JADER DA SILVA JALE

UM ALGORITMO SIMPLES PARA AGRUPAMENTO DE DADOS

Recife – PE – Fevereiro/2011

UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOMETRIA E ESTATÍSTICA APLICADA

UM ALGORITMO SIMPLES PARA AGRUPAMENTO DE DADOS

Dissertação apresentada ao Programa de Pós-Graduação em Biometria e Estatística Aplicada como exigência parcial à obtenção do título de mestre.

Área de Concentração: Desenvolvimento de Métodos Estatísticos e Computacionais

Orientador: Prof. Adauto José Ferreira de Souza

Recife – PE – Fevereiro/2011

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Agrupamento de Dados

Uma das mais básicas e essenciais habilidades dos seres vivos é o agrupamento de objetos similares produzindo uma classificação [34].

A tarefa de agrupamento de dados envolve a organização de um conjunto de objetos em grupos (categorias ou *clusters*) formados com base em suas similaridades [29].

A Análise Multivariada é uma área da estatística que consiste em um conjunto de técnicas que podem ser usadas em situações onde cada objeto de um conjunto de dados possui várias dimensões (atributos), onde cada dimensão (atributo) representa uma variável [25].

Análise de Agrupamento é uma dessas técnicas de Análise Multivariada, e consiste no processo de agrupar um conjunto de objetos físicos ou abstratos em grupos de objetos similares. Um grupo é uma coleção de objetos que são similares entre si (de acordo com algum critério de similaridade fixado *a priori*) e não similares a objetos pertencentes a outros grupos. Na Análise de Agrupamento, o principal objetivo é formar grupos onde os elementos que constituem o grupo sejam os mais homogêneos possíveis entre si, e que os grupos sejam o mais heterogêneos entre si [27].

De acordo com [30], Análise de Agrupamento é a classificação não-supervisionada de dados, formando agrupamentos ou *clusters*. Ela representa uma das principais etapas de processo de análise de dados, denominada análise de *clusters*.

Entende-se por classificação não-supervisionada, o agrupamento de um conjunto de padrões não-rotulados constituindo grupos que possuam algum significado, ou seja, de tal modo que os padrões encontrados apresentem alguma propriedade comum. Sendo assim, uma vez definidos os grupos, os padrões também estarão "rotulados", mas o rótulo, neste caso, é ditado pelos próprios padrões que compõem cada grupo.

Ao contrário de outra técnica multivariada chamada Analise Discriminante, a qual é considerada uma forma de classificação supervisionada, pois classifica os objetos em

grupos pré-rotulados [47].

De acordo com [30], existem algumas etapas necessárias para realização da tarefa de agrupamento de dados:

1. Modelo de representação: opcionalmente inclui extração e/ou seleção de características dos dados;
2. Definição de um modelo de medida de similaridade/dissimilaridade adequado ao dados;
3. Agrupamento;
4. Abstração dos dados (se necessário);
5. Atribuição de saída (se necessário).

Os três primeiros passos estão representados na figura 1.1.

Na etapa 1, têm-se a obtenção e definição dos atributos relevantes dos objetos para a realização do processo de agrupamento de dados. É nessa etapa que se faz opcionalmente, por exemplo, transformações nos atributos originais para homogeneizar os dados tais como normalizações, mudanças de escala, etc.

Na etapa 2, tem-se a definição de um modelo de similaridade/dissimilaridade, o qual se trata de uma métrica usada para quantificar a similaridade/dissimilaridade entre os dados analisados, como por exemplo a distância Euclidiana entre outras.

Na etapa 3, tem-se a aplicação de um algoritmo de agrupamento de dados, sendo que existem vários algoritmos na literatura aplicáveis nessa etapa.

Em seguida, têm-se as etapas de abstração de dados e atribuição de saída, onde se entende por abstração de dados o processo de representação simples de um conjunto de dados, descrevendo a formação dos grupos e proporcionado uma compreensão relativamente fácil ao pesquisador que analisa o agrupamento. Já a atribuição de saída se refere à gráficos e tabelas que trazem informações relevantes sobre a tarefa de agrupamento aplicada nos dados.

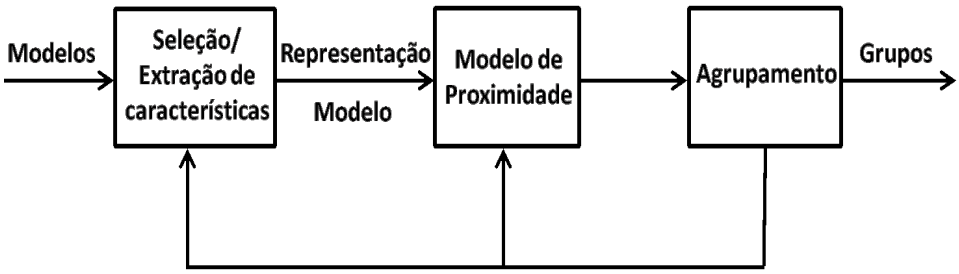

Figura 1.1: Etapas para agrupamento de dados. Fonte [30].

1.1 Definição do Problema

O problema de Agrupamento de Dados pode ser definido formalmente como um problema de otimização, descrito a seguir [20]:

Seja $\mathbf{X}$ um conjunto finito composto por N objetos, tal que $\mathbf{X} = \{x_1, x_2, \cdots, x_N\}$, onde cada $x_i \in \Re^p$ é um vetor $p-$dimensional. Os objetos $x_i, i = \{1, 2, \cdots, N\}$ deverão ser agrupados em k grupos não-vazios e mutuamente exclusivos $\mathbf{C} = \{\mathbf{C_1}, \mathbf{C_2}, \cdots, \mathbf{C_k}\}$, sujeito as seguintes restrições:

i. $\mathbf{C_1} \cup \mathbf{C_2} \cup \cdots \cup \mathbf{C_k} = \mathbf{X}$;

ii. $\mathbf{C_j} \neq \emptyset$ para todo $j = \{1, 2, \cdots, k\}$;

iii. $\mathbf{C_i} \cap \mathbf{C_j} = \emptyset$, para todo $i \neq j$, $i, j = 1, 2, \cdots, k$.

Onde se espera que a partição $\mathbf{C}$ obtida represente o conjunto de dados $\mathbf{X}$ de tal forma que objetos pertencentes ao mesmo grupo $\mathbf{C_i}$ sejam similares entre si e não similares aos objetos pertencentes aos demais grupos $\mathbf{C_j}$, para todo $i \neq j$, em que a medida de similaridade/dissimilaridade é uma métrica obtida em função da distância entre os objetos do conjunto de dados $\mathbf{X}$.

É possível calcular o número de partições distintas de um conjunto de dados $\mathbf{X}$ com N objetos, quando se quer particioná-lo em k grupos, através da sequinte expressão [2, 4, 29]:

$$S(k, N) = \frac{1}{k!} \sum_{i=1}^{k} (-1)^{k-i} \binom{k}{i} (i)^N \tag{1.1}$$

Pode-se observar na tabela 1.1 e na figura 1.2 como cresce rapidamente o número de possíveis soluções em função do número de objetos ($N = 29$) e do número de partições k em que se pretende particionar o conjunto de dados. Com esse rápido crescimento de $S(k, N)$ através do simples aumento de k, o problema de agrupamento de dados é conhecido na literatura como um problema NP completo [5]. Por outro lado, essa dificuldade tem estimulado a pesquisa por novos algoritmos e/ou aprimoramento de métodos computacionais já amplamente utilizados, não só através de heurísticas, mas também na utilização de metaheurísticas com aplicabilidade mais abrangente [41].

A figura ilustra a quantidade de possíveis soluções para um conjunto $\mathbf{X}$ com 29 objetos. Para $k = 1, 2, \cdots, 29$.

Tabela 1.1: Quantidade de possíveis soluções para um conjunto **X** com 29 objetos.

$k = 2$	$k = 3$	$k = 4$	$k = 5$
268 435 455	11 438 127 792 025	11 998 160 744 311 570	1 540 200 411 172 850 688

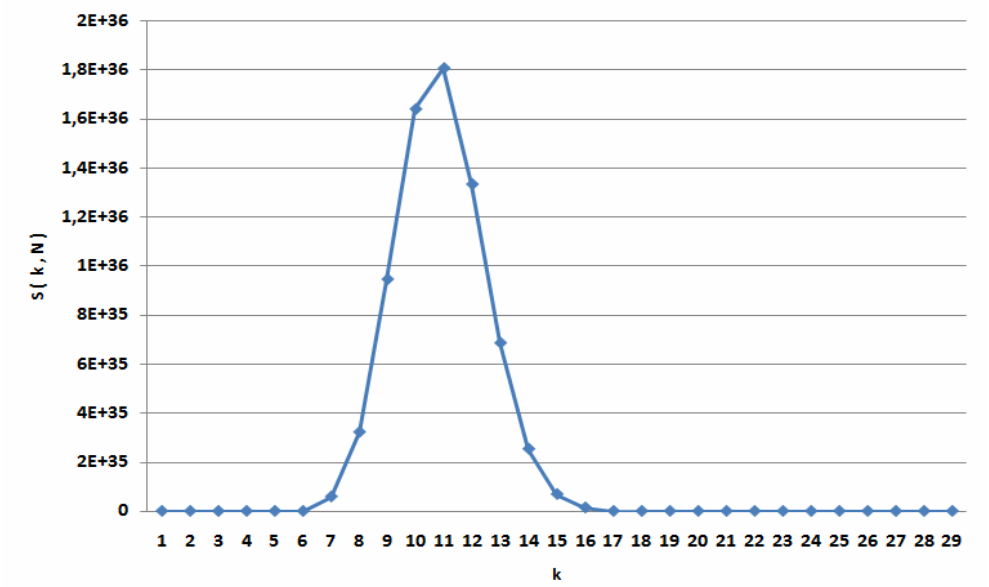

Figura 1.2: Quantidade de possíveis soluções para um conjunto **X** com 29 objetos.

1.2 Aplicações de técnicas de agrupamento

Técnicas de agrupamento têm sido usadas em várias áreas do conhecimento. Algoritmos de agrupamentos de dados podem ser usados em uma grande variedade de aplicações como as seguintes [30]:

i. Segmentação de imagens, que é definida como uma exaustiva partição de imagens em regiões homogêneas de acordo com alguma propriedade de interesse como cor, intensidade, iluminação, textura, etc. A segmentação de imagens é um componente fundamental em muitas aplicações de visão computacional;

ii. Reconhecimento de objetos, que é o uso de agrupamento de dados de imagens para classificar visões de objetos em três dimensões;

iii. Recuperação de informações, cujo objetivo é classificar e armazenar automaticamente documentos e informações relevantes para recuperação futura;

iv. Mineração de dados, cujo objetivo é a busca por informações importantes e inerentes à massa de dados, de tal forma que apresentem algum conhecimento útil e revelem

padrões.

Mais especificamente, algumas aplicações incluem:

i. *Marketing*: Pode auxiliar pessoas ligadas a área de *marketing* a descobrir grupos distintos em sua bases de clientes, para que este conhecimento seja usado para desenvolver programas de marketing direcionados [9].

ii. Uso de terras: Identificação de possibilidade de alocação de uso da terra para fins agrários e/ou urbanos em uma base de dados de observação via satélite de todo o planeta Terra [16].

iii. Identificar grupos de pessoas que possuam seguro de carro com um custo elevado de sinistralidade [49].

iv. *World Wide Web*: agrupamento de documentos de acordo com similaridades semânticas, de forma a melhorar o resultados oferecidos por sites de busca [22].

v. Estudos sísmicos: Análise de dados reais e sintéticos de terremotos para extração de características que permitam a previsão de eventos precursores de abalos sísmicos [13].

1.3 Idéias básicas

De acordo com [29], a tarefa de agrupamento de dados geralmente envolve os seguintes passos:

i. Obtenção dos dados e escolha das variáveis;

ii. Tratamento dos dados;

iii. Definição de um critério de similaridade/dissimilaridade;

iv. Escolha e aplicação de algum algoritmo de agrupamento de dados;

v. Avaliação dos resultados.

1.4 Obtenção dos dados e escolha das variáveis

A obtenção dos dados e a escolha das variáveis (atributos) estão naturalmente inseridas no campo de aplicação de problemas que envolvem a tarefa de agrupamento de dados e depende da experiência, da área de aplicação e do bom senso do pesquisador sobre a obtenção dos dados e a escolha das variáveis de interesse.

A escolha das variáveis é uma fase muito importante, pois a partir delas obtêm-se os resultados da análise de agrupamento. Deve-se levar em consideração que variáveis com valores muito elevados podem mascarar o agrupamento e levar a resultados equivocados, bem como variáveis com valores muito próximos podem influenciar na determinação final do agrupamento [7].

Entende-se por objeto aquilo que se quer classificar em grupos e por variáveis características que descrevem o objeto. Os seres vivos em geral têm a capacidade de distinguir objetos com atributos distintos, porém essa capacidade se torna limitada à medida que aumenta o número de atributos e com isso se torna necessária a busca por métodos computacionais para realizarem a tarefa de agrupamento de dados através de algum algoritmo de agrupamento.

Podem-se representar os dados em uma matriz $N \times P$, onde N representa o número de objetos e P o número de atributos de cada objeto.

$$\begin{bmatrix} x_{11} & \dots & x_{1j} & \dots & x_{1p} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \dots & x_{ij} & \dots & x_{ip} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & \dots & x_{nj} & \dots & x_{np} \end{bmatrix} \tag{1.2}$$

Onde cada observação x_{ij} representa o $j-$ésimo atributo do $i-$ésimo objeto, para todo $i = 1, \cdots, n$ e $j = 1, \cdots p$.

1.5 Tratamento dos Dados

Em [7] os autores afirmam que existem agrupamentos relativamente fáceis de identificar através de simples análise gráfica, e eventualmente um método de agrupamento poderia ter muita dificuldade na identificação de estruturas inerentes a tais dados. Isso se deve ao fato das técnicas de agrupamento partir de suposições implícitas sobre o tipo de estrutura presente nos dados, cabendo ao pesquisador analisar essas suposições.

De acordo com [7] existe um aspecto importante, o qual deve ser sempre analisado, que trata da homogeneidade entre as variáveis (atributos) de diferentes escalas que venham a participar da tarefa de agrupamento, pois a contribuição de uma variável ao coeficiente de similaridade/dissimilaridade adotado dependerá não somente de sua escala, mas também da escala das outras variáveis.

Uma solução para garantir que as variáveis contribuam de forma semelhante para o coeficiente de similaridade/dissimilaridade adotado seria homogeneizar suas variâncias, o que acontecerá se essas variáveis sofrerem transformações com o objetivo padronizá-las. Caso isso não seja feito para variáveis que apresentem variâncias heterogêneas, grupos poderão ser mascarados durante o processo de agrupamento e resultados errôneos poderão ser produzidos [7].

Assim, um forma de resolver esse problema seria fazer uma padronização dos dados através da fórmula 1.3. Ou seja, transformar os dados originais em novos valores com média zero e variância um [29, 15], em que x' são os valores dos dados transformados, $\bar{x}$ e σ_x são respectivamente a média e o desvio padrão da variável original x:

$$x' = \frac{x - \bar{x}}{\sigma_x} \tag{1.3}$$

Como exemplo da padronização da equação 1.3, a tabela 1.2 apresenta duas variáveis (atributos): Peso e Altura, as quais representam Peso(Kg) e Altura(cm) de seis indivíduos hipotéticos [7]:

Tabela 1.2: Peso (Kg) e Altura (cm) de seis indivíduos hipotéticos.

	Dados Originais		Dados Padronizados	
Indivíduo	Altura	Peso	Altura	Peso
A	180	79	1,10	1,31
B	175	75	0,33	0,75
C	170	70	-0,44	0,05
D	167	63	-0,90	-0,93
E	180	71	1,10	0,19
F	165	60	-1,21	-1,35
Média	172,8	69,7	0,0	0,0
D. Padrão	6,5	7,1	1,0	1,0

Observando a tabela 1.3 pode-se verificar o intervalo de variação dos atributos Peso e Altura, os quais estão em escalas distintas de valores, para seis indivíduos hipotéticos:

Tabela 1.3: Intervalo de variação dos atributos Peso (Kg) e Altura (cm).

Atributo	Intervalo (dados originais)	Intervalo (dados padronizados)
Altura	165 — 180	-1,21 — 1,10
Peso	60 — 79	-1,35 — 1,31

A partir dos gráficos das figuras 1.3, 1.4 e 1.5 pode-se observar o quanto a padronização da equação 1.3 torna os dados mais homogêneos.

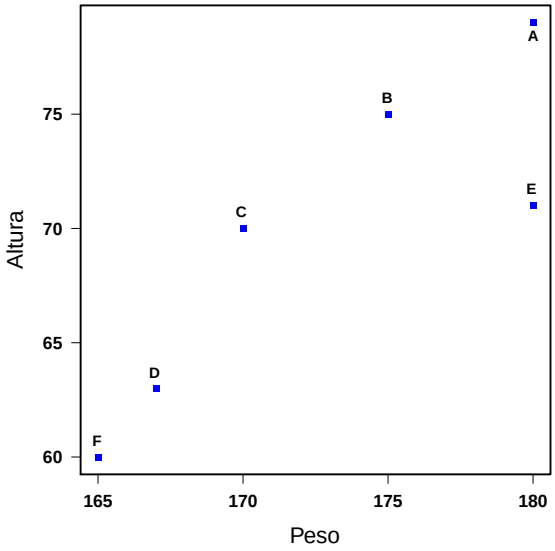

Figura 1.3: Dados originais. Peso x Altura.

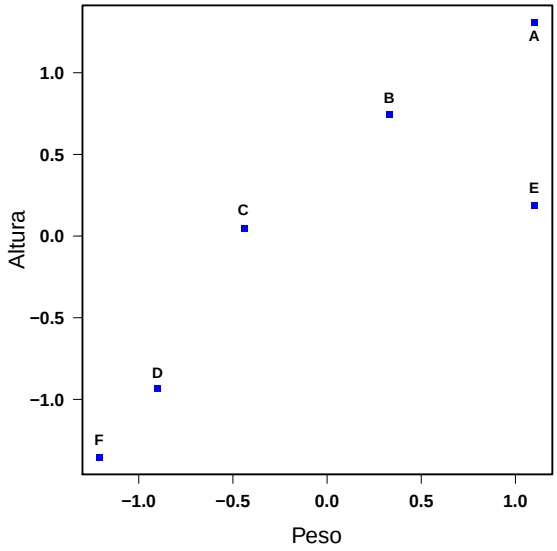

Figura 1.4: Dados Padronizados. Peso x Altura.

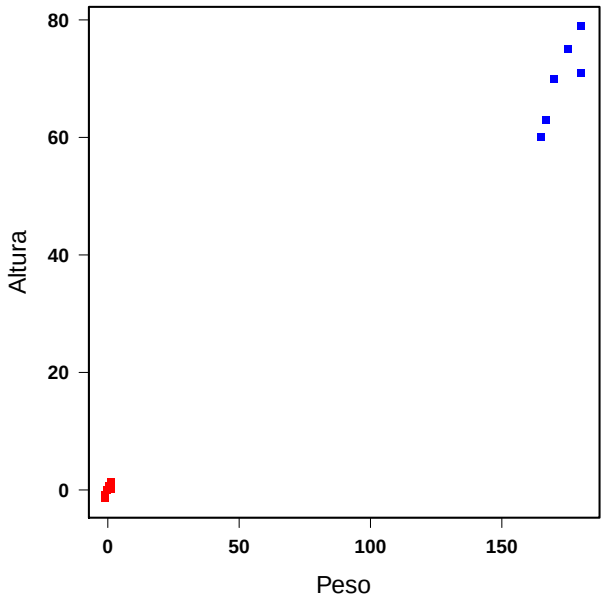

Figura 1.5: Dados originais (azul) e dados padronizados (vermelho). Peso x Altura.

Outra forma de transformação de dados é utilizada em [19]. Essa transformação é realizada de acordo com a equação 1.4, onde x' é novo valor transformado da variável, x é o valor original, $\min(x)$ é o valor mínimo da variável x e o $\max(x)$ é o valor máximo. Essa transformação consiste na normalização dos dados deixando os valores de cada variável entre zero e um.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{1.4}$$

Tomando-se a média como fator normalizador, outra transformação pode ser obtida pela equação 1.5, onde x é o valor da variável original, $\bar{x}$ é média aritmética de x e x' é o novo valor da variável transformada [7].

$$x' = \frac{x}{\bar{x}} \tag{1.5}$$

Com relação a variedade de transformações, recomenda-se que a escala das variáveis seja definida através de transformações sugeridas pelo bom senso do pesquisador e pela área de conhecimento da aplicação [7].

1.6 Medida de Similaridade/Dissimilaridade

Geralmente a tarefa de agrupamento de dados requer uma medida ou critério para quantificar o quanto um objeto é semelhante ou não à outro no conjunto de dados. Afirmar que um objeto A é semelhante a um objeto B mais do que a um objeto C pode se tornar uma tarefa árdua e nem sempre passível de solução correta. Nas medidas de similaridade, quanto maior o valor, mais similares serão os objetos, já nas medidas de dissimilaridade, quanto maior o valor, mais dissimilares serão os objetos.

A correlação, que é uma medida que quantifica o grau de associação, é um bom exemplo de medida de similaridade. Por outro lado, a distância Euclidiana é um exemplo de medida de dissimilaridade, a qual quantifica a distância $p-$dimensional entre dois objetos no espaço Euclidiano.

1.7 Requisitos das Funções de Distâncias

As funções para cálculo distâncias devem satisfazer as seguintes condições [26, 48]:

i. $D(x, y) \geq 0$;

ii. $D(x, x) = D(y, y) = 0$;

iii. $D(x, y) = D(y, x)$;

iv. $D(x, y) \leq D(x, z) + D(z, y)$.

Onde $D(x, y)$ denota a distância entre os objetos x e y de um conjunto de dados.

O primeiro requisito afirma que a distância entre dois objetos deve ser não negativa. O segundo requisito afirma que a distância entre um objeto e ele mesmo é igual a zero. O terceiro requisito afirma que a distância entre dois objetos é sempre simétrica. O quarto requisito afirma que a desigualdade triangular é satisfeita.

A distância Euclidiana, que será adotada nesse trabalho, é uma das medidas mais utilizadas na atividade de agrupamento de dados. Sua popularidade se deve à simplicidade de interpretá-la e ao baixo custo computacional $O(p)$ associado. Onde p refere-se à dimensão do conjunto de dados [15].

Existem diversas métricas utilizadas na tarefa de agrupamento de dados. O escolha de uma métrica é feita com o intuito de representar a estrutura de dados mais adequadamente. Várias métricas utilizadas na tarefa de agrupamento de dados podem ser encontradas em [29, 7].

Em particular, a distância Euclidiana entre dois objetos $p-$dimensionais é dada por:

$$D_{ij} = \left(\sum_{l=1}^{p} (x_{il} - x_{jl})^2 \right)^{\frac{1}{2}} \tag{1.6}$$

1.8 Abordagem Clássica de Técnicas de Agrupamento

De um modo geral, pode-se classificar as técnicas de agrupamento em duas grandes abordagens: Agrupamento Hierárquico e Agrupamento Particional. As duas abordagens são descritas a seguir:

1.8.1 Agrupamento Hierárquico

Agrupamento Hierárquico é uma técnica tradicional de agrupamento de dados que estabelece uma hierarquia entre os grupos formados ao longo das sucessivas iterações. Essa técnica pode ser subdividida em duas abordagens: Agrupamento Hierárquico Divisivo e Agrupamento Hierárquico Aglomerativo.

O Agrupamento Hierárquico Aglomerativo parte do princípio de que cada objeto é exatamente um grupo e prossegue realizando sucessivas aglomerações de acordo com algum critério de dissimilaridade, e ao final do procedimento todos os objetos pertencem a um único grupo, resultando em uma hierarquia de aglomerações. Ao contrário, o Agrupamento Hierárquico Divisivo inicia todos os objetos em um único grupo e realiza sucessivas partições de acordo com algum critério de dissimilaridade, de modo que ao final do procedimento cada objeto é exatamente um grupo, gerando assim uma hierarquia de partições.

Na figura 1.6 tem-se um conjunto de $N = 30$ objetos bi-dimensionais formado por 4 grupos visivelmente separados, enquanto na figura 1.7 tem-se um dendograma que ilustra o agrupamento hierárquico aglomerativo, em que o eixo vertical indica a distância entre as ligações e a base do dendograma os 30 grupos compostos por um único objeto para cada grupo. Nos níveis superiores pode-se observar como os grupos tendem à fusão até formar um único agrupamento. Ainda na figura 1.7, pode-se ver o ponto de parada (21), indicando que nesse ponto há formação de exatamente $k = 4$ grupos.

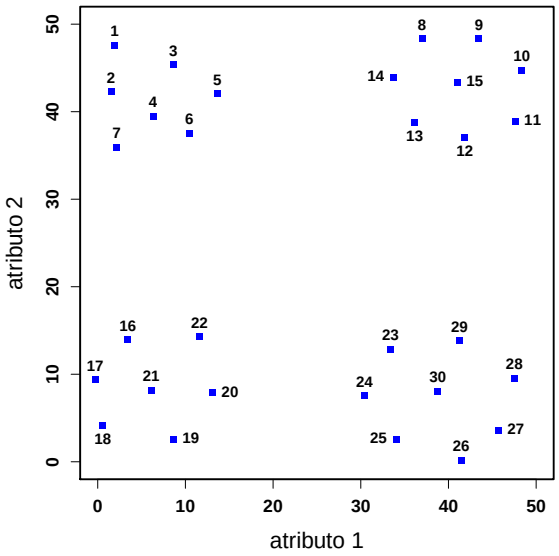

Figura 1.6: Conjunto formado por 30 objetos bi-dimensionais.

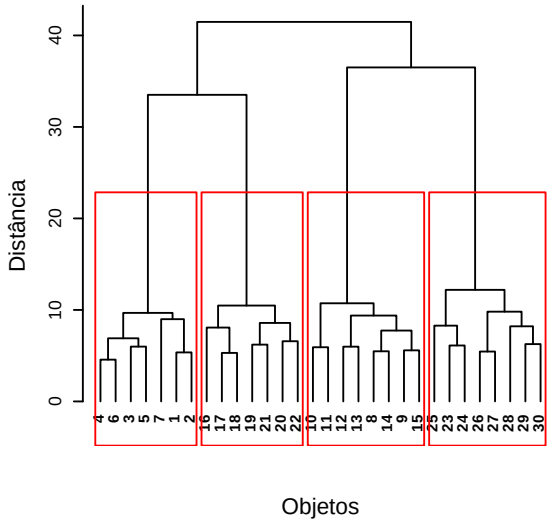

Figura 1.7: Dendograma resultante de um algoritmo hierárquico aglomerativo.

Geralmente se representa o agrupamento hierárquico através de dendrogramas, que são estruturas gráficas em forma de árvore, utilizadas para representar as junções (Hierárquico Aglomerativo) ou divisões (Hierárquico Divisivo) que ocorreram a partir de valores provenientes da matriz de distâncias [31].

Algoritmos hierárquicos de agrupamento de dados são amplamente utilizados [29, 32]. Grande parte das aplicações de agrupamento de dados nas áreas de Zoologia e Biologia utilizam técnicas hierárquicas aglomerativas. Essa técnica é bastante útil para animais e plantas que são hierarquicamente agrupados em relação a características genéticas [14], porém essa técnica se torna inviável para grandes conjuntos de dados, onde prefere-se a utilização de métodos particionais por apresentarem em geral menor custo computacional [30]. Na próxima seção serão vistas as principais características do métodos particionais.

