Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Análise de Agrupamento em Impressões Digitais

Ana Luiza Duarte Araujo
Matheus Fideles Souza de Carvalho

Brasília

2018

Ana Luiza Duarte Araujo **13/0101231**

Matheus F. S. de Carvalho **14/0028439**

Análise de Agrupamento em Impressões Digitais

Relatório apresentado à disciplina Trabalho de Conclusão de Curso II da 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. Dr. Leandro Tavares Correia

Brasília

2018

Agradecimentos

Em primeiro lugar, a Deus por ter nos dado força, saúde e paciência para superar todas as dificuldades. Às nossas famílias, de maneira especial aos nossos pais, pelo incentivo, amor e apoio nos momentos mais complexos. Ao corpo docente que sempre mostrou-se prestativo para nos preparar como profissionais.

Agradecemos também aos colegas da faculdade que em muitos momentos nos ajudou nas inúmeras dúvidas que surgiram, obrigada pelas trocas de conhecimento e pelo apoio mútuo gerado nos momento difíceis da graduação, e que serviram para garantir um melhor resultado final do projeto.

Ao nosso orientador Professor Doutor Leandro Tavares Correia, pela paciência, suporte e confiança transmitidas no decorrer deste projeto. A todos fizeram parte da nossa formação, o nosso muito obrigado.

"Se a boca se cala, falam as pontas dos dedos"

Sigmund Freud

Resumo

As impressões digitais tem um grande papel na identificação de indivíduos. Porém a partir de uma análise química nas digitais é possível permitir identificar determinados agentes contidos (detritos) em uma impressão digital. Para este estudo foram selecionados 20 indivíduos para identificar, usando a técnica de Espectrometria de Massa, os íons presentes em suas impressões digitais e a intensidade destes íons. As digitais foram analisadas quimicamente com o material nanoestruturado (SALDI), o tradicional (MALDI) e também sem nenhum material (LDI). Além disso, foi aplicado um questionário observando aspectos como: o uso de determinados produtos nas últimas 24h, se haviam tomado banho, se fumavam e etc. O principal objetivo foi verificar semelhanças em grupos de indivíduos formados através das Análises de Agrupamentos que se dividiu em: 1.Considerando a intensidade dos íons: Usando o Método Não-Hierárquico K-médias e o Coeficiente Cofenético, que validou as simulações de dendogramas realizadas pelos métodos Hierárquicos. 2.Análise considerando apenas presença e ausência dos íons. (Usando as medidas: Jaccad, Sørensen, Ochiai, Anderberg e Kulezynski II). Percebemos, a partir dos resultados obtidos, que o material MALDI formou grupos que melhor representaram as respostas obtidas no questionário, principalmente nos aspectos que se referem à banho, uso de cosméticos e produtos de limpeza. Levando em consideração que hoje no Brasil são raras as pesquisas estatísticas relacionadas a impressões digitais, infere-se que este trabalho pode ser utilizado futuramente para outros tipos de análises nessa área de identificação.

Palavras-chaves: *Cluster Analysis*, Análise de Agrupamento, Impressões Digitais.

Lista de Tabelas

Lista de Figuras

Sumário

Capítulo 1

Introdução

Os métodos de identificação humana sempre foram objeto de estudos e vêm evoluindo ao decorrer dos séculos. Desde 2000 a.C já eram usadas impressões digitais em barro como uma forma de prevenção de falsificações e identificação de documentos. A partir de 1903 o Brasil adotou a impressão digital como método de identificação de indivíduos na área criminal e desde então, com base em avanços científicos e novas tecnologias surgiram vários estudos que tinham como maior objetivo padronizar e analisar as impressões digitais(Chemello, 2006). A datiloscopia é o estudo das padronizações presentes nas digitais, em especial aquelas presentes nas pontas dos dedos. Apoiando-se neste estudo é possível definir, com destacável precisão, a identidade de digitais presentes na superfície de objetos, por exemplo. Dessa forma, torna-se possível identificar o agente que realizou determinada ação em um meio que contenha vestígios de sua impressão (Chemello, 2006).

Os profissionais em datiloscopia são chamados de papiloscopistas e têm o dever de avaliar objetos em um determinado local que podem revelar vestígios em sua superfície. Os vestígios são chamados de impressões papilares latentes e são divididos em dois grupos: impressões visíveis e impressões ocultas. O primeiro objetivo do trabalho de um papiloscopista é, justamente, tornar visível uma impressão latente oculta. Várias são as técnicas utilizadas para melhor visualização dos vestígios papilares ocultos, dentre elas podemos citar: técnica do pó, vapor de iodo, nitrato de prata, nidrina (Chemello, 2006).

Uma vez que as impressões latentes deixam de ser ocultas e ficam visíveis, torna-se possível fazer análise química das impressões latentes, que são aquelas que consistem em uma mistura de substâncias originárias da pele com outros vestígios (Girod et al., 2012). Existem vários métodos empregados para análise desses vestígios, são

eles: cromatografia líquida de alta eficiência(HPLC) (Dikshitulu et al., 1986), cromatografia gasosa acoplada a espectrometria de massa(GC-MS) (Archer et al., 2005), espectrometria de massa por ionização química à pressão atmosférica (Mountfort et al., 2007), espectrometria de infravermelho com transformada de Fourier(FT-IR) (Mountfort et al., 2007) e espectrometria de massa por ionização a laser assistida por matriz (MALDI-MS) (Wolstenholme et al., 2009).

Quando a análise das impressões é utilizada para a área criminalística o seu papel vai além da indentificação de indivíduos. Na verdade, esse tipo de análise pode verificar de maneira mais profunda a pele do indivíduo podendo encontrar componentes endógenos e exógenos. Os componentes endógenos são aqueles que estão presentes no interior da pele, alguns deles são: aminoácidos, proteínas, ácidos graxos e colesterol. Os componentes exógenos são aqueles que estão na parte exterior a pele, que vêm de fora e passam a constituir as impressões digitais em determinadas situações, estes podem ser vestígios de: nicotina, drogas, explosivos e lubrificantes de preservativos, por exemplo (Kaplan-Sandquist et al., 2014).

A técnica de espectrometria com utilização de uma matriz MALDI-MS permite que íons das impressões sejam transferidos à fase gasosa, tornando possível sua análise (Wanner and Höfner, 2007). Contudo, essa técnica é utilizada em grande parte para análise de biomoléculas grandes de diferentes tipos de analitos, que são os constituintes de interesse da amostra. Nesse caso, seriam polímeros, polisacarídeos, lipídeos e outros. Sendo assim, moléculas orgânicas menores têm maior dificuldade de detecção por conta da alta intensidade dos íons da matriz utilizada. O MALDI-MS é a técnica mais utilizada para realização de análises de impressões latentes. Contudo, foi prosposto uma técnica parecida que visa possibilitar a análise de amostras de analitos [1] que fiquem em menor evidência, ou seja, que são mais difíceis de serem achados. Essa técnica é chamada de espectrometria de massa por ionização/dessorção [2] a laser assistida por superfície (SALDI), ou seja, é um fenômeno onde uma substância é adsorvida ou absorvida por outra possibilitando a análise da amostra dissolvida em uma matriz ou superficie. O material SALDI possui em sua fase sólida um material nanoestruturado, diferentemente da matriz (forma liquida) utilizada em MALDI-MS (Wolstenholme et al., 2009).

[1] Analito: Uma substância ou componente químico, em uma amostra, que é alvo de análise ou tem interesse para uma análise em um ensaio

[2] Dessorção: fenômeno de retirada de substância(s) adsorvida(s) ou absorvida(s) por outra(s)

Alguns estudos apontam vantagens na utilização de nanoestruturas ao invés de matriz (MALDI) como material para realização das análises químicas (Xu et al., 2003). Portanto, através de técnicas estatísticas multivariadas poderemos melhor investigar características em comum nas diferentes amostras e técnicas empregadas para execução da análise. A análise de agrupamento, mais especificamente a análise de *cluster*, que não só investiga, como também por meio de determinadas medida de similaridade agrega observações mais parecidas em um determinado grupo, ou seja, identifica comportamento comuns entre os indivíduos e entre os métodos de tratamento no uso da captação das digitais. Não é muito fácil encontrar na literatura uma análise estatística de agrupamentos em um banco de dados com impressões latentes. Por isso, acredita-se que este tipo de análise pode ajudar a explicar diversas semelhanças entre indivíduos e as vantagens e desvantagens da implementação da nova técnica com materiais nanoestruturados (SALDI-MS) em relação a técnica que é usada com matriz (MALDI-MS), apenas usando as detritos presentes nas digitais.

1.1 Objetivos

O objetivo geral: Verificar semelhanças de indivíduos a partir de análises químicas de uma amostra de impressões latentes por meio do método de agrupamentos.

Os objetivos específicos são:

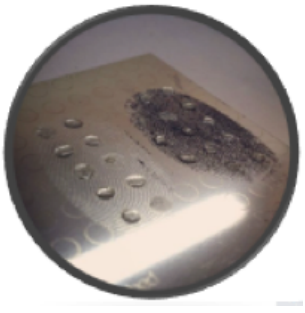
- Compreender os detalhes dos métodos de agrupamento por *clusters*;
- Utilização do *software* estatístico R para realização dos métodos e análises;
- Verificar frequência de íons a partir de análise de agrupamentos;
- Comparar eficácia das 3 técnicas químicas realizadas na amostra;
- Identificar diferentes indivíduos por análise de cluster;
- Identificar relações entre as características dos indivíduos e os agrupamentos;
- Identificar semelhanças entre indivíduos desconsiderando a intensidade dos íons, utilizando somente a presença e ausência dos mesmos;

1.2 Metodologia

A partir de uma amostra de conveniência de 20 indivíduos, 10 homens e 10 mulheres foram coletadas 2 impressões digitais de cada um, em que a primeira foi reservada para ser analisada com a técnica que utiliza nanoestruturas (SALDI-MS), a metade da segunda impressão latente coletada usou-se o método tradicional com alphaciano (MALDI-MS) e a outra metade não foi usada nenhum tipo de material(LDI-MS). Em seguida, as impressões digitais sofreram um procedimento químico onde foi possível captar íons que tornaram-se o objeto de análise.

Figura 1.1: Espectrômetro: Aparelho onde as digitais foram quimicamente tratadas.

Figura 1.2: Lado Esquerdo: Uso do MALDI - Lado Direito: Uso do SALDI



Antes de coletar as impressões implementou-se um questionário coletando informações como: sexo, idade, se o indivíduo era fumante, se fazia uso frequente de algum medicamento e se tomou banho antes do experimento e outras perguntas. É possível encontrar o questionário completo no Apêndice E deste trabalho. Estes dados foram necessários para analisar as semelhanças de cada indivíduo e teve como

principal uso nesse trabalho, a validação dos agrupamentos posteriormente obtidos. Dessa forma, foi possível avaliar as vantagens e desvantagens entre as 03 técnicas realizadas na amostra: LDI-MS, MALDI-MS e SALDI-MS.

