Universidade de Brasília
Departamento de Estatística

# Identificação de Padrões em Fadiga Muscular

**Estevam Caixeta Martins Teixeira - 08/28670**

**Brasília - DF**

**2011**

Estevam Caixeta Martins Teixeira - 08/28670

# Identificação de Padrões em Fadiga Muscular

Relatório apresentado à disciplina Estágio Supervisionado II do curso de graduação em Estatística, Departamento de Estatística, Instituto de Exatas, Universidade de Brasília, como parte dos requisitos necessários para o grau de Bacharel em Estatística.

Orientador: *Prof.º George F. von Borries*

**Brasília - DF**

**2011**

Monografia sob o título "**Identificação de Padrões em Fadiga Muscular**", defendida por *Estevam Caixeta Martins Teixeira* e aprovada em dia de mês de 2011, na Universidade de Brasília - Distrito Federal, pela banca examinadora assim constituída:

---

George Freitas von Borries (Orientador)

PhD. em Estatística, Kansas State University, K.S.U., EUA, 2008

Departamento de Estatística - UnB

---

Lúcio José Vivaldi

PhD. em Estatística, North Carolina State University, U.N.C., EUA, 1982

Departamento de Estatística - UnB

---

Maria Amélia Biagio

Doutora em Engenharia Elétrica, Universidade Estadual de Campinas, UNICAMP, Brasil, 1991

Departamento de Estatística - UnB

Brasília, 2011

Dedico este trabalho

Aos meus pais

*Adão e Edite.*

aos meus irmãos

*Samuel e Bárbara*

e a minha namorada

*Rafaela.*

# Agradecimentos

Primeiramente a Deus pela sua bondade e infinito amor.

À minha família pela compreensão, cooperação e apoio recebido.

À minha namorada Rafaela e sua família pelo imenso apoio e incentivo dados.

Ao meu orientador pela grande paciência, dedicação e prestabilidade na realização deste trabalho.

Aos professores Lúcio e Maria Amélia, participantes da banca examinadora, pela colaboração e contribuição com o trabalho.

Ao *Biopotentials Imaging Laboratory* ($\boldsymbol{BIM}_L$) - laboratório localizado na Universidade do Texas em El Paso (UTEP) - EUA pelo fornecimento do banco de dados de eletromiografia.

Ao departamento de Estatística, sobretudo ao corpo docente que participou na construção da minha trajetória.

# Resumo

A *surface electromyography* é uma ferramenta muito valiosa na indicação de fadiga muscular em estudos ocupacionais. Para este propósito o tempo de captação do sinal de EMG deve ser analisado a fim de se detectar alterações nos sinais típicas de fadiga muscular, tais como, um aumento da amplitude e uma diminuição da frequência do sinal de EMG. Esses procedimentos exigem um conhecimento detalhado sobre a atividade real da pessoa e do músculo sob teste para o tempo total de medição.

Este trabalho tem por intuito utilizar uma nova abordagem desenvolvida para a análise conjunta dos parâmetros dos sinais de EMG (amplitude e frequência) denominado *joint analysis of spectrum and amplitude* (JASA). Este método permite detectar se mudanças no sinal de EMG foram induzidas por fadiga muscular, ou por um estado de força do músculo. Usando este procedimento, as mudanças no sinal de EMG podem ser atribuídas a categorias como a fadiga ou recuperação, bem como ao aumento ou à diminuição na produção de força do músculo sob teste.

A técnica do JASA será abordada conjuntamente com outras ferramentas estatísticas que permitem a análise de dados multidimensionais como, por exemplo, técnicas de agrupamento e *self-organizing maps* (SOM). O objetivo é conciliar o uso destas ferramentas com o JASA de modo a ter-se uma alternativa, estatisticamente eficiente, agregada à representação esquemática proposta pelo JASA.

**Palavras-chaves:** Fadiga ,Eletromiografia, Agrupamento, *Self-organizing map*, SOM, *Joint Analysis of Spectrum and Amplitude*, JASA.

# Lista de Figuras

# Lista de Tabelas

# Lista de Siglas

| | |
|---|---|
| ARI | *Adjusted Rand Index* |
| $\boldsymbol{BIM_L}$ | *Biopotentials Imaging Laboratory* |
| CV | *Conductivity Velocity* |
| EMG | Eletromiografia |
| sEMG | *Surface Electromyography* |
| fdp | *Função Densidade de Probabilidade* |
| FFT | *Fast Fourier Transform* |
| JASA | *Joint Analysis of Spectrum and Amplitude* |
| MDF | *Median Frequency* |
| MDS | *Multidimensional Scaling* |
| MVC | *Maximum Voluntary Contraction* |
| PCA | *Principal Components Analysis* |
| UNB | *Universidade de Brasília* |
| UPGMA | *Unweighted Pair Group Method using arithmetic Averages* |
| UTEP | *University of Texas at El Paso* |

# Sumário

# 1 Introdução

Este trabalho faz parte de um projeto conjunto do *Biopotentials Imaging Lab*($\boldsymbol{BIM_L}$) do Departamento de Engenharia Elétrica e Computacional da *University of Texas at El Paso*(UTEP) na pessoa da aluna de mestrado Fernanda Leite e o Laboratório de Imagens Biomédicas do Departamento de Estatística da Universidade de Brasília (UnB). Assim, alguns resultados práticos e teóricos produzidos em estudos anteriores pela aluna, ou pelo $\boldsymbol{BIM_L}$, concernentes ao objeto de estudo poderão ser utilizados neste trabalho.

## 1.1 Eletromiografia Cinesiológica

A Eletromiografia Cinesiológica, ou simplesmente, Eletromiografia (EMG) é o estudo da função muscular preocupado com o desenvolvimento, captação e análise de sinais mioelétricos oriundos da ativação neuromuscular dos músculos através da postura corporal, dos movimentos funcionais e das atividades físicas. Sua análise se dá através dos sinais mioelétricos.

Outro termo bastante encontrado na literatura científica e que será utilizado neste trabalho é o de *Surface Electromyography*(sEMG) que nada mais é que um procedimento de medição da atividade neuromuscular através de eletrodos posicionados na pele.

Os sinais mioelétricos são provenientes do ciclo de polarização-repolarização na contração muscular que formam um dipolo elétrico transmitido pela superfície da fibra muscular (Konrad, 2005). Esse dipolo permite a geração de corrente elétrica que, por sua vez, é captada por aparelhos próprios denominados amplificadores, que exercem um papel de identificação do sinal interno e exclusão de qualquer interferência externa. Os sinais captados (analógicos) devem ser convertidos em digitais para serem analisados através de um computador. Porém, neste trajeto músculo-amplificador, o sinal de EMG pode ser influenciado por fatores que podem alterar sua forma e características:

- **Características dos Tecidos**
  - O corpo humano contém muita água e sais minerais sendo, portanto, um excelente condutor elétrico. No entanto, características como espessura e temperatura dos tecidos influenciam no sinal o que pode levar-nos a ter dificuldade na comparação dos parâmetros de EMG entre pessoas.
- ***"Cross-Talk"***
  - Músculos "vizinhos" aos músculos monitorados podem produzir uma quantidade significativa de energia, de tal forma que seus sinais acabam captados juntamente com o sinal do músculo de interesse gerando interferência na análise do sinal desse músculo.
- **Mudança de Posição do Eletrodo**
  - Qualquer alteração na posição do eletrodo altera a leitura do sinal de sEMG

A Eletromiografia Cinesiológica além de servir como base de estudos fisiológicos e biomecânicos serve como uma ferramenta de avaliação e pesquisa na fisioterapia (Reabilitação), no treinamento desportivo (Esporte de Alto Rendimento), na ergonomia (Prevenção de Riscos), na medicina (Ortopedia, Cirurgia) e em várias outras áreas.

## 1.2 sEMG na Detecção de Fadiga Muscular

O primeiro registro acerca de alterações dos sinais de sEMG é creditado ao professor de fisiologia Hans Edmund Piper da *Royal Friedrich-Wilhelms-University* , em Berlim, que, em 1912, publicou um artigo chamado *Elektrophysiologic Menschlicher Muskeln* no qual ele relatou uma certa "desaceleração" dos sinais de sEMG durante contrações isométricas[1]. Consoante este fenômeno, em 1923, Stanley Cobb e Alexander Forbes (*Electromyographic Studies of Muscular Fatigue in Man*) concluíram que mudanças na amplitude do sinal de sEMG ocorriam devido à manifestação de fadiga no músculo sob contração isométrica.

[1] Segundo (Guyton and Hall, 2006), a captação de sinais de sEMG se dá através de 2 tipos de contração:

- **Contração Isométrica (Estática)**: é a contração muscular que não provoca movimento ou deslocamento articular. Não há alteração no comprimento do músculo, mas sim, um aumento na tensão máxima do mesmo.
- **Contração Isotônica (Dinâmica)**: é a contração muscular que provoca um movimento articular. Há alteração do comprimento do músculo sem alterar sua tensão máxima.

A partir da década de 50, com o desenvolvimento de aparelhos eletrônicos, as pesquisas sobre a influência da fadiga nas propriedades dos sinais mioelétricos ganharam um grande impulso. Vários autores, ler (Cifrek et al., 2009), começaram a relacionar mudanças na amplitude e na frequência dos sinais como formas de monitoramento da fadiga muscular. Atrelado a esse desenvolvimento tecnológico, o desenvolvimento computacional proporcionou a implementação de vários métodos de processamento do sinal de sEMG . Inicialmente, o método mais utilizado para estimação da frequência do sinal era a transformada de Fourier (FFT) - *Fast Fourier Transform*). Posteriormente, com a utilização da contração isotônica como meio de obtenção dos dados, passou-se à utilização da técnica de transformada de *Wavelets*. [2]

As propriedades dos sinais de EMG estão relacionadas às mudanças biomecânicas e fisiológicas da musculatura esquelética durante a contração muscular num determinado intervalo de tempo. Uma das consequências da contração muscular é o aumento da concentração de ácido lático no músculo relacionada à falta de oxigenação e nutrição das células pela corrente sanguínea. A um certo nível de contração, o fluxo de sangue é interrompido pela pressão intramuscular e com isso o músculo torna-se isquêmico. Um aumento da concentração de ácido lático gera fadiga pois modifica o pH intracelular. Como consequência, a velocidade de condução (CV) elétrica da fibra muscular diminui o que gera um decréscimo na frequência mediana da onda (**MDF** - *Median Frequency*).

Desta forma, a fadiga muscular pode ser determinada pela concentração de ácido lático no músculo através de amostras de sangue retiradas em períodos específicos durante a realização de uma atividade. Entretanto, esta forma de detecção não é eficiente pois, desta maneira, não é possível monitorar exatamente quando a fadiga ocorreu. Assim, um monitoramento contínuo da fadiga muscular seria imprescindível para a determinação exata do momento da fadiga. Portanto, as principais vantagens da utilização da sEMG na detecção de fadiga são, segundo (Cifrek et al., 2009):

1. É um procedimento não invasivo: não rompe a pele, nem penetra objetos no corpo.

[2] "Uma *Wavelet* é uma forma de onda que é limitada em freqüência e duração.A transformada *Wavelet* converte um sinal em uma série de ondas. Em teoria, os sinais processados pela transformação *Wavelet* podem ser armazenados mais eficientemente do que aqueles tratados por transformada de Fourier.

A transformada de Fourier converte um sinal em uma série contínua de ondas senoidais, cada um dos quais é de freqüência constante e de amplitude e de duração infinita. Em contraste, a maioria dos sinais do mundo real (como música ou imagens) têm uma duração limitada e alterações bruscas na freqüência"(Moshou et al., 2005).

2. Aplicação *in situ*: aplicação direta no músculo.

3. Monitoramento em tempo real.

4. Habilidade de monitoramento da fadiga em um músculo específico.

5. Correlação com mudanças bioquímicas e fisiológicas durante a ocorrência de fadiga.

## 1.3 Objetivos

Utilizar técnicas de agrupamento (*cluster*) tradicionais para verificar se existe alguma relação desconhecida entre os sinais de sEMG , características do experimento ou algum relato de desconforto. Comparar o agrupamento de 8 horas de sinais de sEMG e verificar se existe algum padrão de comportamento dos sinais durante este período que permita identificar a fadiga muscular. A idéia é utilizar tantas técnicas quantas forem possíveis e comparar os resultados. Especificar vantagens e desvantagens quando da aplicação de cada uma.

Utilizar, na detecção de fadiga muscular, uma técnica conhecida como *Joint Analysis of EMG Spectrum and Amplitude* (JASA), descrita em (Luttmann et al., 2000), segundo a qual, a frequência e a amplitude dos sinais de sEMG quando analisados conjuntamente, podem permitir detectar quando as mudanças nos parâmetros de sEMG são induzidas por fadiga muscular, ou relacionadas com outros fatores.

Explorar técnicas estatísticas que permitem análises visuais. Como exemplo, cita-se o *Self-Organizing Map* (SOM). O **SOM** (Kohonen, 2001) é uma técnica de rede neural criada pelo professor Teuvo Kohonen (*Helsinki University of Technology Neural Networks Research Centre*) cuja aplicação associada a *Wavelets* na detecção de fadiga foi comentada no artigo (Moshou et al., 2005). Neste caso, utilizando **SOM**, é possível inclusive detectar se o músculo se recuperou temporariamente. Ele mapeia os sinais de entrada de espaços em alta dimensão ($\mathbb{R}^n$) para as redes de dimensão arbitrária. Pela facilidade de visualização e interpretação, os espaços paramétricos mais utilizados são $\mathbb{R}$ e $\mathbb{R}^2$.

### 1.3.1 Análise e Gravação do sinal de sEMG

Foram gravados sinais de sEMG de 3 músculos para ambos os lados:

- *Trapezius* (Lados: Esquerdo/Direito)

- *Splenius Capitis* (Lados: Esquerdo/Direito)
- *Sternocleidomastoid* (Lados: Esquerdo/Direito)

Os dados foram processados por um aparelho com 8 sensores (*Delsys Bagnoli*-8 *DE-2.1 Standard Differential EMG Electrodes*). Os sinais de EMG , como dito anteriormente, são bastante influenciados pelas condições de medição (tecidos, "*cross-talk*"). Uma forma de melhorar essa característica é normalizando os parâmetros do sinal para um valor de referência, no caso, *Maximum Voluntary Contraction* (MVC). A idéia é calibrar os valores para uma unidade com uma relevância maior de interpretação. As vantagens da normalização são:

- Diminui a influência das condições de medição nos valores dos parâmetros do sinal.
- Fornece um entendimento do nível de capacidade de trabalho do músculo.

Tabela 1.1: Estrutura Geral dos Dados

| **Sexo** | **Indivíduo** | **Músculo** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | **$M_1$ (E/D)** | | | **$M_2$ (E/D)** | | | **$M_3$ (E/D)** | | |
| | | **$T_0$** | **$\cdots$** | **$T_7$** | **$T_0$** | **$\cdots$** | **$T_7$** | **$T_0$** | **$\cdots$** | **$T_7$** |
| **Masculino** | **1** | - | - | - | - | - | - | - | - | - |
| | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| | **15** | - | - | - | - | - | - | - | - | - |
| **Feminino** | **1** | - | - | - | - | - | - | - | - | - |
| | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| | **11** | - | - | - | - | - | - | - | - | - |

### 1.3.2 Análise Estatística

Os seguintes passos devem ser seguidos:

- Estudar análise, classificação, procedimentos e aplicações de técnicas de agrupamento utilizando as seguintes referências: (Gan et al., 2007), (Kaufman and Rousseeuw, 1990) e (Hastie et al., 2009) e os seguintes softwares: **SAS**, **JMP** e, eventualmente, o **R**
- Estudar e aplicar procedimentos de decisão para definição do número de *clusters*: **RMSSTD**, $R^2$, **SPRSQ**.

- Estudar procedimentos de validação de *clusters* (Kaufman and Rousseeuw, 1990).
- Analisar exploratoriamente os grupos e identificar possíveis características importantes.
- Aplicar outras técnicas e comparar os resultados: *Complete Linkage*, *Single Linkage*, *Mean Linkage*, Método de Ward.
- Aplicar o Self-organizing map (**SOM**).

Para os dados de sEMG pensa-se em recodificar seus parâmetros de frequência e amplitude em novos índices e, conjuntamente, aplicar o JASA. Talvez utilizar a distância radial [3] nos 4 quadrantes.

[3] Na geometria, uma distância radial é uma coordenada do sistema de coordenadas polares $(r, \theta)$. Basicamente a distância radial é a distância euclidiana escalar entre um ponto e a origem do sistema de coordenadas (0,0).

# 2 *Joint Analysis of Spectrum and Amplitude* (JASA)

Durante contrações musculares, repetitivas alterações típicas na eletromiografia de superfície (sEMG), tais como um aumento na amplitude ou uma diminuição da frequência podem ser observados. Na eletromiografia de superfície, mudanças são comumente interpretadas como sinais de fadiga muscular e usadas para estabelecer a ocorrência de fadiga. No entanto, uma vez que, a amplitude da sEMG, bem como seu conteúdo espectral (frequência) não dependem apenas do estado de fadiga, mas também da produção de força do músculo em teste, as mudanças na amplitude e na frequência não podem ser inequivocamente atribuídas à fadiga muscular.

Em condições de laboratório, é possível controlar a força produzida pelo músculo em teste. Sob tais circunstâncias, a força pode ser mantida constante em um nível conhecido e uma mudança na respectiva sEMG pode ser atribuída a uma mudança no estado de fadiga do músculo. Em condições reais, no entanto, a produção de força é determinada pelas necessidades reais da atividade desempenhada pelos indivíduos e não tem como ser controlada pelo pesquisador. Em geral, não é possível, portanto, decidir se uma variação temporal de uma sEMG é causada por uma mudança na produção de força ou no estado de fadiga. No entanto, é possível conseguir aferir a fadiga muscular sob certas condições.

Uma maneira é comparar os sinais de EMG em situações em que a força aplicada sobre os músculos é idêntica. Dois métodos que atendem a esse critério e que já foram muito explorados em estudos ocupacionais são:

1. A execução de testes de contração com força conhecida sob uma determinada postura;

2. A comparação das EMG's para determinadas atividades com cargas de trabalho semelhantes.

Numa nova abordagem (JASA), mudanças na amplitude e frequência são consideradas simultaneamente, possibilitando diferenciar as mudanças relacionadas à força ou induzidas por fadiga. Usando esta abordagem, mudanças temporais da EMG podem ser atribuídas às diferentes categorias tais como fadiga, recuperação, aumento de força, diminuição de força.

No contexto da ergonomia e da fisiologia do trabalho fadiga muscular é entendida como uma redução na capacidade de geração de força de um músculo. Todas as análises de fadiga através da eletromiografia são baseadas no pressuposto de que a mudança na capacidade de performance mecânica do músculo é refletida em mudanças do sinal mioelétrico do músculo sob teste. Diversos estudos mostraram que a fadiga muscular coincide com mudanças no sinal de EMG como, por exemplo, um aumento da amplitude e uma diminuição da frequência. O método mais utilizado na medição de EMG no ambiente ocupacional é a sEMG, já que este método permite a captação dos sinais mioelétricos de forma não-invasiva.

Devido à dependência dupla da amplitude e da frequência do sinal de EMG na força e na fadiga, os métodos para a determinação de fadiga só serão aplicáveis se a análise for feita levando-se em conta sinais produzidos sob a mesma produção de força.

O JASA é baseado, principalmente, na já conhecida relação entre produção de força muscular e fadiga, por um lado, e a amplitude e a frequência dos sinais de EMG do outro. Com relação à amplitude do sinal de EMG, é experimentalmente comprovado que a amplitude aumenta com o aumento da força assim como na presença de fadiga. Entretanto, uma relação funcional uniforme entre a produção mecânica de força e a amplitude do sinal não está completamente elucidada. A frequência também depende da fadiga e da força produzida. No caso em que ocorre fadiga, uma diminuição na distribuição espectral da frequência foi constantemente verificada. Quanto à dependência da força a distribuição espectral mostrou-se inconsistente e dependente do músculo sob análise. (Luttmann et al., 2000)

A relação entre frequência e amplitude dos sinais de EMG é dado esquematicamente na figura a seguir.

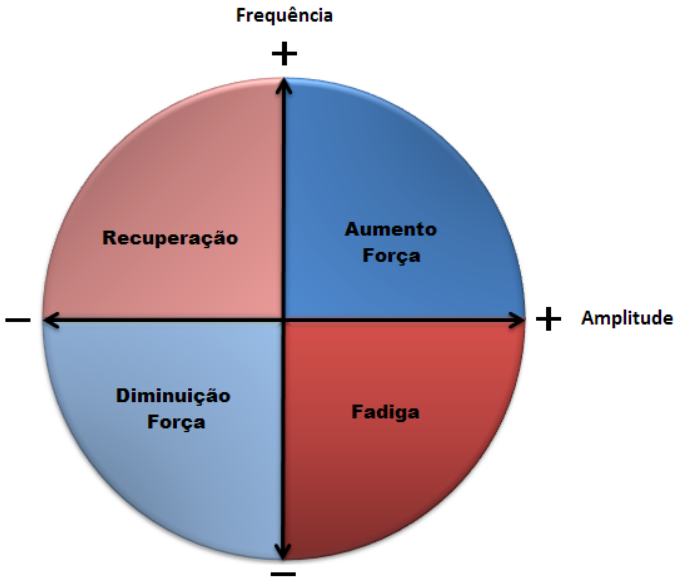



Figura 2.1: Representação Esquemática do método JASA

# 3 Agrupamento de Dados

Neste capítulo, são apresentados alguns conceitos básicos e necessários para a realização de uma análise de *cluster*. Primeiramente, descreve-se o que vem a ser análise de *cluster* e alguns exemplos de sua aplicação em diversas áreas. Posteriormente, alguns vocábulos específicos da análise de *cluster* como atributos, distâncias, similaridades e índices de validação são introduzidos.

## 3.1 Definição

Análise de *cluster*, também denominada análise de agrupamento, *segmentation analysis*, análise de tipologias,*taxonomy analysis* ou *unsupervised classification* é a "arte" de agrupar dados ou dividir os elementos de uma amostra, ou população, de modo que os elementos pertencentes ao mesmo *cluster* sejam homogêneos entre si com respeito às características que neles forem medidas, e os elementos em grupos diferentes, heterogêneos com relação a esta mesma característica.

Em suma, os grupos devem ser formados de maneira que as características dos elementos **dentro** dos *clusters* devem ser homogêneas e as **entre** *clusters* heterogêneas.

A classificação de objetos similares em grupos surge em várias áreas.

- **Psicologia:** classificação de pessoas de acordo com seus perfis de personalidade;
- **Ecologia:** classificação de animais, plantas;
- **Astronomia:** classificação de estrelas com base em características: intensidade de luz emitida, temperatura de sua superfície;
- **Geoquímica:** caracterização de conteúdo dos minerais;
- **Biologia:** análise de expressão genética;

- ***Data Mining*:** análise de grandes bancos de dados para a descoberta de informações relevantes.

## 3.2 Conceitos Básicos

### 3.2.1 Atributos

Durante o estudo de análise de *cluster*, surgem diversos termos para expressar a mesma finalidade. Para um conjunto de dados, as expressões *data*, objeto, observação, item, *tuple*, *record* e *pattern* servem para denotar uma simples observação do banco de dados. Já em espaços $n$-dimensionais, as palavras observação, atributo ou característica denotam um componente escalar (vetor).

### 3.2.2 Distâncias e Similaridades

Um dos conceitos mais importantes no que tange a análise de *cluster* e que devem ser muito bem entendidos é o de distâncias e similaridades.

Para podermos responder à pergunta:

> *Até que ponto dois objetos de um conjunto de dados podem ser considerados semelhantes?*

precisamos de medidas que possam descrever essa similaridade entre elementos amostrais de acordo com as características que neles forem medidas.

Usualmente, medidas de similaridade, medidas de dissimilaridade ou distâncias são utilizadas para descrever quantitativamente a relação entre as observações. Assim, a comparação de diferentes elementos amostrais poderá ser feita através da distância entre as observações.

Similaridade e distância são conceitos intimamente relacionados. Esta relação se dá da seguinte maneira: quanto menor a distância (ou a medida de similaridade), mais homogêneas (similares) são as observações. Estendendo-se essa idéia para dissimilaridade, tem-se que: quanto maior a medida de dissimilaridade (ou a distância) entre duas observações, mais heterogêneas (dissimilares) elas serão.

### 3.2.3 Índices de Validação

A análise de *cluster* é um processo não-supervisionado (*unsupervised process*). Isto significa dizer que ao realizá-la, não sabemos exatamente por quais *clusters* estamos procurando, nem como eles são formados, nem que tipo de relação as observações terão para determinadas características. Assim, visando comparar diversas abordagens possíveis de análise, é necessário estabelecer-se um critério de validação que nos mostre o método de análise mais eficiente. Os índices servem inclusive para fornecer um valor inicial do número de grupos a ser formado.

## 3.3 Procedimento

Normalmente, a análise de *cluster* envolve quatro fases (Gan et al., 2007). Aqui, porém, baseando-se conjuntamente na obra de (Theodoridis and Koutroumbas, 2009) destacamos cinco fases:

1. **Estruturação dos dados:** nesta fase, analisam-se os dados de forma bruta visando identificar algum padrão (característica) de agrupamento natural entre eles. As características devem ser escolhidas de modo a agregar o máximo de informação possível sobre o estudo em questão;

2. **Modelagem:** nesta fase, define-se o conceito de *cluster* e o critério e ser adotado na formação dos grupos.;

   - **Medidas de Distância:** quantifica o quão homogêneas ou similares duas características são;

   - **Critério de Agrupamento:** está intimamente ligado à sensibilidade do pesquisador e ao seu conhecimento a respeito do conjunto de dados.

     *Como ele espera que os dados irão se agrupar?*

3. **Otimização:** está relacionada à escolha adequada do algoritmo computacional que conseguirá desvendar a estrutura de agrupamento do banco de dados;

4. **Validação:** verificar a compatibilidade dos resultados obtidos com relação às expectativas do pesquisador;

5. **Interpretação dos Resultados:** o especialista na área de aplicação deve integrar os resultados do agrupamento com outras evidências experimentais e análise, a fim de tirar as conclusões corretas.

Aqui é importante ressaltar a importância da interação do especialista da área com o pesquisador. Diferentes escolhas de características, medidas de distância, critérios de agrupamento e algoritmos podem levar a resultados completamente diferentes.

> "*Subjetividade é uma realidade com a qual devemos conviver de agora em diante.*"
> (Theodoridis and Koutroumbas, 2009)

O objetivo da análise de *cluster* é separar as observações similares no mesmo grupo e as dissimilares em grupos diferentes. Assim, os problemas de agrupamento dividem-se em duas categorias (Gan et al., 2007):

- ***Hard Clustering*:** as observações pertencem a um e somente um *cluster*;
- ***Fuzzy Clustering*:** as observações podem pertencer a dois ou mais *clusters* com alguma probabilidade.

Matematicamente, essas categorias podem ser expressas assim: (Gan et al., 2007)

Dado um conjunto de dados $\mathbf{V}$ temos que um agrupamento é dado pela aplicação de uma função

$$\begin{aligned} f : \mathbf{V} &\to [0,1]^n \\ x &\mapsto f(x) \end{aligned}$$

sendo $f(x)$ definida como

$$f(x) = \begin{pmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_n(x) \end{pmatrix}$$

com

- $f_i(x) \in [0,1]$ para $i = 1,2,...,n$;

- $x \in \mathbf{V}$;
- $\sum_{i=1}^{n} f_i(x) = 1, \forall\, x \in \mathbf{V}$.

  Se $\forall\, x \in \mathbf{V}$, $f_i(x) \in \{0,1\}$ então $f$ representa um ***Hard Clustering***.

  Se $\forall\, x \in \mathbf{V}$, $f_i(x) \in [0,1]$ então $f$ representa um ***Fuzzy Clustering***.

### 3.3.1 *Missing Values* (NA's)

Saber lidar com *missing values* (valores desconhecidos/não conhecidos) é uma tarefa importante na análise de *cluster* já que na prática eles costumam ocorrer com grande frequência. Segundo (Gan et al., 2007), há três casos em que podem ocorrer *missing values* num banco de dados:

1. *Missing Values* podem ocorrer nas variáveis;
2. *Missing Values* podem ocorrer nas observações;
3. *Missing Values* podem ocorrer aleatoriamente nas variáveis e nas observações.

Se por um acaso existir alguma observação ou variável em que todas as medidas são *missing values*, então esta observação ou variável não possui informação nenhuma, assim ela deve ser retirada do banco de dados (Kaufman and Rousseeuw, 1990). Caso a quantidade de *missing values* não seja tão numerosa, os métodos para lidar com esse problema podem ser classificados em dois grupos (Gan et al., 2007):

1. *Pre-replacing methods*: substituir os *missing values* antes de iniciar o processo de análise.

   (Theodoridis and Koutroumbas, 2009) sugere que isto pode ser feito "completando" os *missing values*:

   - por zeros;
   - pela média incondicional: calculada a partir dos valores disponíveis na respectiva característica;
   - pela média condicional: calculada a partir da função densidade de probabilidade *(fdp)* dos *missing values* obtida com base nos dados observados (imputação). Utiliza argumentos de inferência bayesiana. Um algoritmo que pode ser utilizado na estimação dos parâmetros da *fdp* é o algoritmo EM.

2. *Embedded methods*: lidar com os *missing values* durante o processo de análise.

# 4 Medidas de Similaridade e Dissimilaridade

Neste capítulo, introduzir-se-ão conceitos importantes utilizados na fase de modelagem do processo de agrupamento. Primeiramente, a noção de medidas para diferentes tipos de dados incluindo dados numéricos e binários será discutida[1]. Posteriormente, baseado nestes medidas, várias medidas de similaridade e distância entre *clusters* e observações serão introduzidas.

## 4.1 Similaridade

Seja $s(a,b)$ a medida de similaridade entre duas observações $a$ e $b$. Quanto mais as observações $a$ e $b$ se assemelham, maior é o valor de $s(a,b)$.

- $s(a,b)$ deve satisfazer:

$$\textbf{(S1)}\ 0 \leq s(a,b) \leq 1;$$
$$\textbf{(S2)}\ s(a,a) = 1;$$
$$\textbf{(S3)}\ s(a,b) = s(b,a).$$

para quaisquer observações $a$ e $b$. De **(S1)** infere-se que "0" significa que $a$ e $b$ não se assemelham em quase nenhuma característica, enquanto "1", significa que $a$ e $b$ assemelham-se em todas ou quase todas as características.

[1] Os dados binários terão destaque aqui devido a sua posterior utilização em um exemplo do livro (Kohonen, 2001).
Para medidas relacionadas a outros tipos de dados ler (Gan et al., 2007).

Os números $s(a,b)$ são dispostos numa matriz quadrada $(n \times n)$ denominada matriz de similaridade.

$$M_{sim}(\mathbf{D}) = \begin{pmatrix} 1 & s_{12} & \cdots & s_{1n} \\ s_{21} & 1 & \cdots & s_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n1} & s_{n2} & \cdots & 1 \end{pmatrix}$$

sendo $s_{ab} = s(x_a, x_b)$ com relação a alguma medida de similaridade $s(\cdot,\cdot)$ e $D = \{x_1, x_2, ..., x_n\}$.

## 4.2 Dissimilaridade

Seja $d(a,b)$ a medida de dissimilaridade entre duas observações $a$ e $b$. Quanto mais as observações $a$ e $b$ se assemelham, menor é o valor de $d(a,b)$.

- $d(a,b)$ deve satisfazer:

**(D1)** $d(a,b) \geq 0$ e $d(a,b) = 0$, se e somente se, $x = y$;

**(D2)** $d(a,a) = 0$;

**(D3)** $d(a,b) = d(b,a)$;

**(D4)** $d(a,b) \leq d(a,c) + d(c,b)$.

para quaisquer observações $a$ e $b$. De **(D1)** infere-se que "1" significa que $a$ e $b$ não se assemelham em quase nenhuma característica e "0" significa que $a$ e $b$ assemelham-se em todas ou quase todas as características.

Os números $d(a,b)$ são dispostos numa matriz quadrada $(n \times n)$ denominada matriz de dissimilaridade.

$$M_{diss}(\mathbf{D}) = \begin{pmatrix} 0 & d_{12} & \cdots & d_{1n} \\ d_{21} & 0 & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & 0 \end{pmatrix}$$

sendo $d_{ab} = d(x_a, x_b)$ com relação a alguma medida de dissimilaridade $d(\cdot,\cdot)$ e $D = \{x_1, x_2, ..., x_n\}$.

## 4.3 Medidas para dados Numéricos

### 4.3.1 Distância Euclidiana

A distância Euclidiana é, certamente, a mais utilizada para dados numéricos. Para duas observações x e y em um espaço $p$-dimensional, a distância Euclidiana entre elas é definida como: (Gan et al., 2007)

$$d_{euclid}(\mathrm{x},\mathrm{y}) = \left[\sum_{j=1}^{p}(\mathrm{x_j}-\mathrm{y_j})^2\right]^{1/2} = [(\mathbf{x}-\mathbf{y})(\mathbf{x}-\mathbf{y})^T]^{1/2} \tag{4.1}$$

sendo $\mathrm{x_j}$ e $\mathrm{y_j}$ os valores da $j$-ésima característica de x e y, respectivamente.

Outra medida relacionada à distância Euclidiana é a distância Euclidiana quadrática:

$$d_{sqreuclid} = d^2_{euclid} = \sum_{j=1}^{p}(\mathrm{x_j}-\mathrm{y_j})^2 = (\mathbf{x}-\mathbf{y})(\mathbf{x}-\mathbf{y})^T \tag{4.2}$$

Note que, pelo rigor formal matemático, a distância Euclidiana quadrática não é de fato uma distância (*stricto sensu*).