Em [35], os autores descrevem algumas vantagens e desvantagens dos métodos hierárquicos. Algumas vantagens são: fácil manipulação de qualquer medida de similaridade/dissimilaridade, fácil visualização dos objetos através de dendrogramas e aplicabili-

dade de qualquer tipo de atributo. As desvantagens são: dificuldade de escolha do correto critério de parada e a não relocação de objetos entre grupos já divididos ou aglomerados durante o processo de hierarquização dos objetos.

1.8.2 Agrupamento Particional

Agrupamento Particional é outro método tradicional na área de agrupamento de dados. O Agrupamento Particional tem como objetivo dividir os objetos de um conjunto de dados em k grupos.

Geralmente os agrupamentos são obtidos de forma a minimizar (maximizar) uma função. Assim, a idéia é verificar o ganho dessa função ao se passar de k para $k+1$ grupos [7]. O gráfico do número de grupos contra a função objetivo, permitirá uma visualização da perda (ganho) da função conforme se aumenta o número de grupos. Como o número de grupos é uma informação que está intrísecamente relacionada à estrutura do conjunto de dados, nem todos os valores de k podem levar a partições naturais dos dados. Assim uma das maneiras de investigar o número k de grupos não conhecido *a priori* é obter várias partições para diversos valores de k e avaliá-las através de algum critério numérico ou até mesmo por um especialista da área.

As técnicas de agrupamento particional têm uma vantagem em relação às técnicas de agrupamento hierárquicas: quando a aplicação envolve conjunto de dados com grande número de objetos, onde a construção de dendrogramas é computacionalmente inviável. A desvantagem das técnicas de agrupamento particional está exatamente na escolha do correto número de grupos em conjunto de dados onde esta informação é desconhecida [30].

1.9 Validação dos Resultados

Validação dos resultados se refere a procedimentos que avaliam os resultados da análise de agrupamento de uma forma objetiva e quantitativa [29]. Após coletar os dados, escolher as variáveis, realizar possíveis transformações nos dados originais e executar o algoritmo de agrupamento de dados, surge a necessidade de avaliar se os resultados do agrupamento gerado tem boa qualidade, se é melhor que resultados gerados por outros algoritmos etc. Assim, uma forma de avaliar os resultados é utilizar um conjunto de dados cuja classificação é conhecida e comparar os resultados produzidos pelos algoritmos de agrupamento.

1.9.1 Índice de Rand

O Índice de Rand [40] calcula o grau de similaridade entre o agrupamento $\mathcal{R}$ de um conjunto de dados cuja classificação é conhecida *a priori* e que servirá como referência e o agrupamento $\mathcal{C}$ gerado pelo algoritmo que se pretende avaliar. Esse índice simplesmente avalia os dois agrupamentos $\mathcal{R}$ e $\mathcal{C}$ do mesmo conjunto de dados. O índice de Rand é dado por:

$$I_{\mathbf{R}}(\mathcal{C}, \mathcal{R}) = \frac{a + d}{a + b + c + d} \tag{1.7}$$

Dado $\mathcal{A}$ o conjunto de todos os objetos sem qualquer tipo de classificação prévia e $\mathcal{B}$ o conjunto onde os elementos são todas as possíveis combinações dois a dois dos elementos de $\mathcal{A}$, têm-se que:

- a: Número de pares de objetos do conjunto de dados que pertencem ao mesmo grupo em $\mathcal{R}$ e em $\mathcal{C}$;
- b: Número de pares de objetos do conjunto de dados que pertencem ao mesmo grupo em $\mathcal{R}$ e a diferentes grupos em $\mathcal{C}$;
- c: Número de pares de objetos do conjunto de dados que pertencem a diferentes grupos em $\mathcal{R}$ e ao mesmo grupo em $\mathcal{C}$;
- d: Número de pares de objetos do conjunto de dados que pertencem a grupos ditintos em $\mathcal{R}$ e a grupos distintos em $\mathcal{C}$;

Os termos a e d são medidas de classificações consistentes, enquanto os termos b e c são medidas de classificações inconsistentes. Veja que:

i. $I_{\mathbf{R}} \in [0, 1]$;

ii. $I_{\mathbf{R}} = 0$ se e somente se $\mathcal{C}$ é completamente inconsistente, ou seja, $a = d = 0$;

iii. $I_{\mathbf{R}} = 1$ se e somente se a partição $\mathcal{C}$ sob avaliação corresponde exatamente à partição de referência $\mathcal{R}$, ou seja, $b = c = 0$ ($\mathcal{C} = \mathcal{R}$).

1.9.2 Índice de Rand Ajustado

O Índice de Rand Ajustado foi proposto por Hubert e Arabie (1985) [28]. Estes autores determinaram o valor esperado do Índice de Rand, dando origem ao índice de Rand Ajsutado, visto que no Índice de Rand o valor esperado não é nulo para duas partições completamente aleatórias de um conjunto de dados. O valor esperado para o Índice de Rand é dado por:

$$E\left[I_{\mathbf{R}}(\mathcal{C}, \mathcal{R})\right] = \frac{(a + c)(a + b)}{M} \tag{1.8}$$

em que $M = a + b + c + d$.

Assim, o Índice de Rand Ajustado é dado por:

$$I_{\mathbf{R_{aj}}}(\mathcal{C}, \mathcal{R}) = \frac{a - \frac{(a+c)(a+b)}{M}}{\frac{(a+c)+(a+b)}{2} - \frac{(a+c)(a+b)}{M}} \tag{1.9}$$

Este índice assume valores em $[-1, 1]$, onde o valor 1 indica um perfeito acordo entre as duas partições, enquanto que valores próximos de 0 correspondem a um acordo entre as partições devido ao acaso. As medidas a, b, c, d continuam com o mesmo significado usado no Índice de Rand.

1.9.3 Silhueta

Considerando um objeto $\mathbf{x_i}$ de um conjunto de dados $\mathbf{X}$, tal que $\mathbf{x_i}$ pertence a um dado grupo $\mathbf{C_a}$. Seja $a(\mathbf{x_i})$ a distância média de $\mathbf{x_i}$ para os demais objetos de $\mathbf{C_a}$. Agora, considerando um outro grupo $\mathbf{C_c}$, a distância média do objeto $\mathbf{x_i} \in \mathbf{C_a}$ para todos os objetos do grupo $\mathbf{C_c}$ é denotada por $D(\mathbf{x_i}, \mathbf{C_c})$. Realiza-se o cálculo de $D(\mathbf{x_i}, \mathbf{C_c})$ para todos os grupos $\mathbf{C_c} \neq \mathbf{C_a}$, e então calcula-se o menor valor, ou seja, $b(\mathbf{x_i}) = \min\{D(\mathbf{x_i}, C_c)\}$, para todo $\mathbf{C_c} \neq \mathbf{C_a}$. Onde $b(\mathbf{x_i})$ representa a distância de $\mathbf{x_i}$ para o grupo mais próximo, e a Silhueta $s(\mathbf{x_i})$ é dada por:

$$s(\mathbf{x_i}) = \frac{b(\mathbf{x_i}) - a(\mathbf{x_i})}{\max\{b(\mathbf{x_i}), a(\mathbf{x_i})\}} \tag{1.10}$$

Onde se verifica que $s(\mathbf{x_i}) \in [-1, 1]$. A alocação do objeto $\mathbf{x_i}$ para um dado grupo será tanto melhor quanto maior for o valor de $s(\mathbf{x_i})$. Se $s(\mathbf{x_i}) = 0$, não fica evidente a alocação do objeto $\mathbf{x_i}$ no grupo atual ou no grupo mais próximo. Se o grupo $\mathbf{C_a}$ for unitário (*singleton*), $s(\mathbf{x_i})$ é não definida e deve-se atribuir $s(\mathbf{x_i}) = 0$ [34]. Finalmente, o critério da Silhueta é dado pela média de $s(\mathbf{x_i})$, para todo $i = 1, \cdots, N$. Ou seja:

$$\bar{s} = \sum_{i=1}^{N} s(\mathbf{x_i}) \tag{1.11}$$

Segundo esse critério, a melhor partição dos dados é obtida quando a média das silhuetas atingir o valor máximo.

O critério silhueta é mais adequado para agrupamentos com grupos compactos e separados. Porém, o critério resulta em valores tendenciosos para grupos potencialmente sobrepostos, favorecendo agrupamentos disjuntos. Além disso, o critério nem sempre obtém bons resultados para grupos com formatos arbitrários [43].

Capítulo 2

Equação Logística (Modelo Contínuo)

As curvas logísticas ou curvas sigmóides podem descrever processos de crescimento natural de qualquer sistema [12, 11].

Um processo de crescimento natural consiste em preencher um determinado "nicho" desde o início até a saturação, uma vez que todo o nicho a ser preenchido apresenta limite de capacidade. Assim, o crescimento de uma população (humana ou de qualquer espécie animal), a difusão de uma epidemia ou de uma inovação tecnológica, o crescimento do mercado de um produto, o crescimento de um ser vivo ou de uma população, etc., são considerados como processos de crescimento natural e são descritos por curvas logísticas [12, 11].

Os processos de aprendizagem são também processos de crescimento natural, e por isso as curvas logísticas são também designadas por curvas de aprendizagem. Quando um indivíduo, ou um grupo, ou uma colectividade de pessoas aprende a realizar uma determinada tarefa ou aprende um determinado tema, o que está a acontecer é o crescimento cumulativo de informação, que começa com uma informação inicial e aumenta até ao esgotamento da informação para executar a tarefa ou para dominar o tema em questão [12, 11].

A equação logística contínua, de Pierre Verhulst (1845) é dada por:

$$X_t = \frac{1}{1 + (x_0^{-1} - 1)e^{-\mu t}} \tag{2.1}$$

Onde $x_0 \in (0, 1)$ é o tamanho inicial da população e $\mu \in \mathfrak{R}$ representa a taxa máxima de crescimento ou decaimento populacional.

Para $\mu < 0$ a população se extinguirá com o passar do tempo e para $\mu > 0$ a população sobreviverá, como ilustra a figura 2.1, para tamanho da população $X_0 = 0.5$, $\mu = -0.5$, $\mu = -0.1$, $\mu = 0.1$ e $\mu = 0.5$.

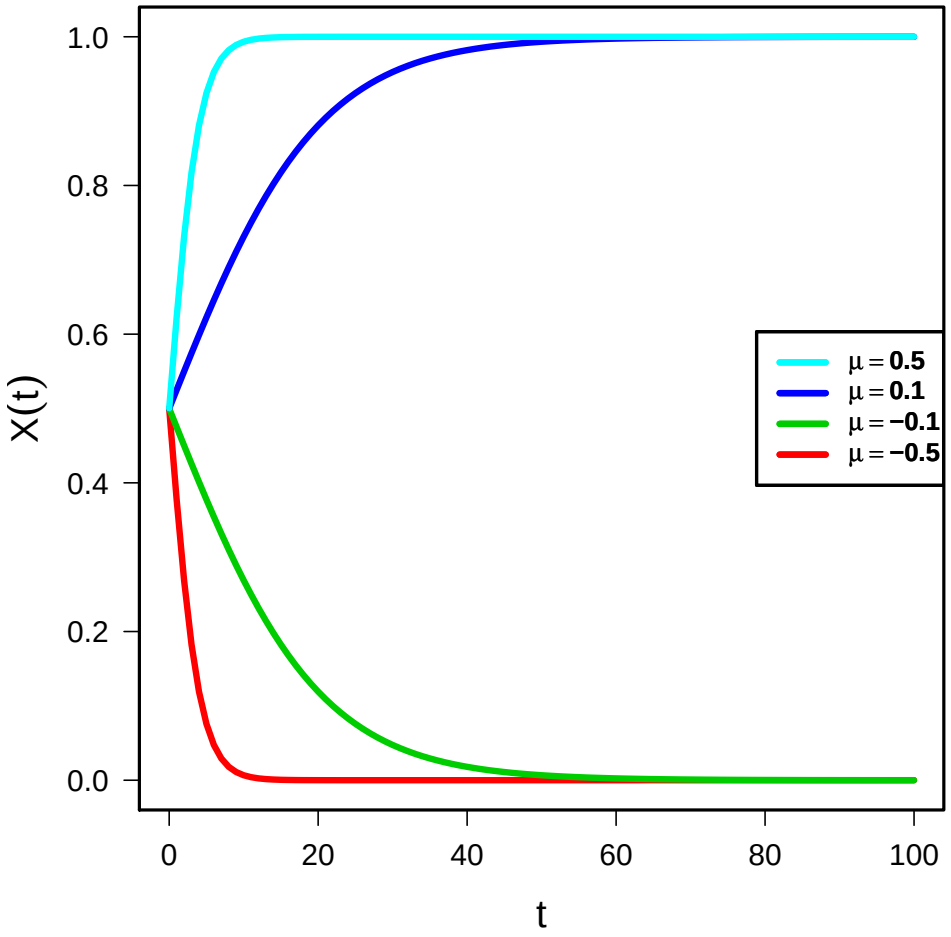

Figura 2.1: Tamanho da população para $X_0 = 0.5$.

As equações e curvas logísticas constituem atualmente uma poderosa e simples ferramenta matemática para fazer previsões sobre o crescimento de sistemas. Este modelo (equação 2.1) é utilizado, por exemplo, para projetar populações futuras, no caso de inobservância de fatalidades, como as provocadas por guerras e epidemias [1]. Outro exemplo, corresponde ao processo de aquisição de vocabulário de uma criança entre o seu nascimento e os seis anos de vida, ou seja, durante o período pré-escolar [12].

A figura 2.2 ilustra o comportamento da curva logística no intervalo $[-10{,}10]$ para diferentes valores do parâmetro μ ($\mu = 0.001$, $\mu = 0.1$, $\mu = 1$ e $\mu = 4$) e tamanho inicial da população $X_0 = 0.5$.

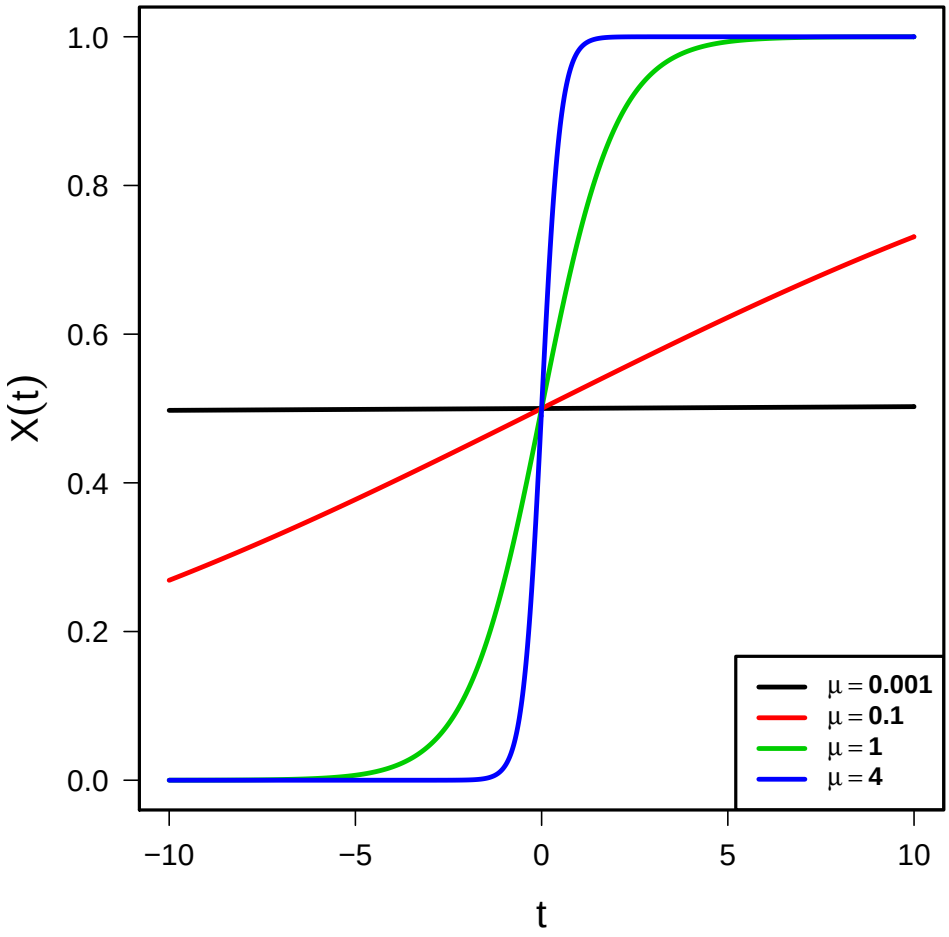

Figura 2.2: Comportamento da curva logística no intervalo $[-10, 10]$.

Capítulo 3

Algoritmos de Agrupamento

3.1 K-médias

O algoritmo k-médias foi proposto por Macqueen [37]. K-médias é uma técnica de agrupamento em que os dados são agrupados de acordo com alguma métrica de distância entre os objetos do conjunto de dados. Geralmente usa-se como métrica a distância Euclidiana.

K-médias é um dos mais simples algoritmos de aprendizado supervisionado propostos para o problema de agrupamento de dados. O procedimento segue uma maneira simples e fácil de classificar um determinado conjunto de dados através de um certo número de grupos K fixado a priori [46]. Segundo Jain et.al. [30] o algoritmo K-médias é popular devido a sua facilidade de implementação e sua ordem de complexidade $O(N)$, onde N é o número de objetos do conjunto de dados.

O termo "médias" refere-se aos centróides dos grupos. Os centróides são selecionados de forma aleatória, sendo recalculados de forma iterativa até o algoritmo atingir seu objetivo. O K-médias tem como objetivo a minimização da distância entre os objetos do mesmo grupo e a maximização em relação aos outros grupos formados [42].

O algoritmo básico do K-médias é dado pelos seguintes passos [18, 3]:

i. Distribua aleatoriamente todos os objetos aos k grupos;
ii. Calcule a centróide de cada grupo;
iii. Para cada objeto calcule a distância entre ele e os centros de todos os grupos, atribuindo esse objeto ao grupo mais próximo;
iv. Caso haja alguma movimentação de objeto de um grupo para o outro no passo iii, volte para o passo ii;
v. Saída com o resultado do agrupamento.

3.1.1 Algumas desvantagens do K-médias

Algumas vantagens do K-médias são [23, 3]:

i. O resultado final é extremamente dependente da partição inicial, o que torna o K-médias propenso a convergir para soluções locais;

ii. O número de grupos K deve ser informado *a priori* como um parâmetro de entrada, o que pode ser um problema em aplicações onde o correto número de grupos não é conhecido;

iii. O K-Médias pode gerar grupos vazios. Se grupos vazios são gerados o algoritmo é reinicializado para forçar a geração de exatamente K grupos;

iv. Devido ao uso do critério de mínima variância, o K-médias apresenta melhor desempenho na identificação de grupos na forma esférica. Em grupos com outras formas o K-médias pode ser completamente ineficaz;

v. O K-médias requer o cálculo do centro dos grupos, isto somente pode ser aplicado em conjunto de dados numéricos;

vi. O algoritmo não é eficiente para conjunto de dados com altas dimensões, uma vez que a distância entre os centros de cada grupo e todos os objetos do conjunto de dados tem que ser recalculada em cada interação.

3.2 Algoritmos Hierárquicos

Os algoritmos hierárquicos criam uma hierarquia de relacionamentos entre os objetos do conjunto de dados.

Existem duas versões: aglomerativa e divisiva. A aglomerativa cria grupos a partir de elementos isolados (*singletons*) e a cada iteração aglutina dois grupos com base na distância entre eles. A divisiva inicia com um grande grupo formado por todos os elementos do conjunto de dados e vai dividindo-os até chegar a elementos isolados. A versão aglomerativa funciona de acordo com o seguinte algoritmo [33]:

- Forme um grupo para cada elemento (*singletons*);
- Encontre os pares de grupos mais similares, de acordo com uma medida de distância e método escolhidos;
- Aglomere-os em um grupo maior e recalcule a distância deste grupo para todos os outros elementos;
- Repita os passos 2 e 3 até formar um único cluster.

Exitem vários algoritmos aglomerativos. Nessa dissertação, serão usados 4 algoritmos: Método Ward, Método da Ligação Simples, Método da Ligação Completa e Método Ligação Média.

Método Ward: o método de Ward é um procedimento em que a medida de similaridade usada para juntar agrupamentos é calculada como a soma de quadrados das distâncias entre os dois agrupamentos feita sobre todas as variáveis. Esse método tende a resultar em agrupamentos de tamanhos aproximadamente iguais devido a sua minimização de variação interna. Em cada iteração, combinam-se os dois agrupamentos que apresentarem menor aumento na soma global de quadrados dentro dos agrupamentos [21]. O método Ward que tem se revelado um dos melhores e mais usados métodos hierárquicos de aglomeração [36, 38].

Ligação Simples: A distância entre dois grupos é dada pela distância entre os seus pontos mais próximos, também chamada de ”agrupamento de vizinhos“.

Ligação Completa: A distância entre grupos é a distância entre seus pontos mais distantes.

Ligação Média: Utiliza-se a distância entre os centróides dos grupos. Os centróides e as distâncias são recalculadas cada vez que um cluster se altera.

A figura 3.1 ilustra a distância entre dois grupos, para os três métodos:

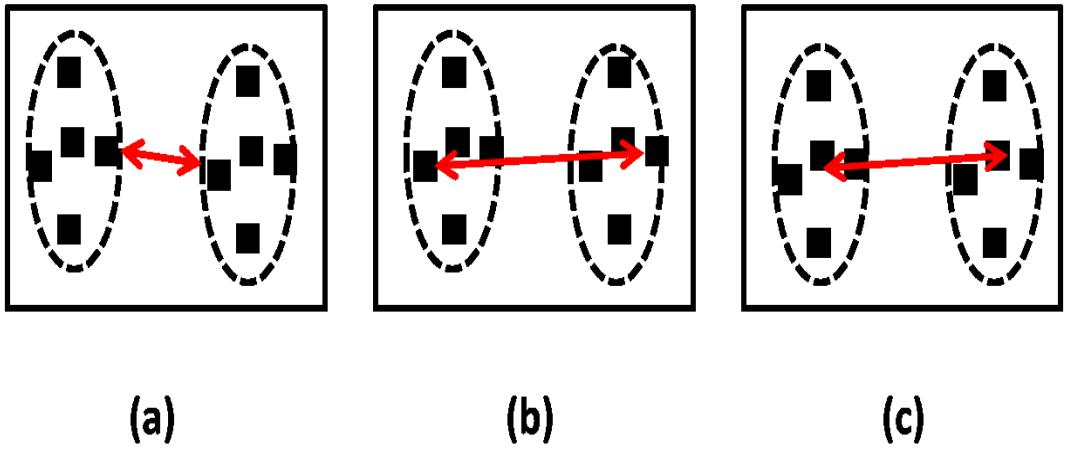

Figura 3.1: (a) Ligação Simples. (b) Ligação Completa. (c) Ligação Média.

Na abordagem de agrupamento aglomerativo, qualquer decisão errada de aglomeração dos grupos no inicio da execução do algoritmo tende a aumentar os erros à medida que o agrupamento é executado [50].

3.3 RGT

A figura 3.2 ilustra os três procedimentos principais do novo método: procedimento de Inicialização, Filtro e Finalização.

No procedimento de Inicialização faz-se a leitura dos atributos dos N objetos do conjunto de dados e calcula-se a distância Euclidiana $p-$dimensional entre eles. Nesse procedimento se dá a inicialização do parâmetro β, necessário à execução do algoritmo RGT. Ainda no procedimento de Inicialização, se necessário, aplica-se transformação nos atributos originais, como por exemplo normalização, etc.

No procedimento de Fitro, faz-se transformações nas distâncias entre os N objetos de acordo com a função de transformação sigmóide descrita na página . O número de transformações é uma informação ditada pelo usuário do algoritmo. Com base na observação do histograma dos valores transformados, toma-se a decisão de proceder com uma nova transformação ou não.

No procedimento de Finalização, informa-se o limiar de ativação das distâncias transformadas (conexões) entre os objetos. Conexões com valores abaixo do limiar são consideradas ativas, enquanto conexões com valores acima do limiar são consideradas inativas. Objetos interligados por conexões ativas, pertencem ao mesmo grupo. Objetos não conectados, pertencem a grupos distintos.

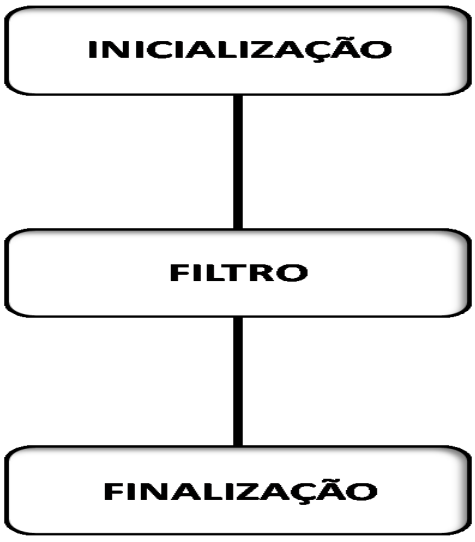

Figura 3.2: Visão geral do algoritmo RGT.

A figura 3.3 ilustra como funciona o procedimento de Inicialização do algoritmo RGT. Primeiramente, informa-se o valor do parâmetro β. Para todas as execuções do algoritmo RGT, usou-se $\beta = 4.0$. A seguir, faz-se a leitura dos a tributos dos N objetos $p-$dimensionais do conjunto de dados. Após essa leitura, se necessário, faz-se uma trasnformação nos dados originais, como padronização, etc. Por último, calcula-se a distância Euclidiana $p-$dimensional entre todos os objetos do conjunto de dados.

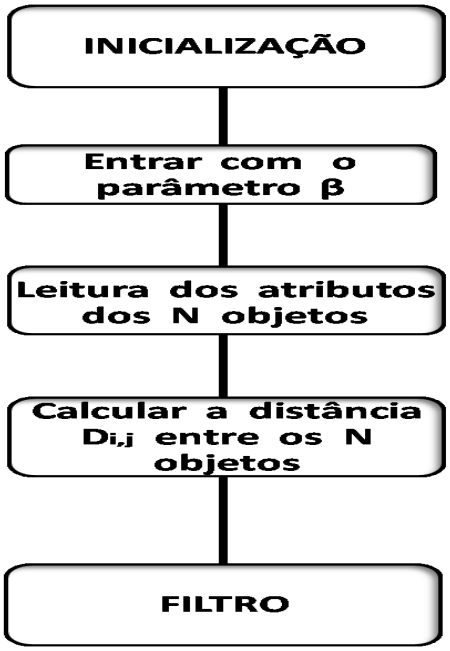

Figura 3.3: Procedimento de Inicialização do algoritmo RGT.

A figura 3.4 ilustra o funcionamento do procedimento de Filtro:

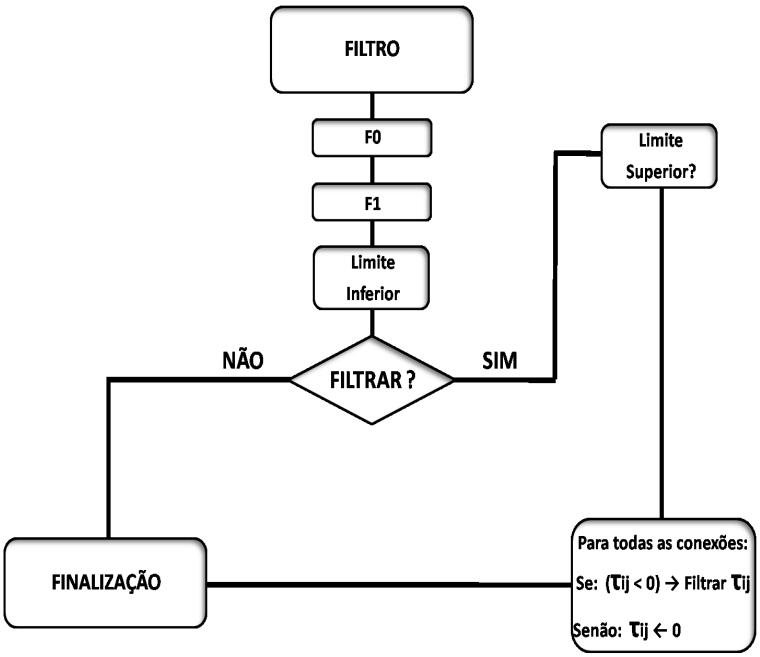

Figura 3.4: Procedimeto de Filtro do algoritmo RGT.