Foi realizada então, uma análise descritiva dos dados para explorar melhor as características dos indivíduos coletada no questionário, além disso, foi analisada descritivamente, também os íons obtidos com o objetivo de comparar as 3 técnicas químicas aplicadas.

Capítulo 2

Análise de Cluster

A análise de agrupamentos ou *cluster* é um conjunto de métodos multivariados que têm objetivo principal de associar determinadas variáveis e/ou indivíduos com base em características em comum que possuem. Além disso, essas técnicas constituem ferramentas indispensáveis no processo de agrupamentos, para que se torne possível realizar o objetivo de reduzir os dados, de forma que seja possível identificar características de uma amostra ou população que estejam divididas em grupos e gerar hipóteses previamente estabelecidas de forma que sejam comprovadas as naturezas relativas aos dados.(Hair, 2009).

Vale lembrar, que todos métodos de análise de agrupamentos seguem um caminho que busca definir a estrutura dos dados separando as observações que possuem características parecidas. Este caminho segue os seguintes passos: Formular o problema, selecionar uma medida de distância, escolher um algoritmo de aglomeração, decidir quanto ao número de agrupamentos, interpretar e perfilar os agrupamentos e avaliar a validade do processo de aglomeração. De forma geral, nenhuma técnica é ideal para a realização deste caminho, por isso é imprescindível entender a nuance de cada técnica, a fim de que a escolha dependa do tipo de banco de dados e do problema enfrentado (Malhotra, 2006).

2.1 Similaridade e Distância

A similaridade entre dois objetos nada mais é do que a busca pelo nível de semelhança de características presentes nas observações de um mesmo agrupamento, ou seja, é uma medição empírica de correspondência. O nível de semelhança pode ser medido através da distância entre duas observações (Tryon, 1939), (Allen and

Goldstein, 2013), (Everitt et al., 2011). Distâncias pequenas podem ser interpretadas como semelhanças entre as observações e portanto, podem, em determinado momento, figurar no mesmo agrupamento. Em resumo, as medidas de distância, na verdade, são medidas de dissimilaridade, ou seja, os maiores valores indicam uma maior dissimilaridade entre as variáveis. A partir dos resultados e pelo uso da relação inversa é possível identificar a medida de similaridade. Existem diversos cálculos na literatura matemática para determinar a distância entre dois pontos, essa cálculo faz parte da definição do critério para dizer que duas observações são semelhantes.

De acordo com o objetivo do pesquisador e com as características do banco de dados, existem diferentes formas de lidar com cada situação para a escolha da forma de tratamento dos dados para análise de agrupamento. Levando em consideração o objetivo de verificar a relação entre os indivíduos a partir da intensidade dos íons e também apenas da presença ou não dos mesmos, foram empregadas diversas formas de cálculo para as medidas de dissimilaridade para os dois tipos de análise. Nesse caso, foram divididas em dados binários e dados quantitativos.

2.1.1 Dados Binários

As medidas de similaridade citadas nesta seção, são indicadas para dados binários, ou seja, dados que têm apenas dois valores possíveis, usualmente convertidos como: presença de em determinado atributo definido por 1, e sua ausência definida por 0. Cada medida irá contabilizar a quantidade da presença e ausência de atributos para cada par de indivíduos e gerar, a partir disso, o grau de similaridade, para isso definimos:

Tabela 2.1: Contagem de atributos 0 e 1 para um par de indivíduos

		Indivíduo$_j$	
		1	0
Indivíduo$_i$	1	a	b
	0	c	d

Em que a,b,c e d, são as frequências de ausências/presenças de 0 ou 1 para ambos os indivíduos analisados. O número de coincidências é representado por: $(a + d)$ e o número de diferenças é representado por: $(b + c)$ (Dunn and Everitt, 1980).

Dos coeficientes de similaridades que serão citados, podemos dividi-los em duas categorias: (1) Aqueles que assumem ausência conjunta de atributos, e (2) aqueles que não consideram a ausência conjunta (Carlini-Garcia, 1998). Podendo também ser representado por considerar ou não zeros duplos (Legendre and Legendre, 1983).

Essa separação na prática pode restringir a aplicação dessas medidas para alguns tipos de problemáticas, por exemplo, para alguns casos não temos interesse em analisar a quantidade de dois zeros pois podem significar falta de informação, e por isso escolhe-se uma medida de similaridade que deve excluir os zeros duplos (Gan et al., 2007).

Para as medidas do tipo (2) podemos citar:

- Jaccard:

$$s_{(X_i,X_j)} = \frac{a}{a+b+c} \qquad 0 \leq s_{(X_i,X_j)} \leq 1 \tag{2.1}$$

  O coeficiente de Jaccard (1908) indica máxima semelhança quando os dois objetos possuem valores idênticos, ou seja, quando o número total de diferenças for nulo ($b+c=0$), Jaccard é sensível à direção da codificação. Ou seja, trocar o significado de "1" e "0" geralmente altera o nível de semelhança entre os indivíduos (Romesburg, 2004).

- Sørensen

$$s_{(X_i,X_j)} = \frac{2a}{2a+b+c} \qquad 0 \leq s_{(X_i,X_j)} \leq 1 \tag{2.2}$$

  Esse coeficiente (1945) pesa duplamente a quantidade de presença conjunta do atributo (a). Este também é chamado de coeficiente de Czekanowski ou coeficiente de Dice, isso pode ser uma consequência de sua reinvenção independente (Romesburg, 2004).

- Ochiai

$$s_{(X_i,X_j)} = \frac{a}{\sqrt{(a+b)(a+c)}} \qquad 0 \leq s_{(X_i,X_j)} \leq 1 \tag{2.3}$$

  Esse coeficiente (1957) significa perfeita similaridade quando $s_{(X_i,X_j)} = 1$, e se obter $s_{(X_i,X_j)} = 0$ indica perfeita dissimilaridade, que nesse caso podemos obter quando $a = 0$

- Anderberg

$$s_{(X_i,X_j)} = \frac{a}{a + 2(b + c)} \qquad 0 \leq s_{(X_i,X_j)} \leq 1 \tag{2.4}$$

- Kulezynski I

$$s_{(X_i,X_j)} = \frac{a}{b + c} \qquad 0 \leq s_{(X_i,X_j)} < \infty \tag{2.5}$$

  Índices que tendem ao infinito geralmente são mais sensíveis a pequenas mudanças (Clifford et al., 1975).

- Kulezynski II

$$s_{(X_i,X_j)} = \frac{a}{2}\left(\frac{1}{a + b} + \frac{1}{a + c}\right) \qquad 0 \leq s_{(X_i,X_j)} \leq 1 \tag{2.6}$$

Para aqueles que consideram ausência conjunta (d), ou seja, medidas do tipo (1) temos:

- *Simple Matching*

$$s_{(X_i,X_j)} = \frac{a + d}{a + b + c + d} \qquad 0 \leq s_{(X_i,X_j)} \leq 1 \tag{2.7}$$

  Esse coeficiente (1958) é insensível à direção da codificação, tem um significado intuitivo, pois pela fórmula vemos que é o cálculo da proporção de ausência conjunta de atributos mais a presença conjunta em relação ao resto.

- Rogers e Tanimoto

$$s_{(X_i,X_j)} = \frac{a + d}{a + d + 2(b + c)} \qquad 0 \leq s_{(X_i,X_j)} \leq 1 \tag{2.8}$$

  Nesta medida (1960) a perfeita similaridade é obtida quando $b = c = 0$, obtendo então, $s_{(X_i,X_j)} = 0$. Aqui vemos que há um duplo peso para o número de diferenças totais $(b + c)$.

- Russell e Rao

$$s_{(X_i,X_j)} = \frac{a}{a + b + c + d} \qquad 0 \leq s_{(X_i,X_j)} \leq 1 \tag{2.9}$$

  Esse coeficiente (1940) expressa a proporção de presenças conjuntas para todas as comparações. Quando $d = 0$, então temos o coeficiente de Jaccard.

- Sokal e Sneath

$$s_{(X_i,X_j)} = \frac{2(a+d)}{2(a+d)+b+c} \qquad 0 \leq s_{(X_i,X_j)} \leq 1 \tag{2.10}$$

Nesta medida (1963), a dissimilaridade perfeita ocorre quando $a = d = 0$, tornado $s_{(X_i,X_j)} = 0$. Podemos perceber também, que esse coeficiente é bem paracido com o *simple matching*, pois se dividirmos a equação por 2, tanto no numerador quando no denominador, obtemos $\frac{a+d}{\frac{1}{2}(a+d)+b+c}$, e assim temos que os totais de coincidências $(a+d)$ entre os indivíduos tem metade do peso.

- Hamann

$$s_{(X_i,X_j)} = \frac{(a+d)-(b+c)}{a+b+c+d} \qquad -1 \leq s_{(X_i,X_j)} \leq 1 \tag{2.11}$$

O coeficiente Hamann (1961) relaciona as variáveis por adição, diferente das demais podemos perceber que essa medida varia de -1 a 1.

O coeficiente mais conhecido e citado na literatura é o coeficiente de Jaccard. Além disso é mais simples de entender, pois compara o números de atributos presentes em comum e o total de atributos restantes, porém nem todas as medidas tem uma compreensão fácil. Basicamente a diferença entre as medidas é o peso e a complexidade que é definido para cada quantidade de atributos presentes/ausentes entre os indivíduos.