### 4.3.2 Distância Manhattan

A distância Manhattan também é conhecida na literatura como "*city block distance*". Para duas observações x e y em um espaço $p$-dimensional, a distância Manhattan entre elas é definida como: (Gan et al., 2007)

$$d_{manhat}(\mathrm{x},\mathrm{y}) = \sum_{j=1}^{p}|\mathrm{x_j}-\mathrm{y_j}| \tag{4.3}$$

sendo $\mathrm{x_j}$ e $\mathrm{y_j}$ os valores da $j$-ésima característica de x e y, respectivamente.

Se as observações x ou y possuem *missing values* em alguma característica, então a distância Manhattan pode ser definida como: (Gan et al., 2007)

$$d_{manhatw}(\mathrm{x},\mathrm{y}) = \sum_{j=1}^{p}\frac{w_j|\mathrm{x_j}-\mathrm{y_j}|}{\sum_{j=1}^{p}w_j} \tag{4.4}$$

sendo

- $w_j = 1$, se ambas observações x e y possuem observações na $j$-ésima característica;
- $w_j = 0$, caso contrário.

### 4.3.3 Distância Minkowski

As distâncias Euclidiana e Manhattan são dois casos particulares da distância Minkowski definida como: (Gan et al., 2007)

$$d_{mink}(\mathrm{x},\mathrm{y}) = \left[\sum_{j=1}^{p} |\mathrm{x_j} - \mathrm{y_j}|^r\right]^{1/r}, \; r \geq 1 \tag{4.5}$$

$r$ é denominado ordem da distância Minkowski.

Note que:

- para $r = 2$, tem-se a distância Euclidiana;
- para $r = 1$, tem-se a distância Manhattan.

A distância Minkowski é menos afetada pela presença de valores discrepantes (*outliers*) na amostra do que a distância Euclidiana (Mingoti, 2005). Se o banco de dados possui *clusters* com valores muito próximos, a distância Minkowski funciona plenamente, caso contrário, as observações discrepantes tendem a influenciar as outras. Para corrigir esse problema, opta-se pela normalização das observações ou pelo uso de pesos que corrijam essa tendência. (Gan et al., 2007)

### 4.3.4 Distância Mahalanobis

A distância Mahalanobis é utilizada quando há combinação linear entre características. (Gan et al., 2007)

$$d_{mahal}(\mathrm{x},\mathrm{y}) = \sqrt{(\mathbf{x}-\mathbf{y})\Sigma^{-1}(\mathbf{x}-\mathbf{y})^T} \tag{4.6}$$

sendo $\Sigma$ a matriz de variância-covariância do banco de dados.

Outra propriedade importante da distância Mahalanobis é que ela é invariante a qualquer transformação não-singular. A desvantagem na utilização da distância de Mahalanobis está no fato de que ela envolve um esforço computacional grande já que

a matrix de variância-covariância é obtida através de todas as observações do banco de dados. (Gan et al., 2007)

## 4.4 Medidas para dados Binários

Dados binários são observações que podem receber exatamente dois valores: "sim/não", "0/1", "masculino/feminino".

Algumas vezes, as observações binárias são tratadas como escalares, ou seja, utiliza-se distância Euclidiana ou Manhattan. Embora isso acarrete a obtenção de bons resultados, existem formas de tratamento das observações especialmente desenvolvidas para elas.

Seguindo a abordagem proposta por Gower (Kaufman and Rousseeuw, 1990), as observações binárias podem ser divididas em dois tipos:

- **Simétricas:** os dois valores assumidos por cada variável são igualmente importantes. Observações binárias simétricas são observações nominais.
  Exemplos: "casado/solteiro", "canhoto/destro", "macho/fêmea";

- **Assimétricas:** um dos valores carrega mais importância do que o outro.
  Exemplo: Para a cor de uma flor, considera-se: "é vermelha"$= 1$, "não é vermelha" $= 0$, sendo que, se $x_{if} = 1$ e $x_{jf} = 1$, significa dizer que as flores $i$ e $j$ possuem a mesma cor, enquanto que, $x_{if} = 0$ e $x_{jf} = 0$ implica que as flores possuem cores completamente diferentes (não, necessariamente, indica que essas cores diferentes do vermelho sejam iguais).

Sejam $\mathbf{x}$ e $\mathbf{y}$ dois vetores binários $p$-dimensionais e sejam A, B, C, D e $\sigma$ definidos como:

$$A = S_{11}(\mathrm{x},\mathrm{y}) = \mathbf{x}.\mathbf{y} = \sum_{i=1}^{p} x_i y_i \tag{4.7a}$$

$$B = S_{01}(\mathrm{x},\mathrm{y}) = \bar{\mathbf{x}}.\mathbf{y} = \sum_{i=1}^{p} (1-x_i) y_i \tag{4.7b}$$

$$C = S_{10}(\mathrm{x},\mathrm{y}) = \mathbf{x}.\bar{\mathbf{y}} = \sum_{i=1}^{p} x_i (1-y_i) \tag{4.7c}$$

$$D = S_{00}(\mathrm{x},\mathrm{y}) = \bar{\mathbf{x}}.\bar{\mathbf{y}} = \sum_{i=1}^{p} (1-x_i)(1-y_i) \tag{4.7d}$$

$$\sigma = \sqrt{(A+B)(A+C)(B+D)(C+D)} \tag{4.7e}$$

sendo $S_{ij}(\mathrm{x},\mathrm{y})$, $i,j \in \{0,1\}$, o número de ocorrências de combinações $i$ em $\mathbf{x}$ e $j$ em $\mathbf{y}$ em cada variável, isto é, $S_{ij}(\mathrm{x},\mathrm{y}) = |\{k : x_k = i \text{ e } y_k = j,\ k = 1,...,p\}|$.

Quando trabalha-se com observações assimétricas, outros coeficientes devem ser utilizados. Por convenção, a característica mais importante recebe "1" e as outras "0". Então, quando uma variável recebe dois $1's$ (par 11) é considerada mais significante do que quando recebe dois $0's$ (par 00). Portanto, valores "0/1" são dados de maneira que A, o número de combinações favoráveis carregue mais importância do que D. No entanto, existem coeficientes que são invariantes a esta característica das observações assimétricas. O mais utilizado deles é o de Jaccard.

Tabela 4.1: Medidas de Similaridade para observações binárias. $d(\mathbf{x},\mathbf{y})$ é a respectiva medida de dissimilaridade.

| Medida | $s(\mathbf{x},\mathbf{y})$ | Amplitude de $s(\mathbf{x},\mathbf{y})$ | $d(\mathbf{x},\mathbf{y})$ |
|---|---|---|---|
| Jaccard | $\frac{A}{A+B+C}$ | $[0,1]$ | $\frac{B+C}{A+B+C}$ |
| Russel-Rao | $\frac{A}{d}$ | $[0,1]$ | $1-\frac{A}{d}$ |
| Kulzinsky | $\frac{A}{B+C}$ | $[0,\infty]$ | $\frac{B+C-A+d}{B+C+d}$ |

# 5 Técnicas Hierárquicas de Agrupamento

***Hard clusterings*** (seção 3.3) podem ser divididos em métodos hierárquicos e não-hierárquicos. Métodos hierárquicos dividem as observações em uma sequência (rede) de partições, enquanto que os não-hierárquicos dividem as observações em um única partição.

Os métodos hierárquicos dividem-se em dois tipos:

- **Aglomerativo:** parte do princípio de que no início do agrupamento, tem-se um *cluster* para cada observação, ou seja, cada observação é considerada como sendo um *cluster* isolado. A cada passo, as observações vão sendo agrupadas ao seu par mais similar de acordo com algum critério de similaridade, previamente escolhido, até o momento em que todas as observações encontram-se num único *cluster*;

- **Divisivo:** é justamente o oposto do aglomerativo. No início tem-se um único *cluster* com todas as observações, ou seja, todas as observações constituem um *cluster*. A cada passo, as observações vão se separando de acordo com algum critério de similaridade, previamente escolhido, até o momento em que cada observação forma um *cluster* isoladamente.

Existem algumas desvantagens quando da utilização de métodos hierárquicos(Gan et al., 2007)

- observações agrupadas de forma indevida em estágios anteriores não podem ser realocadas;

- Medidas de similaridade diferentes levam a resultados diferentes.

Devido à propriedade de hierarquia, é possível construir dendrogramas ou "árvores" para representar o histórico do agrupamento. O dendrograma é um gráfico em forma de

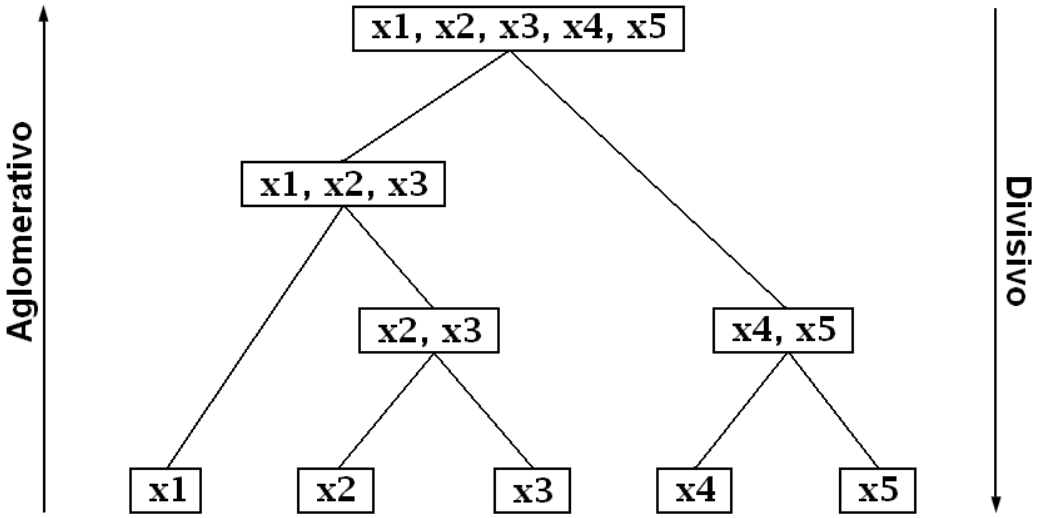



Figura 5.1: Método Hierárquico aglomerativo e divisivo

árvore, em que cada altura indica o nível de similaridade (ou dissimilaridade) em que as observações foram consideradas semelhantes, isto é, o coeficiente de similaridade.

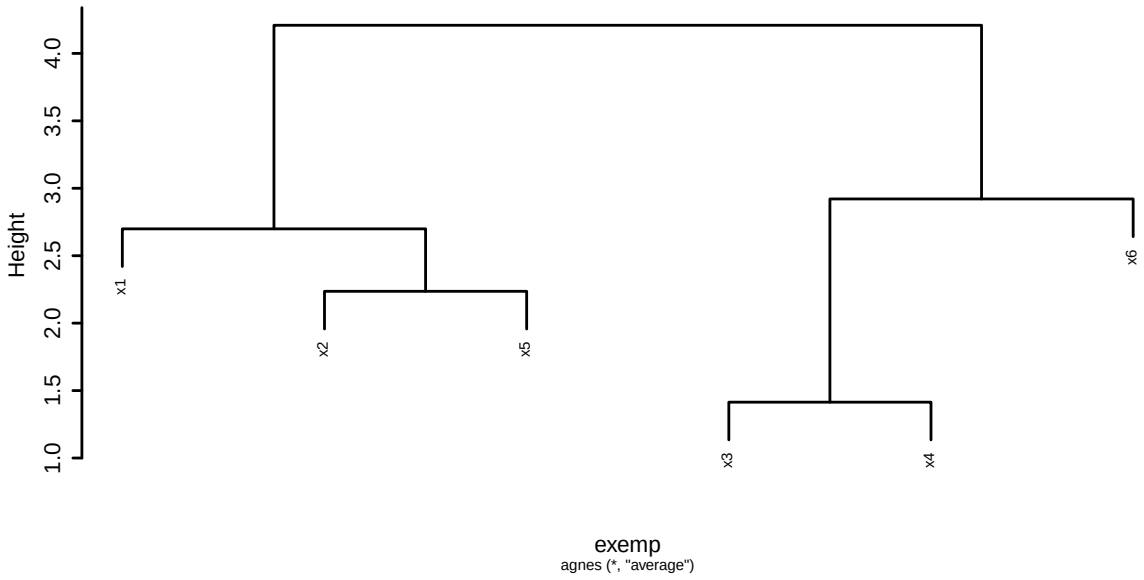



Figura 5.2: Exemplo de um dendrograma

Como ilustração, a figura 5.2 mostra um dendrograma com seis observações. Para cada par de observações $(\mathbf{x_i}, \mathbf{x_j})$, seja $h_{ij}$ a altura do nó especificando o menor *cluster* ao

qual pertencem $(\mathbf{x_i}, \mathbf{x_j})$. Então, um valor pequeno de $h_{ij}$ indica que os elementos $(\mathbf{x_i}, \mathbf{x_j})$ são altamente homogêneos.

## 5.1 Técnicas Hierárquicas Aglomerativas

Partem do princípio de que no início do agrupamento, tem-se um *cluster* para cada observação, ou seja, cada observação é considerada como sendo um *cluster* isolado.

Passos para aplicação das técnicas aglomerativas: (Mingoti, 2005)

1. Cada elemento consitui um *cluster*, tem-se portanto, $n$ *clusters*;

2. A cada estágio do agrupamento, pares de observações são combinados e passam a formar um novo *cluster*. A cada passo, somente um *cluster* pode ser formado. Assim, em cada estágio, o número de *clusters* vai diminuindo até o momento em que todas as observações encontram-se num único *cluster*.

### 5.1.1 Método de Ligação Simples (*Single-Link Method*)

O método da ligação simples é um dos métodos hierárquicos de mais simples aplicação. Neste método, a similaridade entre dois *clusters* é definida pelas duas observações mais homogêneas entre si.

Ele emprega a distância entre os elementos amostrais mais próximos como medida de dissimilaridade entre dois grupos (Gan et al., 2007)

$$\begin{aligned} D(C_k, C_i \cup C_j) &= \frac{1}{2}D(C_k, C_i) + \frac{1}{2}D(C_k, C_j) - \frac{1}{2}|D(C_k, C_i) - D(C_k, C_j)| \\ &= min\{D(C_k, C_i), D(C_k, C_j)\} \\ &= \min_{\mathbf{x}\in C, \mathbf{y}\in C'} d(\mathbf{x}, \mathbf{y}) \end{aligned} \tag{5.1}$$

sendo $D(\cdot,\cdot)$ a distância entre dois *clusters*. $C_k$, $C_i$ e $C_j$ três *clusters* não-vazios, $C' = C_i \cup C_j$ e $d(\cdot,\cdot)$ a medida de dissimilaridade.

Considere o seguinte conjunto de dados:

**Passo 1:** Temos que para esses indivíduos, aparentemente, a relação entre renda e idade deve se dar da seguinte forma: (figura 5.3)

Tabela 5.1: Banco de dados hipotético contendo seis observações

| Indivíduos | Renda | Idade |
|---|---|---|
| x1 | 9 | 25 |
| x2 | 7.4 | 32 |
| x3 | 2.5 | 40 |
| x4 | 18.2 | 38 |
| x5 | 3.8 | 27 |
| x6 | 5 | 43 |

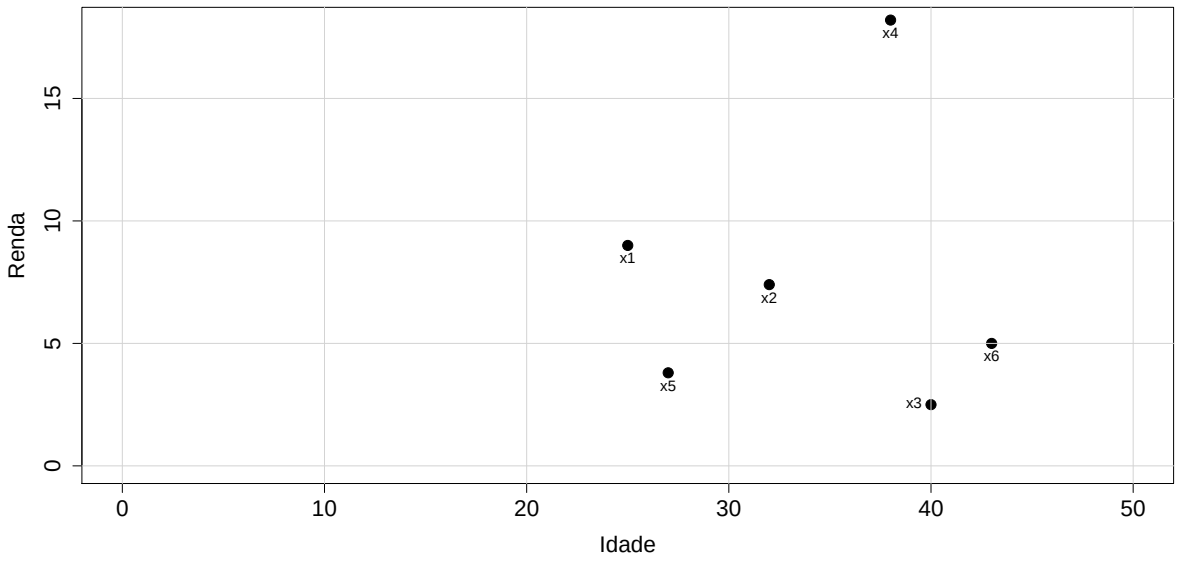



Figura 5.3: Dispersão das observações

Tabela 5.2: Matriz de Dissimilaridade para os dados da tabela 5.1

| | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ |
|---|---|---|---|---|---|---|
| $x_1$ | 0.00 | 7.18 | 16.35 | 15.93 | 5.57 | 18.44 |
| $x_2$ | 7.18 | 0.00 | 9.38 | 12.35 | 6.16 | 11.26 |
| $x_3$ | 16.35 | 9.38 | 0.00 | 15.83 | 13.06 | **3.91** |
| $x_4$ | 15.93 | 12.35 | 15.83 | 0.00 | 18.12 | 14.12 |
| $x_5$ | 5.57 | 6.16 | 13.06 | 18.12 | 0.00 | 16.04 |
| $x_6$ | 18.44 | 11.26 | **3.91** | 14.12 | 16.04 | 0.00 |

Se o método da ligação simples é aplicado a estes dados, então $x_3$ e $x_6$ serão agrupados para formar um *cluster* maior no primeiro estágio, já que eles possuem a menor distância na matriz de dissimilaridade.

A distância entre $\{x_3,x_6\}$ e $x_1$, $x_2$, $x_4$ e $x_5$ será:

$$D(\{x_3,x_6\},x_1) = min\{d(x_3,x_1),d(x_6,x_1)\} = min\{16.35,18.44\} = 16.35$$
$$D(\{x_3,x_6\},x_2) = min\{d(x_3,x_2),d(x_6,x_2)\} = min\{9.38,11.26\} = 9.38$$
$$D(\{x_3,x_6\},x_4) = min\{d(x_3,x_4),d(x_6,x_4)\} = min\{15.83,14.12\} = 14.12$$
$$D(\{x_3,x_6\},x_5) = min\{d(x_3,x_5),d(x_6,x_5)\} = min\{13.06,16.04\} = 13.06$$

**Passo 2:** Depois de agrupar $x_3$ e $x_6$, a matriz de dissimilaridade fica:

| | $\{x_3,x_6\}$ | $x_1$ | $x_2$ | $x_4$ | $x_5$ |
|---|---|---|---|---|---|
| $\{x_3,x_6\}$ | 0.00 | 16.35 | 9.38 | 14.12 | 13.06 |
| $x_1$ | 16.35 | 0.00 | 7.18 | 15.93 | **5.57** |
| $x_2$ | 9.38 | 7.18 | 0.00 | 12.35 | 6.16 |
| $x_4$ | 14.12 | 15.93 | 12.35 | 0.00 | 18.12 |
| $x_5$ | 13.06 | **5.57** | 6.16 | 18.12 | 0.00 |

No segundo estágio, $x_1$ e $x_5$ serão agrupados, já que a distância entre elas é a menor. Então, a distância entre $\{x_1,x_5\}$ e as observações restantes é

$$D(\{x_1,x_5\},\{x_3,x_6\}) = min\{d(\{x_3,x_6\},x_1),d(\{x_3,x_6\},x_5)\} = 13.06$$
$$D(\{x_1,x_5\},x_2) = min\{d(x_2,x_1),d(x_2,x_5)\} = 6.16$$
$$D(\{x_1,x_5\},x_4) = min\{d(x_4,x_1),d(x_4,x_5)\} = 15.93$$

**Passo 3:** Depois de agrupar $x_1$ e $x_5$, a matriz de dissimilaridade fica:

| | $\{x_3,x_6\}$ | $\{x_1,x_5\}$ | $x_2$ | $x_4$ |
|---|---|---|---|---|
| $\{x_3,x_6\}$ | 0.00 | 13.06 | 9.38 | 14.12 |
| $\{x_1,x_5\}$ | 13.06 | 0.00 | **6.16** | 15.93 |
| $x_2$ | 9.38 | **6.16** | 0.00 | 12.35 |
| $x_4$ | 14.12 | 15.93 | 12.35 | 0.00 |

No terceiro estágio, $\{x_1,x_5\}$ e $x_2$ serão agrupados, já que possuem a menor distância. Assim, a distância entre $\{x_1,x_2,x_5\}$ e as observações restantes é

$$D(\{x_1,x_2,x_5\},\{x_3,x_6\}) = 9.38$$
$$D(\{x_1,x_2,x_5\},x_4) = 12.35$$

**Passo 4:** Depois de agrupar $\{x_1, x_5\}$ e $x_2$, a matriz de dissimilaridade fica:

| | $\{x_1, x_2, x_5\}$ | $\{x_3, x_6\}$ | $x_4$ |
|---|---|---|---|
| $\{x_1, x_2, x_5\}$ | 0.00 | **9.38** | 12.35 |
| $\{x_3, x_6\}$ | **9.38** | 0.00 | 14.12 |
| $x_4$ | 12.35 | 14.12 | 0.00 |

No quarto estágio, $\{x_1, x_2, x_5\}$ e $x_3, x_6$ serão agrupados. Com isso, a distância entre $\{x_1, x_2, x_3, x_5, x_6\}$ e as observações restantes é

$$D(\{x_1, x_2, x_3, x_5, x_6\}, x_4) = 12.35$$

**Passo 5:** Depois de agrupar $\{x_1, x_2, x_5\}$ e $x_3, x_6$, a matriz de dissimilaridade fica:

| | $\{x_1, x_2, x_3, x_5, x_6\}$ | $x_4$ |
|---|---|---|
| $\{x_1, x_2, x_3, x_5, x_6\}$ | 0.00 | **12.35** |
| $x_4$ | **12.35** | 0.00 |

No quinto estágio, todos os pontos se unem em um único *cluster*. O respectivo dendrograma deste cluster é mostrado na figura 5.4

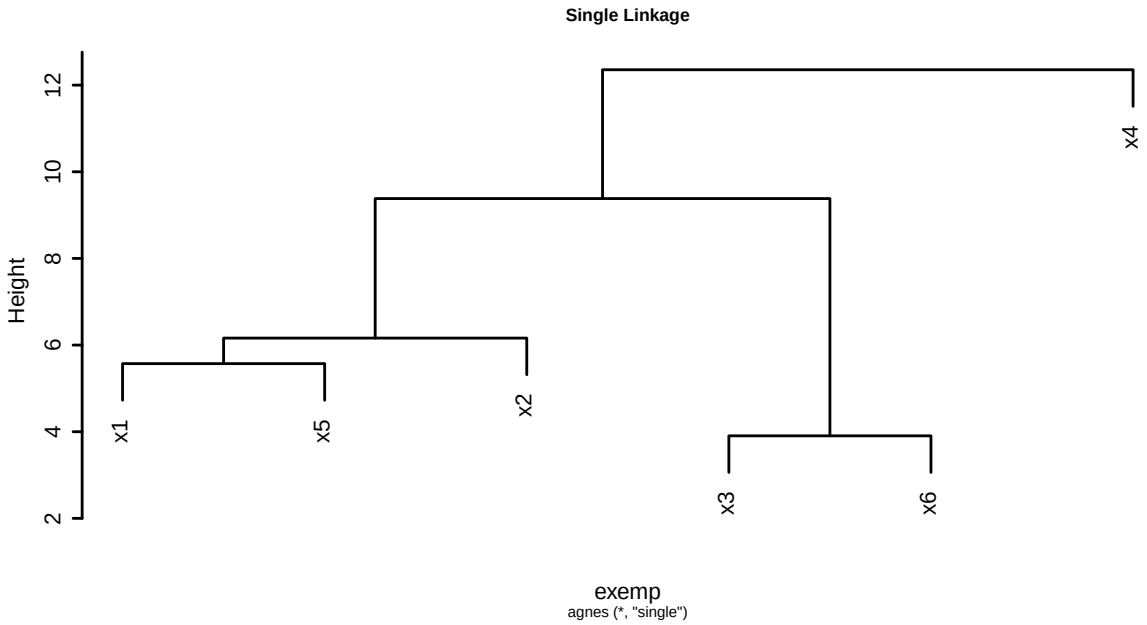



Figura 5.4: Dendrograma produzido ao aplicar o método de ligação simples

### 5.1.2 Método de Ligação Completa (*Complete Link Method*)

Diferentemente do método de ligação simples, neste método, a similaridade entre dois *clusters* é definida pelas duas observações mais heterogêneas entre si. Em cada estágio a distância é calculada para todos os pares de *cluster*, sendo agrupados aqueles que tiverem o menor valor da distância. O método de ligação completa, também é invariante à transformações monótonas. Sejam $C_k$, $C_i$ e $C_j$ três *clusters* não-vazios. A distância entre $C_k$ e $C_i \cup C_j$ é dada por: (Gan et al., 2007)

$$\begin{aligned} D(C_k, C_i \cup C_j) &= \frac{1}{2}D(C_k, C_i) + \frac{1}{2}D(C_k, C_j) + \frac{1}{2}|D(C_k, C_i) - D(C_k, C_j)| \\ &= max\{D(C_k, C_i), D(C_k, C_j)\} \\ &= \max_{\mathbf{x}\in C, \mathbf{y}\in C'} d(\mathbf{x}, \mathbf{y}) \end{aligned} \tag{5.2}$$

sendo $D(\cdot,\cdot)$ a distância entre dois *clusters*, $C' = C_i \cup C_j$ e $d(\cdot,\cdot)$ a medida de dissimilaridade.

**Passo 1:** Aplicando o método de ligação completa à matriz de dissimilaridade dada na tabela 5.4, no primeiro estágio, agrupa-se $x_3$ e $x_6$, assim como no método anterior. A distância entre $\{x_3, x_6\}$ e $x_1$, $x_2$, $x_4$ e $x_5$ será:

$$D(\{x_3,x_6\},x_1) = min\{d(x_3,x_1), d(x_6,x_1)\} = max\{16.35, 18.44\} = 18.44$$
$$D(\{x_3,x_6\},x_2) = min\{d(x_3,x_2), d(x_6,x_2)\} = max\{9.38, 11.26\} = 11.26$$
$$D(\{x_3,x_6\},x_4) = min\{d(x_3,x_4), d(x_6,x_4)\} = max\{15.83, 14.12\} = 15.83$$
$$D(\{x_3,x_6\},x_5) = min\{d(x_3,x_5), d(x_6,x_5)\} = max\{13.06, 16.04\} = 16.04$$

**Passo 2:** Depois de agrupar $x_3$ e $x_6$, a matriz de dissimilaridade fica:

| | $\{x_3, x_6\}$ | $x_1$ | $x_2$ | $x_4$ | $x_5$ |
|---|---|---|---|---|---|
| $\{x_3, x_6\}$ | 0.00 | 18.44 | 11.26 | 15.83 | 16.04 |
| $x_1$ | 18.44 | 0.00 | 7.18 | 15.93 | **5.57** |
| $x_2$ | 11.26 | 7.18 | 0.00 | 12.35 | 6.16 |
| $x_4$ | 15.83 | 15.93 | 12.35 | 0.00 | 18.12 |
| $x_5$ | 16.04 | **5.57** | 6.16 | 18.12 | 0.00 |

No segundo estágio, $x_1$ e $x_5$ serão agrupados, já que possuem a menor distância. Depois de agrupá-los, a distância entre o *cluster* $\{x_1,x_5\}$ e os *clusters* restantes $\{x_3,x_6\}$, $x_2$ e $x_4$ será:

$$D(\{x_1,x_5\},\{x_3,x_6\}) = max\{d(\{x_3,x_6\},x_1),d(\{x_3,x_6\},x_5) = 18.44$$
$$D(\{x_1,x_5\},x_2) = max\{d(x_1,x_2),d(x_5,x_2)\} = 7.18$$
$$D(\{x_1,x_5\},x_4) = max\{d(x_1,x_4),d(x_5,x_4)\} = 18.12$$

**Passo 3:** Depois de agrupar $x_2$ e $x_5$, a matriz de dissimilaridade fica:

| | $\{x_3,x_6\}$ | $\{x_1,x_5\}$ | $x_2$ | $x_4$ |
|---|---|---|---|---|
| $\{x_3,x_6\}$ | 0.00 | 18.44 | 9.38 | 15.83 |
| $\{x_1,x_5\}$ | 18.44 | 0.00 | **7.18** | 18.12 |
| $x_2$ | 9.38 | **7.18** | 0.00 | 12.35 |
| $x_4$ | 15.83 | 18.12 | 12.35 | 0.00 |

No terceiro estágio, $\{x_1,x_5\}$ e $x_2$ serão agrupados, já que possuem a menor distância. Assim, a distância entre $\{x_1,x_2,x_5\}$ e as observações restantes é

$$D(\{x_1,x_2,x_5\},\{x_3,x_6\}) = 18.44$$
$$D(\{x_1,x_2,x_5\},x_4) = 18.12$$

**Passo 4:** Depois de agrupar $\{x_1,x_5\}$ e $x_2$, a matriz de dissimilaridade fica:

| | $\{x_1,x_2,x_5\}$ | $\{x_3,x_6\}$ | $x_4$ |
|---|---|---|---|
| $\{x_1,x_2,x_5\}$ | 0.00 | 18.44 | 18.12 |
| $\{x_3,x_6\}$ | 18.44 | 0.00 | **15.83** |
| $x_4$ | 18.12 | **15.83** | 0.00 |

No quarto estágio, $\{x_3,x_6\}$ e $x_4$ serão agrupados. Com isso, a distância entre $\{x_3,x_4,x_6\}$ e as observações restantes é

$$D(\{x_1,x_2,x_5\},\{x_3,x_4,x_6\}) = 18.44$$

**Passo 5:** Depois de agrupar $\{x_3,x_6\}$ e $x_4$, a matriz de dissimilaridade fica:

| | $\{x_3, x_4, x_6\}$ | $\{x_1, x_2, x_5\}$ |
|---|---|---|
| $\{x_3, x_4, x_6\}$ | 0.00 | **18.44** |
| $\{x_1, x_2, x_5\}$ | **18.44** | 0.00 |

No quinto estágio, todos os pontos se unem em um único *cluster*. O respectivo dendrograma deste cluster é mostrado na figura 5.5

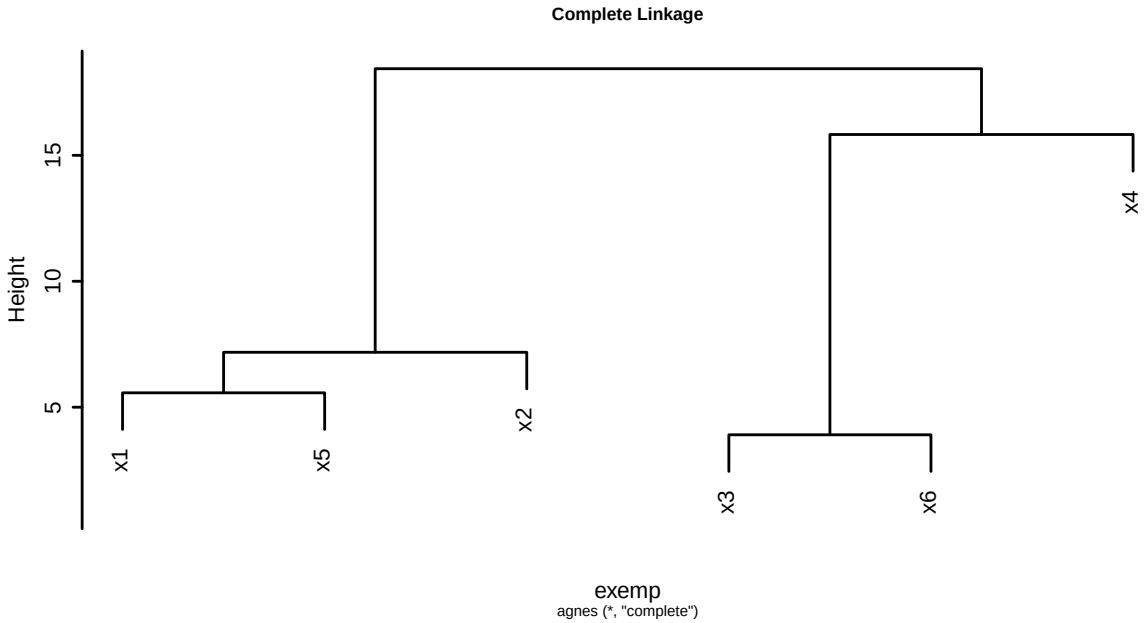



Figura 5.5: Dendrograma produzido ao aplicar o método de ligação completa

### 5.1.3 Método da Média dos Grupos (*Group Average Method*)

O método da média dos grupos é também conhecido por UPGMA (*unweighted pair group method using arithmetic averages*). Nele, a distância entre dois grupos é definida como sendo a média da distância entre todos os possíveis pares de observações. Sejam $C_k$, $C_i$ e $C_j$ três *clusters* não-vazios. A distância entre $C_k$ e $C_i \cup C_j$ é dada por: (Gan et al., 2007)

$$\begin{aligned} D(C_k, C_i \cup C_j) &= \frac{|C_i|}{|C_i| + |C_j|} D(C_k, C_i) + \frac{|C_j|}{|C_i| + |C_j|} D(C_k, C_j) \\ &= \frac{1}{|C||C'|} \sum_{x \in C, \mathbf{y} \in C'} d(\mathbf{x}, \mathbf{y}) \end{aligned} \tag{5.3}$$

sendo $D(\cdot,\cdot)$ a distância entre dois *clusters* e $d(\cdot,\cdot)$ a medida de dissimilaridade e $C' = C_i \cup C_j$ não-vazio.