Primeiramente, aplica-se o Filtro 0 à todas conexões, ou seja, $\forall i, j = 1, \cdots, N$, com i, através da equação 3.1:

$$\tau_{ij} = \frac{D_{ij} - D_{min}}{D_{max} - D_{min}} - 0.5 \tag{3.1}$$

onde $\tau_{ij} \in [-0.5, 0.5]$. O Filtro 0 é simplesmente a transformação inicial das distâncias.

A seguir, aplica-se o Filtro 1 à todas conexões, através da equação 3.2:

$$\tau_{ij} = \frac{2}{1 + e^{-\beta(\tau_{ij})}} - 1 \tag{3.2}$$

onde $\tau_{ij} \in (-1, 1)$.

A seguir, aplica-se uma transformação no limite inferior, através da equação 3.3:

$$\tau_{min} = \frac{2}{1 + e^{-\beta(-0.5)}} - 1 \tag{3.3}$$

Se for necessário aplicar mais um filtro, deve-se informar um novo limite superior para os valores das conexões. Agora, para todas as conexões, testa-se a seguinte condição:

se $\tau_{ij} < 0$:

$$\tau_{ij} = \frac{2}{1 + exp^{-\beta(\triangle)}} - 1 \tag{3.4}$$

onde:

$$\triangle = \frac{\tau_{ij} - \tau_{min}}{\tau_{max} - \tau_{min}} - 0.5 \tag{3.5}$$

senão, $\tau_{ij} \leftarrow 0$.

Após aplicação dos Filtros, têm-se o procedimento de Finalização do algoritmo RGT, como ilustra a figura 3.5. Informa-se o limiar de ativação das conexões. Conexões com valor inferior ao limiar, são consideradas ativas. Conexões com valor acima do limiar, são consideradas inativas. Objetos interligados por conexões ativas pertencem ao mesmo grupo, enquanto objetos não conectados pertencem a grupos distintos.

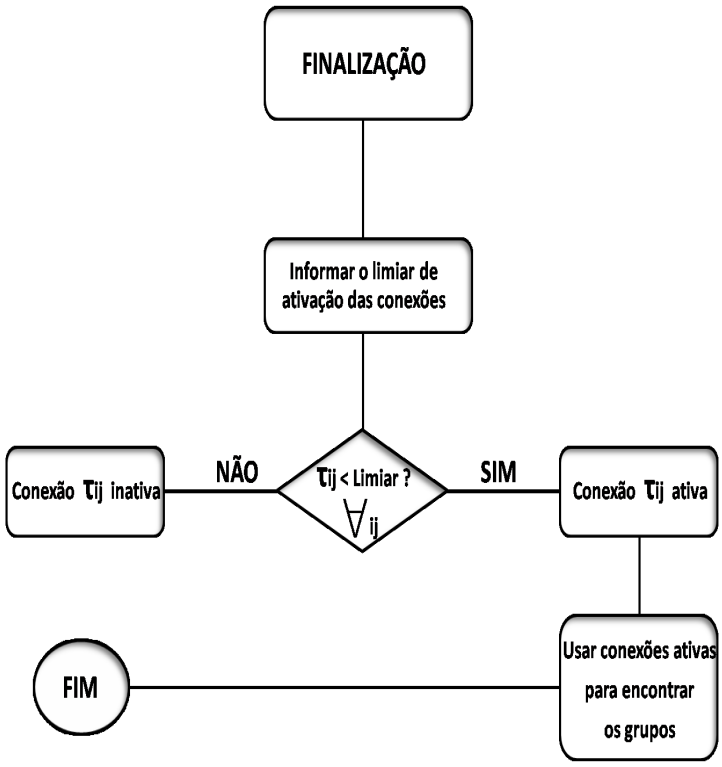

Figura 3.5: Procedimento de Finalização do algoritmo RGT.

Capítulo 4

Resultados

Este capítulo trata dos resultados produzidos pelos métodos utilizados em alguns conjuntos de dados avaliados.

Em [24], os autores afirmam que diferentes algoritmos de agrupamento de dados geram diferentes resultados. Uma maneira para avaliar os diferentes resultados dos algoritmos é comparar a eficiência desses resultados através de conjuntos de dados que já possuem classificação conhecida.

Neste capítulo serão utilizados alguns conjuntos de dados artificiais e reais para comparar o método K-médias e alguns Algoritmos Hierárquicos como o novo método: o algoritmo RGT.

Os conjuntos de dados que foram analisados estão representados na tabela 4.1 a seguir:

Tabela 4.1: Conjuntos de dados analisados.

	Nome	Instância	Dimensão	N^o Grupos
Artificial	Ruspini	75	2	4
	Espiral222-2D-2C	222	2	2
Real	Sobreviventes	306	3	2
	Íris	150	4	3
	Wreath	1000	2	14
	Ionosfera	351	34	2

4.1 Ruspini

O conjunto de dados Ruspini é composto por 75 objetos. Trata-se de um conjunto determinístico. Cada objeto possui dois atributos como mostra a figura 4.1. Esse conjunto de dados foi gerado para realizar testes de agrupamento e é composto por quatro grupos de 20, 23, 17 e 15 objetos em cada grupo [44].

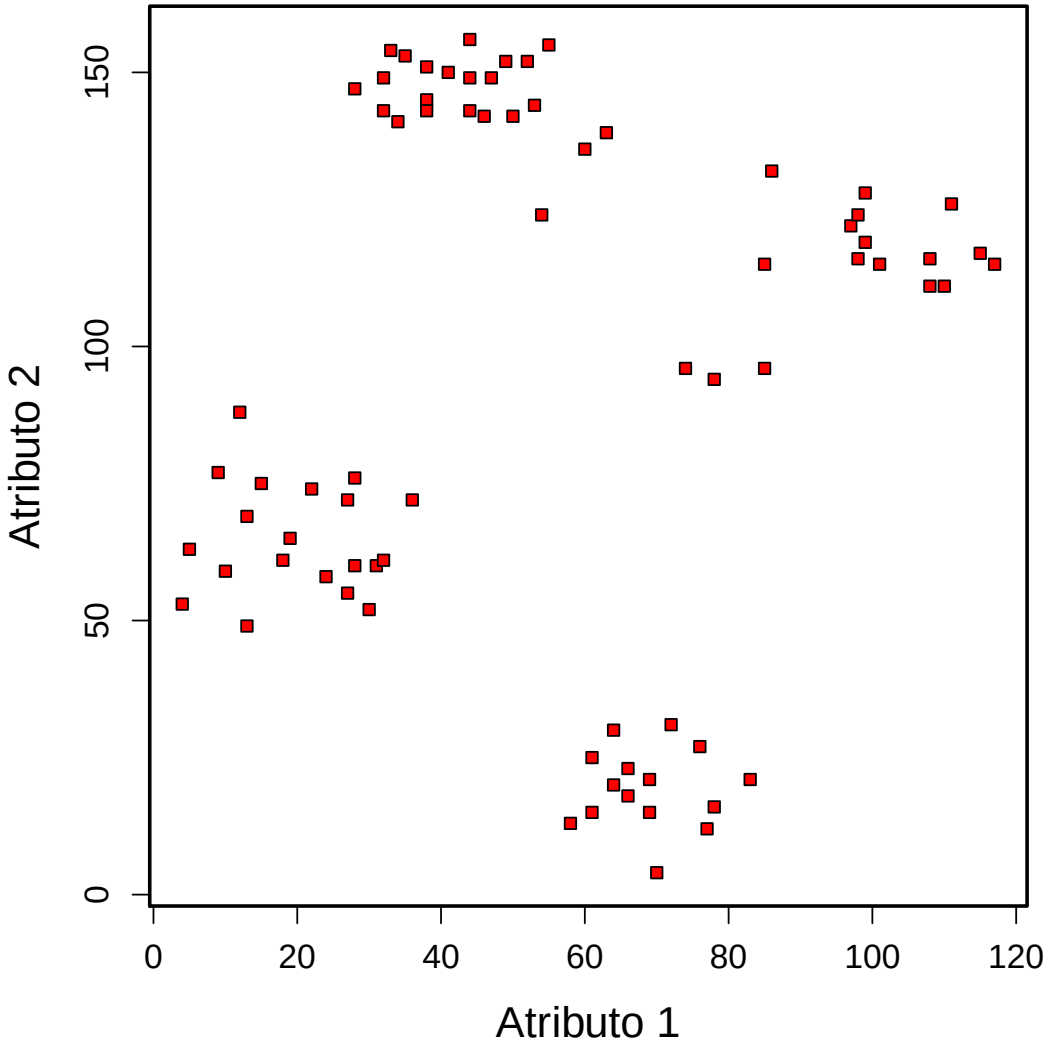

Figura 4.1: Representação do conjunto de dados Ruspini.

A seguir, pode-se ver nas figuras 4.2 e 4.3 os histogramas dos Filtros 0 e Filtro 1 aplicados ao conjunto de dados Ruspini:

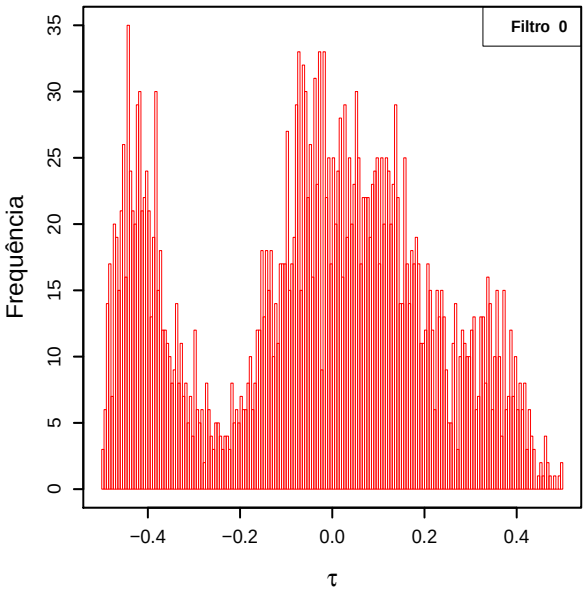

Figura 4.2: Ruspini: filtro 0.

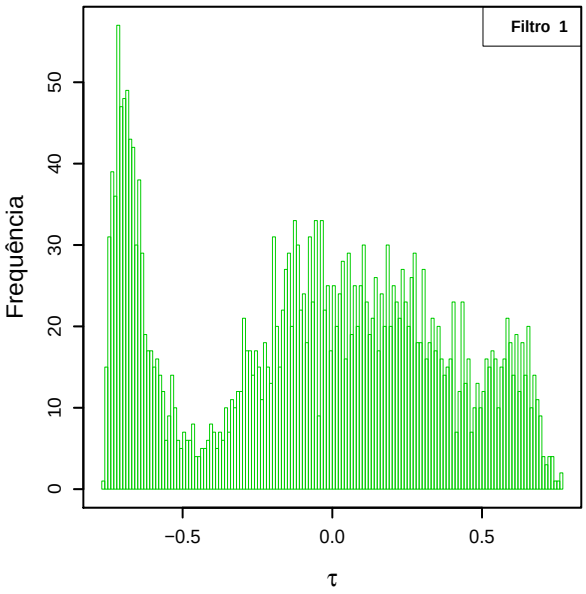

Figura 4.3: Ruspini: filtro 1.

Na figura 4.4 pode-se ver o histograma do filtro 2 e o limiar de ativação das conexões que interligam os objetos do conjunto de dados Ruspini. A figura 4.5 ilustra cada objeto do conjunto de dados Ruspini e seu respectivo rótulo:

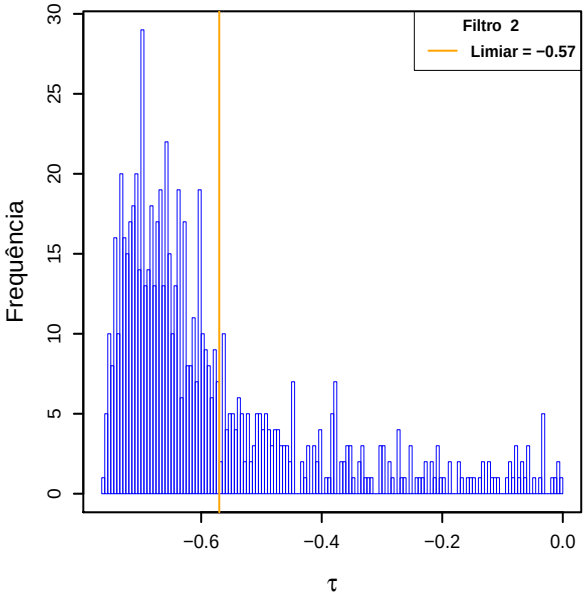

Figura 4.4: Ruspini: filtro 2.

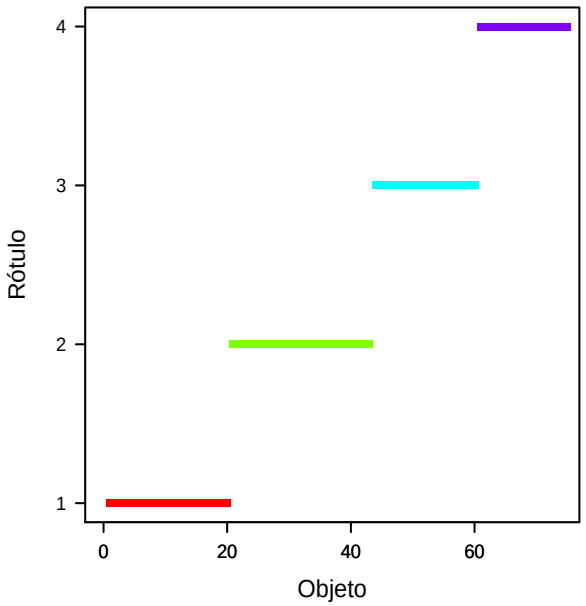

Figura 4.5: Ruspini: rótulos dos objetos, formando 4 grupos.

Pode-se ver a partir da figura 4.6 à figura 4.11 a variação do número de grupos (1 à 6) e o valor da média das silhuetas ($\bar{s}$) dada pela equação 1.11:

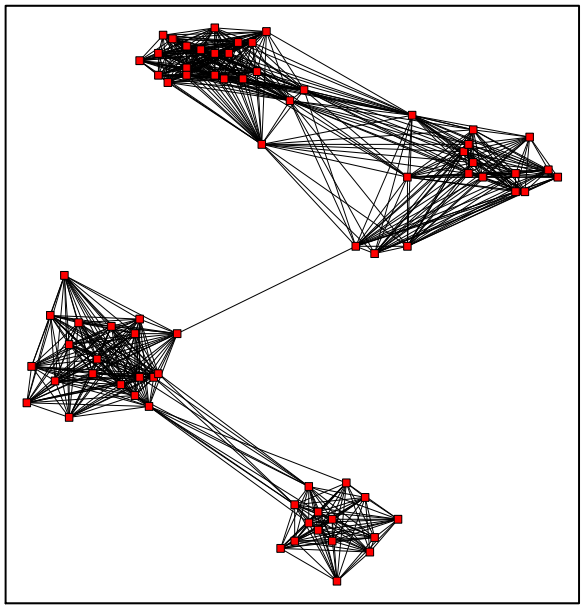

Figura 4.6: Ruspini: resultado com 1 grupo.

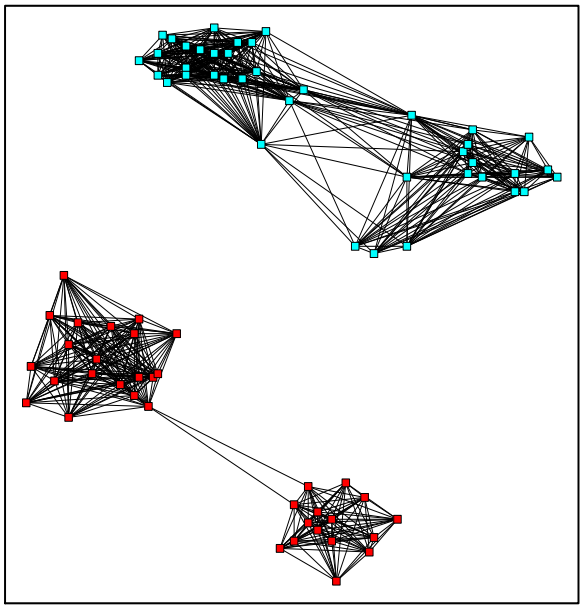

Figura 4.7: Ruspini: resultado com 2 grupos. $\bar{s} = 0.582726$.

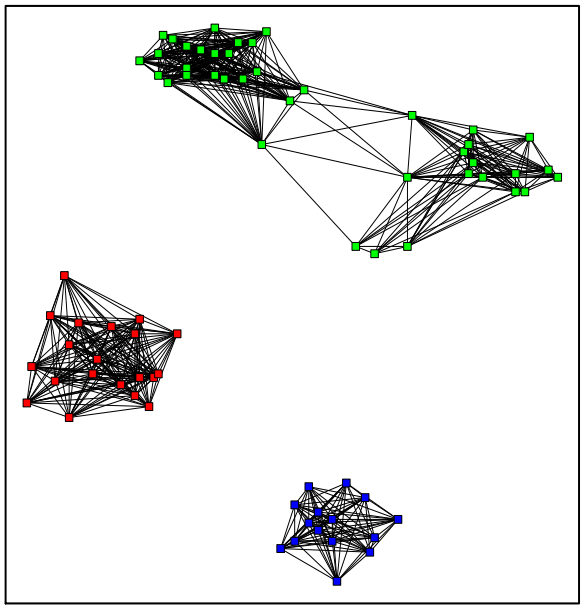

Figura 4.8: Ruspini: resultado com 3 grupos. $\bar{s} = 0.641392$.

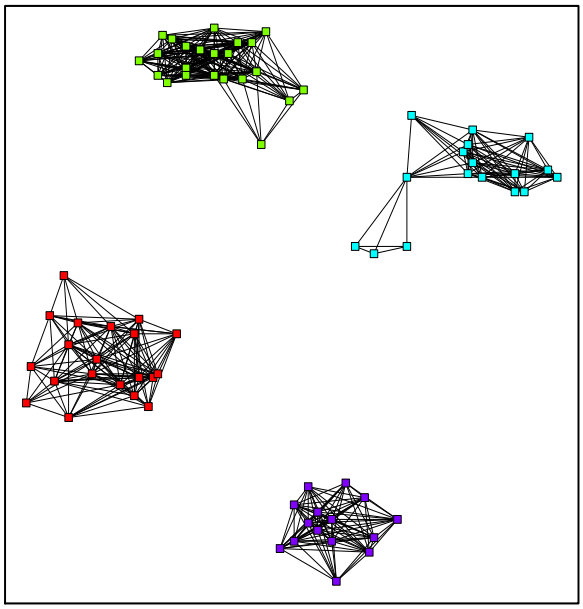

Figura 4.9: Ruspini: resultado com 4 grupos. $\bar{s} = 0.737657$.

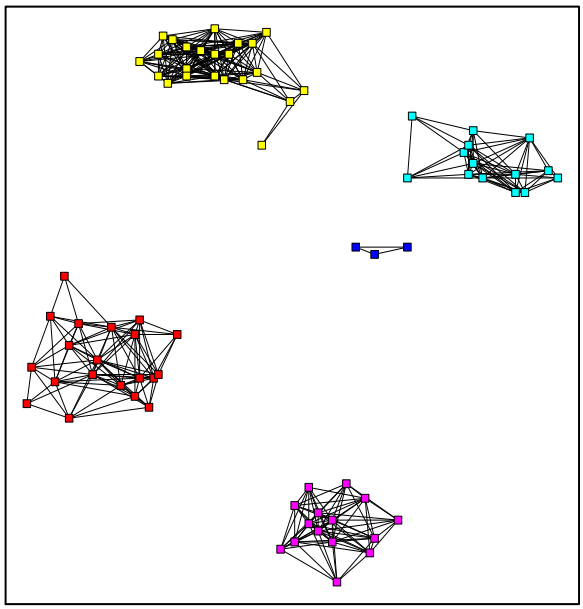

Figura 4.10: Ruspini: resultado com 5 grupos. $\bar{s} = 0.713479$.

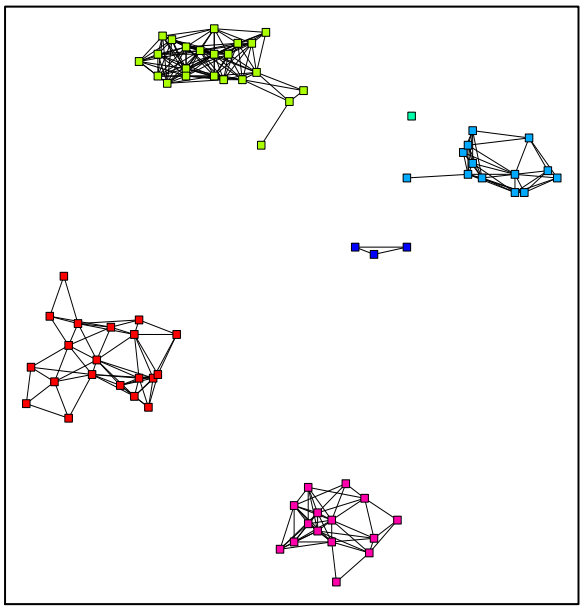

Figura 4.11: Ruspini: resultado com 6 grupos. $\bar{s} = 0.627393$.

Pelo critério da média das silhuetas, a melhor partição é aquela que produz a maior média das silhuetas. Assim, a melhor partição para o conjunto de dados Ruspini é obtida quando há formação de 4 grupos. O Índice de Rand e o Índice de Rand Ajustado confirmam os resultados, pois são iguais a 1 para 4 grupos, como mostra a tabela 4.2. O algoritmo RGT classificou corretamente o conjunto de dados Ruspini. O K-médias teve um bom desempenho em 20 execuções, mas mostrou-se inferior aos demais. Os números entre parênteses representam a média e o desvio padrão dos índices de Rand e Rand Ajustado nas 20 experiências do algoritmo K-médias. Os algoritmos hierárquicos classificaram corretamente o conjunto de dados Ruspini, com exceção do método Ligação Completa. O número entre parênteses representa o ponto de parada para a formação dos grupos para os algoritmos hierárquicos. Os melhores resultados aparecem destacados em negrito.

Tabela 4.2: Ruspini: número de grupos.

			Silhueta			
	N^o Grupos	Obj/Grupo	Média	Variância	Rand	Rand Aj.
RGT	1	75	–	–	0.2465	0
	2	35,4	0.5827	0.0902	0.7510	0.5
	3	20,40,15	0.6414	0.1377	0.8591	0.6817
	4	**20,23,17,15**	**0.7376**	**0.10091**	**1**	**1**
	5	20,23,14,3,15	0.7135	0.1343	0.9850	0.9584
	6	20,23,1,13,3,15	0.6274	0.2826	0.9452	0.9452
K-médias	4				0.948 (0.0734)	0.8248 (0.2608)
Ward	**4 (500)**	20,23,17,15			**1**	**1**
L.Simples	**4 (20)**	20,23,17,15			**1**	**1**
L.Completa	4 (74)	20,20,20,15			0.96	0.8918
L.Média	**4 (47)**	20,23,17,15			**1**	**1**

Pode-se ver na figura 4.12 os objetos do conjunto de dados Ruspini e suas respectivas silhuetas de acordo com a partição gerada pelo algoritmo RGT. Símbolos de mesma cor representam objetos pertencentes ao mesmo grupo.

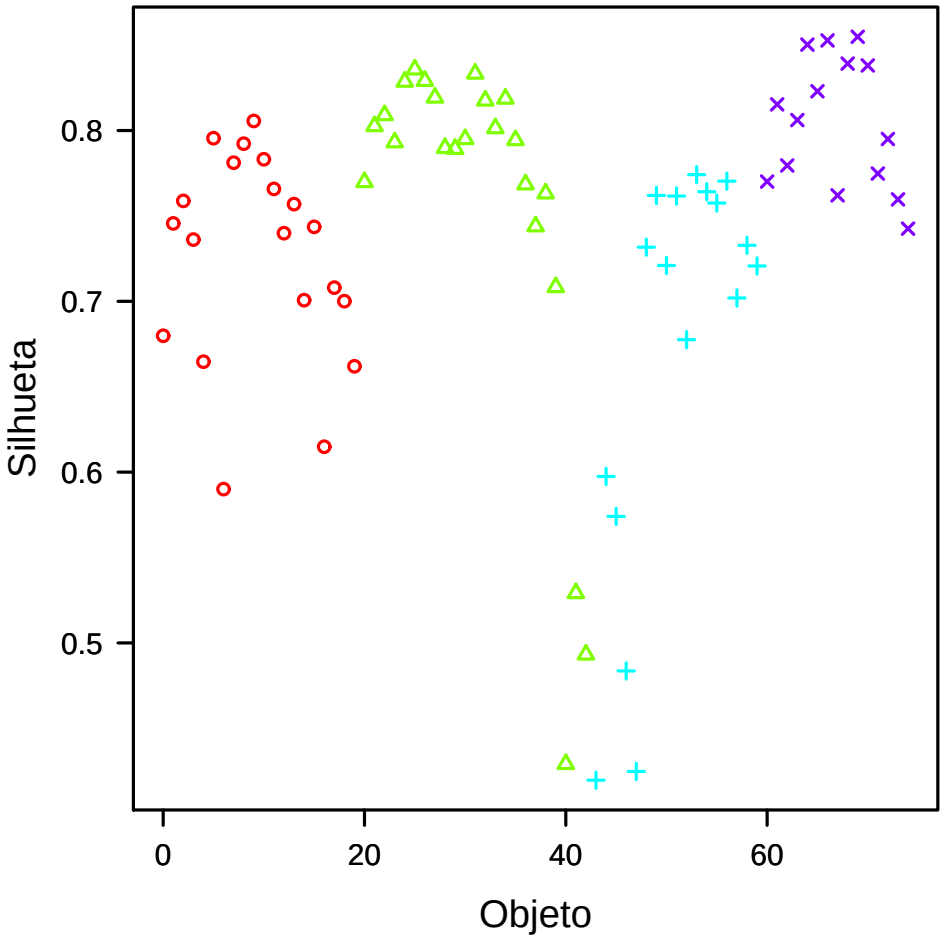

Figura 4.12: Ruspini: silhueta dos objetos para a partição ótima (4 grupos).

Pode-se ver na figura 4.13 a representação da partição ótima gerada pelo algoritmo RGT para o conjunto de dados Ruspini. Objetos pertencentes ao mesmo grupo têm mesma cor e estão conectados. Objetos de grupos distintos têm cores diferentes e não estão conectados.