2.1.2 Dados Quantitativos

- Distância Euclidiana: É uma medida de dissimilaridade que mede a distância entre dois pontos baseado em suas coordenadas no plano cartesiano, medida mais utilizada e exemplificada no calculo da hipotenusa de um triângulo retângulo, ou seja é a distância geométrica entre os indivíduos (Hair, 2009), (Mingoti, 2005), (Souza and Vicini, 2005), para dois elementos seguimos a fórmula:

$$d_{(X_i,X_j)} = \sqrt{\sum_{k=1}^{p}(x_{ik}-x_{jk})^2} \tag{2.12}$$

- Distância Euclidiana Quadrada: É muito parecida com a distância euclidiana normal, contudo é realizada a soma utilizado sem a raiz quadrada. Dessa

forma, o tempo de computação é acelerado. Se adéqua mais facilmente a alguns métodos hierárquicos como Ward e Centróide (Malhotra, 2012).

$$d_{(X_i,X_j)} = \sum_{k=1}^{p}(x_{ik} - x_{jk})^2 \tag{2.13}$$

- Distância city-block (Manhattan): Emprega somas das diferenças absolutas entre as variáveis, comparando a um triângulo retângulo é a medição de seus lados em vez da hipotenusa, como a distância euclidiana sugere (Hair, 2009).

$$d_{(X_i,X_j)} = \sum_{k=1}^{p}|(x_{ik} - x_{jk})| \tag{2.14}$$

- Distância Canberra: É obtida pelo somatório das diferenças entre dois pontos. Essa medida corresponde a versão ponderada da distância city-block ou Manhattan. (Lance and Williams, 1966).

$$d_{(X_i,X_j)} = \frac{\sum_{k=1}^{p}|(x_{ik} - x_{jk})|}{|x_{ik}| + |x_{jk}|} \tag{2.15}$$

- Distância Máxima (Chebychev): A distância é a maior diferença entre todas as variáveis, ou seja:

$$d_{(X_i,X_j)} = \max_{k=1}^{p}|(x_{ik} - x_{jk})| \tag{2.16}$$

- Distância de Minkowski

  Mede a distância entre duas variáveis X_j e X_j usando o parâmetro $\lambda \geq 1$ e pode ser representada por:

$$d_{(X_i,X_j)} = \left(\sum_{k=1}^{p}|x_{ik} - x_{ik}|^{\lambda}\right)^{\frac{1}{\lambda}} \tag{2.17}$$

  - Se $\lambda = 1$ teremos a Distância city-block (Manhattan).
  - Se $\lambda = 2$ teremos a Distância Euclidiana.
  - Se $\lambda \to \infty$ teremos a Distância Máxima.

  Em geral, para valores altos de λ, temos peso relativo maiores de indivíduos muito dissemelhantes dos restantes. Em (Aggarwal et al., 2001) é sugerido maior significância dos resultados para $0 < \lambda < 1$.

- Matriz de Distâncias

  As distâncias são registradas em uma matriz de dimensão n x n, chamada de matriz de similaridade, como mostra a matriz D no exemplo abaixo, na qual d_{ij} representa a distância entre elemento amostral X_i e X_j, podemos observar também, que a matriz é simétrica.

$$D = \begin{bmatrix} 0 & d_{12} & \dots & d_{1n} \\ d_{21} & 0 & \dots & d_{2n} \\ \vdots & & \ddots & \\ d_{p1} & d_{p2} & \dots & 0 \end{bmatrix}$$

Após a escolha da medida de distância é necessário selecionar algoritmo do método para formação dos agrupamentos, esses métodos serão descritos nas sessões abaixo e serão divididos em dois tipos: *Métodos hierárquicos* e *não-hierárquicos*

2.2 Métodos hierárquicos

Os métodos hierárquicos consistem em um conjunto de sucessivos agrupamentos ou divisões de elementos.Os métodos hierárquicos são divididos em duas classes **divisivos** e **aglomerativos** (Doni, 2004).

2.2.1 Divisivo

Os métodos divisivos, em geral é um procedimento onde todas as observações começam em um único agrupamento (raiz), e em seguida são sucessivamente divididas de forma que cada observação se torne um agrupamento unitário (folhas), ou seja,o processo é iterado até que todos os objetos estejam em seu próprio *cluster* (Hair, 2009).

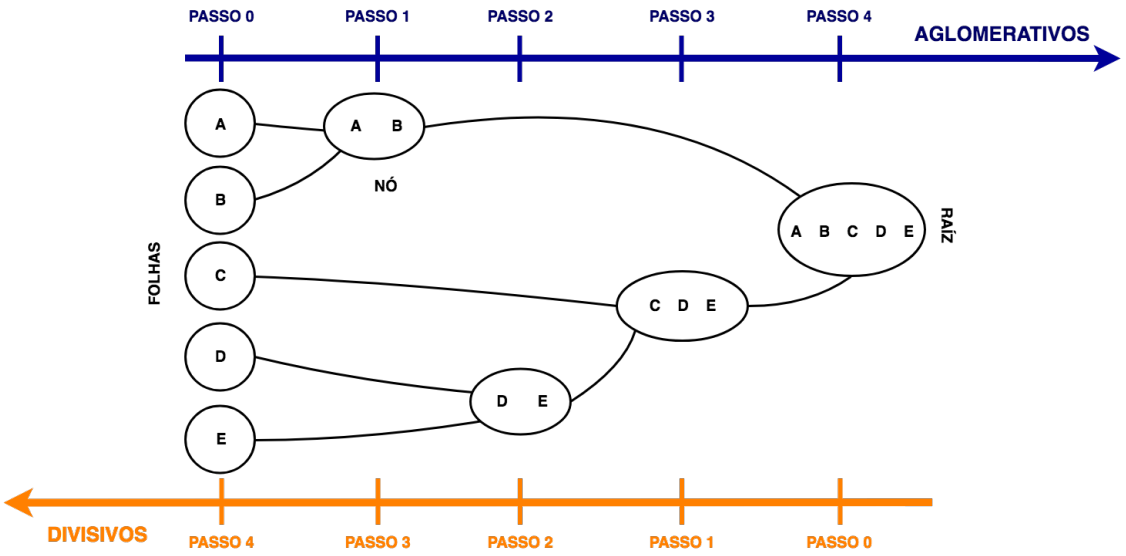
Geralmente na análise de agrupamento, é usado técnicas aglomerativas, esse trabalho também irá seguir essa linha, mesmo sabendo que exitem diferentes formas de dividir e aglomerar, os métodos divisivos, em sua maioria, seguem um padrão inversamente parecido com o dos métodos aglomerativos.

2.2.2 Aglomerativos

Nos métodos aglomerativos cada observação já inicia como um agrupamento unitário. Neste algoritmo cada objeto é considerado inicialmente como um *clus-*

ter de elemento único (folha), ou seja, a quantidade de observações, de início, é igual a quantidade de agrupamentos. A cada etapa do que se passa, os dois *clusters* mais semelhantes são combinados em um novo *cluster* maior, esse procedimento é iterado até que todos os pontos sejam membros de apenas um único grupo (raíz), gerando finalmente, uma árvore de agrupamentos também chamada dendograma (Hair, 2009).

Figura 2.1: Metodologia dos Algoritmos



Fonte: Adaptação da Figura 11 - Cap. 1 de Kaufman e Rousseeuw (2009).

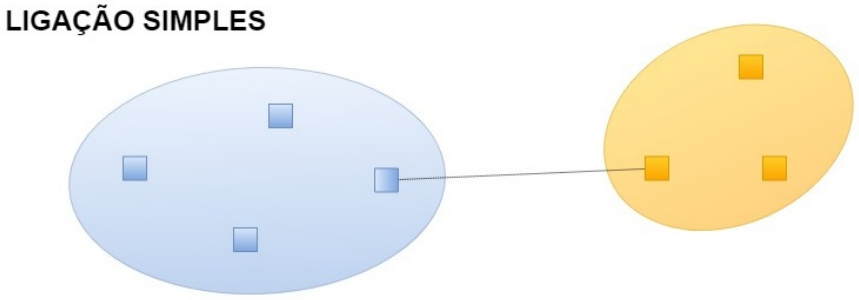
Existem diversas metodologias que têm o objetivo de medir similaridade entre *clusters* ou agrupamentos. Os cinco algoritmos aglomerativos mais utilizados são:

Método de Ligação Simples *(Single linkage)*

Este método, também conhecido como método do vizinho mais próximo, compara as observações e assume como medida de similaridade a menor distância de entre dois pontos, formando então, uma agrupamento a partir dessa definição. Tende a produzir *clusters* longos e "soltos" (Doni, 2004), (Pereira et al., 1993), (Lattin et al., 2011).

$$d_{(AB)X} = min(d_{AX}, d_{BX})$$

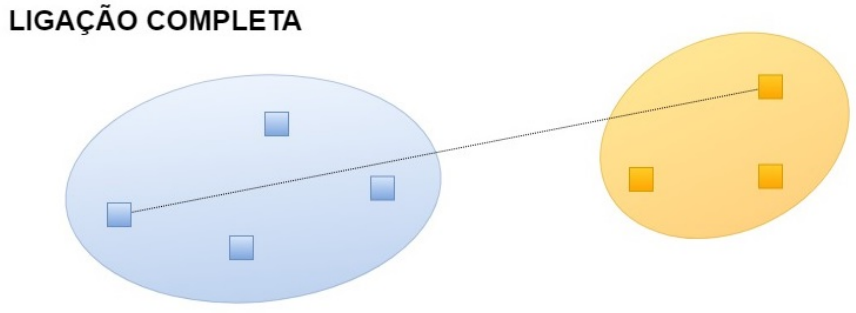
$d_{(AX)} \rightarrow$ distância entre os elementos A e X
$d_{(BX)} \rightarrow$ distância entre os elementos B e X
A, B e $X \rightarrow$ *cluster* da variável quantitativa.



Método de Ligação Completa *(Complete linkage)*

Também conhecido como método do vizinho mais distante. Neste caso, a similaridade é definida a partir da distância máxima entre observações dos agrupamentos analisados. Assume-se que todas as observações de um *cluster* são conectadas umas as outras a partir de uma distância máxima, igualando então a similaridade interna do *cluster* com o diâmetro do grupo. Tende a produzir *cluster* mais compactos (Doni, 2004), (Lattin et al., 2011).

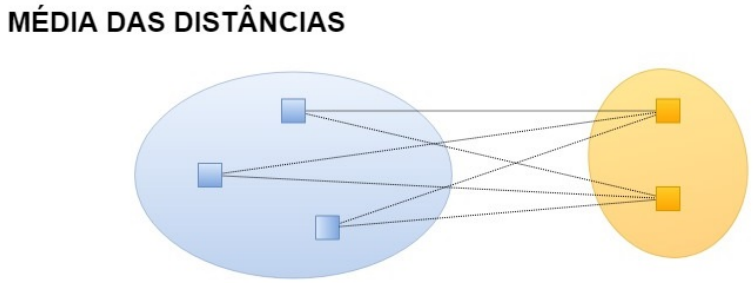
$$d_{(AB)X} = min(d_{AX}, d_{BX})$$



Método de Ligação Média *(Average linkage)*

Nesse caso a similaridade é medida a partir da média de todos os indivíduos presentes nos *clusters*. Dessa forma, essa metodologia não necessita de valores extremos como os dois métodos anteriores (Doni, 2004), (Pereira et al., 1993).

