (0,0162)	0,9622 (0,0224)	0,9629 (0,0216)	74,0%
	$\text{PMLG}_{\text{Proto3}}$	0,9500 (0,0180)	0,9771 (0,0186)	0,9270 (0,0287)	12,2%
	IDPC-PP	0,9078 (0,0242)	0,9088 (0,0309)	0,9088 (0,0321)	0,6%
	Teste de Friedman: p-valor$< 1 \times 10^{-300}$				
	Post-hoc Wilcoxon: todas comparações $p < 0,001$ exceto PMLG *vs* $\text{PMLG}_{\text{Proto3}}$ (p-valor$= 0,00009$)				

A Tabela 17 resume o desempenho dos modelos avaliados em cenários com classes desbalanceadas e sobrepostas. À medida que a variabilidade aumenta ($\sigma = 1$ até $\sigma = 10$), todos os modelos apresentam queda no desempenho. Contudo, o modelo PMLG_{PSO} alcança a maior acurácia e domina as taxas de vitória, ultrapassando 93% nos casos mais difíceis, confirmando sua capacidade de generalização neste cenário.

Em contraste, o modelo padrão PMLG mantém uma acurácia moderada, porém é superado pelo PMLG_{PSO}. O $\text{PMLG}_{\text{Proto3}}$ obtém alta precisão para a classe 0, mas sofre uma queda na precisão para a classe 1. O método intervalar IDPC-PP apresenta o pior desempenho geral, com quedas significativas em acurácia e precisão à medida que σ aumenta. Testes estatísticos (Friedman e Wilcoxon) confirmam diferenças significativas entre os métodos, ressaltando a vantagem de estratégias otimizadas como o PMLG_{PSO} em problemas de classificação complexos e desbalanceados.

A Tabela 18 apresenta os resultados de classificação para conjuntos simbólicos gerados a partir de distribuições Gama sob diferentes cenários de dispersão $[\sigma_1, \sigma_2]$. Em todas as configurações, o modelo PMLG_{PSO} apresenta a maior acurácia, com valores variando de 0,9850 (0,0099) para $[1, 1]$ até 0,9626 (0,0154) para $[25, 2]$. Sua superioridade é reforçada pelas taxas de vitória, que ultrapassam 70% nos cenários com variâncias altamente desbalanceadas entre classes, como $[25, 2]$, $[8, 1]$ e $[16, 2]$.

O modelo PMLG demonstra desempenho estável, com pequena redução na acurácia conforme a dispersão aumenta, embora se mantenha inferior ao PMLG_{PSO} em todos os cenários. O $\text{PMLG}_{\text{Proto3}}$ mantém a melhor precisão para a classe 0 em quase todas as configurações, mas apresenta queda na precisão para a classe 1 à medida que a variabilidade das classes aumenta, indicando sensibilidade ao ruído assimétrico e menor robustez em cenários mais complexos.

O método IDPC-PP tem desempenho inferior aos demais, especialmente em configurações de alta variância, com acurácias abaixo de 0,92 e taxas de vitória inferiores a 3% nos cenários mais complexos. Os testes de Friedman revelam diferenças significativas entre os modelos (p-valor$< 10^{-140}$), enquanto os testes *post-hoc* de Wilcoxon confirmam a significância na quase totalidade das comparações. As exceções ocorrem em comparações entre IDPC-PP e $\text{PMLG}_{\text{Proto3}}$ para configurações de baixa dispersão (ex.: $[1, 3]$, $[4, 2]$). De modo geral, o PMLG_{PSO} demonstra o comportamento mais robusto e preciso, validando sua eficácia para lidar com dados simbólicos de distribuição Gama e dispersões variadas entre classes.

Tabela 17 – Normal: Média e Desvio Padrão da Acurácia e da Precisão para Cenário de Dados 2.

σ	**Modelo**	**Acurácia**	**Precisão 0**	**Precisão 1**	**Vitórias (%)**
1	PMLG	0,8815 (0,0258)	0,9030 (0,0268)	0,8305 (0,0522)	3,1%
	PMLG_{PSO}	**0,9071** (0,0234)	0,9341 (0,0243)	0,8482 (0,0480)	53,1%
	$\text{PMLG}_{\text{Proto3}}$	0,8894 (0,0258)	0,9820 (0,0145)	0,7472 (0,0487)	16,8%
	IDPC-PP	0,8991 (0,0245)	0,9333 (0,0244)	0,8263 (0,0492)	27,0%
	Teste de Friedman: p-valor$= 2,31 \times 10^{-267}$		*Post-hoc* Wilcoxon: todas comparações p-valor$< 0,001$		
2	PMLG	0,8786 (0,0265)	0,8988 (0,0261)	0,8304 (0,0566)	2,7%
	PMLG_{PSO}	**0,9049** (0,0233)	0,9311 (0,0238)	0,8479 (0,0515)	65,1%
	$\text{PMLG}_{\text{Proto3}}$	0,8847 (0,0273)	0,9807 (0,0152)	0,7396 (0,0512)	13,5%
	IDPC-PP	0,8902 (0,0251)	0,9230 (0,0240)	0,8189 (0,0538)	18,7%
	Teste de Friedman: p-valor$= 3,68 \times 10^{-259}$		*Post-hoc* Wilcoxon: todas comparações p-valor$< 0,001$		
3	PMLG	0,8719 (0,0279)	0,8922 (0,0274)	0,8222 (0,0577)	2,5%
	PMLG_{PSO}	**0,8989** (0,0240)	0,9247 (0,0245)	0,8412 (0,0503)	72,0%
	$\text{PMLG}_{\text{Proto3}}$	0,8779 (0,0278)	0,9789 (0,0155)	0,7279 (0,0496)	15,5%
	IDPC-PP	0,8753 (0,0260)	0,9057 (0,0251)	0,8057 (0,0538)	10,0%
	Teste de Friedman: p-valor$= 7,75 \times 10^{-266}$ *Post-hoc* Wilcoxon: todas comparações p-valor$< 0,001$ exceto IDPC-PP *vs* $\text{PMLG}_{\text{Proto3}}$ (p-valor$= 0,00207$)				
4	PMLG	0,8649 (0,0287)	0,8843 (0,0274)	0,8160 (0,0592)	3,8%
	PMLG_{PSO}	**0,8923** (0,0260)	0,9187 (0,0253)	0,8330 (0,0538)	78,6%
	$\text{PMLG}_{\text{Proto3}}$	0,8664 (0,0298)	0,9768 (0,0166)	0,7086 (0,0500)	11,9%
	IDPC-PP	0,8563 (0,0282)	0,8834 (0,0268)	0,7891 (0,0574)	5,7%
	Teste de Friedman: p-valor$= 1,83 \times 10^{-310}$ *Post-hoc* Wilcoxon: todas comparações p-valor$< 0,001$ exceto PMLG *vs* $\text{PMLG}_{\text{Proto3}}$ (p-valor$= 0,10205$)				
6	PMLG	0,8487 (0,0279)	0,8676 (0,0272)	0,7992 (0,0626)	3,6%
	PMLG_{PSO}	**0,8771** (0,0259)	0,9028 (0,0264)	0,8174 (0,0544)	84,2%
	$\text{PMLG}_{\text{Proto3}}$	0,8431 (0,0325)	0,9721 (0,0186)	0,6724 (0,0506)	10,7%
	IDPC-PP	0,8191 (0,0303)	0,8370 (0,0275)	0,7651 (0,0700)	1,5%
	Teste de Friedman: p-valor$< 1 \times 10^{-300}$ *Post-hoc* Wilcoxon: todas comparações p-valor$< 0,001$ exceto PMLG *vs* $\text{PMLG}_{\text{Proto3}}$ (p-valor$= 0,00001$)				
8	PMLG	0,8257 (0,0289)	0,8441 (0,0267)	0,7716 (0,0649)	2,6%
	PMLG_{PSO}	**0,8553** (0,0271)	0,8814 (0,0285)	0,7910 (0,0558)	89,4%
	$\text{PMLG}_{\text{Proto3}}$	0,8109 (0,0342)	0,9630 (0,0219)	0,6276 (0,0478)	7,5%
	IDPC-PP	0,7825 (0,0278)	0,7934 (0,0234)	0,7385 (0,0794)	0,5%
	Teste de Friedman: p-valor$< 1 \times 10^{-300}$		*Post-hoc* Wilcoxon: todas comparações p-valor$< 0,001$		
9	PMLG	0,8154 (0,0291)	0,8334 (0,0270)	0,7606 (0,0687)	3,4%
	PMLG_{PSO}	**0,8457** (0,0275)	0,8712 (0,0274)	0,7811 (0,0602)	89,5%
	$\text{PMLG}_{\text{Proto3}}$	0,7935 (0,0359)	0,9600 (0,0222)	0,6046 (0,0468)	6,1%
	IDPC-PP	0,7699 (0,0281)	0,7786 (0,0236)	0,7304 (0,0898)	1,0%
	Teste de Friedman: p-valor$< 1 \times 10^{-300}$		*Post-hoc* Wilcoxon: todas comparações p-valor$< 0,001$		
10	PMLG	0,8062 (0,0308)	0,8246 (0,0273)	0,7470 (0,0728)	3,0%
	PMLG_{PSO}	**0,8369** (0,0294)	0,8630 (0,0286)	0,7686 (0,0623)	93,8%
	$\text{PMLG}_{\text{Proto3}}$	0,7763 (0,0367)	0,9573 (0,0235)	0,5830 (0,0443)	2,8%
	IDPC-PP	0,7558 (0,0263)	0,7631 (0,0206)	0,7187 (0,1045)	0,4%
	Teste de Friedman: p-valor$< 1 \times 10^{-300}$		*Post-hoc* Wilcoxon: todas comparações p-valor$< 0,001$		

Tabela 18 – Gama: Média e Desvio Padrão da Acurácia e da Precisão para Cenário de Dados 1.