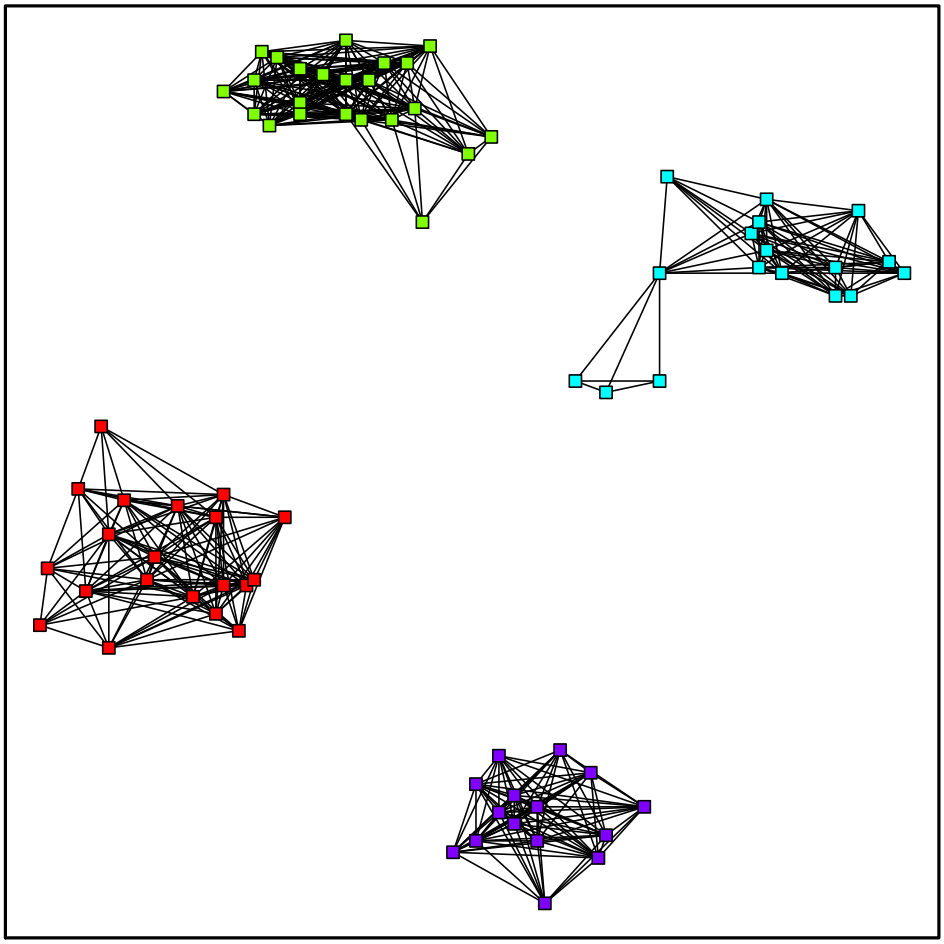

Figura 4.13: Ruspini: algoritmo RGT formando a partição ótima (4 grupos).

A seguir, podem-se ver algumas partições do conjunto de dados Ruspini geradas pelo algoritmo K-médias. O símbolo * representa os centróides dos grupos:

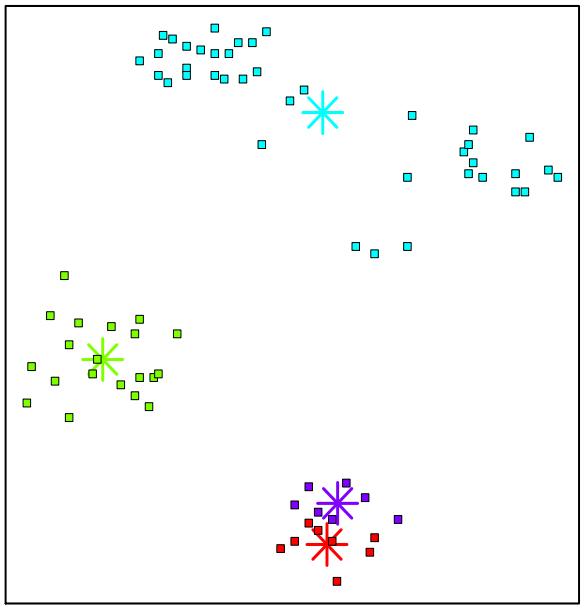

Figura 4.14: Ruspini: partição (1) gerada pelo algoritmo K-médias com 4 grupos.

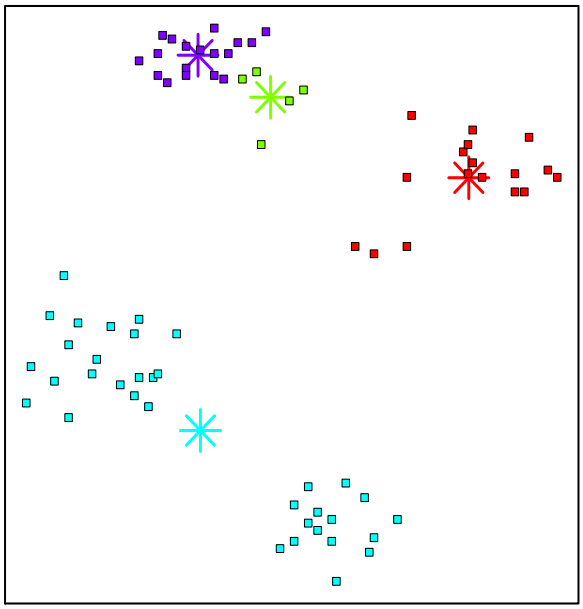

Figura 4.15: Ruspini: partição (2) gerada pelo algoritmo K-médias com 4 grupos.

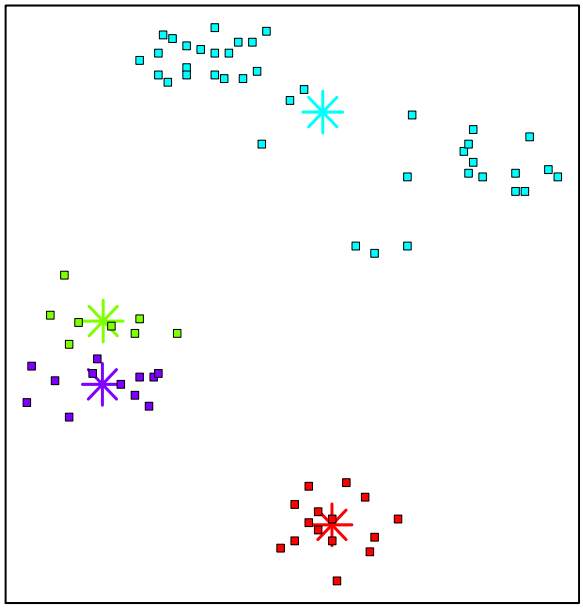

Figura 4.16: Ruspini: partição (3) gerada pelo algoritmo K-médias com 4 grupos.

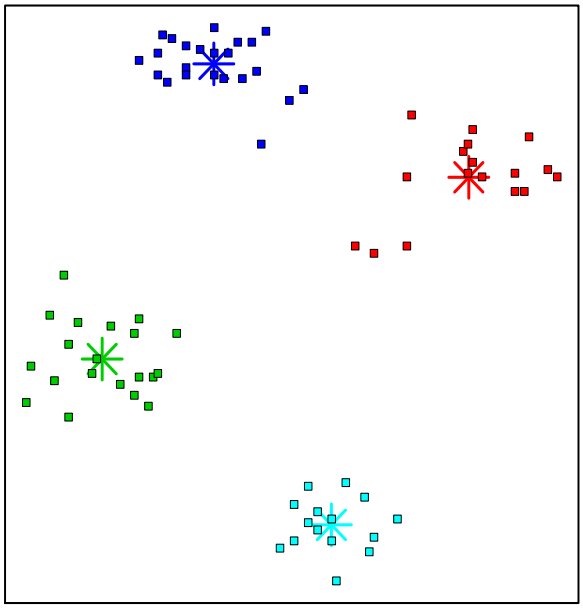

Figura 4.17: Ruspini: partição (4) gerada pelo algoritmo K-médias com 4 grupos.

A partir das figuras 4.18 à 4.21 têm-se dendogramas que ilustram o agrupamento hierárquico aglomerativo para o conjunto de dados Ruspini:

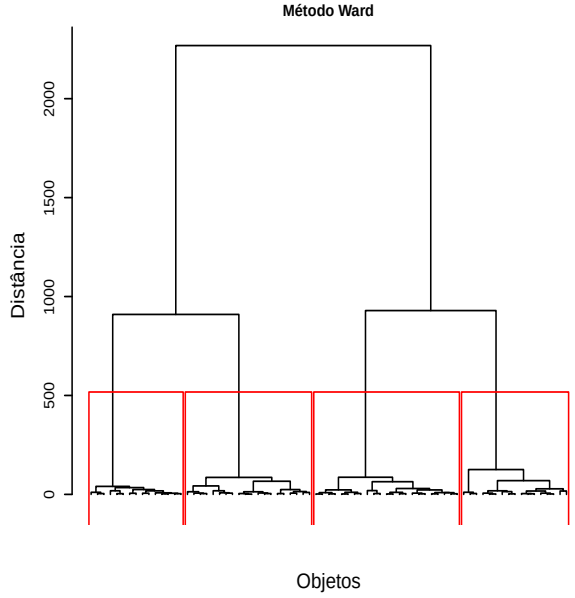

Figura 4.18: Ruspini: ponto de parada 500.

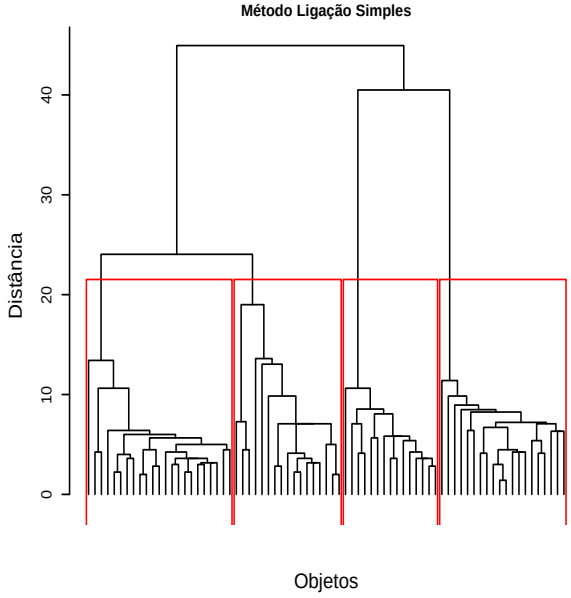

Figura 4.19: Ruspini: ponto de parada 20.

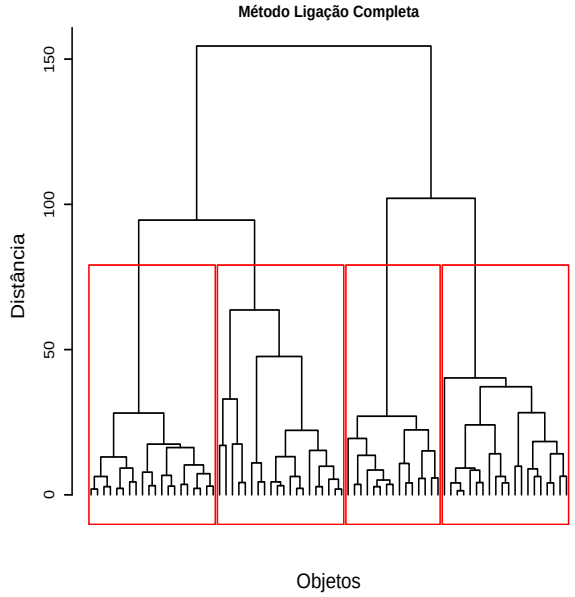

Figura 4.20: Ruspini: ponto de parada 74.

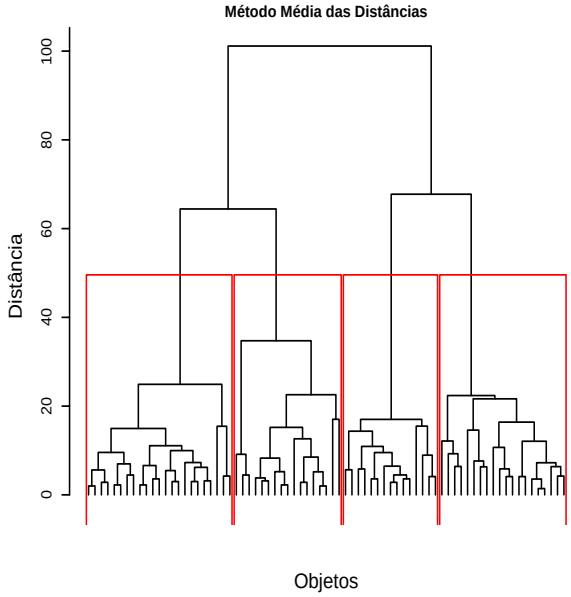

Figura 4.21: Ruspini: ponto de parada 47.

4.2 Espiral222-2D2C

O conjunto de dados Espiral222-2D2C é composto por 222 objetos bi-dimensionais. Trata-se de um conjunto determinístico. Dois grupos compõem sua estrutura. Cada grupo é formado por 111 objetos como ilustra a figura 4.22:

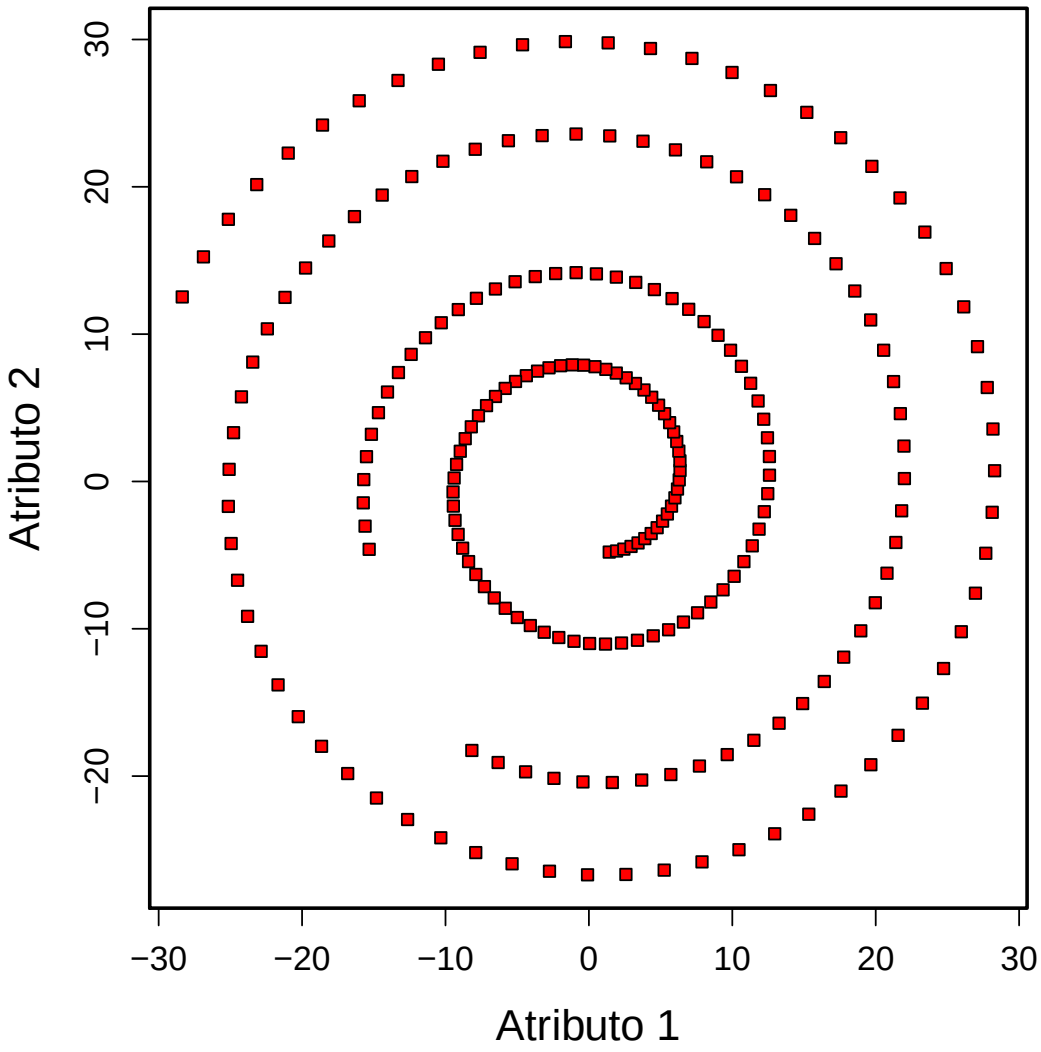

Figura 4.22: Representação do conjunto de dados Espiral222–2D2C.

A seguir, podem-se ver nas figuras 4.23 e 4.24 os histogramas dos Filtros 0 e Filtro 1 aplicados ao conjunto de dados Espiral222–2D2C:

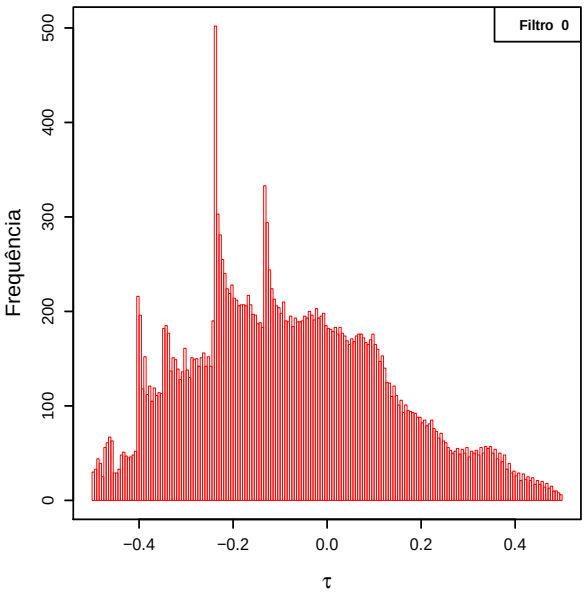

Figura 4.23: Espiral222–2D2C: filtro 0.

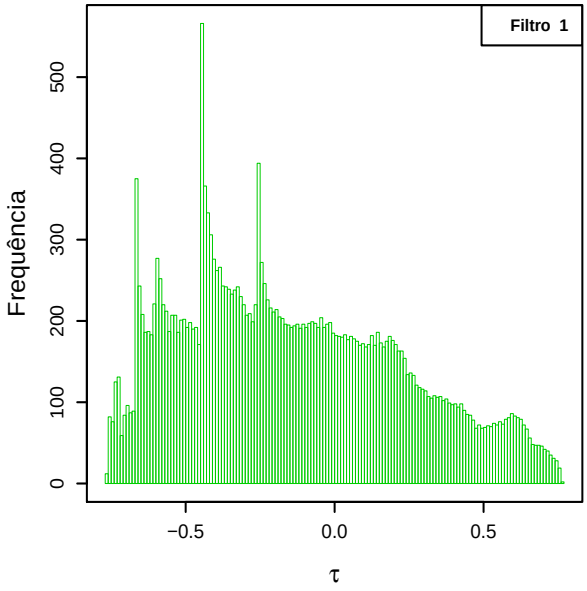

Figura 4.24: Espiral222–2D2C: filtro 1.

Na figura 4.25 pode-se ver o histograma do filtro 2 e o limiar de ativação das conexões que interligam os objetos do conjunto de dados Espiral222–2D2C. A figura 4.26 ilustra cada objeto do conjunto de dados Espiral222–2D2C e seu respectivo rótulo:

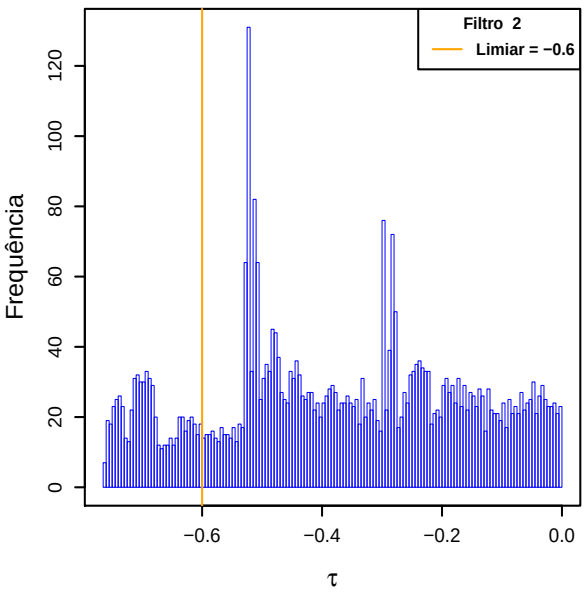

Figura 4.25: Espiral222–2D2C: filtro 2.

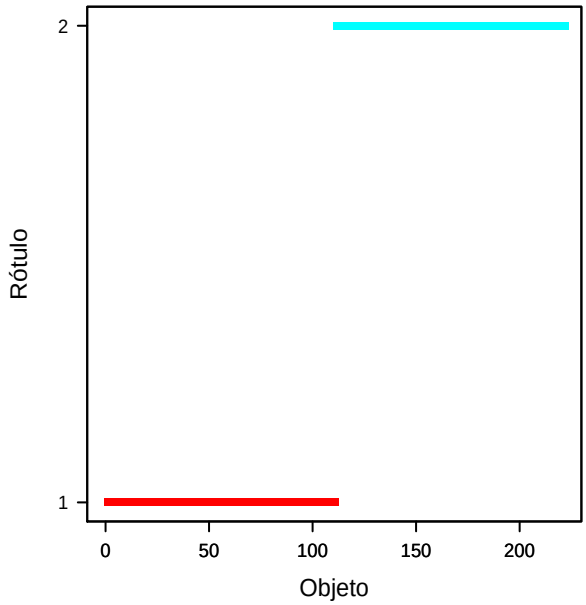

Figura 4.26: Espiral222–2D2C: rótulos dos objetos, formando 2 grupos.

Podem-se ver na tabela 4.3 as variações da Silhueta Média, Índice de Rand e Índice de Rand Ajustado em função da partição gerada pelo algoritmo RGT. O algoritmo RGT classificou corretamente o conjunto de dados Espiral222–2D2C, enquanto o algoritmo K-médias não. Entre os algoritmos de agrupamento hierárquico, apenas o método da Ligação Simples apresentou resultado relevante, para ponto de parada em 6.

Tabela 4.3: Espiral222–2D2C: número de grupos.

			Silhueta			
	N^o Grupos	Obj/Grupo	Média	Variância	Rand	Rand Aj.
RGT	1	222	–	–	0.4977	0
	2	**111,111**	**0.1451**	**0.3366**	**1**	**1**
	5	109,110,1,1,1	-0.0962	0.4043	0.9866	0.9732
	32	82,110,1, ... ,1	-0.3041	0.4933	0.8820	0.7638
K-médias	2				0.4990	-0.0019
					(0.0002)	(0.0006)
Ward	2 (863)	143,79			0.5069	0.014
L.Simples	**2 (6)**	**111,111**			**1**	**1**
L.Completa	2 (56)	142,80			0.511	0.022
L.Média	2 (32)	183,39			0.5597	0.1211

Podem-se ver na figura 4.27 os objetos do conjunto de dados Espiral222–2D2C e suas respectivas silhuetas de acordo com a partição gerada pelo algoritmo RGT. Símbolos de mesma cor representam objetos pertencentes ao mesmo grupo.

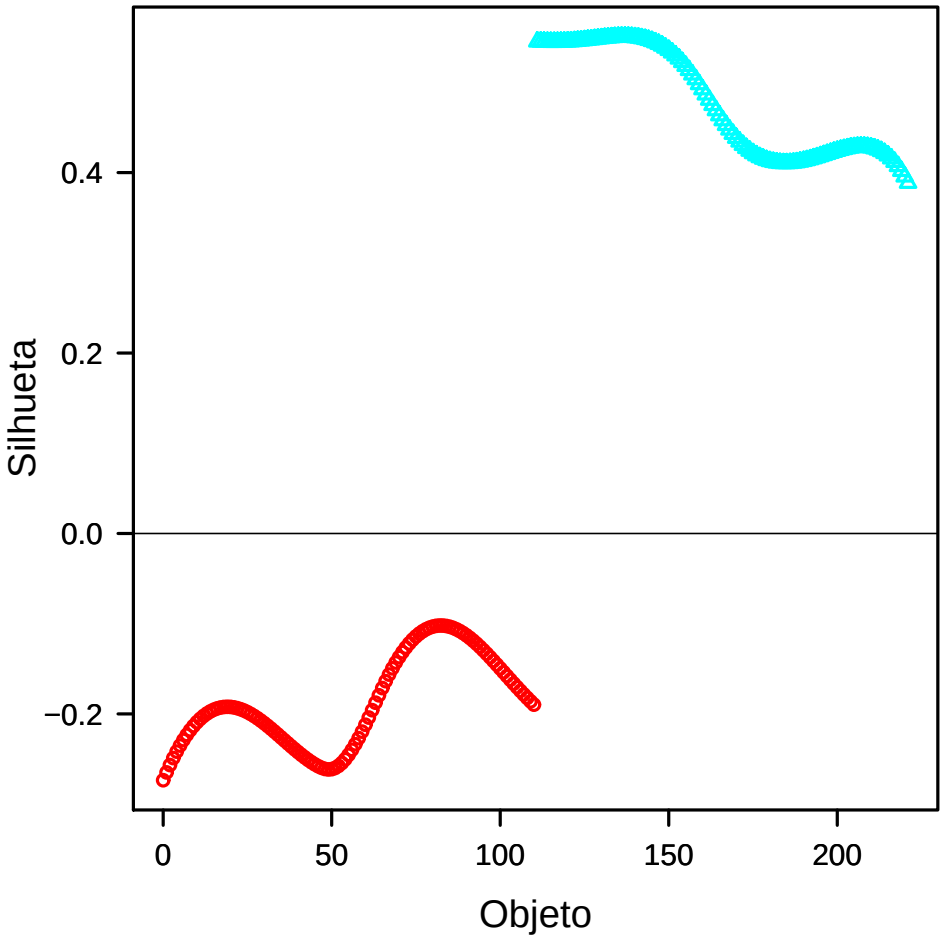

Figura 4.27: Espiral222–2D2C: silhueta dos objetos para a partição ótima (2 grupos).

Pode-se ver na figura 4.28 a representação da partição ótima gerada pelo algoritmo RGT para o conjunto de dados Espiral222–2D2C. Objetos pertencentes ao mesmo grupo têm mesma cor e estão conectados. Objetos de grupos distintos têm cores diferentes e não estão conectados.

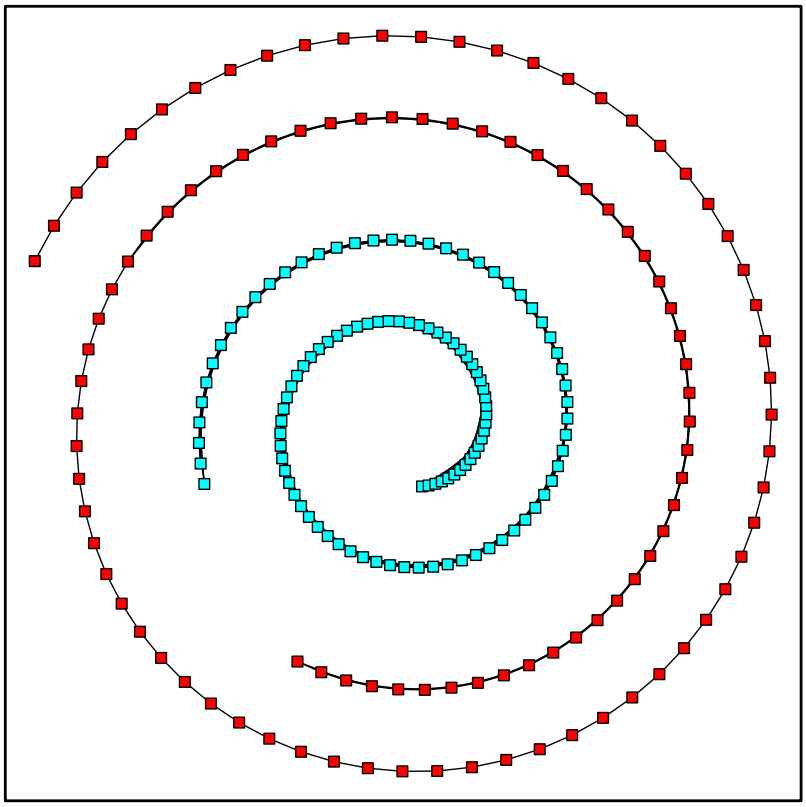

Figura 4.28: Espiral222–2D2C: algoritmo RGT formando a partição ótima (2 grupos).