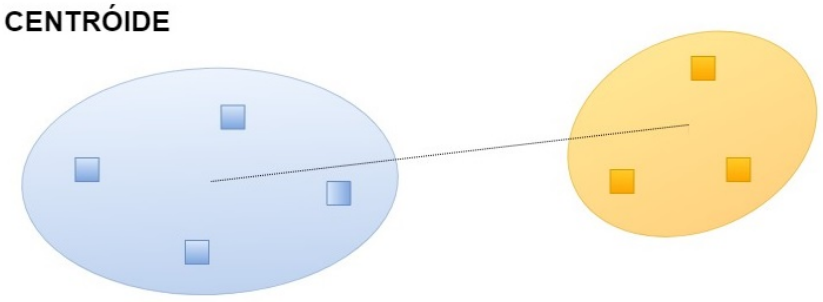
$$d_{(AB)X} = \frac{N_A.d_{AX}+N_B.d_{BX}}{N_A+N_B}$$



Método Centroide *(Centroid method)*

Toda vez que os indivíduos são reunidos em um determinado agrupamento é definido, a partir dos valores médios, um elemento centroides. Este elemento é alterado a cada entrada ou saída de uma nova observação do *cluster*. Portanto, a similaridade, nesse caso é medida a partir da distância entre os centroides dos agrupamentos (Pereira et al., 1993).

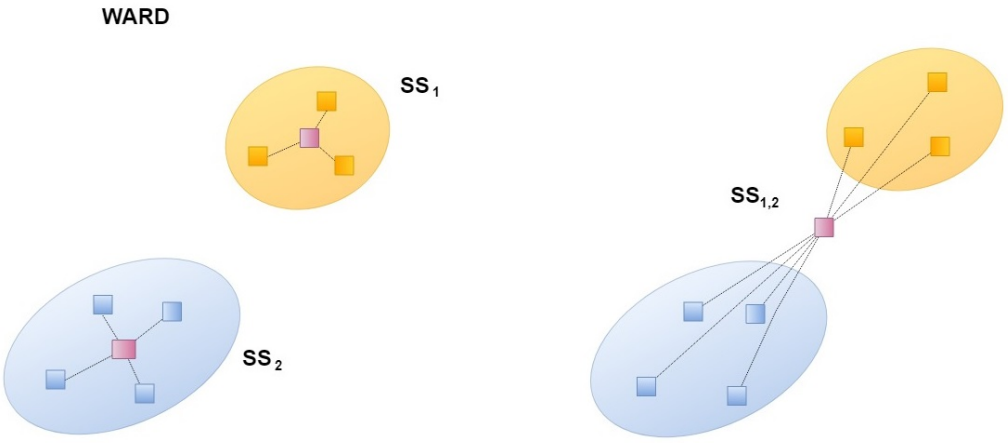
$$d_{(AB)X} = \frac{N_A.d_{AX}+N_B.d_{BX}}{N_A+N_B} - \frac{N_A.N_B.d_{(AB)}}{(N_A+N_B)^2}$$



Método de Ward

Também denominado método da mínima variância. Medida de distância entre dois clusters é a soma das distâncias ao quadrado entre os dois clusters. Em determinados resultados é comum encontrar dois valores de ward. A diferença ward1 e ward2 refere-se ao elevar as distâncias ao quadrado de ward2 antes mesmo de atualizar os agrupamentos (Doni, 2004).

$$d_{(AB)X} = \frac{(N_A+N_X).d_{AX}+(N_B+N_X).d_{BX}-N_X.d_{AB}}{N_A+N_B+N_X}$$



Método de McQuitty

Esse método é considerado por alguns autores como uma espécie de extensão do método de ligação média. É definido quando dois agrupamentos ligam-se entre um novo agrupamento e outro elemento já existente, dessa forma é então calculado como a média das distâncias dos agrupamentos que estão sendo definidos (Pereira et al., 1993), (Carvalho et al., 2009).

$$d_{(AB)X} = \frac{(d_{BX}+d_{AB})}{2}$$

Método de Ligação Mediana *(Median linkage)*

Esse método é considerado é definido quando a distância entre dois agrupamentos distintos é a mediana entre a observação de um agrupamento e a observação de outro. (Marques, 2017)

$$d_{(UV)W} = \frac{(d_{UW}+d_{VW})}{2} - \frac{(d_{UV})}{4}$$

Matriz Cofenética

Em uma matriz de distância dos indivíduos (D), os algoritmos aglomerativos são utilizados para que estes componentes formem grupos, de forma que seja possível a realização de dendogramas. Imagine uma matriz de distâncias D com 4 indivíduos, representados como (a, b, c e d), em que foi utilizado o algoritmo de ligação simples(*(Single linkage)*) para a aglomeração dos indivíduos:

$$D = \begin{array}{c} \\ a \\ b \\ c \\ d \end{array} \begin{array}{c} \begin{array}{cccc} a & b & c & d \end{array} \\ \begin{pmatrix} 0 & 0,85 & 0,67 & 0,25 \\ 0,85 & 0 & 0,55 & 0,98 \\ 0,67 & 0,55 & 0 & 0,77 \\ 0,25 & 0,98 & 0,77 & 0 \end{pmatrix} \end{array}$$

Nesse exemplo os primeiros indivíduos a serem aglomerados seriam o a e d, uma vez que possuem a menor distância $(0,25)$. O processo é refeito realizando a distância entre os indivíduos b e c entre si e também com o novo grupo aglomerado ad. Veja o exemplo:

$$D = \begin{array}{c} \\ ad \\ b \\ c \end{array} \begin{array}{c} \begin{array}{ccc} ad & b & c \end{array} \\ \begin{pmatrix} 0 & 0,37 & 0,94 \\ 0,37 & 0 & 0,55 \\ 0,94 & 0,55 & 0 \end{pmatrix} \end{array}$$

Em seguida, o indivíduo b se juntaria ao grupo ad, uma vez que a menor distância da matriz seria $0,37$. Por fim, sobraria o indivíduo c, e mais uma vez as distâncias seriam calculadas. Veja o passo seguinte do exemplo:

$$D = \begin{array}{c} \\ adb \\ c \end{array} \begin{array}{c} \begin{array}{cc} adb & c \end{array} \\ \begin{pmatrix} 0 & 0,85 \\ 0,85 & 0 \end{pmatrix} \end{array}$$

Dessa forma, como foi mostrado a cada iteração e a cada aglomeração dos grupos, novas distâncias sâo obtidas e ao final do procedimento é obtida uma matriz de distâncias recuperadas que é conhecida também como matriz cofenética. Nesse exemplo como o primeira aglomeraçao (a,d) foi obtida com a distância $0,25$, a segunda (a,d,b) com a distância $0,37$ e última com $0,85$, a matriz cofenética ficaria da seguinte forma:

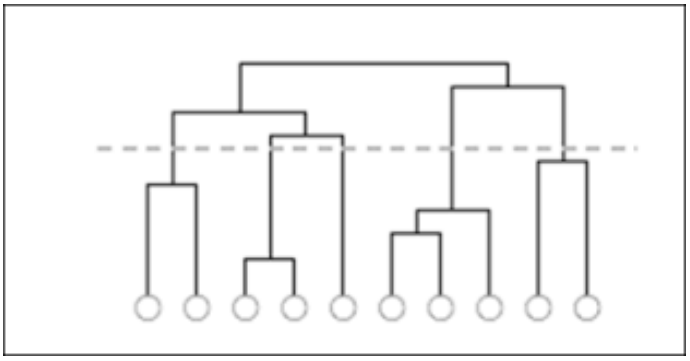
$$C = \begin{array}{c} \\ a \\ b \\ c \\ d \end{array} \begin{array}{c} \begin{array}{cccc} a & b & c & d \end{array} \\ \begin{pmatrix} 0 & 0,37 & 0,85 & 0,25 \\ 0,37 & 0 & 0,85 & 0,37 \\ 0,85 & 0,85 & 0 & 0,85 \\ 0,25 & 0,37 & 0,85 & 0 \end{pmatrix} \end{array}$$

Essa matriz cofenética C é utilizada para o cálculo de um coeficiente que busca verificar a correlação entre os as distâncias recuperadas e distâncias originais dos dados que é a matriz D.

2.2.3 Dendograma

O dendograma permite observar em forma de uma árvore hierárquica a relação entre indivíduos e grupos de indivíduos. Nesse caso, esse tipo de representação gráfica se assemelha aos ramos de uma árvore que vão se dividindo de acordo com o grau de semelhança e/ou diferença entre os objetos analisados. O tamanho das distâncias entre os dados são utilizados como ferramentas para a realização de agrupamentos, bem como para a construção dos dendrogramas. Segue abaixo um exemplo de dendograma genérico, exemplificando um banco de dados onde indivíduos estão em um mesmo grupo inicialmente e após a realização do processo, começam a ser alocados em novos grupos definidos a partir de suas semelhanças e diferenças (Hair, 2009), (Marques, 2017).

Figura 2.2: Dendograma genérico



Como relatado anteriormente na Seção 2 deste trabalho, há pelo menos 6 métodos de ligação ou algoritmos de *cluster* hierárquicos que podem ser utilizados nas mais diferentes situações e tipos de dados em pesquisas. Contudo, assumindo que a avaliação e validação dos agrupamentos é uma etapa de suma importância na formação de uma pesquisa que utiliza o método de agrupamentos, torna-se necessário utilizar abordagens que visem escolher criteriosamente os melhores métodos para a realização dos agrupamentos (Hair, 2009), (Marques, 2017).

Sabendo que em vários estudos podem ser empregados pelo menos 6 métodos para realização dos grupos e que existem para cada método pelo menos 6 medidas de distância, pode-se perceber que há várias formas de realizar agrupamentos. Realizando uma simulação de todos os métodos e medidas de dissimilaridade pode ser possível verificar o melhor tipo de método e medidas a serem utilizadas para determinado tipo de dados. Na Tabela 2.2 é possível verificar como a simulação será

organizada:

Tabela 2.2: Simulações: modelos x medidas

	Método	Euclidiana	Euclidiana^2	Maximo	Manhratan	Caberra	Minkowski
1	Single	x	x	x	x	x	x
2	Completo	x	x	x	x	x	x
3	Ward 1	x	x	x	x	x	x
4	Ward 2	x	x	x	x	x	x
5	Average	x	x	x	x	x	x
6	MCquitty	x	x	x	x	x	x
7	Median	x	x	x	x	x	x
8	Centroid		x				

Vale lembrar, que existem outros métodos e medidas de distância, porém estes são os mais empregados na prática e mais utilizados em programas computacionais (Mingoti, 2005), (Metz, 2006),(Hair, 2009).