$[\sigma_1, \sigma_2]$	**Modelo**	**Acurácia**	**Precisão 0**	**Precisão 1**	**Wins (%)**
$[1, 1]$	PMLG	0,9776 (0,0129)	0,9786 (0,0167)	0,9772 (0,0181)	31,6%
	PMLG_{PSO}	**0,9850** (0,0099)	0,9855 (0,0137)	0,9849 (0,0144)	29,5%
	$\text{PMLG}_{\text{Proto3}}$	0,9796 (0,0117)	0,9896 (0,0120)	0,9706 (0,0202)	10,2%
	IDPC-PP	0,9807 (0,0109)	0,9808 (0,0154)	0,9813 (0,0163)	28,7%
	Teste de Friedman: p-valor$= 5,02 \times 10^{-143}$				
	Post-hoc Wilcoxon: todas comparações p-valor$< 0,001$ exceto IDPC-PP *vs* $\text{PMLG}_{\text{Proto3}}$ (p-valor$= 0,0019$)				
$[1, 2]$	PMLG	0,9762 (0,0135)	0,9762 (0,0181)	0,9770 (0,0187)	25,9%
	PMLG_{PSO}	**0,9844** (0,0104)	0,9843 (0,0147)	0,9850 (0,0149)	39,0%
	$\text{PMLG}_{\text{Proto3}}$	0,9786 (0,0124)	0,9888 (0,0128)	0,9695 (0,0212)	10,6%
	IDPC-PP	0,9783 (0,0119)	0,9782 (0,0171)	0,9790 (0,0170)	24,5%
	Teste de Friedman: p-valor$= 3,61 \times 10^{-162}$				
	Post-hoc Wilcoxon: todas comparações p-valor$< 0,001$ exceto IDPC-PP *vs* $\text{PMLG}_{\text{Proto3}}$ (p-valor$= 0,7624$)				
$[1, 3]$	PMLG	0,9750 (0,0132)	0,9756 (0,0179)	0,9751 (0,0189)	21,7%
	PMLG_{PSO}	**0,9837** (0,0104)	0,9840 (0,0145)	0,9838 (0,0153)	40,0%
	$\text{PMLG}_{\text{Proto3}}$	0,9774 (0,0129)	0,9893 (0,0122)	0,9669 (0,0225)	9,3%
	IDPC-PP	0,9779 (0,0121)	0,9790 (0,0163)	0,9774 (0,0179)	29,0%
	Teste de Friedman: p-valor$= 2,58 \times 10^{-160}$				
	Post-hoc Wilcoxon: todas comparações p-valor$< 0,001$ exceto IDPC-PP *vs* $\text{PMLG}_{\text{Proto3}}$ (p-valor$= 1,0000$)				
$[4, 2]$	PMLG	0,9734 (0,0139)	0,9740 (0,0197)	0,9735 (0,0187)	24,6%
	PMLG_{PSO}	**0,9823** (0,0109)	0,9827 (0,0156)	0,9824 (0,0151)	42,2%
	$\text{PMLG}_{\text{Proto3}}$	0,9752 (0,0126)	0,9880 (0,0134)	0,9638 (0,0212)	8,4%
	IDPC-PP	0,9752 (0,0127)	0,9755 (0,0184)	0,9755 (0,0173)	24,8%
	Teste de Friedman: p-valor$= 7,15 \times 10^{-166}$				
	Post-hoc Wilcoxon: todas comparações p-valor$< 0,001$ exceto IDPC-PP *vs* $\text{PMLG}_{\text{Proto3}}$ (p-valor$= 1,0000$)				
$[9, 2]$	PMLG	0,9666 (0,0153)	0,9668 (0,0214)	0,9674 (0,0210)	16,8%
	PMLG_{PSO}	**0,9774** (0,0123)	0,9775 (0,0179)	0,9780 (0,0172)	56,4%
	$\text{PMLG}_{\text{Proto3}}$	0,9696 (0,0148)	0,9854 (0,0147)	0,9557 (0,0250)	12,5%
	IDPC-PP	0,9630 (0,0158)	0,9636 (0,0220)	0,9634 (0,0221)	14,3%
	Teste de Friedman: p-valor$= 5,38 \times 10^{-231}$		*Post-hoc* Wilcoxon: todas comparações p-valor$< 0,001$		
$[16, 2]$	PMLG	0,9592 (0,0166)	0,9598 (0,0231)	0,9598 (0,0229)	15,2%
	PMLG_{PSO}	**0,9716** (0,0134)	0,9721 (0,0187)	0,9718 (0,0194)	65,7%
	$\text{PMLG}_{\text{Proto3}}$	0,9621 (0,0160)	0,9826 (0,0154)	0,9443 (0,0266)	12,7%
	IDPC-PP	0,9453 (0,0189)	0,9463 (0,0257)	0,9456 (0,0260)	6,4%
	Teste de Friedman: p-valor$< 1 \times 10^{-300}$		*Post-hoc* Wilcoxon: todas comparações p-valor$< 0,001$		
$[8, 1]$	PMLG	0,9534 (0,0180)	0,9534 (0,0250)	0,9546 (0,0243)	14,0%
	PMLG_{PSO}	**0,9664** (0,0146)	0,9667 (0,0208)	0,9671 (0,0207)	69,7%
	$\text{PMLG}_{\text{Proto3}}$	0,9557 (0,0173)	0,9798 (0,0169)	0,9351 (0,0281)	13,9%
	IDPC-PP	0,9250 (0,0221)	0,9260 (0,0300)	0,9260 (0,0299)	2,4%
	Teste de Friedman: p-valor$< 1 \times 10^{-300}$				
	Post-hoc Wilcoxon: todas comparações p-valor$< 0,001$ exceto PMLG *vs* $\text{PMLG}_{\text{Proto3}}$ (p-valor$= 0,0000$)				
$[25, 2]$	PMLG	0,9488 (0,0185)	0,9492 (0,0239)	0,9496 (0,0255)	14,8%
	PMLG_{PSO}	**0,9626** (0,0154)	0,9631 (0,0212)	0,9630 (0,0222)	72,4%
	$\text{PMLG}_{\text{Proto3}}$	0,9503 (0,0183)	0,9775 (0,0177)	0,9273 (0,0296)	11,4%
	IDPC-PP	0,9199 (0,0220)	0,9208 (0,0297)	0,9210 (0,0298)	1,4%
	Teste de Friedman: p-valor$< 1 \times 10^{-300}$				
	Post-hoc Wilcoxon: todas comparações p-valor$< 0,001$ exceto PMLG *vs* $\text{PMLG}_{\text{Proto3}}$ (p-valor$= 0,0156$)				

A Tabela 19 apresenta o desempenho dos modelos avaliados sob condições de classes com sobreposição e desbalanceamento, oriundas da distribuição Gama. Em todos os cenários, o modelo PMLG_{PSO} supera os demais métodos, atingindo as maiores acurácias (por exemplo, 0,9069 (0,0232) para $[1, 1]$ e 0,8372 (0,0286) para $[25, 2]$) e dominando na taxa de vitórias, superando 80% nos casos de maior variância como $[8, 1]$ e $[25, 2]$. Seus valores de precisão equilibrada entre as classes indicam melhor capacidade de generalização e robustez frente a dados ruidosos e desbalanceados.

Em contraste, os modelos PMLG e IDPC-PP apresentam desempenho relativamente estável, porém inferior, com taxas de vitória raramente superiores a 4% em cenários com alta variância. O modelo $\text{PMLG}_{\text{Proto3}}$ continua obtendo as maiores precisões para a classe 0 (por exemplo, acima de 0,96 em todos os cenários), mas sofre redução da precisão da classe 1 (caindo para menos de 0,60 em $[25, 2]$), o que compromete a acurácia geral. Testes estatísticos (Friedman e Wilcoxon) confirmam diferenças significativas entre os modelos em todos os cenários (p-valor$< 10^{-300}$), reforçando a superioridade do PMLG_{PSO} na abordagem de dados simbólicos desbalanceados e com sobreposição. Observa-se também que, com o aumento da variabilidade, os ganhos obtidos em dados poligonais com distribuição assimétrica são maiores do que os observados no cenário de distribuição normal.

Tabela 19 – Gama: Média e Desvio Padrão da Acurácia e da Precisão para Cenário de Dados 2.