Pode-se ver uma partição gerada pelo algoritmo K-médias na figura 4.29. O símbolo * representa os centróides dos grupos:

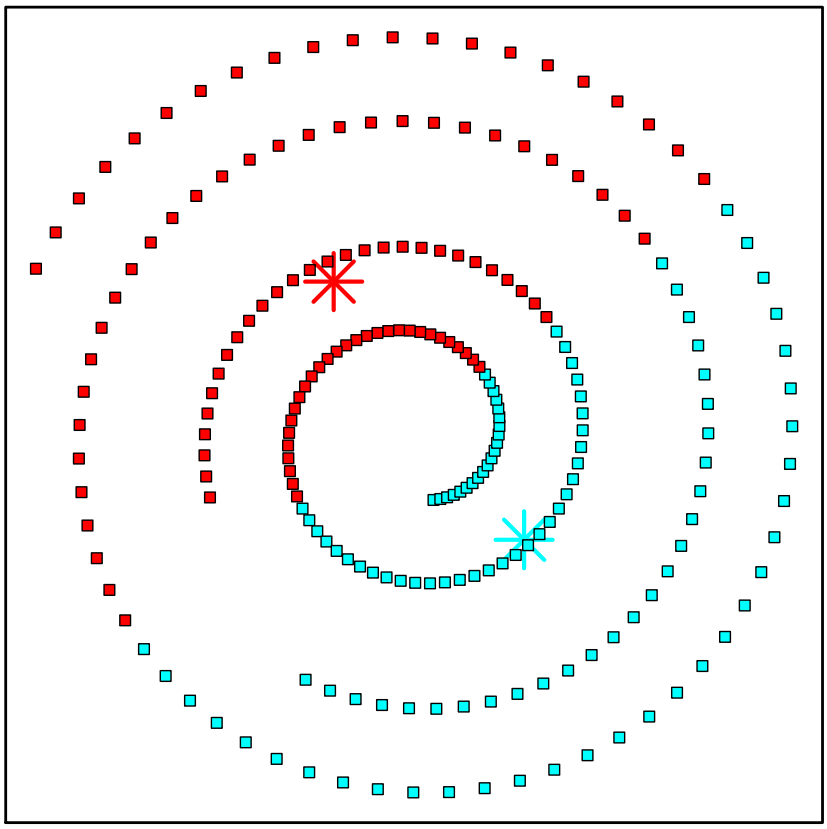

Figura 4.29: Espiral222–2D2C: partição gerada pelo algoritmo K-médias com 2 grupos.

A partir das figuras 4.30 à 4.33 têm-se dendogramas que ilustram o agrupamento hierárquico aglomerativo para o conjunto de dados Espiral222-2D2C:

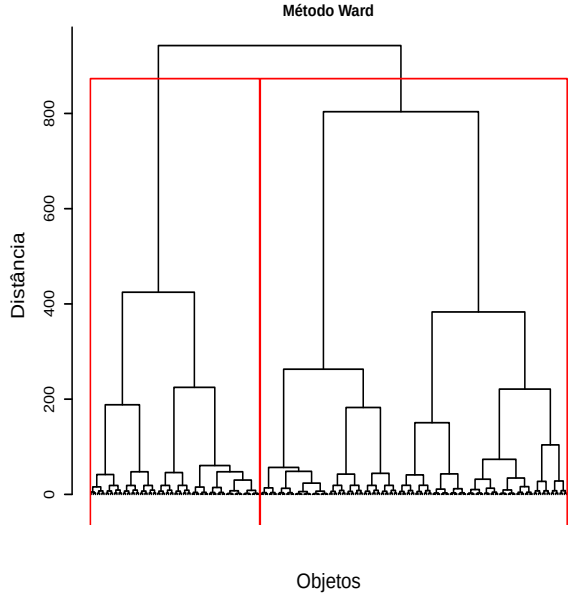

Figura 4.30: Espiral222-2D2C: ponto de parada 863.

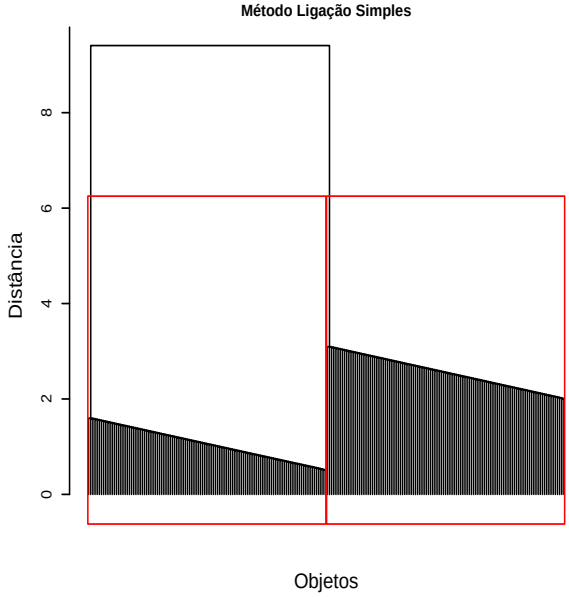

Figura 4.31: Espiral222-2D2C: ponto de parada 6.

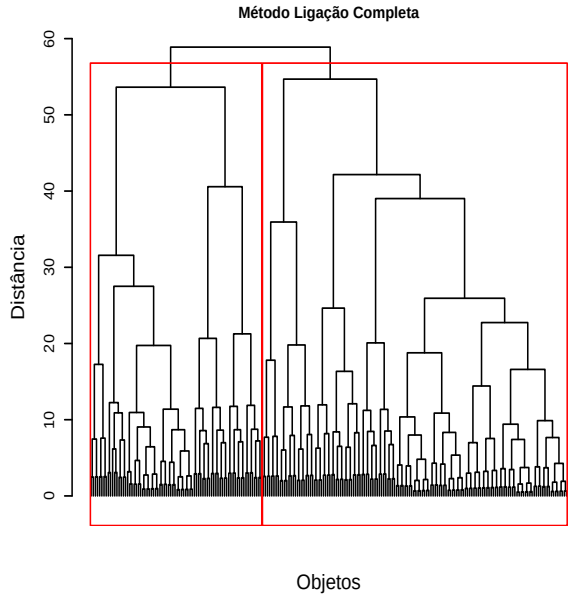

Figura 4.32: Espiral222-2D2C: ponto de parada 56.5.

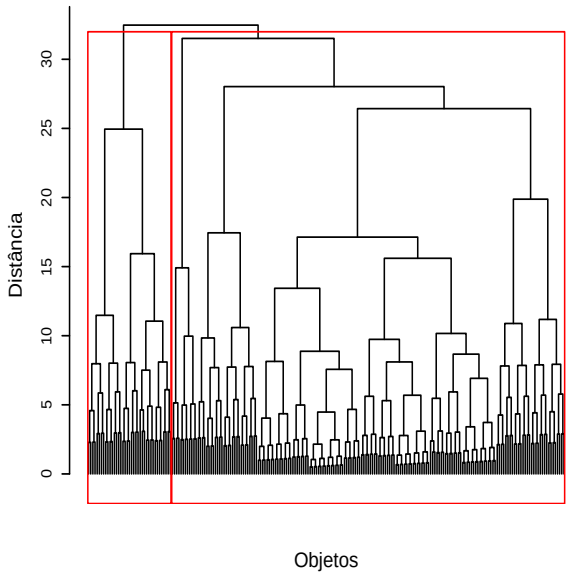

Figura 4.33: Espiral222-2D2C: ponto de parada 32.

4.3 Sobreviventes

O conjunto de dados Sobreviventes é composto por 306 objetos tri-dimensionais. Trata-se de um conjunto de dados de estudo de caso que foi realizado entre os anos de 1958 e 1970 no Hospital Billings da Universidade de Chicago. O estudo é relativo à sobrevida de pacientes submetidos à cirurgia de câncer de mama. Dois grupos compõem sua estrutura. O primeiro grupo é composto pelos pacientes que sobreviveram pelo menos 5 anos após a cirurgia (225 pacientes). O segundo grupo é composto pelos pacientes que morreram antes do quinto ano após a cirurgia (81 pacientes) [10].

Os três atributos desse conjunto de dados estão descritos a seguir:

i. Idade do paciente no momento da operação;

ii. Ano de operação do paciente;

iii. Número de nódulos detectados.

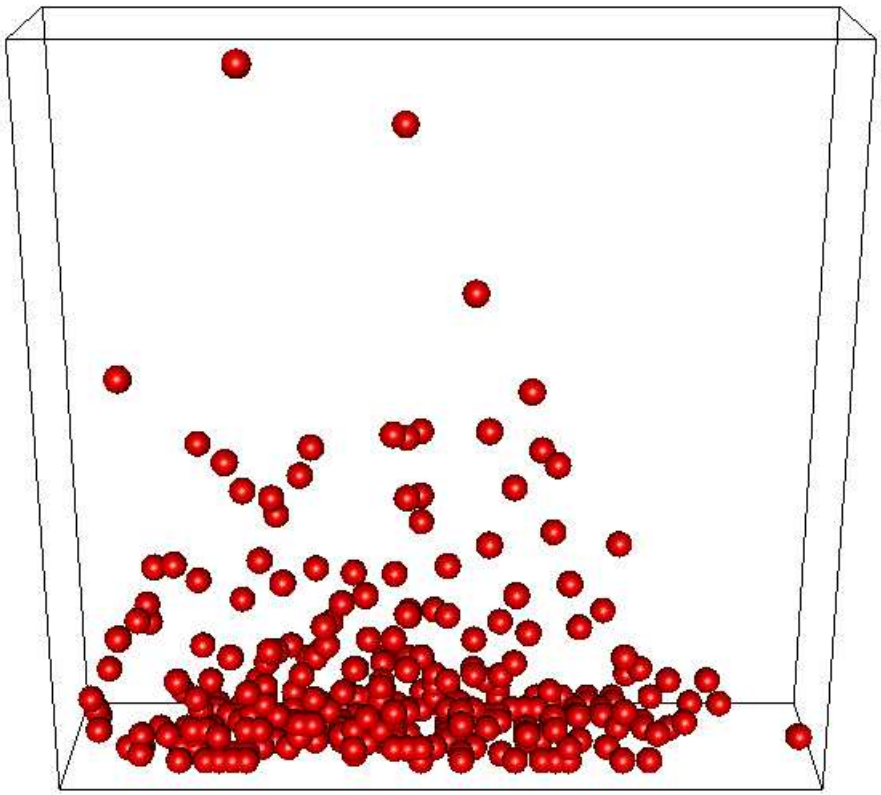

Figura 4.34: Representação do conjunto de dados Sobreviventes.

A seguir, podem-se ver nas figuras 4.35 à 4.42 os histogramas dos Filtros aplicados ao conjunto de dados Sobreviventes:

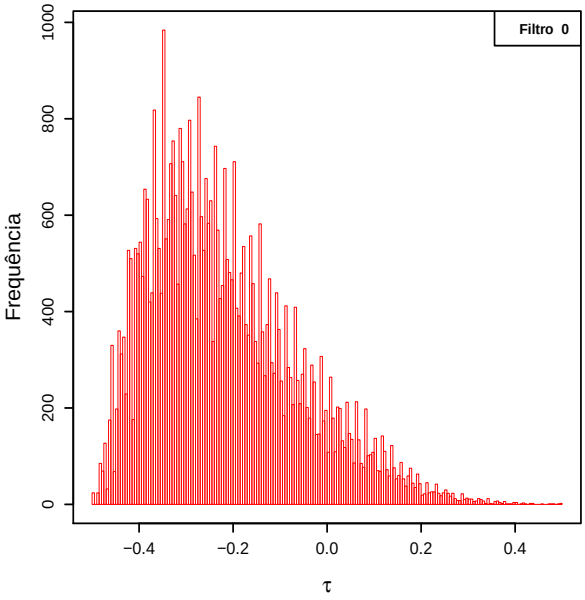

Figura 4.35: Sobreviventes: filtro 0.

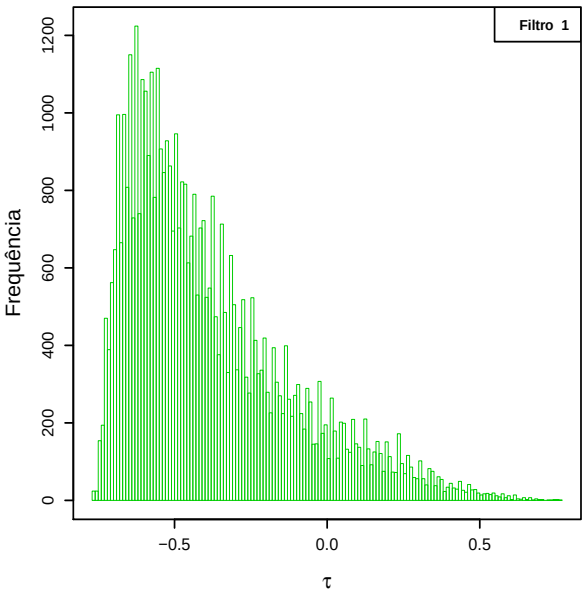

Figura 4.36: Sobreviventes: filtro 1.

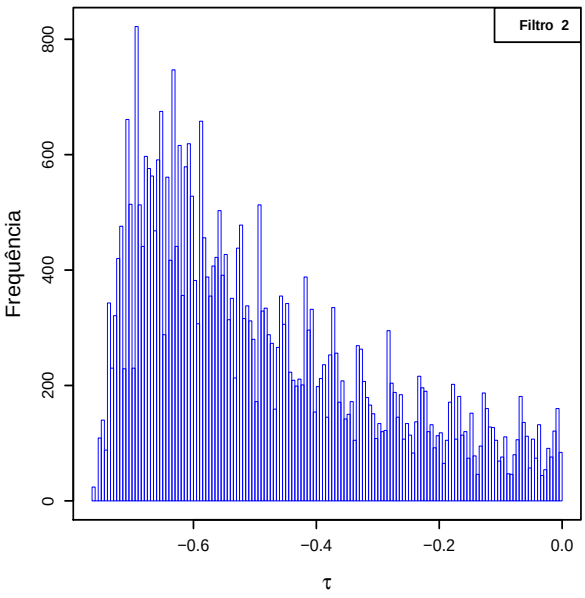

Figura 4.37: Sobreviventes: filtro 2.

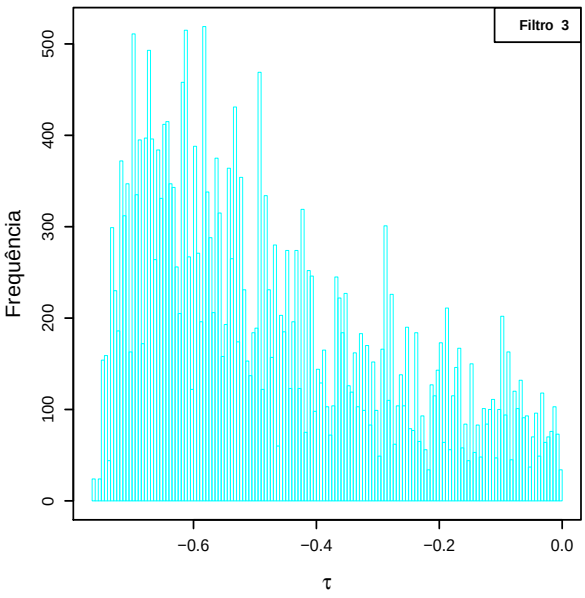

Figura 4.38: Sobreviventes: filtro 3.

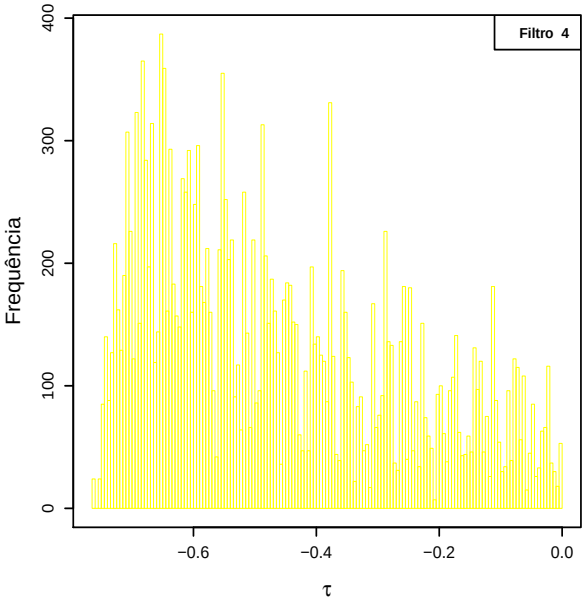

Figura 4.39: Sobreviventes: filtro 4.

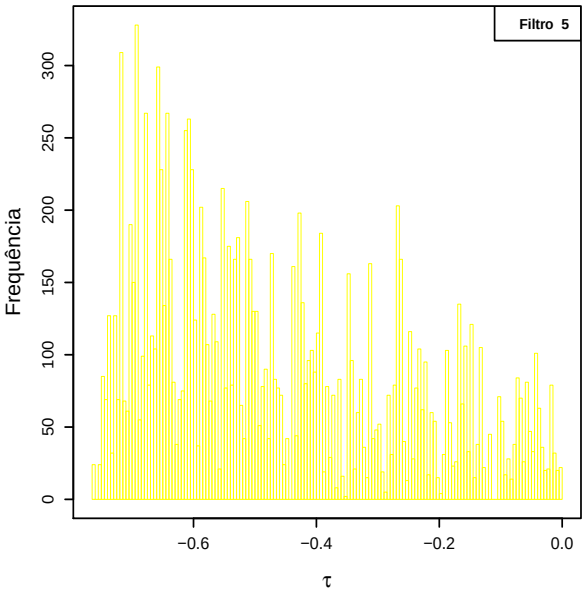

Figura 4.40: Sobreviventes: filtro 5.

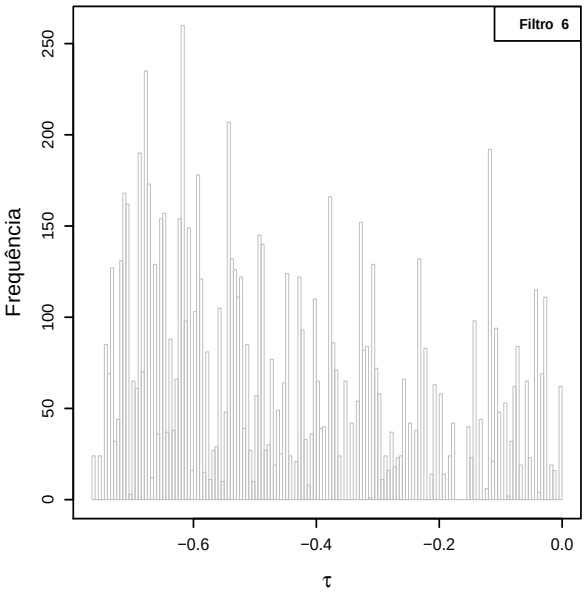

Figura 4.41: Sobreviventes: filtro 6.

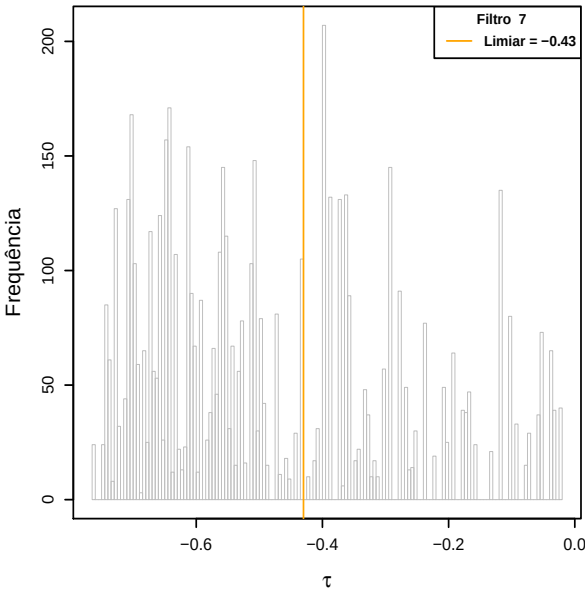

Figura 4.42: Sobreviventes: filtro 7.

A figura 4.43 ilustra cada objeto do conjunto de dados Sobreviventes e seu respectivo rótulo, através da partição gerada pelo algoritmo RGT:

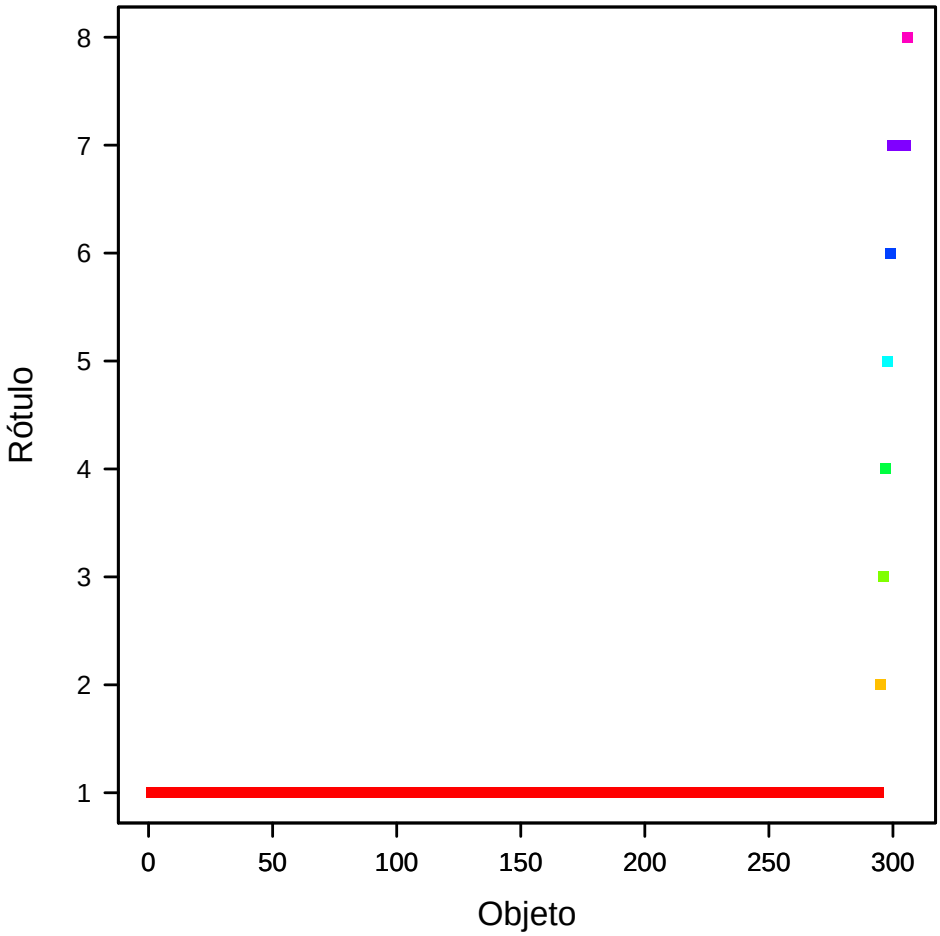

Figura 4.43: Sobreviventes: rótulos dos objetos formando 8 grupos.

Podem-se ver na figura 4.44 os objetos do conjunto de dados Sobreviventes e suas respectivas silhuetas de acordo com a partição gerada pelo algoritmo RGT. Símbolos de mesma cor representam objetos pertencentes ao mesmo grupo.

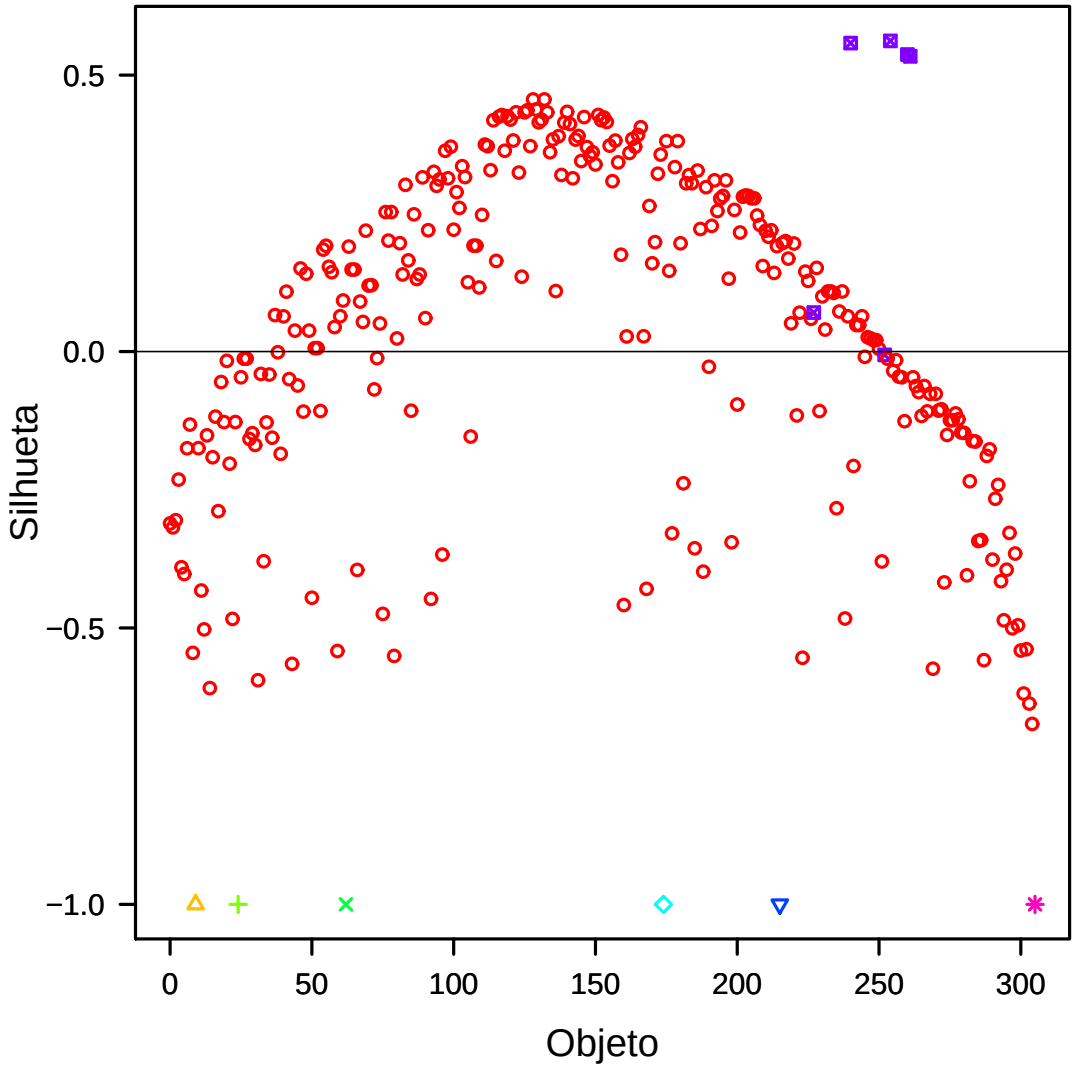

Figura 4.44: Sobreviventes: silhueta dos objetos pelo algoritmo RGT (8 grupos).