Coeficiente de Correlação Cofenético

Buscando escolher os melhores entre os métodos empregados para a realização dos *clusters* na literatura especializada, podemos encontrar índices numéricos orientando a escolha de um método de ligação específico, como o "coeficiente de correlação cofenética" proposto inicialmente por Sokal e Rohlf em 1962. Essa medida é calculada através da equação:

$$c_{cofenetico} = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(c_{ij}-\bar{c})(s_{ij}-\bar{s})}{\sqrt{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(c_{ij}-\bar{c})^2}\sqrt{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(s_{ij}-\bar{s})^2}} \tag{2.18}$$

Em que:

c_{ij}: é o valor da similaridade entre os indivíduos i e j, resultante da matriz cofenética.

d_{ij}: é o valor da similaridade entre os indivíduos i e j, resultante da matriz de similaridade.

Observa-se que essa equação se equivale a uma correlação de pearson entre a matriz de similaridade original (Matriz D) dos dados e a matriz cofenética que é utilizada para a realização do dendograma (Matriz C). Nesse caso, esta equação é:

$$c_{cofenetico} = \frac{Cov(D, C)}{\sqrt{Var(D)}\sqrt{Var(C)}} \tag{2.19}$$

O coeficiente cofenético tem como objetivo validar o grau de ajuste entre a matriz de similaridade cofenética com a matriz resultante da simplificação do método de agrupamento, isso quer dizer que o coeficiente cofenético estabelece uma correlação entre a matriz de distâncias e o dendograma. Dessa forma, entende-se que se o dendograma for adequado, a matriz cofenética C deve ter uma alta correlação com a matriz de distâncias D original (Marques, 2017).

O valor do coeficiente cofenético varia entre 0 e 1, sendo o menor grau de distorção refletido no maior valor do coeficiente, este que indica o melhor dendograma assumindo determinada estrutura de dados. Um valor superior a 0,7 indica adequabilidade do método de agrupamento utilizado (Rohlf, 1970).

De qualquer forma, o $c_{cofenetico}$ superior a 0,7 não é argumento suficiente para validar um determinado tipo de método. Dessa forma, é necessária a inspeção visual dos dendogramas construídos a partir dos métodos validados. Isso quer dizer, que a escolha do melhor dendograma depende não só do valor obtido no $c_{cofenetico}$, mas também da análise do dendograma em relação ao banco de dados determinado em estudo (Metz, 2006).

2.3 Método Não Hierárquico

O método não hierárquico, diferentemente do método hierárquico, não envolve qualquer tipo de processo de construção em árvore. Na verdade, para realizar um procedimento não hierárquico é preciso, primeiramente, definir a quantidade de agrupamentos a serem realizados. Esse procedimento segue dois passos (Everitt et al., 2011):

- Identificar as sementes de agrupamento: As sementes de agrupamento são os pontos de partida a serem identificados, esse é o primeiro passo do procedimento.
- Designação: Com as sementes definidas o próximo passo é designar as observações às sementes do agrupamento a partir da similaridade.

A ideia principal do método é escolher uma parte inicial dos elementos e, só então, alterar os elementos desse grupo de forma que obtenha-se a melhor partição possível.

Os métodos mais utilizados por particionamento são: *K-means* e *K-medoids*. Vale lembrar que os não hierárquicos tem alguns pontos de fraqueza. O primeiro envolve a escolha a priori das posições iniciais dos centroides. Isso geralmente é resolvido na literatura (MacQueen et al., 1967), (Stewart et al., 2012) repetindo o procedimento de agrupamento para vários valores das condições iniciais e selecionando aqueles que levam aos valores mínimos das distâncias entre cada centroide e os elementos do agrupamento. Além disso, no início do procedimento, é necessário definir arbitrariamente o número K de *clusters.*

Método *K-means*

A partir de um parâmetro K acontece a partição de um conjunto de N elementos em K grupos. Sendo, então a entrada o número K de grupos e a base de dados com N elementos. Dessa forma seguem os seguinte passos para realização do procedimento:

- Escolher K elementos para que sejam os centros iniciais dos grupos
- Repetir o passo anterior
- Sublocar cada elemento ao grupo em que o elemento é mais similar a partir do valor médio dos elementos
- Calcular o valor médio dos elementos de cada grupo
- Atualizar as médias dos grupos
- Repetir o processo de forma que os elementos não mudem mais em seus grupos

Este modelo foi proposto pela primeira vez por MacQueen em 1963 (MacQueen et al., 1967) e apresenta tendências a formar grupos esféricos, sem contar que a quantidade de grupos na entrada do processo é o mesmo número de saída. Não é muito indicado para descobrimento de grupos com formas de tamanhos diferentes.

Método K-medoids

A fim de que seja corrigida a sensibilidade do método *K-means* em relação a ruídos, o método *K-medoids* usa o *medoids*, que é o elemento que possui a menor distância entre os outros elementos do grupo e o que está em localidade mais central, como ponto de referência ao invés do valor médio dos objetos. Dessa forma, o método não

hierárquico pode ocorrer com o objetivo de minimizar a soma de dissimilaridades entre cada objeto. O principal objetivo do algoritmo de agrupamento *K-medoids* é um objeto representativo para cada agrupamento. Em seguida, cada objeto remanescente é sublocado junto ao *medoid* que representa maior grau de similaridade. Dessa forma, esse processo troca um dos *medoids* por um dos não *medoids* até que a qualidade dos agrupamentos sejam melhoradas (Kaufman and Rousseeuw, 2009).

Capítulo 3

Resultados e Discussão

3.1 Análises Descritivas

Antes da coleta das impressões digitais, para os três tipos de materiais, todos os indivíduos responderam questões sobre:

- Sexo do indivíduo.
- Se são fumantes ou não.
- Se utilizaram medicamento nas 24h antecedentes a doação.
- Se tomaram banho nas últimas 24h antecedentes a doação.
- Se utilizaram cosméticos nas últimas 24h antecedentes a doação.
- Se utilizaram produtos de limpeza nas últimas 24h antecedentes a doação.
- Se tomaram café nas últimas 24h antecedentes a doação.

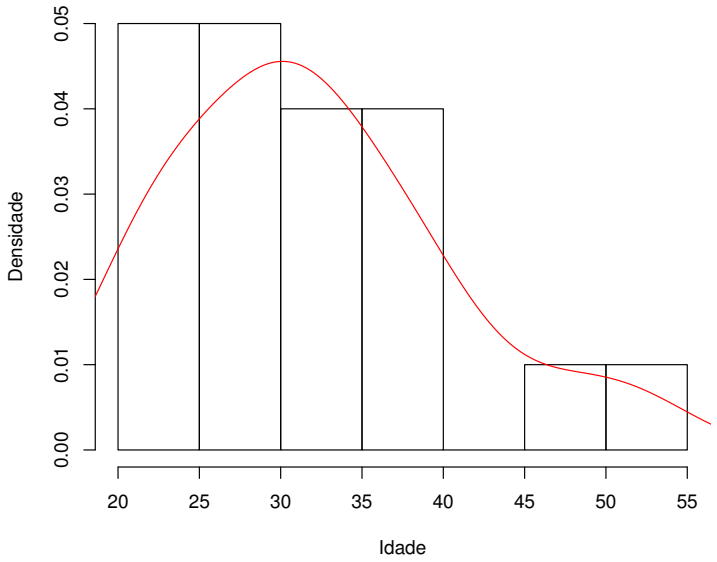
		Fumantes	Medicamento	Banho	Cosméticos	Prod. de Limpeza	Café
Quantidade	Sim	18	13	10	18	15	10
	Não	2	7	10	2	5	10

Das 20 pessoas, temos 10 mulheres e 10 homens, onde 10% dos indivíduos são fumantes, 65% das pessoas fizeram o uso frequente de algum medicamento nas 24 horas antecedentes a coleta de impressões, metade das pessoas tomaram banho antes de doarem suas impressões digitais, apenas duas pessoas não utilizam quaisquer cosméticos, 15 das 20 pessoas confessaram ter contato frequente com algum produto de limpeza específico e 10 das 20 pessoas ingeriram café.

Tabela 3.1: Distribuição das frequências das Idades

Idade	**Frequência**	**Frequência Relativa**
22 - 25	5	0,25
26 - 30	5	0,25
31 - 35	4	0,2
36 - 40	4	0,2
41 - 45	0	0
46 - 50	1	0,05
51 - 55	1	0,05
Total	20	1

Figura 3.1: Histograma das Idades dos Indivíduos



- Idade

Podemos observar que a distribuição das idades é assimétrica à direita, ou seja há mais pessoas jovens, principalmente entre 22 e 30 anos, de acordo com a tabela e com o gráfico anteriores.

- Banho, Cosméticos e Produtos de Limpeza

Levando em consideração que, a partir de resultados de pesquisas com este mesmo banco de dados usando modelos mistos, obteve-se uma relevância destacável referente às perguntas sobre banho, cosmético e produtos de limpeza no questionário, tomou-se a decisão de cruzar estas informações (Cavalcante and Vasconcelos, 2018).

Tabela 3.2: Cruzamento entre as variáveis: banho, cosméticos e produto de limpeza

		Cosméticos		Prod. Limpeza	
		Sim	Não	Sim	Não
Banho	Sim	9	1	9	1
	Não	9	1	7	3

Como pode ser observado, para os indivíduos que tomaram banho a maioria também utiliza cosméticos, somente um dos 10 indivíduos que tomou banho não teve este tipo de contato direto. O mesmo ocorre para os indivíduos que confessaram não ter tomado banho antes do experimento, dos 10, 9 tiveram contato com cosméticos e apenas 1 não obteve este contato. Quando há um cruzamento de dados acerca do banho e utilização de produtos de limpeza, o resultado para aqueles que tomam banho é o mesmo que ocorreu para os 10 indivíduos que tomaram banho. Contudo, dentre os 10 que não tomaram banho, 7 utilizaram produtos de limpeza e apenas 3 não utilizaram. Essa análise é importante, pois o banho pode ter retirado vestígios dos cosméticos e produto de limpeza usados presentes nas mãos, e que então não poderão ser identificados na análise química feitas nas impressões digitais.

- Íons

Como um dos nossos objetivos é analisar a eficácia entre os métodos de tratamento químico das digitais, então foi proposto o cálculo do número médio de íons por indivíduo em cada método, os resultados se encontram apresentados na tabela abaixo:

Tabela 3.3: Número médio de íons para cada indíviduo em cada método.

Método	**Média**	**Variância**	**C.V.**
MALDI	44,5	881,9	0,67
SALDI	32	77,2	0,27
LDI	29,9	143	0,39

Verificamos que em média no método MALDI há 44,5 íons por indivíduo, 32 íons por indivíduo no SALDI, no LDI em média temos 29,9 íons por indivíduo, portanto ao olhar a variabilidade entre cada método, podemos ver que o MALDI tem variabilidade 11 vezes maior que o SALDI, podemos ver que a variabilidade da intensidade dos íons é bem grande e pode ser influenciado na conclusão da eficacia

do tratamento químico, isso nos leva a refletir se podemos considerar a intensidade dos íons como critério para formação dos agrupamentos.

