Pós-Graduação em Ciência da Computação

Kássio Camelo Ferreira da Silva

CLUSTERWISE REGRESSION PARA DADOS TIPO-INTERVALO

Universidade Federal de Pernambuco
posgraduacao@cin.ufpe.br
http://cin.ufpe.br/~posgraduacao

Recife
2019

Kássio Camelo Ferreira da Silva

CLUSTERWISE REGRESSION PARA DADOS TIPO-INTERVALO

Este trabalho foi apresentado à Pós- Graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Mestre Profissional em Ciência da Computação.

Área de Concentração: Inteligência Computacional

Orientador(a): Francisco de Assis Tenório de Carvalho
Co-orientador(a): Eufrásio de Andrade Lima Neto

Recife
2019

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

S586c Silva, Kássio Camelo Ferreira da
Clusterwise regression para dados tipo-intervalo / Kássio Camelo Ferreira da Silva. – 2019.
121 f.: il., fig., tab.

Orientador: Francisco de Assis Tenório de Carvalho.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2019.
Inclui referências e apêndice.

1. Inteligência computacional. 2. Regressão não-linear. 3. Regressão clusterwise. I. Carvalho, Francisco de Assis Tenório de (orientador). II. Título.

006.3 CDD (23. ed.) UFPE- MEI 2019-076

Kássio Camelo Ferreira da Silva

“Clusterwise Regression para Dados Tipo-Intervalo”

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Ciência da Computação.

Aprovado em: 17/04/2019.

BANCA EXAMINADORA

_____

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

_____

Prof. Dr. Marcelo Rodrigo Ferreira Portela
Departamento de Estatística / UFPB

_____

Prof. Dr. Francisco de Assis Tenório de Carvalho
Centro de Informática/UFPE
(**Orientador**)

Este trabalho é dedicado à minha família.

AGRADECIMENTOS

Agradeço a todos que colaboraram, direta e indiretamente, para a elaboração desta dissertação de mestrado:

Meu orientador, Francisco de Assis Tenório de Carvalho, cujo conhecimento e dedicação permitiram que o trabalho tivesse andamento. Devo citar também a presença do Professor Eufrásio de Andrade (DE/UFPB), que deu importantes contribuições e sugestões a este trabalho, na condição de coorientador.

Meus familiares, em particular, minha mãe Lenira, meu pai Hermano e minha irmã Joyce, por todo o apoio e compreensão. Agradeço também aos colegas de trabalho da PROPLAN-UFERSA, pelos conselhos relacionados à vida acadêmica.

Em especial, agradeço a minha esposa, Aline, que contribuiu com suas palavras de incentivo mesmo nos momentos mais difíceis.

RESUMO

Em diversas aplicações, conjuntos de dados podem ser agrupados de modo a formar intervalos, histogramas, distribuições e outras formas de representação de dados. Para esta categoria de dados, conhecida como Dados Simbólicos, apresenta-se a necessidade de técnicas estatísticas adaptadas da análise de dados clássicos. O modelo de Regressão *Clusterwise* tem como objetivo lidar com a heterogeneidade dos dados, isto é, a presença de subgrupos onde a relação entre os regressores e a variável resposta é diferente do resto da amostra. Este trabalho apresenta um modelo de Regressão Clusterwise Não Linear para o Centro e Amplitude para dados tipo-intervalo (*Interval Center and Range Clusterwise Non-Linear Regression* - iCRCNLR), baseado no algoritmo de agrupamento dinâmico (DIDAY; SIMON, 1980) e nos modelos de regressão linear e não-linear para dados tipo-intervalo (NETO; CARVALHO, 2008; NETO; CARVALHO, 2017). O método expande o caso linear de regressão clusterwise para, automaticamente, selecionar o melhor par de modelos (linear e/ou não linear) para centro e meia amplitude dos intervalos, baseado em um critério de otimização. Foram realizados estudos de simulação objetivando avaliar o desempenho do método para estimação e predição considerando 24 cenários, com diferentes estruturas de grupos para centro e amplitude dos intervalos. O estudo sobre estimação avaliou a precisão das estimativas dos parâmetros em um modelo dado, ajustados pelo algoritmo iCRCNLR. No que diz respeito à predição, um esquema de validação cruzada K-folds foi utilizado para avaliar a acurácia do iCRCNLR considerando a estimação para 1, 2 e 3 *clusters*. Três métodos foram comparados para alocar observações de teste a apenas um cluster: *k-nearest neighbors* (KNN) com distância de Hausdorff, Stacked Regressions e alocação aleatória. Por fim, foram feitas aplicações em seis conjuntos de dados reais para comparar a acurácia do iCRCNLR com a regressão *clusterwise* linear para dados tipo-intervalo, iCRCLR. Os resultados obtidos sugerem que o método iCRCNLR é adequado para uso tanto nos dados simulados quanto nos dados reais.

Palavras-chaves: Regressão Não-Linear. Regressão Clusterwise. Dados Tipo-Intervalo. Agrupamento Dinâmico.

ABSTRACT

In several applications, data sets can be grouped together to form intervals, histograms, distributions, and other forms of data representation. For this category of data, known as Symbolic Data, the need for statistical techniques adapted from the analysis of classical data is presented. The Clusterwise Regression model is intended to deal with data heterogeneity ie the presence of subgroups where the relationship between the regressors and the response variable is different from the rest of the sample. This dissertation presents a Non-Linear Center and Range Clusterwise Regressions for interval-valued data, Interval Center and Range Clusterwise Non-Linear Regression (iCRCNLR) , based on the dynamic grouping algorithm (DIDAY; SIMON, 1980) and linear and nonlinear regression models for interval-valued data (NETO; CARVALHO, 2008; NETO; CARVALHO, 2017). The method expands the linear clusterwise regression case to automatically select the best pair of models (linear and/or nonlinear) for center and half range, based on an optimization criterion. Simulation studies were performed aiming to evaluate the performance of the method for estimation and prediction considering 24 scenarios, with different structures of groups for center and range amplitude. The estimation study evaluated the accuracy of the parameter estimates of the models adjusted by the iCRCNLR algorithm. With respect to prediction, a K-folds cross-validation scheme was used to evaluate the accuracy of the iCRCNLR considering the estimation for 1, 2 and 3 clusters. Three methods were compared to allocate test observations to only one cluster: *k-nearest neighbors* (KNN) with Hausdorff distance, Stacked Regressions and random allocation. Finally, applications were made in six real datasets to compare the accuracy of iCRCNLR with the linear case, iCRCLR. The results obtained suggest that the iCRCNLR method is suitable for use in both simulated and real data.

Keywords: Non linear regression. Clusterwise regression. Interval-valued data. Dynamic clustering.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BFGS	Broyden–Fletcher–Goldfarb–Shanno Algorithm
CG	Conjugate Gradient
CRM	Center and Range Method
iCRCNLR	Interval Center and Range Clusterwise Non-Linear Regression
iCRCLR	Interval Center and Range Clusterwise Linear Regression
KNN	*K*-Nearest Neighbors
NLM	Non-Linear Method
SANN	Simulated Annealing
SR	Stacked Regressions

LISTA DE SÍMBOLOS

β	Beta (parâmetros dos modelos de regressão)
$\hat{\beta}$	Valores estimados para Beta
γ	Gama
$\in$	Pertence
δ	Delta
θ	Teta
σ	Sigma
μ	Mi
$\sum$	Somatório
ξ	Xi

SUMÁRIO

1 INTRODUÇÃO

Em análise de dados, as variáveis (numéricas ou categóricas) utilizadas para descrever os objetos usualmente apresentam valores simples. Isto significa que, para um dado objeto, uma variável assume um simples valor quantitativo ou qualitativo. No entanto, em muitas situações, a utilização de variáveis com um único valor pode ser bastante restritiva, especialmente ao analisar grupos de indivíduos, caso em que a variabilidade inerente ao grupo deve ser considerada. Desta forma, dados podem ser agregados em variáveis com múltiplos valores, intervalos, conjuntos de categorias ou histogramas. Este tipo de dado é objeto de estudo da Análise de Dados Simbólicos (*Symbolic Data Analysis* - SDA), um domínio da área de extração de conhecimentos e gerenciamento de dados (BILLARD; DIDAY, 2003; BILLARD, 2006). Este trabalho tem foco um tipo de modelo de regressão que utiliza agrupamento dinâmico para lidar como heterogeneidade nos dados, conhecida como regressão *clusterwise*, aplicados a dados simbólicos do tipo-intervalo. Este tipo de dado pode representar a imprecisão e/ou incerteza existente devida a erros de medida ou a variabilidade natural presente na natureza. Apenas o segundo caso será considerado.

A análise de regressão é uma área da Estatística desenvolvida para estimar a forma de uma relação entre uma variável dependente (Y) e um conjunto de variáveis independentes $(X_1, \ldots, X_p)$ (SEARLE; GRUBER, 2016) com base em uma função que depende de $Y, X_1, \ldots, X_p$ e um conjunto de parâmetros $\beta_0, \beta_1, \ldots, \beta_p$ que devem ser estimados. Estimar a relação entre variáveis significa (i) selecionar um modelo para descrever o tipo de relação presente nos dados e (ii) estimar seus parâmetros utilizando algum critério de otimização. A escolha de um modelo linear do tipo $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon$ no passo (i) leva ao modelo de regressão linear, em que o vetor de parâmetros $\boldsymbol{\beta}$ será estimado utilizando o vetor de dados $\boldsymbol{y}$ e a matriz de posto completo $\boldsymbol{X}$. A estimação por Mínimos Quadrados de $\boldsymbol{\beta}$, que consiste em minimizar a soma dos quadrados dos erros, é um método simples e elegante que não requer nenhum pressuposto probabilístico para o modelo.

O modelo de regressão linear é a forma mais simples de representar o relacionamento entre variáveis. No entanto, em muitos problemas, as variáveis possuem uma relação não-linear, tornando o uso da regressão linear não adequado. O procedimento algébrico para obtenção das estimativas de Mínimos Quadrados em modelos não lineares é semelhante ao adotado para o caso linear, mas em muitos casos, dependendo da função não-linear selecionada para o ajuste, as equações normais obtidas não possuem solução em forma fechada, demandando o uso de métodos de otimização iterativos para encontrar estimativas para $\boldsymbol{\beta}$.

Diversas abordagens para regressão utilizando dados tipo-intervalo tem surgido

desde o artigo seminal de Billard e Diday (BILLARD; DIDAY, 2000), em que a proposta consiste em ajustar um modelo de regressão linear para os pontos médios dos intervalos. Lima Neto e De Carvalho (NETO; CARVALHO, 2008) propuseram um modelo de regressão linear em que são feitos ajustes individuais para o centro e meia amplitude dos intervalos, independentemente. Os valores preditos $\hat{y}$ são obtidos por meio de uma combinação das estimativas para o centro e meia amplitude. Com objetivo de garantir que o limite inferior das estimativas seja menor ou igual ao limite superior, foram propostos modelos com a introdução de restrições para os parâmetros relativos à amplitude. Lima Neto and De Carvalho (NETO; CARVALHO, 2010) propuseram um modelo de regressão linear para dados tipo-intervalo com restrição utilizando técnicas de programação quadrática para estimar os parâmetros do modelo.

Gonzalez e Lin (GONZÁLEZ-RIVERA; LIN, 2013) introduziram um sistema geral dinâmico bivariado para os limites superior e inferior dos intervalos e apresentaram uma rotina em dois passos para a estimação de parâmetros. Giordani (GIORDANI, 2015) abordou o problema introduzindo retrições baseadas em Least Absolute Shrinkage and Selection Operator (LASSO) ao modelo de regressão para dados tipo-intervalo. Um modelo que adota retrições sobrepostas com objetivo de melhorar a capacidade de predição foi apresentado por Hao e Guo (HAO; GUO, 2017), sendo construído como um caso especial de um problema de Mínimos Quadrados com desigualdade (least squares with inequality - LSI).

Alguns estudos abordaram técnicas de inferência estatística considerando um suporte probabilístico da variável resposta tipo-intervalo $[Y_L, Y_U]$ (AHN et al., 2012; BRITO; SILVA, 2012; NETO; ANJOS, 2015; NETO; CORDEIRO; CARVALHO, 2011). Outros avanços neste campo abordam a adaptação de técnicas para dados tipo-intervalo, como regressão logística (SOUZA; QUEIROZ; CYSNEIROS, 2011) e uso do algoritmo *Expectation-Maximization* (EM) para estimação em dados tipo-intervalo.

Relativamente a técnicas não lineares e não paramétricas para variáveis tipo-intervalo, Fagundes *et al* (FAGUNDES; SOUZA; CYSNEIROS, 2014) introduziram um modelo de regressão kernel para dados tipo-intervalo. Jeon *et al* (JEON; AHN; PARK, 2015) propõe o uso de um estimador tipo kernel Gaussiano para aproximar a distribuição dos hiperretângulos dos intervalos de um modo não-paramétrico. Um modelo de regressão não-linear para dados tipo-intervalo (NLM) foi apresentado por Lima Neto e De Carvalho (NETO; CARVALHO, 2017), cuja abordagem é utilizada neste trabalho para ajustar os modelos não-lineares.

A regressão *clusterwise* é uma técnica útil quando os dados apresentam heterogeneidade. O objetivo é identificar tanto a partição dos dados em um número previamente especificado de *clusters* e os modelos de regressão correspondentes, um para cada *cluster*. Os modelos em cada *cluster* podem ser vistos como um modelo de mistura que utiliza estimação de máxima verossimilhança (DESARBO; CRON, 1988). A regressão *clus-*

terwise também tem sido estudada numa abordagem *fuzzy* (D'URSO; SANTORO, 2006). De um ponto de vista da análise exploratória de dados, a regressão *clusterwise* pode ser vista como uma combinação de análise de regressão e agrupamento (SPÄTH, 1979). A partir desta perspectiva, De Carvalho *et al* (CARVALHO; SAPORTA; QUEIROZ, 2010) utilizou a abordagem de centro e amplitude para adaptar a regressão linear *clusterwise* para dados simbólicos do tipo-intervalo.

Este trabalho propõe um modelo de regressão *clusterwise* não-linear para variáveis tipo-intervalo, denominada iCRCNLR (*Interval Clusterwise Non-Linear Regression*). O método proposto combina os métodos de regressão para o centro e amplitude linear (CRM (NETO; CARVALHO, 2008)) e/ou não linear (NLM (NETO; CARVALHO, 2017)) para dados tipo-intervalo com o algoritmo de agrupamento dinâmico (DIDAY; SIMON, 1980). Em comparação com o modelo de regressão *clusterwise* linear para o centro e amplitude (iCRCLR (CARVALHO; SAPORTA; QUEIROZ, 2010)), iCRCNLR é capaz de fornecer uma partição de dados tipo-intervalo em um número fixo de *clusters* e selecionar, a partir de um conjunto de modelos não lineares (com o modelo linear como caso particular), o melhor par de modelos ajustados para o centro e amplitude dos intervalos em cada *cluster*.

Investiga-se a capacidade do modelo proposto em estimar corretamente os parâmetros dos modelos de regressão dentro dos *clusters*. Além disso, é avaliada a capacidade preditiva do modelo proposto para novas observações, em comparação ao modelo de regressão *clusterwise* linear para dados tipo-intervalo (iCRCLR). No que diz respeito à predição, é colocado o problema da alocação de novas observações em um dos *clusters* obtidos pelo modelo, em termos de um problema de classificação supervisionada. Em suma, a pesquisa tem por objetivo avaliar o efeito obtido na predição ao ajustar modelos de não-lineares para os dados intervalo. Além disso, pretende-se introduzir um algoritmo que obtenha o melhor par de modelos para o centro e a amplitude em cada cluster, a partir de uma lista de funções pré-definida.

O conteúdo deste trabalho está estruturado da seguinte maneira: o Capítulo 2 apresenta os modelos de regressão linear simples e múltipla, o modelo de regressão não-linear e a regressão *clusterwise* para dados reais; o Capítulo 3 apresenta detalhes do modelo de centro e amplitude (CRM) para regressão de dados tipo-intervalo e a adaptação deste mesmo modelo para o ajuste de modelos não-lineares (NLM), bem como o modelo de regressão *clusterwise* linear para dados tipo-intervalo. No Capítulo 4, é apresentado o método proposto, iCRCNLR, bem como um algoritmo para execução do método e a demonstração de sua convergência. Ainda, são apresentados os métodos de alocação (KNN, *Stacked Regressions* ou alocação aleatória) para observações de teste em um dos *clusters* obtidos pelo algoritmo. A análise do desempenho do algoritmo proposto, comparativamente ao caso linear, é apresentada no Capítulo 5 onde experimentos feitos com dados sintéticos são apresentados e discutidos na Seção 5.1

objetivando medir a capacidade de estimação e predição em 24 cenários com diferentes características; na Seção 5.2 é feita uma comparação do desempenho de predição do novo método com o caso linear, iCRCLR. Finalmente, no Capítulo 6 são apresentadas as Conclusões e Trabalhos Futuros.

2 ANÁLISE DE REGRESSÃO: MODELOS LINEAR, NÃO LINEAR E *CLUSTERWISE* PARA DADOS REAIS

Este capítulo tem como objetivo apresentar o arcabouço geral da análise de regressão, começando pelo caso mais simples, em que há apenas uma variável explicativa. Em seguida, é apresentada a contrução do modelo de regressão múltipla, com p variáveis explicativas. A presença de relacionamento não linear entre variáveis leva à necessidade de ajustar funções não lineares. Tal procedimento é apresentado na Seção 3.2. Finalmente, é apresentada a regressão *clusterwise*, que combina elementos de regressão e *clustering*, tendo como objetivo agrupar conjuntos de dados a princípio heterogêneos em grupos homogêneos onde serão ajustados modelos lineares.

Em problemas de diversas áreas é de interesse determinar os efeitos que algumas variáveis exercem sobre outras. Pode-se obter uma função cujo exame permita ao pesquisador aprender mais sobre a verdadeira relação funcional entre os dados e os efeitos individuais ou conjuntos produzidos por mudanças em tais variáveis (SEARLE; GRUBER, 2016). Neste sentido, podem-se distinguir dois tipos de variáveis. As variáveis *preditoras* podem ser obtidas em procedimentos controlados ou observadas, mas não controladas. As mudanças nas variáveis preditoras transmitem efeitos para as *variáveis resposta*. O interesse é saber como mudanças na variável preditora afetam a resposta. A forma dos relacionamentos tratados pela análise de regressão consiste em explicar a variável resposta como a soma entre a função ajustada e um erro aleatório.

Para uma relação linear simples, ou seja, considerando apenas uma variável preditora, tem-se que o valor esperado da variável resposta y, $\mathbb{E}(y)$, é dado pelo modelo linear $\beta_0 + \beta_1 x$, ou seja, uma combinação linear de valores desconhecidos, denominados *parâmetros*, β_0 e β_1. O modelo linear possui a vantagem de ser matematicamente tratável e, de um ponto de vista prático, aplicável a diversos problemas. Não obstante, qualquer função não linear em β_0 e β_1 poderia ser utilizada para descrever a relação entre as variáveis.

2.1 REGRESSÃO LINEAR SIMPLES E MÚLTIPLA

Para o i-ésimo elemento de uma amostra, tem-se que

$$\mathbb{E}(y_i) = \beta_0 + \beta_1 x_i \tag{2.1}$$

Seja $y_i - \mathbb{E}(y_i)$ a diferença entre o valor observado y_i e o seu valor esperado. A partir disto, utilizando (2.1), tem-se que

$$e_i = y_i - \mathbb{E}(y_i) = y_i - \beta_0 - \beta_1 x_i \tag{2.2}$$

e portanto, podemos obter a equação do modelo de regressão linear simples como

$$y_i = \beta_0 + \beta_1 x_i + e_i, \ \ i = 1, \ldots, n. \tag{2.3}$$

Os desvios definidos em (2.2) medem o quanto os valores observados da variável resposta diferem do valor esperado $\mathbb{E}(y_i) = \beta_0 + \beta_1 x_i$. Desta forma, os e's incluem erros tanto de medida quanto resultantes das deficiências do modelo em si, isto é, o modelo poderia ter variáveis preditoras que não foram incluídas, mas que descrevem bem a variável resposta. Os erros e são considerados variáveis aleatórias.

Até aqui foi apresentada apenas a equação do modelo. Para que ele esteja completo, é necessário definir algumas características dos erros e. Para cada i, define-se que o valor esperado do erro e é zero, sua variância é σ^2 e a covariância entre os pares e_j e e_k, $j, k \in 1, \ldots, n$, $j \neq k$ é zero:

$$\mathbb{E}(e_i) = 0, \ \ \forall i \tag{2.4}$$

$$Var(e_i) = \mathbb{E}[e_i - \mathbb{E}(e_i)]^2 = \mathbb{E}(e_i^2) = \sigma^2, \ \ \forall i \tag{2.5}$$

$$Cov(e_j, e_k) = \mathbb{E}[e_j - \mathbb{E}(e_j)][e_k - \mathbb{E}(e_k)] = \mathbb{E}(e_j e_k) = 0, \ \ \forall j \neq k. \tag{2.6}$$

Assim, o modelo é composto pelas equações (2.1) - (2.6). O proximo passo é, a partir destas definições, obter as estimativas dos parâmetros do modelo. O procedimento mais utilizado para este fim é o Método dos Mínimos Quadrados, que consiste em minimizar a soma dos quadrados dos desvios dos y_i observados em relação aos seus valores esperados $\mathbb{E}(y_i)$:

$$\boldsymbol{e}^\top \boldsymbol{e} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - \mathbb{E}(y_i)]^2 = \sum_{i=1}^{n} [y_i - \beta_0 - \beta_1 x_i]^2 \tag{2.7}$$

O procedimento consiste em derivar a Equação (2.7) em relação a β_0 e β_1 e igualar a zero, obtendo um sistema de equações normais, cuja solução para β_0 e β_1 corresponde às suas estimativas $\hat{\beta_0}$ e $\hat{\beta_1}$:

$$\frac{\partial \boldsymbol{e}^\top \boldsymbol{e}}{\partial \beta_0} = -2 \left(\sum_{i=1}^{n} y_i - N\beta_0 - \beta_1 \sum x_i \right) \tag{2.8}$$

$$\frac{\partial \boldsymbol{e}^\top \boldsymbol{e}}{\partial \beta_1} = -2 \left(\sum_{i=1}^{n} x_i y_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 \right) \tag{2.9}$$

Igualando (2.8) e (2.9) a zero, temos

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \hat{\beta}_0 \sum_{i=1}^{n} x_i y_i.$$

Finalmente, a solução para as equações normais é dada por:

$$\hat{\beta_1} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum_{i=1}^{n} x^2 - n\overline{x}^2} \tag{2.10}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = (y - \hat{\beta}_1 x)/n. \tag{2.11}$$

em que

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$
$$\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}.$$

A extensão do modelo de regressão linear para o caso em que o número de variáveis preditoras, k, seja maior que 1, é apresentado a seguir. Para k variáveis, são feitas as seguintes representações em forma matricial:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e} \tag{2.12}$$

com $\mathbb{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta}$. Em que

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{bmatrix}, \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \boldsymbol{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$$

Em relação ao modelo de regressão linear múltiplo, são necessárias as seguintes suposições sobre o vetor $\boldsymbol{e}$:

$$\mathbb{E}(\boldsymbol{e}) = \boldsymbol{0} \tag{2.13}$$

$$Var(\boldsymbol{e}) = \mathbb{E}[\boldsymbol{e} - \mathbb{E}(\boldsymbol{e})][\boldsymbol{e} - \mathbb{E}(\boldsymbol{e})]^\top = \mathbb{E}(\boldsymbol{e}^\top \boldsymbol{e}) = \sigma^2 \boldsymbol{I}_N \tag{2.14}$$

A suposição de uma distribuição de probabilidade para o erro faz-se necessária para que testes de hipóteses e intervalos de confiança sejam considerados sob as estimativas dos parâmetros β.

O procedimento para obter as estimativas de mínimos quadrados para o vetor $\boldsymbol{\beta}$ é análogo ao caso do modelo de regressão linear simples, ou seja, deriva-se a soma de quadrados da soma dos erros em relação ao vetor de parâmetros $\boldsymbol{\beta}$ de modo a minimizá-la. Tem-se então

$$\boldsymbol{e}^\top \boldsymbol{e} = [\boldsymbol{y} - \mathbb{E}(\boldsymbol{y})]^\top [\boldsymbol{y} - \mathbb{E}(\boldsymbol{y})] = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{2.15}$$

$$= \boldsymbol{y}^\top \boldsymbol{y} - 2\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{y} + \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} \tag{2.16}$$

Derivando (2.16) em relação ao vetor de parâmetros $\boldsymbol{\beta}$, igualando a zero e reescrevendo em termos de $\hat{\boldsymbol{\beta}}$, temos:

$$\boldsymbol{X}^\top \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^\top \boldsymbol{y} \tag{2.17}$$

e portanto,

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y} \tag{2.18}$$

em que $\hat{\boldsymbol{\beta}}$ é o estimador do vetor de parâmetros do modelo linear $\boldsymbol{\beta}$. Cabe salientar que a estimativa obtida em (2.18) é obtida desde que a matriz $\boldsymbol{X}^\top \boldsymbol{X}$ seja positiva definida, ou seja, inversível de modo que X tem posto k maior que n.

2.2 REGRESSÃO NÃO-LINEAR

Em diversas situações nas áreas de engenharias, ciências biológicas, ecologia, entre outras, a relação entre as variáveis preditoras e a resposta são regidas por uma relação não-linear conhecida, em relação aos parâmetros do modelo. Nestes casos, é mais apropriado ajustar modelos não-lineares para descrever este relacionamento do que um modelo linear menos realista. Qualquer modelo que não possui a forma dada pela Equação (2.10) é chamado de modelo de regressão não-linear.

A estimação de parâmetros para modelos não-lineares também pode se dar pelo Método dos Mínimos Quadrados. Entretanto, as equações normais obtidas são, em geral, analiticamente difíceis de tratar, o que torna necessária a aplicação de métodos iterativos de otimização.

A forma geral de um modelo não-linear é dada por

$$Y = f(\boldsymbol{\xi}; \boldsymbol{\theta}) + \boldsymbol{\epsilon} \tag{2.19}$$

em que $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_K)$ é o conjunto de variáveis preditoras, $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)$ é o vetor de parâmetros do modelo escolhido e ϵ é o vetor de erros aleatórios. Ao assumir que $\mathbb{E}(\boldsymbol{\epsilon}) = 0$, tem-se que $\mathbb{E}(Y) = f(\boldsymbol{\xi}; \boldsymbol{\theta})$. Ainda, analogamente ao caso linear, assume-se que os erros são não correlacionados e que $Var(\boldsymbol{\epsilon}) = \sigma^2$.

Uma observação u pode ser representada pela notação $Y_u, \xi_{1u}, \xi_{2u}, \ldots, \xi_{ku}$. Portanto, para n observações, a Equação (2.19) pode ser escrita como

$$Y_u = f(\xi_u, \boldsymbol{\theta}) + \boldsymbol{\epsilon}_u \tag{2.20}$$

em que $\boldsymbol{\xi}_u = (\xi_{1u}, \xi_{2u}, \ldots, \xi_{ku})^\top$. A soma dos quadrados dos erros para o modelo não-linear é dada por:

$$S(\boldsymbol{\theta}) = \sum_{u=1}^{n} \{Y_u - f(\boldsymbol{\xi_u}, \boldsymbol{\theta})\}^2 \tag{2.21}$$

Pelo Método dos Mínimos Quadrados, para obter as estimativas $\hat{\boldsymbol{\theta}}$ é necessário minimizar (2.21) em relação a $\boldsymbol{\theta}$. Ao derivar em relação aos parâmetros e igualar a zero, as K equações normais tomam a seguinte forma:

$$\sum_{u=1}^{n} \{Y_u - f(\boldsymbol{\xi_u}, \boldsymbol{\theta})\} \left[\frac{\partial f(\boldsymbol{\xi_u}, \boldsymbol{\theta})}{\partial \theta_i}\right]_{\theta=\hat{\theta}} = 0 \tag{2.22}$$

No entanto, a derivada apresentada em (2.22) será função de parâmetros desconhecidos assumindo uma forma não linear, o que implica que a solução das equações normais pode ser difícil de encontrar. A solução consiste na utilização de métodos iterativos para obtenção das estimativas. Dentre estes métodos, o mais conhecido é o Método de Linearização, que utiliza os resultados de Mínimos Quadrados linear em sucessivos estágios (DRAPER; SMITH, 1966).

O primeiro passo do Método de Linearização é expandir $f(\xi_u, \boldsymbol{\theta})$ em Séries de Taylor em torno do ponto $\boldsymbol{\theta}_0 = (\theta_{10}, \theta_{20}, \ldots, \theta_{p0})$ até a primeira ordem, obtendo

$$f(\xi_u, \boldsymbol{\theta}) = f(\xi_u, \boldsymbol{\theta}_0) + \sum_{i=1}^{p} \left[\frac{\partial f(\xi_u, \boldsymbol{\theta})}{\partial \theta_i}\right]_{\theta=\theta_0} (\theta_i - \theta_{i0}) \tag{2.23}$$

Ao definir

$$f_u^0 = f(\xi_u, \boldsymbol{\theta}_0) \tag{2.24}$$

$$\beta_i^0 = \theta_i - \theta_{i0} \tag{2.25}$$

$$Z_{iu}^0 = \left[\frac{\partial f(\xi_u, \boldsymbol{\theta})}{\partial \theta_i}\right] \tag{2.26}$$

pode-se reescrever a Equação (2.20) na forma linear abaixo:

$$Y_u - f_u^0 = \sum_{i=1}^{p} \beta_i^0 Z_{iu}^0 + \epsilon_u. \tag{2.27}$$

Em notação matricial, tem-se que

$$\boldsymbol{y}_0 = \boldsymbol{Z}_0 \beta_0 + \epsilon \tag{2.28}$$

em que

$$\boldsymbol{Z}_0 = \begin{bmatrix} Z_{11}^0 & Z_{21}^0 & \ldots & Z_{p1}^0 \\ Z_{12}^0 & Z_{22}^0 & \ldots & Z_{p2}^0 \\ \vdots & \vdots & & \vdots \\ Z_{1u}^0 & Z_{2u}^0 & \ldots & Z_{pu}^0 \\ \vdots & \vdots & & \vdots \\ Z_{1n}^0 & Z_{2n}^0 & \ldots & Z_{pn}^0 \end{bmatrix}, \hat{\boldsymbol{\beta}} = \begin{bmatrix} , \hat{\beta}_1^0 \\ \hat{\beta}_2^0 \\ \vdots \\ \hat{\beta}_p^0 \end{bmatrix}, \boldsymbol{y}_0 = \begin{bmatrix} Y_1 - f_1^0 \\ Y_2 - f_2^0 \\ \vdots \\ Y_u - f_u^0 \\ \vdots \\ Y_n - f_n^0 \end{bmatrix} = \boldsymbol{Y} - \boldsymbol{f}^0$$

Uma vez que o modelo está linearizado, a estimativa $\hat{\boldsymbol{\beta}}_0$ na iteração 0 será dada por

$$\hat{\boldsymbol{\beta}}_0 = (\boldsymbol{Z}_0^\top \boldsymbol{Z}_0)^{-1} \boldsymbol{Z}_0^\top (\boldsymbol{Y} - \boldsymbol{f}^0) \tag{2.29}$$

Uma vez que $\beta_0 = \theta - \theta_0$, pode-se definir a estimativa revisada de θ como $\hat{\theta}_1 = \hat{\beta}_0 + \theta_0$. Este processo iterativo deve continuar até que

$$\left| \frac{\hat{\theta}_{i(j+1)} - \hat{\theta}_{ij}}{\hat{\theta}_{ij}} \right| \leq \delta \tag{2.30}$$

em que $\delta > 0$ é um valor arbitrariamente pequeno.

No entanto, alguns problemas com essa abordagem devem ser pontuados: (i) a convergência pode ser lenta, exigindo um número grande de iterações; (ii) pode apresentar grande oscilação até que a solução se estabilize e (iii) o método pode não convergir ou até mesmo divergir, fazendo a soma dos quadrados aumentar ilimitadamente.

2.3 REGRESSÃO *CLUSTERWISE*

Muitas vezes, o pesquisador depara-se com situações em que há heterogeneidade nos dados. Por heterogeneidade, no contexto deste trabalho, entende-se a presença de subgrupos numa população em que o relacionamento entre os preditores e a resposta se dá de modo diverso. No caso específico de análise de regressão *clusterwise*, a heterogeneidade se manifesta através do potencial de diferentes preditores para explicar a variável resposta em diferentes grupos de indivíduos.

O método de regressão *clusterwise*, entendido como a combinação das técnicas de regressão e *cluster*, tem sido amplamente discutido na literatura (DESARBO; CRON, 1988; DESARBO; OLIVER; RANGASWAMY, 1989; VEAUX, 1989; LAU; LEUNG; TSE, 1999; HENNIG, 2000). Aplicações do método têm sido desenvolvidas nos campos de segmentação de mercado (AURIFEILLE; QUESTER, 2003; WEDEL; KISTEMAKER, 1989; BRUSCO; CRADIT; TASHCHIAN, 2003; WEDEL, 1990), finanças (CHIRICO, 2013), psicologia (DESARBO; EDWARDS, 1996), transportes (LUO; CHOU, 2006), entre outros.

Tome-se como exemplo uma situação em que uma amostra pode ser classificada em três grupos, A, B e C. Dentro de cada grupo é ajustado um modelo de regressão múltiplo, onde a tupla $(+, -, +)$ indica que o efeito dos preditores 1 e 3 são positivos e o do preditor 2 é negativo. Supondo-se que as tuplas para os grupos B e C sejam $(-, +, +)$ e $(+, +, -)$, respectivamente, temos uma situação em que, no caso de um modelo de regressão ajustado para a amostra inteira, nenhum dos três preditores seja estatisticamente significativo (BRUSCO et al., 2008).

Assim, em muitas situações, ajustar um modelo de regressão para todo o conjunto de dados poderá não exprimir, de modo preciso, o relacionamento entre a resposta e os preditores devido a presença de grupos na amostra em estudo. Neste contexto, a regressão *clusterwise*, método baseado nos trabalhos seminais de Bock (BOCK, 1969) e

Späth (SPÄTH, 1979; SPÄTH, 1982), surge como uma opção, particionando o conjunto de dados em um dado número de classes e ajustando um modelo de regressão dentro de cada classe.

Note-se que o agrupamento tem como critério de alocação de um indivíduo em um grupo a minimização de uma função objetivo. Existem propostas de regressão *clusterwise* em que a alocação dos objetos aos *clusters* é feita dentro de uma lógica *fuzzy*, isto é, cada observação possui um grau de pertencimento em relação aos *clusters* contribuindo, consequentemente, com as estimativas dos modelos em mais de um *cluster* (YANG; KO, 1997; D'URSO; SANTORO, 2006; D'URSO; MASSARI; SANTORO, 2010; TAN et al., 2013; DOTTO et al., 2016). No caso em tela, a função objetivo está relacionada com a minimização da soma dos quadrados dos erros em cada grupo e a alocação dos objetos aos *clusters* é feita de modo *hard*, i.e, cada objeto é alocado a um e somente um *cluster*.

A minimização da função objetivo é feita por meio de um algoritmo de alocação que transfere os indivíduos entre os grupos até que não seja mais possível reduzir o valor da soma dos quadrados dos erros de todos as classes. No entanto, tal algoritmo converge para mínimos locais em relação ao conjunto de todas as trocas possíveis, motivo pelo qual é recomendável a execução do algoritmo a partir de diferentes inicializações aleatórias. A formulação do problema relacionado à presença de grupos num conjunto de dados pode ser colocada nos seguintes termos:

Seja $E = (e_1, \ldots, e_n : e_i = (y_i, \boldsymbol{x}_i) \in \mathbb{R} \times \mathbb{R}^p)$ um conjunto de objetos indexados em $i = 1, \ldots, n$. Encontrar uma partição $P = P_1, \ldots, P_K$ de E tal que a função objetivo dada por

$$\begin{aligned} J_{clr}(P, \boldsymbol{\beta}) &= \sum_{k=1}^{K} J_{P_k, \boldsymbol{\beta}_k} \\ &= \sum_{k=1}^{K} \sum_{e_i \in P_k} \left(y_i - \beta_{0k} - \sum_{j=1}^{p} x_{ij} \beta_{jk} \right)^2 \end{aligned} \tag{2.31}$$

seja minimizada em relação aos coeficientes de regressão $\boldsymbol{\beta}_k = \beta_{0k}, \beta_{1k}, \ldots, \beta_{pk}$, k=1, ..., K.

Vale apontar as condições para que se tenha uma partição de E, a saber,

$$\begin{aligned} P_k &\in E; \\ |P_k| &\geq 0; \\ P_k \cap P_l &\neq \emptyset; \\ \bigcup_{k=1}^{K} P_k &= E; \end{aligned}$$

Além destas condições existe aquela em que a cardinalidade dos grupos deve ser superior ao número de variáveis preditoras dos modelos p (SPÄTH, 1979), isto é, $|P_k| \geq p$.

Note-se que, apesar do ajuste de modelos de regressão dentro dos grupos ser trivial (resolvido por mínimos quadrados), o problema de agrupamento representa um desafio combinatorial. O número de partições possíveis de serem formados por N observações em K grupos é expressa pela seguinte quantidade (HAND, 1981):

$$\frac{1}{K!}\sum_{k=0}^{K}\binom{K}{k}(K-k)^N$$

Portanto, mesmo em problemas modestos o espaço de busca das partições é significativo, o que torna necessária a aplicação de heurísticas que busquem ótimos locais. Para aplicação no método de regressão *clusterwise* é utilizada uma variação de um algoritmo partitivo sequencial baseado em trocas (MACQUEEN et al., 1967), cuja aplicação remonta ao método *K-means*.

O algoritmo é inicializado a partir de uma partição inicial, que pode ser aleatória ou utilizar alguma heurística, como o *K-means*. A partir de então, cada objeto é alocado em grupos diferentes ao que ele pertence, verificando-se o comportamento da função objetivo. Se a realocação do objeto em outros grupos não causa redução na função objetivo, ele é mantido no grupo em que se encontra, por outro lado, se a realocação reduz o critério de adequação, o objeto será finalmente realocado ao grupo que produz a maior redução. No caso da regressão *clusterwise*, o critério a ser minimizado é representado pela Equação (2.31).