Podem-se ver na tabela 4.4 as variações da Silhueta Média, Índice de Rand e Índice de Rand Ajustado em função da partição gerada pelo algoritmo RGT. O melhor resultado é dado pelo algoritmo RGT, quando se formam 8 grupos. O primeiro grupo é formado por 294 objetos. O segundo, por 6 objetos e os 6 grupos restantes são *singletons*. As 20 experiências do algoritmo K-médias mostraram um desempenho inferior ao algoritmo RGT. Entre os algoritmos hierárquicos, apenas o método de Ligação Simples produziu resultados próximos, porém inferiores aos resultados do algoritmo RGT.

Tabela 4.4: Sobreviventes: número de grupos.

			Silhueta			
	N^o Grupos	Obj/Grupo	Média	Variância	Rand	Rand Aj.
	7	300,1,1,...,1	0.1016	0.3547	0.6157	0.037
RGT	**8**	**294,6,1,...,1**	**0.026**	**0.3215**	**0.6157**	**0.0566**
	19	269,6,7,5,2, 3,2,1, ... ,1	-0.2582	0.3502	0.6068	0.1066
K-médias	2				0.4990 (0.0003)	-0.0032 (0.001)
Ward	2 (802.7)	67,239			0.5117	-0.0485
L.Simples	2 (14.6)	305,1			0.6125	0.0116
L.Completa	2 (60.6)	229,77			0.5344	0.0162
L.Média	2 (32.7)	302,4			0.6094	0.0143

A seguir, pode-se ver na figura 4.45 a partição gerada pelo algoritmo RGT:

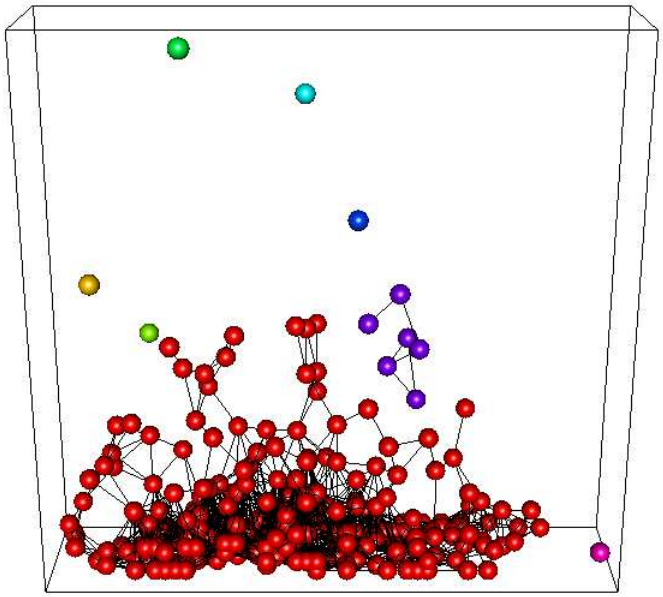

Figura 4.45: Sobreviventes: partição do algoritmo RGT formando 8 grupos.

A figura 4.46 ilustra uma partição gerada através do algoritmo K-médias. O símbolo * representa os centróides dos grupos:

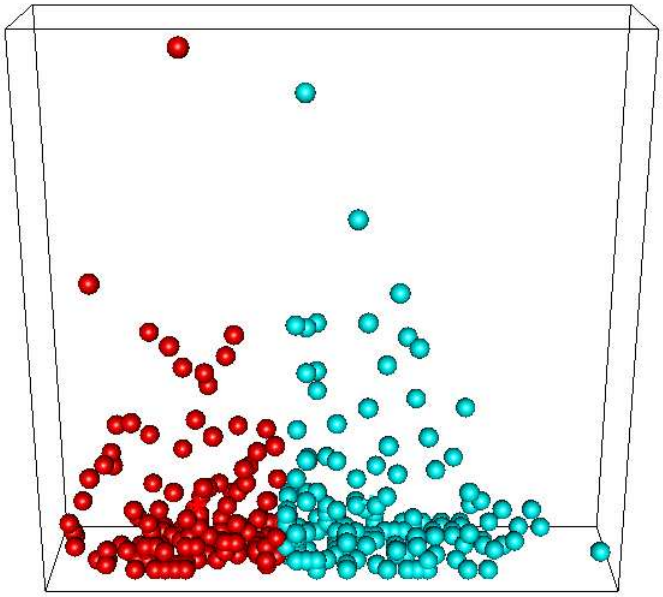

Figura 4.46: Sobreviventes: partição do algoritmo K-médias com 2 grupos.

A partir das figuras 4.47 à 4.50 têm-se dendogramas que ilustram o agrupamento hierárquico aglomerativo para o conjunto de dados Sobreviventes:

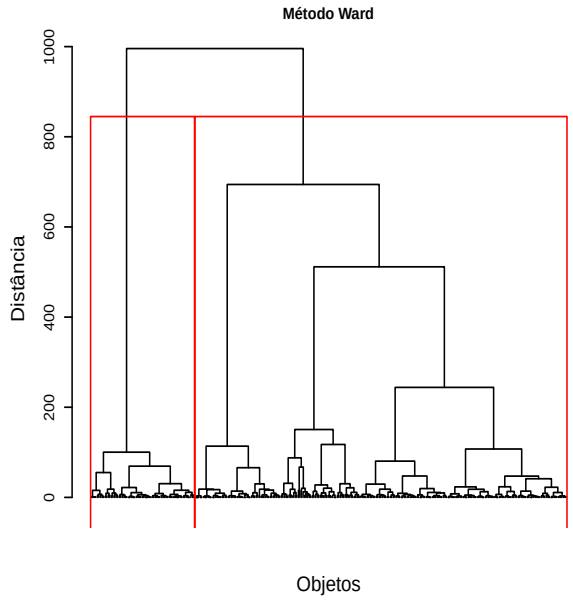

Figura 4.47: Sobreviventes: ponto de parada 802.7.

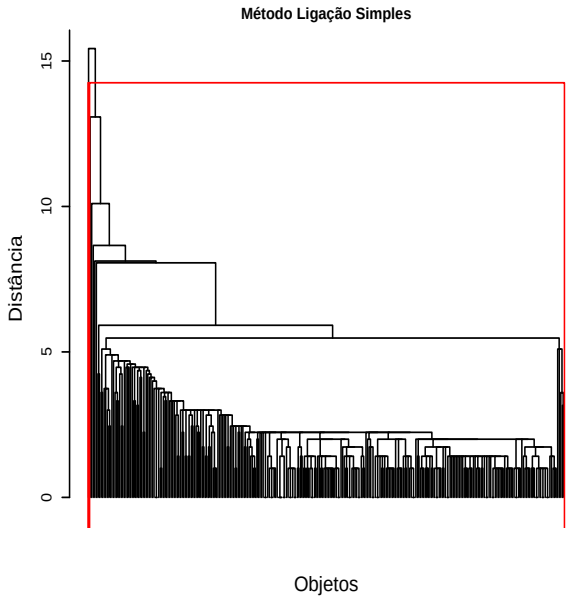

Figura 4.48: Sobreviventes: ponto de parada 14.6.

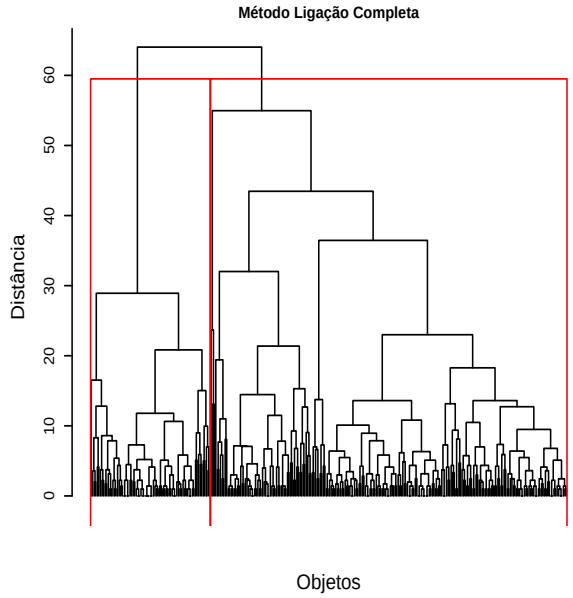

Figura 4.49: Sobreviventes: ponto de parada 60.6.

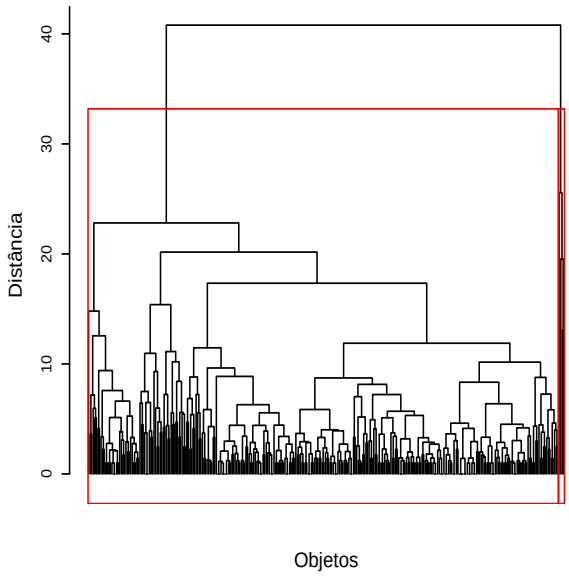

Figura 4.50: Sobreviventes: ponto de parada 32.7.

4.4 Íris

O conjunto de dados Íris [17] contém 3 grupos de 50 objetos cada, totalizando 150 objetos. Cada grupo representa um tipo diferente de planta Íris que são chamadas de Íris Setosa, Íris Versicolor e Íris Virginica. O grupo da Íris Setosa é linearmente separado dos grupos das outras duas. Porém os grupos das Íris Versicolor e Íris Virginica não são separados linearmente. Os quatro atributos de cada objeto são o comprimento da sépala, a largura da sépala, o comprimento da pétala e a largura da pétala [10].

Figura 4.51: Íris Setosa.

Figura 4.52: Íris Versicolor.

Figura 4.53: Íris Virginica.

A figura 4.54 ilustra três dos quatro atributos do conjunto de dados Íris e a linearidade entre os grupos Íris Versicolor (verde) e Íris Virginica (azul).

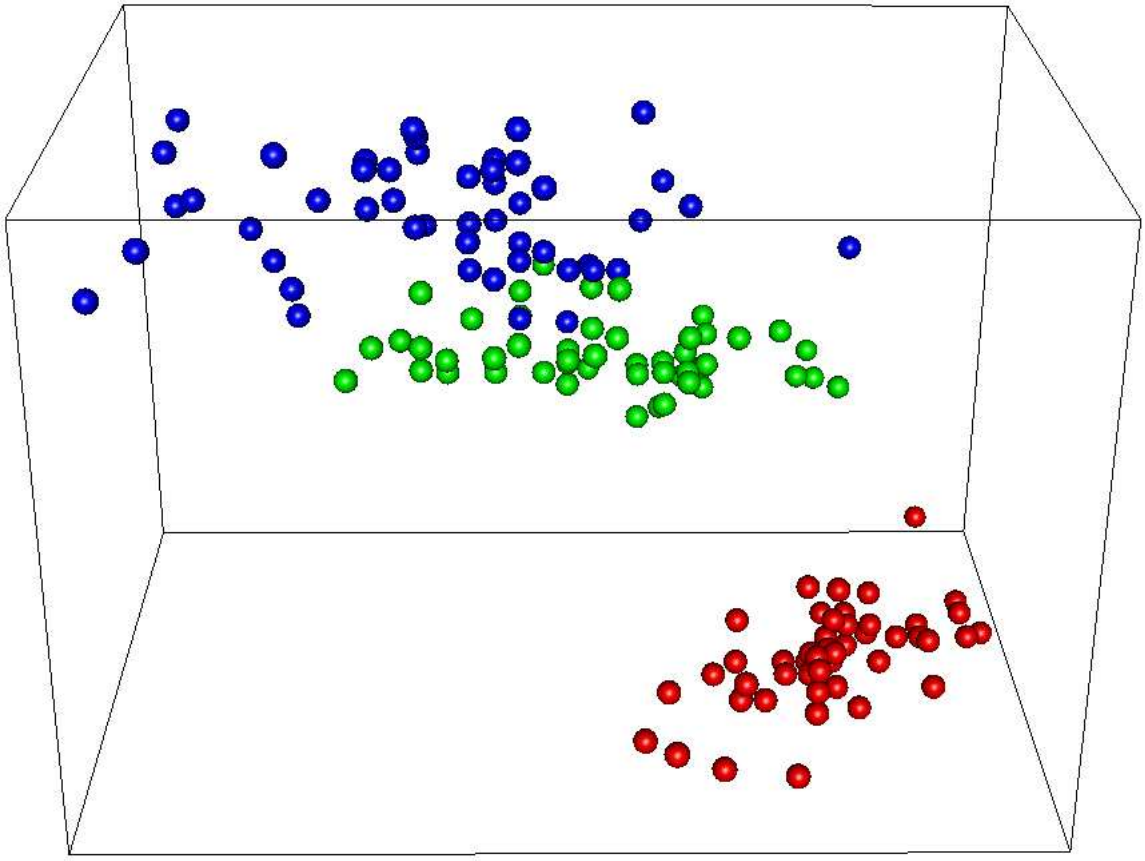

Figura 4.54: Íris: Representação de 3 dos 4 atributos.

A seguir, podem-se ver nas figuras 4.55 à 4.59 os histogramas dos Filtros 0, 1, 2, 3 e 4 aplicados ao conjunto de dados Íris:

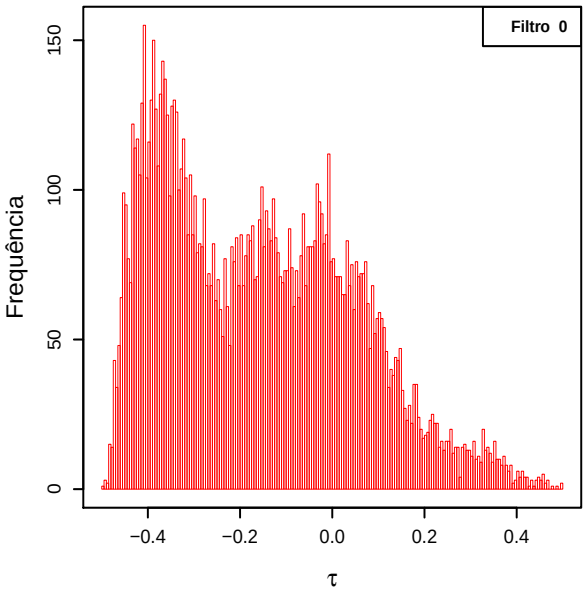

Figura 4.55: Íris: filtro 0.

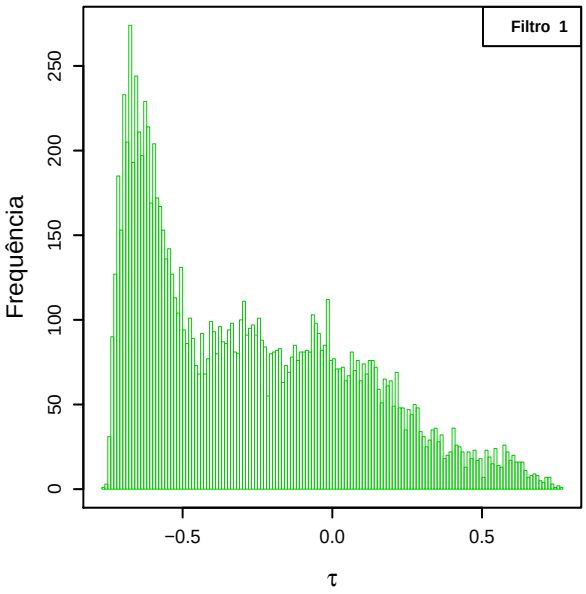

Figura 4.56: Íris: filtro 1.

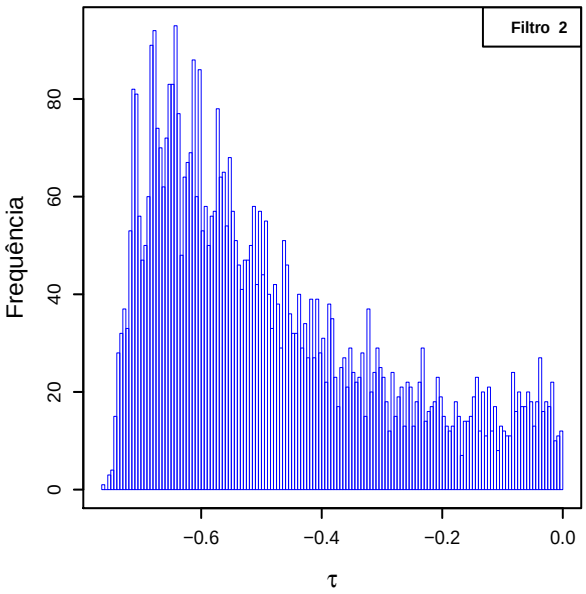

Figura 4.57: Íris: filtro 2.

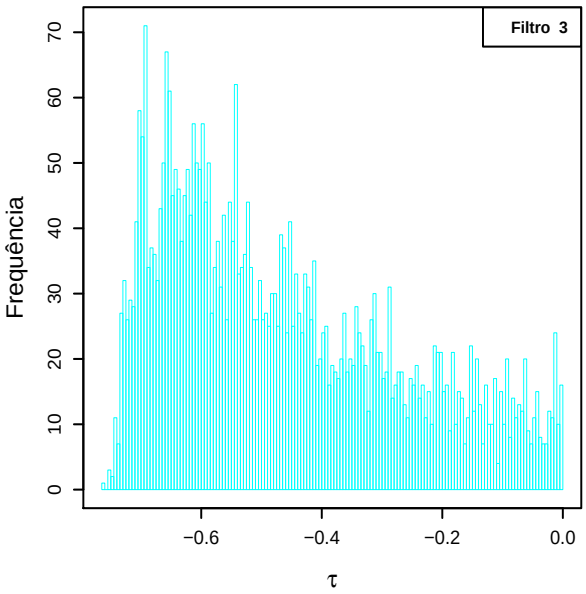

Figura 4.58: Íris: filtro 3.

A figura 4.60 ilustra cada objeto do conjunto de dados Íris e seu respectivo rótulo:

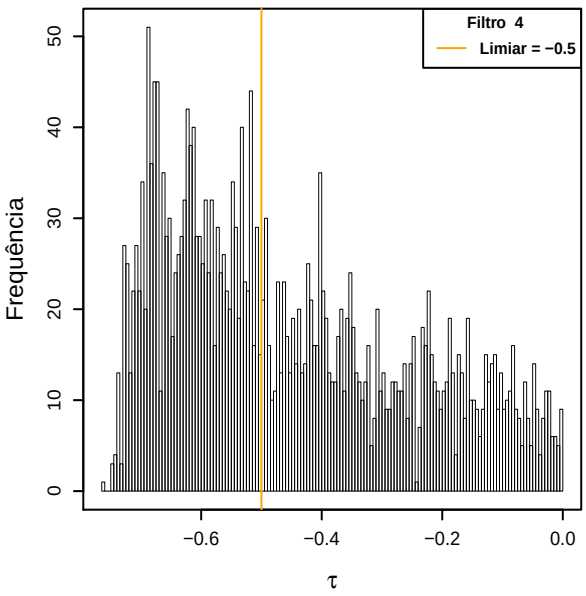

Figura 4.59: Íris: filtro 4.

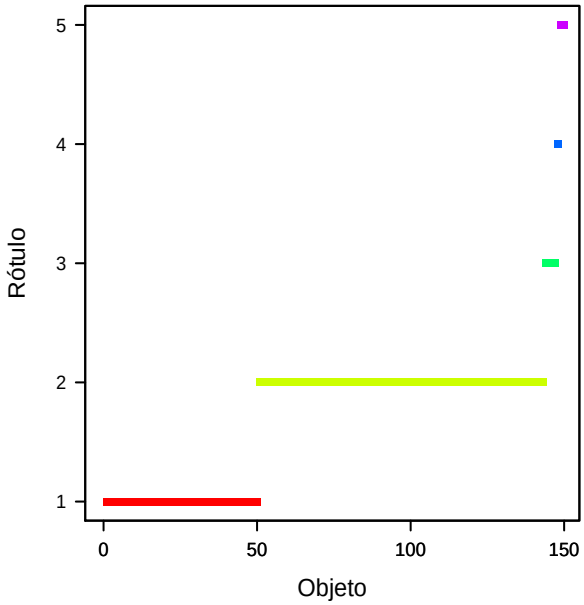

Figura 4.60: Íris: rótulos dos objetos formando 5 grupos.

Podem-se ver na tabela 4.5 as variações da Silhueta Média, Índice de Rand e Índice de Rand Ajustado em função da partição gerada pelo algoritmo RGT. O melhor resultado é dado pelos métodos hierárquicos Ward e Ligação Média. O algoritmo RGT se mostrou inferior devido aos grupos das Íris Versicolor e Íris Virginica não serem separados linearmente. Devido a sobreposição desses dois grupos, o critério da silhueta não funciona bem, e atinge valor máximo exatamente quando há dois grupos, o primeiro com 50 objetos e o segundo com 100 objetos, resultado da aglutinação desses dois grupos não separados linearmente.

Tabela 4.5: Íris: número de grupos.

			Silhueta			
	N^o Grupos	Obj/Grupo	Média	Variância	Rand	Rand Aj.
RGT	2	50,100	0.6216	0.2020	0.7763	0.5681
	3	50,98,2	0.4915	0.3484	0.7766	0.5637
	4	49,1,97,3	0.40	0.285	0.7727	0.5522
	5	50,93,4,1,2,	0.3176	0.3821	0.7773	0.5524
	6	50,92,4,1,1,2	0.2229	0.4230	0.7773	0.5499
K-médias	3				0.8639	0.7005
					(0.0487)	(0.0915)
Ward	**3 (31.9)**	**50,64,36**			**0.8922**	**0.7592**
L.Simples	3 (0.78)	50,98,2			0.7766	0.5637
L.Completa	3 (3.6)	50,72,28			0.8367	0.6422
L.Média	**3 (1.9)**	**50,64,36**			**0.8922**	**0.7592**

Podem-se ver na figura 4.61 os objetos do conjunto de dados Íris e suas respectivas silhuetas de acordo com a partição gerada pelo algoritmo RGT. Objetos de mesma cor e mesmo símbolo pertencem ao mesmo grupo.

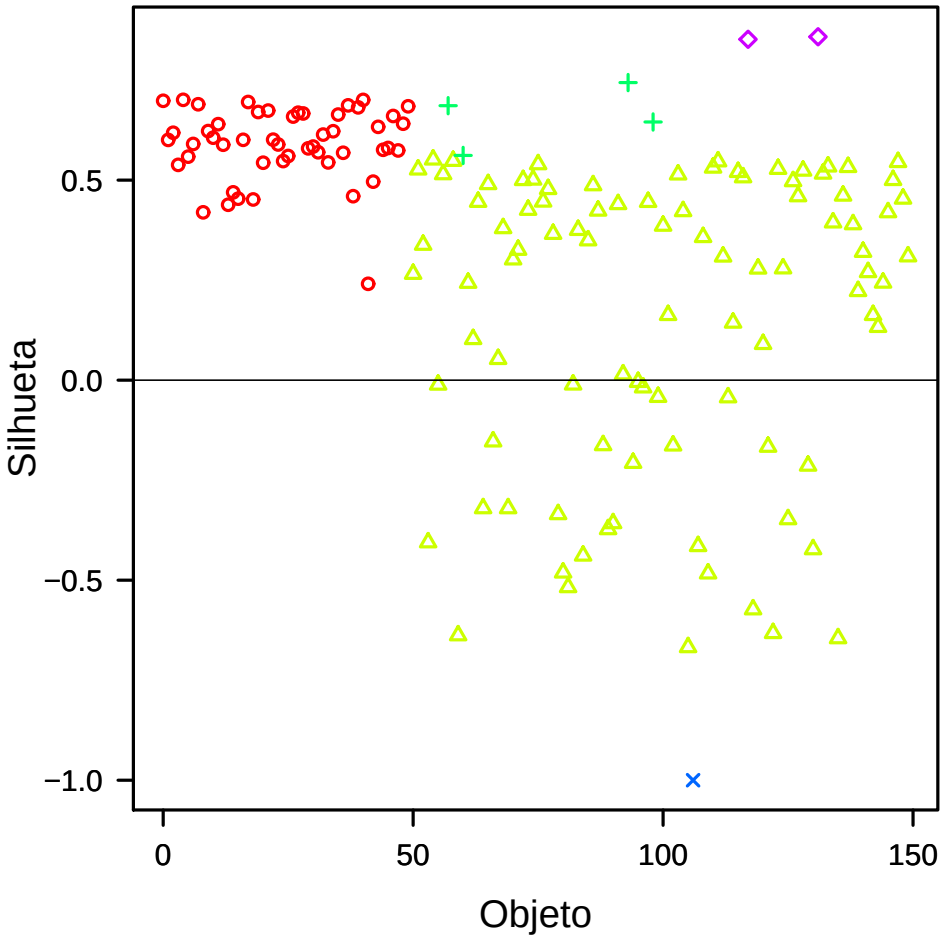

Figura 4.61: Íris: silhueta dos objetos pelo algoritmo RGT (5 grupos).

A partir das figuras 4.62 à 4.65 têm-se dendogramas que ilustram o agrupamento hierárquico aglomerativo para o conjunto de dados Íris:

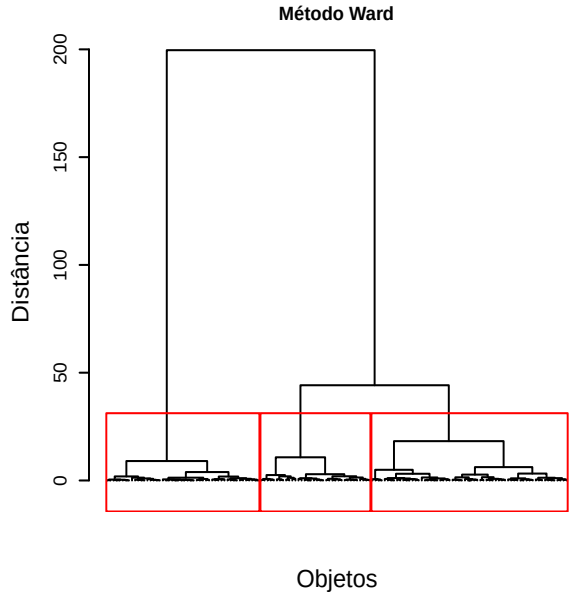

Figura 4.62: Íris: ponto de parada 31.9.

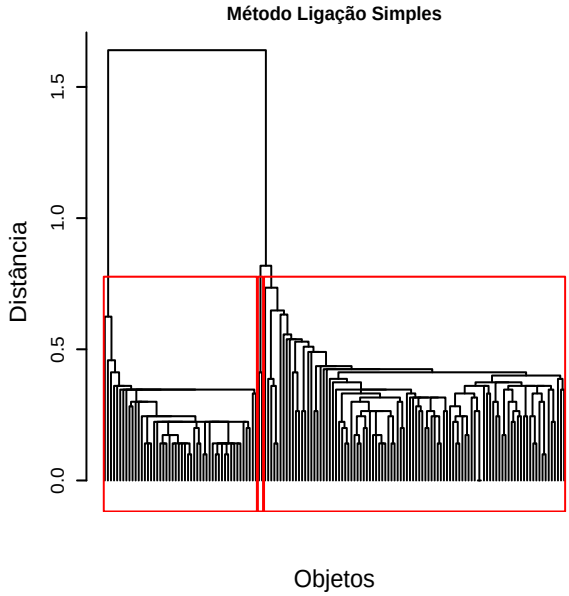

Figura 4.63: Íris: ponto de parada 0.78.