**Passo 1:** Aplicando o método da média dos grupos à matriz de dissimilaridade dada na tabela 5.4, no primeiro estágio, agrupa-se $x_3$ e $x_6$, assim como no método anterior. A distância entre $\{x_3,x_6\}$ e $x_1$, $x_2$, $x_4$ e $x_5$ será:

$$D(\{x_3,x_6\},x_1) = \frac{1}{2}d(x_3,x_1) + \frac{1}{2}d(x_6,x_1) = 17.395$$
$$D(\{x_3,x_6\},x_2) = \frac{1}{2}d(x_3,x_2) + \frac{1}{2}d(x_6,x_2) = 10.320$$
$$D(\{x_3,x_6\},x_4) = \frac{1}{2}d(x_3,x_4) + \frac{1}{2}d(x_6,x_4) = 14.975$$
$$D(\{x_3,x_6\},x_5) = \frac{1}{2}d(x_3,x_5) + \frac{1}{2}d(x_6,x_5) = 14.550$$

**Passo 2:** Depois de agrupar $x_3$ e $x_6$, a matriz de dissimilaridade fica:

| | $\{x_3,x_6\}$ | $x_1$ | $x_2$ | $x_4$ | $x_5$ |
|---|---|---|---|---|---|
| $\{x_3,x_6\}$ | 0.00 | 17.395 | 10.320 | 14.975 | 14.550 |
| $x_1$ | 17.395 | 0.00 | 7.18 | 15.93 | **5.57** |
| $x_2$ | 10.320 | 7.18 | 0.00 | 12.35 | 6.16 |
| $x_4$ | 14.975 | 15.93 | 12.35 | 0.00 | 18.12 |
| $x_5$ | 14.500 | **5.57** | 6.16 | 18.12 | 0.00 |

No segundo estágio, $x_1$ e $x_5$ serão agrupados, já que possuem a menor distância. Depois de agrupá-los, a distância entre o *cluster* $\{x_1,x_5\}$ e os *clusters* restantes $\{x_3,x_6\}$, $x_2$ e $x_4$ será:

$$D(\{x_1,x_5\},\{x_3,x_6\}) = \frac{1}{2}d(\{x_3,x_6\},x_1) + \frac{1}{2}d(\{x_3,x_6\},x_5) = 15.973$$
$$D(\{x_1,x_5\},x_2\}) = \frac{1}{2}d(x_1,x_2) + \frac{1}{2}d(x_5,x_2) = 6.670$$
$$D(\{x_1,x_5\},x_4\}) = \frac{1}{2}d(x_1,x_4) + \frac{1}{2}d(x_5,x_4) = 17.395$$

**Passo 3:** Depois de agrupar $x_1$ e $x_5$, a matriz de dissimilaridade fica:

| | $\{x_3,x_6\}$ | $\{x_1,x_5\}$ | $x_2$ | $x_4$ |
|---|---|---|---|---|
| $\{x_3,x_6\}$ | 0.00 | 15.973 | 10.320 | 14.975 |
| $\{x_1,x_5\}$ | 15.973 | 0.00 | **6.670** | 17.395 |
| $x_2$ | 10.320 | **6.670** | 0.00 | 12.35 |
| $x_4$ | 14.975 | 17.395 | 12.35 | 0.00 |

No terceiro estágio, $\{x_1,x_5\}$ e $x_2$ serão agrupados, já que possuem a menor distância. Assim, a distância entre $\{x_1,x_2,x_5\}$ e as observações restantes é

$$D(\{x_1,x_2,x_5\},\{x_3,x_6\}) = \frac{1}{6}(d_{13}+d_{16}+d_{23}+d_{26}+d_{53}+d_{56}) = 14,089$$
$$D(\{x_1,x_2,x_5\},x_4\}) = \frac{1}{3}(d_{14}+d_{24}+d_{54}) = 15.467$$

**Passo 4:** Depois de agrupar $\{x_1,x_5\}$ e $x_2$, a matriz de dissimilaridade fica:

| | $\{x_1,x_2,x_5\}$ | $\{x_3,x_6\}$ | $x_4$ |
|---|---|---|---|
| $\{x_1,x_2,x_5\}$ | 0.00 | **14.089** | 15.467 |
| $\{x_3,x_6\}$ | **14.089** | 0.00 | 14.975 |
| $x_4$ | 15.467 | 14.975 | 0.00 |

No quarto estágio, $\{x_1,x_2,x_5\}$ e $\{x_3,x_6\}$ serão agrupados. Com isso, a distância entre $\{x_1,x_2,x_3,x_5,x_6\}$ e as observações restantes é

$$D(\{x_1,x_2,x_3,x_5,x_6\},x_4) = 15.270$$

**Passo 5:** Depois de agrupar $\{x_1,x_2,x_5\}$ e $\{x_3,x_6\}$, a matriz de dissimilaridade fica:

| | $\{x_1,x_2,x_3,x_5,x_6\}$ | $x_4$ |
|---|---|---|
| $\{x_1,x_2,x_3,x_5,x_6\}$ | 0.00 | **15.270** |
| $x_4$ | **15.270** | 0.00 |

No quinto estágio, todos os pontos se unem em um único *cluster*. O respectivo dendrograma deste *cluster* é mostrado na figura 5.6

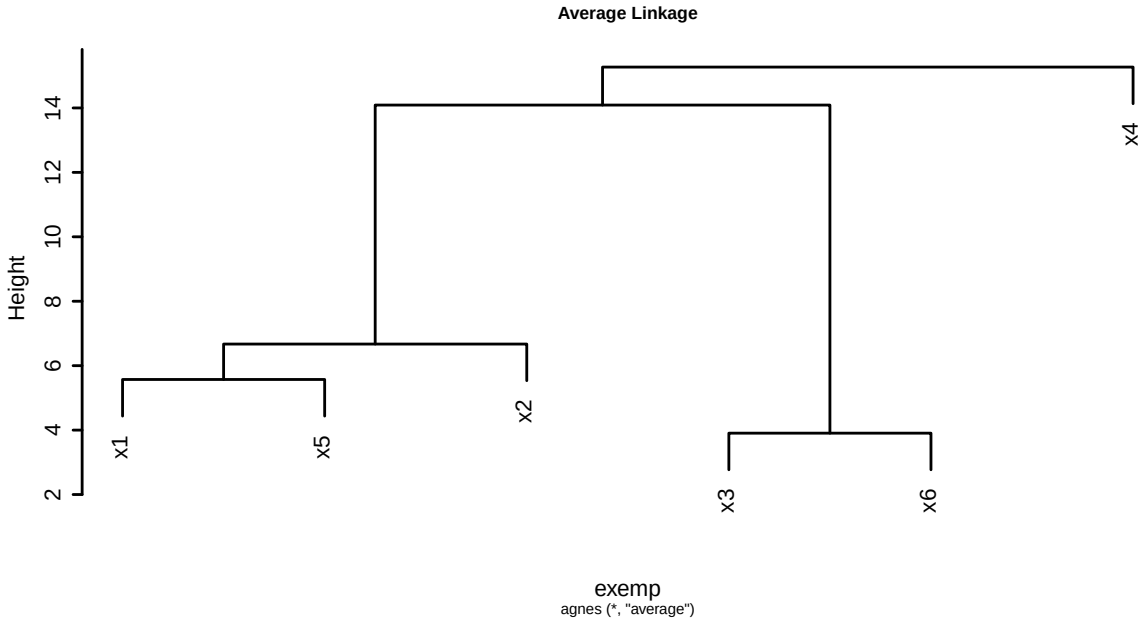



Figura 5.6: Dendrograma produzido ao aplicar o método da média dos grupos

### 5.1.4 Método da Média Ponderada dos Grupos (*Weighted Group Average Method*)

O método da média ponderada é também conhecido por (*weighted pair group method using arithmetic averages*). Nele, a distância entre dois grupos é definida como sendo a média ponderada da distância entre todos os possíveis pares de observações. Sejam $C_k$, $C_i$ e $C_j$ três *clusters* não-vazios e no mesmo nível de agrupamento. A distância entre $C_k$ e $C_i \cup C_j$ é dada por: (Gan et al., 2007)

$$D(C_k, C_i \cup C_j) = \frac{1}{2} D(C_k, C_i) + \frac{1}{2} D(C_k, C_j) \tag{5.4}$$

sendo $D(\cdot,\cdot)$ a distância entre dois clusters e $d(\cdot,\cdot)$ a medida de dissimilaridade.

**Passo 1:** Aplicando o método da média ponderada à matriz de dissimilaridade dada na tabela 5.4, no primeiro estágio, agrupa-se $x_3$ e $x_6$, assim como no método anterior. A distância entre $\{x_3,x_6\}$ e $x_1$, $x_2$, $x_4$ e $x_5$ será:

$$D(\{x_3,x_6\},x_1) = \frac{1}{2}d(x_3,x_1) + \frac{1}{2}d(x_6,x_1) = 17.395$$
$$D(\{x_3,x_6\},x_2) = \frac{1}{2}d(x_3,x_2) + \frac{1}{2}d(x_6,x_2) = 10.320$$
$$D(\{x_3,x_6\},x_4) = \frac{1}{2}d(x_3,x_4) + \frac{1}{2}d(x_6,x_4) = 14.975$$
$$D(\{x_3,x_6\},x_5) = \frac{1}{2}d(x_3,x_5) + \frac{1}{2}d(x_6,x_5) = 14.550$$

**Passo 2:** Depois de agrupar $x_3$ e $x_6$, a matriz de dissimilaridade fica:

| | $\{x_3,x_6\}$ | $x_1$ | $x_2$ | $x_4$ | $x_5$ |
|---|---|---|---|---|---|
| $\{x_3,x_6\}$ | 0.00 | 17.395 | 10.320 | 14.975 | 14.550 |
| $x_1$ | 17.395 | 0.00 | 7.18 | 15.93 | **5.57** |
| $x_2$ | 10.320 | 7.18 | 0.00 | 12.35 | 6.16 |
| $x_4$ | 14.975 | 15.93 | 12.35 | 0.00 | 18.12 |
| $x_5$ | 14.500 | **5.57** | 6.16 | 18.12 | 0.00 |

No segundo estágio, $x_1$ e $x_5$ serão agrupados, já que possuem a menor distância. Depois de agrupá-los, a distância entre o *cluster* $\{x_1,x_5\}$ e os *clusters* restantes $\{x_3,x_6\}$, $x_2$ e $x_4$ será:

$$D(\{x_1,x_5\},\{x_3,x_6\}) = \frac{1}{2}d(\{x_3,x_6\},x_1) + \frac{1}{2}d(\{x_3,x_6\},x_5) = 15.973$$
$$D(\{x_1,x_5\},x_2\}) = \frac{1}{2}d(x_1,x_2) + \frac{1}{2}d(x_5,x_2) = 6.670$$
$$D(\{x_1,x_5\},x_4\}) = \frac{1}{2}d(x_1,x_4) + \frac{1}{2}d(x_5,x_4) = 17.395$$

**Passo 3:** Depois de agrupar $x_1$ e $x_5$, a matriz de dissimilaridade fica:

| | $\{x_3,x_6\}$ | $\{x_1,x_5\}$ | $x_2$ | $x_4$ |
|---|---|---|---|---|
| $\{x_3,x_6\}$ | 0.00 | 15.973 | 10.320 | 14.975 |
| $\{x_1,x_5\}$ | 15.973 | 0.00 | **6.670** | 17.395 |
| $x_2$ | 10.320 | **6.670** | 0.00 | 12.35 |
| $x_4$ | 14.975 | 17.395 | 12.35 | 0.00 |

No terceiro estágio, $\{x_1, x_5\}$ e $x_2$ serão agrupados, já que possuem a menor distância. Assim, a distância entre $\{x_1, x_2, x_5\}$ e as observações restantes é

$$D(\{x_1, x_2, x_5\}, \{x_3, x_6\}) = \frac{1}{2}[d(\{x_1, x_2, x_5\}) + d(\{x_3, x_6\})] = \frac{1}{2}[15.973 + 10.320] = 13.147$$
$$D(\{x_1, x_2, x_5\}, x_4\}) = \frac{1}{2}[d(\{x_1, x_5\}, x_4\}) + d(\{x_2, x_4\})]\frac{1}{2}[17.395 + 12.35] = 14.873$$

**Passo 4:** Depois de agrupar $\{x_1, x_5\}$ e $x_2$, a matriz de dissimilaridade fica:

| | $\{x_1, x_2, x_5\}$ | $\{x_3, x_6\}$ | $x_4$ |
|---|---|---|---|
| $\{x_1, x_2, x_5\}$ | 0.00 | **13.147** | 14.873 |
| $\{x_3, x_6\}$ | **13.147** | 0.00 | 14.975 |
| $x_4$ | 14.873 | 14.975 | 0.00 |

No quarto estágio, $\{x_1, x_2, x_5\}$ e $\{x_3, x_6\}$ serão agrupados. Com isso, a distância entre $\{x_1, x_2, x_3, x_5, x_6\}$ e as observações restantes é

$$D(\{x_1, x_2, x_3, x_5, x_6\}, x_4) = 14.924$$

**Passo 5:** Depois de agrupar $\{x_1, x_2, x_5\}$ e $\{x_3, x_6\}$, a matriz de dissimilaridade fica:

| | $\{x_1, x_2, x_3, x_5, x_6\}$ | $x_4$ |
|---|---|---|
| $\{x_1, x_2, x_3, x_5, x_6\}$ | 0.00 | **14.924** |
| $x_4$ | **14.924** | 0.00 |

No quinto estágio, todos os pontos se unem em um único *cluster*. O respectivo dendrograma deste *cluster* é mostrado na figura 5.7

### 5.1.5 Método de Ward (*Ward's Method*)

A medida em que os passos do algoritmo são realizados a qualidade da partição decresce. Logo, o nível de similaridade decresce, ou seja, a variação **entre** grupos diminui e **dentro** dos grupos aumenta.

O método de Ward foi criado de forma a minimizar a perda de informação associada a cada agrupamento. Geralmente, a perda de informação é quantificada em termos da soma de quadrados do erro (SSE). Esta soma de quadrado é o quadrado da distân-

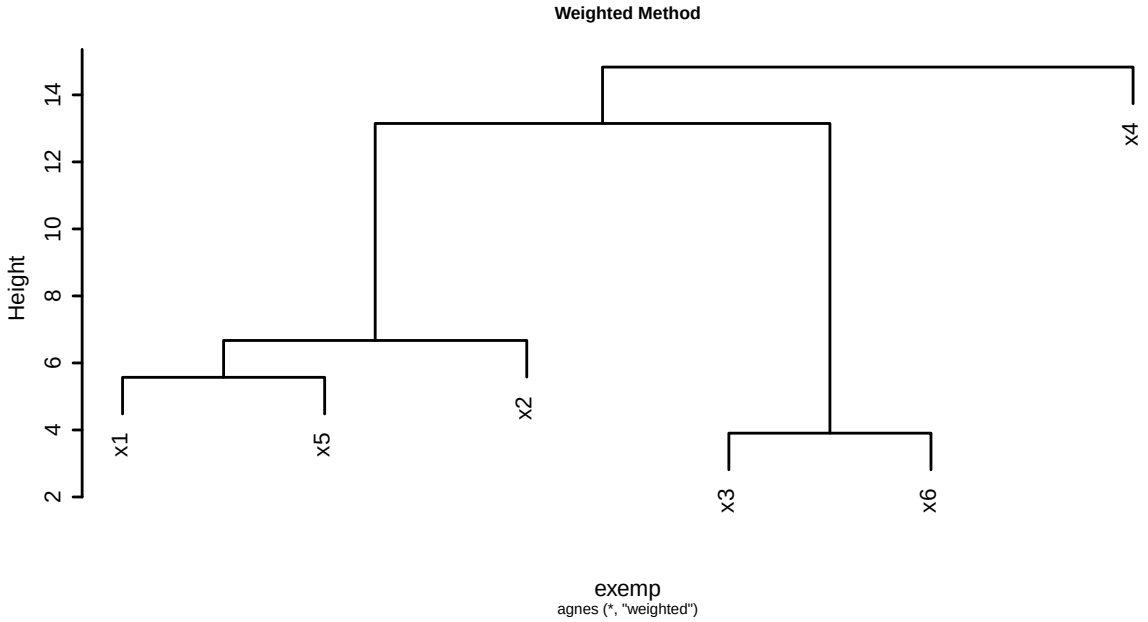



Figura 5.7: Dendrograma produzido ao aplicar o método da média ponderada

cia euclidiana de cada observação pertencente ao *cluster* em relação à média de cada característica.

Em cada passo do algoritmo, todas as possíveis combinações de pares são consideradas e duas observações cuja fusão resulta no menor aumento na perda de informação são agrupados. Se a distância euclidiana quadrática for utilizada no cálculo da matriz de dissimilaridade, então a distância entre duas observações é dada por: (Gan et al., 2007)

$$d_{ij}^2 = d(\mathbf{x_i}, \mathbf{x_j}) = (\mathbf{x_i} - \mathbf{x_j})(\mathbf{x_i} - \mathbf{x_j})^T = \sum_{l=1}^{p} (\mathbf{x_{il}}, \mathbf{x_{jl}})^2 \tag{5.5}$$

sendo $p$ a dimensão do conjunto de dados $D$ e $d(\cdot,\cdot)$ a medida de dissimilaridade. Se $C_i = \{\mathbf{x_i}\}$ e $C_j = \{\mathbf{x_j}\}$, o aumento da SSE resultante da fusão entre $\mathbf{x_i}$ e $\mathbf{x_j}$ é

$$\Delta SSE_{ij} = \frac{1}{2} d_{ij}^2$$

Como o objetivo do método de Ward é encontrar o estágio em que dois *clusters* cuja fusão fornece o menor aumento na SSE dentro do *cluster*, as duas observações com o menor valor no quadrado da distância euclidiana serão agrupados.

Agora seja $C_k = \{\mathbf{x_k}\}$ uma outra observação. O aumento na SSE que resultará na fusão de $C_k = \{\mathbf{x_k}\}$ e $C_i \cup C_j\}$ é dada por por: (Gan et al., 2007)

$$D(C_k, C_i \cup C_j) = \frac{|C_k| + |C_i|}{\sum_{ijk}} D(C_k, C_i) + \frac{|C_k| + |C_j|}{\sum_{ijk}} D(C_k, C_j) - \frac{|C_k|}{\sum_{ijk}} D(C_i, C_j) \tag{5.6}$$

$$\begin{aligned} \Delta SSE_{k(ij)} &= \frac{2}{3} d_{ki}^2 + \frac{2}{3} d_{kj}^2 - \frac{1}{3} d_{ij}^2 \\ &= \frac{1}{2} D(C_k, C_i \cup C_j) \end{aligned} \tag{5.7}$$

Aplicando o quadrado da distância euclidiana à tabela de dados 5.1, obtem-se a seguinte matriz de dissimilaridade:

Tabela 5.3: Matriz de Dissimilaridade para os dados da tabela 5.1 - Quadrado da Distância Euclidiana

| | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ |
|---|---|---|---|---|---|---|
| $x_1$ | 0.00 | 51.56 | 267.25 | 253.64 | 31.04 | 340.00 |
| $x_2$ | 51.56 | 0.00 | 88.01 | 152.64 | 37.96 | 126.76 |
| $x_3$ | 267.25 | 88.01 | 0.00 | 250.49 | 170.69 | **15.25** |
| $x_4$ | 253.64 | 152.64 | 250.49 | 0.00 | 328.36 | 199.24 |
| $x_5$ | 31.04 | 37.96 | 170.69 | 328.36 | 0.00 | 257.44 |
| $x_6$ | 340.00 | 126.76 | **15.25** | 199.24 | 257.44 | 0.00 |

**Passo 1:** Inicialmente, cada observação forma um *cluster* e SSE total é $SSE_0 = 0$. Aplicando o método de Ward à matriz de dissimilaridade dada na tabela 5.3, no primeiro estágio, agrupa-se $x_3$ e $x_6$, e o aumento na SSE resultante dessa fusão é: $\Delta SSE_{12} = \frac{1}{2}(15.25) = 7.625$. Portanto,

$$SSE_1 = SSE_0 + \Delta SSE_{12} = 7.625$$

As distâncias deste novo cluster em relação às variàveis que restaram é:

$$D(\{x_3,x_6\},x_1) = \frac{2}{3}[d(x_3,x_1)+(x_6,x_1)] - \frac{1}{3}d(x_3,x_6) = 399.75$$
$$D(\{x_3,x_6\},x_2) = \frac{2}{3}[d(x_3,x_2)+d(x_6,x_2)] - \frac{1}{3}d(x_3,x_6) = 138.10$$
$$D(\{x_3,x_6\},x_4) = \frac{2}{3}[d(x_3,x_4)+d(x_6,x_4)] - \frac{1}{3}d(x_3,x_6) = 294.74$$
$$D(\{x_3,x_6\},x_5) = \frac{2}{3}[d(x_3,x_5)+d(x_6,x_5)] - \frac{1}{3}d(x_3,x_6) = 280.34$$

**Passo 2:** Depois de agrupar $x_3$ e $x_6$, a matriz de dissimilaridade fica:

| | $\{x_3,x_6\}$ | $x_1$ | $x_2$ | $x_4$ | $x_5$ |
|---|---|---|---|---|---|
| $\{x_3,x_6\}$ | 0.00 | 399.75 | 138.10 | 294.74 | 280.34 |
| $x_1$ | 399.75 | 0.00 | 51.56 | 253.64 | **31.04** |
| $x_2$ | 138.10 | 51.56 | 0.00 | 152.64 | 37.96 |
| $x_4$ | 294.74 | 253.64 | 152.64 | 0.00 | 328.36 |
| $x_5$ | 280.34 | **31.04** | 37.96 | 328.36 | 0.00 |

No segundo estágio, $x_1$ e $x_5$ serão agrupados, já que o aumento na SSE é igual a $\Delta SSE_{15}\frac{1}{2}(31.04) = 15.52$. Portanto,

$$SSE_2 = SSE_1 + \Delta SSE_{15} = 23.145$$

Depois de agrupá-los, a distância entre o *cluster* $\{x_1,x_5\}$ e os *clusters* restantes $\{x_3,x_6\}$, $x_2$ e $x_4$ será:

$$D(\{x_1,x_5\},\{x_3,x_6\}) = \frac{3}{4}[d\{x_3,x_6\},x_1)+d(\{x_3,x_6\},x_5)] - \frac{2}{4}d(x_1,x_5) = 494.55$$
$$D(\{x_1,x_5\},x_2) = \frac{2}{3}[d(x_1,x_2)+d(x_5,x_2)] - \frac{1}{3}d(x_1,x_5) = 49.33$$
$$D(\{x_1,x_5\},x_4) = \frac{2}{3}[d(x_1,x_4)+d(x_5,x_4)] - \frac{1}{3}d(x_1,x_5) = 377.65$$

**Passo 3:** Depois de agrupar $x_1$ e $x_5$, a matriz de dissimilaridade fica:

| | $\{x_3,x_6\}$ | $\{x_1,x_5\}$ | $x_2$ | $x_4$ |
|---|---|---|---|---|
| $\{x_3,x_6\}$ | 0.00 | 494.55 | 138.10 | 294.74 |
| $\{x_1,x_5\}$ | 494.55 | 0.00 | **49.33** | 377.65 |
| $x_2$ | 138.10 | **49.33** | 0.00 | 152.64 |
| $x_4$ | 294.74 | 377.65 | 152.64 | 0.00 |

No terceiro estágio, $\{x_1,x_5\}$ e $x_2$ serão agrupados, já que o aumento na SSE é igual a $\Delta SSE_{(15)2}\frac{1}{2}(49.33) = 15.52$. Portanto,

$$SSE_3 = SSE_2 + \Delta SSE_{(15)2} = 23.145$$

Depois de agrupá-los, a distância entre o *cluster* $\{x_1,x_2,x_5\}$ e os *clusters* restantes $\{x_3,x_6\}$, e $x_4$ será:

$$D(\{x_1,x_2,x_5\},\{x_3,x_6\}) = \frac{4}{5}d\{x_3,x_6\},\{x_1,x_5\}) + \frac{3}{5}d(\{x_3,x_6\},x_2) - \frac{2}{5}d(x_1,x_2,x_5) = 478.5$$
$$D(\{x_1,x_2,x_5\},x_4) = \frac{3}{4}d(\{x_1,x_5\},x_4) + \frac{2}{4}d(x_2,x_4) - \frac{1}{4}d(x_1,x_2,x_5) = 347.23$$

**Passo 4:** Depois de agrupar $\{x_1,x_5\}$ e $x_2$, a matriz de dissimilaridade fica:

| | $\{x_3,x_6\}$ | $\{x_1,x_2,x_5\}$ | $x_4$ |
|---|---|---|---|
| $\{x_3,x_6\}$ | 0.00 | 478.5 | **294.74** |
| $\{x_1,x_2,x_5\}$ | 478.5 | 0.00 | 347.23 |
| $x_4$ | **294.74** | 347.23 | 0.00 |

No quarto estágio, $\{x_3,x_6\}$ e $x_4$ serão agrupados, já que o aumento na SSE é igual a $\Delta SSE_{(36)4}\frac{1}{2}(294.74) = 147.37$. Portanto,

$$SSE_4 = SSE_3 + \Delta SSE_{(36)4} = 170.52$$

Depois de agrupá-los, a distância entre o *cluster* $\{x_3,x_6,x_4\}$ e os *clusters* restantes $\{x_1,x_2,x_5\}$ será:

$$D(\{x_1,x_2,x_5\},\{x_3,x_4,x_6\}) = \frac{5}{6}d\{x_3,x_6\},\{x_1,x_2,x_5\}) + \frac{4}{6}d(\{x_1,x_2,x_5\},x_4) - \frac{3}{6}d(x_3,x_4,x_6) = 482.87$$

**Passo 5:** Depois de agrupar $\{x_3,x_6\}$ e $x_4$, a matriz de dissimilaridade fica:

| | $\{x_3, x_4, x_6\}$ | $\{x_1, x_2, x_5\}$ |
|---|---|---|
| $\{x_3, x_4, x_6\}$ | 0.00 | **482.87** |
| $\{x_1, x_2, x_5\}$ | **482.87** | 0.00 |

No quinto estágio, todos os pontos se unem em um único *cluster*. Quando isso ocorre o aumento na SSE será $\Delta SSE_{(125)(346)} \frac{1}{2}(482.87) = 241.44$. Portanto,

$$SSE_5 = SSE_4 + \Delta SSE_{(125)(346)} = 411.96$$

O respectivo dendrograma deste *cluster* é mostrado na figura 5.8

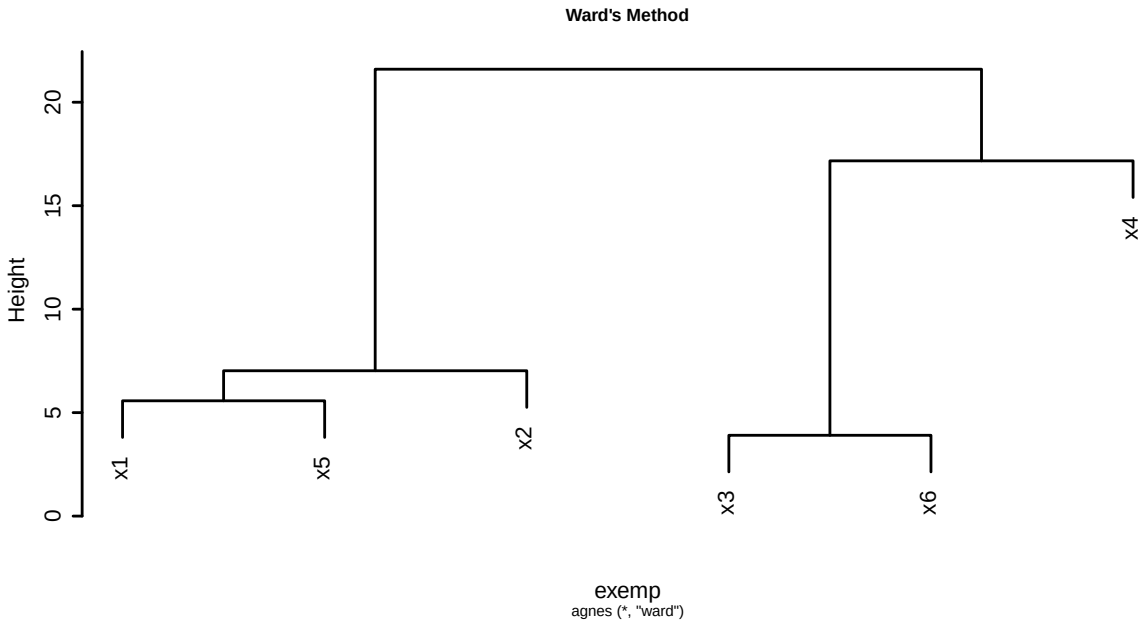



Figura 5.8: Dendrograma produzido ao aplicar o método de ward

## 5.2 Técnicas Hierárquicas Divisivas

O método hierárquico divisivo funciona, justamente, de forma oposta ao aglomerativo. Inicialmente, tem-se um único *cluster* com todas as observações, ou seja, todas as observações constituem um *cluster*. A cada passo do algoritmo, as observações vão se separando de acordo com algum critério de similaridade previamente escolhido. Métodos hierárquicos divisivos dividem-se dois tipos:

- **Monotético:** divide os dados com base em uma específica característica;

- **Politético:** divide os dados com base em valores obtidos em múltiplas características.

### 5.2.1 Método DIANA (DIANA Method)

Um algoritmo para a utilização de técnicas hierárquicas divisivas é denominado DIANA (DIvisive ANAlysis) descrita em (Kaufman and Rousseeuw, 1990). Ele se aplica ao mesmo tipo de dados que permitem a aplicação de técnicas hierárquicas aglomerativas.

O algoritmo DIANA executa sucessivas divisões. A cada passo, o maior *cluster*, ou seja, aquele que apresenta maior índice de dissimilaridade entre duas observações é separado do *cluster* até o momento em que cada observação represente um *cluster* cada.

Seja $C$ um *cluster*, cada passo do algoritmo divide C em dois *clusters* ($A$ e $B$). Considere $A = C$ e $B = \emptyset$. No passo 1, deve-se remover uma observação do *cluster* $A$ para o *cluster* $B$. Para cada observação $v$ de $A$, computa-se a dissimilaridade média desta observação em relação a todas as outras observações de $A$:

$$d(v, A - \{v\}) = \frac{1}{|A| - 1} \sum_{t \in A, t \neq v} d(v,t) \tag{5.8}$$

A observação $v'$ para a qual 5.8 atinge seu valor máximo será retirada de $A$ para $B$.

$$A_{novo} = A_{velho} - \{v'\} \tag{5.9}$$

$$B_{novo} = B_{velho} \cup \{v'\} \tag{5.10}$$

Nos próximos passos, repete-se o passo anterior e, enquanto $A$ possuir uma observação que deve ser movida para $B$, realizar-se-á:

$$d(v, A - \{v\}) - d(v, B) = \frac{1}{|A| - 1} \sum_{t \in A, t \neq v} d(v,t) - \frac{1}{|B|} \sum_{l \in B} d(v,l) \tag{5.11}$$

para cada observação $v$ de $A$ e separaremos aquela observação $v''$ que maximiza 5.11.

- Quando o valor que maximiza 5.11 é positivo, move-se a observação $v''$ de $A$ para $B$ como mostrado em 5.9 e recomeça-se novamente a busca pela observação que deve deixar $A$ em direção a $B$
- Caso contrário, se o valor que maximiza 5.11 for $\leq 0$, para-se o processo e a divisão de $C$ em $A$ e $B$ está finalizada.

A cada passo deve-se decidir qual *cluster* separar. Para isso utiliza-se o diâmetro para cada *cluster*

$$Diam(C) = \max_{\mathbf{x},\mathbf{y} \in C} d(\mathbf{x},\mathbf{y}) \tag{5.12}$$

que é calculado após o passo anterior, e escolhe-se o *cluster* para o qual 5.12 é maior.

Os valores do diâmetro aparecem como *heights* na representação gráfica deste tipo de técnica.

Aplicando o algoritmo DIANA à tabela de dados 5.1, obtem-se a seguinte matriz de dissimilaridade:

Tabela 5.4: Matriz de Dissimilaridade para os dados da tabela 5.1

| | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ |
|---|---|---|---|---|---|---|
| $x_1$ | 0.00 | 7.18 | 16.35 | 15.93 | 5.57 | 18.44 |
| $x_2$ | 7.18 | 0.00 | 9.38 | 12.35 | 6.16 | 11.26 |
| $x_3$ | 16.35 | 9.38 | 0.00 | 15.83 | 13.06 | 3.91 |
| $x_4$ | 15.93 | 12.35 | 15.83 | 0.00 | 18.12 | 14.12 |
| $x_5$ | 5.57 | 6.16 | 13.06 | 18.12 | 0.00 | 16.04 |
| $x_6$ | 18.44 | 11.26 | 3.91 | 14.12 | 16.04 | 0.00 |

**Passo 1:** A primeira coisa a ser feita é procurar qual elemento que destoa dos demais, isto é, procura-se a observação que é mais heterogênea em relação a todas as outras. Para este propósito, utiliza-se a dissimilaridade média. Logo, procura-se pela observação cuja dissimilaridade média é maior em relação às demais observações.