Outra possível análise proposta é a identificação de íons que estão mais evidentes entre os indivíduos, esse resultado também norteará na análise da eficácia de cada método, uma vez que esses íons podem ser identificados posteriormente e associados à alguma substância química.

- Método MALDI
  - Os íons com massa/carga:172, 178, 212, 228, 250, 354, 379 aparecem para todas as pessoas.
  - 95% das pessoas tiveram os íons:164, 190, 234, 284, 522.
  - E os íons 266, 294 e estiveram presentes em 18 das 20 pessoas.
- Método SALDI
  - Em 100% das pessoas foram identificados os ions 113,522.
  - Os 19 das 20 pessoas obtiveram 295,550.
  - E foi detectado os íons 128, 321, 494 em 85% dos indivíduos.
- Método LDI
  - Para todas as pessoas foram verificados os íons de massa/carga: 113, 494, 522, 550.
  - O íon 284 foi detectado em 18 pessoas das 20.
  - Os íons 115, 295, 321 foram identificados em 85% das pessoas.

De modo geral o íon 522 aparecem para maiorias das pessoas independente do método, para os métodos SALDI e LDI, são identificados em quase todos os indivíduos os 550, 494, 321, 294, 113, 104.

3.1.1 Métodos Hierárquicos

Levando em consideração que no banco de dados haviam muitos íons que apareciam exclusivamente para alguns indivíduos, tomou-se a decisão de cortar os dados para que os íons pouco presentes não desequilibrassem a análise. Dessa forma, para a realização dos métodos de agrupamentos hierárquicos, não hierárquicos e binários a seguir, foram considerados apenas os íons que estão presentes em pelo menos 5 dos 20 indivíduos.

Como foi pontuado na Seção 2, os métodos hierárquicos, em especial os métodos aglomerativos, começam com o um número de *clusters* igual ao número de objetos estudados. A partir da realização de cada iteração os indivíduos começam a se aglutinar àqueles que apresentem menor distância. Esse procedimento continua até o momento em que a distância entre os novos *clusters* aglutinados tonam-se apenas um grupo.

Nesse sentido, assumindo que existem diferentes métodos de ligação e diferentes medidas de distância para realizar os agrupamentos e construir os dendogramas, tomou-se a decisão de simular os métodos de ligação com as distâncias de similaridade e assim, calculando cada coeficiente cofenético de cada relação. Ainda referindo-se à Seção 2.2 esse coeficiente tem como objetivo relacionar as medidas de dissimilaridade e os métodos de aglomerativos a fim de que sejam considerados, os agrupamentos realizados no banco de dados. Quando o $C_{cofenetico}$ entre os métodos e medidas é um valor abaixo de 0,7 assume-se que aquele determinado método com aquela distância específica não são válidos para a realização do *clustering* e para a construção do dendograma(Metz, 2006).

Tabela 3.4: Coeficiente Cofenético - MALDI

	Método	Euclidiana	Euclidiana[2]	Maximo	Manhratan	Canberra	Minkowski
1	Single	0,91	0,83	0,94	0,93	0,96	0,95
2	Completo	0,92	0,80	0,95	0,95	0,95	0,96
3	Ward 1	0,88	0,77	0,83	0,88	0,94	0,88
4	Ward 2	0,90	0,79	0,91	0,90	0,96	0,90
5	Average	0,98	0,96	0,99	0,97	0,97	0,97
6	MCquitty	0,97	0,94	0,98	0,97	0,96	0,97
7	Median	0,97	0,95	0,98	0,97	0,79	0,96
8	Centroid		0,96				

[1]Para a distância Minkowski considerou λ=0,5

A partir dos resultados obtidos na tabela 3.4 após o cálculo do $C_{cofeneteico}$ foi possível identificar que não há qualquer método de agrupamento e medida de dissimilaridade que não sejam válidos para realização dos agrupamentos no banco de dados do tratamento MALDI. Todos os coeficientes cofenéticos são superiores a 0,7, diferentemente dos outros tratamentos. Isso quer dizer que não há qualquer método e medida de dissimilaridade que seja inválido segundo do coeficiente cofenético para realização dos agrupamentos e construção dos dendogramas.

Tabela 3.5: Coeficiente Cofenético - SALDI

	Método	Euclidiana	Euclidiana^2	Maximo	Manhratan	Canberra	Minkowski
1	Single	0,93	0,90	0,92	0,86	0,63	0,77
2	Complete	0,93	0,91	0,94	0,74	0,70	0,74
3	Ward 1	0,85	0,86	0,75	0,81	0,59	0,43
4	Ward 2	0,88	0,90	0,90	0,81	0,64	0,58
5	Average	0,95	0,91	0,96	0,89	0,79	0,79
6	MCquitty	0,84	0,91	0,95	0,83	0,77	0,78
7	Median	0,90	0,92	0,96	0,78	0,49	0,66
8	Centroid		0,92				

O novo material nanoestruturado SALDI obteve resultados diferentes em relação ao método tradicional MALDI no que se refere ao coeficiente cofenético. A partir dos resultados obtidos na tabela 3.5 de simulação entre métodos de agrupamento e medidas de distâncias foi percebido que a medida *Canberra* não é uma boa medida para a maioria dos métodos de ligação, como *Single, Complete, Ward e Median.* Além disso, para alguns métodos a medida de dissimilaridade *Minkowski* também teve coeficiente cofenético abaixo de 0,7. Isso quer dizer que estas distâncias e métodos quando simulados não obtiveram um nível de coeficiente suficiente para adequar os prováveis dendogramas. Sendo assim, as demais medidas de similaridade e métodos parecem ser bons para realização dos agrupamentos, pois o coeficiente cofenético supera o valor 0,7.

Os resultados do $C_{cofeneteico}$ presentes na Tabela 3.6 indicam que para o banco de dados de impressões digitais coletadas sem qualquer material (LDI) a medida de dissimilaridade *Canberra* não é muito indicada. Os valores do $C_{cofeneteico}$ dessa medida está inferior a 0,7 para praticamente todos os métodos de agrupamento, excetuando-se apenas o coeficiente referente ao método *Average* que supera por pouco o valor 0,7. Sendo assim, é possível indicar que os possíveis dendogramas construídos com os métodos e medidas que possuem coeficiente cofenético inferiores

Tabela 3.6: Coeficiente Cofenético - LDI

	Método	Euclidiana	Euclidiana^2	Maximo	Manhratan	Canberra	Minkowski
1	Single	0,93	0,91	0,81	0,94	0,59	0,94
2	Complete	0,79	0,73	0,77	0,75	0,56	0,80
3	Ward 1	0,77	0,73	0,76	0,87	0,61	0,82
4	Ward 2	0,78	0,73	0,77	0,88	0,62	0,82
5	Average	0,95	0,93	0,88	0,93	0,72	0,95
6	MCquitty	0,92	0,92	0,81	0,93	0,66	0,89
7	Median	0,92	0,91	0,88	0,89	0,39	0,91
8	Centroid		0,93				

a 0,7, possivelmente não conseguiriam expressar muito bem os dados estudados.

Como indicado anteriormente na seção 2.3 deste trabalho, o valor do $C_{cofeneteico}$ não é argumento suficiente para validar um determinado método ou distância que serão utilizados para a construção de um dendograma que consiga expressar real e graficamente a estrutura de dados em questão. Dessa forma optou-se por escolher os $C_{cofeneteico}$ superiores a 0,9, a fim de que os agrupamentos tenham uma melhor validação e maior possibilidade de serem ideais para as análises. No tratamento MALDI, tabela 3.4, foi perceptível que os métodos *Average* e *MCquitty* possuem coeficiente superior para todas as medidas de dissimilaridade. Enquanto, que o método *Ward 1* só possui coeficiente cofenético superior a 0,9 na medida de dissimilaridade *Canberra*.

De modo geral, foi percebido a partir do coeficiente que a medida de dissimilaridade de *Canberra* para todos os materiais é a medida com maior índice de invalidação segundo o coeficiente cofenético. As demais medidas, surgem então como melhores opções para realização dos agrupamentos.

3.1.2 MALDI

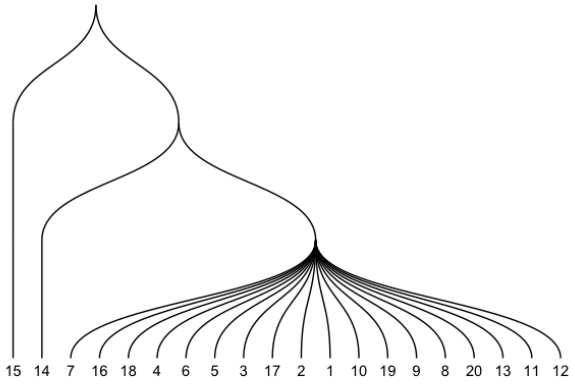
A partir dos resultados obtidos nas simulações realizadas pelos métodos aglomerativos e medidas de distância o tratamento MALDI, obteve a maior quantidade de $C_{cofeneteico}$ válidos em relação aos dois tratamentos. Na verdade, em nenhuma simulação o $C_{cofeneteico}$ ficou abaixo de 0,7. A fim de que houvesse uma maior rigidez na validação e observação dos agrupamentos analisou-se, exclusivamente, os métodos e distâncias simuladas que obtiveram $C_{cofeneteico}$ maior ou igual a 0,9. Dessa forma, foram realizados 34 dendogramas. Houve 3 tipos de organização de agrupamentos que mais apareceram nos 34 dendogramas realizados. Como indicado anteriormente,

faz parte da validação dos dendogramas a análise visual do gráfico a fim de que seja verificado se o dendograma em questão consegue expressar verdadeiramente as características dos dados estudados.

O dendograma que mais apareceu ficou da seguinte forma:

- Cluster 1: Indivíduo 15.
- Cluster 2: Indivíduos 14.
- Cluster 3: Restante dos indivíduos.

Figura 3.2: Dendograma 1 - Tratamento MALDI para 3 grupos



Vale destacar que os indivíduos 15 e 14 apareceram separados dos demais indivíduos. A partir do questionário é possível perceber que ambos fazem o uso de medicamentos, não tomaram banho, fazem uso de cosméticos, usam frequentemente produtos de limpeza e não ingeriram café antes do experimento.