Portanto, pode-se executar a regressão *clusterwise* na forma algorítmica em dois passos: (i) estimação, onde a partição dos dados é fixa e os modelos de regressão em cada grupo são ajustados, e (ii) alocação, em que são fixados os modelos de regressão e o algoritmo particional é aplicado. Estes passos devem ser executados alternadamente até que o algoritmo atinja a convergência, que se dá quando não ocorrem mais mudanças na alocação dos objetos nos grupos. O algoritmo 1 apresenta estes passos:

Algoritmo 1 Regressão *Clusterwise*

Entrada: $E = (e_1, \ldots, e_n)$, número de grupos K; representação dos objetos na forma $\mathcal{D} = \{(y_1, \mathbf{x}_1), \ldots, (y_1, \mathbf{x}_N)\}$

Saída: Uma lista dos modelos estimados (protótipos) e uma partição ótima $\mathcal{P} = (P_1, \ldots, P_K)$

1: **Inicialização:**
2: Defina $t \leftarrow 0$
3: Atribua aleatoriamente os objetos $\boldsymbol{e}_i$ para o *cluster* P_k para formar a partição inicial $\mathcal{P} = (P_1, \ldots, P_K)$;
4: **Passo 1:** Ajuste do Modelo
5: Defina $t \leftarrow t + 1$
6: **Para** $1 \leq k \leq K$ **Faça**
7: Calcule estimativas dos parâmetros dos modelos lineares $\hat{\beta_{0k}}, \ldots, \hat{\beta_{pk}}$.
8: **fim Para**
9: **Passo 2:** Alocação
10: $\mathcal{P}^{(t)} = \mathcal{P}^{(t-1)}$;
11: $test \leftarrow 0$;
12: **Para** $1 \leq i \leq n$ **Faça**
13: Seja $m : \boldsymbol{x}_i \in P_m^{(t)}$
14: Encontre o *cluster* vencedor tal que

$$k = \underset{1 \leq h \leq K}{\arg\min} \left(y_i - \hat{\beta}_{0h} - \sum_{j=1}^{p} x_{ij} \hat{\beta}_{jh} \right)$$

15: **Se** $k \neq m$ **Então**
16: $test \leftarrow 1$
17: $P_k = P_k \cup \{x_i\}$
18: $P_m = P_m \setminus \{x_i\}$
19: **fim Se**
20: **fim Para**
21: Compute o valor atual de J de acordo com a Equação (2.31)
22: **Critério de parada:**
23: **Se** $test == 0$ **Então**
24: pare;
25: **Senão**
26: $t \leftarrow t + 1$ e volte para o passo 1;
27: **fim Se**

Spath (SPÄTH, 1979) afirma que, empiricamente, o algoritmo tende a convergir em aproximadamente seis iterações. A seguir, apresenta-se a prova da convergência do algoritmo:

Teorema 2.3.1. *O algoritmo 1 decrementa monotonicamente o critério J_{clr} a cada iteração.*

Demonstração. Seja a representação da função objetivo do algoritmo *clusterwise* na ite-

ração t:

$$J_{clr}(P^t, \boldsymbol{\beta}^t) = \sum_{k=1}^{K} \sum_{e_i \in P_k^t} \left(y_i - \beta_{0k}^t - \sum_{j=1}^{p} x_{ij} \beta_{jk}^t \right)^2 . \tag{2.32}$$

Dada uma partição $P^t = \{P_1^t, \ldots, P_K^t\}$, um conjunto de coeficientes $\boldsymbol{\beta}_k^t = \{\hat{\beta}_{0k}, \ldots, \hat{\beta}_{pk}\}$ com $k = 1, \ldots, K$ obtidos por mínimos quadrados, de modo que

$$\boldsymbol{\beta}_k^{t+1} = \arg\min_{\boldsymbol{\beta}_k} \sum_{e_i \in P_k} \left(y_i - \beta_{0k}^t - \sum_{j=1}^{p} x_{ij} \beta_{jk}^t \right)^2 .$$

Então, tem-se que

$$\sum_{e_i \in P_k^t} \left(y_i - \beta_{0k}^t - \sum_{j=1}^{p} x_{ij} \beta_{jk}^t \right)^2 \geq \sum_{e_i \in P_k^t} \left(y_i - \beta_{0k}^{t+1} - \sum_{j=1}^{p} x_{ij} \beta_{jk}^{t+1} \right)^2 ,$$

com $k = 1, \ldots, K$. Com isto, para os K grupos,

$$\sum_{k=1}^{K} \sum_{e_i \in P_k^t} \left(y_i - \beta_{0k}^t - \sum_{j=1}^{p} x_{ij} \beta_{jk}^t \right)^2 \geq \sum_{k=1}^{K} \sum_{e_i \in P_k^t} \left(y_i - \beta_{0k}^{t+1} - \sum_{j=1}^{p} x_{ij} \beta_{jk}^{t+1} \right)^2 . \tag{2.33}$$

Finalmente, mostra-se que no passo de estimação, o valor da função objetivo já sofre redução:

$$J_{clr}(P^t, \boldsymbol{\beta}^t) \geq J_{clr}(P^t, \boldsymbol{\beta}^{t+1}).$$

A partir dos novos coeficientes de regressão encontrados, ou seja, $\boldsymbol{\beta}^{t+1}$, a nova partição $P^{t+1} = \{P_1^{t+1}, \ldots, P_K^{t+1}\}$ deve ser tal que

$$P_k^{t+1} = \left\{ e_i \in E : k = \arg\min_{1 \leq h \leq K} \left(y_i - \beta_{0h}^{t+1} - \sum_{j=1}^{p} x_i \beta_{jh}^{t+1} \right)^2 \right\} .$$

Com esta nova partição $\boldsymbol{\beta}^{t+1}$, pode-se representar $J_{clr}(P^{t+1}, \boldsymbol{\beta}^{t+1})$ como

$$J_{clr}(P^{t+1}, \boldsymbol{\beta}^{t+1}) = \sum_{k=1}^{K} \sum_{e_i \in P_k^{t+1}} \left(y_i - \beta_{0k}^{t+1} - \sum_{j=1}^{p} x_{ij} \beta_{jk}^{t+1} \right)^2 ,$$

que pode ser reescrita como:

$$J_{clr}(P^{t+1}, \boldsymbol{\beta}^{t+1}) = \sum_{h,k=1}^{K} \sum_{e_i \in P_h^t \cap P_k^{t+1}} \left(y_i - \beta_{0k}^{t+1} - \sum_{j=1}^{p} x_{ij} \beta_{jk}^{t+1} \right)^2 ,$$

uma vez que $P_h^t (h = 1, \ldots, K)$ forma uma partição sobre E. Assim, para todo $e_i \in P_k^{t+1}$, $h \neq k$, temos:

$$\left(y_i - \beta_{0k}^{t+1} - \sum_{j=1}^{p} x_{ij} \beta_{jk}^{t+1} \right)^2 \leq \left(y_i - \beta_{0h}^{t+1} - \sum_{j=1}^{p} x_{ij} \beta_{jh}^{t+1} \right)^2 ,$$

também válido para qualquer $e_i \in P_k^{t+1} \cap P_h^t$. Assim,

$$\begin{aligned} J_{clr}(P^{t+1}, \boldsymbol{\beta}^{t+1}) &= \sum_{h,k=1}^{K} \sum_{e_i \in P_h^t \cap P_k^{t+1}} \left(y_i - \beta_{0k}^{t+1} - \sum_{j=1}^{p} x_{ij} \beta_{jk}^{t+1} \right)^2 \\ &\leq \sum_{h,k=1}^{K} \sum_{e_i \in P_h^t \cap P_k^{t+1}} \left(y_i - \beta_{0h}^{t+1} - \sum_{j=1}^{p} x_{ij} \beta_{jh}^{t+1} \right)^2 \\ &= \sum_{h=1}^{K} \sum_{e_i \in P_h^{t+1}} \left(y_i - \beta_{0h}^{t+1} - \sum_{j=1}^{p} x_{ij} \beta_{jh}^{t+1} \right)^2 \\ &= J_{clr}(P^t, \boldsymbol{\beta}^{t+1}). \end{aligned} \tag{2.34}$$

Combinando os resultados (2.33) e (2.34), mostra-se que $J_{clr}(P^t, \boldsymbol{\beta}^t) \geq J_{clr}(P^t, \boldsymbol{\beta}^{t+1}) \geq J_{clr}(P^{t+1}, \boldsymbol{\beta}^{t+1})$. Isto quer dizer que, a cada passo do algoritmo 1 ocorre uma redução monotônica da função objetivo (2.32).

□

Os métodos apresentados neste capítulo são aplicados a dados usuais, isto é, aqueles representados por números reais. O uso de qualquer método estatístico em dados simbólicos requer a adaptação destes métodos. No entanto, este trabalho trata de uma categoria especial de dados simbólicos, a saber, dados tipo-intervalo. O capítulo seguinte apresenta os modelos de regressão linear e não linear, além da regressão clusterwise aplicados a dados tipo-intervalo.

3 MODELOS DE REGRESSÃO PARA DADOS TIPO-INTERVALO

Este capítulo tem como objetivo apresentar uma discussão sobre os modelos de regressão para dados tipo-intervalo que serão utilizados no desenvolvimento do método proposto, iCRCNLR. São apresentados o modelo de regressão linear baseado no centro e amplitude dos intervalos (NETO; CARVALHO, 2008), sua expansão para o caso não-linear (NETO; CARVALHO, 2017) e a regressão *clusterwise* linear para dados tipo-intervalo (CARVALHO; SAPORTA; QUEIROZ, 2010).

Os modelos de regressão para dados tipo-intervalo tiveram como marco inicial o método baseado na minimização do erro relativo aos centros dos intervalos (BILLARD; DIDAY, 2000). Neste caso, a predição dos limites superior e inferior era feita utilizando os valores observados destes limites e os coeficientes estimados $\hat{\beta}$ para o modelo do centro. Uma segunda abordagem consiste em ajustar dois modelos de regressão independentes para os limites inferior e superior dos intervalos (BILLARD; DIDAY, 2002), minimizando a quantidade $\sum_{i=1}^{n}(\epsilon_i^L)^2 + \sum_{i=1}^{n}(\epsilon_i^U)^2$, em que ϵ_i^L representa o erro na i-ésima observação no limite inferior e ϵ_i^U representa o erro na i-ésima observação no limite superior. O método CRM (NETO; CARVALHO, 2008) apresenta uma abordagem diferente, ao considerar modelos de regressão para o centro e meia amplitude dos intervalos, representados por $(a+b)/2$ e $(b-a)/2$, respectivamente, para um intervalo $[a, b]$.

Com objetivo de evitar a inversão dos limites dos intervalos preditos, modelos de regressão com restrições foram desenvolvidos. Lima Neto e Carvalho desenvolveram um método em que são ajustados modelos para o centro e amplitude dos dados, com a restrição de que os coeficientes de regressão sejam positivos, isto é, $\beta_i^r > 0, i = 1, \ldots, p$, em que p é o número de variáveis do modelo (NETO; CARVALHO; NETO, 2007; NETO; CARVALHO, 2010). González e Lin utilizaram o cabedal teórico de séries temporais para dados tipo-intervalo para gerar um modelo de regressão com restrições. Tal método torna as estimativas de mínimos quadrados inconsistentes, sendo proposto um algoritmo em dois passos combinando os métodos de máxima verossimilhança e mínimos quadrados para estimação. (GONZÁLEZ-RIVERA; LIN, 2013). Além destes, há o modelo de regressão para dados tipo-intervalo com restrições utilizando o método de seleção de variáveis LASSO (*Least Absolute Shrinkage and Selection Operator*) (GIORDANI, 2015).

A partir destes modelos de regressão linear para dados tipo-intervalo, uma série de métodos já conhecidos para regressão em dados usuais passaram a ser adaptadas, destacando-se a utilização de modelos não-paramétricos (LIM, 2017), métodos de estatística robusta (DOMINGUES; SOUZA; CYSNEIROS, 2010; FAGUNDES; SOUZA; CYSNEIROS, 2013) e regressão kernel (FAGUNDES; SOUZA; CYSNEIROS, 2014), além da expansão do modelo linear para o caso não-linear (NETO; CARVALHO, 2017) e a adaptação da regres-

são *clusterwise* para dados tipo-intervalo (CARVALHO; SAPORTA; QUEIROZ, 2010), apresentados nas seções seguintes.

3.1 MÉTODO DO CENTRO E AMPLITUDE (CRM)

No método CRM, o centro e a amplitude dos intervalos são usados para ajustar um modelo linear, com o objetivo de melhorar a predição do modelo quando comparado com o ajuste de um modelo linear para os limites inferior e superior dos intervalos. Seja $E = \{e_1, \ldots, e_n\}$ um conjunto de exemplos descritos por $p+1$ variáveis tipo-intervalo $Y, X_1, \ldots, X_p$. Os pontos médios destas variáveis são representados por $X_j^c (j = 1, 2, \ldots, p)$ e Y^c, enquanto os valores da meia amplitude são representados por $X_j^r (j = 1, 2, \ldots, p)$ e Y^r.

Neste caso, portanto, um exemplo $e_i \in E (i = 1, \ldots, n)$ é representado por $\boldsymbol{w}_i = (\boldsymbol{x}_i^c, y_i^c)$ e $\boldsymbol{r}_i = (\boldsymbol{x}_i^r, y_i^r)$, em que $\boldsymbol{x}_i^c = (x_{i1}^c, \ldots, x_{ip}^c)$ e $\boldsymbol{x}_i^r = (x_{i1}^r, \ldots, x_{ip}^r)$ com $x_{ij}^c = (a_{ij} + b_{ij})/2$, $x_{ij}^r = (b_{ij} - a_{ij})/2$, $y_i^c = (y_{L_i} + y_{U_i})/2$ e $y_i^r = (y_{U_i} - y_{L_i})/2$. As variáveis dependentes Y^c, Y^r e as variáveis independentes X_j^c, $X_j^r (j = 1, \ldots, p)$ se relacionam de acordo com a regressão linear abaixo:

$$y_i^c = \beta_0^c + \sum_{j=1}^{p} \beta_j x_{ij}^c + \epsilon_i^c \tag{3.1}$$

$$y_i^r = \beta_0^r + \sum_{j=1}^{p} \beta_j x_{ij}^r + \epsilon_i^r \tag{3.2}$$

em que ϵ_i^c e ϵ_i^r são erros aleatórios com $\mathcal{E}(\epsilon_i^c) = \mathcal{E}(\epsilon_i^r) = 0$, $Var(\epsilon_i^c) = \sigma_c^2$, $Var(\epsilon_i^r) = \sigma_r^2$, $Cor(\epsilon_i^c, \epsilon_j^c) = 0$ e $Cor(\epsilon_i^r, \epsilon_j^r) = 0$, $\forall i \neq j$. Das equações (3.1) e (3.2), pode-se obter a soma dos quadrados dos desvios, dada por:

$$\begin{aligned} S_{crm} &= \sum_{i=1}^{n} \left((\epsilon_i^c)^2 + (\epsilon_i^r)^2 \right) \\ &= \sum_{i=1}^{n} \left(y_i^c - \left(\beta_0^c + \sum_{j=1}^{p} \beta_j x_{ij}^c \right) \right)^2 + \sum_{i=1}^{n} \left(y_i^r - \left(\beta_0^r + \sum_{j=1}^{p} \beta_j x_{ij}^r \right) \right)^2 \\ &= \left(\mathbf{X}^c \boldsymbol{\beta}^c - \mathbf{y}^c \right)^\top \left(\mathbf{X}^c \boldsymbol{\beta}^c - \mathbf{y}^c \right) + \left(\mathbf{X}^r \boldsymbol{\beta}^r - \mathbf{y}^r \right)^\top \left(\mathbf{X}^r \boldsymbol{\beta}^r - \mathbf{y}^r \right), \end{aligned} \tag{3.3}$$

em que

- $\underset{(n\times(p+1))}{\mathbf{X}^c} = \begin{pmatrix} 1 & x_{11}^c & \ldots & x_{1p}^c \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1}^c & \ldots & x_{np}^c \end{pmatrix}$, $\underset{((p+1)\times 1)}{\boldsymbol{\beta}^c} = \begin{pmatrix} \beta_0^c \\ \vdots \\ \beta_p^c \end{pmatrix}$, $\underset{(n\times 1)}{\mathbf{y}^c} = \begin{pmatrix} y_1^c \\ \vdots \\ y_n^c \end{pmatrix}$;

- $\underset{(n\times(p+1))}{\mathbf{X}^r} = \begin{pmatrix} 1 & x_{11}^r & \dots & x_{1p}^r \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1}^r & \dots & x_{np}^r \end{pmatrix}$, $\underset{((p+1)\times 1)}{\boldsymbol{\beta}^r} = \begin{pmatrix} \beta_0^r \\ \vdots \\ \beta_p^r \end{pmatrix}$, $\underset{(n\times 1)}{\mathbf{y}^r} = \begin{pmatrix} y_1^r \\ \vdots \\ y_n^r \end{pmatrix}$.

Para encontrar os valores de $\beta_1^c, \dots, \beta_p^c$ e $\beta_1^r, \dots, \beta_p^r$ que minimizam S_{crm} é necessário obter as equações normais, i.e. derivar a equação (3.3) em relação aos parâmetros e igualar a zero. As estimativas de Mínimos Quadrados de $\beta_0^c, \beta_1^c, \dots, \beta_p^c$ e $\beta_0^r, \beta_1^r, \dots, \beta_p^r$ é a solução dos sistema com $2(p+1)$ equações normais. Em notação de matrizes, tem-se:

$$\begin{aligned} \hat{\boldsymbol{\beta}}^c &= (\hat{\beta}_0^c, \hat{\beta}_1^c, \dots, \hat{\beta}_p^c) \\ \hat{\boldsymbol{\beta}}^r &= (\hat{\beta}_0^r, \hat{\beta}_1^r, \dots, \hat{\beta}_p^r). \end{aligned} \tag{3.4}$$

Estas estimativas são obtidas pelas expressões abaixo:

$$\begin{aligned} \hat{\boldsymbol{\beta}}^c &= \left((\boldsymbol{X}^c)^\top \boldsymbol{X}^c\right)^{-1} (\boldsymbol{X}^c)^\top \boldsymbol{y}^c \\ \hat{\boldsymbol{\beta}}^r &= \left((\boldsymbol{X}^r)^\top \boldsymbol{X}^r\right)^{-1} (\boldsymbol{X}^r)^\top \boldsymbol{y}^r. \end{aligned} \tag{3.5}$$

Finalmente, os valores preditos para um novo exemplo, descrito por $\boldsymbol{z} = (\boldsymbol{x}, y)$, são dados por $\hat{y}_L = \hat{y}^c - \hat{y}^r$ e $\hat{y}_U = \hat{y}^c + \hat{y}^r$, em que $\hat{y}^c = (\tilde{\boldsymbol{x}}^c)^\top \hat{\boldsymbol{\beta}}^c$, $\hat{y}^r = (\tilde{\boldsymbol{x}}^r)^\top \hat{\boldsymbol{\beta}}^r$, $(\tilde{\boldsymbol{x}}^c)^\top = (1, x_1^c, \dots, x_p^c)$ e $(\tilde{\boldsymbol{x}}^r)^\top = (1, x_1^r, \dots, x_p^r)$.

3.2 MÉTODO DE REGRESSÃO NÃO LINEAR PARA DADOS TIPO-INTERVALO (NLM)

O método NLM consiste em ajustar dois modelos de regressão não-linear para o centro e meia amplitude dos intervalos. A relação de regressão entre variáveis tipo-intervalo Y_c, Y_r e X_j^c, $X_j^r (j = 1, \dots, p)$ é descrita por:

$$\begin{aligned} y_i^c &= f_c(\boldsymbol{x}_i^c, \boldsymbol{\theta}^c) + \epsilon_i^c \\ y_i^r &= f_r(\boldsymbol{x}_i^r, \boldsymbol{\theta}^r) + \epsilon_i^r \end{aligned}$$

em que $\boldsymbol{\theta}^c$ e $\boldsymbol{\theta}^r$ são vetores de parâmetros de dimensões $q_c \times 1$ e $q_r \times 1$ das funções não lineares f_c e f_r. A soma dos quadrados dos erros no método NLM é dada por:

$$\begin{aligned} S_{NLM} &= \sum_{i=1}^{n} (\epsilon_i^c)^2 + \sum_{i=1}^{n} (\epsilon_i^r)^2 \\ &= \sum_{i=1}^{n} \left[y_i^c - f_c(\boldsymbol{x}_i^c, \boldsymbol{\theta}^c)\right]^2 + \sum_{i=1}^{n} \left[y_i^r - f_r(\boldsymbol{x}_i^r, \boldsymbol{\theta}^r)\right]^2 \end{aligned}$$

O sistema de equações normais para a regressão não linear frequentemente não tem forma fechada. Para encontrar os valores de $\boldsymbol{\theta}^c$ e $\boldsymbol{\theta}^r$ que minimizam S_{NLM}, métodos iterativos de otimização, como Broyden-Fletcher-Goldfarb-Shanno (BFGS) (BROYDEN, 1970; FLETCHER, 1970; GOLDFARB, 1970; SHANNO, 1970), Gradiente Conjugado (CG) (HESTENES; STIEFEL, 1952) ou *Simulated Annealing* (SANN) podem ser utilizados.

BFGS é um método quasi-Newton que consiste em aproximar a matriz Hessiana usando atualizações especificadas por avaliações de gradiente. Este método converge apenas se a função objetivo tem uma expansão de Taylor quadrática próxima do ótimo. O método CG não requer avaliações da matriz Hessiana, armazenamento ou inversão de matrizes.

SANN é uma heurística que executa uma busca probabilística local inspirada pela Termodinâmica. SANN substitui a solução atual por outra em sua vizinhança baseado na função objetivo e no valor de uma variável de temperatura T. À medida em que as iterações ocorrem, o valor de T é decrementado e o método converge para uma solução local.

Finalmente, para uma nova observação e descrita por $\boldsymbol{z} = (\boldsymbol{x}, y)$, o valor $y = [y_L, y_U]$ de Y será predito da seguinte maneira: $\hat{y}_L = min\{\hat{y}^c - \hat{y}^r, \hat{y}^c + \hat{y}^r\}$, $\hat{y}_U = max\{\hat{y}^c - \hat{y}^r, \hat{y}^c + \hat{y}^r\}$, em que $\hat{y}^c = f_c(\boldsymbol{x}^c, \hat{\boldsymbol{\theta}}^c)$, $\hat{y}^r = f_r(\boldsymbol{x}^r, \hat{\boldsymbol{\theta}}^r)$, $(\boldsymbol{x}^c)^\top = (x_1^c, \ldots, x_p^c)$, $(\boldsymbol{x}^r)^\top = (x_1^r, \ldots, x_p^r)$, $\hat{\boldsymbol{\theta}}^c = (\hat{\theta}_1^c, \ldots, \hat{\theta}_p^c)$, $\hat{\boldsymbol{\theta}}^r = (\hat{\theta}_1^r, \ldots, \hat{\theta}_p^r)$. A expressão para os valores preditos de Y previne a inversão dos limites do intervalo, i.e. $\hat{y}_U \geq \hat{y}_L$.

3.3 REGRESSÃO CLUSTERWISE PARA CENTRO E AMPLITUDE DE DADOS TIPO-INTERVALO (ICRCLR)

A adaptação da regressão clusterwise para dados tipo-intervalo foi proposta por De Carvalho *et al* (CARVALHO; SAPORTA; QUEIROZ, 2010). O método propõe uma combinação do algoritmo de agrupamento dinâmico (DIDAY; SIMON, 1980) e o método regressão para dados tipo-intervalo de centro e amplitude (NETO; CARVALHO, 2008).

O método tem como objetivo encontrar uma partição de um conjunto de observações E em K *clusters* $P_1, \ldots, P_K$. Cada *cluster* possui um protótipo representado por um hiperplano definido pela relação de regressão entre as variáveis tipo-intervalo dependentes e independente Y e $X_j (j = 1, \ldots, p)$, respectivamente. Assim, a equação do modelo de regressão *clusterwise* para dados tipo intervalo é dada por:

$$\boldsymbol{y}_{i(k)} = \boldsymbol{\beta}_{0(k)} + \sum_{j=1}^{p} \boldsymbol{\beta}_{j(k)} \boldsymbol{x}_{j(k)} + \boldsymbol{\epsilon}_{i(k)} (\forall i \in P_k), \tag{3.6}$$

em que $\boldsymbol{\beta}_{0(k)} = \begin{pmatrix} \beta^c_{0(k)} \\ \beta^r_{0(k)} \end{pmatrix}$, $\boldsymbol{\beta}_{j(k)} = \begin{pmatrix} \beta^c_{j(k)} & 0 \\ 0 & \beta^r_{j(k)} \end{pmatrix} (j = 1, \ldots, p)$, $\boldsymbol{x}_{j(k)} = \begin{pmatrix} x^c_{j(k)} \\ x^r_{j(k)} \end{pmatrix}$ e

$$\boldsymbol{\epsilon}_{i(k)} = \begin{pmatrix} \epsilon^c_{i(k)} \\ \epsilon^r_{i(k)} \end{pmatrix} = \begin{pmatrix} y^c_i - \left(\beta^c_{0(k)} + \sum_{j=1}^{p} \beta_{j(k)} x^c_{ij}\right) \\ y^r_i - \left(\beta^r_{0(k)} + \sum_{j=1}^{p} \beta_{j(k)} x^r_{ij}\right) \end{pmatrix} (\forall i \in P_k).$$

Adicionalmente, cada protótipo é obtido de acordo com um critério de adequação que mede o ajuste entre os objetos alocados ao *cluster* e o protótipo, i.e. a minimização dos resíduos da regressão. Desta forma, o critério de adequação é dado por:

$$\begin{aligned} J &= \sum_{k=1}^{K} \sum_{e_i \in P_k} (\boldsymbol{\epsilon}_{i(k)})^\top \boldsymbol{\epsilon}_{i(k)} = \sum_{k=1}^{K} \sum_{e_i \in P_k} \left[(\epsilon^c_{i(k)})^2 + (\epsilon^r_{i(k)})^2 \right] \\ &= \sum_{k=1}^{K} \sum_{e_i \in P_k} \left\{ \left[y^c_i - \left(\beta^c_{0(k)} + \sum_{j=1}^{p} \beta^c_{j(k)} x^c_{ij} \right) \right]^2 + \left[y^r_i - \left(\beta^r_{0(k)} + \sum_{j=1}^{p} \beta^r_{j(k)} x^r_{ij} \right) \right]^2 \right\}. \end{aligned} \tag{3.7}$$

O algoritmo executa dois passos até atingir convergência, ponto em que não ocorrem mudanças na alocação dos objetos nos *clusters*. Os passos são definidos abaixo.

3.3.1 Passo 1: definição dos melhores protótipos

Neste estágio, a partição de E em K *clusters* é fixada. O protótipo do *cluster* P_k é representado por:

$$\hat{\boldsymbol{y}}_{i(k)} = \begin{pmatrix} \hat{y}^c_{i(k)} \\ \hat{y}^r_{i(k)} \end{pmatrix} = \begin{pmatrix} \hat{\beta}^c_{0(k)} + \sum_{j=1}^{p} \hat{\beta}^c_{j(k)} x^c_{ij} \\ \hat{\beta}^r_{0(k)} + \sum_{j=1}^{p} \hat{\beta}^r_{j(k)} x^r_{ij} \end{pmatrix} (\forall i \in P_k) \tag{3.8}$$

em que $\hat{\beta}^c_j(k)$ e $\hat{\beta}^r_j(k)$ são as estimativas de Mínimos Quadrados de $\beta^c_j(k)$ e $\beta^r_j(k)$, as quais minimizam o critério de adequação J. De Carvalho *et al* (CARVALHO; SAPORTA; QUEIROZ, 2010) mostra que estas estimativas podem ser obtidas como solução do sistema de $2(p+1)$ equações:

$$\begin{aligned} \hat{\boldsymbol{\beta}}^c &= \left((\boldsymbol{X}^c)^\top \boldsymbol{X}^c \right)^{-1} (\boldsymbol{X}^c)^\top \boldsymbol{y}^c \\ \hat{\boldsymbol{\beta}}^r &= \left((\boldsymbol{X}^r)^\top \boldsymbol{X}^r \right)^{-1} (\boldsymbol{X}^r)^\top \boldsymbol{y}^r \end{aligned} \tag{3.9}$$

3.3.2 Passo 2: definição da melhor partição

Neste passo, os protótipos $\hat{\boldsymbol{y}}_{i(k)} (k = 1, \ldots, K)$ são fixados. Os *clusters* ótimos $P_k = (k = 1, \ldots, K)$ que minimizam o critério de adequação J, são obtidos de acordo com a seguinte regra de alocação:

$$P_k = \left\{ i \in E : (\boldsymbol{\epsilon}_{i(k)})^\top \boldsymbol{\epsilon}_{i(k)} \leq (\boldsymbol{\epsilon}_{i(h)})^\top \boldsymbol{\epsilon}_{i(h)}, \forall h \neq k \, (h = 1, \ldots, K) \right\} \tag{3.10}$$

Para uma nova observação $\boldsymbol{e} = (w_1, \ldots, w_p, z)$, descrita por $\boldsymbol{t} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p, \boldsymbol{y})$, um vetor de vetores bivariados, $z = [z^L, z^U]$ é predito a partir da estimativa $\hat{\boldsymbol{y}}_{(k)} = (\hat{y}^c_{(k)}, \hat{y}^r_{(k)})$ como abaixo:

$$\hat{z}^L_{(k)} = \hat{y}^c_{(k)} - \hat{y}^r_{(k)} \text{ e } \hat{z}^U_{(k)} = \hat{y}^c_{(k)} + \hat{y}^r_{(k)}$$

em que $\hat{y}^c_{(k)} = \hat{\beta}^c_{0(k)} + \sum_{j=1}^{p} \hat{\beta}^c_{j(k)} x^c_j$ e $\hat{y}^r_{(k)} = \hat{\beta}^r_{0(k)} + \sum_{j=1}^{p} \hat{\beta}^r_{j(k)} x^r_j$.

Neste capítulo foram apresentados os modelos de regressão para dados tipo-intervalo utilizados no desenvolvimento do algoritmo iCRCNLR. O novo método, diferentemente da regressão *clusterwise* linear, fornece, dentre uma lista de modelos, o par que minimiza o erro no centro e amplitude dos intervalos. Assim, o modelo de regressão *clusterwise* linear para dados tipo-intervalo constitui-se num caso especial do iCRCNLR. A estimação de parâmetros em modelos não lineares pode ser feita por mínimos quadrados, mas cada estimador precisaria ser computado algebricamente, o que torna a automatização da estimação por um só método impraticável.

Ainda que fosse possível obter analiticamente os estimadores para os parâmetros de cada função em uma lista de funções não lineares, a estimação por mínimos quadrados nem sempre apresentaria forma fechada. Por este motivo, foram utilizados métodos de otimização para encontrar as estimativas de parâmetros no método iCRCNLR. Tal modelo é formalmente definido no capítulo subsequente.

4 MODELO DE REGRESSÃO *CLUSTERWISE* GERAL PARA DADOS TIPO-INTERVALO

Este capítulo apresenta a regressão *clusterwise* geral para dados tipo-intervalo (iCRCNLR), baseada no algoritmo de agrupamento dinâmico (DIDAY; SIMON, 1980) e nos modelos de regressão linear (NETO; CARVALHO, 2017) e não-linear (NETO; CARVALHO, 2008) de centro e amplitude.

O objetivo do algoritmo proposto é fornecer um par de modelos que minimize o erro no centro e amplitude dos intervalos, bem como uma partição dos dados em K grupos. Os modelos são selecionados a partir de um conjunto pré-definido, incluindo o modelo linear. A estimação dos parâmetros dos modelos é feito por meio de heurísticas de otimização, como SANN, Gradiente Conjugado e BFGS.

Além de apresentar o algoritmo iCRCNLR, neste capítulo são apresentados os métodos de alocação de novas observações aos *clusters* obtidos. Uma forma trivial de alocar uma observação é fazer isso de modo aleatório, sorteando um dentre os K *clusters*. Os outros métodos de alocação aqui apresentados são: (i) alocar a nova observação segundo o voto majoritário dos k-vizinhos mais próximos no eixo x, ou (ii) propor uma ponderação das predições nos K *clusters* utilizando o método de *Stacked Regressions*.

4.1 ALGORITMO ICRCNLR

Seja $E = \{e_1, e_2, \ldots, e_n\}$ o conjunto de objetos com $e_i = (\mathbf{w}_i, z_i)$. Cada objeto tem um vetor de p variáveis tipo-intervalo independentes, i.e. $\mathbf{w}_i = (w_{i1}, w_{i2}, \ldots, w_{ip})$, e uma variável dependente z_i. Em termos dos limites inferior e superior dos intervalos, estas variáveis podem ser definidas como $w_{ij} = [w_{ij}^L, w_{ij}^U]$ ($1 \leq i \leq n$ e $1 \leq j \leq p$) e $z_i = [z_i^L, z_i^U]$.

Representando os mesmo objetos em termos dos centros e amplitudes dos intervalos, pode-se escrever $\mathbf{t}_i = (\mathbf{x}_i, \mathbf{y}_i)$, em que

$$\mathbf{x}_i = \begin{bmatrix} x_{ij}^c \\ x_{ij}^r \end{bmatrix} \; ; \; \mathbf{y}_i = \begin{bmatrix} y_{ij}^c \\ y_{ij}^r \end{bmatrix}$$

O centro e a meia amplitude para o objeto i relativo à variável independente j são computados como segue:

$$x_{ij}^c = \frac{w_{ij}^L + w_{ij}^U}{2} \tag{4.1}$$

$$x_{ij}^r = \frac{w_{ij}^U - w_{ij}^L}{2} \tag{4.2}$$

Seja o modelo de regressão não-linear $y_i = f_{(k)}(\mathbf{x}_i, \boldsymbol{\beta}_{(k)}) + \boldsymbol{\epsilon}_i$, em que $\boldsymbol{\epsilon}$ são erros não-correlacionados com distribuição assintótica $\mathcal{N}(0, \sigma^2)$; $\boldsymbol{\beta}$ é um vetor de parâmetros

desconhecidos de uma função não linear e diferenciável. O índice k refere-se ao *cluster* k.

Utilizando a notação de centro e amplitude, o modelo pode ser representado a partir do seguinte par de equações:

$$y_i^c = f_{(k)}^c \left(\mathbf{x}_i^c, \boldsymbol{\beta}_{(k)}^c \right) + \boldsymbol{\epsilon}_i^c \tag{4.3}$$

$$y_i^r = f_{(k)}^r \left(\mathbf{x}_i^r, \boldsymbol{\beta}_{(k)}{}^r \right) + \boldsymbol{\epsilon}_i^r \tag{4.4}$$

Seja $\mathcal{D} = (\mathbf{t}_1, \ldots, \mathbf{t}_n)$ um conjunto de dados tipo-intervalo. Considere-se a partição destes dados em K *clusters* $\mathcal{P} = (P_1, \ldots, P_K)$ e um vetor de protótipos K-dimensional $\boldsymbol{G} = \left(\boldsymbol{g}_{(1)}, \cdots, \boldsymbol{g}_{(K)} \right)$, em que $\boldsymbol{g}_{(k)} = \left(f_{(k)}^c \left(\boldsymbol{x}_i, \boldsymbol{\beta}_{(k)}^c \right), f_{(k)}^r \left(\boldsymbol{x}_i^r, \boldsymbol{\beta}_{(k)}^r \right) \right)$ em que $f_{(k)}^c$ e $f_{(k)}^r$ são funções contínuas com vetores de parâmetros dados por $\boldsymbol{\beta}_{(k)}^c$ e $\boldsymbol{\beta}_{(k)}^r$, respectivamente. Assim, a função objetivo J de iCRCNLR é

$$\begin{aligned} J &= \sum_{k=1}^{K} \sum_{e_i \in P_k} \left(\boldsymbol{\epsilon}_{i(k)} \right)^\top \boldsymbol{\epsilon}_{i(k)} \\ &= \sum_{k=1}^{K} \sum_{e_i \in P_k} \left[\left(\epsilon_{i(k)}^c \right)^2 + \left(\epsilon_{i(k)}^r \right)^2 \right] \\ &= \sum_{k=1}^{K} \sum_{e_i \in P_k} \left\{ \left[y_i^c - f_{(k)h^c o}^c (\mathbf{x}_i^c, \boldsymbol{\beta}_{(k)}^c) \right]^2 + \left[y_i^r - f_{(k)h^r o}^r (\mathbf{x}_i^r, \boldsymbol{\beta}_{(k)}^r) \right]^2 \right\} \end{aligned} \tag{4.5}$$

em que $f_{(k)h^c o}^c$ e $f_{(k)h^r o}^r$ são, respectivamente, funções para o centro h^c e meia amplitude h^r selecionadas dentre H funções para o *cluster* k. As estimativas dos parâmetros destas funções são obtidas pelo método de otimização $o \in \{1, 2, \ldots, O\}$. Objetivando obter as estimativas de parâmetros que minimizam a função objetivo em um *cluster* fixo k, pode-se computar as equações normais para o centro e amplitude como

$$\sum_{i=1}^{n} \left[y_{i(k)}^c - f_{(k)}^c \left(\boldsymbol{x}_i^c, \boldsymbol{\beta}_{(k)}^c \right) \right] \left[\frac{\partial f_{(k)}^c \left(\boldsymbol{x}_i^c, \boldsymbol{\beta}_{(k)}^c \right)}{\partial \beta_{j(k)}} \right]_{\beta^c = \hat{\beta}^c} = 0 \tag{4.6}$$

$$\sum_{i=1}^{n} \left[y_{i(k)}^r - f_{(k)}^r \left(\boldsymbol{x}_i^r, \boldsymbol{\beta}_{(k)}^r \right) \right] \left[\frac{\partial f_{(k)}^r \left(\boldsymbol{x}_i^r, \boldsymbol{\beta}_{(k)}^c \right)}{\partial \beta_{j(k)}} \right]_{\beta^r = \hat{\beta}^r} = 0 \tag{4.7}$$

As soluções das equações normais 4.6 e 4.7 podem ser difíceis de obter analiticamente, ou seja, pode não haver forma fechada para os estimadores. Para contornar esta dificuldade, alguns métodos de otimização foram considerados: BFGS, Gradiente conjugado (CG) e *Simulated Annealing* (SANN). Em algumas situações, estes métodos podem não convergir para um ótimo local, oferecendo estimativas ruins. Neste caso, o algoritmo iCRCNLR utiliza o primeiro destes métodos a apresentar resultados consistentes. Para a implementação do iCRCNLR, o primeiro algoritmo de otimização a não falhar é utilizado para computar as estimativas dos parâmetros. O método iCRCNLR é descrito passo a passo no algoritmo 2.

Algoritmo 2 Algoritmo iCRCNLR

Entrada: $\mathcal{D} = (\boldsymbol{t}_1, \ldots, \boldsymbol{t}_n)$, número de *clusters* K, H funções contínuas candidatas e O heurísticas de otimização;

Saída: Uma lista contendo os melhores dentre os H modelos para centro e amplitude (protótipos) e uma partição ótima $\mathcal{P} = (P_1, \ldots, P_K)$.