$[\sigma_1, \sigma_2]$	**Modelo**	**Acurácia**	**Precisão 0**	**Precisão 1**	**Vitórias (%)**
$[1, 1]$	PMLG	0,8819 (0,0260)	0,9033 (0,0267)	0,8314 (0,0545)	3,4%
	PMLG_{PSO}	**0,9069** (0,0232)	0,9337 (0,0243)	0,8487 (0,0496)	59,4%
	$\text{PMLG}_{\text{Proto3}}$	0,8901 (0,0262)	0,9818 (0,0148)	0,7487 (0,0491)	17,5%
	IDPC-PP	0,8958 (0,0246)	0,9291 (0,0241)	0,8244 (0,0519)	19,7%
	Teste de Friedman: p-valor$= 1,68 \times 10^{-258}$		*Post-hoc* Wilcoxon: todas comparações p-valor$< 0,001$		
$[1, 2]$	PMLG	0,8803 (0,0266)	0,9017 (0,0271)	0,8293 (0,0542)	2,8%
	PMLG_{PSO}	**0,9060** (0,0234)	0,9327 (0,0238)	0,8474 (0,0493)	65,7%
	$\text{PMLG}_{\text{Proto3}}$	0,8856 (0,0269)	0,9817 (0,0142)	0,7401 (0,0494)	16,9%
	IDPC-PP	0,8887 (0,0251)	0,9171 (0,0253)	0,8248 (0,0516)	14,6%
	Teste de Friedman: p-valor$= 6,94 \times 10^{-257}$ *Post-hoc* Wilcoxon: todas comparações p-valor$< 0,001$ exceto IDPC-PP *vs* $\text{PMLG}_{\text{Proto3}}$ (p-valor$= 0,0044$)				
$[1, 3]$	PMLG	0,8723 (0,0278)	0,8935 (0,0280)	0,8207 (0,0565)	3,4%
	PMLG_{PSO}	**0,8992** (0,0247)	0,9264 (0,0251)	0,8393 (0,0512)	65,0%
	$\text{PMLG}_{\text{Proto3}}$	0,8780 (0,0271)	0,9796 (0,0154)	0,7277 (0,0495)	15,1%
	IDPC-PP	0,8821 (0,0268)	0,9040 (0,0275)	0,8304 (0,0542)	16,5%
	Teste de Friedman: p-valor$= 1,91 \times 10^{-259}$ *Post-hoc* Wilcoxon: todas comparações p-valor$< 0,001$ exceto IDPC-PP *vs* $\text{PMLG}_{\text{Proto3}}$ (p-valor$= 0,00001$)				
$[4, 2]$	PMLG	0,8646 (0,0277)	0,8850 (0,0278)	0,8131 (0,0558)	2,5%
	PMLG_{PSO}	**0,8927** (0,0248)	0,9198 (0,0257)	0,8319 (0,0502)	67,5%
	$\text{PMLG}_{\text{Proto3}}$	0,8687 (0,0283)	0,9782 (0,0160)	0,7114 (0,0484)	14,5%
	IDPC-PP	0,8718 (0,0265)	0,8943 (0,0257)	0,8167 (0,0552)	15,5%
	Teste de Friedman: p-valor$= 4,91 \times 10^{-250}$ *Post-hoc* Wilcoxon: todas comparações p-valor$< 0,001$ exceto PMLG *vs* $\text{PMLG}_{\text{Proto3}}$ (p-valor$= 0,00029$)				
$[9, 2]$	PMLG	0,8441 (0,0284)	0,8642 (0,0278)	0,7903 (0,0606)	2,7%
	PMLG_{PSO}	**0,8736** (0,0252)	0,9002 (0,0268)	0,8116 (0,0528)	80,6%
	$\text{PMLG}_{\text{Proto3}}$	0,8420 (0,0313)	0,9718 (0,0182)	0,6708 (0,0492)	11,8%
	IDPC-PP	0,8326 (0,0269)	0,8491 (0,0262)	0,7851 (0,0606)	4,9%
	Teste de Friedman: p-valor$= 7,21 \times 10^{-313}$ *Post-hoc* Wilcoxon: todas comparações p-valor$< 0,001$ exceto PMLG *vs* $\text{PMLG}_{\text{Proto3}}$ (p-valor$= 0,36082$)				
$[16, 2]$	PMLG	0,8271 (0,0299)	0,8446 (0,0276)	0,7756 (0,0667)	3,0%
	PMLG_{PSO}	**0,8569** (0,0273)	0,8819 (0,0275)	0,7951 (0,0581)	88,0%
	$\text{PMLG}_{\text{Proto3}}$	0,8123 (0,0343)	0,9651 (0,0212)	0,6288 (0,0479)	8,0%
	IDPC-PP	0,7953 (0,0292)	0,8065 (0,0261)	0,7548 (0,0755)	1,0%
	Teste de Friedman: p-valor$< 1 \times 10^{-300}$		*Post-hoc* Wilcoxon: todas comparações p-valor$< 0,001$		
$[8, 1]$	PMLG	0.8157 (0.0303)	0.8339 (0.0265)	0.7602 (0.0718)	3.7%
	PMLG_{PSO}	**0.8462** (0.0278)	0.8719 (0.0278)	0.7816 (0.0602)	91.9%
	$\text{PMLG}_{\text{Proto3}}$	0.7941 (0.0356)	0.9602 (0.0218)	0.6053 (0.0466)	4.0%
	IDPC-PP	0.7704 (0.0279)	0.7799 (0.0228)	0.7272 (0.0871)	0.4%
	Teste de Friedman: p-valor$< 1 \times 10^{-300}$		*Post-hoc* Wilcoxon: todas comparações p-valor< 0.001		
$[25, 2]$	PMLG	0.8079 (0.0303)	0.8254 (0.0267)	0.7524 (0.0737)	3.8%
	PMLG_{PSO}	**0.8372** (0.0286)	0.8621 (0.0281)	0.7724 (0.0639)	91.9%
	$\text{PMLG}_{\text{Proto3}}$	0.7777 (0.0373)	0.9577 (0.0234)	0.5850 (0.0458)	3.5%
	IDPC-PP	0.7654 (0.0284)	0.7729 (0.0232)	0.7293 (0.0956)	0.8%
	Teste de Friedman: p-valor$< 1 \times 10^{-300}$		*Post-hoc* Wilcoxon: todas comparações p-valor< 0.001		

6.3 CONSIDERAÇÕES SOBRE O CAPÍTULO

Este capítulo apresentou uma avaliação experimental abrangente do PMLG e suas variantes em cenários de classificação com dados poligonais, baseados na regressão logística. Os

resultados obtidos destacam o seu bom desempenho, especialmente em sua versão otimizada ($PMLG_{PSO}$), que demonstrou superioridade em termos de acurácia, precisão e capacidade de generalização em comparação aos demais métodos testados.

Nos cenários com classes balanceadas e bem separadas, o $PMLG_{PSO}$ manteve altas taxas de acurácia mesmo com o aumento da variabilidade dos dados. Essa capacidade de adaptação a diferentes níveis de dispersão é um destaque, superando não apenas o PMLG padrão, mas também o modelo baseado em protótipos ($PMLG_{Proto3}$) e o método de comparação intervalar (IDPC-PP). Essa vantagem foi confirmada por testes estatísticos. Em situações com classes desbalanceadas e sobrepostas, o $PMLG_{PSO}$ manteve os melhores valores de desempenho.

Em síntese, os experimentos realizados neste capítulo validam o $PMLG_{PSO}$ como uma ferramenta para classificação de dados poligonais, especialmente em contextos onde a variabilidade e o desbalanceamento está presente. Os resultados obtidos também abrem novas perspectivas para o desenvolvimento de métodos avançados de análise de dados simbólicos.

No capítulo seguinte, os modelos desenvolvidos são aplicados a um conjunto de dados reais com o objetivo de avaliar sua capacidade de generalização fora do ambiente simulado. Essa etapa é fundamental para verificar a utilidade prática das abordagens propostas, especialmente em cenários com variabilidade e complexidade inerentes aos dados clássicos.

7 APLICAÇÃO EM DADOS REAIS DE DISTRIBUIÇÃO BINOMIAL

Notícias falsas (*fake news*) são informações que não correspondem à realidade, mas circulam amplamente na internet devido a seu conteúdo sensacionalista, apelativo e controverso. Dada a crescente disseminação desse tipo de conteúdo, torna-se essencial identificar padrões linguísticos e estilísticos que possam auxiliar em sua detecção. Utiliza-se a base de dados *Fake.BR Corpus*, apresentada em (MONTEIRO et al., 2018), que oferece um conjunto de textos rotulados como Notícias *Fake* e Notícias Reais. Essa base foi utilizada em (SILVA et al., 2020) para ampliar sua aplicação e relatar experimentos envolvendo técnicas clássicas de aprendizado de máquina, incluindo diversas estratégias como combinação de modelos (*ensemble*).

Em (LIMA et al., 2023), o objetivo é caracterizar o comportamento das Notícias Fake e mitigar seu impacto social por meio do desenvolvimento de um modelo estatístico parcimonioso e preditivo. A abordagem baseia-se em dados estruturados e técnicas de regressão para avaliar a significância das variáveis envolvidas na detecção de notícias *fake*. Neste trabalho, considera-se os resultados apresentados por (LIMA et al., 2023) para definir as variáveis relevantes para este estudo. Utiliza-se a Análise de Dados Simbólicos (SDA) para aplicar técnicas de aprendizado estatístico e extrair conhecimento relevante.

7.1 CENÁRIO DE APLICAÇÃO: DADOS DE NOTÍCIAS *FAKE*

O conjunto de dados clássico é composto por 7.200 notícias (3.600 Notícias *Fake* e 3.600 notícias verdadeiras) publicadas entre janeiro de 2016 e janeiro de 2018 em sete fontes jornalísticas brasileiras. A base de dados original possui 26 variáveis, das quais (LIMA et al., 2023) indicam as mais relevantes, resultando em um modelo parcimonioso contendo quatro variáveis explicativas: *Tipos*, *Verbos no subjuntivo e imperativo (SI)*, *Verbos modais* e *Comprimento médio das sentenças*.