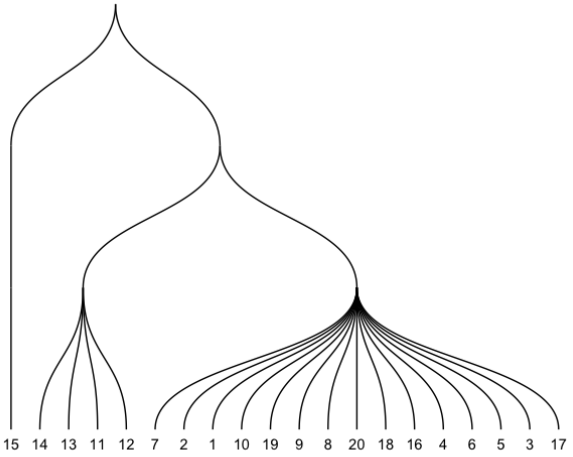
Foi identificado que o indivíduo 14 possui 25 íons que aparecem para outros indivíduos porém estes têm intensidade muito maior no indivíduo 14, são eles: 147, 164, 178, 189, 190, 191, 192, 194, 208, 227, 261, 274, 278, 291, 294, 309, 335, 337, 379, 400, 413, 433, 441, 443 e 471. No indivíduo 15, também não há qualquer íon que seja exclusivo do indivíduo, porém há 11 íons que aparecem com intensidade muito maior em relação aos demais indivíduos, são eles: 228, 250, 256, 266, 284, 304, 354, 494, 522, 550 e 691. Entende-se que a intensidade destacável desses íons possa estar influenciando no resultado explicitado no dendograma.

Todos os dendrogramas com o coeficiente cofenético maior que 0,9, e com agrupamentos em que os tem exclusividade para os indivíduos 14 e 15, podem ser visto no Apêndice A na Seção A.1

O segundo dendograma que precisa ser destacado ficou agrupado da seguinte forma:

- Cluster 1: Indivíduo 15.
- Cluster 2: Indivíduos 14, 13, 11, 12.
- Cluster 3: Restante dos indivíduos.

Figura 3.3: Dendograma 2 - Tratamento MALDI para 3 grupos



Baseando-se no questionário feito antes da obtenção das impressões digitais, foi possível identificar que no *Cluster* 2 desse dendograma os indivíduos tiveram respostas parecidas. Estas foram as respostas dos indivíduos do *Cluster* 2:

Doador	Sexo	Idade	Fumante	Medicamento	Banho	Cosméticos	Prod. Limpeza	Café
14	M	40	N	S	N	S	S	N
13	F	29	N	S	N	S	S	N
12	M	24	N	N	N	N	S	N
11	F	32	N	S	N	S	S	N

A partir das respostas do questionário observa-se que os indivíduos do *Cluster* 2 obtém praticamente as mesmas respostas quando perguntados sobre banho, contato com cosméticos, medicamento, cigarro, café e produtos de limpeza. Na verdade, somente o indivíduo 12 obtém duas respostas diferentes nas perguntas, que é quando perguntado se teve contato com algum produto cosmético ou medicamento. De certa forma, espera-se que os íons sejam mais influenciados por práticas em que aconteçam contato direto com a pele. Sendo assim, imagina-se que as perguntas sobre banho,

cosméticos e produtos de limpeza possam apresentar relevância na construção dos agrupamentos.

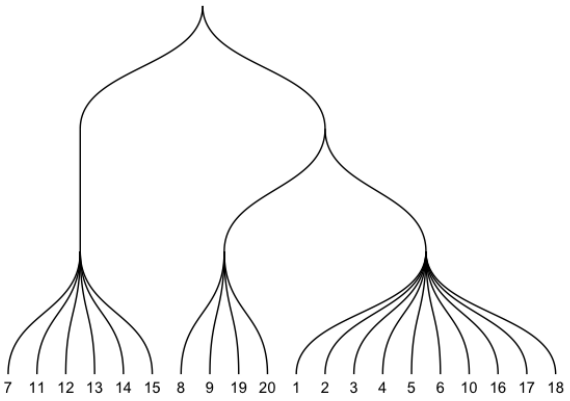
Além disso, como comentado anteriormente, a impressão coletada pelo indivíduo 15 indicou, depois do corte do banco de dados, que este indivíduo possui 11 íons que aparecem para outros indivíduos porém com muito menos intensidade. Para o *Cluster* 2 realizado nesta segunda simulação, que é constituído pelos indivíduos 14, 13, 12 e 11, há exatamente 37 íons que aparecem para os 4 indivíduos com muito mais intensidade do que para os demais indivíduos, são eles: 104, 107, 121, 122, 123, 128, 130, 139, 146, 147, 152, 164, 172, 178, 190, 191, 192, 194, 208, 227, 247, 268, 274, 278, 282, 291, 294, 295, 309, 335, 337, 362, 379, 395, 413, 441, 443.

Todos os dendrogramas com o coeficiente cofenético maior que 0,9, e com agrupamentos em que os tem exclusividade para os indivíduos 15 e 14, 13, 11, 12, podem ser visto no Apêndice A, seção A.2.

Por fim, o último tipo de dendograma com destaque ficou agrupado da seguinte forma:

- Cluster 1: Indivíduos 7, 14, 15, 13, 11, 12
- Cluster 2: Indivíduos 8, 20, 9, 19
- Cluster 3: Indivíduos 10, 1, 18, 16, 2, 4, 17, 5, 3, 6

Figura 3.4: Dendograma 3 - Tratamento MALDI para 3 grupos



Baseando-se no questionário feito antes da obtenção das impressões digitais, foi possível identificar que no *Cluster* 1 desse dendograma os indivíduos 7, 14, 15, 13, 12

e 11 tiveram respostas iguais em perguntas consideradas relevantes para os resultados obtidos nos dendogramas. Eles indicaram que não tomaram banho e responderam sim quando perguntados sobre a utilização de cosméticos e produtos de limpeza, com exceção ao indivíduo 12 que responde que não utilizou cosméticos, diferenciando-se dos demais indivíduos do agrupamento.

Doador	Sexo	Idade	Fumante	Medicamento	Banho	Cosméticos	Prod.Limpeza	Café
14	M	40	N	S	N	S	S	N
15	F	20	S	S	N	S	S	N
7	F	31	N	S	N	S	S	N
13	F	29	N	S	N	S	S	N
11	F	32	N	S	N	S	S	N
12	M	24	N	N	N	N	S	N

Em comparação ao *Cluster* 2 é possível perceber que, justamente, nas perguntas sobre banho, cosméticos e produtos de limpeza, as respostas se diferenciam em relação ao *Cluster* 1. Veja:

Doador	Sexo	Idade	Fumante	Medicamento	Banho	Cosméticos	Prod.Limpeza	Café
8	M	38	N	N	S	S	S	S
20	M	36	S	N	S	S	S	S
9	F	26	N	N	N	S	N	N
19	M	47	N	S	S	S	S	S

Os indivíduos 8, 20 e 19 tomaram banho, diferentemente da maioria dos indivíduos no *Cluster* 1. Além disso, o indivíduo 9 não teve contato com produtos de limpeza, diferenciando, mais uma vez, individuos do *Cluster* 1 e *Cluster* 2. O mesmo acontece no agrupamento 3, nas três perguntas sobre banho, cosméticos e produtos de limpeza, as respostas se diferenciam.

Ao avaliar o banco de dados com a intensidade dos íons coletados nas impressões digitais é possível perceber que para os indivíduos do agrupamento 1 (7, 15, 14, 13, 12, 11) os seguintes íons são exclusivos em relação aos demais indivíduos: 107, 121, 122, 123, 152, 278, 291, 309, 333, 362, 395, 413, 433, 439, 455, 471, 506, 666. Isso quer dizer que para os demais indivíduos fora desse agrupamento não há ocorrência dos supracitados íons. Dessa forma, é possível observar que a ocorrência desses íons para determinados indivíduos e não para outros podem justificar a organização do agrupamento realizado.

Para o agrupamento 2, que é constituído pelos indivíduos 8, 9, 19 e 20, não foram encontrados íons exclusivos ou com intensidade maior em relação aos demais indivíduos. O mesmo acontece para indivíduos do agrupamento 3.

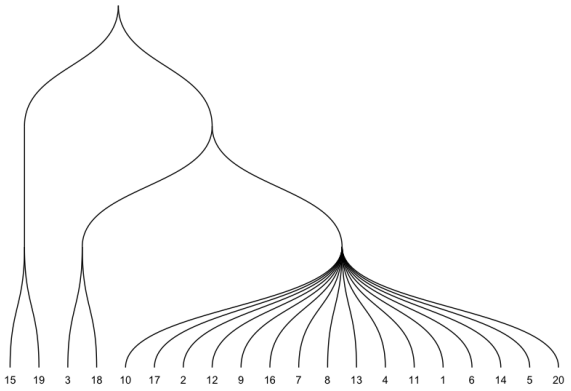
Todos os dendrogramas com o coeficiente cofenético maior que 0,9, e com agrupamentos em que os tem exclusividade para os indivíduos 15, 14, 13, 11, 12 e 7, podem ser visto no Apêndice A, seção A.3.

3.1.3 SALDI

No tratamento SALDI, foram analisados os dendogramas validados que foram construídos a partir das simulações dos métodos de ligação e medidas de dissimilaridade, estes que obtiveram $C_{cofeneteico}$ igual ou superior a 0,9. Foram realizados, portanto, para o tratamento SALDI por meio do *software* R, 16 dendogramas. Além disso, tomou-se a decisão de escolher 3 *clusters* para agrupar os indivíduos. Foi possível perceber que a maioria dos dendogramas indicaram que os 3 grupos foram divididos em:

- Cluster 1: Indivíduos 15 e 19
- Cluster 2: Indivíduos 3 e 18
- Cluster 3: Restante dos indivíduos

Figura 3.5: Dendograma - Tratamento SALDI para 3 grupos



Como foi indicado anteriormente, a maioria dos agrupamentos indicam que existem quatro indivíduos que mais aparecem. Dos 19 dendogramas realizados 9 ficaram com os agrupamentos separados da forma acima e 6 ficaram com o *Cluster* 2 contendo apenas o indivíduo 3, enquanto que o indivíduo 18 ficou agrupado no *Cluster* 3 com os demais indivíduos restantes. Porém no agrupamento realizado a partir da

medida de similaridade Euclidiana e método *mediano* dendograma fica totalmente diferente. Neste caso específico o *Cluster* 1 apenas com o indivíduo 3 e *Cluster* 2 é constituído apenas pelo indivíduo 10, o *Cluster* 3 agrupamento fica com o restante dos indivíduos.