1: **Inicialização:** Defina $t \leftarrow 0$ Atribua aleatoriamente os objetos $\boldsymbol{t}_i$, $i = 1, \ldots, n$, para o *cluster* P_k, $k = 1, \ldots, K$ para formar a partição inicial $\mathcal{P}^{(0)} = \left(P_1^{(0)}, \ldots, P_K^{(0)}\right)$;

2: **Passo 1:** Ajuste do Modelo

3: Defina $t \leftarrow t + 1$

4: **Para** $1 \leq k \leq K$ **Faça**

5: **Para** $1 \leq h \leq H$ **Faça**

6: **Para** $1 \leq o \leq O$ **Faça**

7: Compute, utilizando o método de otimização iterativo o, $\hat{\boldsymbol{\beta}}^c_{(k)h^c o}$ and $\hat{\boldsymbol{\beta}}^r_{(k)h^r o}$;

8: Armazene as estimativas da primeira das O heurísticas que convergir e siga; Se nenhuma das O heurísticas convergir, armazene o último modelo e suas estimativas e siga;

9: **fim Para**

10: **fim Para**

11: Selecione $h^c o$ e $h^r o$, tais que

$$f^c_{(k)h^c o} = \min_{1 \leq h \leq H} \sum_{e_i \in P_k} \left[y^c_i - f^c_{(k)h}(\mathbf{x}^c_i, \boldsymbol{\beta}^c_{(k)}) \right]^2$$

e

$$f^c_{(k)h^r o} = \min_{1 \leq h \leq H} \sum_{e_i \in P_k} \left[y^r_i - f^r_{(k)h}(\mathbf{x}^r_i, \boldsymbol{\beta}^r_{(k)}) \right]^2$$

12: **fim Para**

13: **Passo 2:** Alocação

14: $\mathcal{P}^{(t)} = \mathcal{P}^{(t-1)}$;

15: $test \leftarrow 0$;

16: **Para** $1 \leq i \leq n$ **Faça**

17: Seja $m : \boldsymbol{x}_i \in P_m^{(t)}$

18: Encontre o *cluster* vencedor tal que

$$k = \arg\min_{1 \leq h \leq K} \left\{ \left[\left(\hat{\boldsymbol{e}}^c_{i(h)} \right)^{(t)} + \left(\hat{\boldsymbol{e}}^r_{i(h)} \right)^{(t)} \right]^2 \right\}$$

19: **Se** $k \neq m$ **Então**

20: $test \leftarrow 1$, $P_k = P_k \cup \{\boldsymbol{x}_i\}$, $P_m = P_m \setminus \{\boldsymbol{x}_i\}$;

21: **fim Se**

22: **fim Para**

23: Compute o valor atual de J de acordo com a equação (4.5)
24: **Critério de parada:**
25: **Se** $test == 0$ **Então**
26: pare;
27: **Senão**
28: $t \leftarrow t + 1$ e volte para o passo 1;
29: **fim Se**

4.2 CONVERGÊNCIA DO ALGORITMO ICRCNLR

O algoritmo iCRCNLR busca por uma partição $\mathcal{P}^* = (P_1^*, \ldots, P_K^*)$ de E em K *clusters* não vazios e um vetor de protótipos K-dimensional $\boldsymbol{G}^* = \left(\boldsymbol{g}_{(1)}^*, \ldots, \boldsymbol{g}_{(K)}^*\right)$, em que

$$\boldsymbol{g}_{(k)}^* = \left(f_{(k)}^{c*}\left(\boldsymbol{x}_i^c, \boldsymbol{\beta}_{(k)}^{c*}\right), f_{(k)}^{r*}\left(\boldsymbol{x}_i^r, \boldsymbol{\beta}_{(k)}^{r*}\right)\right)$$

em que $f_{(k)}^{c*}$ e $f_{(k)}^{r*}$ são funções contínuas com vetores de parâmetros dados por $\boldsymbol{\beta}_{(k)}^{c*}$ e $\boldsymbol{\beta}_{(k)}^{r*}$, respectivamente. Assim, $\boldsymbol{G}^*$ e $\mathcal{P}^*$ representam os valores que minimizam J:

$$J\left(\boldsymbol{G}^*, \mathcal{P}^*\right) = min\left\{J\left(\boldsymbol{G}, \mathcal{P}\right) : \boldsymbol{G} \in \mathbb{L}^K, \mathcal{U} \in \mathbb{P}_K\right\} \tag{4.8}$$

em que $\mathbb{P}_K$ é o conjunto de todas as partições de E em K *clusters* não vazios tais que $P_k \in (p(E) - \emptyset)$, em que $p(E)$ é o conjunto de partes de E e $P_k \in \mathbb{P}$ e onde $\mathbb{L}$ representa o espaço de protótipos contendo H funções contínuas.

As propriedades de convergência deste tipo de algoritmo podem ser definidas a partir do estudo de duas séries (DIDAY; SIMON, 1980): $v_t = \left(\boldsymbol{G}^t, \mathcal{P}_K^t\right) \in \mathbb{L}^K \times \mathbb{P}_K$ e $u_t = J(v_t) = J\left(\boldsymbol{G}^t, \mathcal{P}^t\right)$, $t = 0, 1, \ldots$. Partindo do termo inicial $v_0 = \left(\boldsymbol{G}^0, \mathcal{P}_K^0\right)$, o algoritmo computa os termos da série até a convergência, em que o critério J assume um valor estacionário.

Proposição 4.2.1. *A série $u_t = J(v_t)$ diminui a cada iteração e converge.*

Demonstração. Primeiramente, demonstra-se que as desigualdades abaixo valem e decrescem a cada iteração:

$$J\left(\boldsymbol{G}^t, \mathbb{P}_K^t\right) \geq J\left(\boldsymbol{G}^{(t+1)}, \mathbb{P}_K^t\right) \geq J\left(\boldsymbol{G}^{(t+1)}, \mathbb{P}_K^{(t+1)}\right) \tag{4.9}$$

o lado esquerdo vale, uma vez que, para f^c and f^r *fixos*,

$$\hat{\boldsymbol{\beta}}_{(k)}^{c,(t+1)} = \arg\min_{\boldsymbol{\beta}} \sum_{e_i \in P_k^t} \left(\boldsymbol{y}_{i(k)}^{c,t} - f_{(k)}^{c,t}(\boldsymbol{x}_{i(k)}^{c,t}, \boldsymbol{\beta}_{(k)})\right)^2 \tag{4.10}$$

$$\hat{\boldsymbol{\beta}}_{(k)}^{r,(t+1)} = \arg\min_{\boldsymbol{\beta}} \sum_{e_i \in P_k^t} \left(\boldsymbol{y}_{i(k)}^{r,t} - f_{(k)}^{r,t}(\boldsymbol{x}_{i(k)}^{r,t}, \boldsymbol{\beta}_{(k)})\right)^2 \tag{4.11}$$

as funções selecionadas $f^{c(t+1)}_{(k)h^co}$ e $f^{r(t+1)}_{(k)h^ro}$ minimizam as seguintes somas de erros:

$$f^{c(t+1)}_{(k)h^co} = \arg\min_{1\leq h\leq H} \sum_{e_i\in P^t_k} \left[y^c_i - f^c_{(k)h^c}(\mathbf{x}^c_i, \boldsymbol{\beta}^c_{(k)})\right]^2 \tag{4.12}$$

$$f^{r(t+1)}_{(k)h^ro} = \arg\min_{1\leq h\leq H} \sum_{e_i\in P^t_k} \left[y^r_i - f^r_{(k)h^r}(\mathbf{x}^r_i, \boldsymbol{\beta}^r_{(k)})\right]^2 \tag{4.13}$$

Consequentemente, na iteração $t+1$, o valor da função objetico reduz a partir da iteração precedente

$$\sum_{k=1}^{K}\sum_{e_i\in P_k}\left\{\left[\boldsymbol{y}^{c,t}_{i(k)} - f^{c,t}_{(k)h^co}\left(\boldsymbol{x}^{c,t}_{i(k)}, \boldsymbol{\beta}^{c,t}_{(k)}\right)\right] + \left[\boldsymbol{y}^{r,t}_{i(k)} - f^{r,t}_{(k)h^ro}\left(\boldsymbol{x}^{r,t}_{i(k)}, \boldsymbol{\beta}^{r,t}_{(k)}\right)\right]\right\} \geq$$

$$\sum_{k=1}^{K}\sum_{e_i\in P_k}\left\{\left[\boldsymbol{y}^{c,t}_{i(k)} - f^{c,(t+1)}_{(k)h^co}\left(\boldsymbol{x}^{c,t}_{i(k)}, \boldsymbol{\beta}^{c,(t+1)}_{(k)}\right)\right] + \left[\boldsymbol{y}^{r,t}_{i(k)} - f^{r,(t+1)}_{(k)h^ro}\left(\boldsymbol{x}^{r,t}_{i(k)}, \boldsymbol{\beta}^{r,(t+1)}_{(k)}\right)\right]\right\} \tag{4.14}$$

Assim, uma vez que a função selecionada e as estimativas de seus parâmetros minimizam os erros,

$$J\left(\boldsymbol{G}^t, \mathcal{P}^{(t)}_K\right) \geq J\left(\boldsymbol{G}^{(t+1)}, \mathcal{P}^t_K\right)$$

O lado direito da desigualdade em 4.9 vale porque

$$\begin{aligned}\mathcal{P}^{(t+1)}_K = \arg\min_{\mathcal{P}=(P_1,\ldots,P_K)\in\mathbb{P}_K} \sum_{k=1}^{K}\sum_{e_i\in P_k}\Big\{&\left[\boldsymbol{y}^{c,t}_{i(k)} - f^{c,t}_{(k)h^co}\left(\boldsymbol{x}^{c,t}_{i(k)}, \boldsymbol{\beta}^{c,(t+1)}_{(k)}\right)\right] + \\ &\left[\boldsymbol{y}^{r,t}_{i(k)} - f^{r,t}_{(k)h^ro}\left(\boldsymbol{x}^{r,t}_{i(k)}, \boldsymbol{\beta}^{r,(t+1)}_{(k)}\right)\right]\Big\}\end{aligned} \tag{4.15}$$

Não há outra partição $\mathcal{P}$ que faça J decrescer mais do que $\mathbb{P}^{(t+1)}$, que é única. Com isso, tem-se que a segunda desigualdade em 4.9 vale:

$$J\left(\boldsymbol{G}^{(t+1)}, \mathbb{P}^t_K\right) \geq J\left(\boldsymbol{G}^{(t+1)}, \mathbb{P}^{(t+1)}_K\right)$$

Assim, pode-se concluir que a série u_t decresce e é limitada ($J \geq 0$), e portanto, converge. □

Proposição 4.2.2. *A série $v_t = \left(\boldsymbol{G}^t, \mathcal{P}^t\right)$ converge.*

Demonstração. Assumindo que $\mathcal{P}_t$ alcança estacionariedade na iteração T, $\mathcal{P}_T = \mathcal{P}_{(T+1)}$ e $J(v_T) = J(v_{(T+1)})$. Baseado em 4.9, tem-se que

$$J\left(\boldsymbol{G}^T, \mathcal{P}^T\right) = J\left(\boldsymbol{G}^{T+1}, \mathcal{P}^T\right) = J\left(\boldsymbol{G}^T, \mathcal{P}^{T+1}\right) \tag{4.16}$$

Para o lado esquerdo das igualdades, $\boldsymbol{G}^{(T+1)} = \boldsymbol{G}^{(T)}$. Para a partição fixada $\mathcal{P}^T$, as estimativas de mínimos quadrados $\hat{\boldsymbol{\beta}}^{cT}$, $\hat{\boldsymbol{\beta}}^{rT}$, as funções selecionadas e os métodos de otimização são únicos e o critério de otimização J sustenta o mesmo valor. Para o lado direito da equação, $\mathcal{P}^{(T+1)} = \mathcal{P}^{(T)}$ ocorre porque $\mathcal{P}$ é única a cada iteração, ou seja, para $\boldsymbol{G}$ não há duas ou mais partições que minimizam J. □

4.3 MÉTODOS DE ALOCAÇÃO

A saída de uma regressão *clusterwise* consiste num conjunto de K superfícies (hiperplanos no caso linear). No entanto, não fornece nenhuma forma de utilizar estes planos para fazer predições a respeito de observações de teste (MANWANI; SASTRY, 2015). Para fins de predição, é necessário a adoção de métodos *ad hoc* para alocar as observações a serem preditas em um cluster ou, alternativamente, fornecer uma predição resultante de uma combinação das predições individuais dos K modelos ajustados nos *clusters*.

Nesta seção, descrevem-se os métodos utilizados para alocação de uma nova observação em um único *cluster* obtido pelo algoritmo iCRCNLR. Dado que o melhor par de modelos para o centro e amplitude foi obtido, um dos passos para a predição consiste em alocar as novas observações em um dos *clusters* criados pelo algoritmo. A predição do algoritmo iCRCNLR se dá da seguinte maneira:

Para uma nova observação $e = (w_1, \ldots, w_p, z)$, descrita a por $t = (x_1, \ldots, x_p, y)$, pode-se prever $Z = [Z^L, Z^U]$ a partir da estimativa $\hat{y}_{(k)} = (\hat{y}^c_{(k)}, \hat{y}^r_{(k)})$ como $Z^L_{(k)} = \hat{y}^c_{(k)} - \hat{y}^r_{(k)}$ e $Z^U_{(k)} = \hat{y}^c_{(k)} + \hat{y}^r_{(k)}$, em que $\hat{y}^c_{(k)} = f^c_{(k)}(x^c_{(k)}, \hat{\beta}^c_k)$ e $\hat{y}^r_{(k)} = f^r_{(k)}(x^r_{(k)}, \hat{\beta}^r_k)$, em que k é o k-ésimo *cluster* obtido pelo algoritmo.

Vale ressaltar que este problema de alocação é, de fato, um problema de classificação supervisionada, em que os rótulos atribuídos aos dados foram dados pela regressão *clusterwise*. Desta forma, qualquer método de classificação pode ser utilizado a esta altura. Os métodos de alocação devem ser capazes de selecionar um *cluster* k para uma nova observação e, ou fornecer uma combinação das predições dos modelos em cada *cluster*, consequentemente, esta etapa deve ocorrer antes da predição. São apresentadas aqui três abordagens para alocação: (i) KNN para dados tipo-intervalo; (ii) uma abordagem alternativa, conhecida como *Stacked Regressions*, que não seleciona um único *cluster*, mas apresenta uma predição resultante de uma ponderação entre as predições obtidas nos diferentes *clusters* e (iii) alocação aleatória, onde uma observação é alocada a um *cluster* por meio de sorteio. Tais métodos são apresentados com mais detalhamento nas seções seguintes.

4.3.1 KNN para Dados Tipo-Intervalo

Para além da alocação aleatória dos exemplos de treinamento a um *cluster*, pode-se fazer esta atribuição por meio do método não-paramétrico conhecido por K-*Nearest Neighbors* (FIX E., 1951; COVER; HART, 2006). Este é um dos algoritmos mais simples e intuitivos da área de Aprendizagem de Máquina, podendo ser utilizada tanto para classificação, caso deste trabalho, quanto para regressão.

Em resumo, a saída do KNN, no caso de classificação, indica a classe à qual um indivíduo pertence. Tanto no caso de regressão quanto de classificação, a entrada do

algoritmo consiste na descrição do objeto a ser classificado. A classificação ocorre de acordo com o voto majoritário dos seus vizinhos, com o objeto sendo atribuído à classe mais frequente entre os k vizinhos mais próximo. Esta votação também pode ser ponderada, com os pesos atribuídos a cada vizinho sendo proporcionais à sua distância em relação ao objeto a ser classificado. O KNN é um algoritmo de aprendizagem baseada em instâncias. Isto significa que o algoritmo classifica uma nova instância comparando-a com as instâncias que possui armazenadas em memória.

O treinamento do algoritmo KNN resume-se ao armazenamento dos exemplos de treinamento, que por sua vez são formados por vetores em um espaço multidimensional de características e seus respectivos rótulos, representação da sua classe. Para a classificação, uma constante k, definida pelo usuário ou por validação cruzada, é definida, representando o número de vizinhos que serão considerados para classificação do indivíduo não observado, e portanto, sem rótulo.

Para computar as distâncias entre o exemplo de teste e os exemplos de treino, uma métrica de distância deve ser considerada. Para o caso de variáveis contínuas, a mais utilizada é a distância Euclidiana, dada por

$$d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x}) = \sqrt{(x_1 - y_1)^2 + \ldots + (x_n - y_n)^2} \tag{4.17}$$

em que $\boldsymbol{x}$ e $\boldsymbol{y}$ são dois vetores no espaço Euclidiano n-dimensional. No entanto, dados do tipo-intervalo representam hiper-retângulos no espaço Euclidiano, necessitando de medidas de distâncias particulares. Algumas medidas de similaridade e dissmilaridade foram introduzidas na literatura de dados simbólicos (GOWDA; DIDAY, 1991; GOWDA; DIDAY, 1992). Dentre as distâncias existentes para dados tipo-intervalo, a escolhida foi a distância de Hausdorff (CHAVENT, 2004). Originalmente utilizada em processamento de imagens, a distância de Hausdorff é formulada para comparar dois conjuntos de objetos A e B. A distância de Hausdorff é definida como

$$d_H(A, B) = max(h(A, B), h(B, A)) \tag{4.18}$$

em que $h = sup_{u \in A} inf_{v \in B} ||u - v||$. Definindo A e B como os intervalos $x_i = [a_i, b_i]$ e $x_{i'} = [a_{i'}, b_{i'}]$, pode-se obter a seguinte distância para dados tipo-intervalo:

$$d_H(x_i, x_{i'}) = \max\{|a_i - a_{i'}|, |b_i - b_{i'}|\} \tag{4.19}$$

Finalmente, no contexto de regressão *clusterwise* para dados tipo-intervalo, a alocação de um novo objeto (não rotulado) a um *cluster* seria feita da seguinte maneira:

1. A saída da regressão *clusterwise* fornece uma partição dos dados. Cada indivíduo pertence a apenas um *cluster*. Em outras palavras, todo o conjunto de treinamento é rotulado após o ajuste da regressão;

2. Uma nova observação será classificada de acordo com o KNN para dados tipo-intervalo utilizando a distância de Hausdorff em relação às variáveis independentes X, dado que o regressando Y não é observado no indivíduo a ser classificado;

3. Finalmente, o *cluster* ao qual será alocado o novo indivíduo será decidido por voto majoritário por parte dos k-vizinhos mais próximos.

4.3.2 Stacked Regressions

Uma alternativa à escolha de um único *cluster* consiste em prover a predição para uma nova observação como uma ponderação das predições nos diferentes *clusters*. O método de *Stacked Regressions* (BREIMAN, 1996), originalmente desenvolvido para oferecer uma combinação linear de diferentes preditores para oferecer obter maior acurácia das predições, pode ser adaptado ao contexto de *Clusterwise Regression*. Neste caso, o lugar dos preditores é assumido pelos modelos obtidos pelo algoritmo em cada *cluster*.

Em linhas gerais, o que se obtém é uma predição a partir da combinação linear das predições em cada *cluster*. Para obter os valores preditos, a ideia central é utilizar validação cruzada e o método de Mínimos Quadrados sob a restrição de não negatividade para obter os coeficientes da combinação linear dos preditores (*clusters*).

Suponha-se que estejam disponíveis K preditores $v_1(\boldsymbol{x}), \ldots, v_K(\boldsymbol{x})$ para uma variáveis dependente y em termos do vetor $\boldsymbol{x}$. Wolpert (WOLPERT, 1992) propôs a seguinte abordagem: dispondo de K preditores, um preditor com maior acurácia pode ser obtido não escolhendo um dos preditores, mas combinando $v_1, \ldots, v_k$. O método pode ser resumido da seguinte forma: reajustar os preditores, utilizando o mesmo conjunto de dados, excluindo a n-ésima observação. Como notação, $v_k^{(-n)}(\boldsymbol{x})$ representa os K preditores obtidos ao excluir o n-ésimo caso. Assim, obtém-se o vetor com K entradas z_n dado por

$$z_{kn} = v_k^{(-n)}(\boldsymbol{x}_n) \tag{4.20}$$

Desta forma, um novo conjunto de dados é criado, representado por $\{(y_n, z_n), n = 1, \ldots, N\}$. Uma alternativa para selecionar um único preditor seria escolher k que minimize $\sum_k (y_n - z_{kn})^2$. No entanto, o novo conjunto de dados possui informação adicional que pode ser utilizada para combinar os preditor e melhorar a acurácias da predição. A predição final seria dada pela seguinte combinação linear:

$$v(\boldsymbol{x}) = \sum_k \alpha_k v_k(\boldsymbol{x})$$

Obter os coeficientes α de modo a minimizar o erro de predição apresenta-se como um problema de regressão linear. O objetivo, portanto, seria minimizar

$$\sum_n \left(y_n - \sum_k \alpha_k v_k(\boldsymbol{x}_n) \right)^2 \tag{4.21}$$

No entanto, como os preditores foram treinados no mesmo conjunto de aprendizagem e os coeficientes α foram obtidos minimizando o erro sobre este mesmo conjunto, estes coeficientes podem causar *overfit*, fazendo com que a generalização seja pobre. Este problema pode ser atacado por meio de validação cruzada, excluindo-se do conjunto de treinamento a observação a ser predita. Assim, a função objetivo (4.21) seria modificada para

$$\sum_n \left(y_n - \sum_k \alpha_k z_{kn} \right)^2 \tag{4.22}$$

em que $z_{kn} = v_k^{(-n)}(\boldsymbol{x}_n)$. Há um segundo problema, a saber, os preditores $v_k(\boldsymbol{x}_n)$ serão altamente correlacionados, uma vez que tentam predizer os mesmos valores. O método para estimar coeficientes de regressão para variáveis altamente correlacionadas é a *ridge regression*, que consiste em minimizar (4.22) adicionando a restrição de que $\sum \alpha_k^2 = s$ em que o valor de s seria escolhido por validação cruzada.

Em seu trabalho, Breiman obteve resultados consistentes minimizando (4.22) sob as restrições $\alpha_k \geq 0$, $k = 1, \ldots, K$. Além disso, também foi apresentada evidência em dados sintéticos de que a a validação cruzada 10-folds, muito menos custosa em termos computacionais, é mais efetiva do que o método *leave-one-out*. Estas recomendações foram seguidas neste trabalho.

Para o caso em tela, em que se trata da alocação de observações em *clusters* de dados tipo-intervalo, algumas adaptações ao método de *Stacked Regressions* são necessárias. Em primeiro lugar, o protótipo de cada *cluster* é considerado um possível preditor para a amostra. Estes protótipos são os modelos selecionados pelo algoritmo iCRCNLR. Além disso, como são tratados dados do tipo-intervalo, as predições devem ser feitas tanto para o centro quanto para a amplitude.

Considere-se que o algoritmo fornece K *clusters*, cada um deles com um modelo ajustado para o centro $f^c_{(1)}(\boldsymbol{x}_i), \ldots, f^c_{(K)}(\boldsymbol{x}_i)$ e para a amplitude $f^r_{(1)}(\boldsymbol{x}_i), \ldots, f^r_{(K)}(\boldsymbol{x}_i)$. Assim, utilizando validação cruzada N-*folds* [1], o conjunto de dados (4.20) torna-se

$$z^c_{kn} = f^{c(-n)}_k(\boldsymbol{x}^c_n) \tag{4.23}$$

$$z^r_{kn} = f^{r(-n)}_k(\boldsymbol{x}^r_n) \tag{4.24}$$

[1] Normalmente utiliza-se a notação K-folds, foi utilizada a letra N para não causar confusão com o K referente ao número de *clusters* fornecidos pelo modelo iCRCNLR

para o centro e amplitude, respectivamente, em que $1 \leq n \leq N$ Note que cada z_n^c e z_n^r é uma matriz de dimensão $|f| \times K$, em que $|f|$ representa a cardinalidade do *fold* excluído n.

Seguindo o método de Breiman, o objetivo final será encontrar $\boldsymbol{\alpha}^c = (\alpha_1^c, \ldots, \alpha_K^c)$ e $\boldsymbol{\alpha}^r = (\alpha_1^r, \ldots, \alpha_K^r)$ que minimizem

$$\sum_n \left[\left(y_n^c - \sum_k \alpha_k^c z_{kn}^c \right)^2 + \left(y_n^r - \sum_k \alpha_k^r z_{kn}^r \right)^2 \right] \tag{4.25}$$

Desta forma, a predição final será dada por

$$f^c(\boldsymbol{x}) = \sum_k \alpha_k^c f_k^c(\boldsymbol{x}) \tag{4.26}$$

$$f^r(\boldsymbol{x}) = \sum_k \alpha_k^r f_k^r(\boldsymbol{x}) \tag{4.27}$$

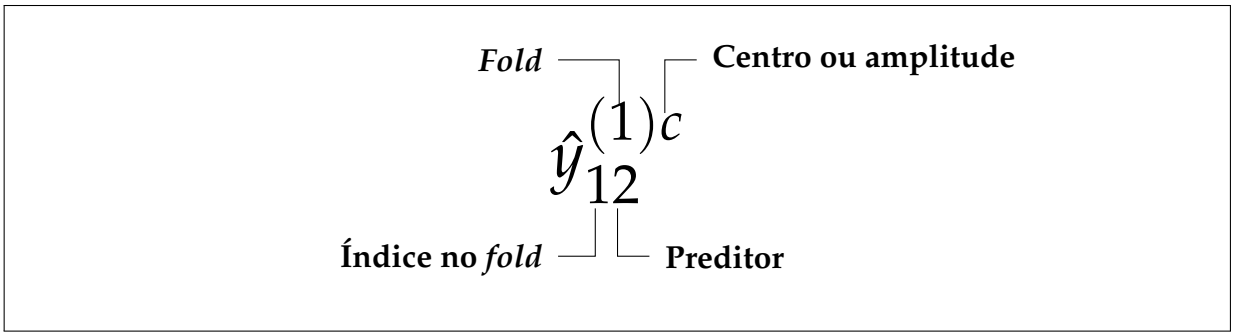
Em termo de intervalos, tem-se que $Z^L = f^c(\boldsymbol{x}) - f^r(\boldsymbol{x})$ e $Z^U = f^c(\boldsymbol{x}) + f^r(\boldsymbol{x})$. Em um experimento de validação cruzada para a regressão *clusterwise* para dados intervalo, os seguintes passos devem ser adotados para executar *Stacked Regressions*:

1. Particionar o conjunto de dados em L-*folds*;
2. Separar um dos *folds* para teste e o restante para treino;
3. Fazer outra validação cruzada dentro dos dados de treino, ou seja, particionar o treino em L_{treino} *folds*;
4. Utilizar $L_{treino} - 1$ *folds* para treinar o modelo e prever os resultados do *fold* excluído.
5. Para cada observação do *fold* excluído, obter $\hat{y_n}_1^c, \ldots, \hat{y_n}_K^c, y_n$ e $\hat{y_n}_1^r, \ldots, \hat{y_n}_K^r, y_n, n = 1, \ldots, N_f$, em que N_f é o número de elementos dentro do *fold* excluído;
6. Repetir o procedimento acima para todos os L_{treino}*folds* do conjunto de treino;
7. A partir dos dados obtidos em todos os *folds*, estimar $\boldsymbol{\alpha}^c$ e $\boldsymbol{\alpha}^r$;
8. Utilizar as equações (4.26) e (4.27) para prever os valores do *fold* separado para teste no passo 2.
9. Repetir o procedimento para os outros $L - 1$ *folds* do conjunto de dados.

O conjunto de dados obtido pelo método de *Stacked regressions* para estimar os parâmetros $\boldsymbol{\alpha}^c$ e $\boldsymbol{\alpha}^r$ é representado na Tabela 1, em que $\sum_{i=1}^{L} N_{fi} = N$. As entradas na Tabela 1 possuem índices que devem ser interpretados como no exemplo abaixo:

Tabela 1 – Conjunto de dados gerado pelo método *Stacked Regressions* para estimar $\boldsymbol{\alpha}^c$ e $\boldsymbol{\alpha}^r$.

y	$\hat{y}_1$	$\hat{y}_2$	$\cdots$	$\hat{y}_K$	Fold
$y_1^{(1)}$	$\left[\hat{y}_{11}^{(1)c}, \hat{y}_{11}^{(1)r}\right]$	$\left[\hat{y}_{12}^{(1)c}, \hat{y}_{12}^{(1)r}\right]$	$\cdots$	$\left[\hat{y}_{1K}^{(1)c}, \hat{y}_{1K}^{(1)r}\right]$	
$y_2^{(1)}$	$\left[\hat{y}_{21}^{(1)c}, \hat{y}_{21}^{(1)r}\right]$	$\left[\hat{y}_{22}^{(1)c}, \hat{y}_{22}^{(1)r}\right]$	$\cdots$	$\left[\hat{y}_{2K}^{(1)c}, \hat{y}_{2K}^{(1)r}\right]$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
$y_{N_{f1}}^{(1)}$	$\left[\hat{y}_{N_{f1}1}^{(1)c}, \hat{y}_{N_{f1}1}^{(1)r}\right]$	$\left[\hat{y}_{N_{f1}2}^{(1)c}, \hat{y}_{N_{f1}2}^{(1)r}\right]$	$\cdots$	$\left[\hat{y}_{N_{f1}K}^{(1)c}, \hat{y}_{N_{f1}K}^{(1)r}\right]$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$y_1^{(L)}$	$\left[\hat{y}_{11}^{(L)c}, \hat{y}_{11}^{(L)r}\right]$	$\left[\hat{y}_{12}^{(L)c}, \hat{y}_{12}^{(L)r}\right]$	$\cdots$	$\left[\hat{y}_{1K}^{(L)c}, \hat{y}_{1K}^{(L)r}\right]$	
$y_2^{(L)}$	$\left[\hat{y}_{21}^{(L)c}, \hat{y}_{21}^{(L)r}\right]$	$\left[\hat{y}_{22}^{(L)c}, \hat{y}_{22}^{(L)r}\right]$	$\cdots$	$\left[\hat{y}_{2K}^{(L)c}, \hat{y}_{2K}^{(L)r}\right]$	L
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
$y_{N_{fL}}^{(L)}$	$\left[\hat{y}_{N_{fL}1}^{(L)c}, \hat{y}_{N_{fL}1}^{(L)r}\right]$	$\left[\hat{y}_{N_{fL}2}^{(L)c}, \hat{y}_{N_{fL}2}^{(L)r}\right]$	$\cdots$	$\left[\hat{y}_{N_{fL}K}^{(L)c}, \hat{y}_{N_{fL}K}^{(L)r}\right]$	



Neste capítulo foi apresentado o método proposto, iCRCNLR, seu algoritmo e prova de convergência da série por ele originada. Além disso, foram definidos os métodos de alocação de novas observações, já que a tarefa de predição pressupõe que o (i) objeto seja alocado a um *cluster* - Alocação aleatório ou KNN - ou que (ii) seja possível fornecer uma predição para o objeto resultante de uma combinação linear de K predições - *Stacked Regressions*.

Com o algoritmo iCRCNLR definido, procede-se a análise experimental tendo como objetivo mensurar seu desempenho frente ao caso linear, aqui chamado de iCRCLR. Tal tarefa é feita no capítulo seguinte. São executadas duas simulações em 24 cenários sintéticos, construídos de acordo com diferentes estruturas de *clusters*.

A primeira simulação mensura a capacidade de estimação do método, ao utilizar heurísticas de otimização, dado uma função conhecida. Deve-se salientar que a convergência do método depende da qualidade das estimativas, pois são elas quem minimizam a função objetivo no primeiro passo do algoritmo. A segunda simulação tem

como função mensurar a capacidade de predição do método iCRCNLR comparativamente ao iCRCLR. Aplicações seis conjuntos de dados tipo-intervalo são executadas, também comparando a capacidade de predição dos algoritmos em questão.

5 ANÁLISE EXPERIMENTAL

Neste capítulo, é investigado o comportamento do algoritmo iCRCNLR no que diz respeito à estimação de parâmetros e capacidade de predição de novas observações. Foram feitas simulações em 24 diferentes cenários, cada um representando uma estrutura de *clusters* de centro e amplitudes para dados intervalo. Estes cenários foram construídos de acordo com as seguintes características: a posição relativa dos *clusters*, denominado *configuração*; o tipo de função utilizado para gerar os dados, podendo ser linear ou não linear e o número de classes $k = 2, 3$. Em relação à configuração, foram gerados cenários com classes disjuntas (D-D), com interseção (I-I) e sobrepostas (U-U) na variável dependente X.

5.1 EXPERIMENTOS COM DADOS SINTÉTICOS

Uma importante tarefa consiste em realizar a predição de novas observações baseado em amostras de treinamento. Avaliar o desempenho do iCRCNLR nesta tarefa é o objetivo do segundo estudo de simulação. Neste caso, há o problema de alocação de novas amostras, a saber, uma vez que os *clusters* e protótipos estão definidos, como alocar uma nova observação a apenas um destes *clusters*. Foram utilizados três métodos de alocação: *k*-vizinhos mais próximos (KNN) para dados intervalo com distância de *Hausdorff*, *Stacked Regressions* (BREIMAN, 1996) e Alocação aleatória (Random Allocation). As funções apresentadas na tabela 2 foram ajustadas para os dados sintéticos no algoritmo iCRCNLR.

O primeiro passo para executar o experimento com dados sintéticos baseia-se em definir o modo de geração de amostras que representarão os cenários desejados de dados intervalo. As amostras geradas têm um número de 50 observações por classe. A função utilizada para gerar os dados não lineares de centro e amplitude é

$$f(x) = x^{\alpha_0 - 1} exp\left\{\frac{-x}{\alpha_1}\right\} \tag{5.1}$$

O uso desta função se justifica pela sua flexibilidade, ilustrada na Figura 1. Isto permite que sejam gerados dados segundo uma relação não linear em qualquer parte do eixo x, facilitando a construção de cenários com *clusters* não lineares sobrepostos, disjuntos ou com interseção.

Uma vez que os parâmetros da função (α_0, α_1) são definidos, os dados são gerados a partir do seguinte processo: (i) selecionar a função para representar a relação entre as variáveis (linear ou não linear (5.1)); (ii) escolher o intervalo no qual X irá variar; (iii) escolher valores dentro deste intervalo, considerando uma distribuição uniforme, ou seja, selecionar n valores $x_1, \ldots, x_n$ de $X \sim \mathcal{U}(a, b)$, em que a e b são os limites do

Tabela 2 – Funções utilizadas para ajustar os dados.

Rótulo	Função
1	$f(x) = x^{(\alpha_0 - 1)} e^{-x/\alpha_1}$
2	$f(x) = \alpha_0 - \frac{\alpha_1}{\alpha_2 + x}$
3	$f(x) = \frac{\alpha_1 x}{\alpha_0 + x}$
4	$f(x) = \alpha_0 + \alpha_1 x$
5	$f(x) = \alpha_2 + \frac{1-\alpha_2}{1+e^{-\alpha_0 - \alpha_1 \log x}}$
6	$f(x) = \frac{1}{1+e^{-\alpha_0 - \alpha_1 \log x}}$
7	$f(x) = \alpha_2 + \frac{1-\alpha_2}{1+e^{-\alpha_0 - \alpha_1 x}}$
8	$f(x) = \frac{1}{1+e^{-\alpha_0 - \alpha_1 x}}$
9	$f(x) = \alpha_0 + (1 - \alpha_0)(1 - e^{-\alpha_1 x})$
10	$f(x) = 1 - e^{-\alpha_0 x}$
11	$f(x) = \alpha_0 + (1 - \alpha_0)(1 - e^{-\alpha_1 x - \alpha_2 x^2})$
12	$f(x) = 1 - e^{-\alpha_0 x - \alpha_1 x^2}$
13	$f(x) = 1 - e^{-\alpha_0 x - \alpha_1 x^2 - \alpha_2 x^3}$

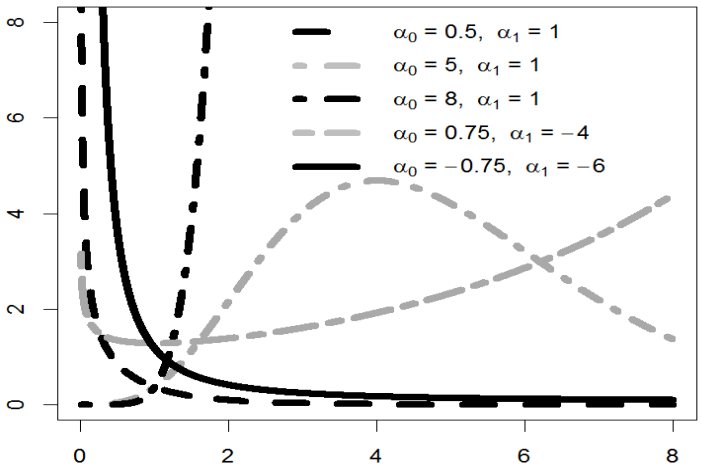


Figura 1 – Formas assumidas pela função (5.1) com diferentes valores dos parâmetros α_0 e α_1

intervalo escolhido em (ii); (iv) computar os n valores aplicados à função selecionada no passo (i), gerando n observações da variável dependente $y_i = f(x_i)$; (v) adicionar um ruído aleatório $\epsilon \sim \mathcal{U}(e_1, e_2)$ a estas observações. O resultado final é a variável de-

pendente y. Para os cenário gerados neste trabalho, foi definidos que $\epsilon \sim \mathcal{U}(0.01, 0.1)$. Esta pequena variação foi adicionada aos dados independentemente da escala no eixo y, de modo que o erro inerente aos dados seja semelhante em todos os cenários. Assim, pode-se ter uma ideia melhor do erro cometido pelo modelo, não sendo confundido com esta perturbação adicionada aos dados. Este procedimento foi utilizado para gerar tanto os dados de centro quanto de amplitude.

Os cenários podem ser divididos em quatro conjuntos, de acordo com a presença de não linearidade no centro e na amplitude. Em cada conjunto foram gerados cenários com configuração disjunta (D-D), interseção (I-I) e sobreposição (U-U). O primeiro conjunto de cenários possui centro e amplitude lineares, o segundo tem centro linear e amplitude não linear; o terceiro, centro não linear e amplitude linear; finalmente, o último conjunto de cenários possui centro e amplitude não lineares.

As Tabelas 3 até 6 apresentam detalhes destes cenários, isto é, a configuração das posições relativas do centros e amplitudes das classes, o tipo de função que define a relação entre as variáveis dependente e independente, os parâmetros da função utilizada para gerar os dados, os valores aleatórios tomados dentro de um intervalo $[a, b]$ para simular realizações da variável independente X e o número de classes geradas. A representação gráfica dos 24 cenários gerados pode ser vista no apêndice A.

Os cenários 1 ao 6, apresentados na Tabela 3, simulam uma situação em que tanto centro quanto amplitude dos intervalos são definidos por funções lineares. Este é um cenário favorável à aplicação do algoritmo iCRCLR, já que as funções que definem centro e amplitude dos dados são lineares. Já o algoritmo iCRCNLR deve apresentar desempenho semelhante ao caso linear, com diferenças devidas à utilização de métodos heurísticos de otimização, enquanto o iCRCLR computa as estimativas de parâmetros algebricamente.

A Tabela 4 apresenta cenários de 7 a 12, com centro definido por funções lineares e amplitude definida por função não-linear. Estes cenários são favoráveis ao algoritmo iCRCNLR, dado a sua capacidade de ajustar funções não-lineares a partir dos dados. A função não-linear que gera os dados pode ou não ser conhecida. No caso em que a função é desconhecida, o algoritmo busca num conjunto de funções aquela que melhor se ajusta aos dados, com base na otimização do critério de adequação. Tal fato permite que sejam ajustadas funções lineares para o centro e funções não-lineares para a amplitude.

A Tabela 5 apresenta as características dos cenários 13 ao 18, com centro definido por funções não-lineares e amplitude gerada por funções lineares. De fato, este conjunto de cenários é uma inversão dos centros e amplitudes apresentados na Tabela 4. Espera-se que as estimativas sejam semelhantes, porém, a inversão do centro e amplitude gera um novo conjunto de dados onde as predições de novas observações serão diferentes.

Tabela 3 – Cenários gerados com centro e amplitude lineares.

Cenário	Configuração		Função		Parâmetros		$X \sim \mathcal{U}(a,b)$		Classes
	Centro	Amplitude	Centro	Amplitude	Centro	Amplitude	Centro	Amplitude	
1	D	D	Linear	Linear	(3,-1) (0.5,-1)	(3,1) (5,0.75)	(0,3) (4,8)	(0,3) (4,8)	2
2	D	D	Linear	Linear	(4,1) (2,2) (1,3)	(6,2) (6,-1) (1,3)	(0,2) (3,5) (6,9)	(0,2) (3,5) (6,9)	3
3	I	I	Linear	Linear	(3,-1) (0.5,1)	(3,1) (5,-0.5)	(0,3) (2,5)	(0,3) (2,5)	2
4	I	I	Linear	Linear	(4,1) (2,2) (1,3)	(6,2) (6,-1) (1,3)	(0,4) (3,6) (5,8)	(0,4) (3,6) (5,8)	3
5	U	U	Linear	Linear	(3,-1) (0.5,1)	(3,1) (5,-0.5)	(0,3) (0,3)	(0,3) (0,3)	2
6	U	U	Linear	Linear	(4,1) (2,2) (1,3)	(6,2) (6,-1) (1,3)	(0,3) (0,3) (0,3)	(0,3) (0,3) (0,3)	3

Tabela 4 – Cenários gerados com centro linear e amplitude não linear.