Para transformar o conjunto de dados clássico — em que cada registro representa uma única notícia — em um conjunto de dados simbólico, identificaram-se inicialmente variáveis de agregação para definir as classes simbólicas. Duas variáveis principais foram selecionadas para construir a variável poligonal: a *categoria* da notícia e a *data de publicação* (mês/ano). Cada classe simbólica, portanto, corresponde a um grupo de notícias que compartilham a mesma categoria e período de publicação.

Para cada grupo j, atribui-se a classe mais frequente (Notícias *Fake* - 1 ou Notícias Verdadeiras - 0) entre os registros agregados. Em seguida, foi gerada a variável poligonal da classe j, onde o centro do polígono corresponde à média da variável aleatória Z dentro da classe j, e o raio é definido como $2 \times sd(Z_j)$, sendo $sd(Z_j)$ o desvio padrão de Z na classe j, conforme proposto em (SILVA; SOUZA; CYSNEIROS, 2019a). Os vértices ℓ do polígono j são calculados pela Equação 2.16.

Após essa transformação, o conjunto de dados simbólico resultante contém 178 classes simbólicas (89 de Notícias *Fake* e 89 de Notícias Verdadeiras), em que cada registro poligonal representa um grupo de notícias agregadas por categoria e período de publicação, com número de vértices $\ell = 5$.

7.2 ANÁLISE DESCRITIVA

A Tabela 20 apresenta as medidas descritivas das variáveis explicativas poligonais definidas em (SILVA; SOUZA; CYSNEIROS, 2019a).

Tabela 20 – Estatísticas descritivas das variáveis poligonais de notícias por classe.

Variável	**Média**	**Desvio Padrão**
Notícias *Fake*(1)		
Comprimento médio da sentença	(15,55; 15,55)	(34,03; 34,03)
Tipos	(125,68; 125,68)	(3692,01; 3692,01)
Verbos modais	(4,80; 4,80)	(15,66; 15,66)
Verbos SI	(1,40; 1,40)	(2,91; 2,91)
Notícias Verdadeiras (0)		
Comprimento médio da sentença	(21,07; 21,07)	(22,29; 22,29)
Tipos	(511,29; 511,29)	(58317,66; 58317,66)
Verbos modais	(23,30; 23,30)	(247,63; 247,63)
Verbos SI	(6,89; 6,89)	(35,03; 35,03)

As informações extraídas dos dados revelam diferenças importantes entre Notícias *Fake* e Notícias Verdadeiras:

- Sobre a variável *Comprimento médio da sentença*, as notícias verdadeiras apresentam uma média poligonal maior (21,07; 21,07) em comparação às notícias *fake* (15,55; 15,55). A menor variabilidade nas notícias reais indica maior consistência no tamanho das sentenças.

- A variável *Tipos* indica que as notícias reais exibem uma média poligonal mais elevada $(511{,}29; 511{,}29)$ em relação às notícias *fake* $(125{,}68; 125{,}68)$. A alta variância observada na classe 0 sugere maior heterogeneidade no uso de vocabulário, refletindo um repertório lexical mais diverso e complexo. Essas características podem ser observadas na Figura 30(a)-(b), onde a escala das dimensões apresenta valores mais altos.

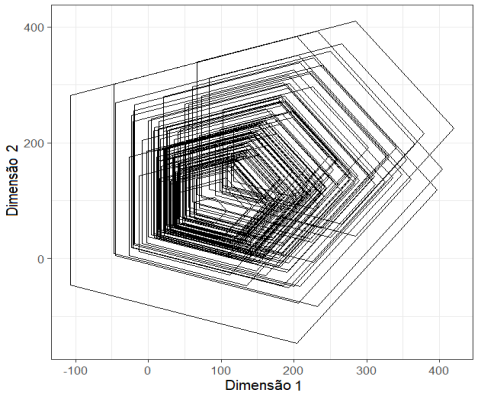
- A variável *Verbos Modais* apresenta maior média na classe Notícias Verdadeiras (23,30; 23,30), indicando construções mais frequentes que expressam possibilidade, necessidade ou permissão, típicas de textos jornalísticos formais. A Figura 30(c)-(d) mostra a distribuição poligonal com 5 vértices da variável em ambas as classes. Por fim, os *Verbos SI* também são mais comuns nas notícias reais (6,89; 6,89), reforçando a tendência das notícias falsas de evitar nuances modais e focar em declarações diretas.

A Tabela 21 apresenta os coeficientes estimados para dois modelos logísticos ajustados às representações simbólicas dos dados: um baseado nos centros e outro nos raios. O intercepto significativamente alto no modelo baseado nos centros indica uma forte tendência inicial para a predição da classe 1 (notícias *fake*), sugerindo que há uma alta probabilidade de classificação como notícia falsa mesmo sem considerar os efeitos das variáveis explicativas.

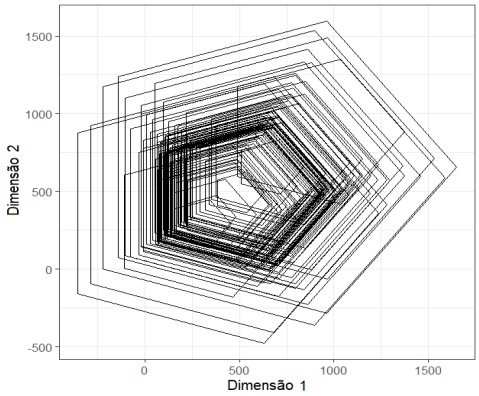
Todos os coeficientes apresentam sinal negativo, indicando que o aumento das variáveis linguísticas reduz a probabilidade de que um texto pertença à classe 1. Os coeficientes do modelo com base nos centros têm magnitude maior que os do modelo baseado nos raios, o que sugere que as médias das características linguísticas (representadas pelos centros) possuem maior influência na classificação do que suas dispersões (representadas pelos raios). Esse resultado reforça que, neste contexto, a posição central das variáveis é mais informativa para a tarefa de classificação nesta base de dados.

Tabela 21 – Coeficientes estimados dos modelos logísticos ajustados aos dados de centro e raio.

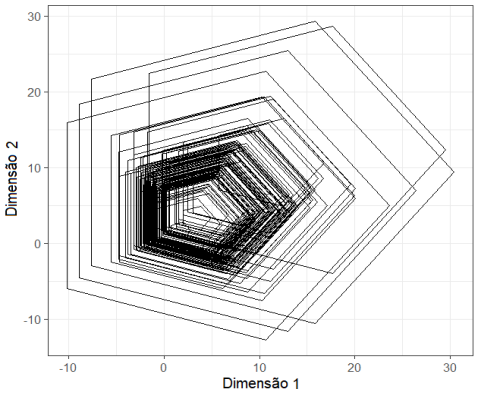
Variável	**Coeficiente (centro)**	**Coeficiente (raio)**
Intercepto	147,7853	12,1272
Tipos	-0,1312	-0,0330
Verbos SI	-11,2930	-0,3653
Verbos modais	-2,1549	-0,1128
Comprimento médio das sentenças	-2,3634	-0,1084



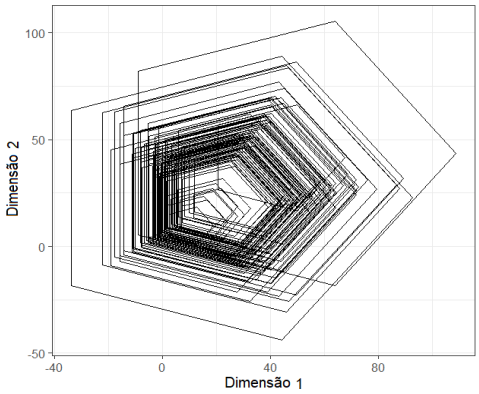
(a) Tipos (1)



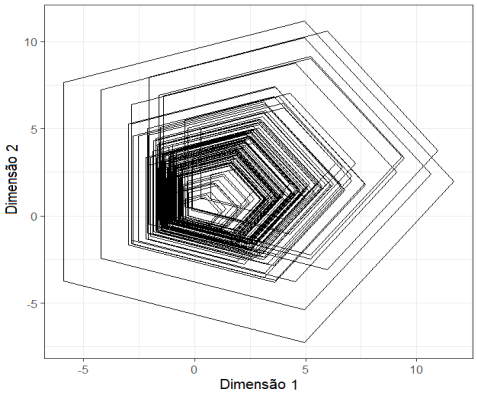
(b) Tipos (0)



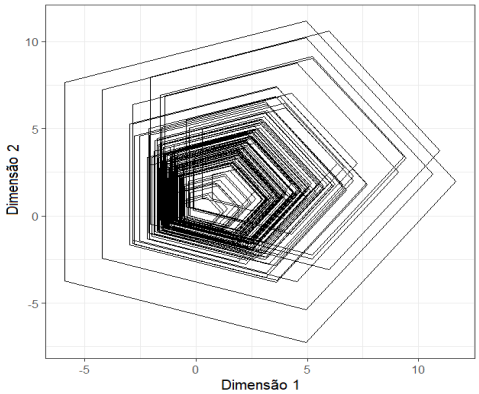
(c) Verbos Modais (1)



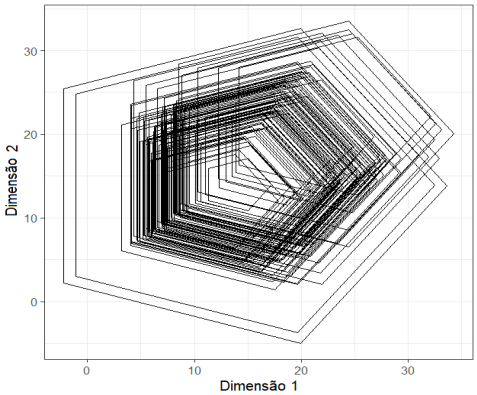
(d) Verbos Modais (0)



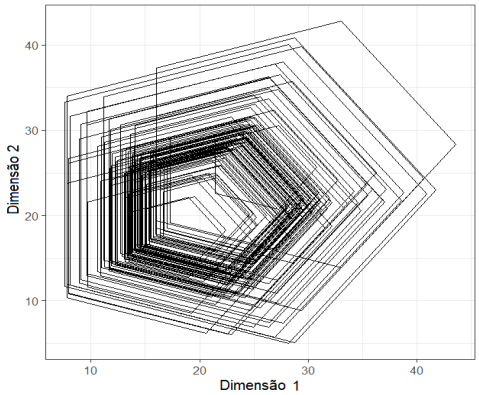
(e) Verbos SI (1)



(f) Verbos SI (0)



(g) Comprimento médio das sentenças (1)



(h) Comprimento médio das sentenças (0)

Figura 30 – Representação das variáveis poligonais *Tipos*, *Verbos modais*, *Verbos SI* e *Comprimento médio das sentenças* nas classes de Notícias *Fake* (1) e Notícias Verdadeiras (0).