Portanto, opta-se pela retirada da observação $x_4$. Neste estágio temos os *clusters*: $\{x_4\}$ e $\{x_1,x_2,x_3,x_5,x_6\}$. O algoritmo prossegue.

Para cada observação do *cluster* maior calcula-se a dissimilaridade média com as observações restantes e compara-se esse valor com o das observações do *cluster* dissidente.

| Observação | Dissimilaridade Média |
|---|---|
| $x_1$ | (7.18+16.35+15.93+5.57+18.44)/5 = 12.694 |
| $x_2$ | (7.18+ 9.38+12.35+6.16+11.26)/5 = 9.266 |
| $x_3$ | (16.35+9.38+15.83+13.06+3.91)/5 = 11.706 |
| $x_4$ | (15.93+12.35+15.83+18.12+14.12)/5 = **15.270** |
| $x_5$ | (5.57+6.16+13.06+18.12+16.04)/5 = 11.790 |
| $x_6$ | (18.44+11.26+3.91+14.12+16.04)/5 = 12.754 |

| Observação | Dissim. Média | Dissim. Dissidente | Diferença |
|---|---|---|---|
| $x_1$ | (7.18+16.35+5.57+18.44)/4 = 11.885 | 15.930 | -4.045 |
| $x_2$ | (7.18+ 9.38+6.16+11.26)/4 = 8.495 | 12.350 | -3.855 |
| $x_3$ | (16.35+9.38+13.06+3.91)/4 = 10.675 | 15.830 | -5.155 |
| $x_5$ | (5.57+6.16+13.06+16.04)/4 = 10.208 | 18.12 | -7.913 |
| $x_6$ | (18.44+11.26+3.91+16.04)/4 =12.413 | 14.12 | -1.708 |

Neste estágio, todas as diferenças são negativas o que significa dizer que nenhuma observação do *cluster* original assemelha-se às observações dissidentes, de maneira suficiente, que permita à elas mudar de *cluster*. Portanto, nenhuma observação deixa o *cluster*. O processo pára e o primeiro passo do algoritmo divisivo está completo. Aqui temos dois *clusters*: $\{x_1,x_2,x_3,x_5,x_6\}$ e $x_4$.

**Passo 2:** Nesta fase, escolhe-se para o processo de divisão o maior **cluster**, isto é, aquele cujo diâmetro é o maior. O diâmetro de $\{x_1,x_2,x_3,x_5,x_6\} = 18.44$ e $x_4 = 0.00$. Portanto, no passo 2 aplicamos o procedimento ao *cluster* $\{x_1,x_2,x_3,x_5,x_6\}$.

Tabela 5.5: Matriz de Dissimilaridade para $\{x_1,x_2,x_3,x_5,x_6\}$

| | $x_1$ | $x_2$ | $x_3$ | $x_5$ | $x_6$ |
|---|---|---|---|---|---|
| | $x_1$ | $x_2$ | $x_3$ | $x_5$ | $x_6$ |
| $x_1$ | 0.00 | 7.18 | 16.35 | 5.57 | 18.44 |
| $x_2$ | 7.18 | 0.00 | 9.38 | 6.16 | 11.26 |
| $x_3$ | 16.35 | 9.38 | 0.00 | 13.06 | 3.91 |
| $x_5$ | 5.57 | 6.16 | 13.06 | 0.00 | 16.04 |
| $x_6$ | 18.44 | 11.26 | 3.91 | 16.04 | 0.00 |

Opta-se pela retirada da observação $x_6$. Neste estágio temos os *clusters*: $\{x_1,x_2,x_3,x_5\}$, $x_4$ e $x_6$. O algoritmo prossegue.

Opta-se pela retirada da observação $x_3$ porque ela apresentou um grau de similaridade mais próximo ao da variável dissidente do que com as outras variáveis do *cluster*. Neste estágio temos os *clusters*: $\{x_1,x_2,x_5\}$, $x_4$ e $x_3,x_6$. O algoritmo prossegue.

| Observação | Dissim. Média | Dissim. Média Dissidente | Diferença |
|---|---|---|---|
| $x_1$ | (7.18+16.35+5.57+18.44)/4 = 11.885 | | |
| $x_2$ | (7.18+9.38+6.16+11.26)/4 = 8.495 | | |
| $x_3$ | (16.35+9.38+13.06+3.91)/4 = 10.675 | | |
| $x_5$ | (5.57+6.16+13.06+16.04)/4 = 10.208 | | |
| $x_6$ | (18.44+11.26+3.91+16.04)/4 = **12.413** | | |

| Observação | Dissim. Média | Dissim. Média Dissidente | Diferença |
|---|---|---|---|
| $x_1$ | (7.18+16.35+5.57)/3 = 9.700 | 18.44 | -8.740 |
| $x_2$ | (7.18+9.38+6.16)/3 = 7.573 | 11.26 | -3.687 |
| $x_3$ | (16.35+9.38+13.06)/3 = 12.930 | 3.91 | **9.020** |
| $x_5$ | (5.57+6.16+13.06)/3 = 8.263 | 16.04 | -7.777 |

| Observação | Dissim. Média | Dissim. Média Dissidente | Diferença |
|---|---|---|---|
| $x_1$ | (7.18+16.35)/2 = 6.375 | 16.35 | -9.975 |
| $x_2$ | (7.18+9.38)/2 = 6.670 | 9.38 | -2.710 |
| $x_5$ | (5.57+6.16)/2 = 5.865 | 13.06 | -7.195 |

Neste estágio, todas as diferenças são negativas o que significa dizer que nenhuma observação do *cluster* original assemelha-se às observações dissidentes, de maneira suficiente, que permita à elas mudar de *cluster*. Portanto, nenhuma observação deixa o *cluster*. O processo pára e o segundo passo do algoritmo divisivo está completo. Aqui temos três *clusters*: $\{x_1,x_2,x_5\}$, $\{x_3x_6\}$ e $x_4$.

**Passo 3:** Nesta fase, escolhe-se para o processo de divisão o maior **cluster**, isto é, aquele cujo diâmetro é o maior. O diâmetro de $\{x_1,x_2,x_5\} = 7.18$, $\{x_3,x_6\} = 3.91$ e $x_4 = 0$. Portanto, no passo 3 aplicamos o procedimento ao *cluster* $\{x_1,x_2,x_5\}$.

Tabela 5.6: Matriz de Dissimilaridade para $\{x_1,x_2,x_5\}$

| | $x_1$ | $x_2$ | $x_5$ |
|---|---|---|---|
| $x_1$ | 0.00 | 7.18 | 5.57 |
| $x_2$ | 7.18 | 0.00 | 6.16 |
| $x_5$ | 5.57 | 6.16 | 0.00 |

| Observação | Dissimilaridade Média |
|---|---|
| $x_1$ | (7.18+5.57)/2 = 6.375 |
| $x_2$ | (7.18+6.16)/2 = **6.670** |
| $x_5$ | (5.57+6.16)/2 = 5.865 |

Opta-se pela retirada da observação $x_2$. Neste estágio temos os *clusters*: $\{x_1,x_5\}$, $x_4$, $x_3$,$x_6$ e $x_2$. O algoritmo prossegue.

| Observação | Dissim. Média | Dissim. Média Dissidente | Diferença |
|---|---|---|---|
| $x_1$ | 5.57 | 7.18 | -1.61 |
| $x_5$ | 5.57 | 6.16 | -0.59 |

Aqui, o processo pára porque todas as diferenças são negativas. Portanto, o **passo 3** divide $\{x_1,x_2,x_5\}$ em $\{x_1,x_5\}$ e $x_2$.

**Passo 4:** Novamente, deve-se escolher qual *cluster* deve ser dividido baseando-se, para isso, no maior diâmetro. O diâmetro de $\{x_1,x_5\} = 5.57$, $\{x_3,x_6\} = 3.91$, $x_4 = 0$ e $x_2 = 0$. Portanto, no passo 4 aplicamos o procedimento ao *cluster* $\{x_1,x_5\}$. Logo, optamos pela divisão de $\{x_1,x_5\}$.

Tabela 5.7: Matriz de Dissimilaridade para $\{x_1,x_5\}$

| | $x_1$ | $x_5$ |
|---|---|---|
| $x_1$ | 0.00 | 5.57 |
| $x_5$ | 5.57 | 0.00 |

Como as dissimilaridades são iguais, devemos escolher qual observação iniciará o processo de divisão. Escolhe-se $x_1$, então obtêm-se os *clusters* $x_1$ e $x_5$. Como $x_1$ é o último remanescente, ele não se une à $x_5$. Portanto, o passo 4 divide o *cluster* $\{x_1,x_5\}$ em $x_1$ e $x_5$. Após o passo 4 temos:$x_1$, $x_5$, $\{x_3,x_6\}$, $x_4$ e $x_2$.

**Passo 5:** Agora, resta dividir apenas o *cluster* $\{x_3,x_6\}$ porque todas as outras observações já estão isoladas em um *cluster*. Aqui, $\{x_3,x_6\}$ é dividido em $x_3$ e $x_6$. Após o quinto passo, só restam *clusters* unitários: $x_1$, $x_2$, $x_3$, $x_4$, $x_5$ e $x_6$. É o fim do processo de divisão.

O respectivo dendrograma deste procedimento é mostrado na figura 5.9

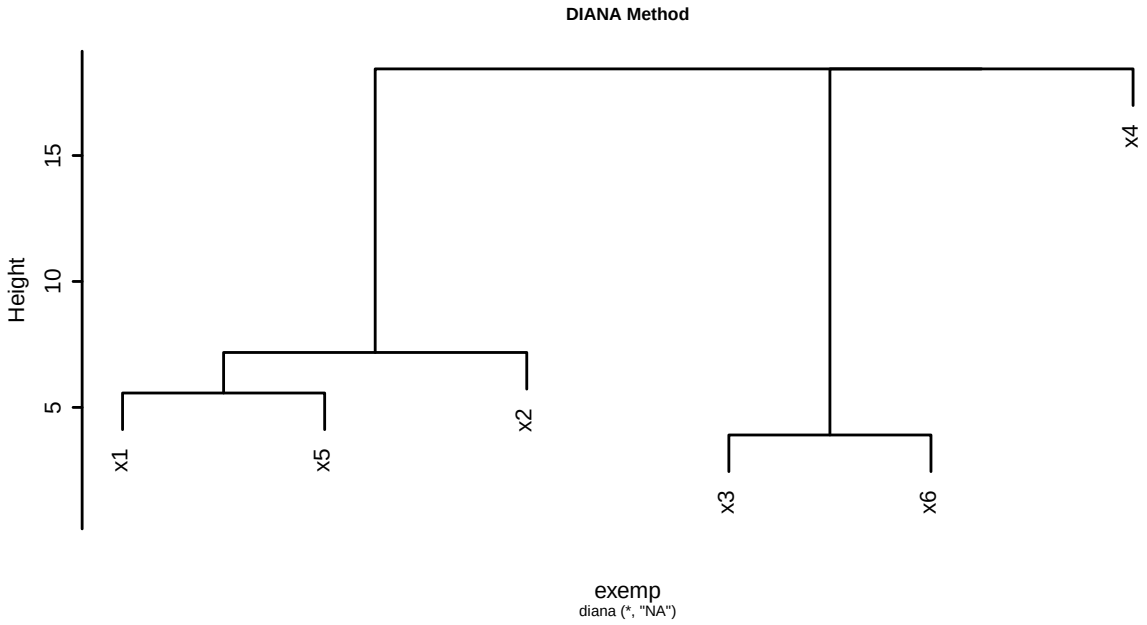



Figura 5.9: Dendrograma produzido ao aplicar o método divisivo DIANA

# 6 Técnicas Não-Hierárquicas de Agrupamento

Técnicas de agrupamento não-hierárquicas têm como objetivo encontrar, diretamente, uma partição de $n$ elementos em $C$ grupos (*clusters*). Diferentemente dos métodos hierárquicos, seus algoritmos baseiam-se na busca de $k$ observações que devem representar vários aspectos da estrutura dos dados, dentre todas as observações do banco de dados.

Geralmente, esses algoritmos possuem suas próprias funções objetivo que definem o quão boa é a partição. *Clusters* formados assim, possuem forma convexa e são representados por um centróide. São muito eficientes no agrupamento de bancos de dados com grandes volumes de informação e bancos de dados multidimensionais.

## 6.1 Método *K-Means* (*The k-Means Algorithm*)

É provavelmente um dos mais conhecidos e mais utilizados em problemas práticos. Foi desenvolvido para agrupar dados numéricos de modo que cada *cluster* tenha um centróide (vetor de médias amostrais). O número $k$ de partições é fixo. O algoritmo do método *k-means* funciona da seguinte forma:

- Escolhem-se $k$ *clusters* iniciais (centróides), chamados de "sementes", para a análise do processo de partição;
- Cada observação restante do conjunto de dados é, então, alocada ao *cluster* mais homogêneo. Este processo é iterativo e prossegue sempre modificando a composição dos *clusters* de acordo com a função erro até o momento em que ela não se altere significativamente ou que nenhuma realocação dos *clusters* seja necessária.

O algoritmo convencional do método *k-means* é: (Gan et al., 2007)

Seja $D$ um conjunto de dados com $n$ observações, e seja $C_1, C_2, ..., C_k$ os $k$ *clusters* disjuntos de $D$. A função erro é definida como

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} d(\mathbf{x}, \mu(C_i)) \tag{6.1}$$

sendo $\mu(C_i)$ o centróide do cluster $C_i$, $d(\mathbf{x}, \mu(C_i))$ denota a distância entre $\mathbf{x}$ e $\mu(C_i)$ que pode ser a distância euclidiana, manhattan, mahalanobis entre outras.

Segundo (Gan et al., 2007) o método *k-means* pode ser dividido em duas fases:

- **Fase de Inicialização:** o algoritmo atribui aleatoriamente as observações em $k$ *clusters*;
- **Fase de Iteração:** o algoritmo calcula a distância entre cada observação e cada *cluster* e atribui as observações ao *cluster* mais próximo.

Ainda segundo ele, o método *k-means* pode ser tratado como um problema de otimização. Neste sentido, o objetivo do algoritmo é minimizar uma função objetivo dada sob certas condições. Para um conjunto de dados $D = \{x_i, i = 1, 2, ..., n\}$ com $n$ observações e $k$ um número inteiro dado. A função objetivo pode ser definida como

$$P(W, Q) = \sum_{l=1}^{k} \sum_{i=1}^{n} w_{il} d(x_i, q_l) \tag{6.2}$$

sendo $Q = \{q_l, l = 1, ..., k\}$ um conjunto de observações, $d(\cdot, \cdot)$ a distância euclidiana, e $W$ um matriz $n \times k$ que satisfaz as seguintes condições:

1. $w_{i1} \in \{0, 1\}$ para $i = 1, 2, ..., k$;
2. $\sum_{l=1}^{k} w_{il} = 1$ para $i = 1, 2, ..., k$.

Ainda segundo (Gan et al., 2007), o método *k-means* possui as seguintes propriedades :

- É eficiente no agrupamento de grandes conjuntos de dados, já que sua complexidade computacional é linearmente proporcional ao tamanho dos conjuntos de dados;
- Muitas vezes termina num ponto ótimo local;
- Os *clusters* tem forma convexa, tal qual uma esfera;

- Funciona com dados numéricos;
- Sua performance depende da escolha inicial dos centróides.

(Mingoti, 2005) alerta que cuidados são necessários na escolha das "sementes" já que isto influencia no agrupamento final. Alguns métodos de seleção são propostos:

- Uso de técnicas hierárquicas de agrupamento;
- Escolha aleatória;
- Escolha prefixada.

Para os dados da tabela 5.1. Baseado nas análises anteriores acerca de agrupamentos utilizando métodos hierárquicos, executamos o método *k-means* para a formação de dois grupos, com seleção aleatória dos centros. Temos o seguinte resultado gráfico:

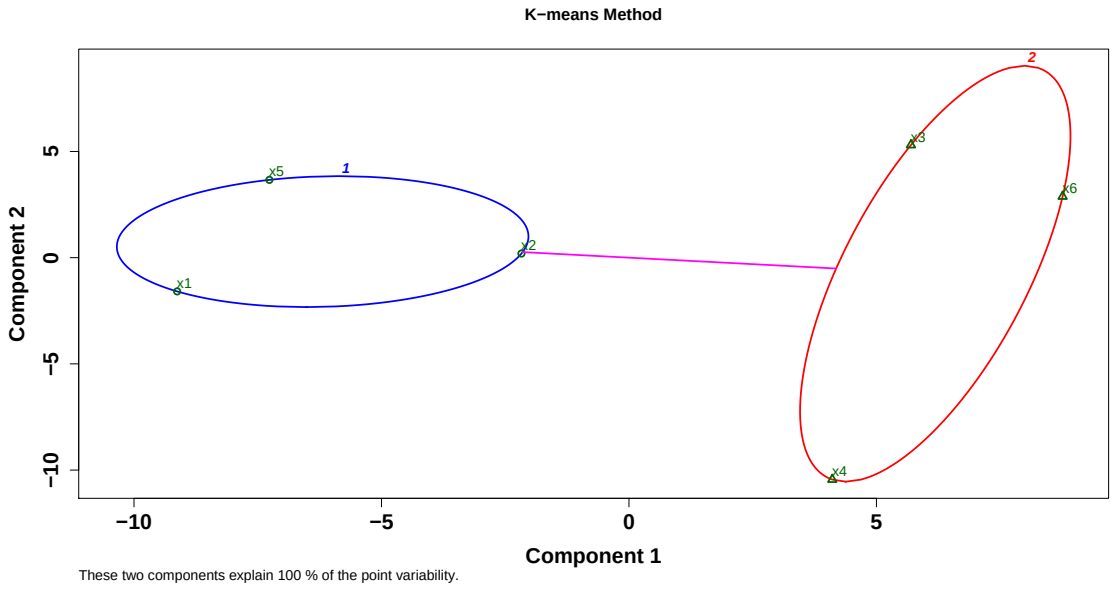



Figura 6.1: Gráfico método *k-means* para os dados da tabela 5.1

Verificamos que o método *k-means* resulta na mesma partição dos métodos anteriores $\{x_3, x_4, x_6\}$ e $\{x_1, x_2, x_5\}$.

## 6.2 Método *K-Medoids* (*The k-Medoids Algorithm*)

Quando constroem-se partições com um número fixo $k$ de *clusters*, geralmente presume-se que exitem funções que medem a qualidade de diferentes agrupamentos no

mesmo conjunto de dados. Esta idéia é a idéia por trás do método *k-medoids*, que é baseada numa medida de posição central a *k*-mediana. Este método segue a seguinte idéia:

> Dado um número finito $n$ de observações, $k$ observações devem ser inicialmente escolhidas como observações representativas do *cluster* (medoid/mediana). Essas observações são selecionadas de tal forma que a distância (dissimilaridade) entre elas e o restante das observações do *cluster* ao qual elas pertencem seja a menor possível, ou seja, deseja-se que observações alocadas nos mesmo *cluster* sejam o mais homogêneas possível.

Matematicamente, o método *k-medoids* pode ser escrito da seguinte maneira: (Kaufman and Rousseeuw, 1990)

Seja $X = \{x_1, x_2, ..., x_n\}$ o conjunto de observações. A dissimilaridade entre observações $x_i$ e $x_j$ é denotada por $d(i, j)$. A resolução do modelo é determinada por dois tipos de decisões:

1. **A seleção de observações como observações representativas de um *cluster*:** $y_i$ é definida como uma observação $\{0,1\}$ da seguinte maneira

$$y_i = \begin{cases} 1, & \text{se, e somente se, a observação} i (i = 1, ..., n) \text{é representativa.} \\ 0, & \text{caso contrário.} \end{cases}$$

2. **A atribuição de cada observação $j$ a uma das observações representativas:** $z_{ij}$ é definida como uma observação $\{0,1\}$ da seguinte maneira

$$z_{ij} = \begin{cases} 1, & \text{se, e somente se, a observação } j \text{ é atribuída a um } \textit{cluster} \text{ em que i} \\ & \text{é a observação representativa (medoid).} \\ 0, & \text{caso contrário.} \end{cases}$$

O algoritmo de otimização proposto por Vinod **??** pode ser escrito como:

$$\text{minimizar} \sum_{i=1}^{n} \sum_{j=1}^{n} d(i, j) z_{ij} \tag{6.3}$$

sob as seguintes restrições:

$$\sum_{i=1}^{n} z_{ij} = 1, \quad j = 1,2,...,n \tag{6.4a}$$

$$z_{ij} \leq y_i, \quad i,j = 1,2,...,n \tag{6.4b}$$

$$\sum_{i=1}^{n} y_i = k, \quad k = \text{número de } \textit{clusters} \tag{6.4c}$$

$$y_i, z_{ij} \in \{0,1\}, \quad i,j = 1,2,...,n \tag{6.4d}$$

- A restrição 6.4a significa que cada observação $j$ deve ser alocada a uma única observação representativa.
- As restrições 6.4a e 6.4d significam, conjuntamente, que para uma dada observação $j$, uma das $z_{ij}$ observações é igual a 1 e todas as outras são iguais a 0.
- A restrição 6.4b implica que uma observação $j$ poderá ser alocada a uma observação $i$ se, e somente se, a observação $i$ foi escolhida como representativa. Caso contrário, então $y_i = 0$ e as restrições 6.4b e 6.4d implicam que todas as observações $z_{ij} = 0$.
- A restrição 6.4c significa que exatamente $k$ observações serão escolhidas como observações representativas.
- A restrição 6.4a implica que a dissimilaridade entre uma observação $j$ e sua observação representativa $i$ é dada por $\sum_{i=1}^{n} d(i,j)z_{ij}$. Quando todas as observações são alocadas, a dissimilaridade total é dada por $\sum_{i=1}^{n}\sum_{j=1}^{n} d(i,j)z_{ij}$ que é a função a ser minimizada no modelo.

Para os dados da tabela 5.1. Baseado nas análises anteriores acerca de agrupamentos utilizando métodos hierárquicos, executamos o método *k-medoids* para a formação de três grupos, com seleção aleatória dos centros. Temos o seguinte resultado gráfico:

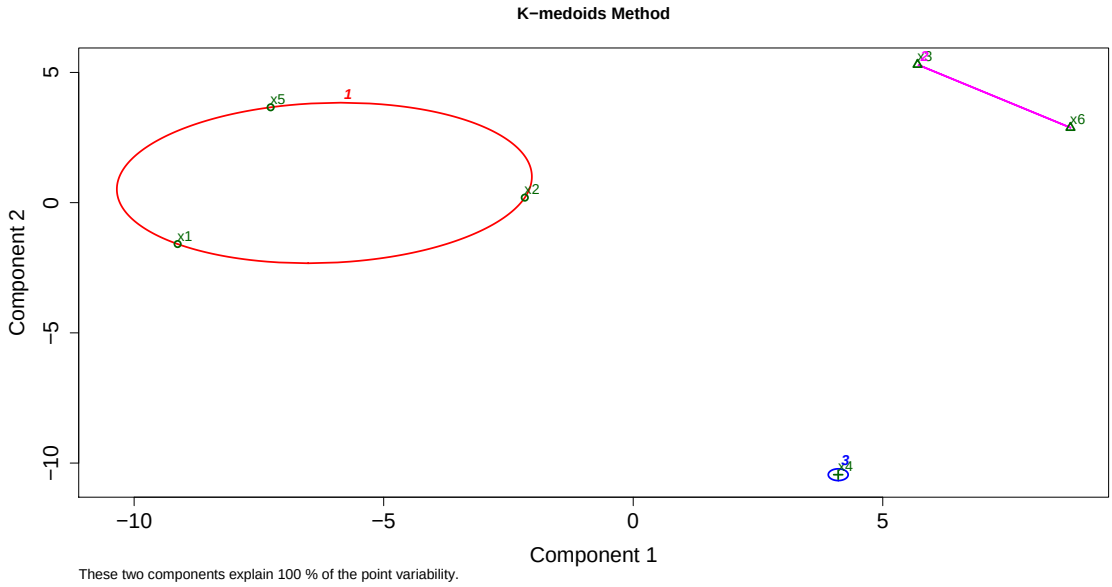



Figura 6.2: Gráfico método *k-medoids* para os dados da tabela 5.1

### 6.2.1 Silhouette

Uma informação de análise disponível no método *k-medoids* e introduzido por Rousseeuw (1987) é o *silhouette*. Cada *cluster* é representado por um *silhouette*, que mostra quais observações estão bem alocadas dentro do *cluster* e quais observações estão numa posição intermediária, ou seja, podem ou não pertencer ao *cluster*.

O agrupamento inteiro é exibido por um diagrama que contém todos os *silhouettes*. Essa medida é muito útil quando as dissimilaridades estão numa escala de razão (é o caso da distância euclidiana) e quando procura-se por *clusters* compactos e bem separados.

*Silhouettes* são construídos da seguinte maneira: (Kaufman and Rousseeuw, 1990)

Seja $C_1$ o *cluster* para o qual a observação $i$ foi designada. Calcule

$$a(i) = \text{dissimilaridade média de } i \text{ em relação a todas as outras observações de } C_1$$

Isso poderá ser feito se, e somente se, $C_1$ possuir outra observação além de $i$. Assim, assume-se que $C_1$ não seja unitário.

Agora seja $G$ um *cluster* diferente de $C_1$, então

$$d(i,C_2) = \text{dissimilaridade média de } i \text{ em relação a todas as observações de } C_2$$

Após calcular $d(i,G)$ para todos os *clusters* $G \neq C_1$, seleciona-se aquele que satisfizer

$$b(i) = \min_{G \neq C_1} d(i,G)$$

O *cluster* $C_2$ para onde este valor é alocado, isto é, $d(i,C_2) = b(i)$ é denominado *neighbor* (vizinho) da observação $i$. Ou seja, se $C_1$ é descartado, então $C_2$ é o *cluster* "próximo" a $i$. É importante ressaltar que *silhouettes* não são definidos para $k = 1$ já que o valor de $b(i)$ depende de pelo menos dois *clusters* $C_1$ e $C_2$ com $C_1 \neq C_2$.

O número $s(i)$ é obtido da combinação de $a(i)$ e $b(i)$ da seguinte forma:

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} = \begin{cases} 1 - \dfrac{a(i)}{b(i)}, & \text{se} \quad a(i) < b(i) \\ 0, & \text{se} \quad a(i) = b(i) \\ \dfrac{b(i)}{a(i)} - 1, & \text{se} \quad a(i) > b(i) \end{cases} \tag{6.5}$$

$$-1 < s(i) < 1$$

Quando $s(i)$ possui valor próximo a 1, isso implica que a dissimilaridade **dentro** ($a(i)$) do *cluster* é menor do que o menor valor de dissimilaridade entre $b(i)$ *clusters*. Isto implica dizer que

> $i$ está bem classificado, isto é, parece haver pouca dúvida de que a observação $i$ foi alocada ao *cluster* apropriado. Em outras palavras, a segunda escolha ($C_2$) não é uma escolha melhor do que a atual ($C_1$).

Situação diferente ocorre quando $s(i) = 0$. Então $a(i)$ e $b(i)$ são aproximadamente iguais e portanto, não está claro para qual *cluster* a observação $i$ deve ser designada. $i$ se encaixaria bem em qualquer *cluster* ($C_1$ ou $C_2$) de modo que este caso é considerado como intermediário.

A pior situação ocorre quando $s(i)$ possui valor próximo a $-1$. Neste caso, $a(i)$ possui valor muito superior ao de $b(i)$. Portanto, $i$, em média, está mais próximo de $C_2$ do que de $C_1$. Assim, podemos concluir que $i$ está mal alocado.

Para os dados da tabela 5.1. O gráfico de *silhouette* é:

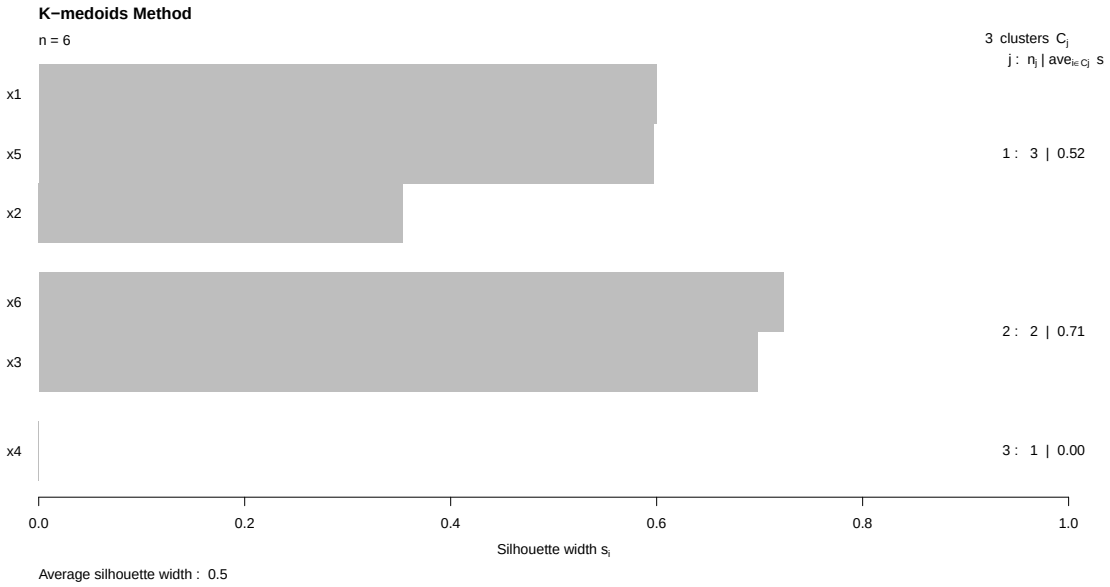



Figura 6.3: Gráfico *Silhouette* método *k-medoids* para os dados da tabela 5.1

A primeira impressão é a de que as *silhouettes* são muito largas, o que indica que a estrutura de agrupamento não é melhor do que razoável. No, 1º *cluster*, $x_1$ possui o maior valor $s(i)_{x_1} = 0.60$, o que indica que $x_1$ está bem alocado, ou seja, foi classificado com o menor grau de incerteza. O 2º *cluster* contêm observações com valores de *silhouette* mais próximos $s(i)_{x_3} = 0.70$ e $s(i)_{x_6} = 0.73$. A observação $x_4$ possui valor de *silhouette* nulo e, portanto, não está claro para qual *cluster* esta observação deve ser designada.

Note que os *silhouettes* dependem apenas da partição atual das observações e não do algoritmo de agrupamento que foi utilizado. Como consequência, *silhouettes* podem ser utilizadas para melhorar os resultados do processo de agrupamento ou para comparar os resultados de diferentes algoritmos aplicados ao mesmo conjunto de dados. *silhouettes* também é muito empregado na escolha de valores adequados para $k$.

# 7 *Self-Organizing Map* (SOM)

O *self-organizing map* (SOM) é uma técnica de visualização gráfica descrita como uma rede neural artificial e idealizada pelo professor Teuvo Kohonen da Universidade de Helsinki na Finlândia e, por isso também é conhecida como mapa de Kohonen. Segundo (Hastie et al., 2009), o SOM pode ser visto como uma versão espacialmente limitada do método *k-means*, em que as observações são dispostas segundo suas características num espaço uni/bidimensional. Nesta analogia, cada unidade corresponde a um *cluster* e o número de *clusters* é definido pelo tamanho do *grid*, que normalmente é organizado numa forma retangular ou hexagonal.

Existem muitas abordagens para o mapeamento de conjuntos de dados multidimensionais em um espaço bidimensional. Uma das mais utilizadas é o PCA (*principal component analysis*). Entretanto, em muitos casos, mais de duas dimensões são necessárias para fornecer uma descrição razoavelmente informativa de modo que a visualização continua sendo o problema principal. Mais, PCA em sua forma pura não incorpora informações sobre como os objetos devem ser comparados, a distância euclidiana padrão nem sempre é a melhor medida de dissimilaridade. Métodos a partir de matrizes de distância ou similaridade, podem ser mais úteis porque ao escolher uma função de distância apropriada para os dados, é possível se ater aos aspectos dos dados que são mais informativos.

Uma abordagem para a visualização de uma matriz de distância em duas dimensões é o MDS (*multidimensional scaling*). Esta técnica tem como objetivo encontrar uma configuração no espaço bidimensional cuja matriz de distância, de alguma forma se aproxima da matriz de distância original, calculada a partir dos dados multidimensionais.

O SOM (Kohonen, 2001) resolve o problema de uma forma semelhante ao MDS, mas ao invés de tentar reproduzir distâncias, seu objetivo é reproduzir topologias, ou em outras palavras, ele tenta manter os mesmos *neighbours*. Portanto, se dois objetos

multidimensionais são muito semelhantes, então suas posições em um plano bidimensional devem ser muito semelhantes também. Ao invés de mapear objetos em um espaço contínuo, o SOM usa um *grid* regular de 'unidades' no qual os objetos são mapeados. As diferenças com o MDS podem ser vistas como pontos fortes e fracos (Wehrens and Buydens, 2007): onde, em um gráfico bidimensional do MDS a distância pode ser diretamente interpretada como uma 'estimativa' da verdadeira distância, em gráfico do SOM isso não ocorre: só se pode dizer que os objetos mapeados nas mesmas unidades são muito semelhantes. Em outras palavras, o SOM concentra-se nas maiores similaridades, enquanto o MDS concentra-se nas maiores dissimilaridades. A aplicação de cada um dependerá do problema e da experiência do pesquisador acerca do assunto.