Cenário	Configuração		Função		Parâmetros		$X \sim \mathcal{U}(a,b)$		Classes
	Centro	Amplitude	Centro	Amplitude	Centro	Amplitude	Centro	Amplitude	
7	D	D	Linear	Non-Linear	(3,-1) (0.5,1)	(0.5,2) (1,-3)	(0,3) (4,8)	(0,3) (4,8)	2
8	D	D	Linear	Non-Linear	(4,1) (2,2) (1,3)	(0.5,1) (0.75,-4) (0.75,-6)	(0,2) (3,5) (6,9)	(0,6) (7,12) (14,20)	3
9	I	I	Linear	Non-Linear	(3,-1) (0.5,1)	(0.5,2) (1,-3)	(0,3) (2,5)	(0,4) (2,5)	2
10	I	I	Linear	Non-Linear	(4,1) (2,2) (1,3)	(-5,1) (0.75,-4) (-0.75,-6)	(0,4) (3,6) (5,8)	(0,4) (2,8) (5,10)	3
11	U	U	Linear	Non-Linear	(3,-1) (0.5,1)	(0.5,2) (1,-3)	(0,3) (0,3)	(0,4) (0,4)	2
12	U	U	Linear	Non-Linear	(4,1) (2,2) (1,3)	(0.5,1) (0.75,-4) (0.75,-6)	(0,3) (0,3) (0,3)	(0,10) (0,10) (0,10)	3

Tabela 5 – Cenários gerados com centro não linear e amplitude linear.

Cenário	Configuração		Função		Parâmetros		$X \sim \mathcal{U}(a,b)$		Classes
	Centro	Amplitude	Centro	Amplitude	Centro	Amplitude	Centro	Amplitude	
13	D	D	Non-Linear	Linear	(0.5,2) (1,-3)	(3,-1) (0.5,1)	(0,3) (4,8)	(0,3) (4,8)	2
14	D	D	Non-Linear	Linear	(0.5,1) (0.75,-4) (0.75,-6)	(4,1) (2,2) (1,3)	(0,6) (7,12) (14,20)	(0,2) (3,5) (6,9)	3
15	I	I	Non-Linear	Linear	(0.5,2) (1,-3)	(3,-1) (0.5,1)	(0,4) (2,5)	(0,5) (2,6)	2
16	I	I	Non-Linear	Linear	(0.5,1) (0.75,-4) (0.75,-6)	(4,1) (2,2) (1,3)	(0,10) (5,15) (10,20)	(0,4) (3,6) (5,8)	3
17	U	U	Non-Linear	Linear	(0.5,2) (1,-3)	(3,-1) (0.5,1)	(0,4) (0,4)	(0,3) (0,3)	2
18	U	U	Non-Linear	Linear	(0.5,1) (0.75,-4) (0.75,-6)	(4,1) (2,2) (1,3)	(0,10) (0,10) (0,10)	(0,3) (0,3) (0,3)	3

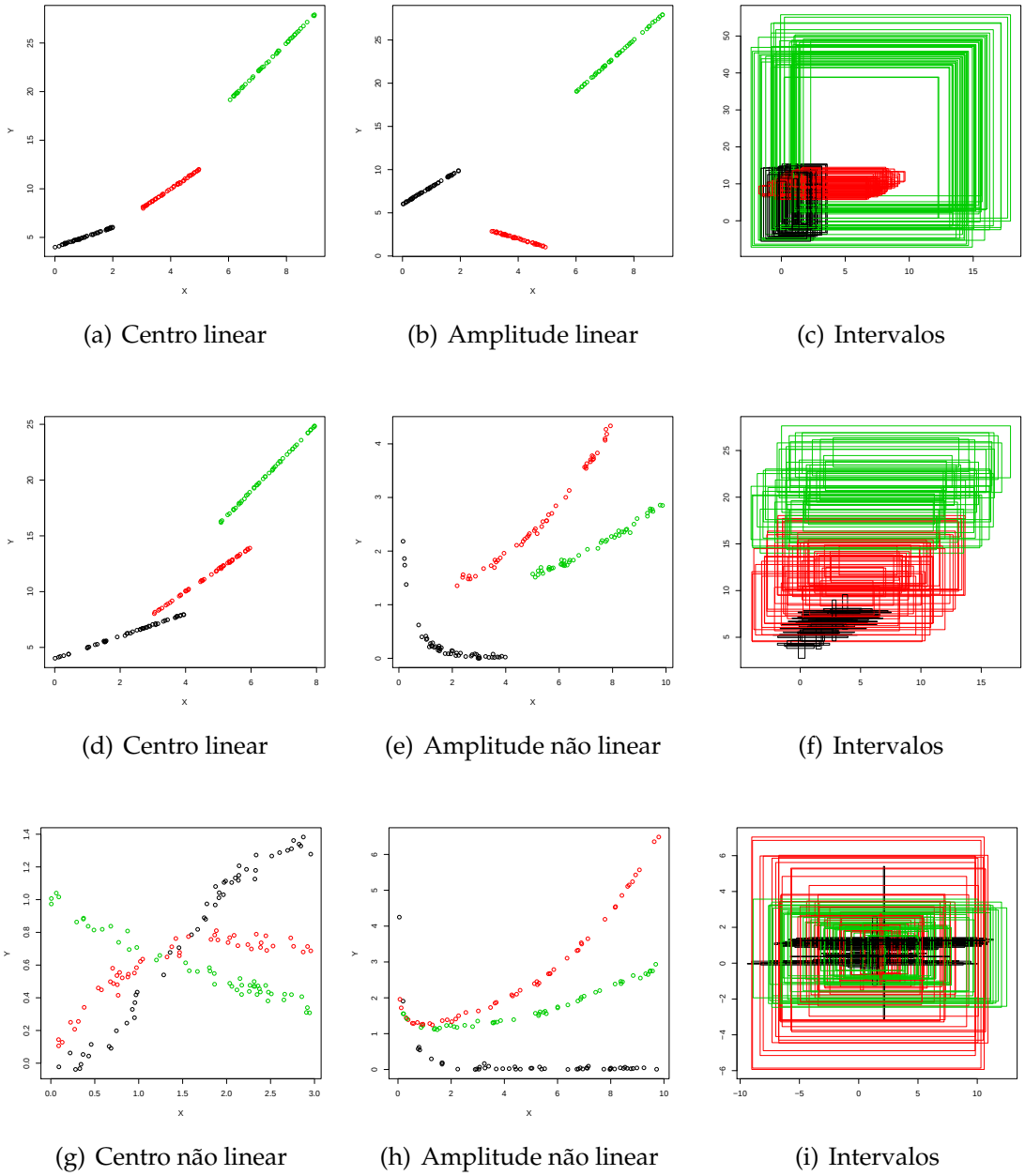
Finalmente, na Tabela 6 são apresentados os cenários 19 a 24, onde tanto o centro quanto a amplitude são definidos por funções não-lineares. Note-se que este conjunto de cenários é o mais propício à aplicação do algoritmo iCRCNLR.

Tabela 6 – Cenários gerados com centro e amplitude não lineares.

Cenário	Configuração		Função		Parâmetros		$X \sim \mathcal{U}(a,b)$		Classes
	Centro	Amplitude	Centro	Amplitude	Centro	Amplitude	Centro	Amplitude	
19	D	D	Non-Linear	Non-Linear	(0.5,1) (0.75,-3)	(0.5,2) (1,-3)	(0,3) (4,8)	(0,3) (4,8)	2
20	D	D	Non-Linear	Non-Linear	(0.5,1) (0.5,-3) (0.75,-4)	(0.5,1) (0.75,-4) (0.75,-6)	(0,6) (7,12) (14,20)	(0,6) (7,12) (14,20)	3
21	I	I	Non-Linear	Non-Linear	(0.5,1) (0.75,-3)	(0.5,2) (1,-3)	(0,4) (2,5)	(0,4) (2,5)	2
22	I	I	Non-Linear	Non-Linear	(0.5,1) (0.5,-3) (0.75,-4)	(0.5,1) (0.75,-4) (0.75,-6)	(0,4) (2,8) (5,10)	(0,4) (2,8) (5,10)	3
23	U	U	Non-Linear	Non-Linear	(0.5,1) (0.75,-3)	(0.5,2) (1,-3)	(0,4) (0,4)	(0,4) (0,4)	2
24	U	U	Non-Linear	Non-Linear	(4,1) (2,2) (1,3)	(0.5,1) (0.75,-4) (0.75,-6)	(0,3) (0,3) (0,3)	(0,10) (0,10) (0,10)	3

A Figura 2 apresenta alguns dos cenários. A primeira fila de figuras representa uma estrutura de 3 classes com relação linear entre X e Y, correspondente ao cenário 2. Neste caso, $X_1^c \sim \mathcal{U}(0,2)$, $X_2^c \sim \mathcal{U}(3,5)$, $X_3^c \sim \mathcal{U}(6,9)$, e $X_1^r \sim \mathcal{U}(0,2)$, $X_2^r \sim \mathcal{U}(3,5)$, $X_3^r \sim \mathcal{U}(6,9)$. O segundo cenário tem relação linear no centro e não linear na amplitude. Neste caso, $X_1^c \sim \mathcal{U}(0,4)$, $X_2^c \sim \mathcal{U}(3,6)$, $X^c 3 \sim \mathcal{U}(5,8)$, and $X_1^r \sim \mathcal{U}(0,4)$, $X_2^r \sim \mathcal{U}(2,8)$, $X_3^r \sim \mathcal{U}(5,10)$. Finalmente, o terceiro cenário possui centro e amplitude não lineares

com $X_1^c \sim \mathcal{U}(0,3)$, $X_2^c \sim \mathcal{U}(0,3)$, $X_3^c \sim \mathcal{U}(0,3)$, e $X_1^r \sim \mathcal{U}(0,10)$, $X_2^r \sim \mathcal{U}(0,10)$, $X_3^r \sim \mathcal{U}(0,10)$. Estes cenários também ilustram estruturas de classes disjuntas (D-D), com interseção (I-I) e sobrepostos (U-U), respectivamente.



(a) Centro linear (b) Amplitude linear (c) Intervalos

(d) Centro linear (e) Amplitude não linear (f) Intervalos

(g) Centro não linear (h) Amplitude não linear (i) Intervalos

Figura 2 – Exemplos para o centro, amplitude e intervalos gerados pelos cenários 2, 10 e 24. Estes cenários são exemplos de configurações de classes disjuntas (D-D), com interseção (I-I) e sobrepostas (U-U), respectivamente.

Os cenários gerados cobrem uma gama de situações com diferentes graus de dificuldade, tanto na estimação quanto na alocação de novas observações. Espera-se que cenários com configuração D-D, ou seja, aqueles em que as classes de centro e amplitude são disjuntas, tendem a facilitar a alocação de uma nova observação ao grupo adequado. Nos cenários em que há interseção (I-I) e sobreposição (U-U) nos centros e

amplitudes no eixo X tendem a apresentar um grau maior de dificuldade na alocação de novas observações. A seguir, são apresentados os resultados das simulações para *estimação*, onde as funções que definem o centro e a amplitude são conhecidas, isto é, a função a ser ajustada é a mesma que foi utilizada para gerar os dados, suprimindo a busca pelo melhor par de funções. Também são apresentados os resultados da simulação referente a *predição*, onde o algoritmo encontra o melhor par de modelos para o centro e amplitude antes de selecionar um grupo para cada observação de teste por meio de três métodos (KNN, *Stacked Regressions* e aleatório) e obter o valor predito para a variável Y.

5.1.1 Estimação

Nesta seção é realizado um estudo para validar a utilização de métodos de otimização numérica para obter as estimativas de parâmetros. A utilização de heurísticas de otimização para obter as estimativas dos modelos de regressão não-linear é justificada por dois fatores: (i) do ponto de vista prático, obter a solução do sistema de equações normais para um modelo de regressão não-linear em forma fechada pode não ser possível e (ii) o método iCRCNLR deve ser capaz de ajustar um conjunto grande de funções não-lineares aos dados, não sendo viável exigir do usuário que se compute todos os sistemas de equações normal exigidos.

No algoritmo iCRCNLR, utiliza-se um esquema para evitar estimativas errôneas decorrentes da não convergência de métodos de otimização, como SANN, CG e BFGS. A não convergência dos métodos de otimização terá como resultado estimativas ruins para os modelos de regressão em cada *cluster*. Estas estimativas, por sua vez, não resultarão na minimização do erros e por fim, no critério J.

O esquema para evitar a obtenção de estimativas de parâmetros de regressão ruins, em decorrência da não convergência de heurísticas de otimização, consiste iterar sobre estas heurísticas, sendo o critério de parada a convergência. Em outras palavras, dado que um método não converge, outro método é acionado e assim sucessivamente. Finalmente, no caso de nenhuma das heurísticas convergir para as estimativas de um modelo de um *cluster* específicos, o modelo da iteração anterior é mantido e o passo seguinte de alocação é executado, sem prejuízo sobre o minimização do critério J.

Para cada cenário, 1000 conjuntos de dados foram gerados. Como a solução final do iCRCNLR depende da partição inicial, convergindo para um ótimo local, o algoritmo foi executado 100 vezes e o melhor resultado de acordo com o critério de adequação, menor valor de J, foi selecionado. As métricas apresentadas nas tabelas correspondem ao valor médio dos 1000 conjuntos de dados.

As tabelas 7 a 10 mostram a raiz quadrada do erro quadrático médio (root mean squared error - RMSE) e a média das estimativas de mínimos quadrados para os 24 cenários de dados sintéticos. A função não-linear a ser estimada (5.1) é mantida fixa,

uma vez que se pretende medir a qualidade da estimação para uma função conhecida. A expressão para o RMSE é dada por:

$$RSME_{\alpha_0} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{\alpha_0} - \alpha_0)^2}{n}} \tag{5.2}$$

$$RSME_{\alpha_1} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{\alpha_1} - \alpha_1)^2}{n}} \tag{5.3}$$

em que $\hat{\alpha_0}$ e $\hat{\alpha_1}$ são as estimativas dos parâmetros α_0 e α_1, respectivamente, e n é o número de observações na amostra. No contexto de uma simulação, o valor dos parâmetros é controlado, o que permite o cálculo da expressão (5.2). Os resultados mostram que, em geral, a metodologia introduzida para obter estimativas de parâmetros utilizando heurísticas de otimização mostrou-se eficiente. A média das estimativas mostra que o viés da estimação é baixo. Além disso, a média do RMSE das estimativas é baixo em todos os cenários, aumentando ou diminuindo em função apenas da escala das variáveis, como nas estimativas de amplitude dos cenários 22 e 24 (Tabela 9). Com a função geradora dos dados conhecida, infere-se a partir da qualidade das estimativas que o algoritmo também consegue agrupar os dados eficientemente.

A partir da conclusão de que o algoritmo iCRCNLR consegue obter boas estimativas de parâmetros para uma dada função, o estudo passa a medir a qualidade da predição fornecida pelo método, em comparação com o caso linear.

5.1.2 Predição

Nesta seção, são apresentados os resultados do experimento que compara a capacidade de predição do algoritmo iCRCNLR em comparação com o caso linear, iCRCLR. Foi executado um esquema de validação cruzada 10 *times* 10 *folds* para investigar qual método de alocação é mais adequado em cada um dos 24 cenários de dados sintéticos apresentados. As Tabelas 11 a 26 apresentam a média, o desvio padrão e o valor mínimo dos RMSE's das 10 repetições da validação cruzada, com o menor valor do RMSE médio em cada cenário destacado. Para testar a hipótese nula de que os RMSE's são oriundos da mesma distribuição, ou possuem diferença de medianas igual a zero, utilizou-se o teste de Mann-Whitney. Este teste é adequado quando o número de amostras é pequeno e não há suposição de normalidade dos dados. Este teste mostra-se adequado, uma vez que é robusto em relação a *outliers* que foram observados nas amostras.

Para cada cenário, o método iCRCNLR é executado 100 vezes em 9 *folds* e o resultado que apresenta o menor critério J no conjunto de treinamento é escolhido. Isto ocorre em função do algoritmo convergir para um ótimo local, dependendo da partição inicial aleatória. Como ocorre nos esquemas de validação cruzada, os indivíduos deixados no *fold* de teste são tratados como novas observações para as quais o algoritmo deverá prever os valores da variável resposta.

Tabela 7 – RMSE médio e média das estimativas dos parâmetros para cenários com centro e amplitude lineares

Scenario	Classes	Média das estimativas				RMSE			
		Centro		Amplitude		Centro		Amplitude	
		α_0	α_1	α_0	α_1	α_0	α_1	α_0	α_1
1	2	3.0003	-1.0003	3.0000	1.0002	0.0145	0.0085	0.0162	0.0089
		0.5016	-1.0002	5.0028	0.7496	0.0325	0.0055	0.0379	0.0063
2	3	3.9961	1.0087	5.9987	2.0012	0.0185	0.0213	0.0177	0.0156
		2.0123	1.9970	5.9450	-0.9866	0.0599	0.0149	0.0991	0.0245
		1.0184	2.9975	1.0238	2.9972	0.0743	0.0099	0.0907	0.0117
3	2	3.0003	-1.0001	2.9991	1.0006	0.0149	0.0090	0.0159	0.0092
		0.4999	1.0002	5.0025	-0.5007	0.0295	0.0082	0.0323	0.0090
4	3	3.9984	1.0008	6.0003	2.0000	0.0155	0.0068	0.0166	0.0073
		1.9992	2.0002	5.9828	-0.9963	0.0418	0.0092	0.0456	0.0099
		1.0105	2.9984	1.0012	2.9998	0.0572	0.0089	0.0571	0.0088
5	2	2.9999	-1.0001	3.0002	1.0001	0.0149	0.0087	0.0170	0.0101
		0.4995	0.9999	5.0004	-0.5004	0.0152	0.0088	0.0166	0.0095
6	3	4.0006	0.9999	5.9997	2.0002	0.0143	0.0083	0.0169	0.0097
		2.0003	2.0000	5.9990	-0.9999	0.0138	0.0079	0.0170	0.0096
		1.0000	2.9997	0.9998	3.0003	0.0140	0.0081	0.0162	0.0095

Tabela 8 – RMSE médio e média das estimativas dos parâmetros para cenários com centro linear e amplitude não linear

Scenario	Classes	Média das estimativas				RMSE			
		Centro		Amplitude		Centro		Amplitude	
		α_0	α_1	α_0	α_1	α_0	α_1	α_0	α_1
7	2	3.0009	-1.0006	0.4998	2.0033	0.0145	0.0083	0.0055	0.0510
		0.5013	0.9999	0.9978	-2.9960	0.0144	0.0027	0.0057	0.0129
8	3	4.0004	0.9992	0.5153	0.7238	0.0142	0.0121	0.0362	0.3551
		1.9953	2.0011	0.7444	-3.9863	0.0764	0.0187	0.0147	0.0496
		1.0388	2.9950	0.7443	-5.9728	0.0983	0.0131	0.0087	0.0447
9	2	2.9995	-0.9998	0.5003	2.0125	0.0144	0.0083	0.0081	0.0580
		0.4996	1.0001	0.9972	-2.9932	0.0266	0.0073	0.0093	0.0261
10	3	3.9999	1.0000	0.5056	1.0176	0.0146	0.0064	0.0128	0.0351
		1.9891	2.0022	0.7732	-4.1711	0.0486	0.0106	0.0381	0.2287
		1.0724	2.9890	0.7474	-5.9861	0.0939	0.0147	0.0108	0.0854
11	2	3.0000	-0.9996	0.5017	2.0145	0.0143	0.0082	0.0087	0.0572
		0.5006	0.9997	0.9979	-2.9962	0.0149	0.0085	0.0129	0.0415
12	3	3.9992	1.0006	0.5266	0.8959	0.0155	0.0091	0.0507	0.2060
		2.0011	1.9992	0.7521	-4.0236	0.0173	0.0098	0.0164	0.0752
		0.9995	3.0003	0.7399	-5.9337	0.0150	0.0089	0.0193	0.1574

Tabela 9 – RMSE médio e média das estimativas dos parâmetros para cenários centro não linear e amplitude linear

Scenario	Classes	Média das estimativas				RMSE			
		Centro		Amplitude		Centro		Amplitude	
		α_0	α_1	α_0	α_1	α_0	α_1	α_0	α_1
13	2	0.5002	2.0007	2.999	-0.9993	0.0057	0.0500	0.0145	0.0082
		0.9958	-2.9924	0.503	0.9995	0.0073	0.0155	0.0157	0.0029
14	3	0.5218	1.0264	4.0007	0.9992	0.0354	0.0477	0.0153	0.0134
		0.7485	-3.9978	2.0021	1.9997	0.0099	0.0346	0.0725	0.0179
		0.7472	-5.9860	1.0300	2.9959	0.0060	0.0322	0.0837	0.0111
15	2	0.5002	2.0105	3.0009	-1.0004	0.0083	0.0633	0.0155	0.0087
		0.9975	-2.9956	0.5011	0.9998	0.0108	0.0308	0.0280	0.0079
16	3	0.5051	1.0202	4.0005	0.9998	0.0130	0.0394	0.0142	0.0064
		0.7711	-4.1510	1.9856	2.0030	0.0325	0.1971	0.0485	0.0105
		0.7477	-5.9917	1.0549	2.9917	0.0111	0.0891	0.0752	0.0118
17	2	0.5017	2.0146	3.0007	-1.0003	0.0086	0.0608	0.0155	0.0089
		0.9989	-2.9978	0.5005	0.9999	0.0119	0.0378	0.0158	0.0089
18	3	0.5253	1.0367	4.0023	0.9990	0.0466	0.0654	0.0148	0.0087
		0.7484	-4.0017	2.0005	1.9997	0.0138	0.0577	0.0149	0.0088
		0.7444	-5.9616	1.0001	2.9998	0.0146	0.1246	0.0150	0.0086

Tabela 10 – RMSE médio e média das estimativas dos parâmetros para cenários com centro e amplitude não linear

Scenario	Classes	Média das estimativas				RMSE			
		Centro		Amplitude		Centro		Amplitude	
		α_0	α_1	α_0	α_1	α_0	α_1	α_0	α_1
19	2	0.4998	1.0009	0.5003	1.9986	0.0045	0.0245	0.0055	0.0529
		0.7490	-2.9981	0.9979	-2.9964	0.0072	0.0179	0.0064	0.0147
20	3	0.4959	1.0013	0.5125	1.0270	0.0175	0.0374	0.0262	0.0484
		0.5068	-3.0166	0.7533	-4.0152	0.0110	0.0249	0.0088	0.0338
		0.7495	-3.9988	0.7474	-5.9880	0.0014	0.0034	0.0055	0.0287
21	2	0.5013	0.9998	0.5003	2.0140	0.0069	0.0281	0.0075	0.0623
		0.7461	-2.9979	0.9971	-2.9939	0.0326	0.0991	0.0097	0.0257
22	3	0.4986	0.9970	0.5070	1.0212	0.0072	0.0282	0.0137	0.0378
		0.5156	-3.0512	0.7705	-4.1425	0.0273	0.0797	0.0352	0.2047
		0.7412	-3.9706	0.7462	-5.9813	0.0164	0.0608	0.0151	0.1256
23	2	0.5001	1.0015	0.5005	2.0076	0.0072	0.0278	0.0083	0.0602
		0.7496	-2.9993	0.9983	-2.9967	0.0068	0.0252	0.0127	0.0407
24	3	4.0035	0.9998	0.5328	1.0381	0.0744	0.0264	0.0611	0.0688
		1.9948	2.0110	0.7430	-3.9790	0.0438	0.0628	0.0163	0.0628
		1.0015	2.9984	0.7392	-5.9237	0.0103	0.0790	0.0186	0.1546

Em cada uma das 10 repetições de validação cruzada, o algoritmo obtém uma partição e um conjunto de funções ajustadas para k *clusters*, $k = 1, 2, 3$. As medidas utilizadas para avaliar a qualidade da predição serão os erros quadráticos médios dos limites inferior ($RMSE_{Lf}$), superior ($RMSE_{Uf}$) e geral ($RMSE_{Of}$), em que o índice f corresponde ao f-ésimo *fold* com $f = 1, \ldots, 10$:

$$RMSE_{Lf} = \sqrt{\frac{\sum_{i=1}^{n_f}(z_i^L - \hat{z}_i^L)^2}{n_f}} \tag{5.4}$$

$$RMSE_{Uf} = \sqrt{\frac{\sum_{i=1}^{n_f}(z_i^U - \hat{z}_i^U)^2}{n_f}} \tag{5.5}$$

$$RMSE_{Of} = \sqrt{\frac{\sum_{i=1}^{n_f}(z_i^L - \hat{z}_i^L)^2 + \sum_{i=1}^{n}(z_i^U - \hat{z}_i^U)^2}{n_f}} \tag{5.6}$$

em que z_i é a i-ésima observação no conjunto de teste, $\hat{z}_i$ é o valor estimado para esta mesma observação e n_f é o número de elementos do conjunto de teste. Estas medidas, por serem dependentes da escala dos dados, devem ser utilizadas apenas para comparações de métodos aplicados num mesmo cenário.

Quando $k \neq 1$, não há alocação de *clusters*, portanto, a avaliação dos métodos de alocação se dá para $k = 2$ e 3. Os métodos de alocação testados foram (i) KNN para dados tipo-intervalo utilizando a distância de Hausdorff, selecionando o número de vizinhos como o valor que minimiza o erro dentre os elementos do conjunto $1, 3, 5, 7, 9$; (ii) *Stacked regressions* e (iii) alocação aleatória.

A distância de Hausdorff foi selecionada em função da ampla utilização na literatura sobre técnicas de agrupamento em dados tipo-intervalo (CHAVENT; LECHEVALLIER, 2002; CARVALHO et al., 2006; CHAVENT et al., 2006). Entretanto, não há impedimentos para utilização de outras distâncias, como *L1*, *L2* (CARVALHO; BRITO; BOCK, 2006) ou *City-Block* (SOUZA; CARVALHO, 2004).

Um dos parâmetros do algoritmo é uma lista de funções candidatas a ajustar os dados de centro e amplitude. A quantidade de funções a serem fornecidas ao algoritmo dependem do conhecimento prévio do problema por parte do pesquisador. O algoritmo deverá indicar o melhor par de funções para o centro e amplitude dos intervalos. O conjunto de funções candidatas utilizadas neste trabalho é apresentado na Tabela 2.

5.1.2.1 Limite inferior

Primeiramente são apresentados os resultados de predição para o limite inferior dos intervalos.

Os métodos de alocação podem ser comparados ao fixar os métodos de ajuste e o número de *clusters*. A Tabela 11 apresenta os resultados para os 6 primeiros cenários, isto é, aqueles em que tanto o centro quanto a amplitude foram gerados a partir de funções lineares. Verifica-se o método KNN apresentou melhores resultados médios, tanto para iCRCNLR quanto para iCRCLR, nos cenários 1 a 4. No cenário 5, com ajuste pelo método iCRCLR, os melhores métodos foram Stacked Regressions e KNN para 2 e 3 *clusters*, respectivamente. No mesmo cenário, com o ajuste feito pelo método iCRCNLR, o método *Stacked Regressions* prevaleceu.

Nos cenários 7 a 12, apresentados na Tabela 12, os dados gerados possuem centro linear e amplitude não linear. O método KNN é o melhor em todos os cenários, exceto no cenário 12 com método de ajuste iCRCNLR, onde o método de Stacked Regressions obteve menor RMSE médio.

Para os cenários 13 a 18 (Tabela 13), com centro não-linear e amplitude linear, o método de alocação KNN tem melhor desempenho nos cenários 13 a 16, além do cenário 17, quando o método de ajuste é o iCRCLR. O método *Stacked Regressions* prevalece no cenário 17 com ajuste iCRCNLR e no cenário 18. Deve-se notar que estes cenários apresentam *clusters* sobrepostos, o que dificulta a alocação por meio do KNN, uma vez que ela é feita com base nos vizinhos mais próximos no eixo x. A presença de interseção ou sobreposição (em relação ao eixo x) faz com que os vizinhos de uma observação de teste possam pertencer a grupos diferentes, de tal maneira que o voto majoritário destes vizinhos indique um *cluster* diferente do qual a observação de teste originalmente pertence.

A Tabela 14 apresenta os resultados dos cenários 19 a 24. Novamente, o método KNN apresenta melhor desempenho nos cenário que são disjuntos (D-D) ou contém uma sobreposição parcial (I-I). No cenário 23, com *clusters* sobrepostos (U-U), o método KNN ainda apresenta melhores resultados, não obstante o fato de que a diferença em relação ao RMSE médio do método *Stacked Regressions* cai acentuadamente. No cenário 24, a combinação linear dos preditores obtida pelo método de *Stacked Regressions* apresenta os melhores resultados

Fixando os métodos de alocação, pode-se comparar o desempenho dos métodos de ajuste dos dados, iCRCLR e iCRCNLR.

Para os cenários de 1 a 6, novamente na Tabela 11, o método proposto iCRCNLR apresenta melhor resultado médio em todos os métodos de alocação, para número de *clusters* maior que 1. Considerando um único *cluster*, o método iCRCNLR também apresentar menor RMSE médio em relação ao caso linear, iCRCLR. No cenário 2, para um *cluster* os métodos apresentam resultado semelhante, enquanto para 2 e 3 *clusters* apresenta-se a seguinte situação: para 2 *clusters*, o método iCRCNLR apresenta melhores resultados, enquanto para 3 *clusters* o método iCRCLR apresenta melhores resultados médios. Entretanto, conforme pode ser verificado na Tabela 15, para o método de

alocação KNN a diferença não foi estatisticamente significativa ($p = 0.078$). Ainda, o valor alto alcançado pela alocação aleatório em 3 *clusters* para o método iCRCNLR no cenário 2 pode ser explicada pelo fato de que tal método ajusta funções não-lineares. Como o crescimento de algumas destas funções se dá de modo exponencial, uma alocação a um *cluster* errado pode causar um *outlier* no que diz respeito ao erro de predição. No cenário 3, o valor do RMSE médio considerando 1 *cluster* foi praticamente o mesmo para os dois métodos, o que supõe que o método iCRCNLR selecionou a função linear para ajustar os dados. Para dois e 3 *clusters*, o método iCRCNLR apresentou melhores resultados médios, sendo a diferença no método Stacked Regressions com 2 *clusters* estatisticamente significativa. No cenário 4, o RMSE médio obtido pelos métodos foi semelhante considerando 1 *cluster*. O método iCRCNLR apresentou melhor desempenho em 2 *clusters* e pior desempenho com 3 *clusters*. No cenário 5, o método linear apresentou melhor desempenho em todos os métodos de alocação, porém, como informado na Tabela 15, as diferenças não foram consideradas estatisticamente significativas. No cenário 6, o desempenho dos métodos foi bastante semelhante, com o único caso estatisticamente significativo ocorrendo no método *Stacked Regressions* em 3 *clusters*, onde o método iCRCNLR apresentou melhor desempenho.

Tabela 11 – RMSE médio da predição do limite inferior dos intervalos utilizando três métodos de alocação nos cenários 1 a 6.

Cenário	Método	Medida	1 Cluster	2 Clusters			3 Clusters		
				KNN	SR	Random	KNN	SR	Random
1	iCRCLR	Média	0.144	0.087	0.895	0.234	0.081	0.935	0.232
		Desvio	0.036	0.028	0.083	0.128	0.022	0.098	0.103
		Min.	0.07	0.034	0.633	0.098	0.031	0.701	0.069
	iCRCNLR	Média	0.087	**0.069**	0.084	0.095	0.071	0.103	0.106
		Desvio	0.067	0.022	0.027	0.024	0.025	0.039	0.004
		Min.	0.039	0.031	0.046	0.039	0.020	0.047	0.054
2	iCRCLR	Média	0.882	0.220	1.109	1.350	**0.057**	0.961	1.269
		Desvio	0.097	0.072	0.201	0.384	0.013	0.089	0.255
		Min.	0.596	0.094	0.799	0.557	0.031	0.792	0.592
	iCRCNLR	Média	0.885	0.165	1.084	1.121	0.059	1.413	1396.159
		Desvio	0.079	0.047	0.468	0.377	0.017	0.991	4774.254
		Min.	0.714	0.062	0.639	0.230	0.023	0.703	0.712
3	iCRCLR	Média	0.399	0.197	0.898	1.311	0.220	0.886	1.380
		Desvio	0.122	0.301	0.061	0.438	0.317	0.076	0.346
		Min.	0.125	0.024	0.751	0.055	0.025	0.702	0.621
	iCRCNLR	Média	0.398	0.077	0.865	1.187	**0.071**	0.861	1.223
		Desvio	0.123	0.044	0.099	0.431	0.035	0.123	0.365
		Min.	0.184	0.035	0.091	0.115	0.035	0.613	0.079
4	iCRCLR	Média	1.188	0.507	1.551	1.360	**0.285**	1.025	1.286
		Desvio	0.134	0.203	0.355	0.262	0.304	0.108	0.303
		Min.	0.855	0.189	0.822	0.789	0.013	0.846	0.148
	iCRCNLR	Média	1.181	0.489	1.051	1.156	0.329	1.153	1.395
		Desvio	0.144	0.221	0.219	0.357	0.336	0.324	0.313
		Min.	0.834	0.175	0.644	0.351	0.030	0.749	0.880
5	iCRCLR	Média	**0.352**	0.797	0.738	0.929	0.811	0.847	2.649
		Desvio	0.071	0.303	0.078	0.289	0.314	0.289	16.074
		Min.	0.162	0.071	0.462	0.049	0.034	0.490	0.388
	iCRCNLR	Média	0.374	0.901	0.744	0.959	0.941	0.759	1.083
		Desvio	0.088	0.217	0.072	0.270	0.253	0.083	0.235
		Min.	0.228	0.061	0.555	0.093	0.074	0.586	0.065
6	iCRCLR	Média	1.296	0.621	0.616	0.688	0.751	0.662	0.819
		Desvio	0.122	0.120	0.104	0.144	0.163	0.124	0.207
		Min.	0.906	0.405	0.376	0.357	0.392	0.386	0.272
	iCRCNLR	Média	1.290	0.640	**0.606**	0.717	0.779	0.613	0.841
		Desvio	0.138	0.145	0.078	0.155	0.187	0.085	0.183
		Min.	0.947	0.316	0.381	0.362	0.341	0.394	0.208

A Tabela 12, apresenta os resultados da simulação para predição do limite inferior dos intervalos nos cenários 7 a 12. No cenário 7, o método iCRCNLR apresentou menor RMSE médio em relação ao caso linear em todas as combinações de métodos de alocação e número de *clusters*, exceto na combinação (*Stacked Regressions* - 3 *clusters*). A Tabela 15 mostra que, neste cenário, há pouca evidência a favor de H_0 no teste de Mann-Whitney, exceto no caso da alocação aleatória com 3 *clusters* ($p = 0.570$). No cenário 8, tem-se que o método iCRCNLR apresenta melhores resultados médios em 1 *cluster*, 2 *clusters* e 3 *clusters* com método de alocação KNN, com todos os valores estatisticamente significativos, exceto em 2 *clusters* com método de alocação aleatória. No cenário 9, detectou-se a presença de um *outlier* que afetou o valor do RMSE médio em 1 *cluster* no método iCRCNLR. Para 2 *clusters* o método iCRCNLR apresentou desempenho significativamente melhor, exceto com alocação aleatória, enquanto para 3 *clusters* o mesmo método foi melhor utilizando os métodos *Stacked Regressions* e Aleatório. No cenário 10, o método iCRCNLR apresentou melhor desempenho com 1 *cluster* e em 3 *clusters* com métodos de alocação KNN e aleatório. No entanto, a Tabela 15 mostra que as diferenças foram estatisticamente significativas para *Stacked Regressions* em 2 e 3 *clusters* e para o KNN em 3 *clusters*. No cenário 11, considerando 1 *cluster*, não houve diferença significativa entre os RMSE's obtidos a partir dos métodos de ajuste. O método iCRCNLR apresentou desempenho melhor nas configurações com 2 e 3 *clusters*. A diferença não foi estatisticamente significativa apenas para o caso com 3 clusters e método de alocação KNN. No cenário 12, as únicas diferenças significativas se deram para o método de *Stacked Regressions* em 2 e 3 *clusters*, em que o método iCRCNLR apresentou melhor desempenho.

Tabela 12 – RMSE médio da predição do limite inferior dos intervalos utilizando três métodos de alocação nos cenários 7 a 12.

Cenário	Método	Medida	1 Cluster	2 Clusters			3 Clusters		
				KNN	SR	Random	KNN	SR	Random
7	iCRCLR	Média	1.764	0.823	5.365	2.988	0.537	5.372	4.256
		Desvio	1.141	0.334	1.160	1.628	0.284	1.268	2.842
		Min.	0.753	0.272	2.699	0.643	0.097	2.650	0.632
	iCRCNLR	Média	1.019	0.464	1.826	1.991	**0.071**	5.450	4.087
		Desvio	0.457	0.268	0.867	1.202	0.025	1.479	2.831
		Min.	0.477	0.058	0.770	0.399	0.030	2.325	0.586
8	iCRCLR	Média	1.734	1.041	2.864	2.304	0.693	3.067	2.679
		Desvio	0.471	0.254	0.582	0.604	0.216	0.604	0.639
		Min.	0.989	0.471	1.617	1.028	0.226	1.542	1.493
	iCRCNLR	Média	1.498	0.6790	2.241	2.218	**0.351**	3.819	7.431
		Desvio	0.300	0.131	0.854	0.529	0.301	1.626	6.689
		Min.	0.975	0.357	1.066	0.733	0.042	1.510	1.496
9	iCRCLR	Média	1.616	0.382	1.785	1.407	**0.332**	1.882	1.658
		Desvio	0.494	0.270	0.398	0.664	0.273	0.444	0.999
		Min.	0.869	0.083	1.111	0.233	0.042	0.984	0.148
	iCRCNLR	Média	20.955	0.332	0.901	1.186	0.361	1.093	1.308
		Desvio	137.234	0.346	0.306	0.549	0.392	0.335	0.636
		Min.	0.737	0.034	0.489	0.113	0.025	0.477	0.241
10	iCRCLR	Média	0.878	0.662	0.903	1.217	0.592	0.927	1.652
		Desvio	0.383	0.229	0.149	0.299	0.289	0.197	0.564
		Min.	0.314	0.266	0.607	0.429	0.112	0.532	0.609
	iCRCNLR	Média	0.787	0.703	0.974	1.347	**0.503**	1.312	1.569
		Desvio	0.178	0.323	0.192	0.407	0.412	0.446	0.513
		Min.	0.472	0.174	0.606	0.614	0.037	0.576	0.528
11	iCRCLR	Média	1.637	0.846	1.331	1.243	0.793	1.314	1.720
		Desvio	0.683	0.414	0.392	0.681	0.390	0.432	1.506
		Min.	0.985	0.161	0.549	0.138	0.129	0.603	0.357
	iCRCNLR	Média	1.628	**0.721**	0.796	1.000	0.744	0.894	1.061
		Desvio	0.655	0.483	0.271	0.488	0.492	0.373	0.407
		Min.	0.869	0.035	0.310	0.020	0.036	0.309	0.113
12	iCRCLR	Média	1.658	1.267	1.655	1.505	1.435	1.500	1.689
		Desvio	0.329	0.389	0.414	0.438	0.486	0.369	0.609
		Min.	0.916	0.524	0.770	0.656	0.605	0.672	0.394
	iCRCNLR	Média	1.724	1.294	1.294	1.424	1.484	**1.254**	1.516
		Desvio	0.350	0.446	0.385	0.572	0.417	0.375	0.566
		Min.	1.070	0.542	0.663	0.670	0.282	0.513	0.328

A Tabela 13, apresenta os resultados da simulação para predição do limite inferior dos intervalos nos cenários 13 a 18. No cenário 13, o método iCRCNLR apresenta melhor desempenho em todas as combinações de (*cluster* - método de alocação). Os resultados foram todos estatisticamente significativos neste cenário. No cenário 14, o método iCRCNLR só não apresentou melhor resultado médio na configuração de 3 *clusters* com alocação aleatória, ainda assim, a diferença não foi estatisticamente significativa ($p = 0.211$). No cenário 15, o desempenho do método iCRCNLR são não foi melhor que o do caso linear ao considerar 1 *cluster*, mesmo assim, o teste de Mann-Whitney apresentou evidência a favor H_0. O cenário 16 apresenta situação semelhante, onde a única diferença estatisticamente significativa aparece na configuração com 1 *cluster*, nos demais casos, os valores médios do RMSE para o método iCRCNLR são menores. No cenário 17, todos os valores do RMSE médio são menores para o algoritmo iCRCNLR em comparação ao iCRCLR. No entanto, são estatisticamente signficativos em quatro configurações: *Stacked Regressions* e alocação aleatória em 2 e 3 *clusters*. O cenário 18 apresenta resultados médios melhores para o método iCRCLR, no entanto, apenas as configurações com KNN e Aleatório em 3 *clusters* são estatisticamente significativas para o teste de Mann-Whitney.