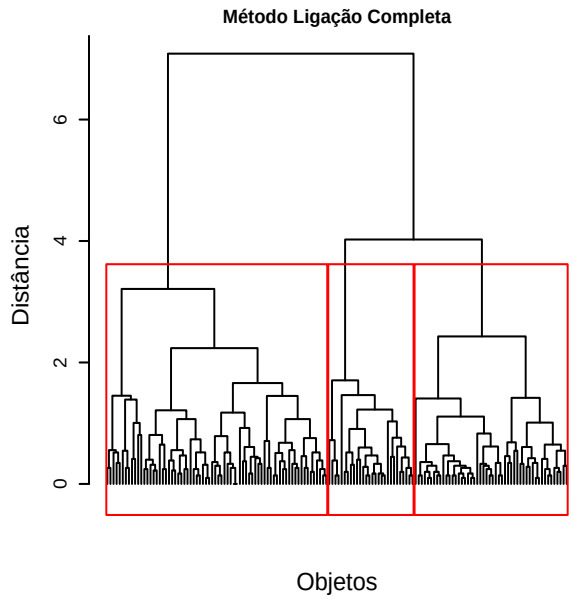

Figura 4.64: Íris: ponto de parada 3.6.

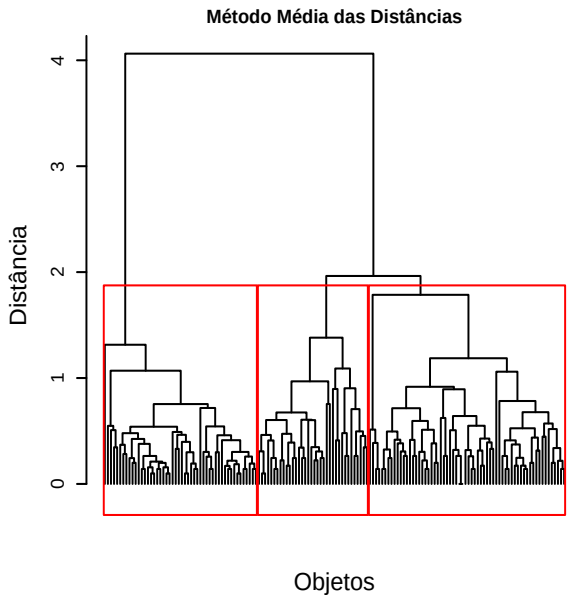

Figura 4.65: Íris: ponto de parada 1.9.

4.5 Wreath

O conjunto de dados Wreath é composto por 1000 objetos bi-dimensionais simulados a partir de um modelo de 14 grupos em que as matrizes de covariâncias são de mesmo tamanho e forma, mas diferem na orientação [39]. Cada grupo é formado por 74,69,63,74,68,70,71,66,83,77,66,77,61,81 objetos, como ilustra a figura 4.66:

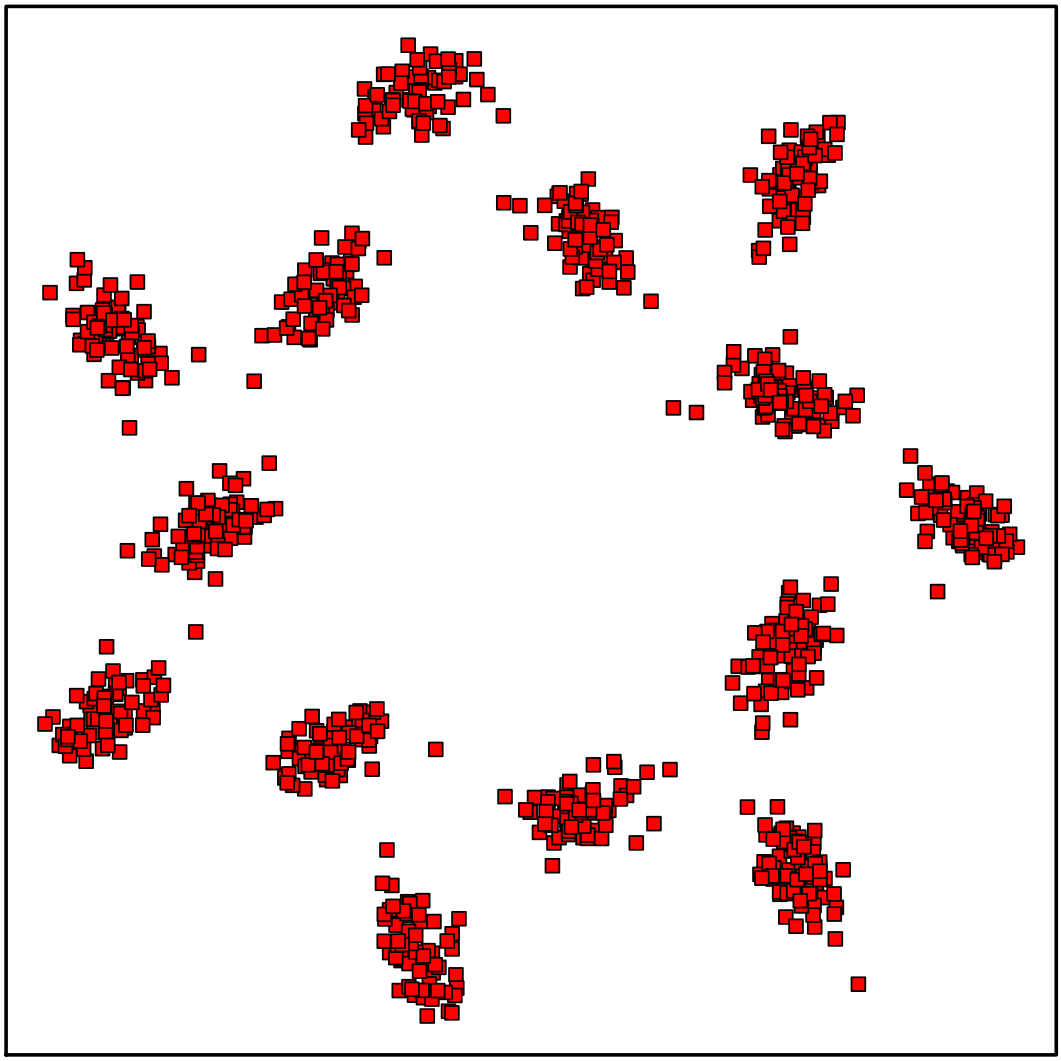

Figura 4.66: Representação do conjunto de dados Wreath.

A seguir, podem-se ver nas figuras 4.67 e 4.68 os histogramas dos Filtros 0 e Filtro 1 aplicados ao conjunto de dados Wreath:

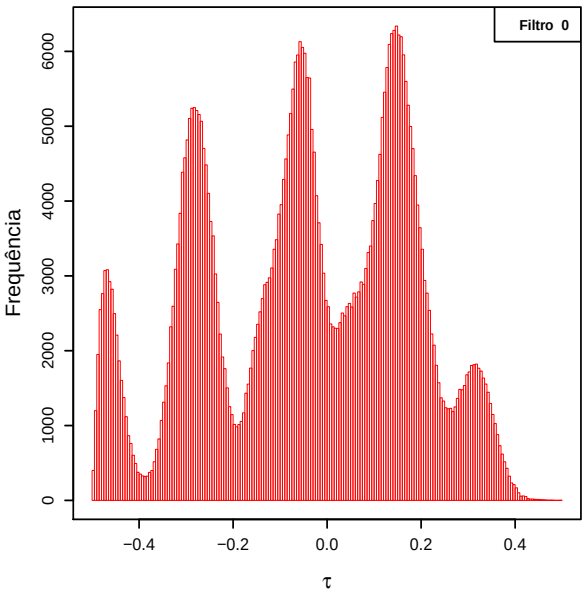

Figura 4.67: Wreath: filtro 0.

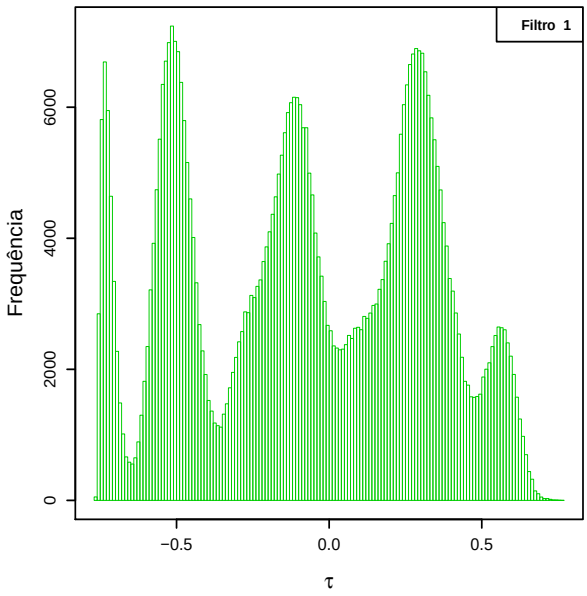

Figura 4.68: Wreath: filtro 1.

A seguir, podem-se ver nas figuras 4.69 e 4.70 os histogramas dos Filtros 2 e Filtro 3 aplicados e o limiar de ativação das conexões que interligam os objetos do conjunto de dados Wreath:

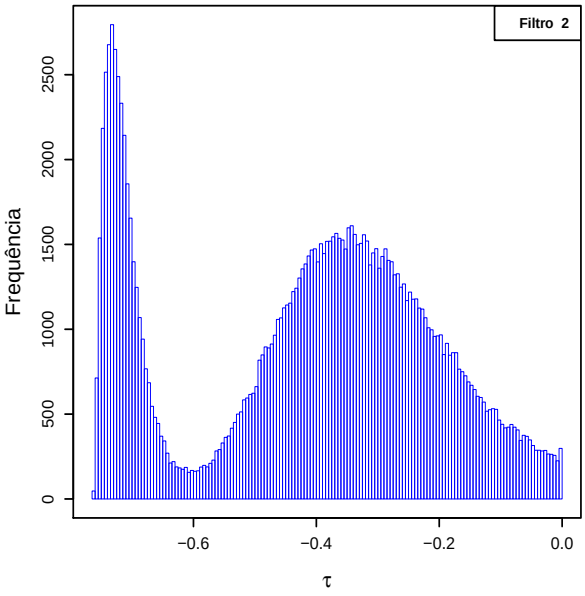

Figura 4.69: Wreath: filtro 2.

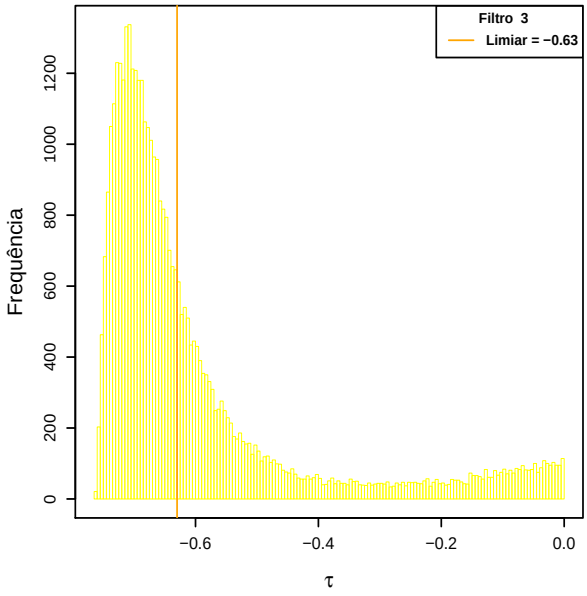

Figura 4.70: Wreath: filtro 3.

A figura 4.71 ilustra cada objeto do conjunto de dados Wreath e seu respectivo rótulo:

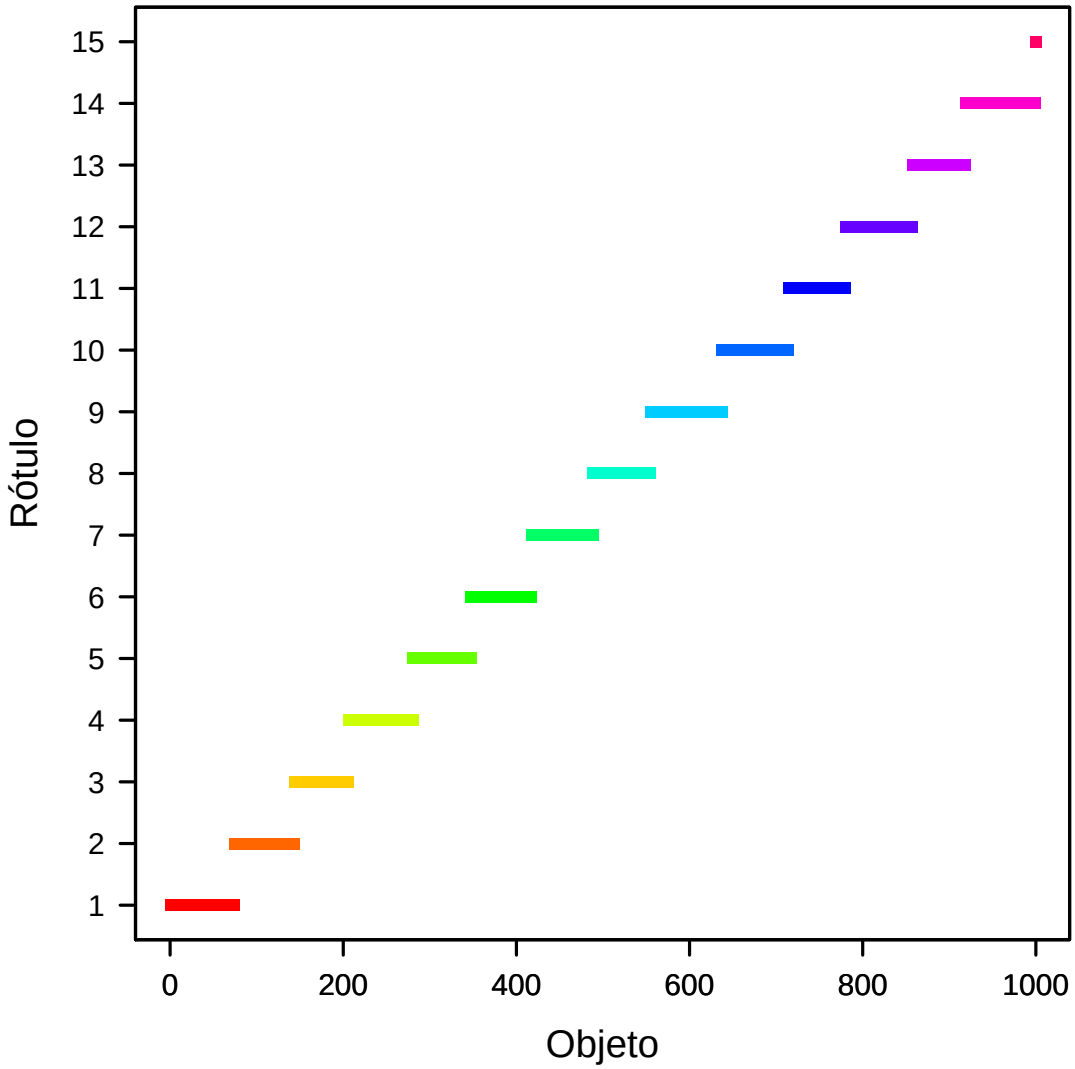

Figura 4.71: Representação do conjunto de dados Wreath.

Podem-se ver na tabela 4.6 (página 65) as variações do valor da Silhueta Média, Índice de Rand e Índice de Rand Ajustado em função da partição gerada pelo algoritmo RGT. Nas 20 experiências, o algoritmo K-médias apresentou, em média, bons resultados. Porém, o algoritmo RGT apresentou um ótimo resultado classificando corretamente 999 objetos. Apenas um objeto ficou mal classificado, ficando isolado (*singleton*) do grupo que pertence, como ilustra a figura 4.73 da página 67. Os algoritmos hierárquicos tiveram um bom desempenho na classificação do conjunto de dados Wreath.

Tabela 4.6: Wreath: número de grupos.

			Silhueta			
	N^o Grupos	Obj/Grupo	Média	Variância	Rand	Rand Aj.
RGT	1	1000	–	–	0.0710	0
	12	140,69,133,74, 68,71,83,77,66, 77,61,81	0.6830	0.1779	0.9811	0.8725
	13	140,69,63,74,68, 70,71,83,77,66, 77,71,81	0.7159	0.1578	0.9899	0.9285
	14	140,69,63,74, 67,70,71,83,77, 66,77,61,81,1	0.6939	0.1749	0.9901	0.9294
	15	**74,69,63,74,67, 70,71,66,83,77, 66,77,61,81,1**	**0.7435**	**0.1309**	**0.9999**	**0.9990**
	19	74,69,62,74,67, 70,71,65,82,77, 66,77,1,61,80, 1,1,1,1	0.6461	0.2170	0.9993	0.9946
K-médias	14				0.9733 (0.0146)	0.8211 (0.0868)
Ward	14 (310.6)	74,69,63,74,68, 70,71,66,83,77, 66,77,61,81			0.9997	0.998
L.Simples	14 (2.17)	140,69,63,74, 67,1,70,71,83, 77,66,77,61,81			0.99	0.929
L.Completa	14 (10.81)	74,69,60,80,68, 73,65,66,83,77, 66,77,61,81			0.997	0.979
L.Média	14 (7.05)	75,69,63,74,68, 70,71,65,83,77, 66,77,61,81			0.9994	0.9959

Podem-se ver na figura 4.72 os objetos do conjunto de dados Wreath e suas respectivas silhuetas de acordo com a partição gerada pelo algoritmo RGT. Símbolos de mesma cor representam objetos pertencentes ao mesmo grupo. Em destaque, na parte inferior do gráfico pode-se observar a silhueta do objeto *singleton*, ou seja, um grupo constituído de apenas um único objeto.

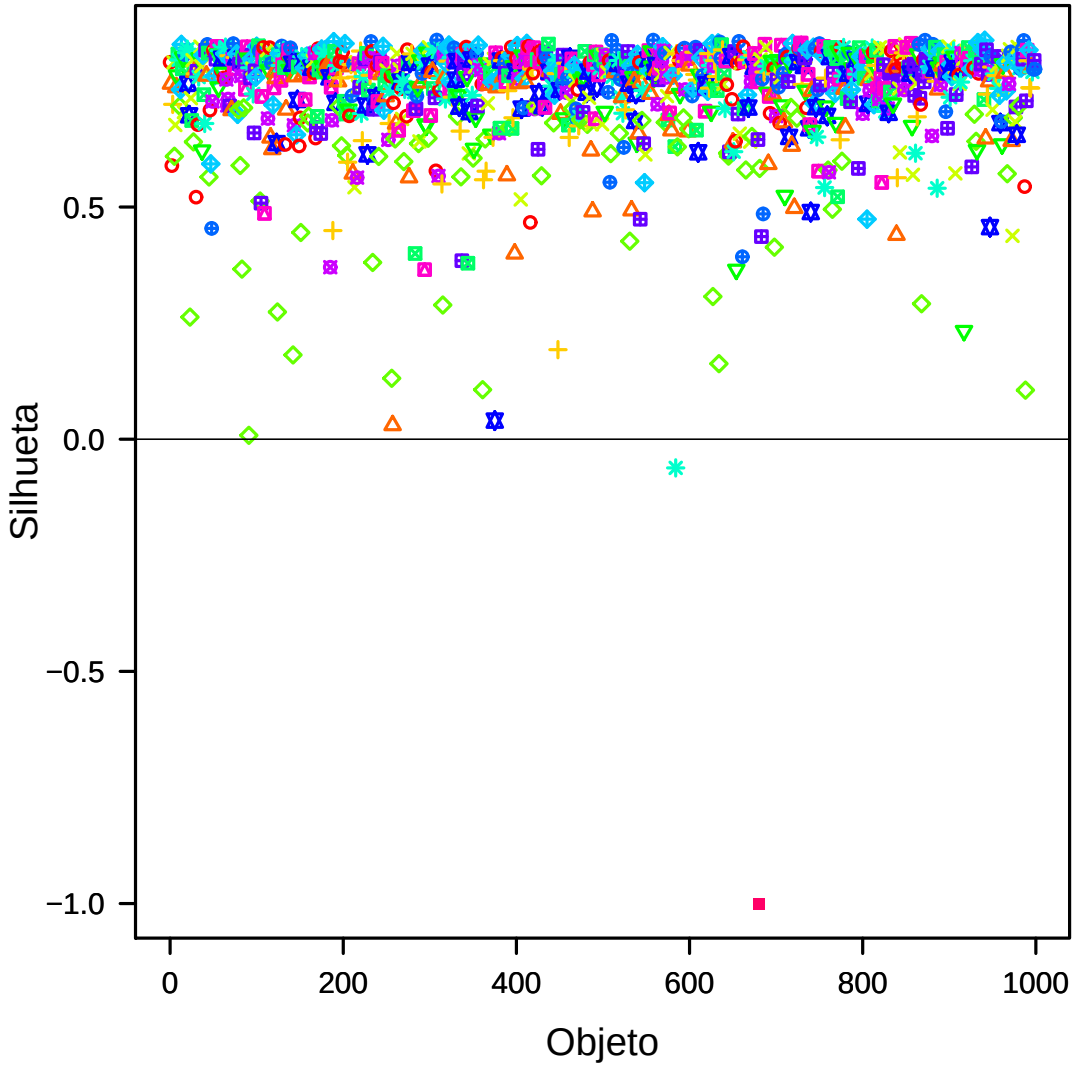

Figura 4.72: Wreath: silhueta dos objetos pelo algoritmo RGT (15 grupos).

A seguir, pode-se ver na figura 4.73 a partição gerada pelo algoritmo RGT e outra partição gerada através do algoritmo K-médias na figura 4.74:

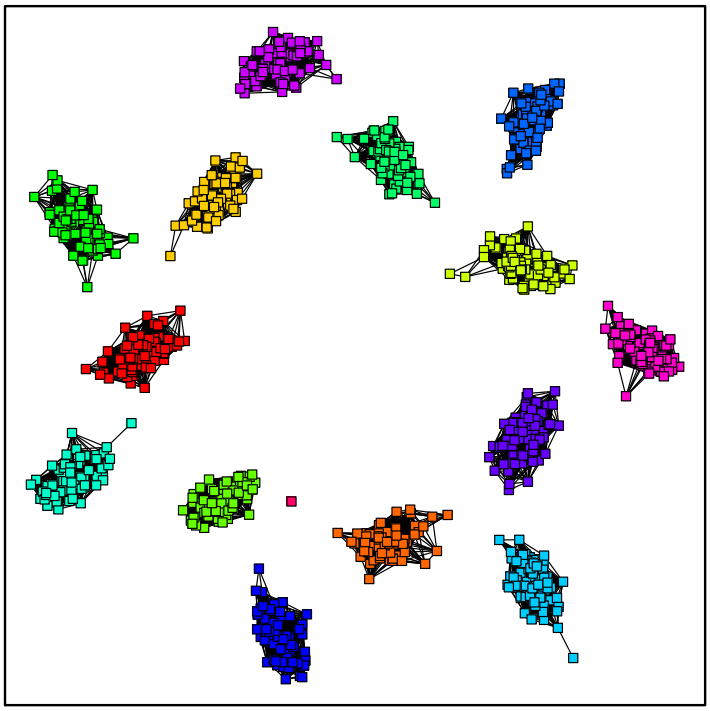

Figura 4.73: Wreath: partição do algoritmo RGT formando 15 grupos.

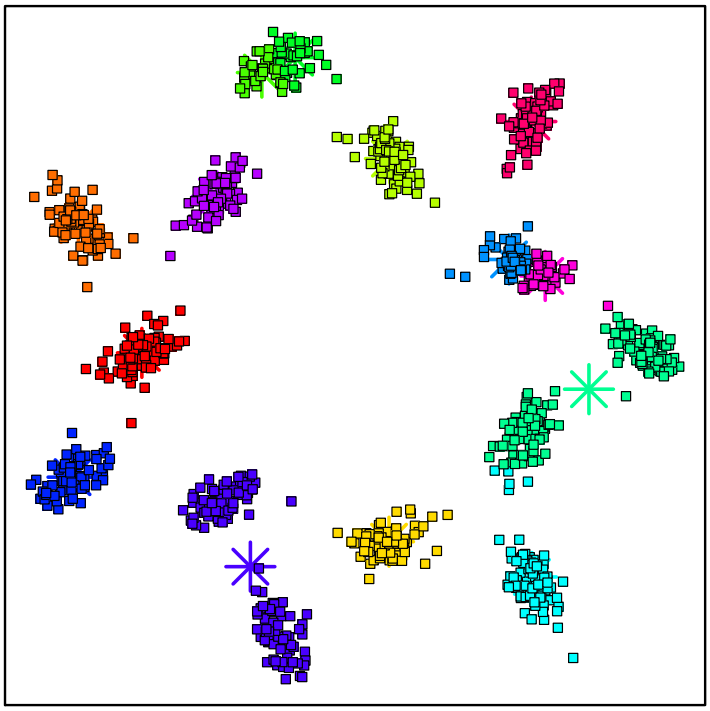

Figura 4.74: Wreath: partição do algoritmo K-médias com 14 grupos.

A partir das figuras 4.75 à 4.78 tem-se dendogramas que ilustram o agrupamento hierárquico aglomerativo para o conjunto de dados Wreath:

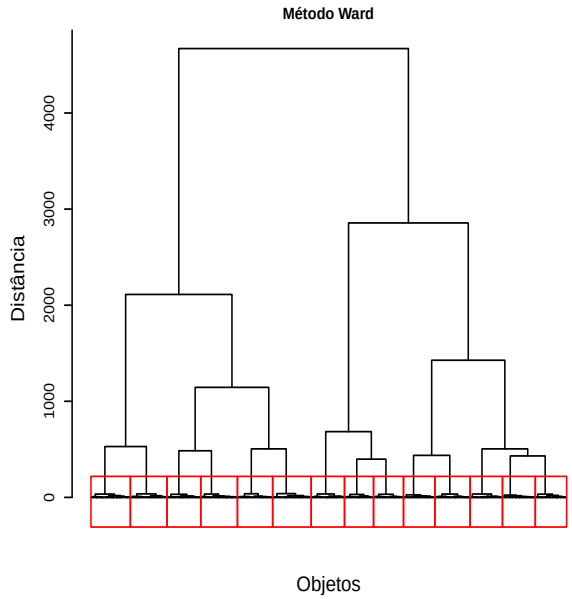

Figura 4.75: Wreath: ponto de parada 310.6.

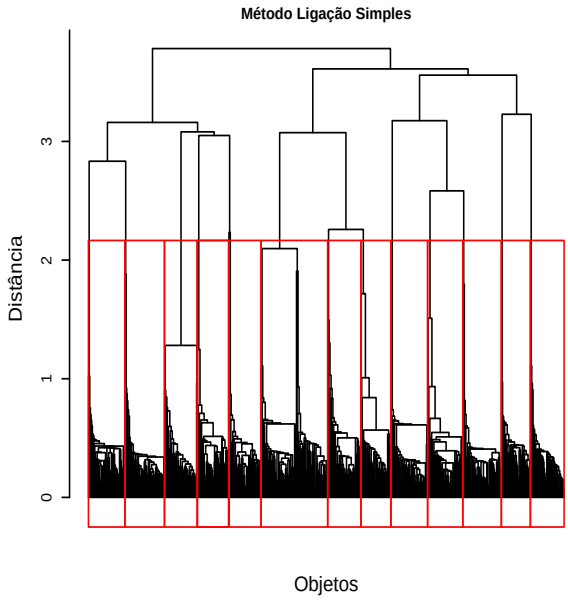

Figura 4.76: Wreath: ponto de parada 2.17.