O SOM é muito útil para visualização de dados multivariados, análises de agrupamento, busca de padrões. Seu uso associado à técnica de análise de sinais (*wavelets*) está descrito no artigo (Moshou et al., 2005) como técnicas avançadas na detecção de fadiga muscular sob condições dinâmicas.

> "Para a detecção de fadiga muscular, sob condições dinâmicas, técnicas mais avançadas baseadas em *wavelets* e redes neurais são propostas no documento atual. Técnicas convencionais de análise de frequência e amplitude não funcionam neste caso. A amplitude do sinal para todos os indivíduos está diminuindo perto do final do teste mostrando que os indivíduos estão fazendo menos força. Isto indica que enquanto a fadiga é claramente presente, de acordo com a experiência dos próprios indivíduos, ela não pode ser detectada devido à condição de força constante que não está sendo aplicada." (Moshou et al., 2005)

Ainda segundo este artigo, devido à capacidade de preservar a topologia, os "neurônios" ativados no caso da presença de fadiga tendem a estar em uma região claramente definida do mapa o que possibilita uma melhor análise e interpretação dos resultados.

## 7.1 Teoria

O SOM consiste em duas camadas totalmente conectadas: uma camada de entrada e uma camada de Kohonen 7.1. Os "neurônios" na camada Kohonen são dispostos numa estrutura uni/bidimensional.

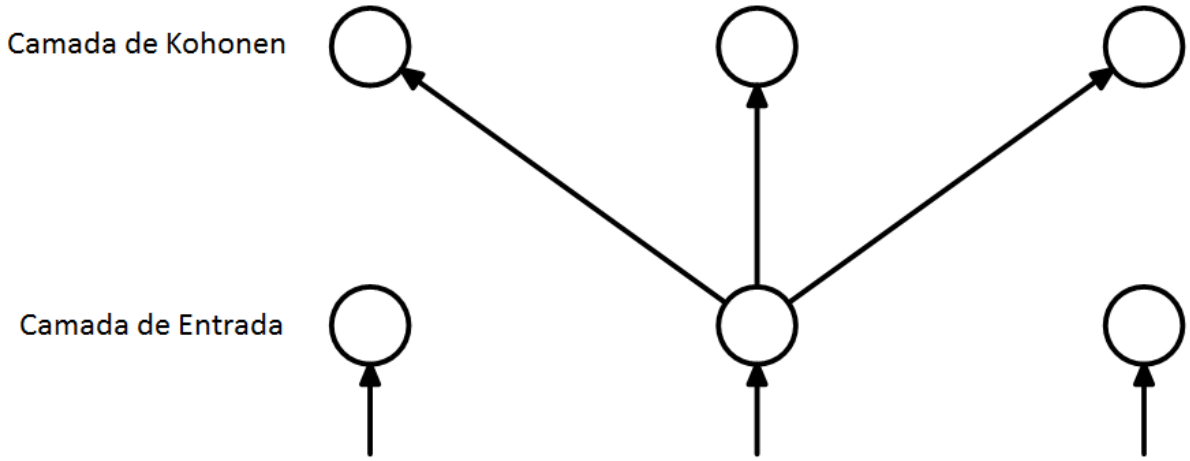



Figura 7.1: *Layout* de um mapa de Kohonen unidimensional (Gan et al., 2007).

O número de "neurônios" na camada de entrada correspondem ao número de características das variáveis. Cada "neurônio" na camada de entrada, relaciona-se com um "neurônio" na camada de Kohonen. Pressupõe-se que as variáveis de entrada estejam normalizadas, ou seja, $\|x\| = 1$. As entradas na camada de Kohonen podem ser calculadas como

$$y_j = \sum_{i=1}^{p} w_{ji} x_i \tag{7.1}$$

sendo $w_{ji}$ o peso do "neurônio" de entrada $i$ para o "neurônio" de saída $j$ e $p$ a dimensão da camada de entrada.

O algoritmo computacional do SOM funciona da seguinte maneira:

- Inicializam-se os pesos da rede atribuindo-lhes pequenos valores aleatórios;
- O algoritmo prossegue na realização de três processos essenciais: competição, cooperação e adaptação.
  - **Processo de Competição:** no processo de competição, é escolhida a melhor combinação de variáveis de entrada e os respectivos pesos que satisfaçam 7.1. Isso se dá da seguinte maneira:

Seja $\mathbf{x} = (x_1, x_2, ..., x_p)^T$ o vetor de variáveis escolhido aleatoriamente na camada de entrada, sendo $p$ sua dimensão. Seja $\mathbf{w}_j = (w_{j1}, w_{j2}, ...w_{jd})^T$, $j = 1, 2, ..., d$ o vetor de pesos do "neurônio" $j$ na camada de Kohonen, sendo $d$ o número total de "neurônios" na camada de Kohonen. A melhor combinação do vetor de variáveis

$\mathbf{x}$ e o vetor de pesos $\mathbf{w}_j$ pode ser dada através da comparação do produto interno $< \mathbf{x}, \mathbf{w}_j^T >=< \mathbf{x}\mathbf{w}_1^T, \mathbf{x}\mathbf{w}_2^T, ..., \mathbf{x}\mathbf{w}_d^T >$ e seleção do maior valor. Matematicamente, isso é equivalente a minimizar a distância euclidiana entre os vetores $\mathbf{w}_j$ e $\mathbf{x}$. Desta forma, o índice $i(\mathbf{x})$ dado ao "neurônio" escolhido para a variável de entrada $\mathbf{x}$ é dada por

$$i(\mathbf{x}) = \underset{1 \leq j \leq d}{argmin} \|\mathbf{x} - \mathbf{w}_j\| \tag{7.2}$$

- **Processo de Cooperação:** no processo cooperativo, um *neighborhood* topológico ($Q_j$) é definido de modo que o "neurônio" escolhido no processo anterior localize-se no centro de um *neighborhood* topológico de "neurônios" cooperativos.

Seja $h_{j,t}$ o *neighborhood* topológico centrado no "neurônio" escolhido $t$ e $d_{t,j}$, distância lateral entre $t$ e um "neurônio" $j$. O *neighborhood* topológico $h_{j,t}$, pode ser uma função unimodal da distância lateral $d_{t,j}$, satisfazendo as seguintes condições (Gan et al., 2007):

**(i)** $h_{t,j}$ é simétrico em relação ao ponto máximo definido por $d_{t,j} = 0$;

**(ii)** A amplitude de $h_{t,j}$ diminui monotonicamente com o aumento da distância lateral $d_{t,j}$ e tende a zero quando $d_{t,j} \to \infty$.

Uma das escolhas para $h_{j,t}$ pode ser a função Gaussiana

$$h_{j,t} = exp\left(-\frac{d_{t,j}^2}{2\sigma^2}\right) \tag{7.3}$$

sendo $\sigma$ o parâmetro que mede o nível em que os "neurônios" excitados no *neighborhood* do "neurônio" escolhido participam do processo. No caso de um gráfico unidimensional, a distância lateral $d_{t,j}$ pode ser definida como

$$d_{t,j} = |t - j| \tag{7.4}$$

Já no caso bidimensional, a distância lateral $d_{t,j}$ é

$$d_{t,j} = \|\mathbf{r}_t - \mathbf{r}_j\| \tag{7.5}$$

sendo $\mathbf{r}_t$ e $\mathbf{r}_j$ vetores discretos que definem a posição do "neurônio" $j$ excitado e a posição do "neurônio" escolhido $t$, respectivamente.

- **Processo Adaptativo:** no processo adaptativo, o vetor de pesos do "neurônio" $j$ ($\mathbf{w}_j$) muda de acordo com a variável de entrada $\mathbf{x}$. Dado o vetor de pesos do "neurônio" $j$ na iteração $s$ ($\mathbf{w}_j^{(s)}$), o novo vetor de pesos no tempo $s+1$ é definido como:

$$\mathbf{w}_j^{(s+1)} = \mathbf{w}_j^{(s)} + \eta(s) h_{j,i(\mathbf{x})}(s)(\mathbf{x} - \mathbf{w}_j^{(s)}) \tag{7.6}$$

sendo $\eta(s)$ o parâmetro *learning-rate* definido como

$$\eta(s) = \eta_0 exp\left(-\frac{s}{\tau_2}\right), \quad s = 0,1,2,... \tag{7.7}$$

e $h_{j,i(\mathbf{x})}$ é a função ***neighborhood*** definida como

$$h_{j,i(\mathbf{s})} = exp\left(-\frac{d^2_{i(\mathbf{x}),j}}{2\sigma^2(s)}\right) \tag{7.8}$$

$$\sigma(s) = \sigma_0 \left(-\frac{s}{\tau_1}\right), \quad s = 0,1,2,... \tag{7.9}$$

(Gan et al., 2007) sugere que as constantes $\eta_0$, $\sigma_0$, $\tau_1$ e $\tau_2$ podem ser configuradas assim

$$\eta_0 = 0.1$$
$$\sigma_0 = \text{o raio do gráfico}$$
$$\tau_1 = \frac{1000}{log(\sigma_0)}$$
$$\tau_2 = 1000$$

O SOM se encaixa num tipo de estrutura conhecida como *unsupervised learning*. Neste tipo de estrutura, refere-se a problemas em que não se sabe o padrão de comportamento das variáveis, isto é, não há resultados esperados. Uma das vantagens do SOM é que ele se adapta às características dos dados.

No livro (Kohonen, 2001, p. 164) temos um exemplo de um SOM para os dados apresentados na tabela 7.1

| Animal | Pequeno | Médio | Grande | Patas_2 | Patas_4 | Pêlo | Casco | Juba | Penas | Caçar | Correr | Voar | Nadar |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Pomba | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |
| Galinha | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| Pato | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 |
| Ganso | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 |
| Coruja | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 |
| Falcão | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 |
| Águia | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 |
| Raposa | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| Cão | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| Lobo | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| Gato | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| Tigre | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| Leão | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 |
| Cavalo | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 |
| Zebra | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 |
| Vaca | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |

Tabela 7.1: Self Organizing Maps - pag.164

Cada coluna é uma descrição esquemática de um animal baseado na presença ($=1$) ou ausência ($=0$) de alguma das 13 características fornecidas. Algumas características como “penas” e “e patas” são correlacionadas, indicando diferenças mais significativas do que as outras.

As variáveis do banco de dados foram alocados de forma iterativa e aleatoriamente em um SOM de $4\times4$ “neurônios” sujeitos ao processo de adaptação descrito. O resultado a ser alcançado fornecido pelo livro é:

Começa-se por atribuir um *codebook vector* para cada unidade, que irá desempenhar o papel de um padrão típico, um protótipo, associado a essa unidade. Geralmente, atribui-se aleatoriamente um subconjunto dos dados para as unidades. Durante o processo de formação, os objetos são repetidamente alocados no mapa de forma aleatória. A ‘unidade vencedora’, ou seja, a mais semelhante à que foi alocada, será atualizada para se tornar ainda mais similar. Aqui, uma média ponderada é usada, onde o peso do novo objeto é um dos parâmetros de alocação do SOM. Também referida como a *learning rate* $\alpha$, que geralmente é um valor pequeno na ordem de $0,05$. Durante a alocação, esse valor diminui para que o mapa convergir. A restrição espacial mencionada anteriormente reside no fato de que o SOM requer que unidades vizinhas tenham *codebook vectors* similares. Isto é alcançado não só através da atualização da unidade vencedora, mas também da atualização das unidades imediatamente próximas (vizinhas). O tamanho da vizinhança diminui durante o processo de formação, de modo que, apenas as unidades vencedoras

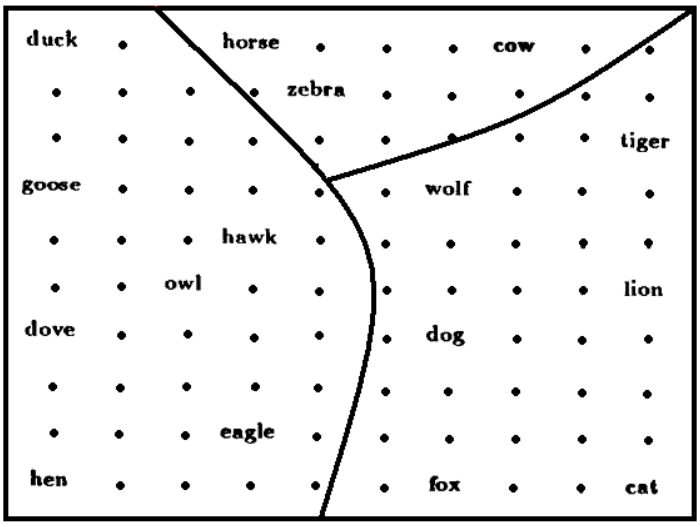



Figura 7.2: (Kohonen, 2001, p. 164)

sejam alocadas. Nessa fase, o procedimento é exatamente igual ao *k-means*. O algoritmo termina após um número pré-definido de iterações. Mais informações podem ser encontradas no livro de (Kohonen, 2001; Wehrens and Buydens, 2007).

O objetivo aqui foi mapear os 16 animais em um SOM de $4 \times 4$ unidades hexagonalmente orientadas utilizando o pacote **kohonen** disponibilizado no software R. O *codebook vectors* são plotados em um *segment plot*, que é o padrão para este tipo de gráfico.

A amostra dos animais projetada na parte inferior à direita do mapa, estão associados à habilidade de nadar, voar, possuem 2 patas e penas, enquanto que animais com 4 patas, pêlos, casco e grandes.

Outro gráfico que pode ser obtido na análise e que pode ajudar na busca de grupos é o *mapping*. Nele é possível visualizar quem são e aonde estão plotados os objetos do banco de dados. Além disso, é possível rodar técnicas de agrupamento (hierárquicas e não-hierárquicas) conjuntamente de modo que consegue-se visualizar os grupos formados e quem são os objetos que formam os grupos (os *clusters* são separados por linhas). Para os dados da tabela 7.1 temos o seguinte *mapping*:

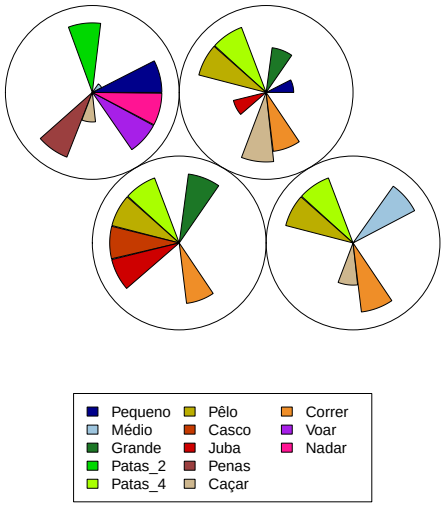



Figura 7.3: *Codebook vectors* do mapeamento $4 \times 4$ dos dados sobre animais.

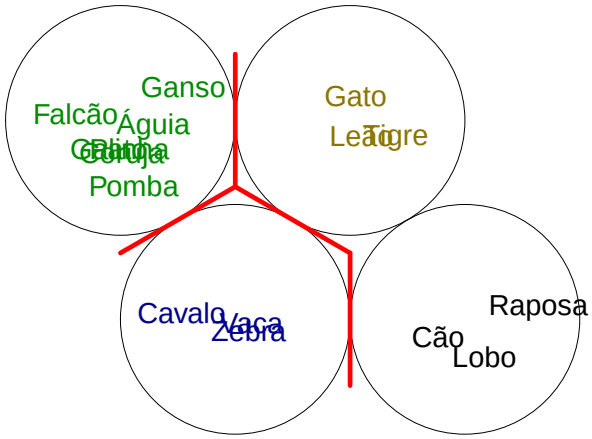



Figura 7.4: *Mapping* para os animais com a técnica de agrupamento *k-medoids* para 3 *clusters*

## 7.2 Simulação

Esta simulação foi criada com o objetivo de auxiliar na leitura dos gráfico fornecidos pelo pacote `kohonen` do *software* R para o SOM. Pretende-se gerar 100 pontos no intervalo $[-1, 1]$, isto é, 100 pontos com as coordenadas:

- $(1, 1)$
- $(-1, 1)$

- $(-1,-1)$
- $(1,-1)$

de modo que cada grupo de 100 pontos sirva como identificação dos extremos dos eixos coordenados. E mais 100 pontos (identificados como 'móveis' ) com a coordenada incial $(0,0)$ que serão deslocados na direção de cada um do extremos dos eixos gerados anteriormente. O objetivo é verificar como o SOM reage aos deslocamentos e se ele será uma ferramenta útil na busca por padrões de fadiga muscular utilizando-se a representação esquemática do JASA (figura 2.1).

Aqui fazer-se-á uso dos gráficos *mapping*, que fornecem a posição de mapeamento dos dados e o *distance neighbours*, que fornece a soma das distâncias em relação a todos os vizinhos imediatos. Todos os gráficos terão acrescidos um recurso denominado *boundaries* que adiciona linhas aos mapas plotados que permitem a visualização das unidades que devem ser agrupadas.

Todos os agrupamentos formados para a inclusão do *boundaries* foi obtido através do método k-*medoids* para no máximo 6 grupos.

### 7.2.1 Etapa I: Geração dos Dados

A primeira coisa a ser feita foi a geração dos pontos e a criação de pares ordenados para cada uma das localizações dos eixos coordenados. Os gráficos *mapping* e *distance neighbours* para o caso em que os pontos 'móveis' situam-se na origem é fornecido abaixo.

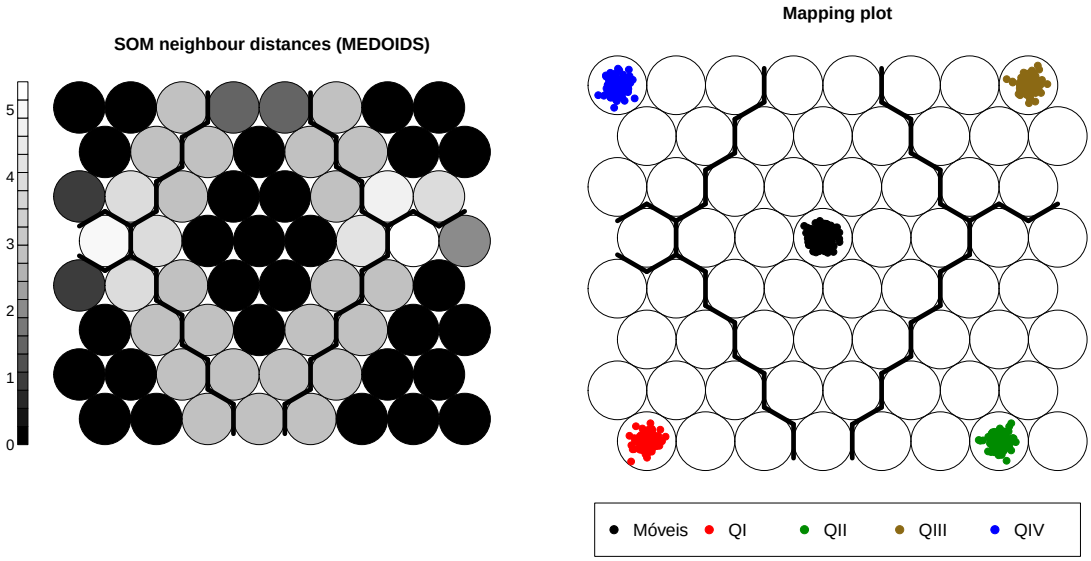



Figura 7.5: *Mapping & Distance Neighbours* - Móveis (0,0)

Cada cor no gráfico *Mapping* indica uma localização do plano coordenado:

- Pontos móveis - preto;
- Pontos do 1º quadrante (1,1) - vermelho;
- Pontos do 2º quadrante (-1,1) - verde;
- Pontos do 3º quadrante (-1,-1) - dourado;
- Pontos do 4º quadrante (1,-1) - azul;

Essas posições devem ser estendidas a todos os gráficos que serão mostrados nesta seção. Perceba como o SOM consegue definir bem o comportamento dos dados. Observando-se ambos os gráficos verifica-se que as posições das nuvens de pontos estão muito bem definidas

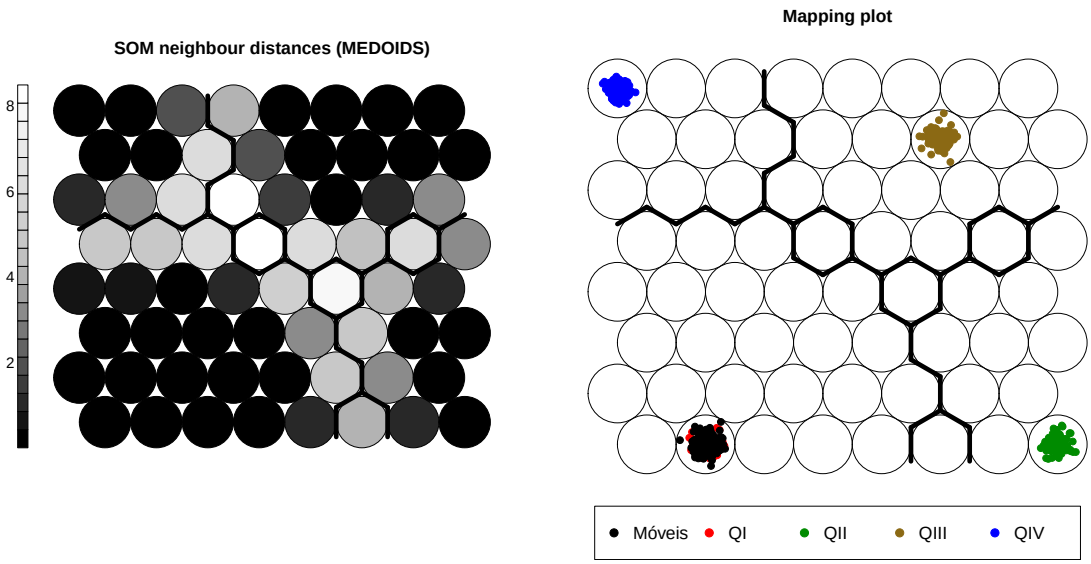



Figura 7.6: *Mapping & Distance Neighbours* - Móveis (1,1)

### 7.2.2 Etapa II: Deslocamento 1º Quadrante

A figura 7.6 corresponde ao deslocamento da nuvem de pontos móveis em direção ao 1ºquadrante. Perceba que a nuvem de pontos móveis é alocada na mesma unidade dos pontos correspondentes ao 1º quadrante (vermelho). Vale ressaltar a forma como a linha de *boundaries* se desloca na mesma direção da nuvem de pontos.

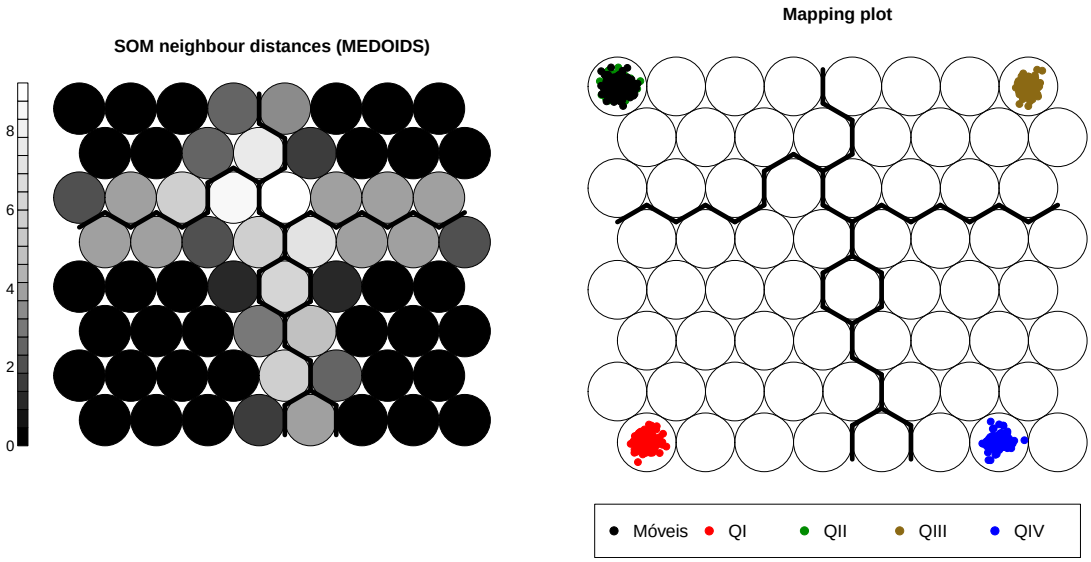



Figura 7.7: *Mapping & Distance Neighbours* - Móveis (-1,1)

### 7.2.3 Etapa III: Deslocamento 2º Quadrante

A figura 7.7 corresponde ao deslocamento da nuvem de pontos móveis em direção ao 1ºquadrante. Perceba que a nuvem de pontos móveis é alocada na mesma unidade dos pontos correspondentes ao 2º quadrante (verde). Vale ressaltar a forma como a linha de *boundaries* se desloca na mesma direção da nuvem de pontos.

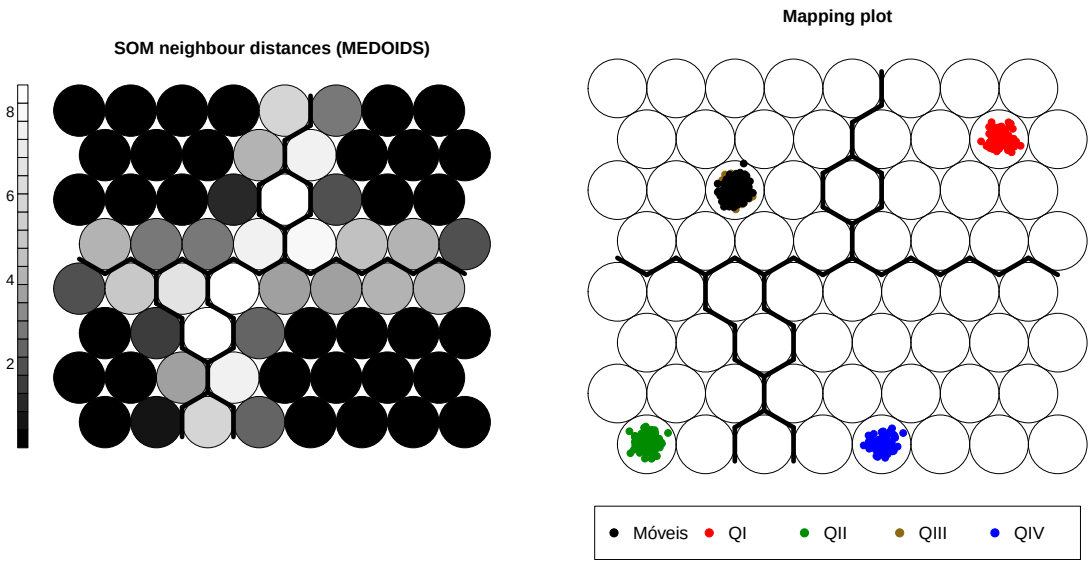



Figura 7.8: *Mapping & Distance Neighbours* - Móveis (-1,-1)

### 7.2.4 Etapa IV: Deslocamento 3º Quadrante

A figura 7.8 corresponde ao deslocamento da nuvem de pontos móveis em direção ao 1ºquadrante. Perceba que a nuvem de pontos móveis é alocada na mesma unidade dos pontos correspondentes ao 3º quadrante (dourado). Vale ressaltar a forma como a linha de *boundaries* se desloca na mesma direção da nuvem de pontos.

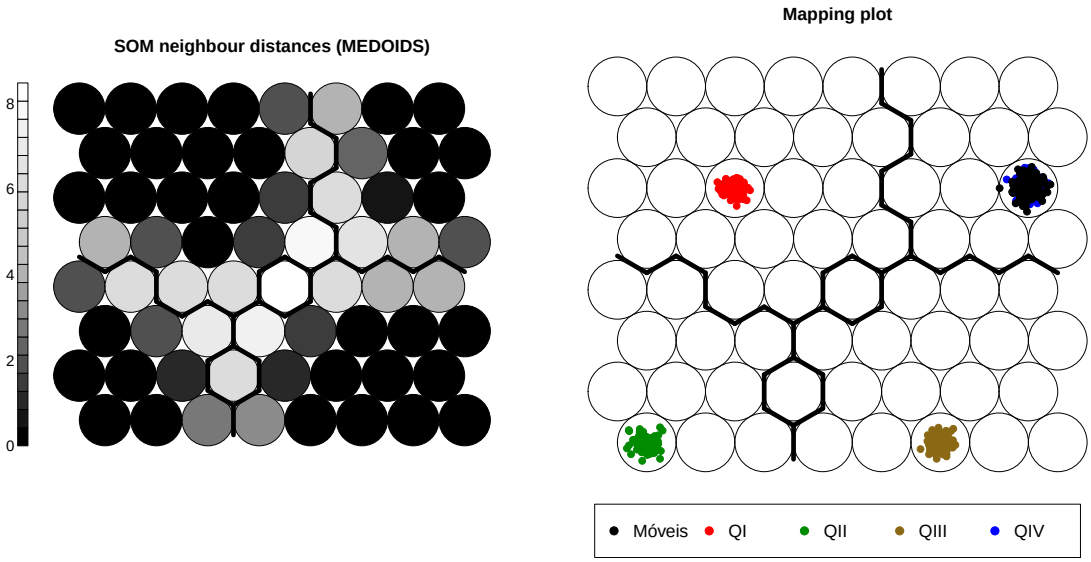



Figura 7.9: *Mapping & Distance Neighbours* - Móveis (1,-1)

### 7.2.5 Etapa V: Deslocamento 4º Quadrante

A figura 7.9 corresponde ao deslocamento da nuvem de pontos móveis em direção ao 1ºquadrante. Perceba que a nuvem de pontos móveis é alocada na mesma unidade dos pontos correspondentes ao 4º quadrante (azul). Vale ressaltar a forma como a linha de *boundaries* não se desloca na mesma direção da nuvem de pontos.

Percebe-se que o SOM pode ser uma ferramenta útil na análise de dados multi-dimensionados. Algo que deve ser destacado é que o comportamento do SOM para o 4º quadrante no gráfico de *distance neighbours* e ao *boundaries* difere um pouco daquilo que era esperado e foi observado para os outros quadrantes. Mesmo assim, o SOM mostrou-se poderoso para auxiliar no estudo.

# 8 Estudo de Caso

## 8.1 Estrutura do Banco de Dados

Os dados foram fornecidos pelo $BIM_L$, do Departamento de Engenharia Elétrica e Computacional da UTEP. Todos os arquivos fornecidos vieram com a extensão `.mat` (Matlab). O primeiro passo, portanto, foi a conversão destes arquivos para `.csv` (Comma Separated Value) já que essa extensão é aceita em diversos *softwares* estatísticos, inclusive os utilizados no trabalho: SAS e R.

Cada indivíduo foi submetido a 5 sessões de teste (uma para cada configuração de peso). Cada sessão tinha duração de 8 horas sendo que o tempo mínimo de espera entre cada sessão era de 48 horas. Os dados foram processados por um aparelho com 8 sensores (*Delsys Bagnoli*-8 *DE-2.1 Standard Differential EMG Electrodes*). Os sinais de EMG , como dito na seção 1.1, são bastante influenciados pelas condições de medição. Uma forma de melhorar essa característica é normalizando os parâmetros do sinal para um valor de referência, no caso, MVC. A idéia é calibrar os valores para uma unidade com uma relevância maior de interpretação. O sinal de cada indivíduo possui como valor de referência sua contração de 100% MVC, que era obtida antes e após as 8 horas de teste. Outra contração utilizada foi a de 70% MVC que era obtida ao final de toda hora (da $2^a$ a $7^a$). Sua duração era baseada na capacidade do indivíduo de manter, ou não, o nível de esforço com um tempo máximo de 3 minutos.

Em cada pasta de arquivo analisada, havia informações sobre 4 tipos de sinais:

- Frequência instantânea (`instfreq`);
- Inclinação da frequência instantânea (`slopefreq`);
- Amplitude instantânea (`instpower`);

- Inclinação da amplitude instantânea (`slopepower`)

Para as 8 horas de informação (H0-H7), dos 3 músculos sob análise (*splenius, trapezius e sternoclidomastoid*), para ambos os lados (esquerdo, direito) e para as duas contrações (70% MVC,100% MVC). A informação em cada pasta era referente a apenas um dentre os 5 pesos utilizados no experimento (A, B, C, D, E). Um exemplo de como os arquivos estavam estruturados segue abaixo.

Tabela 8.1: Identificação dos Arquivos e Pesos

| Subject ID | NFF (Test Number) | CELL (Peso) | 70% MVC | 100% MVC |
|---|---|---|---|---|
| F-1 | 0001 | A | H1-H7 | H0, H7 |
| | 0002 | B | H1-H7 | H0, H7 |
| | 0024 | C | H1-H7 | H0, H7 |

Da tabela 8.1 pode-se observar que a pasta NFF0001 continha informações para o peso A de um indivíduo do sexo feminino, onde (F-1) significa que ela foi a 1ª mulher a passar pelo teste, para as horas H1-H7 com contração de 70% MVC e para as horas H0 e H7 com contração de 100% MVC. Na pasta NFF0002 temos as mesmas informações a respeito de horas, contração e indivíduo, porém, para o peso B. Já para o peso C a pasta passa a ser a de número NFF0024. Perceba que a sequência das pastas para cada indivíduo testado não segue nenhum padrão preestabelecido.