Tabela 13 – RMSE médio da predição do limite inferior dos intervalos utilizando três métodos de alocação nos cenários 13 a 18.

Cenário	Método	Medida	1 Cluster	2 Clusters			3 Clusters		
				KNN	SR	Random	KNN	SR	Random
13	iCRCLR	Média	2.122	1.035	3.255	3.552	0.851	3.528	4.023
		Desvio	0.958	0.495	1.177	1.565	0.475	1.458	1.693
		Min	0.686	0.255	0.999	0.670	0.179	0.624	0.444
	iCRCNLR	Média	1.547	0.079	1.221	2.008	**0.067**	1.859	1.955
		Desvio	0.830	0.029	0.968	1.177	0.024	1.269	1.229
		Min	0.726	0.033	0.299	0.100	0.021	0.197	0.063
14	iCRCLR	Média	7.045	4.459	18.653	11.958	2.799	21.612	12.062
		Desvio	1.935	1.114	3.984	3.009	0.983	5.906	3.857
		Min	3.051	1.558	10.029	5.669	0.881	7.637	5.629
	iCRCNLR	Média	3.088	0.780	6.125	6.751	**0.123**	6.543	13.025
		Desvio	1.250	0.294	5.481	5.602	0.068	4.531	7.366
		Min	1.076	0.300	1.581	0.972	0.050	1.248	1.650
15	iCRCLR	Média	2.069	0.525	1.170	1.431	0.426	1.577	1.621
		Desvio	0.544	0.284	0.387	0.528	0.233	0.596	0.769
		Min	1.150	0.094	0.334	0.292	0.109	0.437	0.285
	iCRCNLR	Média	2.174	0.221	0.669	0.857	**0.193**	1.143	1.136
		Desvio	1.814	0.198	0.175	0.417	0.172	0.512	0.442
		Min	1.125	0.044	0.366	0.051	0.039	0.377	0.296
16	iCRCLR	Média	1.325	0.837	1.421	1.674	0.729	2.162	2.661
		Desvio	0.467	0.302	0.481	0.693	0.290	1.276	1.940
		Min	0.580	0.350	0.562	0.492	0.209	0.590	0.770
	iCRCNLR	Média	1.540	0.680	0.924	1.197	**0.349**	0.860	1.257
		Desvio	1.511	0.275	0.366	0.511	0.244	0.297	0.462
		Min	0.874	0.287	0.352	0.295	0.045	0.510	0.509
17	iCRCLR	Média	2.145	0.962	0.979	1.111	0.904	1.400	1.907
		Desvio	2.410	0.399	0.383	0.534	0.431	0.801	1.427
		Min	0.737	0.273	0.262	0.131	0.175	0.372	0.326
	iCRCNLR	Média	1.782	0.802	**0.773**	0.958	0.810	1.037	0.998
		Desvio	0.646	0.376	0.286	0.504	0.374	0.436	0.510
		Min	0.853	0.064	0.380	0.074	0.092	0.293	0.102
18	iCRCLR	Média	1.323	1.481	1.516	1.874	1.547	1.438	1.820
		Desvio	0.379	0.520	0.390	0.604	0.560	0.350	0.628
		Min	0.510	0.417	0.702	0.686	0.332	0.779	0.563
	iCRCNLR	Média	**1.319**	1.624	1.529	1.968	1.774	1.514	2.094
		Desvio	0.353	0.459	0.350	0.552	0.515	0.269	0.646
		Min	0.564	0.591	0.798	0.904	0.815	0.635	0.717

A Tabela 14, apresenta os resultados da simulação para predição do limite inferior dos intervalos nos cenários 19 a 24. No cenário 19, o método iCRCNLR apresenta resultados melhores do que o caso linear, iCRCLR. As configurações que não apresentaram diferença estatisticamente significativa foram aquelas em que o método de alocação é o aleatório. O cenário 20 apresenta melhores resultados médios para o método iCRCNLR com 1 e 2 *clusters*, são estatisticamente significativos os resultados com 1 *cluster* e 2 *clusters* com KNN. Para 3 *clusters*, o método iCRCLR apresenta melhores resultados, inclusive sendo estatisticamente significativos. No cenário 21, os resultados dos RMSE's médios do iCRCNLR são menores do que os obtidos via iCRCLR. O teste de Mann-Whitney fornece evidência contra H_0 para este cenário em todos os pares de *cluster* com método de alocação. No cenário 22, o método iCRCNLR apresenta menor RMSE médio em 1 *cluster* e com 2 *clusters* com KNN e *Stacked Regressions*. No entanto, para este cenário, a única configuração que apresentou evidência contra H_0 no teste de Mann-Whitney foi a alocação aleatória com 3 *clusters*. O cenário 23 apresenta situação em que o método iCRCNLR apresenta RMSE médio menor do que o método iCRCLR em todas as situações. Entretanto, há evidência a favor de H_0 para o teste de Mann-Whitney em 1 *cluster* e 3 *clusters* com KNN. No cenário 24 não se nota diferença entre os valores médios do RMSE entre os métodos. Tal fato é confirmado na Tabela 15, em que há evidência favorável a H_0 em todos os *clusters* em métodos de alocação.

Tabela 14 – RMSE médio da predição do limite inferior dos intervalos utilizando três métodos de alocação nos cenários 19 a 24.

Cenário	Método	Medida	1 Cluster	2 Clusters			3 Clusters		
				KNN	SR	Random	KNN	SR	Random
19	iCRCLR	Média	1.828	0.980	5.383	2.993	0.742	6.545	4.174
		Desvio	0.785	0.359	1.306	1.496	0.331	2.762	2.646
		Min.	0.623	0.155	1.677	0.964	0.163	1.221	0.686
	iCRCNLR	Média	0.826	0.290	1.601	3.164	**0.062**	2.339	2.759
		Desvio	0.308	0.231	0.770	2.448	0.024	1.311	1.464
		Min.	0.349	0.025	0.643	0.314	0.020	0.500	0.260
20	iCRCLR	Média	1.675	1.049	2.171	2.667	0.654	2.232	3.029
		Desvio	0.450	0.270	0.593	0.988	0.159	0.641	0.875
		Min.	0.919	0.526	1.1992	1.172	0.228	1.104	1.264
	iCRCNLR	Média	1.451	0.727	1.941	2.288	**0.415**	3.331	5.692
		Desvio	0.331	0.167	0.611	0.825	0.283	1.112	4.851
		Min.	0.696	0.451	0.860	0.819	0.036	1.312	1.092
21	iCRCLR	Média	2.003	0.556	1.232	1.476	0.438	1.960	2.278
		Desvio	0.721	0.328	0.609	0.597	0.283	0.879	1.951
		Min.	1.108	0.089	0.393	0.224	0.094	0.292	0.182
	iCRCNLR	Média	1.729	**0.357**	0.671	1.037	0.395	0.816	0.936
		Desvio	0.381	0.335	0.211	0.600	0.367	0.236	0.456
		Min.	0.903	0.040	0.229	0.045	0.036	0.224	0.185
22	iCRCLR	Média	3.275	2.054	4.661	6.417	**1.594**	4.536	6.510
		Desvio	1.026	0.835	1.424	1.991	0.930	1.142	2.247
		Min.	1.379	0.687	1.750	1.728	0.401	2.280	1.796
	iCRCNLR	Média	3.152	2.192	4.405	7.097	1.976	4.757	8.818
		Desvio	0.821	1.556	1.375	4.030	1.772	1.274	5.110
		Min.	1.971	0.531	1.817	1.540	0.057	2.604	2.479
23	iCRCLR	Média	1.574	0.854	1.386	0.966	0.843	2.386	3.317
		Desvio	0.570	0.423	0.448	0.445	0.462	1.843	3.638
		Min.	1.095	0.083	0.4454	0.138	0.066	0.485	0.214
	iCRCNLR	Média	1.524	**0.673**	0.692	0.841	0.714	0.790	0.889
		Desvio	0.636	0.443	0.215	0.452	0.449	0.298	0.431
		Min.	0.798	0.052	0.286	0.037	0.064	0.266	0.102
24	iCRCLR	Média	1.717	1.278	1.305	1.609	1.464	1.356	1.712
		Desvio	0.271	0.318	0.354	0.539	0.527	0.317	0.567
		Min.	1.064	0.402	0.612	0.714	0.298	0.645	0.532
	iCRCNLR	Média	1.627	1.380	1.272	1.554	1.627	**1.254**	1.703
		Desvio	0.300	0.441	0.307	0.461	0.588	0.269	0.478
		Min.	0.853	0.582	0.625	0.767	0.381	0.547	0.615

Tabela 15 – Valor-p do teste de Mann-Whitney para o RMSE do limite inferior dos intervalos, algoritmos iCRCLR e iCRCNLR.

Cenário	**1 Cluster**	2 Clusters			3 Clusters		
		KNN	SR	Random	KNN	SR	Random
1	**0.000**	**0.005**	**0.000**	**0.000**	**0.013**	**0.000**	**0.000**
2	0.887	**0.000**	0.390	0.086	0.078	**0.000**	**0.000**
3	0.806	0.100	**0.035**	0.151	0.448	0.247	**0.004**
4	0.812	0.369	**0.000**	0.072	0.945	**0.004**	**0.001**
5	0.166	0.253	0.551	0.665	0.185	0.299	0.499
6	0.544	0.260	0.635	0.114	0.419	**0.009**	0.597
7	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	0.570
8	**0.001**	**0.000**	**0.000**	0.983	**0.000**	**0.003**	**0.000**
9	**0.034**	**0.006**	**0.000**	0.257	**0.041**	**0.000**	0.082
10	0.419	0.872	**0.044**	0.472	**0.027**	**0.000**	0.736
11	0.900	**0.029**	**0.000**	**0.035**	0.253	**0.000**	**0.000**
12	0.240	0.957	**0.000**	0.200	0.806	**0.001**	0.290
13	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
14	**0.000**	**0.000**	**0.000**	**0.001**	**0.000**	**0.000**	0.211
15	0.189	**0.000**	**0.000**	**0.000**	**0.000**	**0.001**	**0.000**
16	0.469	**0.006**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
17	0.062	0.061	**0.004**	**0.020**	0.379	**0.036**	**0.000**
18	0.970	0.075	0.674	0.267	**0.023**	0.076	**0.013**
19	**0.000**	**0.000**	**0.000**	0.177	**0.000**	**0.000**	0.279
20	**0.006**	**0.000**	0.077	0.334	**0.000**	**0.000**	**0.006**
21	**0.005**	**0.000**	**0.000**	**0.000**	**0.032**	**0.000**	**0.000**
22	0.517	0.146	0.340	0.986	0.634	0.190	**0.008**
23	0.090	**0.025**	**0.000**	**0.008**	0.165	**0.000**	**0.000**
24	0.071	0.617	0.551	0.562	0.217	0.051	0.869

A Tabela 16 apresenta os melhores e piores pares de métodos de ajuste e de alocação para predição do limite inferior dos intervalos, de acordo com o RMSE médio. Nota-se que o método iCRCNLR apresenta o melhor desempenho global em 19 dos 24 cenários. O KNN é apresentado como o melhor dentre os métodos de alocação em 19 cenários, seguido pelo *Stacked Regressions*, melhor em 3 cenários. O ajuste com apenas um *cluster* foi melhor em 2 cenários, 5 e 18.

Tabela 16 – Melhores e piores pares Modelo-Método de alocação para predição do limite inferior dos intervalos nos 24 cenários.

Cenário	Melhor Par			Pior Par		
	Clusters	Modelo	Alocação	*Clusters*	Modelo	Alocação
1	2	**iCRCNLR**	KNN	3	iCRCLR	SR
2	3	iCRCLR	KNN	3	**iCRCNLR**	Random
3	3	**iCRCNLR**	KNN	3	iCRCLR	Random
4	3	iCRCLR	KNN	3	**iCRCNLR**	Random
5	1	iCRCLR	-	3	iCRCLR	Random
6	2	**iCRCNLR**	SR	1	iCRCLR	-
7	3	**iCRCNLR**	KNN	3	**iCRCNLR**	SR
8	3	**iCRCNLR**	KNN	3	**iCRCNLR**	Random
9	3	iCRCLR	KNN	1	**iCRCNLR**	-
10	3	**iCRCNLR**	KNN	3	iCRCLR	Random
11	2	**iCRCNLR**	KNN	3	iCRCLR	Random
12	3	**iCRCNLR**	SR	1	**iCRCNLR**	-
13	3	**iCRCNLR**	KNN	3	iCRCLR	Random
14	3	**iCRCNLR**	KNN	3	iCRCLR	SR
15	3	**iCRCNLR**	KNN	1	**iCRCNLR**	-
16	3	**iCRCNLR**	KNN	3	iCRCLR	Random
17	2	**iCRCNLR**	SR	1	iCRCLR	-
18	1	**iCRCNLR**	-	3	**iCRCNLR**	Random
19	3	**iCRCNLR**	KNN	3	iCRCLR	SR
20	3	**iCRCNLR**	KNN	3	**iCRCNLR**	Random
21	2	**iCRCNLR**	KNN	3	iCRCLR	Random
22	3	iCRCLR	KNN	3	**iCRCNLR**	Random
23	2	**iCRCNLR**	KNN	2	iCRCLR	Random
24	3	**iCRCNLR**	SR	1	iCRCLR	-

5.1.2.2 Limite superior

Da mesma forma que no limite inferior, são apresentados os resultados para predição do limite superior dos intervalos. A Tabela 17 apresenta os resultados da simulação referentes à predição do limite superior dos intervalos. Primeiramente, fixando os métodos de ajuste e o número de *clusters*, tem-se que, no cenário 1 o método KNN apresenta menor RMSE médio que os outros métodos, tanto para 2 quanto para 3 *clusters* ajustados. A mesma situação ocorre no cenário 2, 3 e 4. Já no cenário 5, em que os clusters são sobrepostos, o método de *Stacked Regressions* leva vantagem quando o método de ajuste escolhido é o iCRCNLR com 2 clusters. No cenário 6, construído com 3 *clusters* sobrepostos, o método de alocação *Stacked Regressions* apresenta melhor desempenho em todas as configurações. A Tabela 18 apresenta os mesmos resultados para os cenários de 7 a 12. O método de alocação KNN é claramente superior nos cenário 7, 8, 9 e 10, que apresentam *clusters* disjuntos ou com sobreposição parcial. Nos cenários 11 e 12, com *clusters* completamente sobrepostos no eixo das abcissas, os valores do RMSE médio para os métodos de alocação KNN e *Stacked Regressions* tornam-se muito próximos. A Tabela 13 apresenta os resultados para os *clusters* 13 a 18. Nos cenários 13, 14, 15 e 16, o método de alocação KNN se sai melhor em todas as combinações, seguido pelo método de *Stacked Regressions*. Já no cenário 17, o método de alocação *Stacked Regressions* apresenta melhor desempenho, exceto no par (iCRCLR-3 *clusters*), novamente numa situação em que há sobreposição total dos *clusters* gerados. Finalmente, a Tabela 20 apresenta os resultados da predição no limite superior para os cenários 19 a 24. Nestes cenários, os dados tanto de centro e amplitude são gerados por meio de um função não linear. O método de alocação KNN é melhor nos cenários 19 a 22 e no cenário 23 quando o método de ajuste é linear. Neste mesmo cenário, com método de ajuste iCRCNLR, o melhor método passa a ser o *Stacked Regressions*, assim como no cenário 24. Novamente, estes cenários foram gerados de modo a fornecer *clusters* sobrepostos.

Fixando os métodos de alocação, procede-se a comparação dos métodos de ajuste iCRCLR e iCRCNLR. Avaliando os números apresentados na Tabela 17, o algoritmo iCRCNLR apresenta melhor desempenho em todas as combinações de números de *clusters* ajustados e métodos de alocação, no cenário 1. No cenário 2, o método iCRCNLR apresenta melhor desempenho para 2 e 3 *clusters* quando é utilizado o método de alocação KNN e é pior que o iCRCLR quando o método de alocação é o *Stacked Regressions*. Para 1 *cluster*, a diferença entre os métodos de ajuste é muito pequena. No cenário 3, o método iCRCNLR apresenta RMSE médio menor que o iCRCLR quando o método de alocação é o *Stacked Regressions* e maior no KNN. Para 1 *cluster* o método iCRCNLR apresenta resultado melhor, mas não estatisticamente significativo - ver Tabela 21. No cenário 4, a diferença entre os métodos para 1 *cluster* não é significativa. O método iCRCNLR apresenta desempenho melhor nos pares (2 clusters - *Stacked Regressions*), (3 *clusters* - KNN) com diferença significativa, considerando $\alpha = 0.05$ de acordo com o

p-valor para o teste de Mann-Whitney. No cenário 5, foram significativas as diferenças para KNN e Random, com 2 *clusters* e *Stacked Regressions*, com 3 *clusters*, onde o método iCRCNLR apresentou desempenho melhor. No cenário 6, a única diferença estatisticamente significativa foi ocorreu no par (3 *clusters* - *Stacked Regressions*), em que o método iCRCNLR apresentou melhor desempenho.

Tabela 17 – RMSE médio da predição do limite superior dos intervalos utilizando três métodos de alocação nos cenários 1 a 6

Cenário	Método	Medida	1 Cluster	2 Clusters			3 Clusters		
				KNN	SR	Random	KNN	SR	Random
1	iCRCLR	Média	0.134	0.090	0.763	0.199	0.080	0.854	0.221
		Desvio	0.036	0.029	0.110	0.064	0.025	0.101	0.079
		Min.	0.061	0.021	0.552	0.071	0.031	0.655	0.092
	iCRCNLR	Média	0.090	**0.064**	0.091	0.111	0.064	0.102	0.122
		Desvio	0.071	0.016	0.034	0.038	0.017	0.038	0.060
		Min.	0.048	0.023	0.044	0.044	0.019	0.046	0.043
2	iCRCLR	Média	0.880	0.224	1.562	2.201	0.058	1.848	2.541
		Desvio	0.101	0.066	0.477	0.688	0.012	0.157	0.560
		Min.	0.686	0.085	0.918	0.283	0.028	1.517	1.097
	iCRCNLR	Média	0.872	0.152	1.786	1.983	**0.056**	2.518	1397.591
		Desvio	0.098	0.043	0.536	0.487	0.011	0.882	4775.397
		Min.	0.698	0.072	0.940	0.783	0.034	1.610	1.144
3	iCRCLR	Média	0.539	**0.061**	0.481	0.447	0.062	0.443	0.521
		Desvio	0.124	0.016	0.085	0.164	0.017	0.110	0.147
		Min.	0.209	0.031	0.288	0.059	0.024	0.249	0.137
	iCRCNLR	Média	0.501	0.114	0.427	0.558	0.068	0.397	0.579
		Desvio	0.153	0.102	0.095	0.244	0.024	0.103	0.243
		Min.	0.185	0.031	0.255	0.064	0.030	0.161	0.088
4	iCRCLR	Média	0.968	0.417	1.802	2.026	0.0883	1.879	2.346
		Desvio	0.136	0.100	0.329	0.454	0.067	0.088	0.545
		Min.	0.614	0.174	1.055	0.808	0.031	1.090	1.020
	iCRCNLR	Média	0.949	0.557	1.584	2.161	**0.072**	2.053	2.625
		Desvio	0.136	0.493	0.290	0.790	0.032	0.261	0.704
		Min.	0.607	0.118	0.971	0.500	0.027	1.718	1.114
5	iCRCLR	Média	0.722	**0.363**	0.380	0.413	0.369	0.511	0.511
		Desvio	0.079	0.163	0.072	0.171	0.164	0.184	0.195
		Min.	0.515	0.042	0.227	0.041	0.033	0.255	0.110
	iCRCNLR	Média	0.743	0.438	0.372	0.522	0.409	0.418	0.539
		Desvio	0.096	0.199	0.064	0.172	0.173	0.105	0.119
		Min.	0.539	0.066	0.221	0.025	0.065	0.197	0.035
6	iCRCLR	Média	**0.595**	1.450	1.316	1.895	1.550	1.355	1.822
		Desvio	0.107	0.297	0.112	0.315	0.313	0.129	0.357
		Min.	0.300	0.161	0.924	0.194	0.244	0.937	0.608
	iCRCNLR	Média	0.620	1.534	1.289	1.909	1.620	1.303	1.832
		Desvio	0.109	0.345	0.111	0.352	0.359	0.099	0.345
		Min.	0.392	0.165	1.004	0.748	0.142	0.965	0.399

Na Tabela 18 são apresentados os resultados dos cenários 7 a 12. No cenário 7, em todos os pares de método de alocação e número de *clusters* os resultados são estatisticamente significativos. O método iCRCNLR apresentou desempenho melhor, de acordo com a média dos RSME's. O mesmo ocorre no cenário 8, em que todos as combinações são estatisticamente significativas, exceto para 2 *clusters* e método de alocação aleatória. No cenário 9, para 1 *cluster*, o método iCRCNLR apresenta resultado pior devido à presença de um *outlier* dentre os RMSE's. Para 2 e 3 *clusters*, a diferença entre os métodos é estatisticamente significativa para os métodos KNN e *Stacked Regressions* e *Stacked Regressions* e Aleatório, respectivamente. No cenário 10, somente os valores para 3 *clusters* são estatisticamente significativos, neste caso, o método iCRCNLR apresenta menor RMSE médio no método KNN. No cenário 11, as diferenças para o método de *Stacked Regressions*, tanto em 2 quanto em 3 *clusters* são significativas. Além disso, foi verificado que a diferença entre os métodos, para 3 *clusters* e método de alocação aleatória, é estatisticamente suficientes. Nestes casos, o método iCRCNLR apresenta melhor desempenho. Finalmente, no cenário 12, a diferença entre os métodos de ajuste é estatisticamente significativa para *Stacked Regressions* em 2 e 3 *clusters*. Nestes casos, o método iCRCNLR apresenta melhor desempenho, de acordo com o RMSE médio.

Tabela 18 – RMSE médio da predição do limite superior dos intervalos utilizando três métodos de alocação nos cenários 7 a 12.

Cenário	Método	Medida	1 Cluster	2 Clusters			3 Clusters		
				KNN	SR	Random	KNN	SR	Random
7	iCRCLR	Média	1.727	0.903	5.877	3.240	0.568	6.010	4.785
		Desvio	1.156	0.375	1.294	2.075	0.292	1.453	3.457
		Min.	0.737	0.261	2.744	0.778	0.100	2.760	0.776
	iCRCNLR	Média	1.014	0.443	2.152	2.803	**0.075**	6.081	5.037
		Desvio	0.452	0.248	1.002	2.368	0.035	1.607	3.044
		Min.	0.496	0.049	0.893	0.541	0.023	2.540	0.771
8	iCRCLR	Média	1.737	1.028	2.903	2.322	0.694	3.117	2.662
		Desvio	0.476	0.235	0.578	0.617	0.231	0.651	0.592
		Min.	0.776	0.600	1.463	0.982	0.230	1.694	1.611
	iCRCNLR	Média	1.475	0.705	2.129	2.110	**0.364**	3.885	7.376
		Desvio	0.306	0.127	0.749	0.510	0.292	1.551	6.730
		Min.	0.804	0.419	1.049	0.679	0.033	1.566	1.576
9	iCRCLR	Média	1.249	0.781	2.814	2.785	0.784	2.909	3.051
		Desvio	0.521	0.747	0.391	0.880	0.787	0.442	1.407
		Min.	0.593	0.087	1.869	0.156	0.041	1.924	0.273
	iCRCNLR	Média	20.670	**0.765**	1.888	2.794	0.828	1.890	2.468
		Desvio	137.265	0.829	0.287	0.921	0.860	0.383	0.829
		Min.	0.606	0.035	1.272	0.103	0.039	1.211	0.717
10	iCRCLR	Média	0.933	0.597	0.838	0.976	0.522	0.984	1.814
		Desvio	0.350	0.189	0.195	0.261	0.271	0.293	0.823
		Min.	0.553	0.225	0.356	0.360	0.116	0.507	0.388
	iCRCNLR	Média	0.845	0.604	0.858	1.050	**0.379**	1.408	2.213
		Desvio	0.225	0.234	0.209	0.253	0.347	0.437	0.579
		Min.	0.424	0.223	0.475	0.484	0.038	0.576	0.561
11	iCRCLR	Média	1.000	1.651	1.845	2.157	1.694	2.018	2.340
		Desvio	0.867	0.557	0.369	0.776	0.664	0.433	1.164
		Min.	0.343	0.129	1.080	0.309	0.114	1.172	0.484
	iCRCNLR	Média	**0.957**	1.549	1.536	2.044	1.592	1.621	2.202
		Desvio	0.796	0.753	0.244	0.794	0.762	0.351	0.756
		Min.	0.335	0.043	0.983	0.073	0.043	0.950	0.153
12	iCRCLR	Média	**1.250**	1.696	1.943	2.282	1.795	1.849	2.220
		Desvio	0.360	0.536	0.453	0.517	0.524	0.441	0.649
		Min.	0.637	0.586	1.036	1.038	0.601	0.903	0.883
	iCRCNLR	Média	1.272	1.863	1.739	2.162	2.004	1.661	2.188
		Desvio	0.391	0.567	0.344	0.658	0.652	0.353	0.639
		Min.	0.551	0.448	1.043	1.151	0.384	1.016	0.778

Na Tabela 19 estão apresentados os resultados para os cenários 13 a 18. No cenário 13, o método iCRCNLR apresenta desempenho melhor em todos os pares (Quantidade de *clusters* - Métodos de alocação), com significância estatística em todos eles, exceto para 2 *clusters* com alocação aleatória. No cenário 14, o método iCRNLR apresenta menor RMSE médio em todas as combinações, exceto em 3 *clusters* com alocação aleatória. Neste cenário, todas as diferenças são estatisticamente significativas. O cenário 15, assim como o cenário 13, apresenta diferenças estatisticamente significativas em todas as combinações, excetuando-se a alocação aleatória em 2 *clusters*. O método iCRCNLR apresenta menor RMSE médio em todas as combinações. No cenário 16, os resultados estatisticamente significativos aparecem nas combinações em que há método de alocação *Stacked Regressions* e aleatório e com 1 *cluster*. Nestes casos, o método iCRCNLR apresenta menor RMSE médio do que o método iCRCLR. No cenário 17, somente ao utilizar métodos de alocação KNN não foi registrada diferença significativa entre os métodos de ajuste. Nestes cenários, o método iCRCNLR apresentou menor RMSE médio e comparação ao método iCRCLR, que só apresentou desempenho melhor com KNN e 3 *clusters*, mas sem diferença estatisticamente significativa. Finalmente, no cenário 18, não foi registrada diferenças entre os métodos, de acordo com o teste de Mann-Whitney, o que fornece evidência de que o desempenho dos métodos é semelhante neste cenário.

Tabela 19 – RMSE médio da predição do limite superior dos intervalos utilizando três métodos de alocação nos cenários 13 a 18.

Cenário	Método	Método	1 Cluster	2 Clusters			3 Clusters		
				KNN	SR	Random	KNN	SR	Random
13	iCRCLR	Média	2.058	1.166	5.883	7.414	0.904	5.384	8.854
		Desvio	0.899	0.625	3.117	5.419	0.498	2.459	5.788
		Min.	0.660	0.182	1.493	0.208	0.151	1.047	0.932
	iCRCNLR	Média	1.175	0.081	2.693	6.071	**0.065**	2.613	5.252
		Desvio	1.145	0.040	0.865	2.561	0.022	0.946	3.199
		Min.	0.455	0.030	1.632	0.865	0.021	0.820	0.064
14	iCRCLR	Média	8.255	3.860	19.519	10.287	2.685	21.545	10.461
		Desvio	2.035	1.119	4.609	3.006	0.870	11.166	4.417
		Min.	4.360	1.664	9.346	3.683	1.005	5.773	4.165
	iCRCNLR	Média	2.523	1.232	6.321	8.495	**0.118**	9.072	26.770
		Desvio	1.261	0.245	5.037	6.929	0.080	4.662	12.521
		Min.	0.904	0.418	2.346	2.263	0.041	4.579	7.562
15	iCRCLR	Média	1.317	1.393	2.420	3.534	1.370	3.214	4.087
		Desvio	0.620	0.958	0.408	1.055	0.987	0.952	1.504
		Min.	0.496	0.189	1.363	0.313	0.120	1.452	0.743
	iCRCNLR	Média	1.251	**1.127**	2.149	3.269	1.145	2.459	3.116
		Desvio	1.961	1.088	0.288	0.910	1.112	0.699	0.945
		Min.	0.311	0.032	1.632	0.063	0.029	1.329	1.053
16	iCRCLR	Média	1.151	1.004	1.692	1.786	0.932	2.949	3.886
		Desvio	0.662	0.320	0.594	0.714	0.327	2.143	3.762
		Min.	0.388	0.279	0.556	0.852	0.270	0.855	0.906
	iCRCNLR	Média	1.089	1.000	1.298	1.714	**0.831**	1.518	3.043
		Desvio	1.665	0.310	0.292	0.485	0.483	0.295	0.999
		Min.	0.288	0.310	0.859	0.944	0.035	0.936	1.189
17	iCRCLR	Média	1.363	2.021	1.813	2.349	2.026	2.104	2.715
		Desvio	2.602	0.585	0.420	0.922	0.624	0.701	1.116
		Min.	0.268	0.326	0.850	0.334	0.224	0.680	0.289
	iCRCNLR	Média	**0.961**	2.015	1.707	2.159	2.029	1.852	2.364
		Desvio	0.859	0.783	0.255	0.740	0.777	0.357	0.584
		Min.	0.330	0.176	1.181	0.385	0.329	1.087	0.269
18	iCRCLR	Média	1.403	1.407	1.466	1.768	1.491	**1.344**	1.725
		Desvio	0.365	0.502	0.413	0.612	0.566	0.347	0.634
		Min.	0.759	0.332	0.561	0.593	0.281	0.580	0.575
	iCRCNLR	Média	1.442	1.472	1.456	1.835	1.588	1.384	1.837
		Desvio	0.348	0.499	0.436	0.527	0.538	0.319	0.662
		Min.	0.721	0.519	0.646	1.014	0.225	0.531	0.554

Na Tabela 20 são apresentados os resultados para a predição do limite superior dos intervalos nos cenários 19 a 24. Estes cenários são gerados a partir de funções não lineares para o centro e amplitude dos dados. O cenário 19 apresenta valores médios do RMSE médio menores para o método iCRCNLR, exceto onde o método de alocação é aleatória, no entanto, nestes casos, não há diferença estatisticamente significativa entre os métodos. No cenário 20, o método iCRCNLR apresenta valor médio do RMSE menor do que o método iCRCLR em 1 *cluster*, 2 *clusters* com KNN e em 3 *clusters* com KNN e *Stacked Regressions*, com a diferença entre os métodos sendo estatisticamente significativa. Nos outros casos, o método iCRCLR é melhor (3 *clusters* e alocação aleatória), ou não há evidência suficiente para considerar os métodos diferentes. No cenário 21, apenas o método KNN, onde o iCRCNLR é pior, não apresentou diferença estatisticamente significativa entre os métodos. Nos outros casos, o método iCRCNLR apresentou resultados melhores. No cenário 22, apenas as combinações *Stacked Regressions* com 2 e 3 *clusters*, onde o método iCRCNLR teve melhor desempenho, e alocação aleatória com 3 *clusters* apresentaram diferenças estatisticamente significativas entre os métodos. Nas demais combinações, os resultados foram semelhantes entre os métodos. No cenário 23, além das combinações mencionadas para o cenário anterios, também foi estatisticamente significativa a diferença entre os métodos para 1 *cluster*. Nestas combinações, o método iCRCNLR apresentou menor média. Finalmente, no cenário 24, as combinações com método de alocação KNN foram estatísticamente significativas. Nestes casos, o método iCRCNLR apresentou melhores resultados.

Tabela 20 – RMSE médio da predição do limite superior dos intervalos utilizando três métodos de alocação nos cenários 19 a 24.

Cenário	Método	Método	1 Cluster	2 Clusters			3 Clusters		
				KNN	SR	Random	KNN	SR	Random
19	iCRCLR	Média	1.804	1.066	5.432	2.802	0.759	5.861	4.450
		Desvio	0.780	0.391	1.876	1.263	0.374	2.336	3.006
		Min.	0.803	0.160	2.049	1.143	0.140	1.884	0.518
	iCRCNLR	Média	0.816	0.288	2.319	3.587	**0.066**	3.428	4.295
		Desvio	0.307	0.234	1.173	2.563	0.024	1.891	2.689
		Min.	0.328	0.023	1.027	0.446	0.032	1.049	0.728
20	iCRCLR	Média	1.674	1.045	2.422	2.712	0.662	2.740	3.051
		Desvio	0.454	0.285	0.994	1.057	0.184	1.075	0.979
		Min.	0.841	0.534	0.977	1.129	0.243	0.977	1.209
	iCRCNLR	Média	1.435	0.760	1.995	2.277	**0.422**	3.502	5.689
		Desvio	0.378	0.154	0.557	0.669	0.275	1.285	4.708
		Min.	0.459	0.398	0.992	1.082	0.045	1.277	1.307
21	iCRCLR	Média	1.270	1.478	2.304	3.371	1.400	3.071	3.800
		Desvio	0.869	0.939	0.462	0.938	0.991	1.025	2.001
		Min.	0.266	0.087	1.010	0.452	0.066	1.408	0.162
	iCRCNLR	Média	**0.859**	1.635	2.077	3.138	1.649	2.140	2.822
		Desvio	0.304	1.135	0.233	0.741	1.195	0.267	0.746
		Min.	0.463	0.038	1.610	0.063	0.030	1.561	1.196
22	iCRCLR	Média	3.429	1.932	6.671	5.904	**1.560**	6.986	6.083
		Desvio	1.001	0.691	2.169	1.889	0.790	2.395	2.093
		Min.	1.614	0.733	1.901	1.614	0.473	2.051	1.888
	iCRCNLR	Média	3.400	1.917	4.565	5.502	1.684	4.634	8.934
		Desvio	0.844	1.064	1.514	2.554	1.252	1.175	4.664
		Min.	2.135	0.722	2.245	1.518	0.076	2.758	3.164
23	iCRCLR	Média	0.921	1.593	1.833	1.946	1.623	2.894	3.661
		Desvio	0.693	0.679	0.310	0.876	0.713	2.015	2.936
		Min.	0.300	0.193	1.007	0.132	0.103	1.150	0.578
	iCRCNLR	Média	**0.900**	1.573	1.383	1.778	1.646	1.498	1.885
		Desvio	0.860	0.641	0.216	0.737	0.625	0.361	0.624
		Min.	0.257	0.032	0.903	0.032	0.051	0.880	0.483
24	iCRCLR	Média	**1.277**	1.801	1.789	2.373	1.915	1.832	2.307
		Desvio	0.338	0.427	0.305	0.614	0.482	0.306	0.581
		Min.	0.649	0.776	1.178	0.850	0.363	1.192	0.766
	iCRCNLR	Média	1.311	1.834	1.664	2.217	2.007	1.616	2.222
		Desvio	0.316	0.527	0.295	0.545	0.579	0.247	0.584
		Min.	0.735	0.816	0.946	1.264	0.898	0.859	1.074

Tabela 21 – Valor-p do teste de Mann-Whitney para o RMSE do limite superior dos intervalos, algoritmos iCRCLR e iCRCNLR.

Cenário	1 Cluster	2 Clusters			3 Clusters		
		KNN	SR	Random	KNN	SR	Random
1	**0.000**	**0.000**	**0.000**	**0.000**	**0.002**	**0.000**	**0.000**
2	0.689	**0.000**	**0.000**	0.900	0.203	**0.000**	**0.000**
3	0.158	0.162	**0.001**	**0.005**	0.617	**0.029**	0.320
4	0.277	**0.000**	**0.018**	0.476	0.956	**0.000**	**0.000**
5	0.203	**0.020**	0.556	**0.004**	0.109	**0.010**	0.206
6	0.315	0.186	0.105	0.966	0.279	**0.001**	0.931
7	**0.000**	**0.000**	**0.000**	**0.009**	**0.000**	**0.000**	**0.029**
8	**0.000**	**0.000**	**0.000**	0.141	**0.000**	**0.001**	**0.001**
9	0.228	**0.019**	**0.000**	0.606	0.053	**0.000**	**0.035**
10	0.146	0.554	0.834	0.364	**0.000**	**0.000**	**0.002**
11	0.556	0.388	**0.000**	0.497	0.507	**0.000**	**0.031**
12	0.709	0.132	**0.033**	0.260	0.064	**0.037**	0.906
13	**0.000**	**0.000**	**0.000**	0.338	**0.000**	**0.000**	**0.000**
14	**0.000**	**0.000**	**0.000**	**0.025**	**0.000**	**0.000**	**0.000**
15	**0.000**	**0.012**	**0.000**	0.175	**0.023**	**0.000**	**0.000**
16	**0.003**	0.809	**0.001**	0.763	0.386	**0.000**	0.344
17	**0.014**	0.944	**0.045**	**0.032**	0.922	**0.064**	**0.006**
18	0.417	0.504	0.703	0.422	0.269	0.279	0.160
19	**0.000**	**0.000**	**0.000**	0.976	**0.000**	**0.000**	0.276
20	**0.004**	**0.000**	0.081	0.440	**0.000**	**0.000**	**0.002**
21	**0.000**	0.706	**0.000**	**0.023**	0.677	**0.000**	**0.000**
22	0.992	0.208	**0.000**	0.951	0.785	**0.000**	**0.000**
23	**0.034**	0.957	**0.000**	0.097	0.970	**0.000**	**0.000**
24	0.494	0.850	**0.033**	0.197	0.354	**0.000**	0.413

A Tabela 22 apresenta os melhores e piores pares de métodos de ajuste e de alocação para predição do limite superior dos intervalos, de acordo com o RMSE médio. Nota-se que o método iCRCNLR apresenta o melhor desempenho global em 17 dos 24 cenários. O KNN é apresentado como o melhor dentre os métodos de alocação.

Tabela 22 – Melhores e piores pares Modelo-Método de alocação para predição do limite superior dos intervalos nos 24 cenários.