7.3 ANÁLISE PREDITIVA

Com o objetivo de avaliar a capacidade preditiva do modelo PMLG proposto, foi desenvolvido um ambiente experimental, onde foi comparado o desempenho de três regras de classificação para baseados em regressão logística para dados poligonais: PMLG, PMLG_{PSO} e $\text{PMLG}_{\text{Proto}}$. Para o modelo $\text{PMLG}_{\text{Proto}}$, varia-se o número de protótipos com base em diferentes níveis de quantis: a mediana ($\text{PMLG}_{\text{Proto1}}$), os quartis ($\text{PMLG}_{\text{Proto3}}$) e os quintis ($\text{PMLG}_{\text{Proto4}}$).

Além disso, foi incluído o método tradicional para dados intervalares, o IDPC-PP, como base para comparação. A avaliação considera acurácia e precisão, buscando compreender o impacto dos diferentes modelos na tarefa de classificação. Foi realizado 100 iterações de MC, com 75% dos dados para treino e 25% para teste.

A Tabela 22 apresenta os resultados de desempenho dos modelos, com o PMLG_{PSO} apresentando a melhor acurácia (0,9944 $\pm$ 0,0134) e precisão tanto para notícias reais (Classe 0: 0,9905 $\pm$ 0,0257) quanto para notícias *fake* (Classe 1: 0,9976 $\pm$ 0,0112), superando o modelo intervalar IDPC-PP (p-valor $< 0{,}001$). A menor variabilidade (menores desvios padrão) do PMLG_{PSO} sugere maior consistência nas previsões, especialmente para notícias *fake* (Classe 1). O modelo que utiliza otimização, o PMLG_{PSO}, supera o modelo PMLG padrão, que se baseia na média aritmética direta das representações poligonais.

Tabela 22 – Média e Desvio Padrão da Acurácia e Precisão nas Classes 0 e 1 no cenário de notícias $fake$.

Modelo	Acurácia	Precisão 0	Precisão 1
PMLG	0,9891 (0,0202)	0,9849 (0,0298)	0,9929 (0,0218)
PMLG_{PSO}	**0,9944** (0,0134)	**0,9905** (0,0257)	**0,9976** (0,0112)
$\text{PMLG}_{\text{Proto1}}$	0,6823 (0,4414)	0,6664 (0,4572)	0,6958 (0,4310)
$\text{PMLG}_{\text{Proto3}}$	0,9274 (0,2351)	0,9205 (0,2510)	0,9208 (0,2571)
$\text{PMLG}_{\text{Proto4}}$	0,9935 (0,0124)	0,9926 (0,0155)	0,9941 (0,0178)
IDPC-PP	0,9847 (0,0179)	0,9802 (0,0323)	0,9887 (0,0183)

Teste de Friedman: p-valor$< 1 \times 10^{-17}$
Post-hoc Wilcoxon: PMLG_{PSO} *vs* IDPC-PP p-valor$< 1 \times 10^{-6}$

Para investigar o impacto do número de protótipos no desempenho da classificação, avaliam-se três variantes do modelo baseado em protótipos: $\text{PMLG}_{\text{Proto1}}$, $\text{PMLG}_{\text{Proto3}}$ e $\text{PMLG}_{\text{Proto4}}$. Conforme mostrado na Tabela 22, o aumento do número de protótipos resulta em melhorias de desempenho em todas as métricas. Especificamente, o $\text{PMLG}_{\text{Proto1}}$ alcança uma acurácia moderada e apresenta alta variabilidade (0,6823 $\pm$ 0,4414). Ao utilizar três protótipos

($PMLG_{Proto3}$), o desempenho do modelo aumenta, atingindo uma acurácia média de 0,9274 $\pm$ 0,2351. Os melhores resultados são obtidos com quatro protótipos ($PMLG_{Proto4}$), que alcançam alta acurácia e mínima variabilidade (0,9935 $\pm$ 0,0124).

7.4 CONSIDERAÇÕES SOBRE O CAPÍTULO

Este capítulo explorou a aplicação do PMLG e suas variantes na classificação de notícias *fakes*, utilizando dados reais. Os resultados obtidos comprovam a eficácia da abordagem simbólica em problemas de classificação complexos e de impacto na sociedade.

A análise descritiva revelou diferenças consistentes entre as classes. Notícias verdadeiras apresentaram maior diversidade lexical (variável Tipos), uso mais frequente de verbos modais e subjuntivos/imperativos (Verbos SI), além de sentenças mais longas e consistentes. Essas características refletem a natureza formal e elaborada do jornalismo de qualidade, enquanto as notícias *fakes* tenderam a simplificações e estruturas mais diretas, necessitando apelo emocional e grande compartilhamento.

Na análise preditiva, o modelo $PMLG_{PSO}$ destacou-se alcançando acurácia próxima a 99,5% e precisão equilibrada entre as classes. Pode ser observado na etapa descritiva que nessa base de dados os centros possuem maior capacidade discriminativa do que os raios, portanto, o método baseado na média ponderada superou o modelo que utiliza como regra de predição a média aritmética (PMLG). Essa superioridade em relação ao método intervalar IDPC-PP também foi estatisticamente comprovada. A versão baseada em protótipos apresentou resultados promissores, sugerindo que a representação por múltiplos protótipos pode melhorar a predição, especialmente nesse cenário com alta variabilidade interna.

8 CONCLUSÃO

Esta seção tem como objetivo apresentar as considerações finais sobre os principais tópicos abordados, incluindo as contribuições alcançadas e indicações para trabalhos futuros.

8.1 CONSIDERAÇÕES FINAIS

Para obter conhecimento a partir de grandes e complexos conjuntos de dados de diferentes contextos da sociedade, é necessário desenvolver ferramentas específicas. A Análise de Dados Simbólicos (ADS) fornece ferramentas que permitem o processamento e análise de grandes volumes de dados, podendo descrever um grupo ou classes, reduzir a dimensão e manter a confidencialidade dos dados mantendo a diversidade original. Esta abordagem tem desenvolvido diferentes métodos de predição, análise e representação de dados.

Em relação a métodos preditivos, trabalhos introduziram diferentes métodos de regressão linear em diferentes representações de dados. No entanto, consideraram-se apenas os Modelos Lineares Generalizados (MLG) para dados simbólicos tipo intervalar. Percebe-se assim a necessidade de estudos que abordem diferentes distribuições de dados, visto que em muitos contextos os dados não satisfazem as suposições do modelo linear baseado em mínimos quadrados ordinários. As variáveis podem apresentar distribuições assimétricas, contínuas e discretas. Portanto os MLG ampliam a aplicação dos modelo linear a partir de funções de ligação que relacionam a variável resposta com as explicativas.

A representação de dados simbólicos tipo poligonal é introduzida por (SILVA; SOUZA; CYSNEIROS, 2019a), desenvolvendo um modelo linear poligonal além de medidas descritivas como média, desvio padrão e histogramas para as variáveis poligonais. Uma medida de desempenho foi introduzida, a qual é baseada na diferença de áreas dos polígonos observados e preditos. No entanto, esta métrica apresenta-se incompleta pois não considera outras características, como a posição que pode constituir um grande resíduo de centro.

Portanto, a abordagem desenvolvida neste trabalho utiliza os MLG no contexto da ADS, com a variável simbólica tipo poligonal chamada de método PMLG. Os experimentos foram conduzidos utilizando dados de dois conjuntos de distribuições oriundas da exponencial: (a) distribuições contínuas assimétricas, como a Normal Inversa e Gama, e (b) distribuição Binomial.

No primeiro grupo de distribuições, além do PMLG, também é introduzido o resíduo poligonal, baseado na diferença entre os vértices dos polígonos e sendo referência para a definição da métrica de desempenho Erro Médio Quadrático da Distância dos Vértices (EMQDV), que foi comparada com métricas baseadas em centro e raio e área de polígonos. Para avaliar o método PMLG, bases de dados simuladas foram geradas e considerou-se distribuições de dados assimétricos e contínuos, sendo a Gama e Normal Inversa.