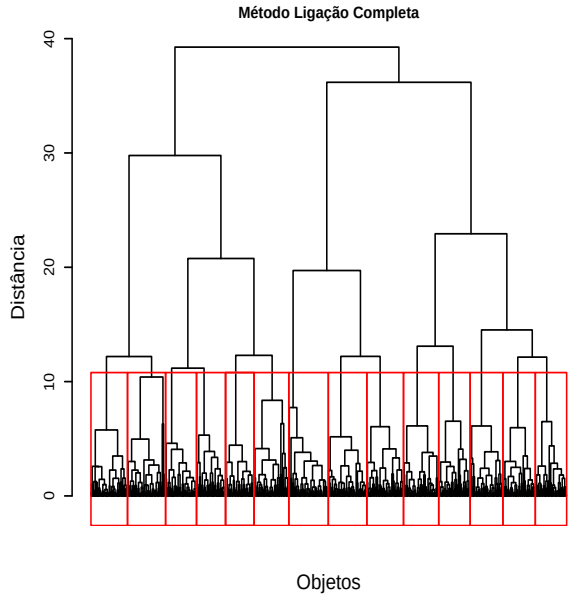

Figura 4.77: Wreath: ponto de parada 10.81.

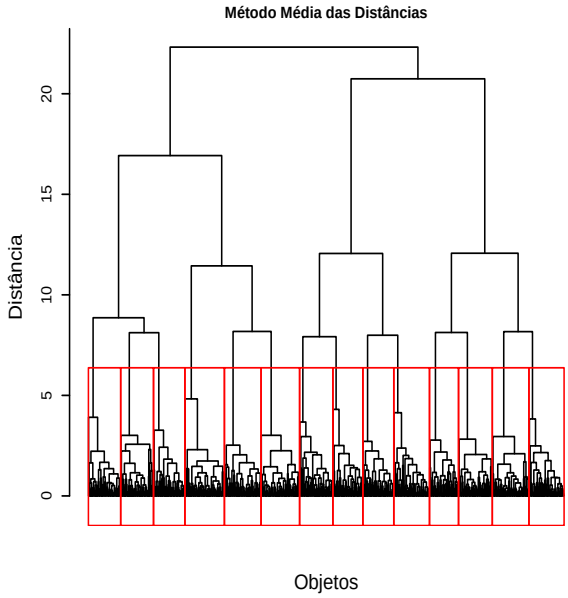

Figura 4.78: Wreath: ponto de parada 7.05.

4.6 Ionosfera

A Ionosfera se localiza entre 60 Km e aproximadamente 400 Km de altitude. Ela é composta de íons, plasma ionosférico, e, devida sua composição, reflete ondas de rádio até aproximadamente 30 MHz. Seu maior agente de ionização é o Sol, cuja radiação nas bandas de raios-X, e ultravioleta, insere grandes quantidades de elétrons livres em seu meio [8]. A Ionosfera se localiza entre 60 Km e aproximadamente 400 Km de altitude. Ela é composta de íons, plasma ionosférico, e, devida sua composição, reflete ondas de rádio até aproximadamente 30 MHz. Seu maior agente de ionização é o Sol, cuja radiação nas bandas de raios-X, e ultravioleta, insere grandes quantidades de elétrons livres em seu meio [8].

O conjunto de dados Ionosfera foi obtido no repositório de banco de dados da Universidade da Califórnia [10, 6, 45]. Esse conjunto de dados corresponde a dados adquiridos pelo laboratório de Física Aplicada, da Universidade John Hopkins. Os dados da base representam elétrons livres na ionosfera, que é ionizada pela radiação solar ultravioleta. Cada objeto possui 34 dimensões (atributos), sendo que a classe positiva corresponde a presença de estruturas na ionosfera (225 objetos), e a negativa (126 objetos) indica ausência, totalizando 351 objetos.

A seguir, podem-se ver nas figuras 4.79 à 4.81 os histogramas dos Filtros 0, 1 e 2 aplicados ao conjunto de dados Wreath. A figura 4.82 ilustra cada objeto do conjunto de dados Ionosfera e seu respectivo rótulo:

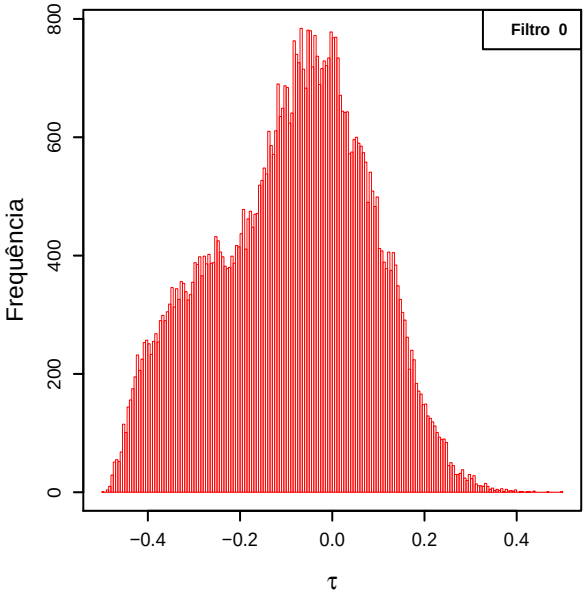

Figura 4.79: Ionosfera: filtro 0.

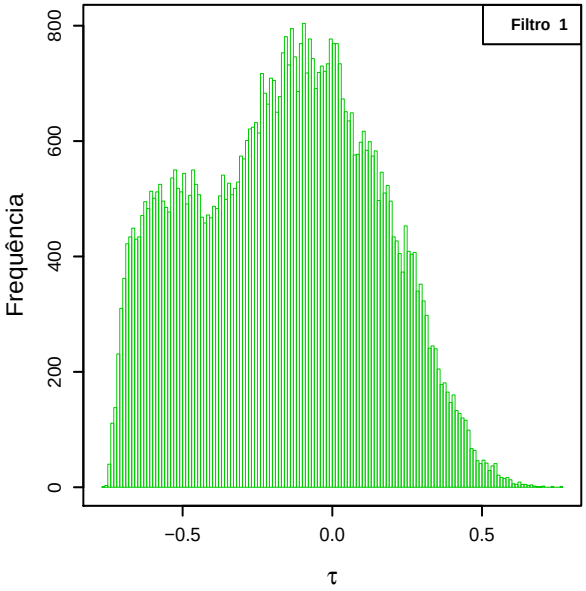

Figura 4.80: Ionosfera: filtro 1.

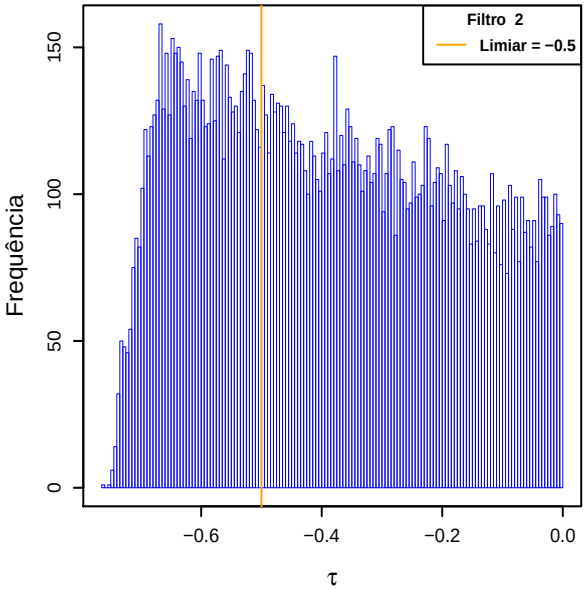

Figura 4.81: Ionosfera: filtro 2.

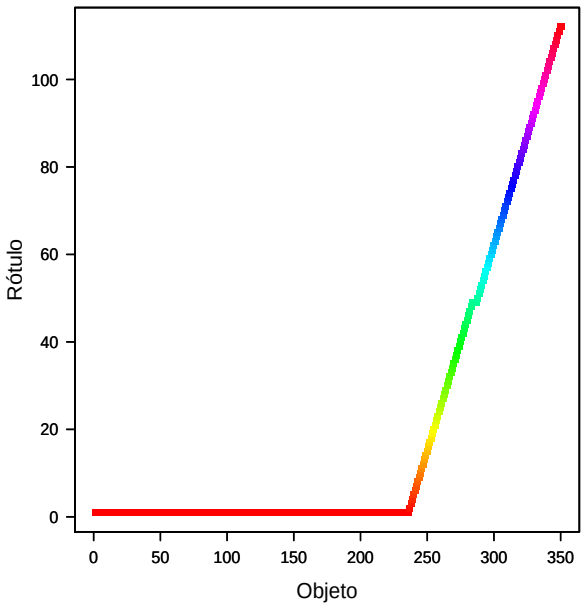

Figura 4.82: Ionosfera: Rótulos dos objetos.

Podem-se ver na tabela 4.7 as variações do valor da Silhueta Média, Índice de Rand e Índice de Rand Ajustado em função da partição gerada pelo algoritmo RGT. Apesar de não encontrar o correto número de grupos (2), o algoritmo RGT classificou em um grupo 225 objetos corretamente e agrupou mais 11 objetos excedentes que não deveriam fazer parte desse primeiro grupo. O restante dos objetos ficaram dispostos em um grupo com 4 objetos, outro grupo com 2 objetos e restante dos objetos formando *singletons*. O fato de conseguir agrupar 225 objetos corretamente elevou os índices de Rand e Rand Ajustado no desempenho do algoritmo RGT. Os demais algoritmos tiveram desempenho inferior.

Tabela 4.7: Ionosfera: número de grupos.

			Silhueta			
	N^o*Grupos*	Obj/Grupo	Média	Variância	Rand	Rand Aj.
	95	253,4,2,1,...,1	-0.3746	0.41145	0.7628	0.5242
RGT	**112**	**236,4,2,1,...,1**	**-0.51**	**0.37**	**0.7789**	**0.561**
	144	177,4,2,28,1,...,1	-0.5842	0.382	0.3382	0.561
K-médias	2				0.5889 (0)	0.1776 (0)
Ward	2 (149.7)	194,157			0.5938	0.1872
L.Simples	2 (5.2)	1,350			0.54	0.0044
L.Completa	2 (9.6)	46,305			0.5684	0.099
L.Média	2 (6.6)	1,350			0.54	0.0044

Podem-se ver na figura 4.83 os objetos do conjunto de dados Ionosfera e suas respectivas silhuetas de acordo com a partição gerada pelo algoritmo RGT. Objetos de mesma cor e mesmo símbolo pertencem ao mesmo grupo.

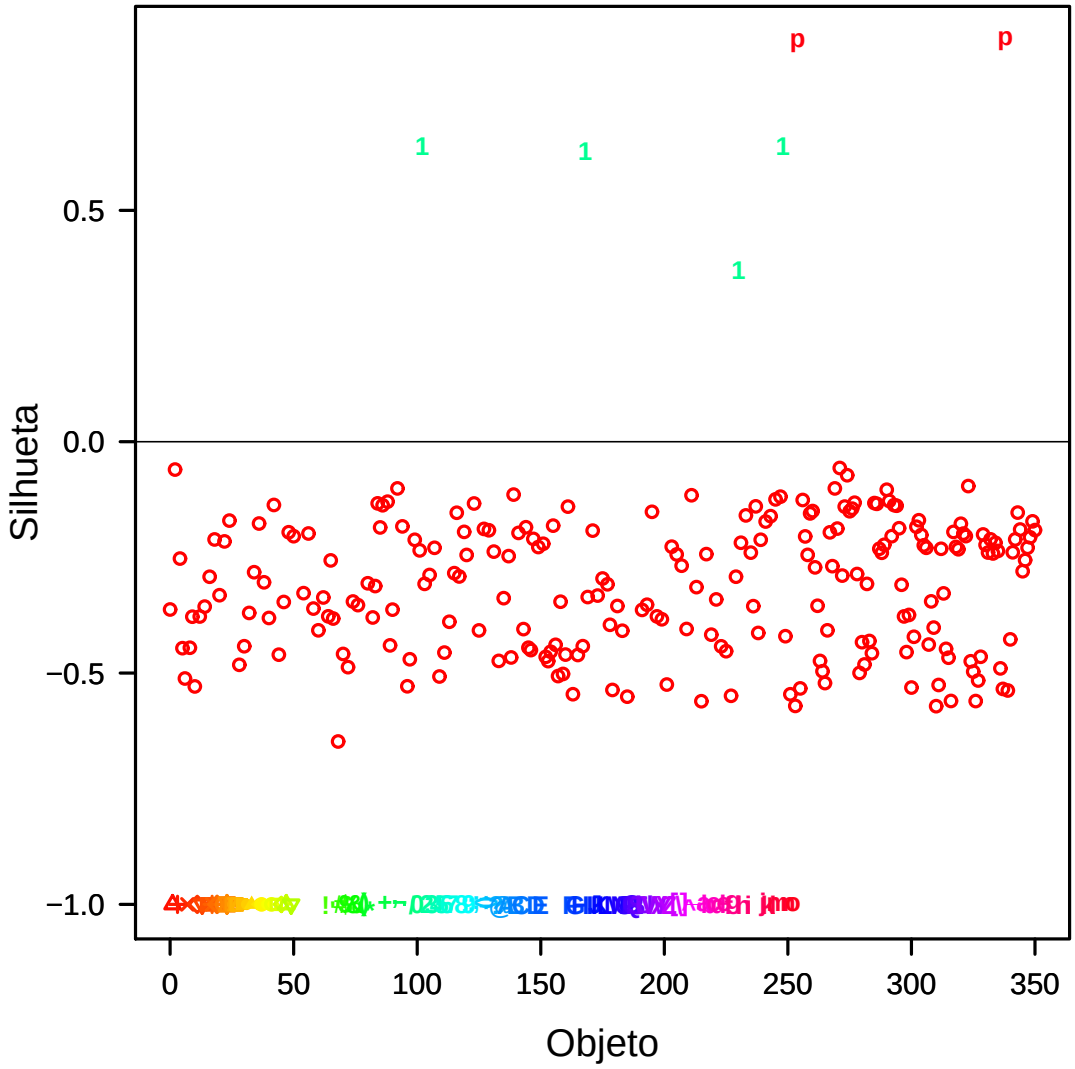

Figura 4.83: Ionosfera: silhueta dos objetos pelo algoritmo RGT (112 grupos).

A partir das figuras 4.84 à **??** tem-se dendogramas que ilustram o agrupamento hierárquico aglomerativo para o conjunto de dados Ionosfera:

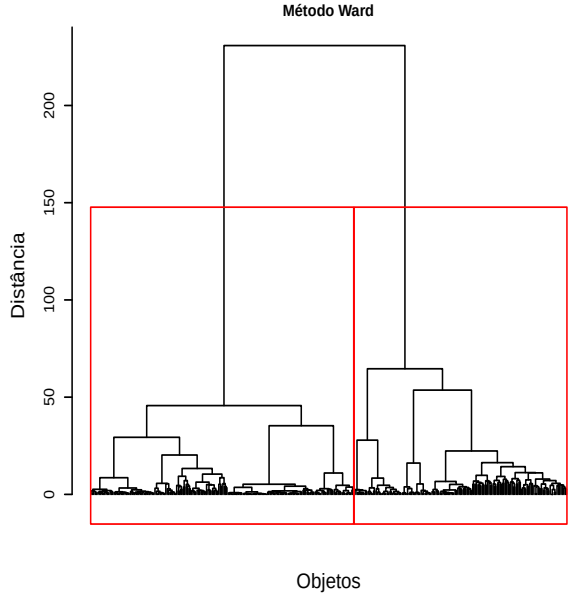

Figura 4.84: Ionosfera: ponto de parada 149.7.

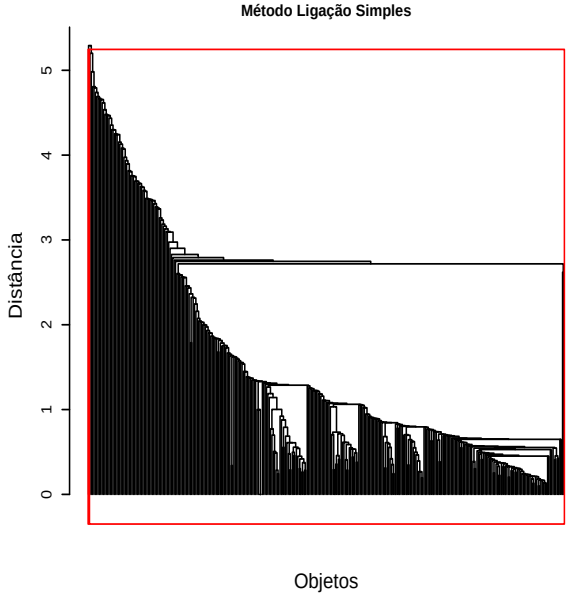

Figura 4.85: Ionosfera: ponto de parada 5.2.

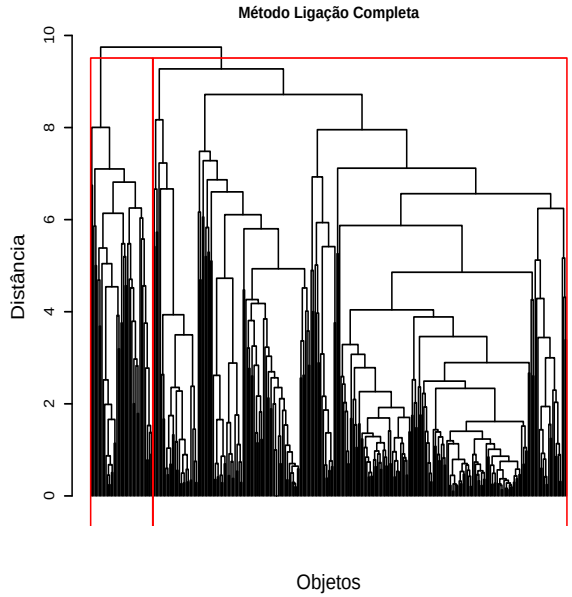

Figura 4.86: Ionosfera: ponto de parada 9.6.

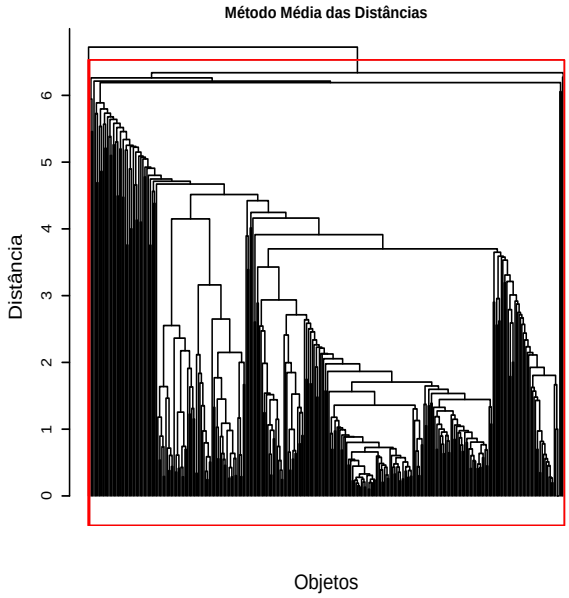

Figura 4.87: Ionosfera: ponto de parada 6.6.

Referências Bibliográficas

[1] *Dinâmica de Populações – Teoria.* Incor - São Paulo - SP, Disponível em: <http://physionet.incor.usp.br/~julio/Cursos/MPT5762/Aula5.pdf>.

[2] A. Abraham, S. Das, and S. Roy. *Swarm Intelligence Algorithms for Data Clustering.* Springer, New York, USA, 2007.

[3] José Domingos Albuquerque Aguiar. *MCAC – Monte Carlo Ant Colony: Um Novo Algoritmo Estocástico de Agrupamento de Dados.* DEINFO, Universidade Federal Rural de Pernambuco, 2008.

[4] N. Arley and K. R. Buch. *Introduction to then theory of probability and Statistics.* Wiley and Sons Publishers, New York, USA, 1950.

[5] J. Bilmes, A. Vahdaty, and W. Hsu. *Empirical Observations of Probabilistic Heuristics for the Clustering Problem.* International Computer Science Institute - Technical Report TR-97-018, 1997.

[6] C.L. Blake and C.J. Merz. Uci repository of machine learning databases. 1989.

[7] W. de O. Bussab, E. S. Miazaki, and D. F. Andrade. *Introdução à Análise de Agrupamentos.* 9ž Simpósio Nacional de Probabilidade e Estatística, São Paulo. Associação Brasileira de Estatística, 1990.

[8] A. H Cardoso. *Análise de Alguns Parâmetros lonosféricos na Anomalia Geomagnética do Atlântico Sul Mediante Ondas "VLF"*, volume 12. Revista Brasileira de Física, 1982. NP 2.

[9] I.W-Y. Chiang, G-S. Liang, and S.Z. Yahalom. *The fuzzy clustering method: Applications in the air transport market in Taiwan*, volume 11. The Journal of Database Marketing & Customer Strategy Management, 2003.

[10] Universidade da Califórnia. Machine learning repository. Disponível em: http://archive.ics.uci.edu/ml/datasets.html.

[11] T. Devezas and J. Corredine. *The Biological Determinants of Long Wave Behavior in Socioeconomic Growth and Development*, volume 68. Technological Forecasting & Social Change, 2001.

[12] Tessaleno. Devezas and George. Modelsky. *Power Law Behavior and World System Evolution: A Millennial Learning Process*, volume 70. Technological Forecasting & Social Change, 2003.

[13] W. Dzwinel, D. A. Yuen, and K. et al Boryczko. *Nonlinear multidimensional scaling and visualization of earthquake clusters over space, time and feature space*. Number 12. Nonlinear Processes in Geophysics, 2005.

[14] B. Everitt. *Cluster Analysis*. Heinemann Educacional Books, Londres.

[15] B. S. Everitt, S. Landau, and M. Leese. *Cluster Analysis*. Arnold Publishers, 2001.

[16] Levia Jr D. F. and Page D. R. *The Use of Cluster Analysis in Distinguishing Farmland Prone to Residential Development: A Case Study of Sterling, Massachusetts*, volume 25. Environ Manage, 2003.

[17] R.A. Fisher. *The use of multiple measurements in taxonomic problems*. Annual Eugenics, 7, parth ii edition, 1936.

[18] A. Fontana and M. C. Naldi. *Estudo de Comparação de Métodos para Estimação de Números de Grupos em Problemas de Agrupamento de Dados*. Universidade de São Paulo, ISSN - 0103 - 2569, 2009.

[19] Z. Güngör and A. Ünler. *K-harmonic Means Data Clustering With Simulated Annealing Heuristic*. Applied Mathematics and Computation 184, 2007.

[20] T. Graepel. *Statistical physics of clustering algorithms*. Technical Report 171822, FB Physik, Institut fur Theoretische Physic, 1998.

[21] J. F Hair. *Análise multivariada de dados*. Bookman, Porto Alegre, RS, 5 edition, 2005. Trad. Adonai S. SantŠAnna e Anselmo C. Neto.

[22] K. M Hammouda. *Web Mining: Identifying Document Structure for Web Document Clustering*. Department of Systems Design Engineering, University of Waterloo, Canada, 2002.

[23] J. Handl. *Ant-Based Methods for Task of Clustering and Topographic Mapping: Improvements, Evaluation and Comparison with Alternative Methods*, volume 2003.130f. Tese (Doutorado em informática), Universidade Erlangen-Nürnberg, 2003.

[24] J. Handl, J. Knowles, and M. Dorigo. *On the performance of ant-based clustering. In: Abraham A., M. Koppen, K. Franke. Design And Application of Hybrid Intelligent System.* IOS Press, 2003.

[25] A. C. Hencher. *Methods of Multivariate Analysis.* Wilwy Interscience, New Jersey, USA, 2002.

[26] C. Hennig and B. Hausdorf. *Design of dissimilarity measures: A new dissimilarity measure between species distribution ranges.* in: V. Batagelj, H.H. Bock, A. Ferligoj, A. Ziberna (Eds.), Data Science and Classification, Studies in Classification, Data Analysis, and Knowledge Organization, SpringerVerlag GmbH, Berlin, Germany, 2006.

[27] W. Härdle and L. Simar. *Applied Multivariate Statistical Analysis.* Springer, New York, USA, 2007.

[28] L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2:193 Ű 218, 1985.

[29] A. K. Jain and R. C. Dubes. *Algorithms for Clustering data.* Prentice-Hall, Inc, Upper Saddle River, New Jersey, USA, 1988.

[30] A. K. Jain, M. N. Murty, and Flynn P.J. *Data Clustering: A Review. ACM Computing Surveys*, volume 31. New York, NY, USA, September 1999.

[31] R. A. Johnson and D. W. Wichern. *Applied Multivariate Statistical Analysis.* Number 3. Prentice-Hall, 1988.

[32] R. A. Johnson and D. W. Wichern. *Applied Multivariate Statistica Analysis.* Prentice-Hall, Inc, USA, 1998.

[33] Richard. A. Johnson and Dean W. Wichern. *Applied Multivariate Statistical Analysis.* Prentice Hall, New Jersey, 4th edition, 1992.

[34] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis, Probability and Mathematical Statistics.* John Wiley, 1990.

[35] J. Kogan, C. Nicholas, and M. Teboulle. *Grouping Multidimensional Data, Recent Advances in Clustering*. Springer, New York, NY, USA, 1998.

[36] L. S. Kubrusly. *Um procedimento para calcular índices a partir de uma base de dados multivariados*, volume 21. Pesquisa Operacional, Rio de Janeiro, 2005.

[37] J. MacQueen. *Some methods for classification and analysis of multivariate observations*, volume 1. University of California Press. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, eds L. M. Le Cam & J. Neyman, Berkeley, CA, 2005.

[38] N. Malhotra. *Pesquisa de marketing: uma orientação aplicada*. Bookman, Poto Alegre, 4 edition, 2006. Trad. Laura Bocco.

[39] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.

[40] W. M. Rand. *Objective criteria for the evaluation of clustering methods*. Journal of the American Statistical Assiciation, 1971.

[41] V.J. Rayward-Smith. *Metaheuristics for clustering in kdd. Proceedings of the IEEE Congress on Evolutionary Computation 2005*. IEE Press, 2005.

[42] S. O. Resende, J. B. Pugliesi, E. A. Melanda, and M. F. de Paula. *Mineração de Dados. Sistemas Inteligentes: fundamentos e aplicações*, volume 1. Manole, São Paulo, SP, 2005.

[43] P. J. Rousseeuw. *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis*. Journal of Computational and Applied Mathematics, 1987.

[44] E. H. Ruspini. *Numerical Methods for Fussy Clustering*, volume 2. Information Sciences, 1970.

[45] V G Sigillito, S P Wing, L V Hutton, and K B Baker. Classification of radar returns from the ionosphere using neural networks. *Johns Hopkins APL Tech. Dig*, vol. 10:262–266, 1989. in.

[46] J. Sun, W. Xu, and B. Ye. *Quantum-Behaved Particle Swarm Optimization Clutering Algorithm. Advanced data mining and aplications: second international conference, ADMA*. Springer, Verlag Berlin Heidelberg, 2006.

[47] N. Timm. *Applied Multivariate Analysis*. Springer, New York, USA, 2002.

[48] W.T. Williams, M.B. Dale, and P. MacnaughtonSmith. *An objective method of weighting in similarity analysis*. Nature.

[49] E. Yeoh, M. E. Ross, and S. A. et al Shurtleff. *Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling*, volume 1. Cancer Cell, 2002.

[50] Y. Zhao and G. Karypis. *Evaluation of hierarchical clustering algorithms for document datasets*. Pesquisa Operacional, ACM, New York, NY, USA, 2002.