Assim, o primeiro passo no trabalho foi agregar todas as informações disponíveis nas pastas, referentes a todos os indivíduos, num único banco de dados de forma a facilitar a leitura e a análise futura[1]. Nesse passo, foram formados dois bancos (um para cada sexo) dentro de cada hora. Um exemplo de como os bancos ficaram estruturados após o primeiro passo é dado pela tabela 8.2.

Tabela 8.2: Estrutura dos Dados - Passo1 - Sexo Feminino

| ID | CELL | Muscle | Side | Contraction | Hour | Freq | Amp | Timeline (s) |
|---|---|---|---|---|---|---|---|---|
| f01 | A | splenius | lt | 100 | H0 | -0.138265 | -0.00018 | 0.00099 |
| f01 | A | splenius | lt | 100 | H0 | -0.157507 | -0.00017 | 0.00299 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| f12 | D | sterno | rt | 100 | H7 | -0.00250 | 0.00126 | 17.10629 |

Um segundo problema verificado durante a etapa de leitura dos dados foi o grande volume de informação. Juntando todo o conteúdo fornecido tínhamos cerca de 40 GB

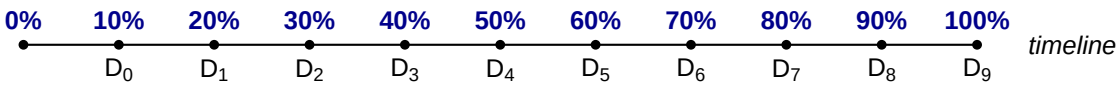
[1] Para essa composição da estrutura dos dados, utilizou-se o `slopefreq` e o `slopepower` como mencionado no artigo de (Luttmann et al., 2000).

de informação que totalizavam mais de 200 milhões de pontos. Assim, necessitávamos realizar uma redução na base de dados a ser analisada, já que, seria inviável a análise do banco completo. Uma forma de reduzirmos a informação pode ser obtida no texto de (Luttmann et al., 2000) durante a explicação do JASA:

> 'Ao aplicar o JASA, o conhecimento sobre o comportamento temporal da amplitude e da frequência ou de suas medidas é necessário. Em estudos anteriores, tais informações foram obtidas a partir de gravações de EMG de longo prazo, calculando-se os valores médios da amplitude e da frequência para períodos curtos e sucessivos de tempo (por exemplo, $5s$ ou $10s$). Este procedimento resulta em séries temporais para ambas as características com o respectivo intervalo amostral. Amplitude e frequência das séries temporais foram resumidos por análises de regressão e os seus coeficientes de regressão foram considerados como indicadores quantitativos da mudança temporal na amplitude e na frequência do sinal de EMG'.

Para a presente análise, as informações foram obtidas a partir de gravações de EMG de longo prazo, calculando-se os valores medianos da amplitude e da frequência para os decis de períodos sucessivos de tempo.

Para melhor entendimento, observe a figura abaixo.




Considere que a mesma represente os dados da hora inicial (H0). Os dados utilizados na análise correspondiam, justamente, à mediana da amplitude e à mediana da frequência para cada decil. Deste modo, para cada indivíduo na hora H0, ter-se-á 10 valores representando a amplitude e 10 valores representando a frequência. Deste modo, o volume de informação deve cair drasticamente[2].

A configuração final do banco de dados utilizado na análise é:

[2] Vale destacar que a utilização do decil é apenas uma opção. Poder-se-ia utilizar os percentis ou qualquer outro critério que o pesquisador julgar necessário. O que deve ser destacado é que qualquer critério deve ser escolhido de maneira criteriosa já que o mesmo poderá interferir nos resultados obtidos.

Tabela 8.3: Estrutura dos Dados - Sexo Feminino

| ID | CELL | Muscle | Side | Hour | Med_Freq | Med_Amp | Med_Timeline (s) |
|---|---|---|---|---|---|---|---|
| f01 | A | splenius | lt | H0 | -0.03768 | 5.89352e-05 | 0.00300 |
| f01 | A | splenius | lt | H0 | -0.15751 | -0.000172 | 0.97490 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| f12 | D | sterno | rt | H7 | -1.14033e-02 | -6.60750e-03 | 37.165780 |

Devemos alertar que o banco de dados possui perdas de informação. Na hora 4, por exemplo, temos a ausência das informações a respeito de uma mulher, que no caso, seria a F-12 ($12^{a}$ mulher a ser testada). Além disso, para muitos indivíduos não temos a informação completa a respeito dos pesos, isto é, existem indivíduos para os quais temos informações a respeito dos pesos A,B,D mas não temos de C e E.

## 8.2 Análise Descritiva

A análise exploratória de dados, cujo objetivo básico é o de sintetizar uma série de valores de mesma natureza, emprega grande variedade de técnicas gráficas e quantitativas que visam maximizar a obtenção de informações ocultas na estrutura dos dados, permitindo:

- que se tenha uma visão global da variação e organização desses valores;
- a descoberta de variáveis importantes e detecção de comportamentos anômalos ao fenômeno.

Esta seção compreende a análise exploratória dos dados referentes ao músculo *splenius capitis*, lado esquerdo e peso B para todos os indivíduos do sexo feminino. Todas as informações que constam a partir desta seção em diante foram desenvolvidas no software R.

### 8.2.1 Informação Geral

Pensando na representação esquemática sugerida pelo JASA, a primeira análise a ser discutida é a respeito do comportamento dos dados longitudinalmente, ou seja, como os indivíduos se comportavam de acordo com o transcorrer do tempo, utilizando-se para tal de um gráfico *amplitude* $\times$ *frequência*.

Devido à grande quantidade de observações a serem plotadas, esta tarefa tornou-se muito complicada de ser realizada, do ponto de vista da qualidade da informação fornecida pelo gráfico, já que a limitação do espaço gráfico dificulta a percepção de mudanças, principalmente, mudanças sutis.[3]

Antes de explicar a figura 8.2, algumas estatísticas básicas e os Boxplots das variáveis sob análise são fornecidos.

Tabela 8.4: Medidas Estatísticas Básicas

| | Frequência Mediana | Amplitude Mediana |
|---|---|---|
| Mínimo | -0.18779 | -4.922e-03 |
| 1º Quartil | -0.00377 | -1.306e-04 |
| Mediana | 0.00237 | -2.816e-05 |
| Média | 0.00168 | -1.011e-04 |
| 3º Quartil | 0.00790 | 7.149e-07 |
| Máximo | 0.12255 | 4.058e-03 |
| Amplitude | 0.31034 | 0.0089797 |
| Distância Interquartílica | 0.01166 | 1.312e-04 |

Pelos dados da tabela 8.4 é possível verificar como a intensidade de variação entre as duas variáveis é diferente. Enquanto que na frequência mediana a intensidade é da ordem de $10^{-1}$, a intensidade da amplitude mediana é de $10^{-3}$. A diferença entre os valores extremos da amplitude mediana também chama a atenção devido à grandeza da variação ser muito pequena. Os boxplots refletem que ambas as variáveis possuem um número elevadíssimo de valores atípicos. Cerca de 25% dos dados sob análise foram considerados atípicos, utilizando-se para a construção dos limites a distância interquartílica multiplicada por 3, tanto para a frequência mediana, quanto para a amplitude mediana. Vale frisar que esses valores atípicos não podem nem devem ser excluídos da análise, já que, não se espera que indivíduos comportem-se de maneira igual e uniforme.

A figura 8.2 representa um gráfico (x,y) com *x=abscissas*, representando a amplitude e *y=ordenadas*, representando a frequência. Cada linha representa, respectivamente, uma das 8 horas sob análise (H0-H7), cada coluna os respectivos decis ($D_0$-$D_9$) e cada ponto uma mulher para aquele decil e aquela hora específica. Frente ao exposto anteri-

[3] Na UTEP, existe um sistema conhecido como *cybershare*. Este sistema permite interligar em rede 45 monitores conduzidos por 45 estações de trabalho. Utilizando-se deste recurso, ter-se-ia a possibilidade de alocação de cada gráfico individualmente, em cada monitor, ou seja, cada um dos gráficos plotados apareceriam cada um em um monitor. Isso permitira uma melhora na capacidade de visualização e na percepção de características de interesse. Para mais informações acessar: *http://cybershare-portal.utep.edu/c2vis*.

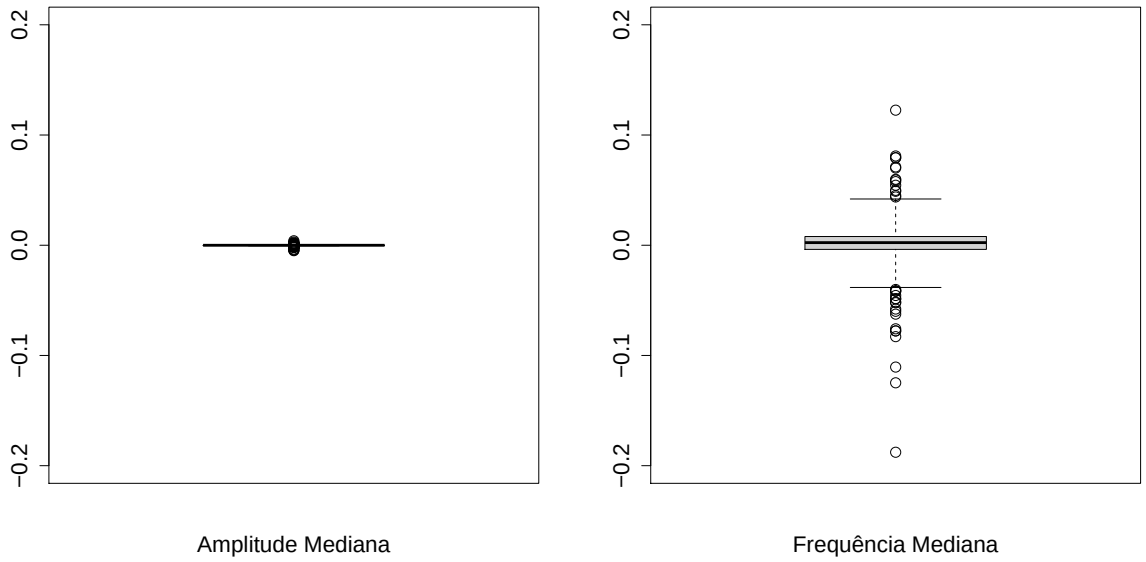



Figura 8.1: BoxPlot das variáveis Amplitude Mediana e Frequência Mediana.

** **Para o cálculo dos limites superiores e inferiores dos BoxPlots foi considerada a distância interquartílica multiplicada por 3.***

ormente, tem-se que a pequena variação entre os valores extremos da amplitude acabam interferindo no aspecto do gráfico 8.2. Perceba como é difícil, a partir da hora 1, verificar a evolução do estado do músculo de cada mulher na hora. Todas parecem não sair da origem (0,0).

Uma sugestão para tentar melhorar o aspecto do gráfico 8.2 foi transformar todos os valores das variáveis amplitude mediana e frequência mediana para o intervalo $[-1, 1]$. Perceba que a transformação manterá as ordens de grandeza das duas variáveis. O objetivo desta transformação é melhorar a relação de proporcionalidade do gráfico fazendo com que ambos os eixos variem no mesmo intervalo.

Considere:

- ANT = antigo valor da variável;
- NOVO = novo valor da variável.

A fórmula para a transformação de variáveis que estejam distribuídas sob qualquer intervalo para um intervalo compreendido entre $-1$ e $1$ é:

$$\frac{\text{NOVO} - (-1)}{1 - (-1)} = \frac{\text{ANT} - \text{MIN(ANT)}}{\text{MÁX(ANT)} - \text{MIN(ANT)}}$$

$$\text{NOVO} = 2\left[\frac{\text{ANT} - \text{MIN(ANT)}}{\text{MÁX(ANT)} - \text{MIN(ANT)}}\right] - 1 \tag{8.1}$$

Outra sugestão discutida e também adotada foi 'forçar' os valores atípicos a assumirem o valor do respectivo limite superior e/ou inferior[4] de modo a atenuar o efeito desses valores atípicos.

[4] $LI = q_1 - (3)d_q$
$LS = q_3 + (3)d_q$

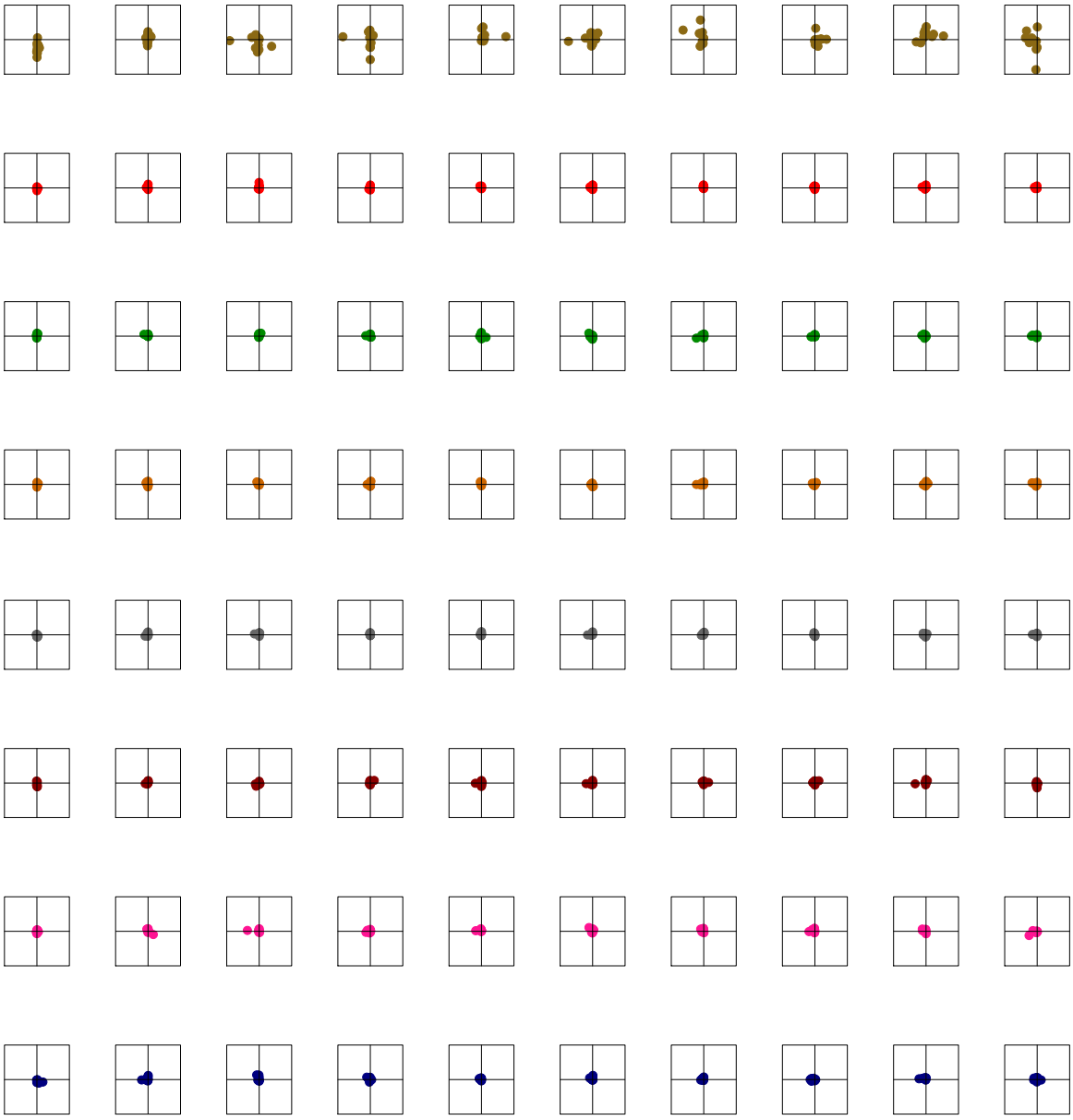


Figura 8.2: Gráficos (x,y) - Indivíduos - Informação Original (H0-H7)

Os novos resultados obtidos seguem abaixo. Perceba pela figura 8.3 que não há mais valores discrepantes em nenhuma das duas variáveis. Ambas comportam-se da mesma maneira, porém os valores da média e da mediana continuam diferentes. Perceba que as transformações sugeridas surtiram efeito e a figura 8.4 passou a refletir melhor o comportamento das variáveis, principalmente da amplitude. Porém, não foi possível, através dessas informações, reconhecer algum padrão de fadiga muscular. Esperava-se que as características dos indivíduos no tempo ficassem melhor definidas e, consequentemente, que a percepção de padrões de comportamento fossem mais facilmente observadas.

Tabela 8.5: Medidas Estatísticas Básicas

| | Frequência Mediana $[-1,1]$ | Amplitude Mediana $[-1,1]$ |
|---|---|---|
| Minimo | -1.00000 | -1.000000 |
| 1º Quartil | -0.14286 | -0.14286 |
| Mediana | 0.00735 | 0.08004 |
| Média | -0.00180 | -0.03212 |
| 3º Quartil | 0.14286 | 0.14286 |
| Máximo | 1.00000 | 1.00000 |
| Amplitude | 2.00000 | 2.00000 |
| Distância Interquartílica | 0.28572 | 0.28572 |

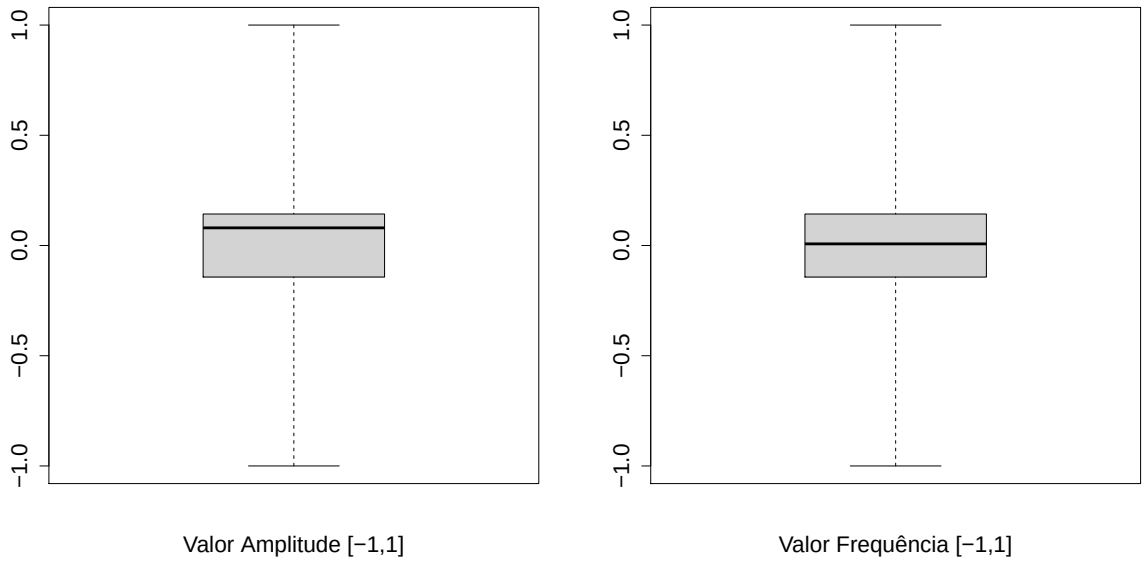



Figura 8.3: BoxPlot das variáveis no intervalo $[-1,1]$.

** **Para o cálculo dos limites superiores e inferiores dos BoxPlots foi considerada a distância interquartílica multiplicada por 3.***

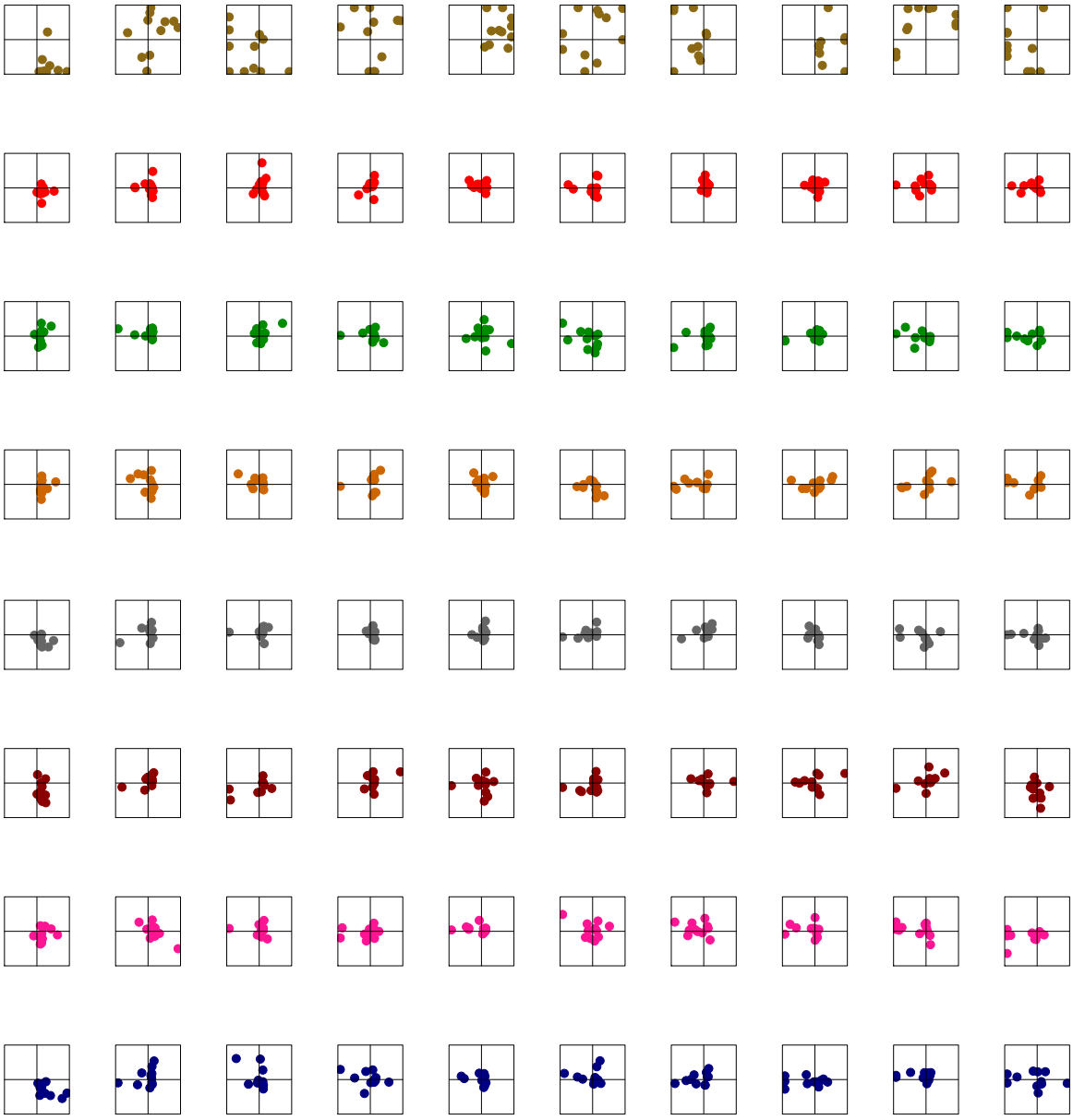


Figura 8.4: Gráficos (x,y) - Indivíduos - Variáveis no Intervalo $[-1, 1]$ (H0-H7)

## 8.3 Análise de Agrupamento e SOM

Nesta seção, deseja-se verificar se através de métodos de agrupamento consegue-se identificar padrões de comportamento entre indivíduos do mesmo grupo durante os testes. Como em todas as análises do trabalho, a configuração esquemática do JASA continua sendo a idéia base.

### 8.3.1 Radial

Segundo a configuração esquemática do JASA, a análise conjunta da amplitude e da frequência funcionam como 'coordenadas' para localização do posicionamento dos indivíduos. Dependendo dos valores desses dois parâmetros consegue-se posicionar os indivíduos no gráfico do JASA e dizer o estado em que se encontra o músculo.

A partir disso e percebendo que a simples representação em um gráfico $(x, y)$ não estava atingindo o objetivo traçado, resolveu-se, utilizar métodos de agrupamento que pudessem levar em consideração o posicionamento do indivíduo durante os testes.

Sabe-se que cada indivíduo, durante o teste, respondia a um questionário informando seu estado antes de iniciar o procedimento e que durante o procedimento esse relatório ia sendo atualizado. Assim sendo, existe no relatório todo o histórico de sintomas que o indivíduo relatou durante os testes.

Desta feita, acredita-se que agrupando os dados e verificando através do relatório se, naquele momento específico, o indivíduo relatou estar sentindo algum desconforto muscular, ou algo que o fizesse ser classificado como em estado de fadiga, ou qualquer outro estado definido no JASA, permitirá relacionar essas características aos indivíduos localizados na mesma região e/ou aos indivíduos do mesmo grupo.

O radial foi idealizado da seguinte maneira:

- dividiu-se cada quadrante em 9 áreas baseando-se para isso na variação do raio e do ângulo (em radianos) correspondente a cada local;
- para o agrupamento, considerou-se o centro de massa de cada local, como a informação de cada indivíduo. Desta forma, todos os indivíduos foram resumidos pelo ponto central da área a que ele pertence. Portanto, cada indivíduo possui como variável um par de coordenadas $(r, \theta)$.

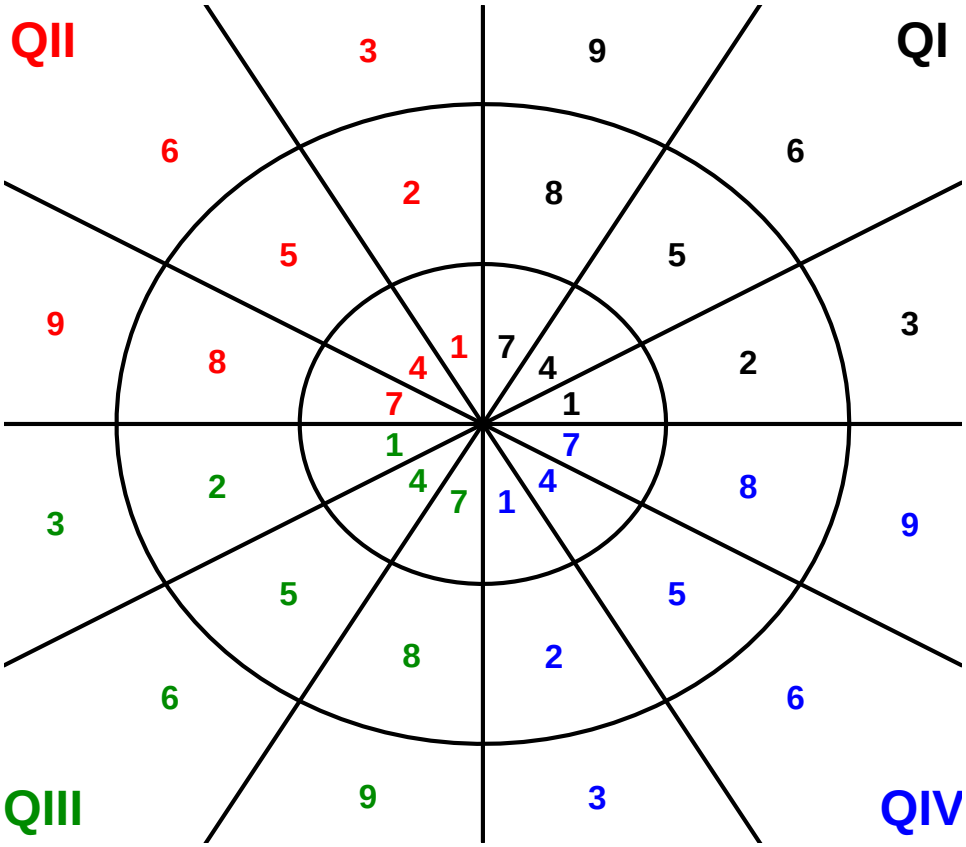



Figura 8.5: Gráfico Radial

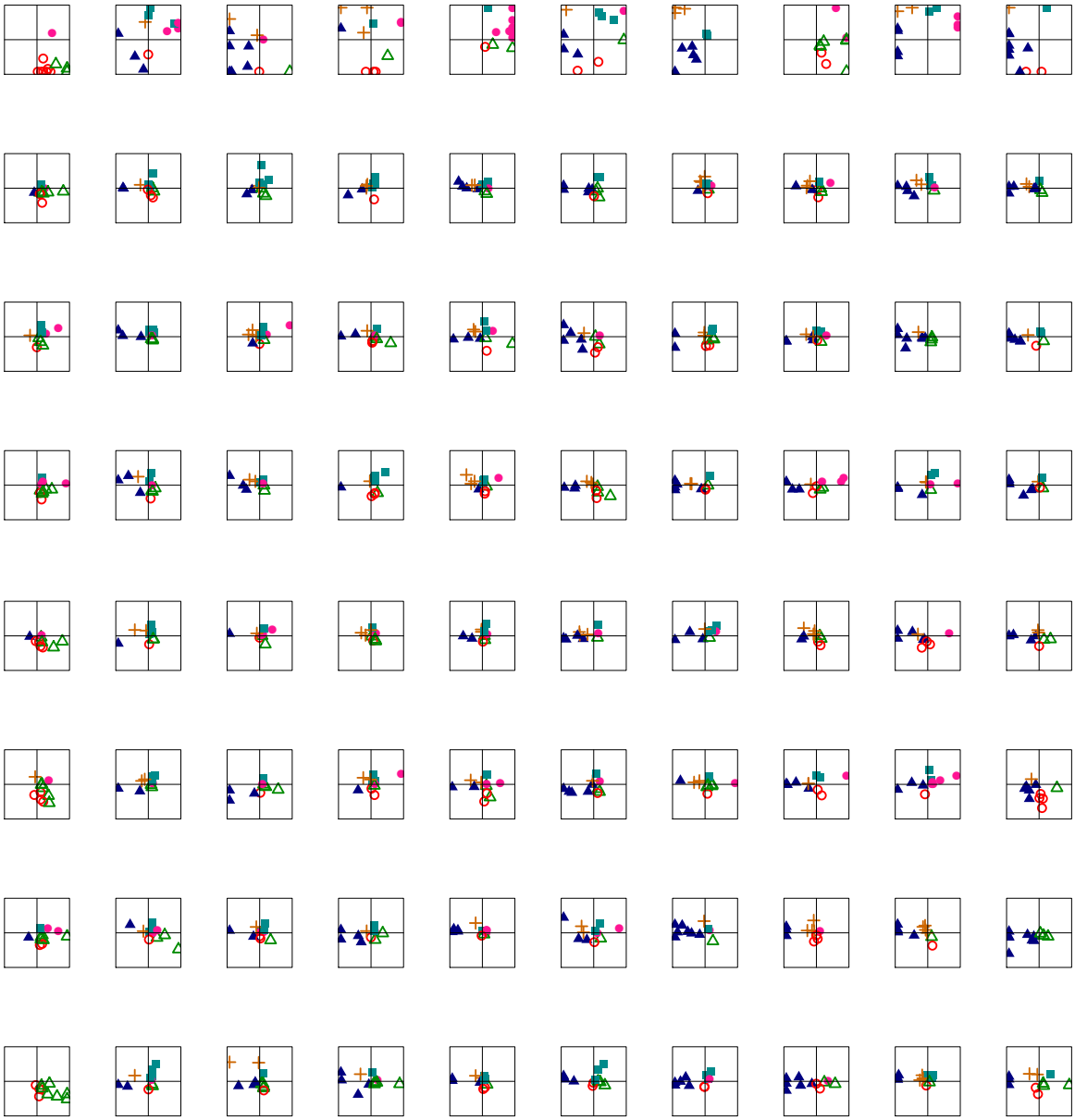


Figura 8.6: Gráfico (x,y) - K-*Medoids* - (H0-H7)

Figura 8.7: Gráfico (x,y) - K-*Means* - (H0-H7)

Os dois métodos de agrupamento utilizados foram o k-*means* e o k-*medoids*. As figuras 8.6 e 8.7 representam um gráfico (x,y) com $x$=*abscissas*, representando a amplitude transformada (intervalo $[-1,1]$ com valores discrepantes iguais aos respectivos limites superior/inferior) e $y$=*ordenadas*, representando a frequência transformada (intervalo $[-1,1]$ com valores discrepantes iguais aos respectivos limites superior/inferior). Cada linha representa, respectivamente, uma das 8 horas sob análise (H0-H7), cada coluna os respectivos decis ($D_0$-$D_9$) e cada ponto uma mulher para aquele decil e aquela hora específica.

Cada cor simbolizada e símbolo diferente estão relacionados intrinsecamente a cada *cluster* e não ao indivíduo. Portanto, é possível que os indivíduos mudem de cor em todos os gráficos dos decis. O número máximo de *clusters* fornecido por cada método foi 6. Importante observar que a alocação dos valores dos *clusters* para estes métodos é aleatória, deste modo, para um mesmo indivíduo, o k-*means* pode ter considerado ele como *cluster* 1, enquanto o k-*medoids* o considerou como *cluster* 3. É por isso que os gráficos apresentam distribuição de cores diferentes. O mais importante aqui é verificar se os dois métodos alocam os indivíduos no mesmo *cluster*.

Relação *cluster*-cor:

- *cluster* 1 - vermelho
- *cluster* 2 - verde
- *cluster* 3 - laranja
- *cluster* 4 - cyan
- *cluster* 5 - rosa
- *cluster* 6 - azul

Mais uma vez, a limitação espacial dificultou a procura por padrões de comportamento. Na hora inicial dá para perceber os grupos seguindo uma tendência através dos decis, porém para as outras horas, o gráfico não foi elucidativo. Os métodos de agrupamento cumprem muito bem o seu papel de alocação das variáveis e são consistentes. Os métodos divergem na formação dos grupos, porém o k-*medoids* possui duas vantagens em relação ao método k-*means*.