Cenário	Melhor Par			Pior Par		
	Clusters	Modelo	Alocação	*Clusters*	Modelo	Alocação
1	2	**iCRCNLR**	KNN	3	iCRCLR	SR
2	3	**iCRCNLR**	KNN	3	**iCRCNLR**	Random
3	2	iCRCLR	KNN	3	**iCRCNLR**	Random
4	3	**iCRCNLR**	KNN	3	**iCRCNLR**	Random
5	2	iCRCLR	KNN	1	**iCRCNLR**	-
6	1	iCRCLR	-	2	**iCRCNLR**	Random
7	3	**iCRCNLR**	KNN	3	**iCRCNLR**	SR
8	3	**iCRCNLR**	KNN	3	**iCRCNLR**	Random
9	2	**iCRCNLR**	KNN	1	**iCRCNLR**	-
10	3	**iCRCNLR**	KNN	3	**iCRCNLR**	Random
11	1	**iCRCNLR**	-	3	iCRCLR	Random
12	1	iCRCLR	-	2	iCRCLR	Random
13	3	**iCRCNLR**	KNN	3	iCRCLR	Random
14	3	**iCRCNLR**	KNN	3	**iCRCNLR**	Random
15	2	**iCRCNLR**	KNN	3	iCRCLR	Random
16	3	**iCRCNLR**	KNN	3	iCRCLR	Random
17	1	**iCRCNLR**	-	3	iCRCLR	Random
18	3	iCRCLR	SR	3	**iCRCNLR**	Random
19	3	**iCRCNLR**	KNN	3	iCRCLR	SR
20	3	**iCRCNLR**	KNN	3	**iCRCNLR**	Random
21	1	**iCRCNLR**	-	3	iCRCLR	Random
22	3	iCRCLR	KNN	3	iCRCLR	Random
23	1	**iCRCNLR**	-	3	iCRCLR	Random
24	1	iCRCLR	-	2	iCRCLR	Random

5.1.2.3 Erro geral

Finalmente, são apresentados os resultados para intervalo como um todo. O $RMSE_{0f}$ é dado pela raiz quadrada da média das somas dos erros nos limites superior e inferior dos intervalos. Fixando o método de ajuste e o número de *clusters*, pode-se comparar os métodos de alocação. A Tabela 23 apresenta os resultados para predição geral dos intervalos 1 a 6. No cenário 1, o método de alocação KNN apresenta os menores RMSE's médios em relação aos outros métodos. Note-se que, para o ajuste linear, o método aleatório alcança um resultado melhor que o *Stacked Regressions*. No cenário 2, o método KNN também apresenta os melhores resultados, seguido pelo *Stacked Regressions* e aleatório. A mesma situação ocorre nos cenários 3 e 4. Nos cenários 5 e 6, onde há sobreposição total de *clusters*, o método de *Stacked Regressions* apresenta os melhores resultados médios em relação aos outros métodos. A Tabela 24 apresenta os mesmos resultados para os cenários 7 a 12. O método KNN apresenta grande diferença em relação aos outros métodos nos cenários 7, 8, 9 e 10. Nos cenários 11 e 12, o método KNN permanece com melhor performance considerando o método iCRCLR, enquanto no método iCRCNLR o melhor método de alocação é o *Stacked Regressions*. A Tabela 25 apresenta os resultados para os cenários 13 a 18. Novamente, nos cenários em que os *clusters* são disjuntos, 13 e 14, o método KNN apresenta resultados melhores que os outros métodos de alocação. O método KNN apresenta também melhor desempenho nos cenários com sobreposição parcial, 15 e 16. No cenário 17, o método *Stacked Regressions* apresenta os melhores resultados, seguido pelo KNN. No cenário 18, o método KNN é melhor, em termos do RMSE médio, apenas com 2 *clusters* com método de ajuste linear. Finalmente, a Tabela 26 apresenta os resultados para os cenários 19 a 24. Nos cenários 19 e 20, onde não há sobreposição de *clusters*, o método KNN apresenta menor RMSE médio. O mesmo ocorre nos cenários 20 e 21, onde a sobreposição é parcial. No cenário 23, o método KNN apresenta melhor resultado médio apenas no modelo de ajuste linear. No método iCRCNLR, o método de alocação que tem melhor desempenho é o *Stacked Regressions*. No cenário 24, o método *Stacked Regressions* é melhor em todas as combinações de modelo de ajuste e número de *clusters*, no entanto, a diferença observada é muito pequena.

Fixando os métodos de alocação, pode-se comparar o desempenho dos dois métodos de ajuste, iCRCLR e iCRCNLR. Novamente, na Tabela 23, estão apresentados os resultados para os cenários de 1 a 6. No cenário 1, o método iCRCNLR apresentou melhor desempenho médio em todas as combinações de métodos de alocação e número de *clusters*. Nestes casos, a hipótese nula no teste de Mann-Whitney foi rejeitada ao nível de 5%, como apresentado na Tabela 27. No cenário 2, as diferenças estatisticamente significativas, de acordo com o teste de Mann-Whitney, se deram nas combinações KNN e *Stacked Regressions* com 2 *clusters* e *Stacked Regressions* e aleatório com 3 *clusters*. Exceto por este último caso, o método iCRCNLR apresentou melhor desempenho mé-

dio. Para 1 *cluster*, os resultados não evidenciam diferenças na qualidade da predição entre os métodos. No cenário 3, só foi observada diferença estatisticamente significativa entre os métodos de ajuste em dois casos: 2 *clusters* ajustados com alocação *Stacked Regressions* e 3 *clusters* ajustados com método de alocação aleatório. Nestes casos, o método de ajuste iCRCNLR apresentou melhor desempenho. Considerando o ajuste para 1 *cluster* os resultados próximos fornecem pouca evidência de que os métodos diferem entre si. No cenário 4, ocorre a mesma situação do cenário 2, mas o método iCRCNLR apresenta resultados piores que o caso linear. O cenário 5 não apresentou diferenças estatisticamente significativas entre os métodos em nenhuma das combinações de métodos de alocação e número de clusters. No cenário 6, a única diferença significativa entre os métodos se dá com 3 *clusters* e método de alocação *Stacked Regressions*.

Tabela 23 – RMSE médio da predição geral dos limites utilizando três métodos de alocação nos cenários 1 a 6.

Cenário	Método	Medida	1 Cluster	2 Clusters			3 Clusters		
				KNN	SR	Random	KNN	SR	Random
1	iCRCLR	Média	0.199	0.1281	1.180	0.280	0.116	1.264	0.321
		Desvio	0.040	0.033	0.097	0.085	0.025	0.108	0.116
		Min.	0.104	0.059	0.931	0.124	0.058	1.014	0.121
	iCRCNLR	Média	0.127	0.099	0.126	0.155	**0.098**	0.141	0.167
		Desvio	0.096	0.023	0.034	0.043	0.021	0.038	0.061
		Min.	0.062	0.061	0.069	0.082	0.062	0.074	0.073
2	iCRCLR	Média	1.249	0.319	1.925	2.564	**0.083**	2.082	2.854
		Desvio	0.115	0.062	0.432	0.714	0.013	0.132	0.548
		Min.	1.005	0.191	1.276	0.843	0.051	1.795	1.395
	iCRCNLR	Média	1.245	0.224	2.013	2.302	0.084	2.856	1975.606
		Desvio	0.097	0.044	0.501	0.515	0.016	1.098	6752.586
		Min.	1.089	0.118	1.315	1.132	0.050	1.772	2.020
3	iCRCLR	Média	0.680	0.088	1.025	1.458	**0.087**	0.988	1.477
		Desvio	0.131	0.016	0.084	0.352	0.016	0.095	0.312
		Min.	0.335	0.056	0.835	0.081	0.050	0.764	0.636
	iCRCNLR	Média	0.646	0.116	0.965	1.361	0.098	0.951	1.352
		Desvio	0.176	0.075	0.124	0.362	0.029	0.143	0.339
		Min.	0.300	0.049	0.345	0.332	0.053	0.633	0.119
4	iCRCLR	Média	1.537	0.611	2.376	2.470	**0.234**	2.136	2.696
		Desvio	0.147	0.121	0.433	0.518	0.289	0.106	0.527
		Min.	1.215	0.304	1.337	1.290	0.046	1.429	1.371
	iCRCNLR	Média	1.521	0.787	1.929	2.404	0.431	2.355	2.987
		Desvio	0.151	0.539	0.362	0.779	0.532	0.354	0.700
		Min.	1.173	0.280	1.280	0.669	0.059	1.901	1.523
5	iCRCLR	Média	**0.807**	0.883	0.840	1.062	0.914	0.924	1.162
		Desvio	0.080	0.326	0.072	0.258	0.319	0.146	0.294
		Min.	0.601	0.096	0.614	0.064	0.059	0.670	0.406
	iCRCNLR	Média	0.834	1.006	0.835	1.113	1.029	0.878	1.220
		Desvio	0.112	0.262	0.078	0.274	0.300	0.105	0.244
		Min.	0.625	0.093	0.635	0.096	0.099	0.666	0.075
6	iCRCLR	Média	1.431	1.586	1.457	2.021	1.733	1.512	2.011
		Desvio	0.116	0.278	0.110	0.316	0.302	0.139	0.359
		Min.	1.088	0.436	1.112	0.564	0.589	1.106	0.762
	iCRCNLR	Média	1.435	1.673	**1.429**	2.031	1.798	1.441	2.014
		Desvio	0.143	0.337	0.113	0.372	0.371	0.103	0.351
		Min.	1.157	0.434	1.160	1.111	0.605	1.114	0.729

A Tabela 24 apresenta os resultados da predição geral dos intervalos nos cenário 7 a 12. No cenário 7, o método iCRCNLR apresenta resultados consistentemente melhores que o caso linear, com resultados médios menores e estatisticamente significativos, de acordo com a Tabela 27. No cenário 8, o único resultado sem significância estatística no que diz respeito aos métodos de ajuste foi a combinação alocação aleatória com 2 *clusters*. Em todos os outros casos, o método iCRCNLR apresenta melhor resultado médio. No cenário 9, apenas os casos onde a alocação é aleatória não são estatisticamente significativos. Considerando 1 *cluster*, o método iCRCLR apresenta menor RMSE médio, pela presença de um *outlier* na amostra dos RMSE's obtidos pelo método iCRCNLR. Nos outros casos, o método iCRCNLR apresenta melhores resultados médios. No cenário 10, os resultados estatisticamente significativos só se apresentam quando o ajuste é feito para 3 *clusters*. Nestes casos, o método iCRCNLR apresenta melhor desempenho médio apenas com o método de alocação KNN. Para 1 *cluster* a diferença entre os métodos não é significativa. No cenário 11, as combinações em que os métodos apresentam diferença significativa são *Stacked Regressions* com 2 *clusters*; *Stacked Regressions* e aleatório com 3 *clusters*, nos quais o método iCRCNLR apresenta menor RMSE médio. Para o cenário 12, apenas os resultados em que o método de alocação é o *Stacked Regressions* apresentam resultado significativo para diferença entre os métodos. Nestes casos, o método iCRCNLR apresenta melhor resultado médio.

Tabela 24 – RMSE médio da predição geral dos limites utilizando três métodos de alocação nos cenários 7 a 12.

Cenário	Método	Medida	1 Cluster	2 Clusters			3 Clusters		
				KNN	SR	Random	KNN	SR	Random
7	iCRCLR	Média	2.478	1.195	7.965	4.426	0.795	8.068	6.561
		Desvio	1.610	0.432	1.737	2.605	0.405	1.901	4.648
		Min.	1.163	0.391	3.849	1.090	0.144	3.827	1.089
	iCRCNLR	Média	1.439	0.632	2.804	3.368	**0.105**	8.172	6.725
		Desvio	0.641	0.346	1.242	2.435	0.037	2.160	4.374
		Min.	0.688	0.076	1.212	0.813	0.046	2.75	1.024
8	iCRCLR	Média	2.457	1.465	4.080	3.272	0.984	4.365	3.795
		Desvio	0.663	0.335	0.813	0.858	0.306	0.857	0.887
		Min.	1.319	0.784	2.232	1.474	0.326	2.291	2.244
	iCRCNLR	Média	2.105	0.983	3.055	3.079	**0.509**	5.475	10.533
		Desvio	0.416	0.160	1.046	0.757	0.414	2.178	9.418
		Min.	1.264	0.606	1.511	0.999	0.070	2.295	2.173
9	iCRCLR	Média	2.054	0.915	3.338	3.106	0.882	3.489	3.492
		Desvio	0.683	0.812	0.525	1.065	0.830	0.587	1.631
		Min.	1.224	0.132	2.232	0.281	0.077	2.195	0.809
	iCRCNLR	Média	29.456	**0.864**	2.092	2.971	0.934	2.241	2.815
		Desvio	194.097	0.916	0.354	1.026	0.958	0.446	0.882
		Min.	1.167	0.062	1.363	0.627	0.066	1.436	1.495
10	iCRCLR	Média	1.289	0.897	1.239	1.544	0.796	1.368	2.515
		Desvio	0.498	0.242	0.186	0.311	0.328	0.260	0.912
		Min.	0.780	0.385	0.861	0.805	0.161	0.887	0.815
	iCRCNLR	Média	1.168	0.919	1.301	1.772	**0.676**	1.932	2.787
		Desvio	0.228	0.281	0.209	0.432	0.480	0.494	0.481
		Min.	0.850	0.462	0.962	0.988	0.061	1.207	0.951
11	iCRCLR	Média	1.942	1.835	2.284	2.455	1.850	2.421	2.894
		Desvio	1.060	0.679	0.500	0.863	0.682	0.559	1.664
		Min.	1.179	0.207	1.278	0.898	0.172	1.395	0.679
	iCRCNLR	Média	1.913	1.805	**1.741**	2.313	1.855	1.857	2.532
		Desvio	0.984	0.861	0.308	0.878	0.866	0.442	7.851
		Min.	1.092	0.055	1.089	0.076	0.056	1.094	0.191
12	iCRCLR	Média	2.082	2.133	2.561	2.752	2.335	2.397	2.810
		Desvio	0.462	0.617	0.581	0.597	0.648	0.547	0.823
		Min.	1.241	1.050	1.468	1.351	1.025	1.374	1.158
	iCRCNLR	Média	2.149	2.314	2.174	2.599	2.464	**2.059**	2.654
		Desvio	0.497	0.644	0.492	0.844	0.592	0.465	7.741
		Min.	1.204	0.703	1.237	1.360	0.737	1.147	1.216

A Tabela 25 apresenta os resultados para os cenários 13 a 18, com centro não linear e amplitude linear. No cenário 13, com exceção da combinação 2 *clusters* e método de alocação aleatório, os resultados apresentaram diferença estatisticamente significativa entre os métodos. Nestes casos, o método iCRCNLR apresenta RMSE médio menor que o método de ajuste linear. Nos cenários 14, todas as combinações foram significativas, com o método iCRCNLR apresentando melhor performance média em todas as situações. No cenário 15, todas as combinações apresentaram diferença significativa entre os métodos de ajuste, com o método iCRCNLR apresentando melhor desempenho nas combinações *Stacked Regressions* e aleatório com 2 *clusters*, e nos 3 métodos de alocação considerando 3 *clusters*. No cenário 16, as combinações com 3 *clusters* apresentam resultados estatisticamente significativos para a diferença entre os métodos, com o iCRCNLR tendo RMSE médio menor, exceto com 1 *cluster*. No cenário 17, são estatisticamente significativos os resultados para *Stacked Regressions* e alocação aleatória em 2 e 3 *clusters* e em 1 *cluster*. Neste cenário, em todas as combinações, tem-se melhor desempenho do método iCRCNLR. No cenário 18, não houve diferença significativa entre os métodos, de acordo com o teste de Mann-Whitney.

Tabela 25 – RMSE médio da predição geral dos limites utilizando três métodos de alocação nos cenários 13 a 18.

Cenário	Método	Medida	1 Cluster	2 Clusters			3 Clusters		
				KNN	SR	Random	KNN	SR	Random
13	iCRCLR	Média	3.032	1.615	7.007	8.434	1.279	6.783	10.003
		Desvio	1.126	0.738	2.916	5.237	0.668	2.372	5.858
		Min.	1.531	0.440	2.643	0.957	0.301	1.757	1.155
	iCRCNLR	Média	1.981	0.116	3.157	6.540	**0.096**	3.478	5.682
		Desvio	1.359	0.045	1.322	2.612	0.027	1.451	3.383
		Min.	0.961	0.057	1.762	1.016	0.045	1.408	0.122
14	iCRCLR	Média	10.871	5.939	27.705	15.677	3.833	29.782	15.896
		Desvio	2.734	1.412	3.755	3.998	1.091	7.138	5.454
		Min.	5.768	3.546	19.541	6.760	1.357	14.550	7.080
	iCRCNLR	Média	4.025	1.474	8.912	11.171	**0.161**	11.392	30.198
		Desvio	1.687	0.302	7.309	9.111	0.086	6.225	13.598
		Min.	1.405	0.762	3.081	2.640	0.074	4.746	7.740
15	iCRCLR	Média	2.508	1.524	2.741	3.879	1.463	3.705	4.560
		Desvio	0.635	0.909	0.473	1.082	0.939	1.070	1.751
		Min.	1.597	0.218	1.737	0.485	0.212	2.278	1.088
	iCRCNLR	Média	2.561	**1.164**	2.273	3.449	1.183	2.759	3.432
		Desvio	2.621	1.089	0.273	0.963	1.117	0.788	0.933
		Min.	1.521	0.065	1.799	0.081	0.061	1.496	1.219
16	iCRCLR	Média	1.790	1.323	2.230	2.557	1.199	3.625	4.775
		Desvio	0.729	0.323	0.603	0.996	0.345	2.108	3.923
		Min.	0.838	0.448	1.087	1.328	0.416	1.412	1.414
	iCRCNLR	Média	1.924	1.239	1.616	2.109	**0.926**	1.762	3.362
		Desvio	2.216	0.312	0.345	0.527	0.497	0.340	1.088
		Min.	1.126	0.533	1.093	1.193	0.058	1.205	1.743
17	iCRCLR	Média	2.599	2.249	2.091	2.712	2.227	2.574	3.419
		Desvio	3.504	0.599	0.428	1.044	0.619	0.920	1.683
		Min.	1.250	0.438	1.242	0.632	0.287	1.329	0.637
	iCRCNLR	Média	2.069	2.186	**1.880**	2.468	2.212	2.151	2.517
		Desvio	0.985	0.679	0.228	0.735	0.650	0.440	0.691
		Min.	1.240	0.740	1.332	0.819	0.741	1.295	0.877
18	iCRCLR	Média	**1.934**	2.054	2.129	2.586	2.156	1.966	2.518
		Desvio	0.507	0.689	0.573	0.829	0.741	0.466	0.864
		Min.	1.061	0.533	0.898	1.146	0.650	1.122	0.916
	iCRCNLR	Média	1.959	2.206	2.087	2.696	2.397	2.069	2.799
		Desvio	0.477	0.632	0.514	0.742	0.694	0.371	0.882
		Min.	0.916	0.850	1.089	1.417	0.875	0.828	0.906

Por fim, a Tabela 26 apresenta os resultados da predição para os intervalos, de modo geral, nos cenários 19 a 24. No cenário 19, a comparação entre os métodos de ajuste apresenta diferença significativa entre eles, exceto quando o método de alocação utilizado é aleatório, com o iCRCNLR apresentando desempenho médio melhor. No cenário 20, há diferença significativa entre os métodos em todas as combinações, exceto onde o método de alocação é aleatório com 2 *clusters*. O método linear, iCRCLR, apresenta desempenho melhor com método *Stacked Regressions* e aleatório em 3 *clusters*. No cenário 21, os resultados com método de alocação KNN não foram significativos em relação à diferença entre os métodos. Nos casos em que houve significância no teste de Mann-Whitney, o método de ajuste iCRCNLR apresentou melhor desempenho médio. No cenário 22, a diferença entre os métodos foi significativa nas seguintes combinações: 2 *clusters* com *Stacked Regressions* e 3 *clusters* com *Stacked Regressions* e aleatório. No cenário 23, as combinações em que há significância estatística ocorreram em 1 *cluster*, 2 *clusters* com alocação KNN e 3 *clusters* com alocação *Stacked Regressions* e aleatória. Em todos estes cenários, o método de ajuste iCRCNLR apresentou melhor desempenho médio, de acordo com o critério RMSE. Finalmente, no cenário 24, houve evidência para rejeitar H_0 no teste de Mann-Whitney apenas na combinação *Stacked Regressions* com 3 *clusters*.

Tabela 26 – RMSE médio da predição geral dos limites utilizando três métodos de alocação nos cenários 19 a 24.

Cenário	Método	Medida	1 Cluster	2 Clusters			3 Clusters		
				KNN	SR	Random	KNN	SR	Random
19	iCRCLR	Média	2.577	1.437	8.187	4.059	1.064	9.266	6.119
		Desvio	1.088	0.478	2.182	1.891	0.472	2.828	3.977
		Min.	1.017	0.223	4.261	1.495	0.215	2.245	0.860
	iCRCNLR	Média	1.163	0.411	2.764	4.905	**0.093**	4.018	5.291
		Desvio	0.431	0.327	1.231	3.649	0.028	1.881	3.256
		Min.	0.565	0.049	1.443	0.651	0.049	1.430	1.117
20	iCRCLR	Média	2.370	1.467	3.324	3.806	0.924	3.547	4.333
		Desvio	0.632	0.361	1.046	1.445	0.234	1.009	1.348
		Min.	1.268	0.795	1.590	1.665	0.333	1.723	1.749
	iCRCNLR	Média	2.044	1.055	2.791	3.204	**0.595**	4.947	8.064
		Desvio	0.486	0.212	0.801	0.980	0.390	1.762	6.740
		Min.	0.833	0.704	1.314	1.559	0.064	1.831	1.703
21	iCRCLR	Média	2.409	1.614	2.670	3.713	**1.471**	3.717	4.505
		Desvio	1.047	0.925	0.649	1.048	0.959	1.216	2.672
		Min.	1.357	0.124	1.552	1.078	0.138	1.586	0.244
	iCRCNLR	Média	1.948	1.685	2.190	3.253	1.709	2.304	3.011
		Desvio	0.410	1.166	0.258	0.710	1.232	0.279	8.046
		Min.	1.098	0.065	1.648	0.078	0.060	1.661	1.285
22	iCRCLR	Média	4.745	2.860	8.156	8.722	**2.242**	8.390	8.916
		Desvio	1.423	1.102	2.326	2.739	1.198	2.419	3.054
		Min.	2.123	1.167	3.566	2.365	0.762	3.806	2.605
	iCRCNLR	Média	4.638	3.028	6.420	9.656	2.720	6.648	12.574
		Desvio	1.170	1.994	2.076	5.591	2.254	1.704	6.880
		Min.	3.032	0.910	3.349	2.162	0.099	4.072	4.274
23	iCRCLR	Média	1.842	1.851	2.341	2.108	1.863	3.824	4.724
		Desvio	0.860	0.754	0.357	0.847	0.772	2.626	4.123
		Min.	1.220	0.251	1.124	0.192	0.138	1.295	0.860
	iCRCNLR	Média	1.796	1.749	**1.557**	1.999	1.795	1.684	2.098
		Desvio	1.024	0.746	0.242	0.786	0.688	0.383	0.672
		Min.	1.001	0.061	1.094	0.052	0.082	1.014	0.556
24	iCRCLR	Média	2.146	2.220	2.224	2.871	2.413	2.285	2.889
		Desvio	0.402	0.481	0.445	0.741	0.574	0.405	0.704
		Min.	1.414	1.174	1.432	1.304	0.602	1.486	0.933
	iCRCNLR	Média	2.098	2.262	2.102	2.743	2.610	**2.032**	2.824
		Desvio	0.388	0.583	0.388	0.695	0.739	0.333	6.574
		Min.	1.389	1.178	1.232	1.564	1.085	1.142	1.492

Tabela 27 – Valor-p do teste de Mann-Whitney para o RMSE geral dos intervalos, algoritmos iCRCLR e iCRCNLR.

Cenário	1 Cluster	2 Clusters			3 Clusters		
		KNN	SR	Random	KNN	SR	Random
1	**0.000**	**0.000**	**0.000**	**0.000**	**0.001**	**0.000**	**0.000**
2	0.913	**0.000**	**0.003**	0.631	0.148	**0.000**	**0.000**
3	0.338	0.098	**0.004**	0.603	0.375	0.051	**0.050**
4	0.536	**0.002**	**0.000**	0.850	0.972	**0.000**	**0.000**
5	0.114	0.080	0.869	0.231	0.146	0.026	0.382
6	0.976	0.155	0.131	0.768	0.198	**0.001**	0.879
7	**0.000**	**0.000**	**0.000**	**0.002**	**0.000**	**0.000**	0.082
8	**0.000**	**0.000**	**0.000**	0.556	**0.000**	**0.001**	**0.000**
9	**0.028**	**0.015**	**0.000**	0.960	**0.043**	**0.000**	0.047
10	0.265	0.944	0.118	0.147	**0.025**	**0.000**	**0.015**
11	0.916	0.265	**0.000**	0.236	0.573	**0.000**	**0.003**
12	0.388	0.314	**0.002**	0.178	0.143	**0.007**	0.745
13	**0.000**	**0.000**	**0.000**	0.199	**0.000**	**0.000**	**0.000**
14	**0.000**	**0.000**	**0.000**	**0.005**	**0.000**	**0.000**	**0.001**
15	**0.000**	**0.005**	**0.000**	**0.050**	**0.014**	**0.000**	**0.000**
16	0.220	0.153	**0.000**	0.113	**0.001**	**0.000**	**0.010**
17	**0.028**	0.709	**0.009**	**0.030**	0.887	**0.027**	**0.000**
18	0.637	0.210	0.989	0.384	0.071	0.118	**0.037**
19	**0.000**	**0.000**	**0.000**	0.648	**0.000**	**0.000**	0.718
20	**0.004**	**0.000**	**0.017**	0.422	**0.000**	**0.000**	**0.003**
21	**0.000**	0.909	**0.000**	**0.002**	0.709	**0.000**	**0.000**
22	0.683	0.184	**0.000**	0.989	0.995	**0.001**	**0.001**
23	**0.028**	0.567	**0.000**	0.060	0.769	**0.000**	**0.000**
24	0.469	0.776	0.151	0.253	0.214	**0.001**	0.551

A Tabela 28 apresenta os melhores e piores pares de método de alocação e de ajuste para predição geral dos intervalos de acordo com o critério do menor RMSE médio. O método iCRCNLR aparece como melhor em 17 dos 24 cenários, com empate no cenário 2. Em relação aos métodos de alocação, o KNN se destaca como melhor em 16 dos 24 cenários, seguido pelo *Stacked Regressions*, melhor em 6 cenários. O ajuste com apenas um *cluster* apresentou menor RMSE médio em dois cenários, 5 e 18.

Tabela 28 – Melhores e piores pares Modelo-Método de alocação para predição geral dos intervalos nos 24 cenários.

Cenário	Melhor Par			Pior Par		
	Clusters	Modelo	Alocação	*Clusters*	Modelo	Alocação
1	3	**iCRCNLR**	KNN	3	iCRCLR	SR
2	3	**iCRCNLR** iCRCLR	KNN	3	**iCRCNLR**	Random
3	3	iCRCLR	KNN	3	iCRCLR	Random
4	3	iCRCLR	KNN	3	**iCRCNLR**	Random
5	1	iCRCLR	-	3	**iCRCNLR**	Random
6	2	**iCRCNLR**	SR	2	**iCRCNLR**	Random
7	3	**iCRCNLR**	KNN	3	**iCRCNLR**	SR
8	3	**iCRCNLR**	KNN	3	**iCRCNLR**	Random
9	2	**iCRCNLR**	KNN	1	**iCRCNLR**	-
10	3	**iCRCNLR**	KNN	3	**iCRCNLR**	Random
11	2	**iCRCNLR**	SR	3	iCRCLR	Random
12	3	**iCRCNLR**	SR	3	iCRCLR	Random
13	3	**iCRCNLR**	KNN	3	iCRCLR	Random
14	3	**iCRCNLR**	KNN	3	**iCRCNLR**	Random
15	2	**iCRCNLR**	KNN	3	iCRCLR	Random
16	3	**iCRCNLR**	KNN	3	iCRCLR	Random
17	2	**iCRCNLR**	SR	3	iCRCLR	Random
18	1	iCRCLR	-	3	**iCRCNLR**	Random
19	3	**iCRCNLR**	KNN	3	iCRCLR	SR
20	3	**iCRCNLR**	KNN	3	**iCRCNLR**	Random
21	3	iCRCLR	KNN	3	iCRCLR	Random
22	3	iCRCLR	KNN	3	**iCRCNLR**	Random
23	2	**iCRCNLR**	SR	3	iCRCLR	Random
24	3	**iCRCNLR**	SR	3	iCRCLR	Random

Os resultados do experimento nos 24 cenários envolvendo o algoritmo proposto,

iCRCNLR, frente ao modelo linear, iCRCLR, mostram que a predição de dados tipo-intervalo por meio de regressão *clusterwise* tem acurácia dependente do método de ajuste e de alocação a um *cluster* ajustado. Desta forma, mostrou-se experimentalmente que a inclusão do ajuste de métodos não lineares para o centro e amplitude no algoritmo de regressão *clusterwise* tem potencial para gerar melhores predições, no sentido de obter menor RMSE. Ainda, dentre os métodos de alocação apresentados, o KNN para dados tipo-intervalo utilizando distância de Hausdorff tem bom desempenho em casos em que os grupos presentes no centro e amplitude dos dados tenha sobreposição parcial ou ausência de sobreposição. No caso de dados em que haja sobreposição total dos grupos de centro e amplitude, o método *Stacked Regressions* mostra-se mais promissor, dado que oferece como valor predito um combinação de predições individuais em cada *cluster* ajustado. Ainda, o método iCRCNLR permite o ajuste de funções mais flexíveis, adequadas para descrever relações não lineares entre as variáveis. No entanto, nota-se que o método iCRCNLR, apesar de apresentar menores valores mínimos de RMSE, está mais exposto a cometer erros de grande monta, uma vez que pode ajustar funções com grande taxa de crescimento ou decrescimento, de tal modo que uma alocação errada pode levar a um valor muito elevado de RMSE.

5.2 EXPERIMENTOS COM DADOS REAIS

Nesta seção, é avaliado o desempenho do método iCRCNLR em conjuntos de dados reais em comparação com o método iCRCLR. São propostas duas avaliações para a adequação do algoritmo aos dados.

Primeiramente, são comparados os critérios de adequação J do iCRCNLR e do iCRCLR para um número fixo de *clusters* $k = \{1, 2, 3\}$ ajustados sobre todo o conjunto de dados. Por definição, para qualquer conjunto de dados, J será reduzido ao passo em que o número de *clusters* aumenta. Além dos métodos iCRCNLR e iCRCLR, uma abordagem alternativa foi proposta para avaliar o valor do critério de adequação J: a utilização independente de agrupamento, representado pelo método K-means para dados intervalo utilizando distância de Hausdorff (CHAVENT et al., 2006), e o ajuste de modelos lineares dentro de cada *cluster* obtido pelo agrupamento. O objetivo deste primeiro experimento é demonstrar que a regressão *clusterwise*, ao fazer o agrupamento e a modelagem alternadamente, obtém uma soma de erros nos *clusters* menor do que a execução do agrupamento e da regressão separadamente.

O segundo experimento compara os métodos iCRCNLR, CRCLR e a abordagem que utiliza agrupamento e regressão linear em relação à predição para novas observações. Com objetivo de avaliar o RMSE da predição de ambos os métodos, utilizando as três técnicas de alocação para a seleção do *cluster* para novas observações, foram feitas 10 execuções de validação cruzada 10-*folds* e computado o RMSE médio das predições nos *folds* de teste. Neste segundo experimento, também foi computada a média do erro

percentual absoluto (*Mean Absolute Percentage Error* - MAPE) para medir a acurácia das predições para os dados reais nos métodos iCRCLR e iCRCNLR. O MAPE é calculado como:

$$MAPE = \frac{100}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|, \tag{5.7}$$

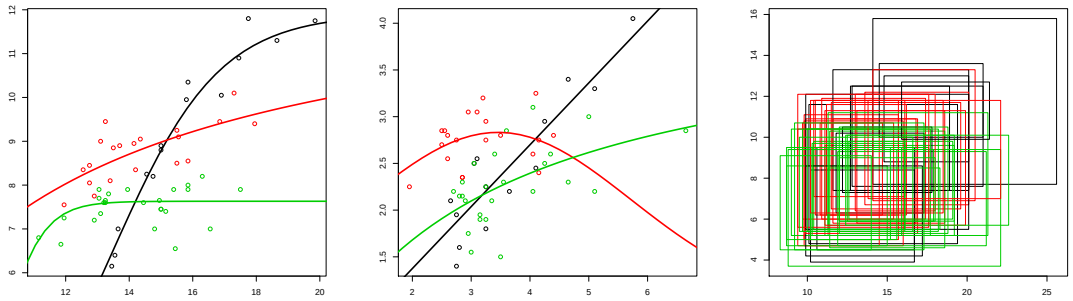
em que n é a quantidade de pontos ajustados, A_t é o valor observado do ponto e F_t é o valor predito pelo método. O MAPE é expresso como um erro percentual pelo qual o valor predito difere do valor observado. Finalmente, são apresentadas tabelas com o *rank* dos métodos em relação ao RMSE médio e mínimo da predição em cada conjunto de dados. O criterio de adequação para esta metodologia será dado pela soma dos erros dos modelos dentro de cada *cluster*. Dado que o agurpamento K-means fornece uma partição do conjunto de dados tipo-intervalo E em K *clusters*, teremos, após os modelos lineares serem ajustados a seguinte expressão para o critério de adequação $J_{K-means}$:

$$J_{k-means} = \sum_{k=1}^{K} \sum_{e_i \in P_k} \left(y_i - \beta_{0(k)} - \beta_{1(k)} x_i \right) \tag{5.8}$$

em que P_k é o k-ésimo *cluster*, e_i é o i-ésimo elemento dentro do cluster k, e $\beta_{0(k)}$ e $\beta_{1(k)}$ são os parâmetros do modelo linear ajustado no *cluster* k. É importante notar que, neste método, o critério de adequação, isto é, o valor a ser minimizado possui duas etapas: (i) no agrupamento, a minimização ocorre em relação aos K centroides dos grupos e (ii) na regressão linear, a minimização ocorre em relação a Equação (5.8). Nos métodos *clusterwise* aqui apresentados, o critério é minimizado sempre tendo em vista a soma dos quadrados dos erros dos modelos ajustados. Com isso, espera-se que estes métodos tenham desempenho médio melhor, porém, também é esperado que estes apresentem mais oscilação no desempenho, uma vez que o agrupamento com base nos centroides dos *clusters* é mais estável do que o agrupamento feito com base em modelos de regressão.

5.2.1 Conjunto de dados *Cardio*

Este conjunto de dados tipo-intervalo reflete o relacionamento entre a pressão sanguínea sistólica (X) e diastólica (Y) para 59 pacientes em um hospital (BLANCO-FERNÁNDEZ; CORRAL; GONZÁLEZ-RODRÍGUEZ, 2011). A medidas foram tomadas ao longo do dia e os valores máximo e mínimo da pressão foram computados. Assim, para identificar e modelar grupos de pacientes, pode-se utilizar um modelo de regressão *clusterwise*. A Figura 3 apresenta os gráficos do centro, amplitude e intervalos para o conjunto de dados Cardio após o ajuste do iCRCNLR para três *clusters*.



(a) Centro (b) Amplitude (c) Intervalos

Figura 3 – Gráficos do centro, amplitude e intervalos com as curvas ajustadas pelo iCRCNLR, considerando três *clusters* para os dados *Cardio*

Os modelos apresentados na Figura 3 são os especificados na Tabela 29 para 3 *clusters*. Também são apresentados os valores do critério J para os métodos K-means + regressão linaer, iCRCLR e iCRCNLR. No primeiro caso, é possível alcançar uma redução na soma dos erros em cada *cluster*.

Tabela 29 – Modelos selecionados, valor final do critério e estimativas de parâmetros sobre o conjunto de dados *Cardio* utilizando os métodos iCRCLR e iCRCNLR.

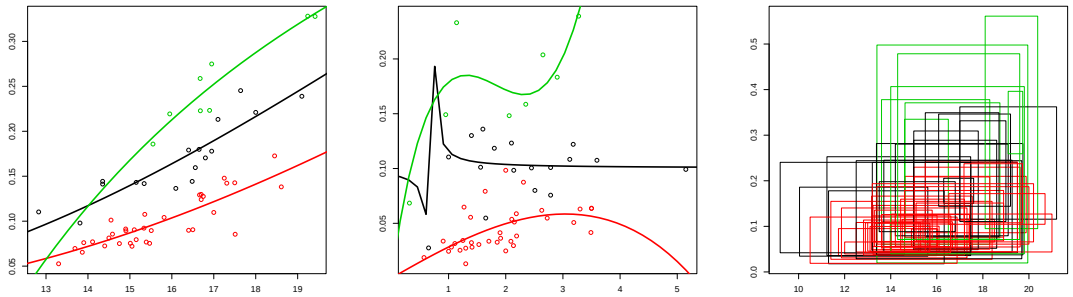
Método	*Clusters*	J	Modelos Selecionados		Estimativas dos parâmetros					
			Centro	Amplitude	Centro			Amplitude		
					α_1	α_2	α_3	α_1	α_2	α_3
K-means + regressão linear	1	65.069	-	-	1.685	0.453	-	1.584	0.257	-
	2	60.960	-	-	6.335	0.102	-	2.118	0.074	-
			-	-	-0.096	0.566	-	1.509	0.283	-
	3	46.07	-	-	10.946	-0.172	-	1.905	0.178	-
			-	-	5.888	0.138	-	2.116	0.080	-
			-	-	1.296	0.523	-	0.397	0.593	-
iCRCLR	1	65.066	-	-	1.685	0.453	-	1.584	0.257	-
	2	25.227	-	-	1.404	0.523	-	1.385	0.391	-
			-	-	4.031	0.242	-	1.501	0.212	-
	3	16.126	-	-	4.997	0.261	-	2.449	0.095	-
			-	-	1.992	0.352	-	1.269	0.248	-
			-	-	-1.346	0.687	-	0.265	0.612	-
iCRCNLR	1	63.195	9	7	3.901	-0.062	1.613	1.865	-0.238	6.608
	2	24.753	9	2	4.264	-0.061	0.431	-0.501	28.055	-12.217
			9	Linear	4.720	-0.038	-1.810	1.501	0.212	-
	3	14.205	7	Linear	7.974	-0.592	11.911	0.050	0.662	-
			2	11	12.974	62.366	0.611	1.571	-0.674	0.097
			7	3	15.007	-1.514	7.631	2.952	4.162	-

No entanto, como o agrupamento é realizado em relação á distância de Hausdorff, e não em relação aos erros dos modelos de regressão, esta redução em J ocorre em menor escala do que nos outros métodos. Ao aplicar o método *clusterwise*, é possível

obter uma maior redução no critério J. À medida em que o método iCRCNLR consegue ajustar modelos não lineares, a redução em J é ainda maior. Os números apresentados na coluna Modelos Selecionados referem-se aos modelos apresentados na Tabela 2.

5.2.2 Conjunto de dados *Tree*

O conjunto de dados Tree (FILHO et al., 2012) contém as medidas máxima e mínima do volume do tronco (Y) e altura (X) de 60 grupos de clones de *Eucalyptus* na região do Araripe, Brasil. A Figura 4 apresenta as curvas ajustadas pelo método iCRCNLR, considerando três *clusters* para os dados Tree. A Tabela 30 apresenta a comparação entre o algoritmo iCRCNLR e o modelo linear, ICRCLR, neste conjunto de dados. Nota-se que a redução no critério de adequação J foi pequena, apesar do algoritmo iCRCNLR selecionar majoritariamente modelos não-lineares.