Os resultados foram comparados aos métodos Modelo de Regressão Linear Poligonal (PRL) (SILVA; SOUZA; CYSNEIROS, 2019a) e Modelo de Regressão Linear Bivariado (PBIVAR) (NETO; CORDEIRO; CARVALHO, 2011). Os resultados mostraram menores valores de erro de predição do PMLG em todos os cenários abordados. Além disso, a métrica EMQDV aumentou o ganho relativo, ou seja, evidenciou a diferença entre os modelos avaliados.

Após a avaliação experimental, considera-se dados da meteorologia que possui como variável resposta a precipitação da chuva. A partir do histograma da variável resposta, percebe-se uma distribuição assimétrica positiva dos centros e os raios dos polígonos formados, portanto, considera-se modelos Gama e função de ligação inversa. Os resultados indicam que o PMLG obteve menor valor de erro na medidas de desempenho EMQDV.

O método poligonal proposto aplicado a cenários com variável resposta de distribuição Binomial, baseia-se na regressão logística e define as regras de classificação por meio da modelagem das probabilidades associadas à variável poligonal do centro e do raio. A primeira abordagem utiliza a média das predições; a segunda implementa uma média baseada em um algoritmo de otimização, denominado PMLG_{PSO}; e a terceira propõe uma representação de classe a partir de protótipos e probabilidades, denominada $\text{PMLG}_{\text{Proto}}$.

A abordagem proposta foi validada por meio de uma série de experimentos realizados com dados sintéticos e reais. Nas simulações, foram considerados cenários com classes bem separadas e sobrepostas, geradas a partir de pontos provenientes de distribuições Gama e Normal Inversa. No cenário com dados reais, aplicou-se o método à base de notícias *fake*. Para a avaliação do desempenho, foram utilizadas as métricas de acurácia e precisão e o modelo intervalar IDPC-PP (SOUZA; QUEIROZ; CYSNEIROS, 2011) foi usado como comparação com as abordagens poligonais. Os resultados demonstram a eficácia dos métodos poligonais em comparação com o modelo intervalar. O PMLG_{PSO} apresentou desempenho superior nos cenários com dados simulados e reais, evidenciando bom ajuste à variabilidade presente nos dados.

De modo geral, os métodos propostos, aplicados a dados simbólicos poligonais, mostraram-

se eficazes na modelagem e avaliação dos modelos. Esta pesquisa contribui para o avanço da ADS, ao explorar estratégias capazes de lidar com representações poligonais oriundas de MLG. Os conhecimentos desenvolvidos podem ser aplicados a diferentes cenários, nos quais os dados são representados por polígonos e as variáveis resposta se originam de distintas distribuições.

8.2 PRINCIPAIS CONTRIBUIÇÕES

Em análise aos resultados obtidos, são elencadas as principais contribuições:

1. Elaboração de uma abordagem em ADS de MLG para dados tipo poligonal. A motivação desta contribuição se deu, principalmente, por estender a aplicabilidade de modelos preditivos nesta representação de dados simbólicos.

2. Introdução do resíduo poligonal ordinário, o qual pode ser aplicado para verificar a adequação de modelos aplicados a dados simbólicos tipo poligonal. Os resíduos podem ser analisados a partir de representação gráfica.

3. Introduzir uma medida de avaliação do erro preditivo baseada na diferença entre vértices dos polígonos. Além disso, compara-se os resultados a métricas baseadas em área e em valores de centro e raio.

4. Com os resultados dos cenários de dados reais para dados de distribuições contínuas e assimétricas (Capítulo 5) evidencia-se a importância de analisar diferentes problemas e questionamentos da sociedade. O cenário estudado considerou dados meteorológicos, no entanto, exemplifica-se outras áreas que possuem dados assimétricos positivos relevantes para estudo.

5. Introduzir e comparar regras de classificação baseadas na probabilidade *a posteriori*, obtidas por meio do modelo logístico, em cenários com dados de distribuição Binomial.

6. Destaca-se que os setores da sociedade tornam-se dependentes do conhecimento oriundo de imensos e complexos conjuntos de dados, portanto a ADS torna-se uma ferramenta de solução. Então introduzir novas representações de dados e dar novos significados a variáveis, torna-se uma das principais contribuições.

8.3 TRABALHOS FUTUROS

Para dar continuidade ao trabalho de pesquisa descrito nesta tese, lista-se, nesta seção, atividades de trabalhos futuros a serem realizadas:

- Introduzir as medidas descritivas para a distribuição de dados poligonais que não foram definidas ou exploradas na tese, como curtose e assimetria empírica.
- Os resultados apresentados mostram a análise poligonal a variáveis contínuas e discretas, os quais compreendem a família de distribuições exponenciais. Como exemplo o estudo considerou a distribuição Gama, Normal Inversa e Binomial. No entanto, outros tipos de distribuições podem ser analisadas, como Poisson. Assim, pode-se introduzir outras ferramentas de análises de dados poligonais.
- Investigar a definição do resíduo padronizado para a abordagem PMLG, visto que foi introduzido a análise a partir dos resíduos ordinários.
- Ampliar o estudo da regra baseada em protótipos, incluindo outras distâncias e estatégias de representação do protótipo.
- Verificar a formação de polígonos irregulares e a aplicação de outras distâncias na equação da EMQDV.
- Ampliar o estudo em dados reais, assim como variar funções de ligações e quantidade de vértices na geração de polígonos.

8.4 ARTIGOS PUBLICADOS DURANTE A TESE

Esta tese está associada à seguinte publicação científica, resultante da pesquisa desenvolvida ao longo do curso de doutorado, na qual foi proposto e analisado o Modelo Linear Generalizado Poligonal (PMLG), com experimentos realizados para dados contínuos assimétricos (ver Anexo A):

- do Nascimento, R.L.S., Souza, R.M.C.R., & Cysneiros, F.J.A. (2024). Generalized linear models for symbolic polygonal data. Knowledge-Based Systems. doi.org/10.1016/j.knosys.2024.111569.

A proposta desta tese também foi apresentada em congressos da área de Estatística e Análise de Dados Simbólicos:

- do Nascimento, R.L.S., Souza, R.M.C.R., & Cysneiros, F.J.A. Stacked Logistic Regression for Interval Data Classification. In: Symbolic Data Analysis Workshop, 2025, Varazdin - Croatia.

- do Nascimento, R.L.S., Souza, R.M.C.R., & Cysneiros, F.J.A. GLM For Symbolic Polygonal Data Applied To School Failure Indicator. In: SINAPE - Simpósio Brasileiro de Probabilidade e Estatística, 2024, Fortaleza - CE.

- do Nascimento, R.L.S., Souza, R.M.C.R., & Cysneiros, F.J.A. PGLM:A Regression Model Class for Symbolic Polygonal Data. In: Symbolic Data Analysis Workshop, 2023, Paris - France.

Algumas das contribuições e resultados apresentados no Capítulo 2, foram publicadas em periódico internacional. A publicação referenciada resultou de um estudo sobre dados simbólicos do tipo intervalar, introduzindo os resíduos intervalar ordinários e padronizados (ver Anexo B). Vale salientar que os dados tipo intervalar é um caso particular dos dados tipo poligonal, os quais serviram como base teórica para definições nesta tese.

- do Nascimento, R.L.S., Fagundes, R.A.A., Souza, R.M.C.R., & Cysneiros, F.J.A. (2022). Interval regression model adequacy checking and its application to estimate school dropout in Brazilian municipality educational scenario. Pattern Analysis and Applications. doi.org/10.1007/s10044-022-01093-0.

Outra publicação realizada durante o período da tese (ver Anexo C) avaliou o uso de métodos de regressão comumente empregados na literatura para estimar a evasão escolar:

- do Nascimento, R.L.S., Fagundes, R.A.A., Souza, R.M.C.R., & Cysneiros, F.J.A. (2021). Statistical Learning for Predicting School Dropout in Elementary Education: A Comparative Study. Annals of Data Science. doi.org/10.1007/s40745-021-00321-4.

REFERÊNCIAS

ABDALLA, H. B. A brief survey on big data: technologies, terminologies and data-intensive applications. *Journal of Big Data*, Springer, v. 9, n. 1, p. 107, 2022.

ARAÚJO, M. C.; SOUZA, R. M.; LIMA, R. C.; FILHO, T. M. S. An interval prototype classifier based on a parameterized distance applied to breast thermographic images. *Medical & Biological Engineering & Computing*, Springer, v. 55, n. 6, p. 873–884, 2017.

BARROS, A. P. de; CARVALHO, F. d. A. T. de; NETO, E. d. A. L. A pattern classifier for interval-valued data based on multinomial logistic regression model. In: IEEE. *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. [S.l.], 2012. p. 541–546. <https://doi.org/10.1109/ICSMC.2012.6377781>.

BERTRAND, P.; GOUPIL, F. Descriptive statistics for symbolic data. In: *Analysis of symbolic data*. Berlin, Heidelberg: Springer, 2000. p. 106–124. <https://doi.org/10.1007/978-3-642-57155-8_6>.

BILLARD, L.; DIDAY, E. Regression analysis for interval-valued data. In: *Data Analysis, Classification, and Related Methods*. Berlin, Heidelberg: Springer, 2000. p. 369–374. <https://doi.org/10.1007/978-3-642-59789-3_58>.

BILLARD, L.; DIDAY, E. Symbolic regression analysis. In: *Classification, Clustering, and Data Analysis*. Berlin, Heidelberg: Springer, 2002. p. 281–288. <https://doi.org/10.1007/978-3-642-56181-8_31>.

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

BILLARD, L.; DIDAY, E. *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. Chichester.: Wiley & Sons, 2006.