1. mais robusto a presença de valores discrepantes;

2. independe da ordem em que os objetos são examinados.

Desta forma, possivelmente a alocação fornecida pelo k-*medoids* é mais confiável do que àquela fornecida pelo k-*means*. Ainda assim, nenhuma conclusão, com relação ao pretendido através do JASA, pode ser tomada. Mais testes devem ser realizados, incluindo até outros métodos de agrupamento como, por exemplo, técnicas hierárquicas.

Perceba mais uma vez como seria interessante, para este caso, a tecnologia do *cybershare*. O poder de visualização que este recurso fornece definitivamente poderia auxiliar no desenvolvimento visual de qualquer projeto.

### 8.3.2 SOM

Outra maneira de realizar o agrupamento, desta vez dando mais ênfase em cada indivíduo, é através do SOM. Uma maneira de encontrar grupos é a realização de um agrupamento dos *codebook vectors*[5] individuais. A vantagem de agrupar os *codebook vectors*, ao invés dos dados originais, é que o número de unidades é geralmente de ordem inferior ao número de objetos.

As figuras 8.8, 8.9, 8.10, 8.11, 8.12, 8.13, 8.14, 8.15 fornecem a estrutura de agrupamento que fora realizada, anteriormente, através da idéia radial e que agora é realizada através do SOM. Perceba o SOM também permite acompanhar, assim como nos gráficos anteriores, a evolução dos indivíduos no decorrer do tempo. Cada gráfico é referente a apenas uma hora sendo que a primeira linha corresponde aos 5 primeiros decis da hora, enquanto que, a terceira aos 5 últimos.

Dois tipos de gráficos foram plotados para cada hora de teste: o primeiro (correspondente às linhas ímpares) é denominado *mapping*. Ele mostra aonde os objetos foram mapeados. Ele necessita de argumentos de classificação, no caso, os *clusters*. O segundo (correspondente às linhas pares) é denominado *distance neighbours*. Ele fornece a soma das distâncias em relação a todos os 'vizinhos' imediatos. Este tipo de visualização também é conhecido como gráfico da matriz-U. Espera-se que unidades com limites de classificação próximos possuam distâncias médias elevadas com relação ao seu 'vizinho' e que indivíduos localizados nas mesmas unidades possuam pequena distância entre si, isto

[5] Os *codebook vectors* devem ser vistos como um resumo conciso dos dados originais. Um exemplo de *codebook vectors* é dado pela figura 7.3.

é, sejam altamente homogêneos. Um bom *mapping* deve mostrar pequenas distâncias por toda a sua extensão.

Quanto às cores segue o exposto para o gráfico anterior. Cores iguais significam que o *cluster* fornecido pelo método de agrupamento foi o mesmo para determinados indivíduos.

Com relação ao *distance neighbours*, a gradação das cores segue a soma das distâncias em relação a todos os 'vizinhos' imediatos. Para ler esse gráfico basta seguir o estabelecido pela régua situada do lado esquerdo do gráfico. Distâncias pequenas indicam que os objetos alocados naquelas unidades são homogêneos. Distâncias grandes indicam que os objetos alocados naquelas unidades são pouco homogêneos e que possivelmente só foram alocados na mesma unidade devido à imposição de uma regra externa ao agrupamento, no caso, um número máximo de grupos. Esse gráfico fornece três informações:

1. fornece a soma das distâncias em relação às observações situadas nas mesmas camadas;

2. fornece a soma das distâncias em relação às observações situadas nas camadas 'vizinhas';

3. fornece informação acerca do agrupamento através do *boundaries*.

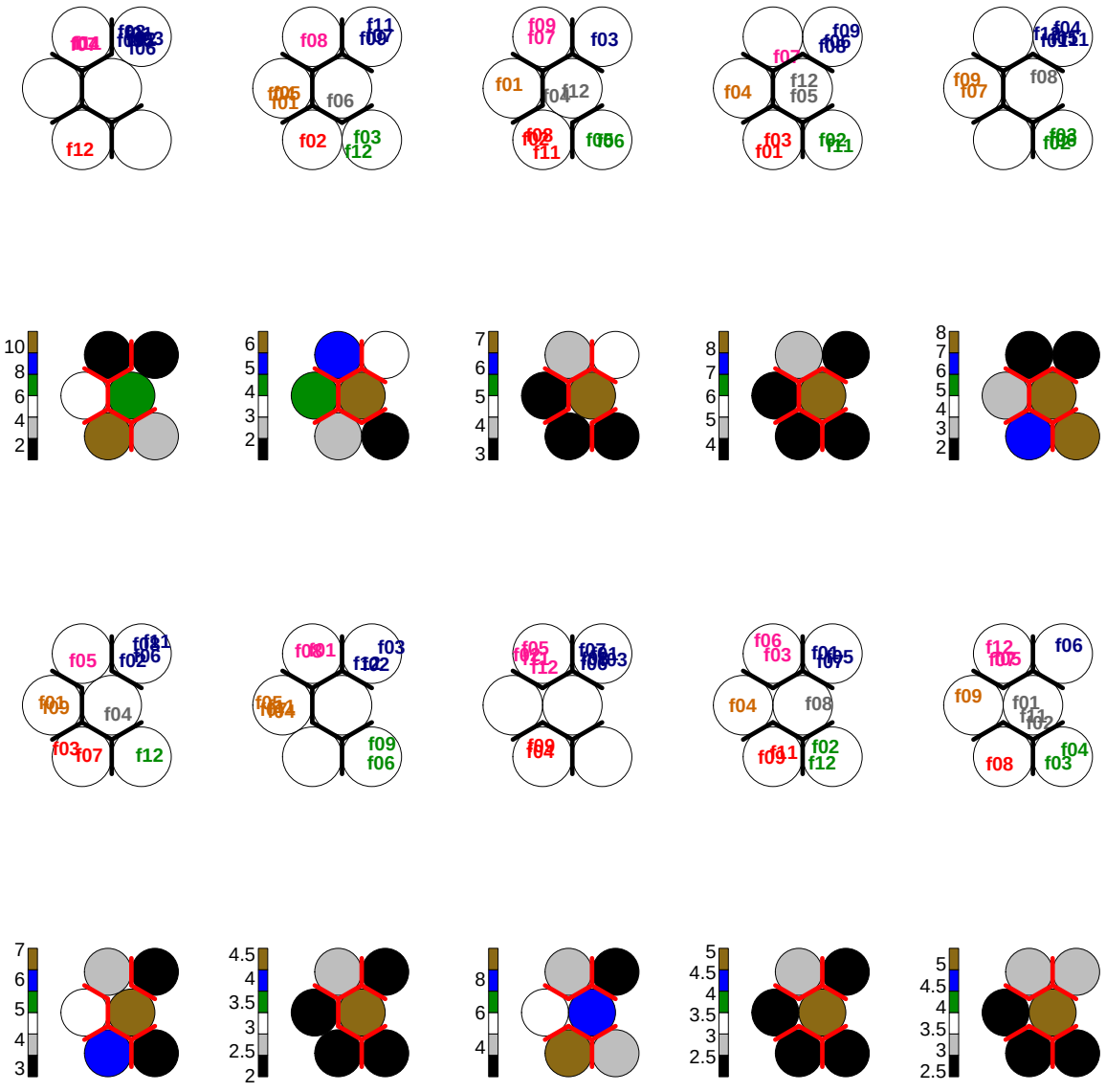



Figura 8.8: SOM - Mapping & Distance Neighbours - (H0)

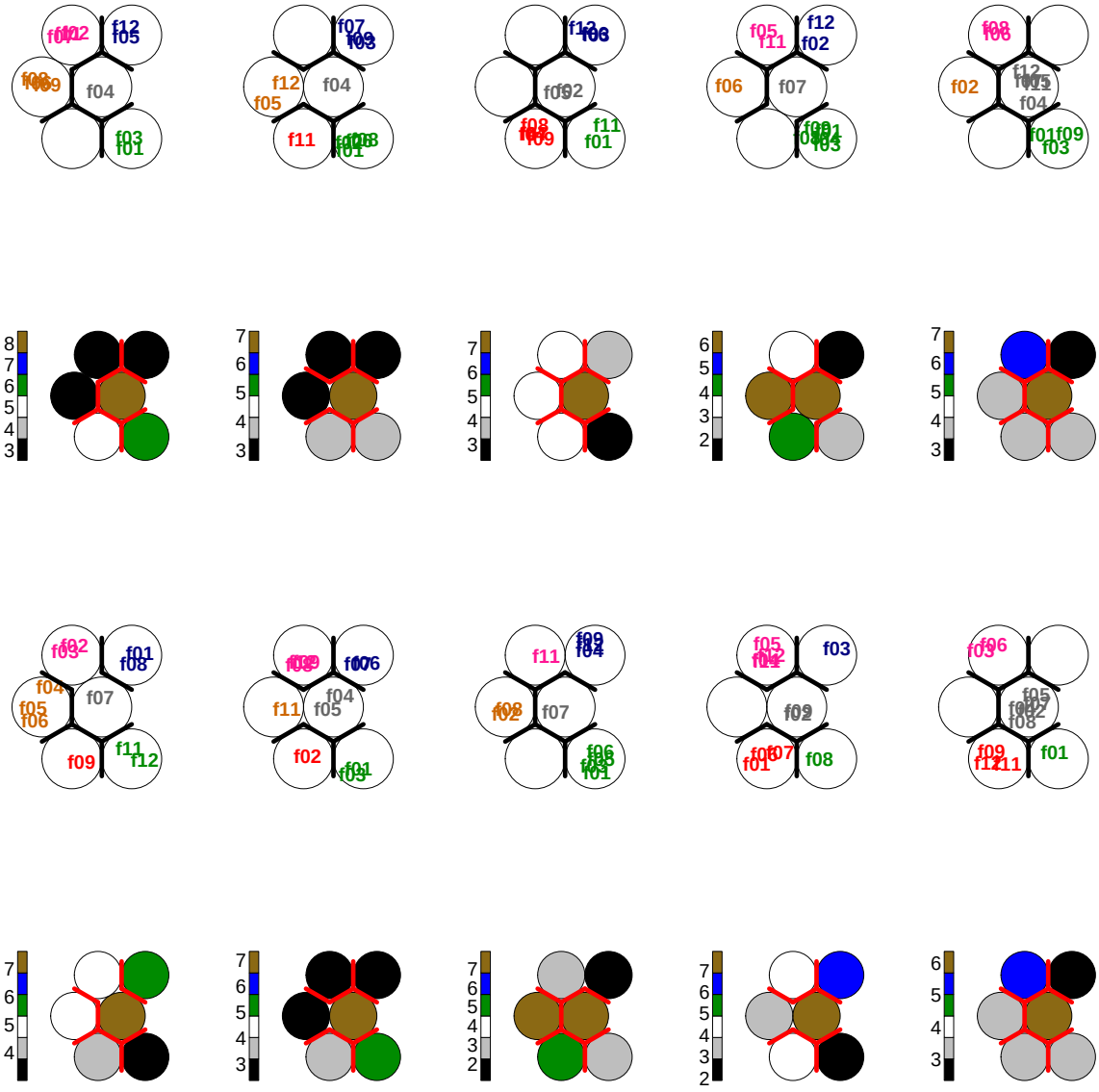


Figura 8.9: SOM - Mapping & Distance Neighbours - (H1)

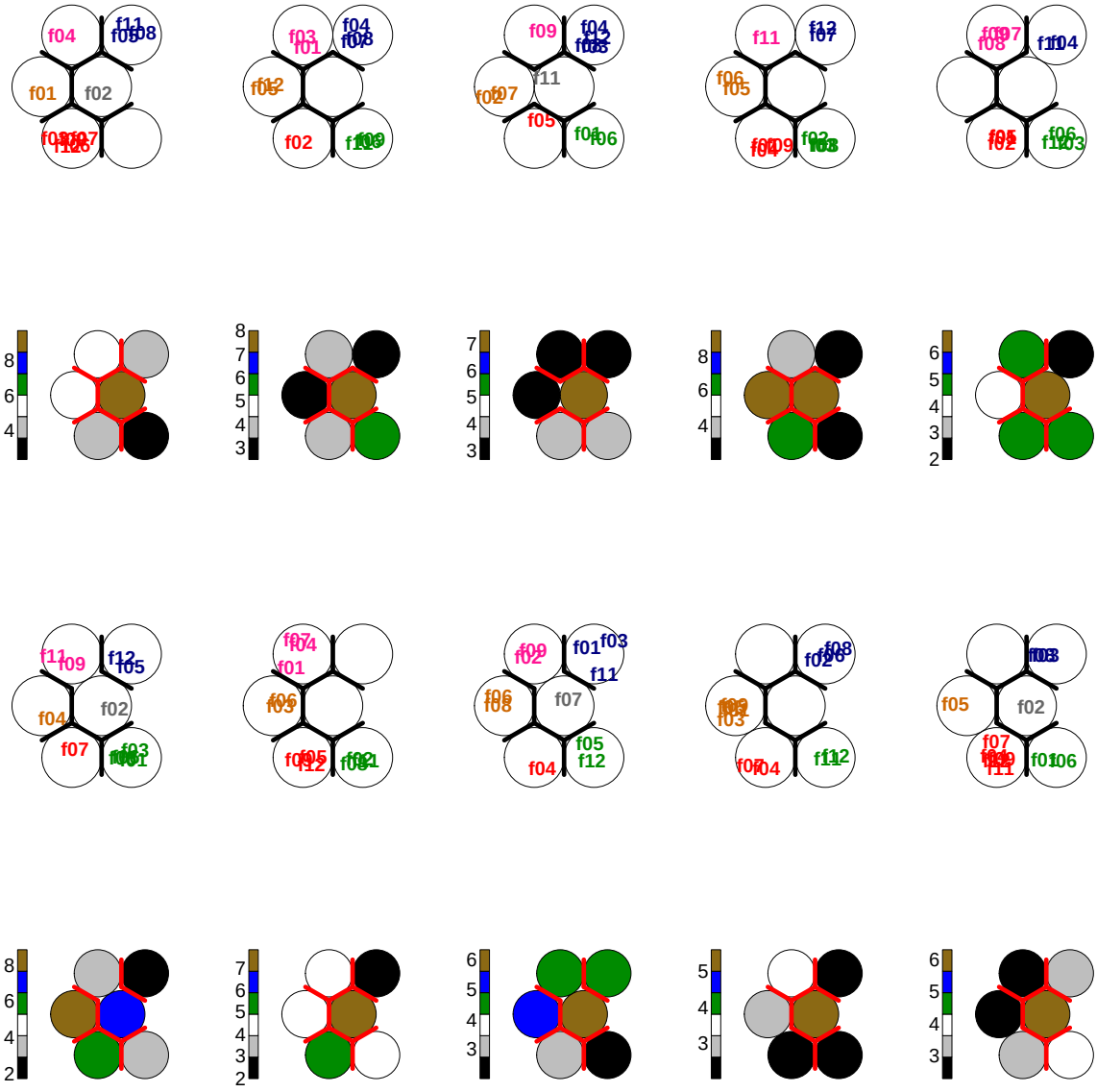


Figura 8.10: SOM - Mapping & Distance Neighbours - (H2)

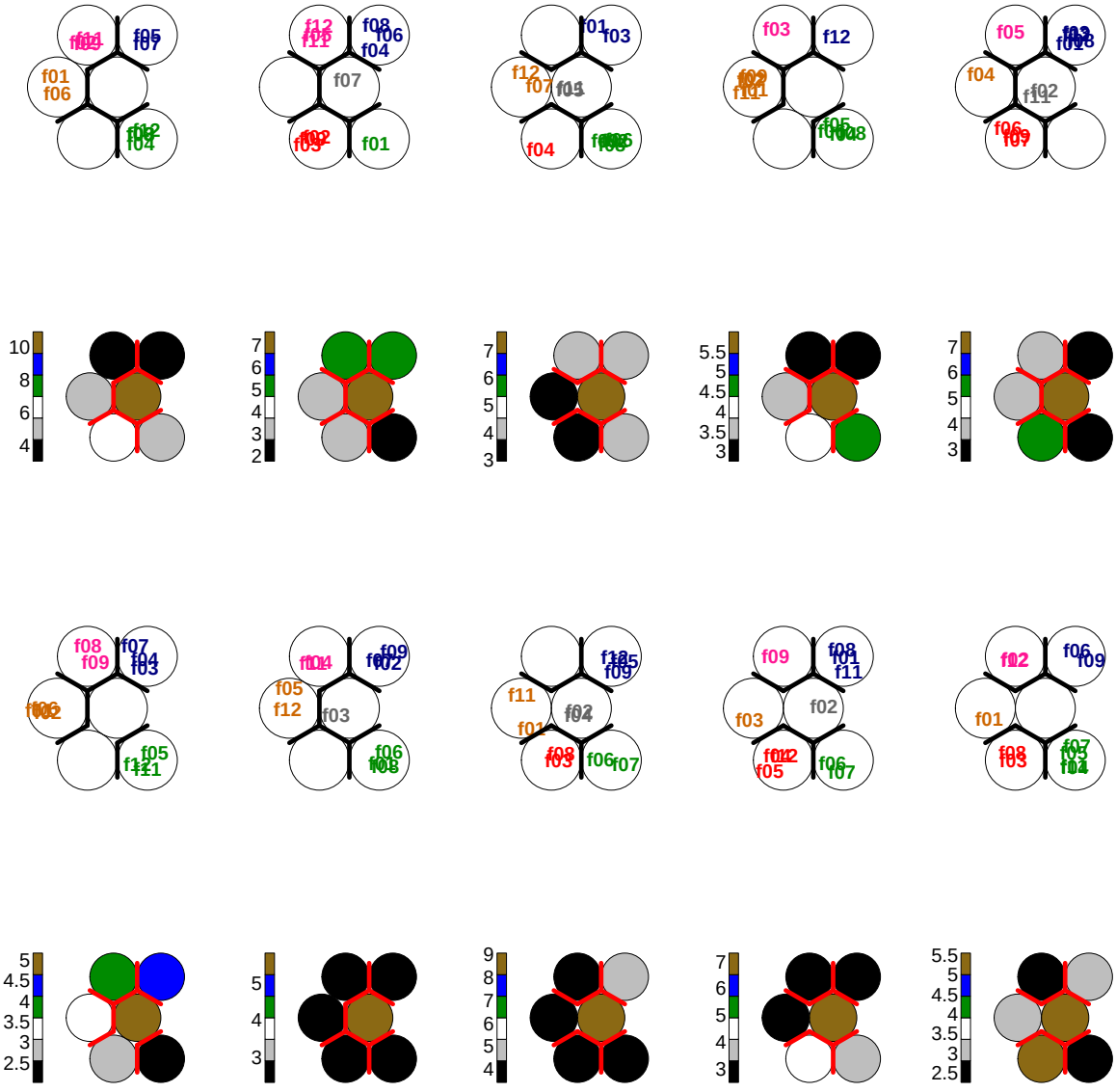


Figura 8.11: SOM - Mapping & Distance Neighbours - (H3)

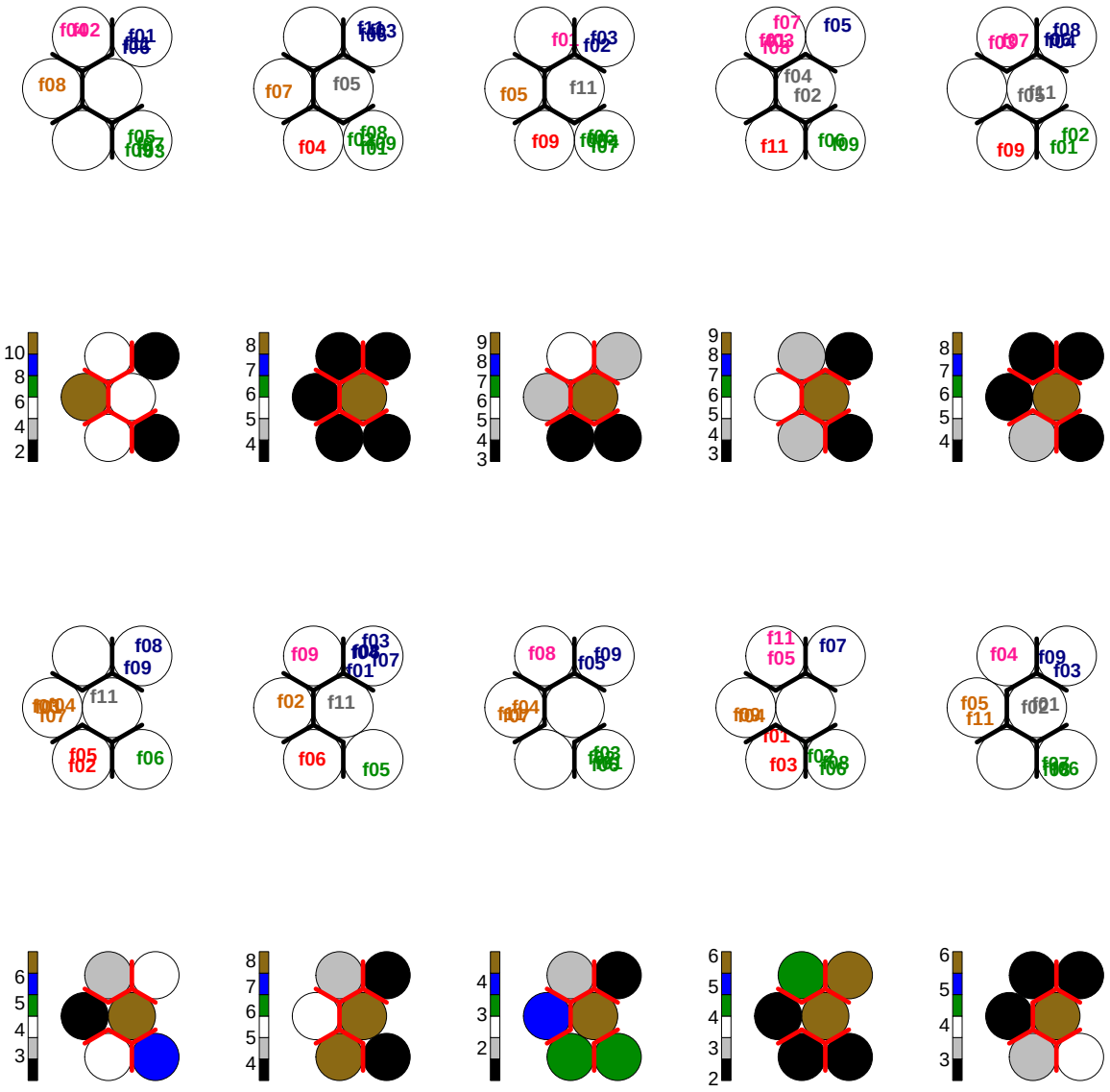



Figura 8.12: SOM - Mapping & Distance Neighbours - (H4)

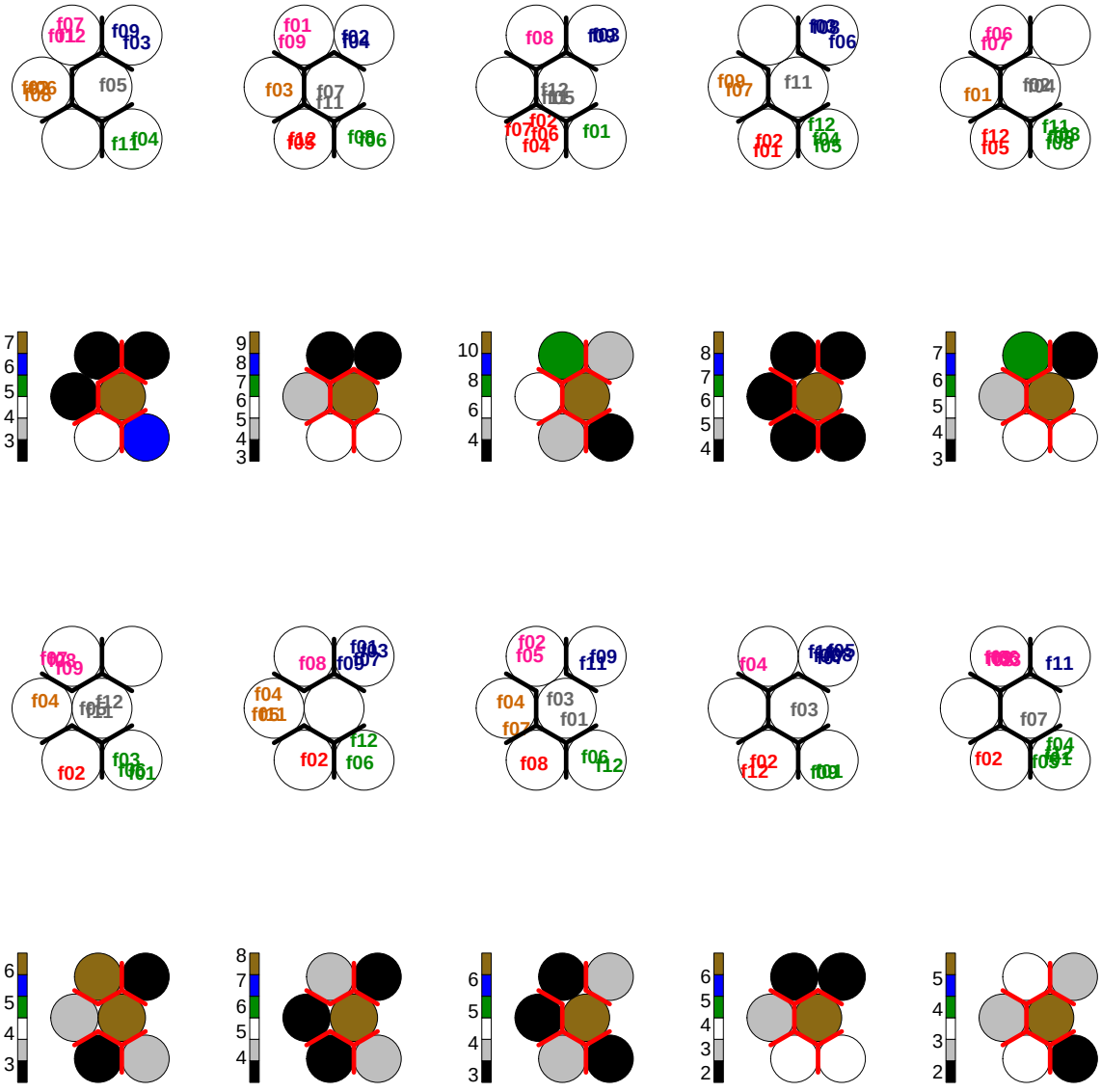


Figura 8.13: SOM - Mapping & Distance Neighbours - (H5)

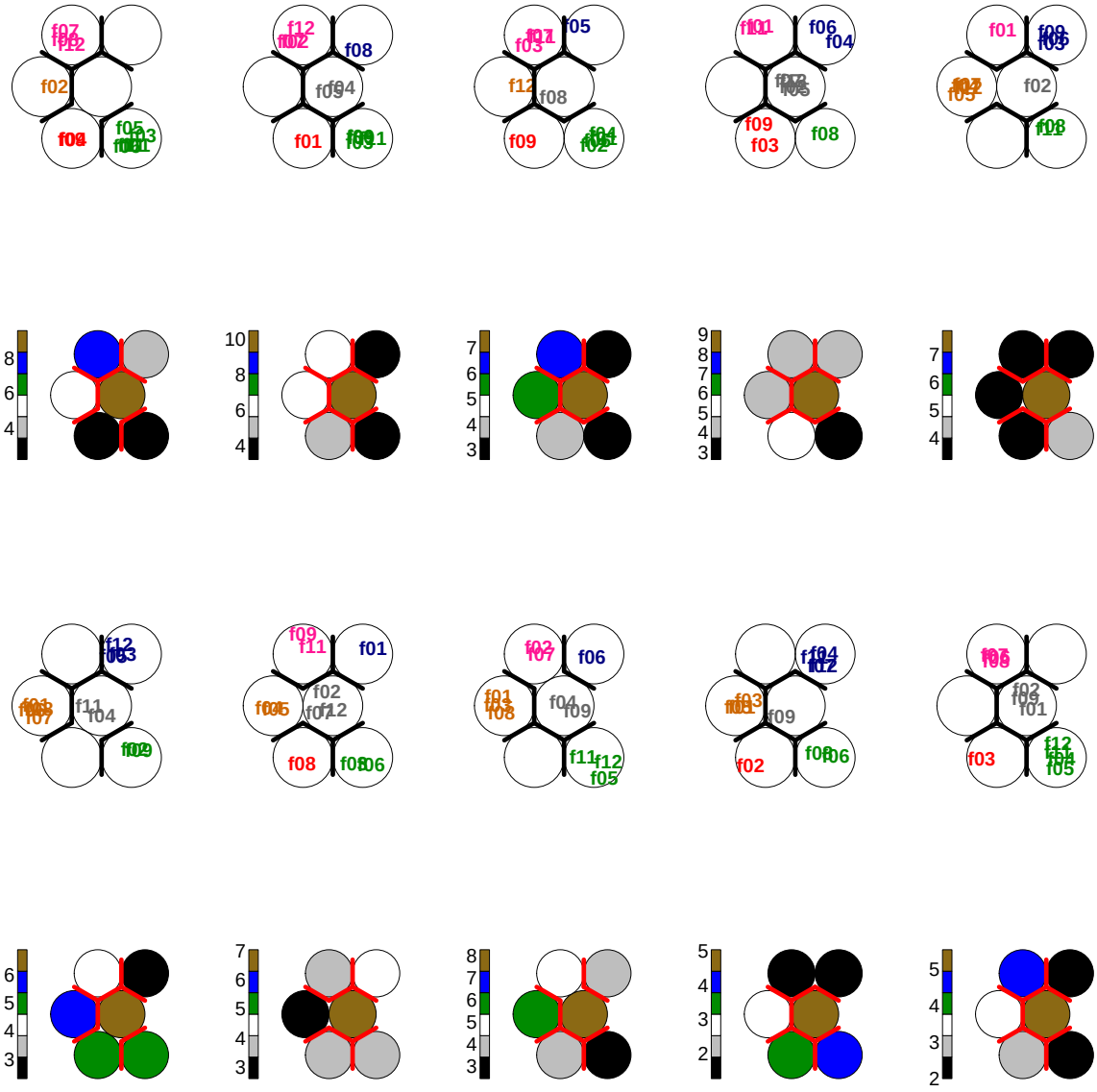


Figura 8.14: SOM - Mapping & Distance Neighbours - (H6)

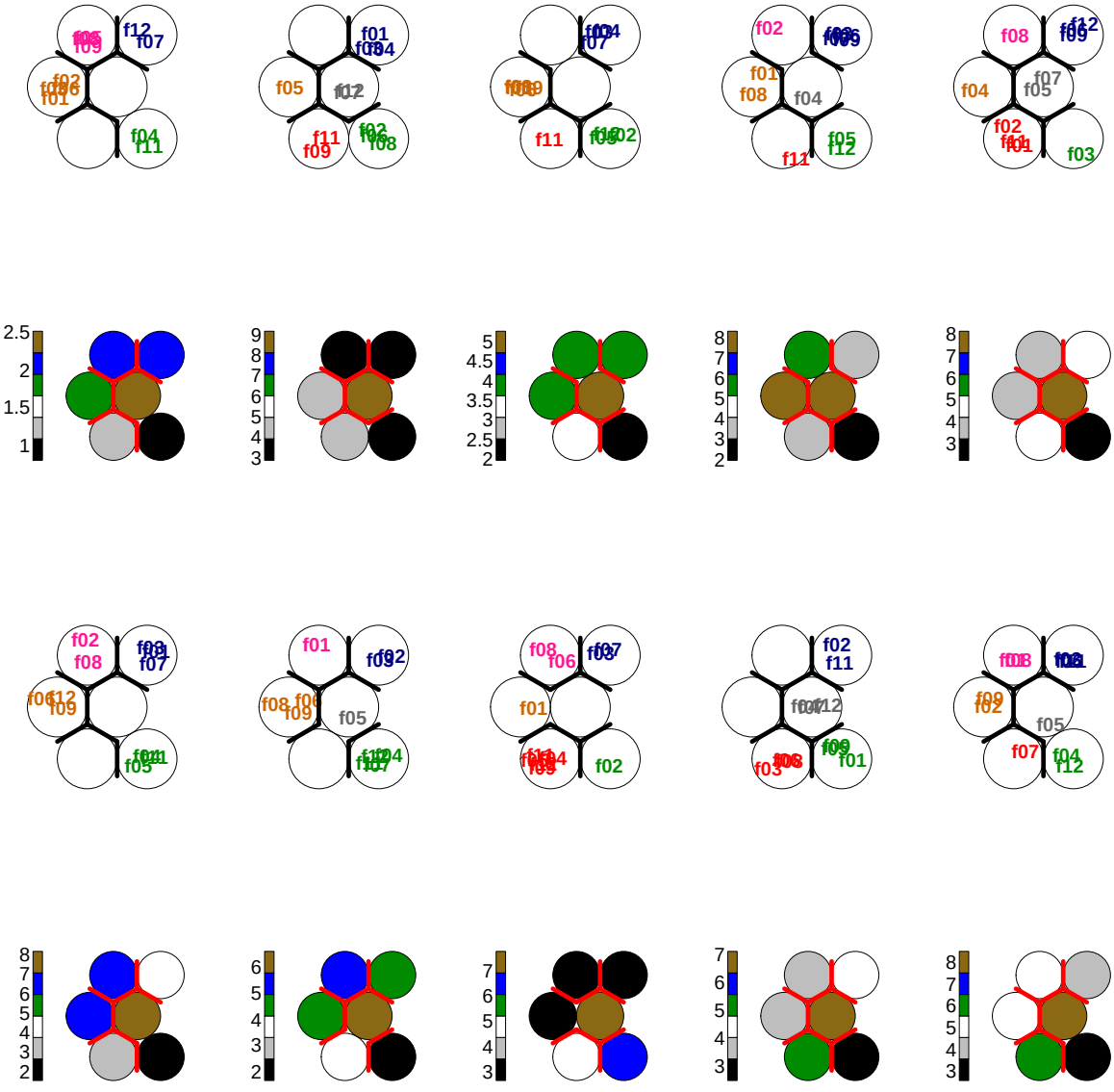


Figura 8.15: SOM - Mapping & Distance Neighbours - (H7)

Através dos gráficos do SOM, fica mais fácil de visualizar quais indivíduos fazem parte de quais grupos. Desta forma, esses gráficos podem auxiliar na análise de fadiga relacionada diretamente ao indivíduo e, continua fornecendo informações acerca do agrupamento. Assim, com esta opção gráfica, fica mais fácil de visualizar a relação as informações do agrupamento ao indivíduo.