(a) Centro (b) Amplitude (c) Intervalos

Figura 4 – Gráficos do centro, amplitude e intervalos com as curvas ajustadas pelo iCRCNLR, considerando três *clusters* para os dados *Tree*

Utilizando a metodologia em que o agrupamento K-means para dados intervalo e o ajuste dos modelos lineares são feitos sequencialmente, nota-se uma redução na soma dos erros das regressões em cada *cluster*, mas novamente, esta redução se dá em menor escala do que nas aplicações de regressão *clusterwise*. Para um número fixo de *clusters*, iCRCNLR e iCRCLR apresentam valores similares, com o primeiro apresentando resultados um pouco melhores, especialmente ao considerar três *clusters*. Novamente, iCRCNLR alcança resultados melhores por meio do uso de funções não lineares, que são mais flexíveis e podem ajustar melhor os dados. No entanto, esta capacidade de ajustar funções não lineares pode levar ao *overfit*, que prejudicaria a predição.

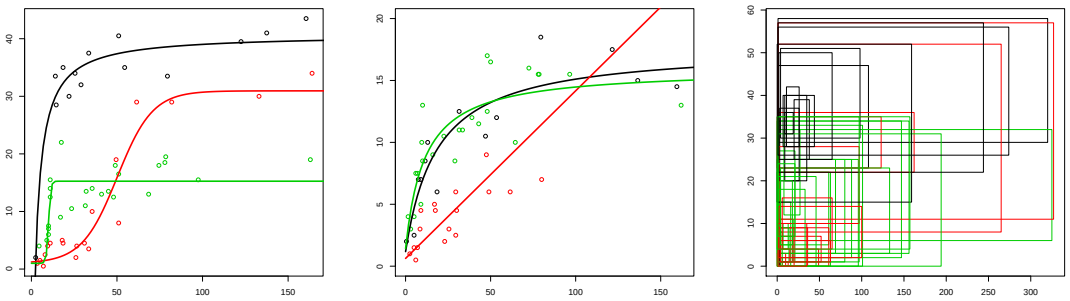
Tabela 30 – Modelos selecionados, valor final do critério e estimativas de parâmetros sobre o conjunto de dados *Tree* utilizando os métodos iCRCLR e iCRCNLR.

Método	*Clusters*	J	Modelos Selecionados		Estimativas dos parâmetros					
					Centro			Amplitude		
			Centro	Amplitude	α_1	α_2	α_3	α_1	α_2	α_3
K-means + regressão linear	1	0.278	-	-	-0.361	0.031	-	0.044	0.016	-
	2	0.237	-	-	-0.067	0.011	-	0.021	0.015	-
			-	-	-0.426	0.035	-	0.077	0.011	-
	3	0.225	-	-	-0.679	0.049	-	0.081	0.007	-
			-	-	0.516	-0.020	-	0.096	0.005	-
			-	-	-0.067	0.011	-	0.021	0.015	-
iCRCLR	1	0.278	-	-	-0.361	0.031	-	0.044	0.016	-
	2	0.086	-	-	-0.128	0.015	-	0.012	0.019	-
			-	-	-0.266	0.029	-	0.083	0.037	-
	3	0.046	-	-	-0.337	0.033	-	0.086	0.039	-
			-	-	-0.129	0.014	-	0.014	0.013	-
			-	-	-0.082	0.014	-	0.058	0.013	-
iCRCNLR	1	0.275	7	11	-10.884	0.507	0.065	0.039	0.023	-0.001
	2	0.084	8	12	-4.539	0.152	-	0.031	-0.002	-
			Linear	13	-0.334	0.032	-	0.333	-0.223	0.045
	3	0.036	6	2	-9.706	2.913	-	0.100	-0.004	-0.710
			6	13	-10.450	2.990	-	0.03	-0.001	-0.001
			2	13	0.884	10.733	-0.021	0.384	-0.228	0.041

5.2.3 Conjunto de dados *Unemployment*

Este conjunto de dados tipo-intervalo informa sobre o desemprego em Portugal (DIAS; BRITO, 2017) baseado no logaritmo do tempo de desemprego X e o tempo em que as pessoas desempregadas trabalharam previamente Y, para 58 classes de indivíduos agrupados de acordo com gênero, região, idade e educação. Neste caso, o objetivo é prever o tempo de experiência de trabalho por meio do tempo em que o indivíduo leva para conseguir realocação. Para cada classe, as variáveis representam os valores mínimo e máximo observados para o conjunto de indivíduos.

A Figura 5 mostra a representação do centro, amplitude e intervalos para este conjunto de dados. Novamente, as curvas e *clusters* apresentados foram gerados pelo método iCRCNLR considerando três *clusters*. Os resultados para o ajuste sobre todo o conjunto de dados, para os métodos iCRCNLR e iCRCLR são apresentados na Tabela 31.



(a) Centro (b) Amplitude (c) Intervalos

Figura 5 – Gráficos do centro, amplitude e intervalos com as curvas ajustadas pelo iCRCNLR, considerando três *clusters* para os dados *Unemployment*

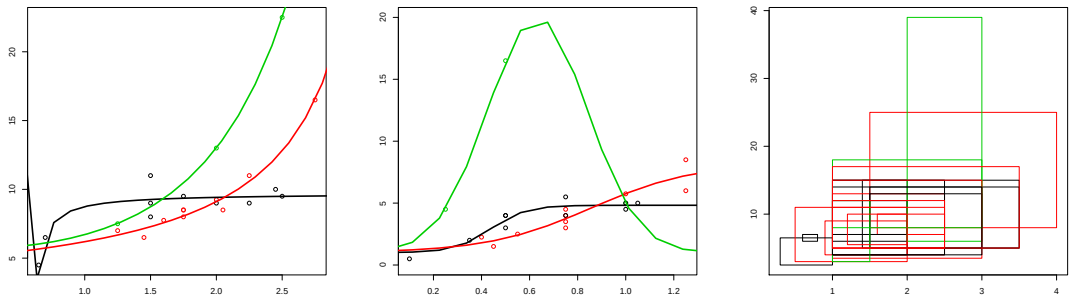
Para um número fixo de *clusters*, a introdução de funções não-lineares para ajustar os dados reduz o valor do critério de adequação J. No método iCRCNLR, em dois dos *clusters* foram ajustadas funções lineares, que consiste num caso especial algoritmo. A queda no valor do critério de adequação J foi mais acentuada para o ajuste com 3 *clusters*. O método *K-means* combinado com regressão linear obtém redução no critério à medida em que o número de *clusters* cresce, no entanto, esta redução não alcança o mesmo patamar dos outros métodos.

Tabela 31 – Modelos selecionados, valor final do critério e estimativas de parâmetros sobre o conjunto de dados *Unemployment* utilizando os métodos iCRCLR e iCRCNLR.

Método	*Clusters*	J	Modelos Selecionados		Estimativas dos parâmetros					
			Centro	Amplitude	Centro			Amplitude		
					α_1	α_2	α_3	α_1	α_2	α_3
K-means + regressão linear	1	6833.88	-	-	9.428	0.137	-	5.101	0.094	-
	2	6564.207	-	-	7.769	0.223	-	4.091	0.129	-
			-	-	16.181	0.111	-	9.735	0.051	-
	3	6487.629	-	-	4.808	0.439	-	4.697	0.050	-
			-	-	11.132	0.147	-	16.230	0.007	-
			-	-	11.153	0.166	-	7.570	0.078	-
iCRCLR	1	6833.88	-	-	9.428	0.137	-	5.101	0.094	-
	2	1955.704	-	-	2.379	0.242	-	4.551	0.102	-
			-	-	33.159	-0.029	-	6.592	0.079	-
	3	1041.804	-	-	2.437	0.193	-	2.993	0.14	-
			-	-	12.570	0.204	-	10.255	0.044	-
			-	-	36.018	-0.086	-	7.051	0.034	-
iCRCNLR	1	6477.532	5	7	3.366	-0.859	41.358	1.354	-0.038	17.098
	2	1250.048	2	Linear	36.583	63.908	-7.917	6.946	0.077	-
			7	7	2.484	-0.075	18.046	2.469	-0.078	14.996
	3	759.405	2	2	40.710	172.879	1.978	17.689	306.317	18.552
			7	Linear	4.870	-0.097	30.945	0.626	0.135	-
			7	2	16.096	-1.597	15.256	15.936	163.263	11.244

5.2.4 Conjunto de dados *Mushroom*

O conjunto de dados *mushroom* (DOMINGUES; SOUZA; CYSNEIROS, 2010) apresenta 23 espécies de uma família de cogumelos *Amanita*. O experimento consiste na predição da espessura do estipe Y por meio da variável resposta tamanho do píleo X. A Figura 6 apresenta os dados de centro, amplitude e intervalos para este conjunto de dados ajustados pelo algoritmo iCRCNLR para com três *clusters*.



(a) Centro (b) Amplitude (c) Intervalos

Figura 6 – Gráficos do centro, amplitude e intervalos com as curvas ajustadas pelo iCRCNLR, considerando três *clusters* para os dados *Mushroom*

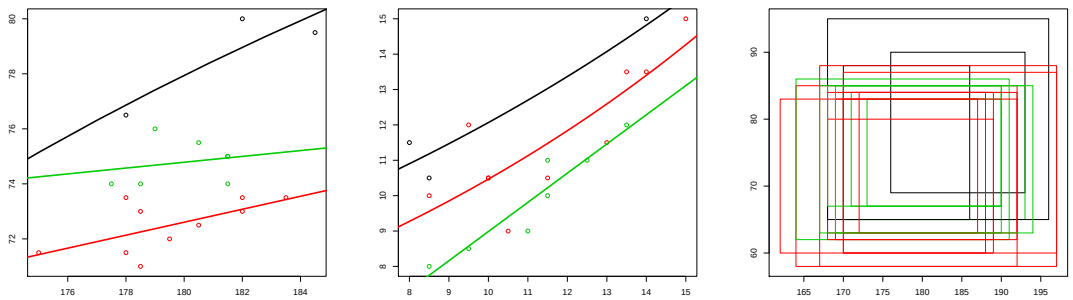
A Tabela 32 apresenta os modelos selecionados para centro e amplitude pelo algoritmo iCRCNLR, estimativas de parâmetros e a redução obtida no valor do critério de adequação J. Novamente, o método iCRCNLR apresenta menor valor do critério J para 1, 2 e 3 *clusters*, seguido pelo método *clusterwise* linear e pelo método que aplica *K-means* sequencialmente. Nota-se que, ao ajustar os modelos *clusterwise* para 2 *clusters* obtém-se uma grande redução do critério de adequação em relação ao método *K-means* + regressão linear.

Tabela 32 – Modelos selecionados, valor final do critério e estimativas de parâmetros sobre o conjunto de dados *Mushroom* utilizando os métodos iCRCLR e iCRCNLR.

Método	*Clusters*	J	Modelos Selecionados		Estimativas dos parâmetros					
			Centro	Amplitude	Centro			Amplitude		
					α_1	α_2	α_3	α_1	α_2	α_3
K-means + regressão linear	1	349.634	-	-	1.264	4.651	-	2.300	3.225	-
	2	280.728	-	-	4.047	2.409	-	1.436	2.793	-
			-	-	0.014	5.392	-	9.360	-3.525	-
	3	31.086	-	-	4.047	2.409	-	1.436	2.793	-
			-	-	10.498	-0.256	-	2.886	2.252	-
			-	-	82.500	-24.000	-	21.833	-10.666	-
iCRCLR	1	349.634	-	-	1.264	4.651	-	2.300	3.225	-
	2	49.346	-	-	82.5	-24	-	21.833	1.167	-
			-	-	4.221	2.586	-	-10.666	3.853	-
	3	18.710	-	-	4.271	2.299	-	-0.086	5.043	-
			-	-	-7.5	12.0	-	-7.5	48.0	-
			-	-	0.589	5.783	-	0.462	5.931	-
iCRCNLR	1	331.555	11	3	5.662	-0.018	-0.152	0.317	6.982	-
	2	43.274	2	Linear	6.198	2.481	-2.998	0.184	5.231	-
			11	11	6.264	0.203	-0.306	1.186	-14.613	11.548
	3	14.318	2	2	9.703	0.410	-0.571	5.980	-13.650	4.824
			7	7	1.895	10.655	-3.478	3.966	-4.616	8.251
			11	11	5.753	-1.742	18.385	7.042	0.096	-0.279

5.2.5 Conjunto de dados *Soccer*

O conjunto de dados *soccer* (NETO; CORDEIRO; CARVALHO, 2011) apresenta dados de peso Y e altura X de 531 jogadores de futebol do Campeonato Francês agrupados em 20 equipes. O agrupamento foi feito tomando os valores mínimo e máximo de cada variável por equipe. A Tabela 7 apresenta o centro, amplitude e intervalos do conjunto de dados para 3 *clusters* ajustados pelo algoritmo iCRCNLR.



(a) Centro (b) Amplitude (c) Intervalos

Figura 7 – Gráficos do centro, amplitude e intervalos com as curvas ajustadas pelo iCRCNLR, considerando três *clusters* para os dados *Soccer*

A redução no valor do critério de adequação J obtido pelo algoritmo iCRCNLR é

apresentado na Tabela 33. Para 1 e 3 *clusters* foram ajustados modelos lineares. As reduções no critério de adequação em relação ao algoritmo iCRCLR foram maiores para 1 e 3 *clusters*. Como esperado, o método *K-means* + regressão linear obtém reduções muito menores do que os métodos que aplicam regressão *clusterwise*.

Tabela 33 – Modelos selecionados, valor final do critério e estimativas de parâmetros sobre o conjunto de dados *Soccer* utilizando os métodos iCRCLR e iCRCNLR.

Método	*Clusters*	J	Modelos Selecionados		Estimativas dos parâmetros					
					Centro			Amplitude		
			Centro	Amplitude	α_1	α_2	α_3	α_1	α_2	α_3
K-means + regressão linear	1	116.808	-	-	-19.270	0.519	-	3.675	0.665	-
	2	100.427	-	-	-62.373	0.762	-	6.018	0.432	-
			-	-	18.338	0.308	-	-2.435	1.165	-
	3	95.056	-	-	-96.434	0.943	-	9.736	0.077	-
			-	-	14.333	0.333	-	12.983	-0.336	-
			-	-	-9.909	0.466	-	-11.285	1.785	-
iCRCLR	1	116.808	-	-	-19.270	0.519	-	3.675	0.665	-
	2	53.300	-	-	11.906	0.341	-	2.105	0.786	-
			-	-	-58.332	0.752	-	6.460	0.467	-
	3	23.513	-	-	-12.083	0.501	-	5.415	0.680	-
			-	-	55.694	0.106	-	0.734	0.825	-
			-	-	30.187	0.235	-	3.530	0.698	-
iCRCNLR	1	108.258	Linear	11	0.101	0.411	-	12.832	0.105	-0.007
	2	47.658	2	11	153.187	6478.557	-95.844	14.282	0.109	-0.007
			2	9	114.724	2918.823	-110.688	4.626	-0.088	5.687
	3	22.302	2	9	117.911	3051.230	-103.657	7.349	-0.055	-4.531
			Linear	9	30.187	0.235	-	5.821	-0.067	1.103
			Linear	Linear	55.694	0.106	-	0.734	0.825	-

5.2.6 Predição em conjuntos de dados reais

Esta seção tem como objetivo mensurar comparativamente a qualidade da predição do algoritmo proposto, iCRCNLR e sua versão linear iCRCLR. São executadas 10 validações cruzadas 10-*folds* para comparar a predição para a variável dependente de indivíduos não utilizados para ajustar os modelos e *clusters* - conjunto de teste. Um conjunto de funções não lineares é dada como entrada ao algoritmo iCRCNLR e o melhor par de funções para o centro e amplitude é escolhida de acordo com o critério de minimização J. É importante notar que, ao ajustar modelos mais flexíveis, incorre-se no risco do *overfit*, problema que consiste em ajustar bem os dados de treino, mas não detectar corretamente o modelo desconhecido a gerar os dados. Consequentemente, numa situação de *overfit* a predição para modelos não lineares tenderia a ser pior do que no caso linear.

Com objetivo de avaliar o desempenho do iCRCNLR na tarefa de predição em dados reais, foram computados o RMSE's da predição, comparando-o com o caso linear, iCRCLR, e com a abordagem *K-means* com regressão linear. Novamente, três métodos

de alocação de novas observações foram utilizados: KNN para dados intervalo, *Stacked regressions* e aleatório. A significância estatística da médias dos RMSE para comparação dos métodos *clusterwise* (iCRCLR e iCRCNLR) foi obtida por meio do teste não-paramétrico de Mann-Whitney. A Tabela 34 apresenta a média, desvio padrão e valor mínimo dos RMSE's nas 10 validações cruzadas para os métodos *K-means* com regressão linear, iCRCLR, iCRCNLR. Note-se que, para 1 *cluster*, os métodos *K-means* com regressão linear e iCRCLR são equivalente ao ajuste de um modelo de regressão para todo o conjunto de dados. Para evitar redundância, apenas o método iCRCLR foi computado neste caso.

No conjunto de dados *Cardio*, o melhor resultado foi obtido utilizando o método iCRCNLR e alocação KNN com 2 *clusters*. A diferença entre o melhor resultado do iCRCNLR e iCRCLR foi estatisticamente significativo, considerando $\alpha = 0.05$, conforme apresentado na Tabela 38. O método iCRCNLR apresentou resultados melhores para 2 e 3 *clusters* com os métodos de alocação KNN e *Stacked Regressions*. No entanto, as diferenças foram estatisticamente significativas, de acordo com o teste não-paramétrico de Mann-Whitney, apenas nas combinações em que o método de alocação é o *Stacked Regressions*. Considerando 1 *cluster*, os RMSE's médios dos métodos apresentaram valores muito próximos. Para a abordagem *K-means* com regressão linear, os resultados foram piores em todos os cenários, exceto para 2 *clusters* com alocação *Stacked Regressions*. Em relação aos valores mínimos obtidos, o método iCRCNLR apresentou RMSE menor relativo ao método iCRCLR, com exceção dos casos com alocação aleatória em 2 e 3 *clusters* e 1 *cluster*. No entanto, a abordagem *K-means* com regressão linear apresentou valores mínimos de RMSE mais baixos em todos os cenários com 2 *clusters* e utilizando alocação aleatória para 3 *clusters*. A abordagem com agrupamento *K-means* e regressão linear apresentou maior RMSE que o método iCRCNLR em todas as combinações de quantidade de *clusters* e métodos de alocação, mas foi competitivo em relação ao método iCRCLR. Para o conjunto de dados *Tree*, a menor média do RMSE para a predição foi obtida para 2 *clusters* utilizando o método de alocação *Stacked Regressions*. Os valores foram quase os mesmos, sem diferenças estatisticamente significativas entre os métodos. Para 1 *cluster*, o método linear apresentou melhor desempenho, incluindo menor média e valor mínimo para o RMSE. O método *K-means* + regressão linear apresentou RMSE médio maior do que o método iCRCNLR em todas as configurações. Ainda nesta abordagem, os valores mínimos obtidos para 2 *clusters* mostraram-se competitivos em relação aos demais métodos.

O conjunto de dados *Unemployment* apresentou diferenças estatisticamente significativas para 1 *cluster*, 2 *clusters* com alocação KNN e *Stacked Regressions* e 3 *clusters* com alocação KNN e aleatória. Dentre estes casos, o método iCRCNLR apresentou menor RMSE médio em 2 e 3 *clusters* com KNN. Em todos os casos, o valor mínimo do RMSE obtido pelo método iCRCNLR é menor do que o dado pelo caso linear. Em relação ao

Tabela 34 – RMSE médio, desvio padrão e mínimo da predição utilizando KNN, *Stacked Regressions* e alocação aleatória nos conjuntos de dados apresentados.

Dataset	Método	Medida	1 Cluster	2 Clusters			3 Clusters		
				KNN	SR	Random	KNN	SR	Random
Cardio	K-means +regressão linear	Média	-	1.930	2.541	2.064	3.605	4.004	3.726
		Desvio	-	0.763	0.831	1.235	1.584	1.032	1.235
		Min.	-	0.933	0.977	0.996	1.530	2.115	0.996
	iCRCLR	Média	1.516	1.512	2.788	1.873	1.605	2.418	2.06
		Desvio	0.026	0.068	0.197	0.16	0.096	0.292	0.139
		Min.	1.459	1.419	2.541	1.566	1.466	2.073	1.85
	iCRCNLR	Média	1.508	**1.469**	1.573	1.986	1.634	2.003	2.308
		Desvio	0.023	0.072	0.062	0.134	0.059	0.767	0.588
		Min.	1.48	1.347	1.482	1.723	1.509	1.594	1.804
Tree	K-means +regressão linear	Média	-	0.151	0.178	0.191	0.178	2.390	0.245
		Desvio	-	0.033	0.048	0.052	0.024	0.499	0.074
		Min.	-	0.081	0.093	0.132	0.141	1.681	0.122
	iCRCLR	Média	0.036	**0.032**	0.217	0.05	0.034	0.19	0.053
		Desvio	0.001	0.002	0.008	0.003	0.001	0.012	0.003
		Min.	0.034	0.03	0.209	0.045	0.031	0.172	0.049
	iCRCNLR	Média	0.097	0.098	0.103	0.135	0.096	0.708	0.147
		Desvio	0.002	0.007	0.006	0.012	0.009	1.538	0.011
		Min.	0.093	0.086	0.094	0.12	0.083	0.099	0.128
Unemployment	K-means +regressão linear	Média	-	18.571	31.935	32.447	18.762	23.486	25.644
		Desvio	-	0.752	2.666	3.839	4.687	4.950	4.485
		Min.	-	17.337	27.836	25.959	10.688	15.062	18.502
	iCRCLR	Média	15.567	14.111	24.348	22.519	14.017	21.938	22.358
		Desvio	0.207	0.519	1.772	1.718	0.656	0.759	0.775
		Min.	15.31	13.293	20.709	19.601	12.976	20.805	21.347
	iCRCNLR	Média	15.94	**13.032**	25.607	33.776	13.55	21.561	27.437
		Desvio	0.429	0.707	16.509	27.308	0.775	5.566	8.573
		Min.	15.519	11.864	18.08	21.878	12.552	17.834	19.204
Mushroom	K-means +regressão linear	Média	-	6.070	4.988	5.089	8.916	8.981	8.287
		Desvio	-	3.112	1.819	1.507	1.802	3.652	2.004
		Min.	-	3.221	2.449	2.264	5.429	4.831	5.551
	iCRCLR	Média	4.957	5.54	5.551	12.843	7.126	9.441	21.02
		Desvio	0.165	0.377	0.711	2.959	0.391	2.663	8.577
		Min.	4.569	4.797	4.498	10.055	6.361	5.632	8.746
	iCRCNLR	Média	4.674	**4.493**	9.942	20.362	5.341	102.46	18.833
		Desvio	0.257	0.351	2.073	7.352	0.915	220.924	23.066
		Min.	4.335	4.044	7.48	9.788	3.823	6.644	7.625
Soccer	K-means +regressão linear	Média	-	4.418	4.151	5.445	6.795	6.633	7.354
		Desvio	-	1.496	0.917	1.972	1.838	2.586	2.052
		Min.	-	2.484	2.514	2.983	4.316	2.821	4.691
	iCRCLR	Média	6.158	6.842	144.711	7.816	8.763	189.797	9.211
		Desvio	0.199	0.516	25.167	0.827	0.549	77.276	1.874
		Min.	5.875	6.182	110.001	6.334	8.032	98.489	7.627
	iCRCNLR	Média	3.581	3.841	**3.759**	4.3	4.454	4.279	5.248
		Desvio	0.06	0.183	0.228	0.441	0.525	0.647	0.417
		Min.	3.462	3.59	3.392	3.418	3.621	3.601	4.715

desvio padrão, o método iCRCNLR apresenta maiores valores do que o caso linear. Quando o método de alocação é o Aleatório, o método iCRCNLR perde desempenho. Isto pode ser justificado pelo fato de que, dado que as funções não lineares se ajustam melhor aos dados, uma alocação aleatória feita no *cluster* errado pode levar a uma erro maior do que no caso linear. Neste conjunto de dados, a abordagem de agrupamento com regressão apresentou RMSE médio mais elevado do que os métodos *clusterwise*. O método *K-means* + regressão linear apresentou maior RMSE médio neste conjunto de dados, com valores mínimos competitivos para 3 *clusters*.

No experimento para os dados *Mushroom*, o melhor resultado geral foi obtido pelo algoritmo iCRCNLR com 2 *clusters* e método de alocação KNN. A única combinação onde o teste de Mann-Whitney não apresentou diferença estatisticamente significativa é dada por 3 *clusters* e método de alocação *Stacked Regressions*. Considerando o ajuste para 1 *cluster* o método iCRCNLR apresentou menor RMSE médio e valor mínimo, porém, com desvio padrão maior. Em 2 *clusters*, o método iCRCNLR apresenta menor valor para o método de alocação KNN, com menor desvio padrão e valor mínimo. Para 3 *clusters*, o método iCRCNLR apresentou menor RMSE médio no método de alocação KNN, porém com desvio padrão maior. Neste conjunto de dados, apesar do menor RMSE médio ter sido obtido pelo método iCRCNLR, o método que utiliza agrupamento *K-means* e regressão linear apresentou melhores resultados nos métodos de alocação *Stacked Regressions* e aleatório.

No conjunto *soccer*, o algoritmo iCRCNLR foi consistentemente melhor em todos as configurações. O menor RMSE foi obtido com 2 *clusters* e método de alocação *Stacked Regressions*. O método iCRCLR apresentou RMSE médio maior do que a abordagem alternativa de agrupamento e regressão linear com a alocação aleatória e no método *Stacked Regressions*. Nota-se que os valores do RMSE no método iCRCLR com alocação *Stacked Regressions* foram elevados. Uma possível explicação para isto é que as estimativas de $\boldsymbol{\alpha}$ tenham sido afetadas pela presença de um *outlier* em pelo menos um caso nas 10 validações cruzadas. O método *K-means* + regressão linear obteve resultados melhores em relação ao método iCRCLR, mas piores em relação ao método não-linear iCRCNLR.

A Tabela 35 apresenta o MAPE obtido para um experimento de predição com validação cruzada 10-*times* 10-folds. Pela forma como é computada, a equação (5.7) tende ao infinito quando um valor observado na amostra A_t tende a zero. Esta situação ocorre no conjunto de dados *Unemployment*, em que há limites inferiores dos intervalos iguais a zero. Nota-se que os maiores erros percentuais de predição são obtidos no conjunto de dados *unemployment*. O método iCRCNLR apresenta o menor erro em três dos conjuntos de dados: *cardio*, *unemployment* e *soccer*. Em dois conjuntos de dados, *mushroom* e *soccer* o menor erro percentual é obtido quando o ajuste é feito sem agrupamento, ou seja, para todo o conjunto de dados. Como os valores do RMSE médio na Tabela 34

do ajuste sem agrupamento são competitivos, este resultado sugere que não há uma estrutura de grupos nestes dados.

Tabela 35 – MAPE da predição para KNN, *Stacked Regressions* e alocação aleatória nos conjuntos de dados apresentados.

Dataset	Método	1 Cluster	2 Clusters			3 Clusters		
			KNN	SR	Random	KNN	SR	Random
Cardio	K-means + regressão linear	-	23.271	23.996	28.978	35.892	38.919	43.814
	iCRCLR	10.566	12.548	20.911	13.512	13.839	15.888	14.591
	iCRCNLR	10.435	10.578	**10.275**	13.337	11.801	11.281	15.943
Tree	K-means + regressão linear	-	76.197	109.476	70.433	73.103	183.862	73.812
	iCRCLR	51.813	36.460	258.687	66.125	**35.899**	237.808	68.764
	iCRCNLR	112.256	56.755	38.345	112.159	64.178	70.385	941.611
Unemployment	K-means + regressão linear	-	Inf	Inf	Inf	Inf	Inf	Inf
	iCRCLR	-	Inf	Inf	Inf	Inf	Inf	Inf
	iCRCNLR	-	Inf	Inf	Inf	Inf	Inf	Inf
Mushroom	K-means + regressão linear	-	67.641	113.910	47.493	53.839	111.048	97.836
	iCRCLR	**23.618**	24.728	140.162	77.266	29.434	153.748	37.638
	iCRCNLR	31.236	30.807	84.188	136.242	36.616	52.557	57.567
Soccer	K-means + regressão linear	-	4.519	20.424	8.270	9.156	32.143	8.982
	iCRCLR	2.781	3.061	60.452	4.524	3.596	51.499	4.111
	iCRCNLR	**2.597**	2.689	3.281	3.535	3.455	3.591	4.683

As Tabelas 36 a 37 apresentam o ranking das combinações de métodos de ajuste e alocação nos seis conjuntos de dados apresentados. Em relação ao RMSE médio, para 1 *cluster*, o método iCRCNLR apresenta melhor desempenho em relação ao iCRCLR. Em 2 *clusters* o algoritmo iCRCNLR com alocação KNN apresenta menor *rank* médio, seguido pelo iCRCLR, também com alocação KNN. A mesma situação ocorre com 3 *clusters*. A abordagem alternativa apresentou *rank* médio mais alto entre os três métodos.

Em relação ao valor mínimo de RMSE obtido no experimento, dentre os métodos *clusterwise*, o método iCRCNLR apresenta o menor *rank* médio para 2 e 3 *clusters* com alocação KNN. Para 1 *cluster*, ocorre empate no *rank* médio. O método *Stacked Regressions* apresentou *rank* médio pior que a alocação aleatória no método iCRCLR com 2 *clusters* e iCRCNLR com 3 *clusters*. Considerando 1 *cluster*, novamente o iCRCNLR apresentou *rank* médio menor do que o caso linear. No entanto, de modo geral, a abordagem alternativa *K-means* + regressão linear apresentou melhores resultados, em relação ao valor mínimo, para 2 *clusters*.

Os resultados da aplicação dos métodos iCRCLR e iCRCNLR aos conjuntos de dados reais mostraram que (i) o método iCRCNLR é competitivo em termos de predição, em relação ao método linear; (ii) os métodos de alocação de uma nova observação são essenciais para o desempenho da predição, neste caso, o método KNN apresentou melhores resultados; (iii) em relação ao valor mínimo obtido nas 10 repetições da validação cruzada, o método iCRCNLR apresentou menor RMSE mínimo que o caso linear.

Tabela 36 – Desempenho dos métodos baseado no *rank* da média.

Clusters	Método	Alocação	Dataset					Rank médio
			Cardio	Tree	Unemployment	Mushroom	Soccer	
1	iCRCLR	-	2	1	1	2	2	1.6
	iCRCNLR	-	1	2	2	1	1	1.4
2	K-means + regressão linear	KNN	5	6	3	6	5	5.0
		SR	8	7	7	3	3	5.6
		Random	7	8	8	1	6	6.0
	iCRCLR	KNN	2	1	2	4	7	3.2
		SR	9	9	5	5	9	7.4
		Random	4	2	4	8	8	5.2
	iCRCNLR	KNN	1	3	1	2	2	1.8
		SR	3	4	6	7	1	4.2
		Random	6	5	9	9	4	6.6
3	K-means + regressão linear	KNN	7	5	3	4	5	4.8
		SR	9	9	7	5	4	6.8
		Random	8	7	8	3	6	6.4
	iCRCLR	KNN	1	1	2	2	7	2.6
		SR	6	6	5	6	9	6.4
		Random	4	2	6	8	8	5.6
	iCRCNLR	KNN	2	3	1	1	2	1.8
		SR	3	8	4	9	1	5.0
		Random	5	4	9	7	3	5.6

Tabela 37 – Desempenho dos métodos baseado no *rank* do valor minimo.

Clusters	Método	Alocação	Dataset					Rank médio
			Cardio	Tree	Unemployment	Mushroom	Soccer	
1	iCRCLR	-	1	1	1	2	2	1.4
	iCRCNLR	-	2	2	2	1	1	1.6
2	K-means + regressão linear	KNN	1	3	3	3	1	2.2
		SR	2	5	9	2	2	4.0
		Random	3	8	8	1	3	4.6
	iCRCLR	KNN	5	1	2	6	7	4.2
		SR	9	9	6	5	9	7.6
		Random	7	2	5	9	8	6.2
	iCRCNLR	KNN	4	4	1	4	6	3.8
		SR	6	6	4	7	4	5.4
		Random	8	7	7	8	5	7.0
3	K-means + regressão linear	KNN	4	7	1	3	4	3.8
		SR	9	9	4	2	1	5.0
		Random	1	5	6	4	5	4.2
	iCRCLR	KNN	2	1	3	6	8	4.0
		SR	8	8	8	5	9	7.6
		Random	7	2	9	9	7	6.8
	iCRCNLR	KNN	3	3	2	1	3	2.4
		SR	5	4	5	7	2	4.6
		Random	6	6	7	8	6	6.6

Tabela 38 – P-valores do teste de Mann-Whitney para os métodos iCRCLR e iCRCNLR.

Dataset	1 Cluster	2 Clusters			3 Clusters		
Cardio	0.578	0.165	**0.000**	0.123	0.315	**0.018**	0.247
Tree	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Unemployment	**0.035**	**0.000**	**0.043**	0.165	0.190	**0.035**	0.123
Mushroom	**0.008**	**0.000**	**0.000**	**0.014**	**0.000**	0.853	**0.023**
Soccer	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**

Em muitos dos cenários, o RMSE médio associado ao método iCRCNLR foi menor, no entanto, em relação ao desvio padrão, o caso linear é mais estável. Tal estabilidade decorre do fato de que a estimação de parâmetros no caso linear é fechada, enquanto que no caso do iCRCNLR, a estimação depende de heurísticas de otimização, que nem sempre levam aos melhores resultados. Tal fato pode gerar *outliers* nas amostras de RMSE, fazendo com que o desvio padrão seja maior. O método de *Stacked Regressions* por realizar a predição por meio de uma combinação linear de predições dos modelos ajustados em *K clusters* teve um desempenho abaixo do esperado, em muitos casos resultando num RMSE médio maior do que aquele fornecido pela alocação aleatória.

6 CONCLUSÕES E TRABALHOS FUTUROS

Este trabalho propõe um novo método de regressão *clusterwise* para dados intervalo, iCRCNLR, capaz de ajustar modelos não lineares dentro de grupos homogêneos de observações. O objetivo é fornecer o melhor par de funções ajustadas para o centro e amplitude dos dados, de acordo com um critério de otimização. O método representa uma expansão do caso linear para regressão *clusterwise* de dados intervalo, sendo, ao final, uma combinação do algoritmo de agrupamento dinâmico (DIDAY; SIMON, 1980) com o método de regressão não linear para dados intervalo (NETO; CARVALHO, 2017).

Estudos de simulação foram feitos para testar dois aspectos do método proposto: estimação de parâmetros e capacidade de predição. Foram gerados 24 cenários com diferentes estuturas de grupos no centro e na amplitude dos intervalos. Tais grupos poderiam ser lineares ou não-lineares, no que diz respeito à estrutura interna, ou apresentar sobreposição parcial ou completa em relação à varíavel independente.

A importância do estudo de simulação para estimação reside no fato de que o ajuste dos modelos no método iCRCNLR é feita por meio de heurísticas de otimização, como BFGS, *Simulated Annealing* e Gradiente Conjugado. Estes métodos, por uma série de razões, podem convergir para ótimos locais ou mesmo não convergir, levando a estimativas incorretas e distorcendo a capacidade de predição do método. Os resultados mostraram que o método apresenta bom comportamento no que diz respeito à estimação, uma vez que permite a substituição de uma heurística que não converge por outra. Além disso, o método é executado 100 vezes e o melhor resultado, de acordo com o critério J, é considerado. Dado que a qualidade das estimativas de parâmetros possui influência no primeiro passo do modelo, afetando-o de modo geral, escolher o melhor resultado significa descartar estimativas ruins.

O estudo de simulação para medir a capacidade de predição tem como objetivo comparar o desempenho do método iCRCNLR com o caso linear, iCRCLR. O problema de alocação de uma nova observação foi tratado por meio de três métodos, KNN, *Stacked Regressions* e alocação aleatória. Além disso, os métodos foram aplicados no ajuste de 1, 2 ou 3 *clusters*. Os resultados mostraram que o método iCRCNLR apresentou RMSE menor ou próximo aos obtidos pelo método linear para muitos dos cenários e métodos de alocação, exceto quando algum erro de alocação ocorre, levando a ocorrência de um *outlier* dentre os RMSE obtidos.

Finalmente, o método foi aplicado a seis conjuntos de dados reais e o desempenho da predição foi novamente comparado com o método iCRCLR. Para as aplicações em dados reais, também foi implementado um método simples que consiste em aplicar agrupamento *K-means* para dados tipo-intervalo e regressão linear.

Os resultados mostram que, nestes conjuntos, o iCRCNLR apresentou melhores

resultados de predição.

O desempenho dos métodos estudados na predição, em dados sintéticos e reais, é fortemente influenciado pelo método de alocação empregado para novas observações. A adoção de um método de alocação, como o KNN, ou combinação de predições pelos diferentes *clusters* ajustados, como o *Stacked Regressions*, é uma prática que apresenta vantagens em relação à simples alocação aleatória de novas observações. De modo geral, pode-se afirmar que o método de alocação KNN para dados intervalo com distância de Hausdorff apresentou melhor desempenho na predição. No entanto, a diferença entre os métodos diminui quando os cenários gerados apresentam grupos sobrepostos. Em alguns casos, o desempenho da alocação por meio de *Stacked Regressions* foi pior do que a alocação aleatória. Apesar desta questão necessitar de um estudo mais aprofundado, uma das possíveis causas reside no fato de que o método *Stacked Regressions* obtém os pesos da combinação linear das predições nos modelos ajustados por meio de regressão *linear*. No entanto, nada garante que haja uma forte correlação linear entre as predições e os valores da variável dependente utilizados no ajuste.

Em suma, pode-se afirmar que o ajuste de modelos não-lineares em regressão *clusterwise* para dados intervalo pode significar uma melhora em termos de predição, quando comparado ao ajuste apenas de métodos lineares. Obviamente, tal prática deve ser vista sempre com cautela, pois a escolha dos modelo não lineares dependerá, em parte, de conhecimento prévio do especialista sobre o domínio do problema. Além disso, a possibilidade de ocorrer sobreajuste dos dados é maior ao serem utilizados modelos não lineares. A flexibilidade dos modelos escolhidos pode fazer com que o método seja mais sensível a ruídos do que o caso linear. Recomenda-se, portanto, que os ajustes devam ser sempre acompanhados de algum estudo de validação cruzada, com o objetivo de prevenir o sobreajuste dos dados.

Como trabalhos futuros, podem ser desenvolvidas variações do método a partir de outros modelos de regressão para dados intervalo e variações do método de agrupamento dinâmico. Além disso, podem ser feitos testes do método em dados multivariados, o que depende de uma cuidadosa escolha dos modelos e um aumento no custo computacional em função da utilização das heurísticas de otimização para cenários multivariados. Outro campo a ser investigado refere-se ao aprimoramento dos métodos de alocação, em especial, a avaliação do uso de outras distâncias (Euclidiana, *City-Block*) para dados tipo intervalo no método KNN, ou ainda, a utilização de outros modelos de regressão para ajustar os pesos na predição do *Stacked Regressions* - regressão não-linear, robusta, entre outras.

REFERÊNCIAS

AHN, J.; PENG, M.; PARK, C.; JEON, Y. A resampling approach for interval-valued data regression. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, Wiley Online Library, v. 5, n. 4, p. 336–348, 2012.

AURIFEILLE, J.-M.; QUESTER, P. G. Predicting business ethical tolerance in international markets: a concomitant clusterwise regression analysis. *International Business Review*, Elsevier, v. 12, n. 2, p. 253–272, 2003.

BILLARD, L. Symbolic data analysis: what is it? In: *Compstat 2006-Proceedings in Computational Statistics*. [S.l.]: Springer, 2006. p. 261–269.

BILLARD, L.; DIDAY, E. Regression analysis for interval-valued data. In: *Data Analysis, Classification and Related Methods: Proceedings of the Seventh Conference of the International Federation of Classification Societies*. [S.l.: s.n.], 2000. p. 369–374.