BRITO, P. Symbolic data analysis: another look at the interaction of data mining and statistics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Wiley Online Library, v. 4, n. 4, p. 281–295, 2014.

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

DIDAY, E. Thinking by classes in data science: the symbolic data analysis paradigm. *WIREs Comput Stat*, Wiley Online Library, v. 8, n. 5, p. 172–205, 2016.

DUNN, P. K.; SMYTH, G. K. *Generalized linear models with examples in R*. [S.l.]: Springer, 2018. v. 53.

FAGUNDES, R. A. A.; SOUZA, R. M. C. R.; CYSNEIROS, F. J. A. Robust regression with application to symbolic interval data. *Eng Appl Artif Intell*, Elsevier, v. 26, n. 1, p. 564–573, 2013. <https://doi.org/10.1016/j.engappai.2012.05.004>.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. *AI magazine*, v. 17, n. 3, p. 37–37, 1996.

GAD, A. G. Particle swarm optimization algorithm and its applications: a systematic review. *Archives of computational methods in engineering*, Springer, v. 29, n. 5, p. 2531–2561, 2022. <https://doi.org/10.1007/s11831-021-09694-4>.

HAN, J.; PEI, J.; TONG, H. *Data mining: concepts and techniques*. [S.l.]: Morgan Kaufmann, 2011.

HAO, P.; GUO, J. Constrained center and range joint model for interval-valued symbolic data regression. *Comput Stat Data Anal*, Elsevier, v. 116, p. 106–138, 2017. <https://doi.org/10.1016/j.csda.2017.06.005>.

HARRIS, S. J. Thoughts at large. *The Ledger (Lakeland, FL)*, 1978. P. 7D, Column 4.

IWASAKI, M.; TSUBAKI, H. A bivariate generalized linear model with an application to meteorological data analysis. *Statistical Methodology*, Elsevier, v. 2, n. 3, p. 175–190, 2005.

KENNEDY, J.; EBERHART, R. Particle swarm optimization. In: IEEE. *Proceedings of ICNN'95-international conference on neural networks*. [S.l.], 1995. v. 4, p. 1942–1948. <https://doi.org/10.1109/ICNN.1995.488968>.

LIMA, G. B.; CHAVES, T. d. M.; FREITAS, W. W.; SOUZA, R. M. de. Statistical learning from brazilian fake news. *Expert Systems*, Wiley Online Library, v. 40, n. 3, p. e13171, 2023. <https://doi.org/10.1111/exsy.13171>.

MACHADO, F. N. R. *Big data o futuro dos dados e aplicações*. [S.l.]: Saraiva Educação SA, 2018.

MADRID, E. A.; ANTONIO, N. Short-term electricity load forecasting with machine learning. *Information*, MDPI, v. 12, n. 2, p. 50, 2021.

MONTEIRO, R. A.; SANTOS, R. L.; PARDO, T. A.; ALMEIDA, T. A. D.; RUIZ, E. E.; VALE, O. A. Contributions to the study of fake news in portuguese: New corpus and automatic detection results. In: SPRINGER. *Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings 13*. [S.l.], 2018. p. 324–334. <https://doi.org/10.1007/978-3-319-99722-3_33>.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. *Introduction to linear regression analysis*. [S.l.]: John Wiley & Sons, 2012.

MYERS, R. H.; MONTGOMERY, D. C.; VINING, G. G.; ROBINSON, T. J. *Generalized linear models: with applications in engineering and the sciences*. [S.l.]: John Wiley & Sons, 2002.

NASCIMENTO, R. L. do; FAGUNDES, R. A. d. A.; SOUZA, R. M. de; CYSNEIROS, F. J. A. Interval regression model adequacy checking and its application to estimate school dropout in brazilian municipality educational scenario. *Pattern Analysis and Applications*, Springer, p. 1–21, 2022.

NETO, E. A. L.; CARVALHO, F. A. T. D. Centre and range method for fitting a linear regression model to symbolic interval data. *Comput Stat Data Anal*, Elsevier, v. 52, n. 3, p. 1500–1515, 2008. <https://doi.org/10.1016/j.csda.2007.04.014>.

NETO, E. A. L.; CARVALHO, F. A. T. D. Constrained linear regression models for symbolic interval-valued variables. *Comput Stat Data Anal*, Elsevier, v. 54, n. 2, p. 333–347, 2010. <https://doi.org/10.1016/j.csda.2009.08.010>.

NETO, E. A. L.; CARVALHO, F. A. T. D. An exponential-type kernel robust regression model for interval-valued variables. *Inf Sci*, Elsevier, v. 454, p. 419–442, 2018. <https://doi.org/10.1016/j.ins.2018.05.008>.

NETO, E. A. L.; CORDEIRO, G. M.; CARVALHO, F. A. T. D. Bivariate symbolic regression models for interval-valued variables. *J Stat Comput Simul*, Taylor & Francis, v. 81, n. 11, p. 1727–1744, 2011.

NETO, E. d. A. L.; CORDEIRO, G. M.; CARVALHO, F. A. de; ANJOS, U. U. dos; COSTA, A. G. da. Bivariate generalized linear model for interval-valued variables. In: IEEE. *2009 International Joint Conference on Neural Networks*. [S.l.], 2009. p. 2226–2229.

OUSSOUS, A.; BENJELLOUN, F.-Z.; LAHCEN, A. A.; BELFKIH, S. Big data technologies: A survey. *Journal of King Saud University-Computer and Information Sciences*, Elsevier, v. 30, n. 4, p. 431–448, 2018.

PAULA, G. A. *Modelos de regressão: com apoio computacional*. [S.l.]: IME-USP São Paulo, 2013.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria., 2020. <https://www.R-project.org/>.

RAO, T. R.; MITRA, P.; BHATT, R.; GOSWAMI, A. The big data system, components, tools, and technologies: a survey. *Knowledge and Information Systems*, Springer, v. 60, n. 3, p. 1165–1245, 2019.

REYES, D. M. A.; SOUZA, R. M. C. R.; CYSNEIROS, F. J. A. Estimating risk in capital asset pricing for interval-valued data. *Int J Bus Inf Syst*, Inderscience Publishers (IEL), v. 32, n. 4, p. 522–535, 2019. <https://doi.org/10.1504/IJBIS.2019.103795>.

SILVA, R. M.; SANTOS, R. L.; ALMEIDA, T. A.; PARDO, T. A. Towards automatically filtering fake news in portuguese. *Expert Systems with Applications*, Elsevier, v. 146, p. 113199, 2020. <https://doi.org/10.1016/j.eswa.2020.113199>.

SILVA, W. J.; SOUZA, P. J.; SOUZA, R. M.; CYSNEIROS, F. J. A. A clustering algorithm for polygonal data applied to scientific journal profiles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, v. 45, n. 11, p. 13766–13777, 2023. <https://doi.org/10.1109/TPAMI.2023.3297022>.

SILVA, W. J. F.; SOUZA, R. M. C. R.; CYSNEIROS, F. J. A. Polygonal data analysis: A new framework in symbolic data analysis. *Knowl Based Syst*, Elsevier, v. 163, p. 26–35, 2019. <https://doi.org/10.1016/j.knosys.2018.08.009>.

SILVA, W. J. F.; SOUZA, R. M. C. R.; CYSNEIROS, F. J. A. *Symbolic Polygonal Data Analysis*. [S.l.], 2019. R package version 1.3.3. Disponível em: <https://cran.r-project.org/package=psda>.

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 Comput*, Springer, p. 1–17, 2020. <https://doi.org/10.1007/s00500-020-05252-5>.

SKIENA, S. S. *The data science design manual*. [S.l.]: Springer, 2017.

SOARES, Y. M. G.; FAGUNDES, R. A. A. Interval quantile regression models based on swarm intelligence. *Appl Soft Comput*, Elsevier, v. 72, p. 474–485, 2018. <https://doi.org/10.1016/j.asoc.2018.04.061>.

SOUZA, L. C.; SOUZA, R. M. C. R.; AMARAL, G. J.; FILHO, T. M. S. A parametrized approach for linear regression of interval data. *Knowl Based Syst*, Elsevier, v. 131, p. 149–159, 2017. <https://doi.org/10.1016/j.knosys.2017.06.012>.

SOUZA, R. M. C. R.; QUEIROZ, D. C. F.; CYSNEIROS, F. J. A. Logistic regression-based pattern classifiers for symbolic interval data. *Pattern Analysis and Applications*, Springer, v. 14, p. 273–282, 2011. <https://doi.org/10.1007/s10044-011-0222-1>.

SRAKAR, A.; VECCO, M. Classification of entrepreneurial regimes: A symbolic polygonal clustering approach. In: SPRINGER. *Data Analysis and Rationality in a Complex World 16*. [S.l.], 2021. p. 261–271. <https://doi.org/10.1007/978-3-030-60104-1_29>.

STATISTA. *Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2023, with forecasts from 2024 to 2028*. 2025. Acesso em: 25 jun. 2025. Disponível em: <https://www.statista.com/statistics/871513/worldwide-data-created/>.

TIEN, J. M. Internet of things, real-time decision making, and artificial intelligence. *Annals of Data Science*, Springer, v. 4, n. 2, p. 149–178, 2017.

WITTEN, I. H.; FRANK, E.; HALL, M. A.; PAL, C. J. Practical machine learning tools and techniques. In: *Data Mining*. [S.l.: s.n.], 2005. v. 2, n. 4.

WLODARCZAK, P.; ALLY, M.; SOAR, J. Data process and analysis technologies of big data. In: *Networking for Big Data*. [S.l.]: Chapman and Hall/CRC, 2015. p. 103–119.