### 8.3.3 ARI

Analisando-se o comportamento dos indivíduos no processo de agrupamento, para ambas as técnicas utilizadas, percebeu-se que indivíduos posicionados em áreas com características opostas foram classificados no mesmo *cluster*. Desta forma, com base na representação esquemática do JASA, indivíduos com comportamentos musculares distintos estão sendo classificados em grupos iguais. Visando atenuar esta característica do agrupamento, decidiu-se fornecer pesos às posições do indivíduo na hora de acordo com o que foi proposto no JASA de modo a 'forçar' que indivíduos em áreas diferentes não sejam classificados no mesmo grupo. Para avaliar a qualidade do agrupamento, resolveu-se verificar através do ARI o comportamento dos grupos para cada decil da hora e, posteriormente, para toda a hora. Deseja-se verificar se o comportamento dos indivíduos nos grupos muda de acordo com o método de agrupamento e se a adoção dos pesos ajuda na identificação de um padrão de comportamento dos indivíduos.

O ARI (índice de Rand ajustado), proposto por (Hubert e Arabie, 1985), veio para corrigir um problema existente no valor esperado do índice de Rand (o valor esperado do índice de Rand não era constante).

O índice ajustado de Rand assume a distribuição hipergeométrica generalizada como o modelo de aleatoriedade, ou seja, as partições são escolhidos ao acaso de forma que o número de objetos nas classes e grupos sejam fixos.

O Índice Ajustado de Rand (ARI) é, frequentemente, utilizado em validação de *cluster*, uma vez que é uma medida de concordância entre duas partições:

- uma dada pelo processo de agrupamento;
- outras definidas por critérios externos.

Para o agrupamento, considerou-se como variáveis a amplitude e a frequência transformadas, isto é, pertencentes ao intervalo $[-1, 1]$ e a posição do indivíduo exem-

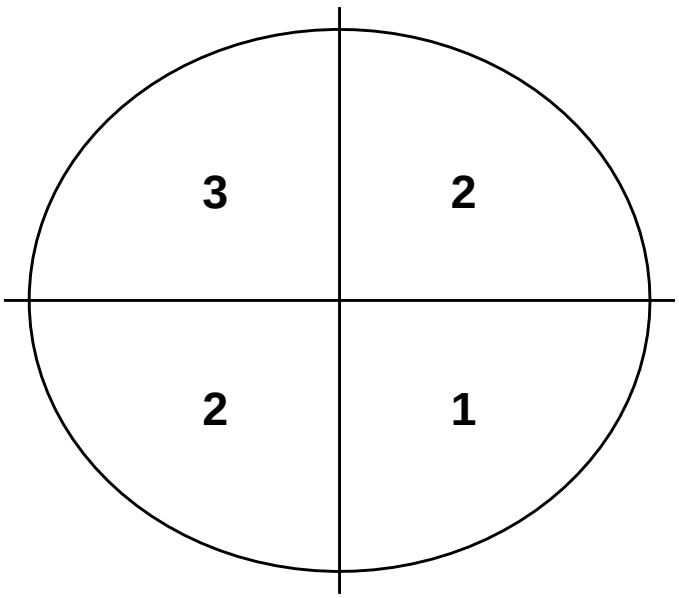



Figura 8.16: Pesos fornecidos aos indivíduos

plificada pela figura 8.16. Os valores fornecidos às posições correspondem levaram em consideração a distância das demais áreas em relação à área de fadiga imposta pela JASA. Perceba que o quadrante mais afastado recebeu valor 3, enquanto, que os outros por serem considerados mais próximos receberam valor 2.

O ARI correspondente à evolução do agrupamento para cada decil da hora (para no máximo 6 grupos) são fornecidos pelas figuras 8.17 e 8.18. O fato de os valores correspondentes ao ARI mudaram a cada decil da hora indica que tanto para o k-*medoids*, quanto para o SOM, os indivíduos vão mudando de grupo. Ou seja, os indivíduos ,de maneira geral, vão mudando de grupo a cada instante do teste o que indica falta de um padrão de comportamento destes indivíduos. Perceba que, caso os indivíduos apresentassem um padrão de comportamento, o ARI para os mesmo permaneceria constante, inalterado.

O ARI correspondente à evolução dos indivíduos considerando toda a hora, sem distinguir os decis, para as mesmas variáveis anteriores também passa a informação de falta de padrão no comportamento dos indivíduos agrupados nas horas.

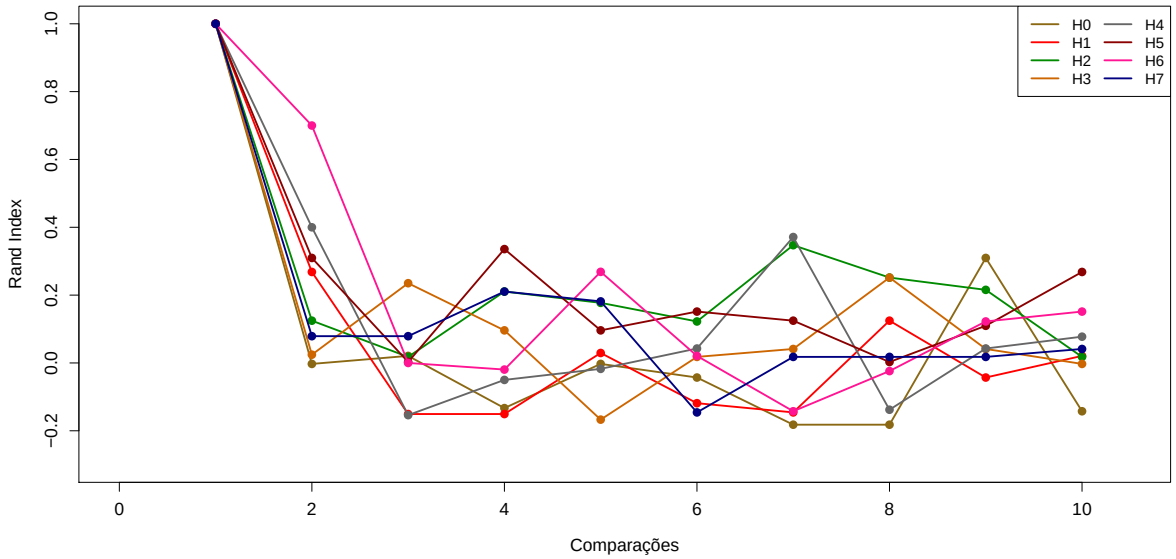



Figura 8.17: ARI - Evolução nos decis das horas - K-*Medoids*

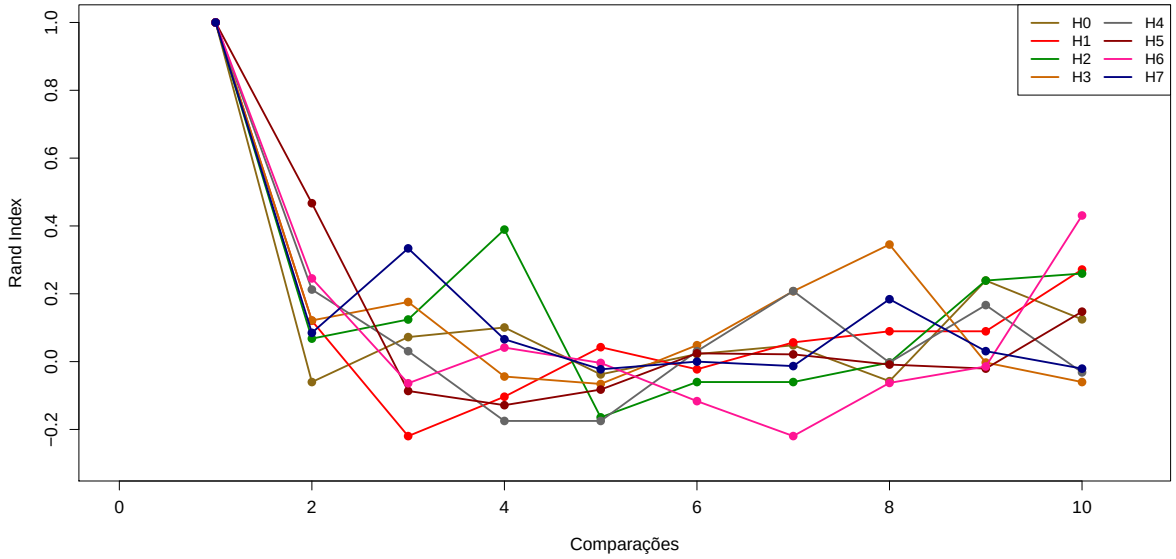



Figura 8.18: ARI - Evolução nos decis das horas - SOM

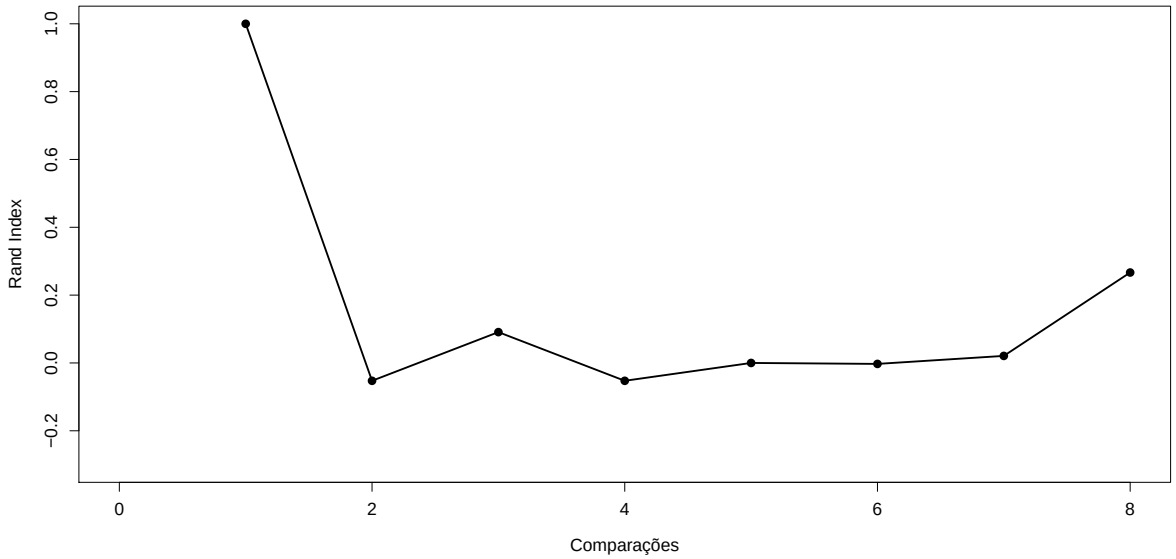



Figura 8.19: ARI - Evolução nas horas - K-*Medoids*

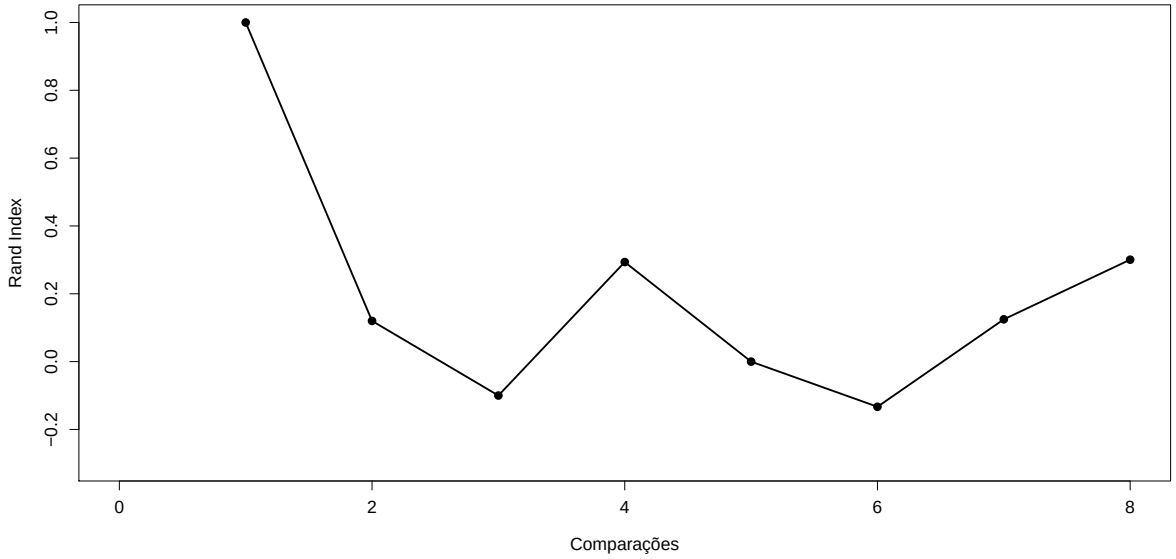



Figura 8.20: ARI - Evolução nas horas - SOM

Como última recurso de análise, desejou-se verificar o posicionamento de cada indivíduo no decorrer da hora. Para isso criou-se a figura 8.21. Para cada indivíduo analisado tem-se o lugar em que ele se encontrava durante cada decil da hora. Cada lugar foi baseado no JASA utilizando-se o esquema adotado na figura 8.16.

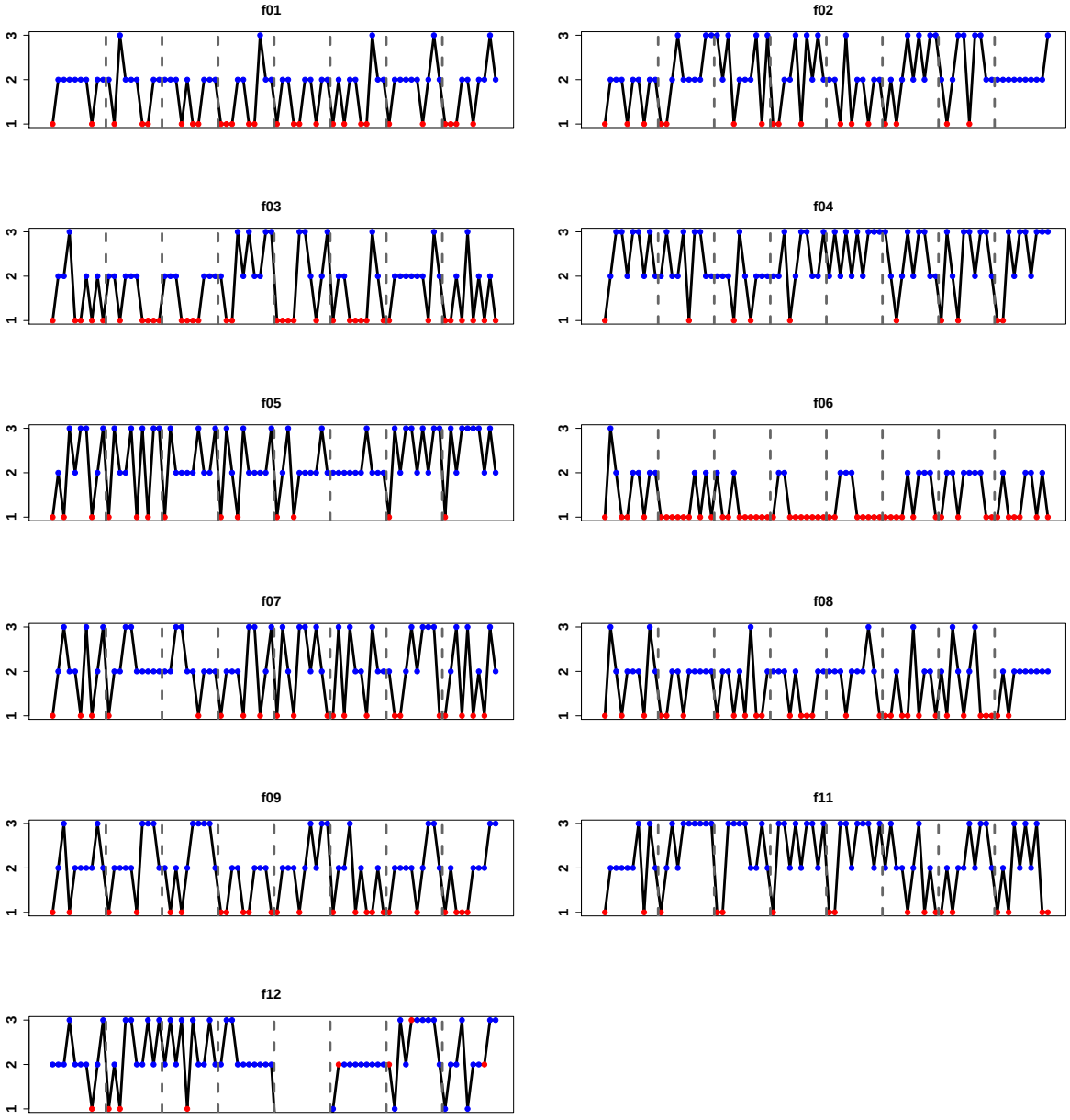



Figura 8.21: Posição no decorrer das 8 horas

Cada ponto vermelho, indica que o indivíduo estava na área 1 que segundo o JASA corresponde a dizer que o indivíduo está em fadiga, as demais área indicam que o indivíduo está sob 'descanso'. Cada traço cinza delimita uma hora de análise ou 10 decis. Ressalta-se que os gráficos não devem ser visualizados como uma série temporal. Os picos não significam nada além do que uma área de conforto para o músculo do indivíduo analisado.

Pela figura 8.21 destacam-se os indivíduos $f04$ e $f06$. Aparentemente, a mulher $f04$ passa a maior parte do teste sem fadiga muscular. Perceba a baixa quantidade de pontos vermelhos existentes no gráfico correspondente a ela. Já a mulher $f06$ aparenta ter um comportamento exatamente oposto ao da mulher $f04$. A grande quantidade de pontos vermelhos indicam que possivelmente esses indivíduo esteve sob fadiga durante grande parte do teste.

# 9 Conclusão

Este trabalho é apenas o início de um estudo que ainda tem muito a evoluir. Os dados fornecidos pelo $BIM_L$ foram analisados pela primeira vez neste projeto após passarem por uma refiltragem para a retirada de ruídos e outros problemas que afetavam a qualidade do sinal.

Durante toda a execução do trabalho, tinha-se em mente, sempre, analisar os dados com base em toda a teoria e esquematização do JASA. No transcorrer das análises ficou claro o quão difícil é o trabalho estatístico de análise de dados, principalmente, de dados fornecidos por terceiros já que não se sabe de maneira detalhada como se deu a coleta dos dados. Os dados de eletromiografia necessitam de extremo cuidado na sua coleta e manuseio devido ao fato de que qualquer problema, por menor que seja, pode afetar a qualidade do sinal o que, posteriormente, afetará a análise e interpretação.1.1

No decorrer da análise, percebe-se que não foi possível identificar um padrão de fadiga através do JASA e alguns fatores podem surgir para explicar esse problema de diagnóstico. Um primeiro problema pode ser relacionado ao início do procedimento, durante a avaliação da estrutura do banco de dados. A grande quantidade de dados exigiu que as informações contidas no banco de dados fossem resumidas. Um problema que pode ter surgido, durante essa redução do banco através dos decis, é que a informação através do tempo tenha sido reduzida em demasia. Pela figura 8.6 é possível verificar um comportamento interessante dos grupos para a hora zero (inicial) mas não para as demais. Sabe-se que a hora zero possui cerca de $12,5\%$ do total de pontos das outras horas. O que pode ter ocorrido é que a aplicação do decil para as demais horas não tenha sido significativa, no sentido em que, características distintas no comportamento do músculo tenham sido alocadas no mesmo decil. Assim, uma característica que, através da análise do banco de dados por completo seria percebida não o foi.

Um segundo problema pode estar relacionado à escolha dos dados para análise. Segundo o professor Ricardo Freitas von Borries, da UTEP, pode ter ocorrido que a escolha do músculo *splenius capitis*, lado esquerdo, peso B para as mulheres tenha sido uma escolha infeliz. Infeliz no sentido de que talvez esse músculo tenha como característica ser pouco sobrecarregado, independentemente da posição dos pesos nos testes. Um terceiro problema que, *a priori* parece ser o menos provável, está relacionado à técnica utilizada para análise (k-*medoids*, *kmeans*, SOM e até o próprio JASA). Como última possibilidade, o problema pode estar relacionado ao próprio banco de dados em si e a qualidade da informação captada.

Porém, nem só pontos negativos devem ser enchergados. Como ponto positivo tem-se o ganho na visualização fornecido pelas técnicas gráficas exploradas. No início da análise era muito complicado verificar qualquer padrão nos dados já que havia muita informação e não se sabia, exatamente, o que fazer com ela. Com o passar do tempo e com a execução de tudo aquilo que foi proposto, a visualização dos resultados foi melhorando ao ponto de, mesmo não possuindo uma tecnologia como o *cybershare*, ser capaz de visualizar múltiplos gráficos com informações relevantes que servem como ponto de partida para trabalhos futuros.

Como sugestão para estudos futuros, sugere-se mudar a abordagem em cima do banco de dados durante o preparo da informação para análise. Aumentar a quantidade de grupos trabalhando com percentis ao invés dos decis talvez seja uma forma de alcançar isso. Posterior a isso, deve-se realizar a mesma análise e verificar se algo de diferente ocorre.

Outra sugestão seria testar a validade do JASA. Utilizar dados de indivíduos em que há uma informação *a priori* da existência de fadiga muscular e realizar o estudo do JASA para os mesmos e verificar se a posição que o indivíduo aparece é realmente aquela preconizada pelo JASA. Caso contrário, pode-se estar tentando utilizar uma técnica que não é assim tão eficiente como esperado.

Uma última sugestão seria o fato de o próprio laboratório da UTEP realizar a captação dos dados. Perceba que a participação da UTEP limita-se apenas à filtragem de ruídos e melhora da qualidade do sinal de EMG. Não se tem a informação detalhada de como foi realizado todo o procedimento de captação dos sinais. O fato é que essa

captação pode estar sendo realizada de maneira incorreta o que pode estar afetando a qualidade do sinal, mesmo que a quantidade de ruídos seja mínima.

# Referências Bibliográficas


Cifrek, M., Medved, V., Tonkovic, S., and Ostojic, S. (2009). Surface emg based muscle fatigue evaluation in biomechanics. *Clinical Biomechanics*, 24(4):327–340.

Gan, G., Ma, C., and Wu, J. (2007). *Data Clustering: Theory, Algorithms, and Applications*. ASA-SIAM Series on Statistics and Applied Probability. SIAM, Philadelphia, ASA, Alexandria, VA.

Guyton, A. C. and Hall, J. E. (2006). *Textbook of Medical Physiology*. Elsevier Saunders, eleventh edition.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Maning, Inference and Prediction*. Springer Series in Statistics. Springer-Verlag.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley-Interscience. John Wiley & Sons, Hoboken, New Jersey.

Kohonen, T. (2001). *Self-Organising Maps*. Springer Series in Information Sciences. Springer-Verlag, third edition.

Konrad, P. (2005). The abc of emg - a practical introduction to kinesiological electromyography.

Luttmann, A., Jager, M., and Laurig, W. (2000). Electromyographical indication of muscular fatigue in occupational field studies. *International Journal of Industrial Ergonomics*, 25(6):645–660.

Mingoti, S. A. (2005). *Análise de Dados Através de Métodos de Estatística Multivariada: Uma Abordagem Aplicada*. UFMG, Belo Horizonte.

Moshou, D., Hostens, I., Papaioannou, G., and Ramon, H. (2005). Dynamic muscle fatigue detection using self-organizing maps. *Applied Soft Computing*, 5(4):391–398.

Theodoridis, S. and Koutroumbas, K. (2009). *Pattern Recognition*. Academic Press, fourth edition.

Wehrens, R. and Buydens, L. M. C. (2007). Self and super-organizing maps in r: The kohonen package. *Journal of Statistical Software*, 21(5).

# Apêndice

## A - Programação em R - Simulação SOM - Seção 7.2

```
## Simulação criada para aprender a ler o gráfico gerado pelo SOM
## conforme as variáveis são deslocadas de posição

# serão geradas variáveis para posterior deslocamento, afim de observar
# as mudanças geradas pelas mesmas no gráfico do SOM
# Comparar-se-ão, também, os algoritmos de agrupamento KMEANS E MEDOIDS(PAM)

# Para rodar a simulação serão necessários estes pacotes.
# Caso não os tenha instalar no R via install.packages()

require(kohonen)
require(RColorBrewer)
require(cluster)
require(mclust)

# gerando dados aleatórios no intervalo [-1 1];
# variáveis deslocadas para o ponto (0,0)

xt <- c(rep(-1,100),rep(1,100),rep(-1,100),rep(1,100),rep(0,100),rep(0,100))
yt <- c(rep(-1,200),rep(1,200),rep(0,100),rep(0,100))

# criando os pares ordenados (xt,yt)
grupo.0 <- as.matrix(cbind(xt,yt))
```

```
## Variável de classificação

zt <- as.factor(c(rep("QIII",100),rep("QIV",100),rep("QII",100),rep("QI",100),
rep("Móveis",100),rep("Móveis",100)))

########################################
## Movimentando em relação ao 1º quadrante
########################################

# gerando dados aleatórios no intervalo [-1 1];
# variáveis deslocadas para o ponto  (1,1)

xt <- c(rep(-1,100),rep(1,100),rep(-1,100),rep(1,100),rep(1,100),rep(1,100))
yt <- c(rep(-1,200),rep(1,200),rep(1,100),rep(1,100))

# criando os pares ordenados (xt,yt) identificados atraves de zt

grupo.1 <- as.matrix(cbind(xt,yt))


########################################
## Movimentando em relação ao 2º quadrante
########################################

#  variáveis deslocadas para o ponto  (-1,1)

xt <- c(rep(-1,100),rep(1,100),rep(-1,100),rep(1,100),rep(-1,100),rep(-1,100))
yt <- c(rep(-1,200),rep(1,200),rep(1,100),rep(1,100))

# criando os pares ordenados (xt,yt)
grupo.2 <- as.matrix(cbind(xt,yt))


########################################
```

```
## Movimentando em relação ao 3º quadrante
########################################

#  variáveis deslocadas para o ponto para o ponto (-1,-1)

xt <- c(rep(-1,100),rep(1,100),rep(-1,100),rep(1,100),rep(-1,100),rep(-1,100))
yt <- c(rep(-1,200),rep(1,200),rep(-1,100),rep(-1,100))

# criando os pares ordenados (xt,yt)
grupo.3 <- as.matrix(cbind(xt,yt))



########################################
## Movimentando em relação ao 4º quadrante
########################################

#  variáveis deslocadas para o ponto para o ponto (1,-1)

xt <- c(rep(-1,100),rep(1,100),rep(-1,100),rep(1,100),rep(1,100),rep(1,100))
yt <- c(rep(-1,200),rep(1,200),rep(-1,100),rep(-1,100))

# criando os pares ordenados (xt,yt)
grupo.4 <- as.matrix(cbind(xt,yt))



############################################
# rodando o SOM juntamente com o agrupamento.
############################################

# escolha da semente: o SOM faz uma escolha aletória para a amostra dos dados
# que iniciarão o algoritmo.

# semente
```

```
set.seed(28670)

## rodando o SOM (grid = 8 X 8)

simulation.som.0 = som(grupo.0, grid = somgrid(8,8,"hexagonal"))
simulation.som.1 = som(grupo.1, grid = somgrid(8,8,"hexagonal"))
simulation.som.2 = som(grupo.2, grid = somgrid(8,8,"hexagonal"))
simulation.som.3 = som(grupo.3, grid = somgrid(8,8,"hexagonal"))
simulation.som.4 = som(grupo.4, grid = somgrid(8,8,"hexagonal"))

## plot das distâncias usando o método MEDOIDS

## cores que serão alocadas a cada valor de classificação nos quadrantes

colors = col =c("black","red","green4","goldenrod4","blue")
classes = as.integer((zt))

## Gradiente de cor
## pesquisando cores no R

colors()[grep('black',colors())]

## Alguns gradientes (devem ser utilizados na opção palette.name em plot)

# colorido

jet.colors = colorRampPalette(c("#00007F", "blue", "#007FFF", "cyan",
             "#7FFF7F", "yellow", "#FF7F00", "red", "#7F0000"))

# monocromatico preto, cinza, branco

mono = colorRampPalette(c("black","gray","white"))

## o SOM possui problemas de resolução. Desta forma, rode esses códigos para os
```

```
## plots com a janela maximizada. Caso contrário, os gráficos aparecerão
## numa dimensão de difícil visualização.

## Plots: distance neighbours e mapping
## add.cluster.boundaries adicionará a um gráfico já existente linhas,
## que permitirao a visualizar quais unidades deverao ser agrupadas juntas

## abre janela grafica no R
windows()

## opcao grafica que permite alocar graficos matricialmente.
## Aqui pede-se para alocar graficos em 1 linha e 2 colunas.
par(mfrow = c(1,2))

## plot dist.neighbours utilizando informacoes dos SOMs gerados anteriormente

plot(simulation.som.0, type='dist.neighbours',
main = "SOM neighbour distances (MEDOIDS)",palette.name=mono)

## rodando medoids nos codes do SOM

som.medoids = pam(simulation.som.0$codes,k=6)
add.cluster.boundaries(simulation.som.0, som.medoids$clustering,col='black')

## plot mapping

plot(simulation.som.0, type = "mapping",pch = 19, col = colors[classes])
add.cluster.boundaries(simulation.som.0, som.medoids$clustering,col='black')
legend("bottom",legend=as.factor(levels(zt)),ncol=5,
col=colors[as.factor(levels(zt))],pch=19,cex=1.2)


# Deslocamento em direção ao 1º quadrante (1,1)
```

```
windows()
par(mfrow = c(1,2))

## plot dist.neighbours

plot(simulation.som.1, type='dist.neighbours',
main = "SOM neighbour distances (MEDOIDS)",palette.name=mono)

## rodando medoids nos codes do SOM

som.medoids = pam(simulation.som.1$codes,k=6)
add.cluster.boundaries(simulation.som.1, som.medoids$cluster,col='black')

## plot mapping

plot(simulation.som.1, type = "mapping",pch = 19, col =colors[classes])
add.cluster.boundaries(simulation.som.1, som.medoids$cluster,col='black')
legend("bottom",legend=as.factor(levels(zt)),ncol=5,
col=colors[as.factor(levels(zt))],pch=19,cex=1.2)


# Deslocamento em direção ao 2º quadrante (-1,1)

windows()
par(mfrow = c(1,2))

## plot dist.neighbours

plot(simulation.som.2, type='dist.neighbours',
main = "SOM neighbour distances (MEDOIDS)",palette.name=mono)

## rodando medoids nos codes do SOM

som.medoids = pam(simulation.som.2$codes,k=6)
```

```
add.cluster.boundaries(simulation.som.2, som.medoids$cluster,col='black')

## plot mapping

plot(simulation.som.2, type = "mapping",pch = 19, col =colors[classes])
add.cluster.boundaries(simulation.som.2, som.medoids$cluster,col='black')
legend("bottom",legend=as.factor(levels(zt)),ncol=5,
col=colors[as.factor(levels(zt))],pch=19,cex=1.2)


# Deslocamento em direção ao 3º quadrante (-1,-1)

windows()
par(mfrow = c(1,2))

## plot dist.neighbours

plot(simulation.som.3, type='dist.neighbours',
main = "SOM neighbour distances (MEDOIDS)",palette.name=mono)

## rodando medoids nos codes do SOM

som.medoids = pam(simulation.som.3$codes,k=6)
add.cluster.boundaries(simulation.som.3, som.medoids$cluster,col='black')

## plot mapping

plot(simulation.som.3, type = "mapping",pch = 19, col =colors[classes])
add.cluster.boundaries(simulation.som.3, som.medoids$cluster,col='black')
legend("bottom",legend=as.factor(levels(zt)),ncol=5,
col=colors[as.factor(levels(zt))],pch=19,cex=1.2)


# Deslocamento em direção ao 4º quadrante (1,-1)
```

```
windows()
par(mfrow = c(1,2))

## plot dist.neighbours

plot(simulation.som.4, type='dist.neighbours',
main = "SOM neighbour distances (MEDOIDS)",palette.name=mono)

## rodando medoids nos codes do SOM

som.medoids = pam(simulation.som.4$codes,k=6)
add.cluster.boundaries(simulation.som.4, som.medoids$cluster,col='black')

## plot mapping

plot(simulation.som.4, type = "mapping",pch = 19, col =colors[classes])
add.cluster.boundaries(simulation.som.4, som.medoids$cluster,col='black')
legend("bottom",legend=as.factor(levels(zt)),ncol=5,
col=colors[as.factor(levels(zt))],pch=19,cex=1.2)

## Vale salientar que outros métodos de agrupamento podem ser utilizados
## para gerar os boundaries (kmeans,métodos hierárquicos)
```