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

BILLARD, L.; DIDAY, E. From the statistics of data to the statistics of knowledge: Symbolic data analysis. *J. Amer. Statist. Assoc.*, v. 98, n. 462, p. 470–487, 2003.

BLANCO-FERNÁNDEZ, A.; CORRAL, N.; GONZÁLEZ-RODRÍGUEZ, G. Estimation of a flexible simple linear model for interval data based on set arithmetic. *Computational Statistics & Data Analysis*, Elsevier, v. 55, n. 9, p. 2568–2578, 2011.

BOCK, H. The equivalence of two extremal problems and its application to the iterative classification of multivariate data. *Mathematisches Forschungsinstitut*, 1969.

BREIMAN, L. Stacked regressions. *Machine learning*, Springer, v. 24, n. 1, p. 49–64, 1996.

BRITO, P.; SILVA, A. P. D. Modelling interval data with normal and skew-normal distributions. *Journal of Applied Statistics*, Taylor & Francis, v. 39, n. 1, p. 3–20, 2012.

BROYDEN, C. G. The convergence of a class of double-rank minimization algorithms: 2. the new algorithm. *IMA journal of applied mathematics*, Oxford University Press, v. 6, n. 3, p. 222–231, 1970.

BRUSCO, M. J.; CRADIT, J. D.; STEINLEY, D.; FOX, G. L. Cautionary remarks on the use of clusterwise regression. *Multivariate Behavioral Research*, Taylor & Francis, v. 43, n. 1, p. 29–49, 2008.

BRUSCO, M. J.; CRADIT, J. D.; TASHCHIAN, A. Multicriterion clusterwise regression for joint segmentation settings: An application to customer value. *Journal of Marketing Research*, American Marketing Association, v. 40, n. 2, p. 225–234, 2003.

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

CARVALHO, F. d. A. de; BRITO, P.; BOCK, H.-H. Dynamic clustering for interval data based on l 2 distance. *Computational Statistics*, Springer, v. 21, n. 2, p. 231–250, 2006.

CARVALHO, F. d. A. de; SAPORTA, G.; QUEIROZ, D. N. A clusterwise center and range regression model for interval-valued data. In: *Proceedings of COMPSTAT'2010*. [S.l.]: Springer, 2010. p. 461–468.

CHAVENT, M. A hausdorff distance between hyper-rectangles for clustering interval data. In: *Classification, clustering, and data mining applications*. [S.l.]: Springer, 2004. p. 333–339.

CHAVENT, M.; CARVALHO, F. d. A. de; LECHEVALLIER, Y.; VERDE, R. New clustering methods for interval data. *Computational statistics*, Springer, v. 21, n. 2, p. 211–229, 2006.

CHAVENT, M.; LECHEVALLIER, Y. Dynamical clustering of interval data: Optimization of an adequacy criterion based on hausdorff distance. In: *Classification, clustering, and data analysis*. [S.l.]: Springer, 2002. p. 53–60.

CHIRICO, P. A clusterwise regression method for the prediction of the disposal income in municipalities. In: *Classification and Data Mining*. [S.l.]: Springer, 2013. p. 173–180.

COVER, T.; HART, P. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theor.*, IEEE Press, Piscataway, NJ, USA, v. 13, n. 1, p. 21–27, set. 2006. ISSN 0018-9448. Disponível em: <https://doi.org/10.1109/TIT.1967.1053964>.

DESARBO, W. S.; CRON, W. L. A maximum likelihood methodology for clusterwise linear regression. *Journal of classification*, Springer, v. 5, n. 2, p. 249–282, 1988.

DESARBO, W. S.; EDWARDS, E. A. Typologies of compulsive buying behavior: A constrained clusterwise regression approach. *Journal of consumer psychology*, Wiley Online Library, v. 5, n. 3, p. 231–262, 1996.

DESARBO, W. S.; OLIVER, R. L.; RANGASWAMY, A. A simulated annealing methodology for clusterwise linear regression. *Psychometrika*, Springer, v. 54, n. 4, p. 707–736, 1989.

DIAS, S.; BRITO, P. Off the beaten track: A new linear model for interval data. *European Journal of Operational Research*, v. 258 (3), p. 1118–1130, 2017.

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

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

DOTTO, F.; FARCOMENI, A.; GARCÍA-ESCUDERO, L. A.; ISCAR, A. M. A fuzzy approach to robust clusterwise regression. 2016.

DRAPER, N.; SMITH, H. *Applied regression analysis*. New York [u.a.]: Wiley, 1966. (Wiley series in probability and mathematical statistics). ISBN 0471221708. Disponível em: <http://gso.gbv.de/DB=2.1/CMD?ACT=SRCHA&SRT=YOP&IKT=1016&TRM=ppn+022791892&sourceid=fbw_bibsonomy>.

D'URSO, P.; MASSARI, R.; SANTORO, A. A class of fuzzy clusterwise regression models. *Information Sciences*, Elsevier, v. 180, n. 24, p. 4737–4762, 2010.

D'URSO, P.; SANTORO, A. Fuzzy clusterwise linear regression analysis with symmetrical fuzzy output variable. *Computational statistics & data analysis*, Elsevier, v. 51, n. 1, p. 287–313, 2006.

FAGUNDES, R. A. A.; SOUZA, R. M. C. R. de; CYSNEIROS, F. J. A. Robust regression with application to symbolic interval data. *Engineering Applications of Artificial Intelligence*, v. 26, p. 563–573, 2013.

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

FILHO, L. M. d. A. L.; SILVA, J. A. A. d.; CORDEIRO, G. M.; FERREIRA, R. L. C. Modeling the growth of eucalyptus clones using the chapman-richards model with different symmetrical error distributions. *Ciência Florestal*, SciELO Brasil, v. 22, n. 4, p. 777–785, 2012.

FIX E., H. J. Discriminatory analysis—nonparametric discrimination: Consistency properties. *Technical Report 4, Project no. 21-29-004*, USAF School of Aviation Medicine, Texas, 1951.

FLETCHER, R. A new approach to variable metric algorithms. *The computer journal*, Oxford University Press, v. 13, n. 3, p. 317–322, 1970.

GIORDANI, P. Lasso-constrained regression analysis for interval-valued data. *Advances in Data Analysis and Classification*, v. 9, n. 1, p. 5–19, 2015.

GOLDFARB, D. A family of variable-metric methods derived by variational means. *Mathematics of computation*, v. 24, n. 109, p. 23–26, 1970.

GONZÁLEZ-RIVERA, G.; LIN, W. Constrained regression for interval-valued data. *Journal of Business & Economic Statistics*, v. 31, n. 4, p. 473–490, 2013.

GOWDA, K. C.; DIDAY, E. Symbolic clustering using a new dissimilarity measure. *Pattern Recognition*, v. 24, n. 6, p. 567 – 578, 1991. ISSN 0031-3203. Disponível em: <http://www.sciencedirect.com/science/article/pii/003132039190022W>.

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

HAND, D. J. Discrimination and classification. *Wiley Series in Probability and Mathematical Statistics, Chichester: Wiley, 1981*, 1981.

HAO, P.; GUO, J. Constrained center and range joint model for interval-valued symbolic data regression. *Computational Statistics and Data Analysis*, v. 116, p. 106–138, 2017.

HENNIG, C. Identifiablity of models for clusterwise linear regression. *Journal of Classification*, Springer, v. 17, n. 2, p. 273–296, 2000.

HESTENES, M. R.; STIEFEL, E. *Methods of conjugate gradients for solving linear systems*. [S.l.]: NBS Washington, DC, 1952. v. 49.

JEON, Y.; AHN, J.; PARK, C. A nonparametric kernel approach to interval-valued data analysis. *Technometrics*, Taylor & Francis, v. 57, n. 4, p. 566–575, 2015.

LAU, K.-n.; LEUNG, P.-l.; TSE, K.-k. A mathematical programming approach to clusterwise regression model and its extensions. *European Journal of Operational Research*, Elsevier, v. 116, n. 3, p. 640–652, 1999.

LIM, C. Interval-valued data regression using nonparametric additive models. *Journal of the Korean Statistical Society*, v. 45, n. 3, p. 358–370, 2017.

LUO, Z.; CHOU, E. Pavement condition prediction using clusterwise regression. *Transportation Research Record: Journal of the Transportation Research Board*, Transportation Research Board of the National Academies, n. 1974, p. 70–77, 2006.

MACQUEEN, J. et al. Some methods for classification and analysis of multivariate observations. In: OAKLAND, CA, USA. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. [S.l.], 1967. v. 1, n. 14, p. 281–297.

MANWANI, N.; SASTRY, P. K-plane regression. *Information Sciences*, Elsevier, v. 292, p. 39–56, 2015.

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

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

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

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

NETO, E. d. A. L.; CARVALHO, F. d. A. de. Nonlinear regression applied to interval-valued data. *Pattern Analysis and Applications*, Springer, v. 20, n. 3, p. 809–824, 2017.

NETO, E. d. A. L.; CARVALHO, F. d. A. de; NETO, J. F. C. Constrained linear regression models for interval-valued data with dependence. In: IEEE. *Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on*. [S.l.], 2007. p. 456–461.

SEARLE, S. R.; GRUBER, M. H. *Linear models*. [S.l.]: John Wiley & Sons, 2016.

SHANNO, D. F. Conditioning of quasi-newton methods for function minimization. *Mathematics of computation*, v. 24, n. 111, p. 647–656, 1970.

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

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

SPÄTH, H. Algorithm 39 clusterwise linear regression. *Computing*, Springer, v. 22, n. 4, p. 367–373, 1979.

SPÄTH, H. A fast algorithm for clusterwise linear regression. *Computing*, Springer, v. 29, n. 2, p. 175–181, 1982.

TAN, T.; SUK, H. W.; HWANG, H.; LIM, J. Functional fuzzy clusterwise regression analysis. *Advances in Data Analysis and Classification*, Springer, v. 7, n. 1, p. 57–82, 2013.

VEAUX, R. D. D. Mixtures of linear regressions. *Computational Statistics & Data Analysis*, Elsevier, v. 8, n. 3, p. 227–245, 1989.

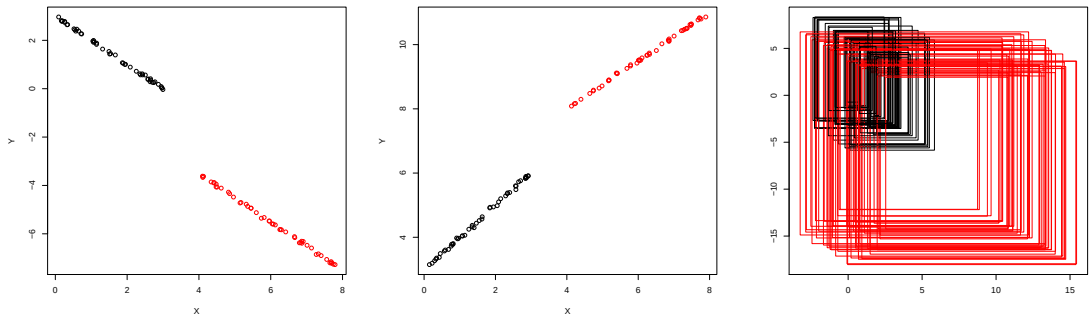
WEDEL, M. *Clusterwise regression and market segmentation: developments and applications*. Tese (Doutorado) — Wedel, 1990.

WEDEL, M.; KISTEMAKER, C. Consumer benefit segmentation using clusterwise linear regression. *International Journal of Research in Marketing*, Elsevier, v. 6, n. 1, p. 45–59, 1989.

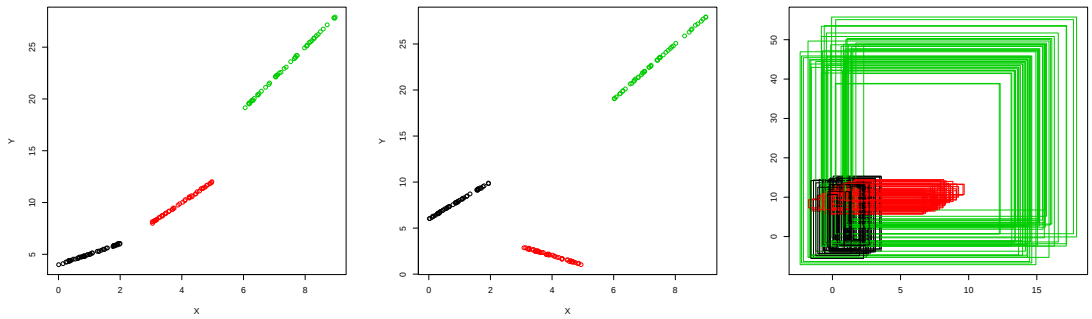
WOLPERT, D. H. Stacked generalization. *Neural networks*, Elsevier, v. 5, n. 2, p. 241–259, 1992.

YANG, M.-S.; KO, C.-H. On cluster-wise fuzzy regression analysis. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, IEEE, v. 27, n. 1, p. 1–13, 1997.

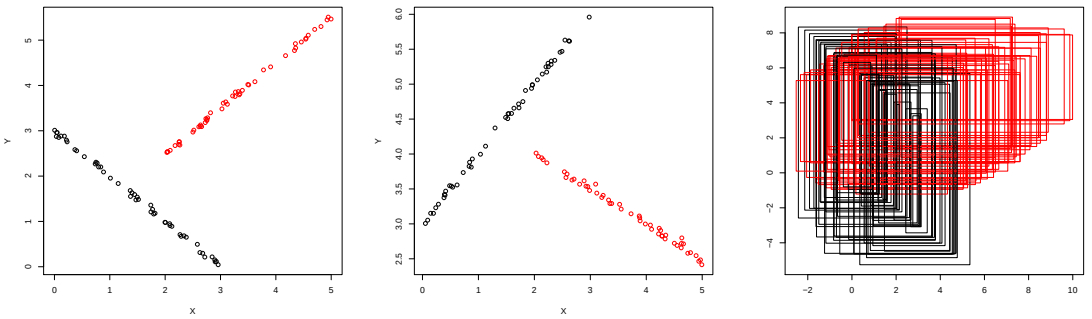
APÊNDICE A – GRÁFICOS DOS CENÁRIOS GERADOS



(a) Centro (b) Amplitude (c) Intervalos

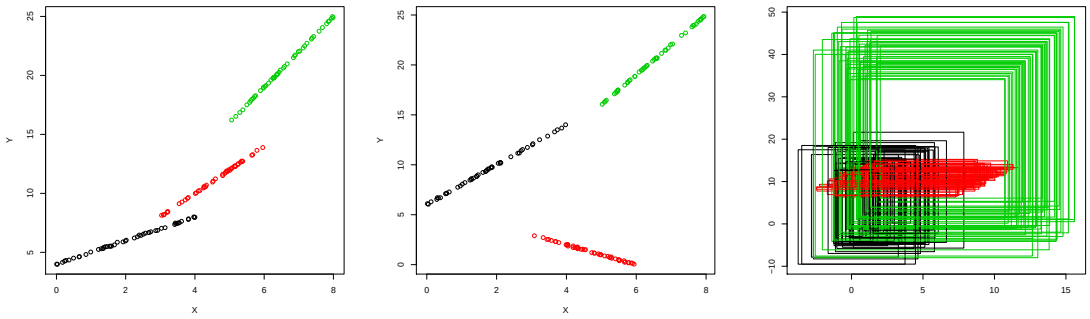


(d) Centro (e) Amplitude (f) Intervalos

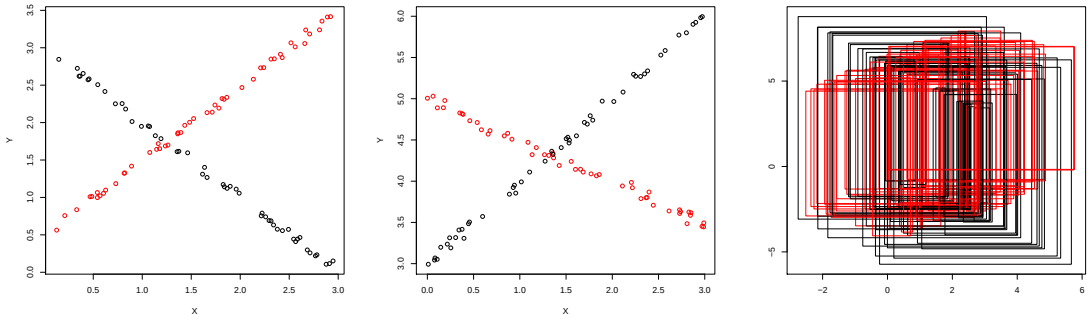


(g) Centro (h) Amplitude (i) Intervalos

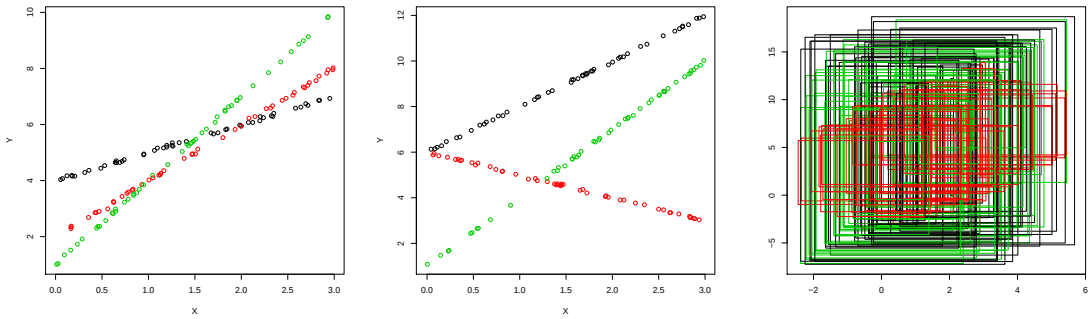
Figura 8 – Exemplos dos cenários 1, 2 e 3.



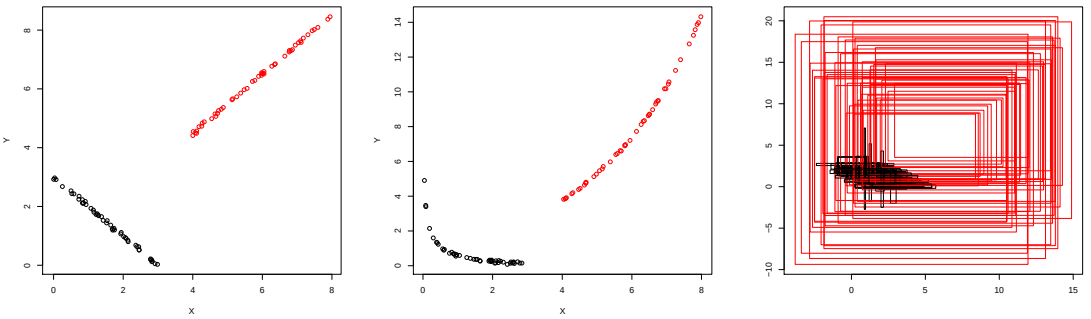
(a) Centro (b) Amplitude (c) Intervalos



(d) Centro (e) Amplitude (f) Intervalos

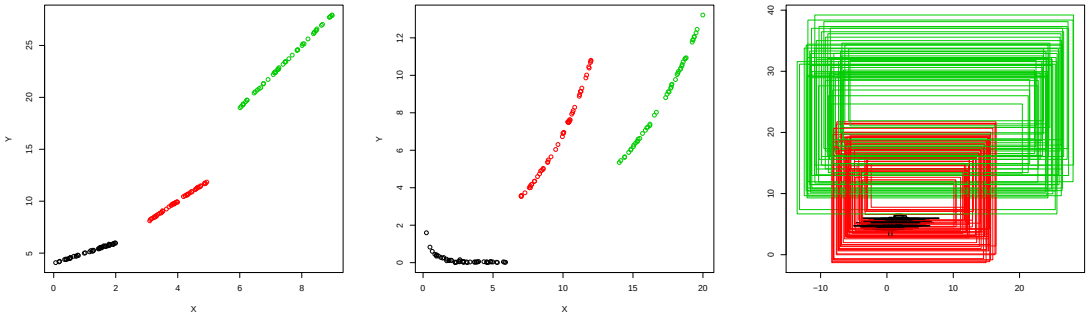


(g) Centro (h) Amplitude (i) Intervalos

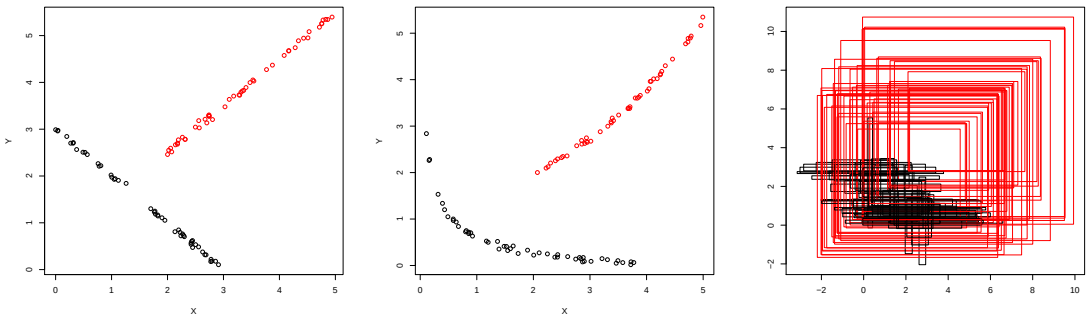


(j) Centro (k) Amplitude (l) Intervalos

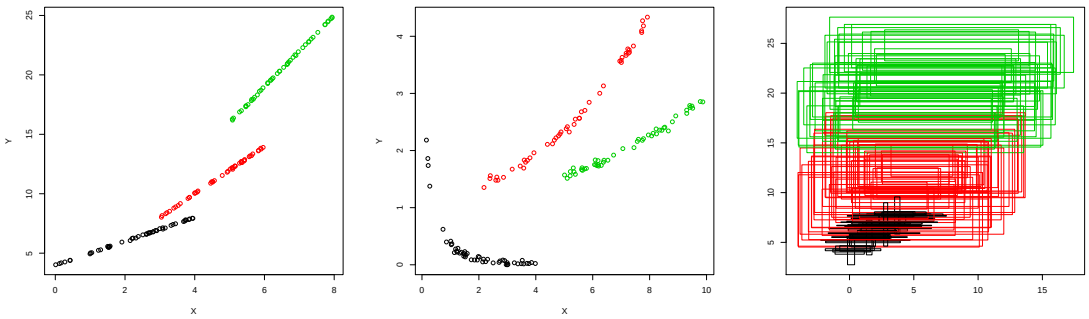
Figura 9 – Exemplos dos cenários 4, 5, 6 e 7.



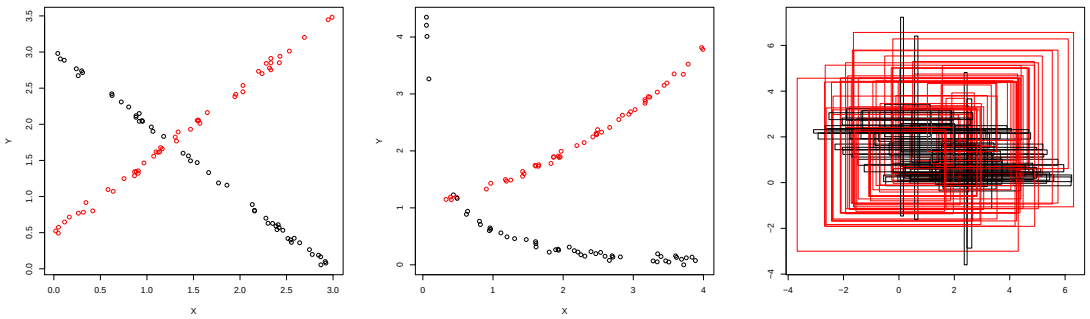
(a) Centro (b) Amplitude (c) Intervalos



(d) Centro (e) Amplitude (f) Intervalos

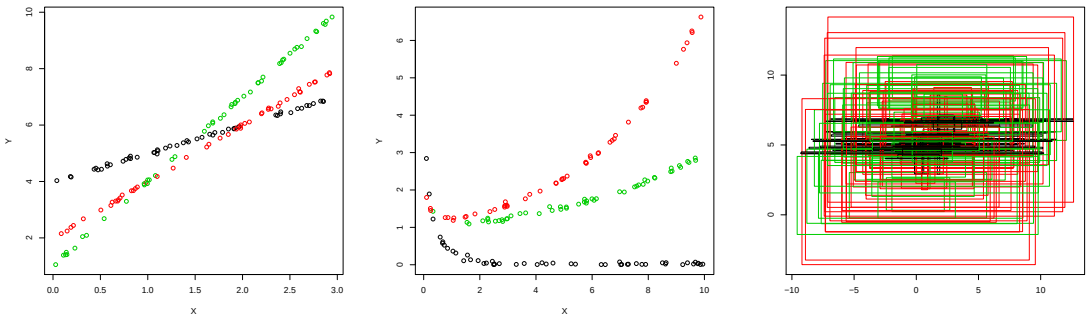


(g) Centro (h) Amplitude (i) Intervalos

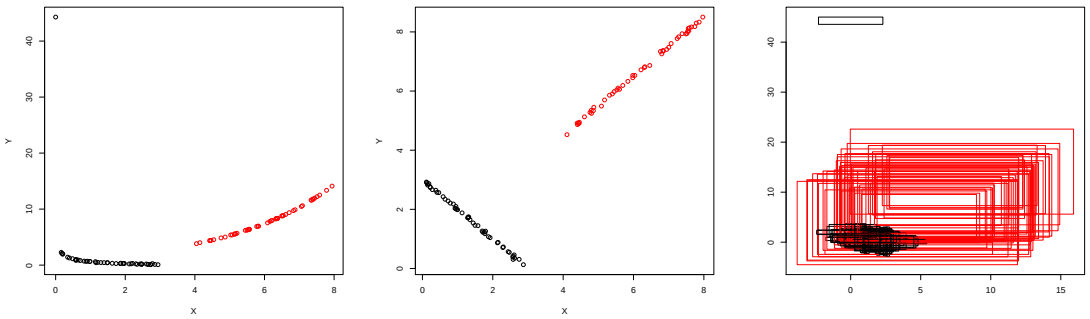


(j) Centro (k) Amplitude (l) Intervalos

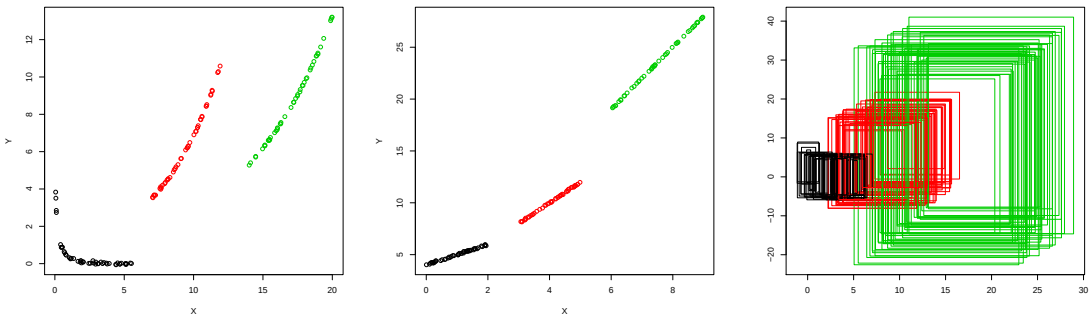
Figura 10 – Exemplos dos cenários 8, 9, 10 e 11.



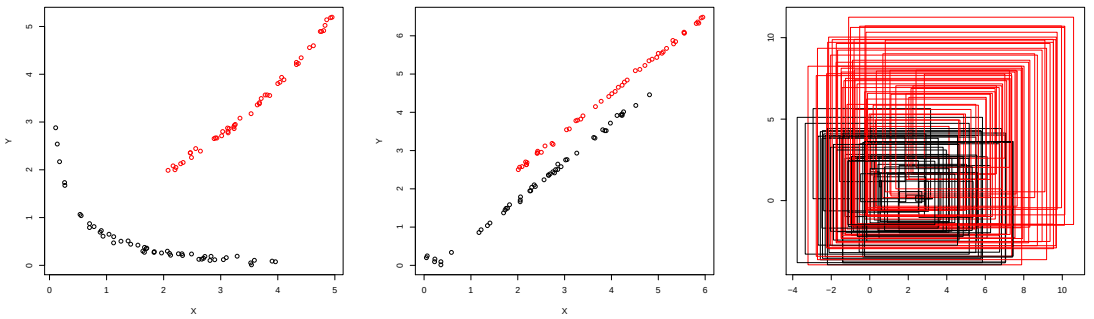
(a) Centro (b) Amplitude (c) Intervalos



(d) Centro (e) Amplitude (f) Intervalos

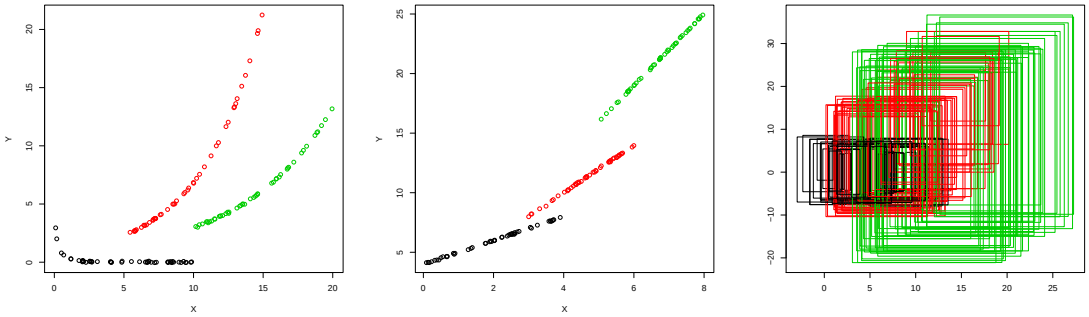


(g) Centro (h) Amplitude (i) Intervalos

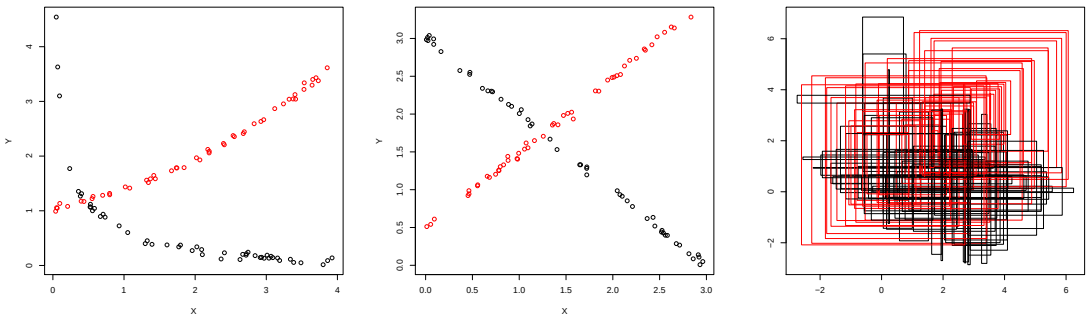


(j) Centro (k) Amplitude (l) Intervalos

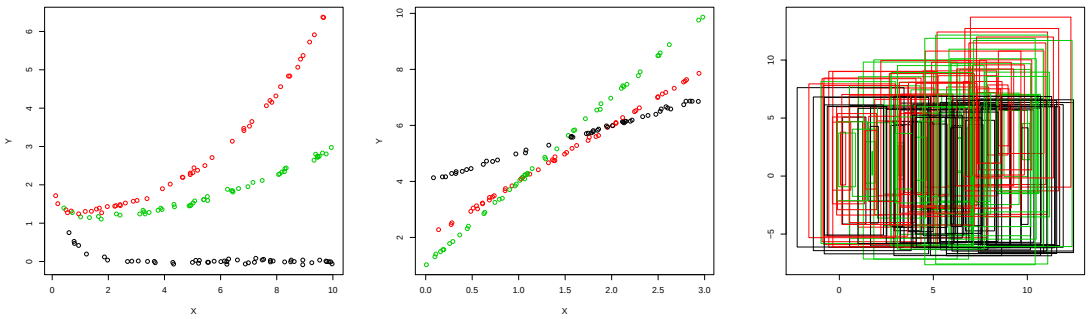
Figura 11 – Exemplos dos cenários 12, 13, 14 e 15.



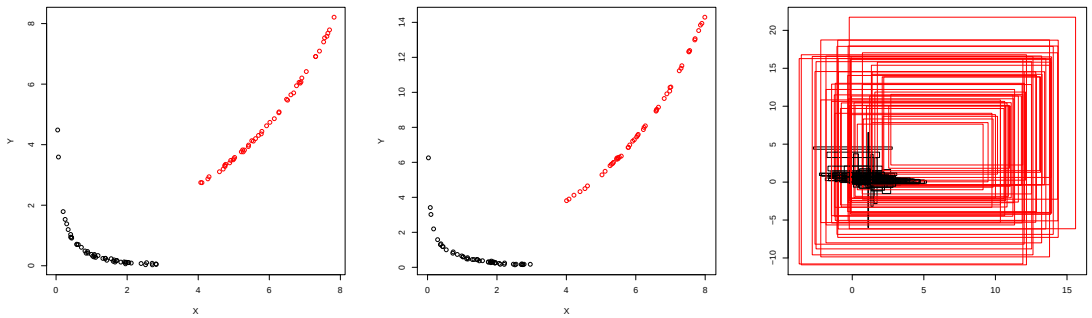
(a) Centro (b) Amplitude (c) Intervalos



(d) Centro (e) Amplitude (f) Intervalos

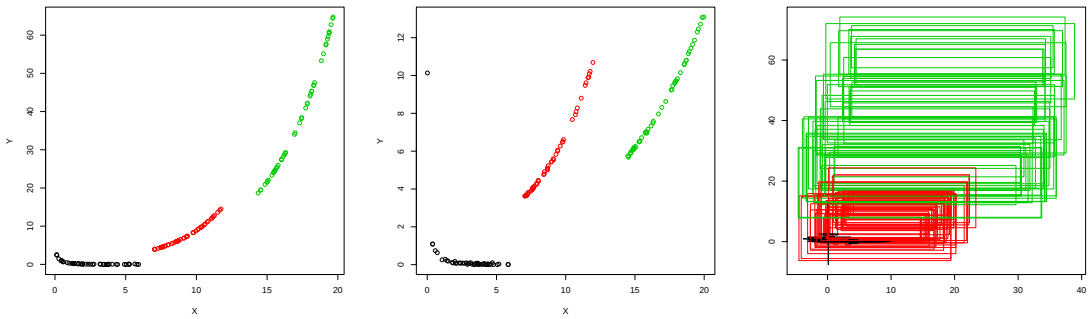


(g) Centro (h) Amplitude (i) Intervalos

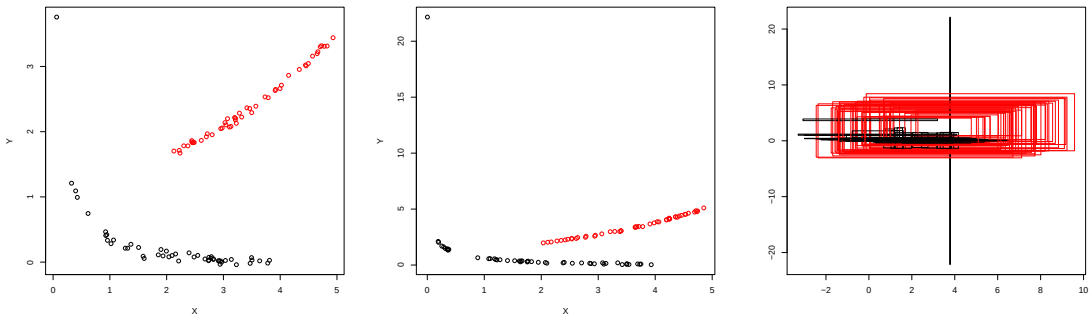


(j) Centro (k) Amplitude (l) Intervalos

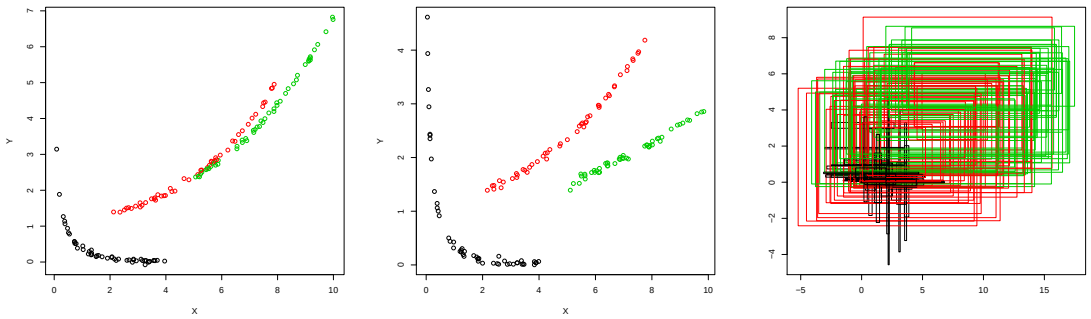
Figura 12 – Exemplos dos cenários 16, 17, 18 e 19.



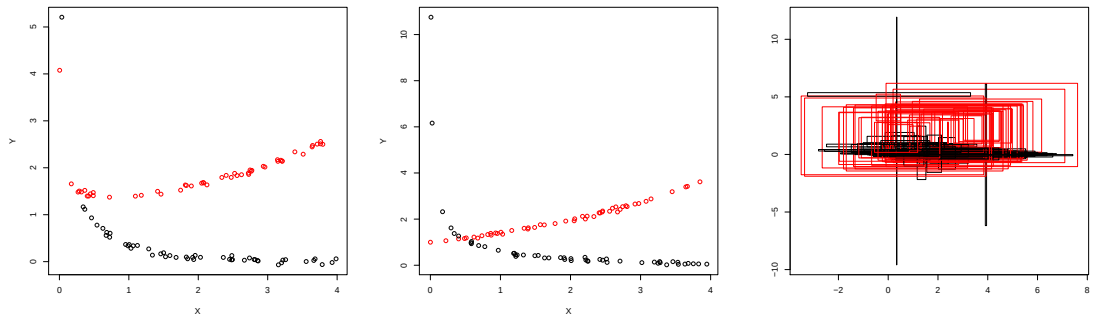
(a) Centro (b) Amplitude (c) Intervalos



(d) Centro (e) Amplitude (f) Intervalos

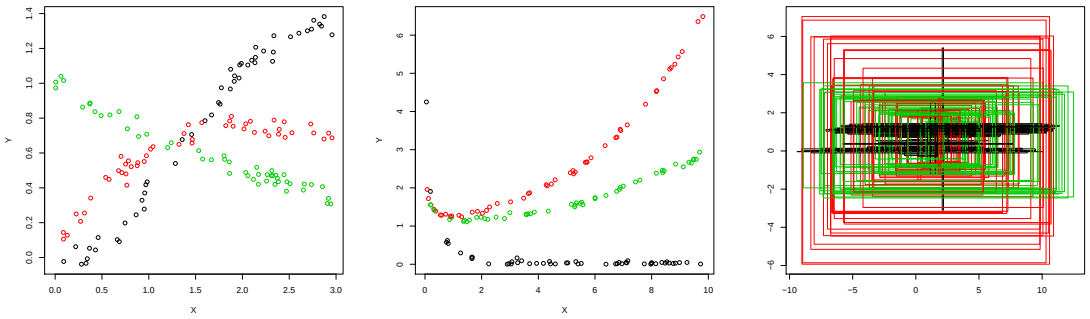


(g) Centro (h) Amplitude (i) Intervalos



(j) Centro (k) Amplitude (l) Intervalos

Figura 13 – Exemplos dos cenários 20, 21, 22 e 23.



(a) Centro (b) Amplitude (c) Intervalos