XU, W. *Symbolic data analysis: interval-valued data regression*. Tese (Doutorado) — University of Georgia Athens, GA, 2010.

ANEXO A – PUBLICAÇÃO 1

Knowledge-Based Systems 290 (2024) 111569

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Generalized linear models for symbolic polygonal data

Rafaella L.S. do Nascimento [a], Renata M.C.R. de Souza [a,*], Francisco José de A. Cysneiros [b]

[a] Centro de Informática, Universidade Federal de Pernambuco, Brazil
[b] Departamento de Estatística, Universidade Federal de Pernambuco, Brazil

ARTICLE INFO

Keywords:
Generalized linear models
Symbolic data analysis
Polygonal data
Residual analysis

ABSTRACT

Symbolic data analysis data has provided several advances in regression models concerning the type of symbolic variable. Due to the advantages of using symbolic polygonal data, this paper introduces a linear regression approach for polygonal data based on the generalize linear model theory that provides a unified method to broad range of modeling problems for different types of response as asymmetric continuous and discrete. Ordinary polygonal residuals and a way for finding model inadequacies are presented. Moreover, a quality measure of fit for polygons is also proposed in this paper. Experimental evaluation results illustrate the usefulness of the proposed approach regarding synthetic and real polygonal data.

1. Introduction

Data sets are generated daily in various social contexts (e.g., health, government, education, finance) [1,2] and organizations from different sectors of society are increasingly dependent on the knowledge extracted from these large volumes of data. It becomes necessary to use complex models and algorithms to produce reliable and repeatable decisions and results besides discovering hidden insights through analysis of correlated data [3].

Data mining techniques provide some of the most explicit illustrations of data science principles, which is the intersection of computer science, statistics, and domains of study [4]. When the entities under analysis are aggregated units based on a specific criteria, the variability between the members of each entity may to be effectively considered and better expressed by intervals, histograms, probability distributions, lists of categorical or numerical values. In symbolic data analysis (SDA) [5] these kinds of data are called symbolic and the aggregated units are called classes.

SDA aims to extract new knowledge from data that allow to take into account variability by extending data science methods and tools to symbolic data. In this context, polygonal data have been considered as symbolic in [6] due to valuable advantages such as open new possibilities for grain change in data mining through a structure able to store more information and preserve internal variability of the entities in analysis.

Studies involving regression techniques have been widely developed in SDA, considering different types of symbolic data: interval-valued [7–16] and histogram-valued [17–20]. Most of these regression models in SDA are fitted based on the ordinary least squares method. Concerning polygonal-valued data, [6] introduced the first regression model that also uses the least squares method applied to center and radius of the polygons and this model is evaluated with Brazilian educational data [21].

It is known in the classic literature of regression analysis that the assumptions considered by the least squares method can be violated in many data behavior contexts and some of them are: homoscedastic and continuous response variable. The theory of generalized linear models (GLM) consists a class of regression models that permit to fit a linear model for response variable following different distributions [22] as, Binomial, Normal, Gamma, Poisson, Inverse Gaussian, among others.

GLM can be suitable in situations such as medical expense data (continuous data with nonconstant variance), presence of disease (binary response), degree of severity of disease (ordinal data), number of recorded disease cases (count values), and treatment duration data (skewed and positive data). These situations of response variable can be also found when using symbolic data. In this context, the main goal of this paper is to introduce a regression approach applied to polygonal data that allows an appropriate choice for modeling continuous and discrete symbolic polygonal response.

The contributions of this work are: (i) to propose a linear regression model for polygons that takes into account the random nature of the response variable; (ii) to present a definition of ordinary polygonal residual and a way for finding the model inadequacies since residual analysis plays an essential role in validating regression models; (iii) to introduce a prediction quality measure of the model for polygonal data based on the Euclidean distance between the vertices of the residual polygons. (iv) to conduct an experimental analysis to evaluate the

* Corresponding author.
E-mail addresses: rlsn@cin.ufpe.br (Rafaella L.S. do Nascimento), rmcr@cin.ufpe.br (Renata M.C.R. de Souza), cysneiros@de.ufpe.br (F.J.d.A. Cysneiros).

https://doi.org/10.1016/j.knosys.2024.111569
Received 8 June 2023; Received in revised form 29 December 2023; Accepted 23 February 2024
Available online 24 February 2024

ANEXO B – PUBLICAÇÃO 2

Pattern Analysis and Applications (2023) 26:39–59
https://doi.org/10.1007/s10044-022-01093-0

THEORETICAL ADVANCES

Interval regression model adequacy checking and its application to estimate school dropout in Brazilian municipality educational scenario

Rafaella L. S. do Nascimento[1] · Roberta A. de A. Fagundes[2] · Renata M. C. R. de Souza[1] · Francisco José A. Cysneiros[3]

Received: 14 July 2021 / Accepted: 17 June 2022 / Published online: 18 July 2022

Abstract
Interval-valued data have been commonly encountered in practice, and Symbolic Data Analysis provides a solution to the statistical treatment of these data. Regression analysis for interval-valued symbolic data is a topic that has been widely investigated in the literature of symbolic data analysis, and several models from different paradigms have been proposed. There are basic regression assumptions, and it is essential to validate them. This paper introduces an approach to check interval regression model adequacy based on residual analysis. Concepts of ordinary and standardized interval residual are presented, and graphical analysis of these residuals is also proposed. To show the usefulness of the proposed approach, an application for estimating school dropout in the scenario of Brazilian municipalities is performed. We observed some outliers from the interval residuals analysis, and interval robust regression models are more suitable for estimating school dropout.

Keywords Symbolic data analysis · Educational data · Residual · Interval-valued symbolic data · Regression

1 Introduction

In many real experiences, data can have internal variation. These data can arise in two situations. First, the original data may be naturally collected as lists, intervals or histograms. For example, by recording air temperature changes in meteorological stations throughout the day, the result is not a single value but a range of values, i.e., an interval. Second, original data can be processed, and lists, intervals or histograms can be produced. With the advent of modern computer science, the ability to generate, store and collect massive size data sets is expected in the most varied scenarios. Often, the importance of analyze these massive data sets can require the use of specific methodologies. A example is to aggregate individual observations into groups of interests, especially when characteristics of groups are of higher interest to an analyst than those of individual observations. For example, data about scientific production for analyzing research groups and not individual researchers [23]. The result is not a single value as mean or median but can also be an interval for each variable. To represent data taking into account internal variability within each observation, variables have allowed assuming new forms.

Symbolic data analysis (SDA) provides a framework where the variability observed may effectively be considered in the data representation, and methods that take it into account. Symbolic data values can be intervals, histograms, distributions, lists of values, taxonomies, etc. This kind of data is called symbolic because it is not purely numerical to express the internal variation of each concept. Symbolic data can be induced from classical data, and this type of data allows to take into account more complete and complex

✉ Renata M. C. R. de Souza
rmcrs@cin.ufpe.br

Rafaella L. S. do Nascimento
rlsn@cin.ufpe.br

Roberta A. de A. Fagundes
roberta.fagundes@upe.br

Francisco José A. Cysneiros
cysneiros@de.ufpe.br

[1] Centro de Informática, Universidade Federal de Pernambuco, Recife, Brazil

[2] Departamento de Engenharia da Computação, Universidade de Pernambuco, Recife, Brazil

[3] Departamento de Estatística, Universidade de Federal de Pernambuco, Recife, Brazil

ANEXO C – PUBLICAÇÃO 3

Annals of Data Science (2022) 9(4):801–828
https://doi.org/10.1007/s40745-021-00321-4

Statistical Learning for Predicting School Dropout in Elementary Education: A Comparative Study

Rafaella L. S. do Nascimento[1] · Roberta A. de A. Fagundes[2] · Renata M. C. R. de Souza[1]

Received: 17 June 2020 / Revised: 7 January 2021 / Accepted: 3 February 2021 /
Published online: 22 March 2021

Abstract
School dropout is a significant challenge for the education system. This phenomenon is present in different environments, modalities, and stages of education. In the Brazilian scenario, despite advances in some respects as a reduction of indexes, combating evasion is still one of the significant efforts. Identifying the factors that involve school dropout is supported by different decision support techniques such as Statistical Learning. Statistical learning consists of a method set for exploring and understanding data to establish an association between explanatory and response variables and develop an accurate model. We propose to examine the use of some regression methods commonly used in the Statistical Learning literature for estimating school dropout in the context of elementary school from the state of Pernambuco. The data involves educational indicators, and we defined phases in the study to understand, prepare, and model the data. For prediction, we apply models for estimating school dropout using kernel-based and linear regression methods. We measured the performance by the prediction error from the test data set using Mean Absolute Error and Root Mean Square Error. We considered Statistical tests to confirm the results. The findings show that kernel-based models are effective alternatives to provide greater precision in the estimation of school dropout in scope studied. The reason to explore more accurate predictive models is supporting intervening and targeting the most at-risk students of scholar dropout. The study provides knowledge about the applied scenario supporting policies to mitigate the problem.

✉ Renata M. C. R. de Souza
rmcrs@cin.ufpe.br

Rafaella L. S. do Nascimento
rlsn@cin.ufpe.br

Roberta A. de A. Fagundes
roberta.fagundes@upe.br

[1] Centro de Informática, Universidade Federal de Pernambuco, Recife, Brazil

[2] Departamento de Engenharia da Computação, Universidade de Pernambuco, Recife, Brazil