Doador	Sexo	Idade	Fumante	Medicamento	Banho	Cosméticos	Prod.Limpeza	Café
3	F	22	N	S	S	S	S	S
18	M	30	N	N	S	N	N	N
15	F	20	S	S	N	S	S	N
19	M	47	N	S	S	S	S	S

Quando analisadas as respostas dos questionários não foi possível identificar um motivo para que os indivíduos 3, 18 e 15, 19 estarem agrupados em um cluster diferente dos demais indivíduos. Pode-se observar que há diferenças de respostas entre os indivíduos 3 e 18 no que se refere às perguntas sobre medicamento, cosméticos e produtos de limpeza. O mesmo acontece com os indivíduos 15 e 19. Estes até têm a mesma respostas quando perguntados sobre a utilização de cosméticos e produtos de limpeza. Contudo, quando a pergunta é tomar banho os indivíduos tem respostas diferentes. Sendo assim, é difícil relacionar as respostas obtidas no questionário com os dendogramas construídos a partir dos métodos hierárquicos para o tratamento SALDI.

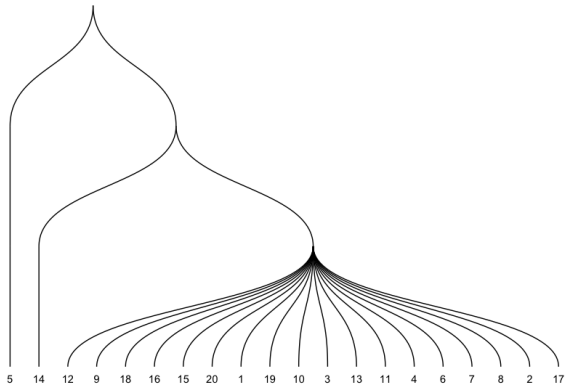
Todos os dendrogramas com o coeficiente cofenético maior que 0,9, e com agrupamentos indivíduos citados [15,19] e [3,18] podem ser visto no Apêndice B.

3.1.4 LDI

Para o banco de dados com impressões digitais coletadas sem qualquer tipo de material, foram analisados os dendogramas validados que foram construídos a partir das que obtiveram validação via $C_{cofeneteico}$ igual ou superior a 0,9. No quadro de simulção de métodos de ligação e medidas de dissimilaridade, apenas 15 coeficientes superaram o $C_{cofeneteico}$ igual ou superior a 0,9. Praticamente, todos os dendogramas realizados agruparam os indivíduos da seguinte forma:

- Cluster 1: Indivíduos 5
- Cluster 2: Indivíduos 14
- Cluster 3: Restante dos indivíduos

Figura 3.6: Dendograma - Tratamento LDI para 3 grupos



Somente o dendograma construído a partir do método de ligação Mediana (*Median*) com a medida de distância *Minkowski* obteve resultado diferente. No caso, ao invés do indivíduo 14 estar agrupado sozinho no segundo agrupamento, este é substituído pelo indivíduo 15.

A partir das respostas obtidas pelo questionário aplicado aos indivíduos antes da coleta das impressões digitais não foi possível identificar o motivo dos indivíduos 14 e 5 terem sido agrupados em um cluster diferente dos demais. Abaixo, segue a resposta de ambos:

Doador	Sexo	Idade	Fumante	Medicamento	Banho	Cosméticos	Prod.Limpeza	Café
14	M	40	N	S	N	S	S	N
5	F	30	N	S	S	S	S	S

Realizando a análise do banco de dados das impressões digitais coletadas sem qualquer material, foi possível identificar para os indivíduos 14 e 5 não existem íons exclusivos. Isso quer dizer que não há quaisquer íons que aparecem apenas para estes indivíduos. Contudo, para o indivíduo 14, os íons 113, 115 e 125 aparecem com intensidade muito maior do que quando aparecem para outros indivíduos. Já para o indivíduo 5 os íons 12, 141, 157, 167, 368, 494 e 522 aparecem com intensidade bem maior do que quando aparecem para outros indivíduos.

Todos os dendogramas obtidos com o $C_{cofenetico} > 0,9$, e exclusivo para agrupamento com indivíduos 5 e 14, podem ser vistos no Apêndice C.

3.2 Métodos Não-Hierárquicos

O método escolhido para a realização dos métodos não hierárquicos no banco de dados estudado foi o *K-means*. O primeiro passo para a realização de um método de particionamento é a definição da quantidade de grupos a serem agrupados determinado número N de elementos. A fim de que seja possível realizar uma comparação de métodos hierárquicos com métodos não hierárquicos, foi escolhido a quantidade de 3 *clusters* para a realização da rotina do *K-means*. Como informado anteriormente, a ideia do método é, justamente, escolher uma quantidade de sementes para que sirvam de pontos de partida para que os N elementos sejam sublocados à semente mais similar. Foram realizadas, portanto 10 simulações com $K = 3$ para cada tipo de tratamento.

Ao analisar as 10 simulações de *K-means* realizadas para o banco de dados do tratamento MALDI, foi possível perceber que todas as simulações agruparam os indivíduos da seguinte forma:

- Cluster 1: Indivíduo 15
- Cluster 2: Indivíduos 14, 13, 11, 12
- Cluster 3: Restante dos indivíduos

Estes agrupamentos são exatamente iguais a forma como o dendograma na Figura 3.3 agrupou os indivíduos nos métodos hierárquicos. Isso quer dizer que para o tratamento MALDI ambas as metodologias de agrupamentos obtiveram resultados muito parecidos, indicando portanto que os dados coletados por este material teve bom desempenho na obtenção de agrupamentos para os dois métodos. Como comentado na análise de resultados dos métodos hierárquicos para o tratamento MALDI. Os indivíduos agrupados no *Cluster* 2 possuem respostas parecidas para os fatores já analisados na seção anterior.

O mesmo foi realizado para o método SALDI. Nesse caso, a metodologia não hierárquica *K-means*, nas 10 simulações, agrupou os indivíduos predominantemente da seguinte maneira:

- Cluster 1: Indivíduos 15 e 19
- Cluster 2: Indivíduos 3 e 18

- Cluster 3: Restante dos indivíduos

Mais uma vez os métodos hierárquicos e não hierárquicos obtiveram a mesma resposta no que se refere aos agrupamentos dos 20 indivíduos. Dessa forma, esse é mais um motivo que contribui para relacionar os agrupamento obtidos com o questionário realizado. As perguntas sobre banho, cosméticos e produtos de limpeza parecem obter relevância no resultados da clusterização.

Por fim, a mesma rotina de 10 simulações foi aplicada ao tratamento LDI, que é a coleta de impressões sem qualquer material. O resultado obtido foi:

- Cluster 1: Indivíduos 5, 9, 14 e 18
- Cluster 2: Indivíduos 8, 10, 12 e 16
- Cluster 3: Restante dos indivíduos

Para o tratamento LDI, os agrupamentos predominantes nas 10 simulações realizadas pelo método *K-means* são completamente diferentes dos agrupamentos observados nos dendogramas dos métodos hierárquicos.

3.3 Abordagem Binária

Das medidas de similaridade citadas na Seção 2.1.1, aplicamos no banco de dados as citadas na tabela abaixo:

Tabela 3.7: Medidas para dados binários utilizadas

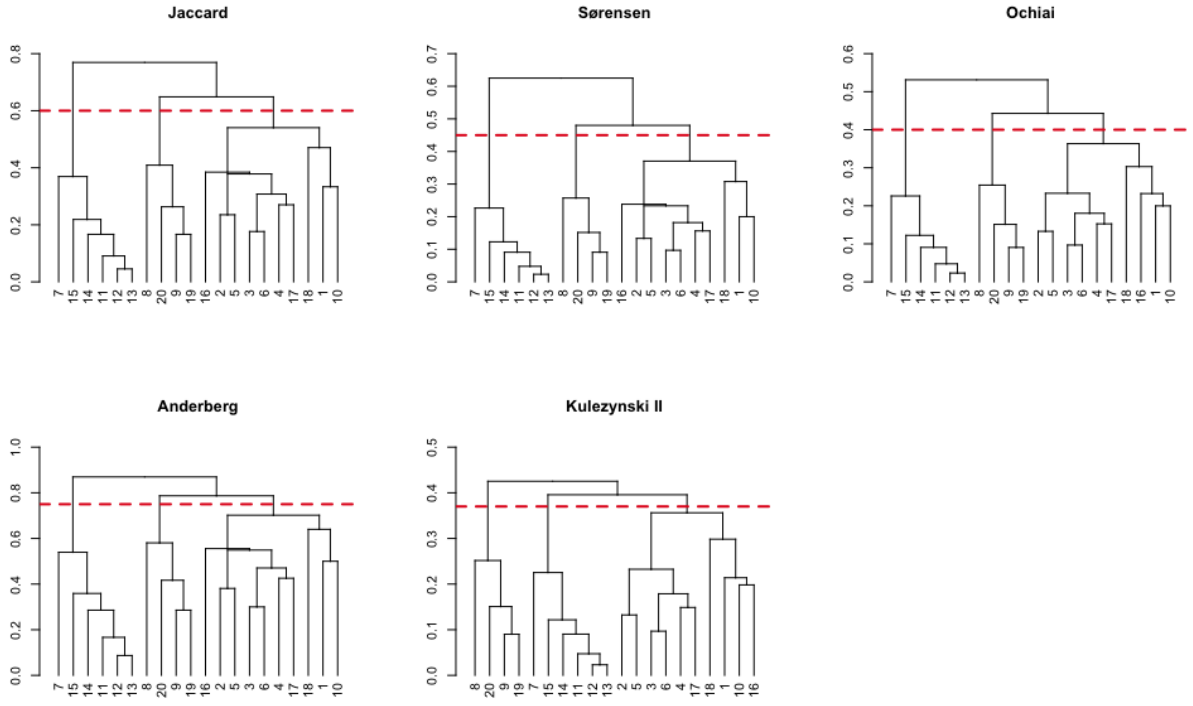
Nome	$d_{(X_i,X_j)}$	Suporte
Jaccard	$\frac{a}{a+b+c}$	[0,1]
Sørensen	$\frac{2a}{2a+b+c}$	[0,1]
Ochiai	$\frac{a}{\sqrt{(a+b)(a+c}}$	[0,1]
Anderberg	$\frac{a}{a+2(b+c)}$	[0,1]
Kulezynski II	$\frac{a}{2}\left(\frac{1}{(a+b)}\frac{1}{(a+c)}\right)$	[0,1]

Foram aplicadas essas medidas, já que não é levado em consideração a ausência conjunta (d), ou seja, caso dois indivíduos não tenham a presença de um determinado

íon, isso não será critério para mensurar o nível de semelhança entre eles. Para gerar os resultados para cada tratamento químico, foi registrado para cada método descrito na primeira coluna da Tabela 3.7 uma Matriz de Distância D citada na Seção 2.1.2 e a partir disso construído os dendogramas que serão apresentados a seguir. Vale ressaltar que esses resultados tiveram o auxilio do *Software* R, cuja a programação do cálculo das distâncias se encontra no Apêndice D.

MALDI

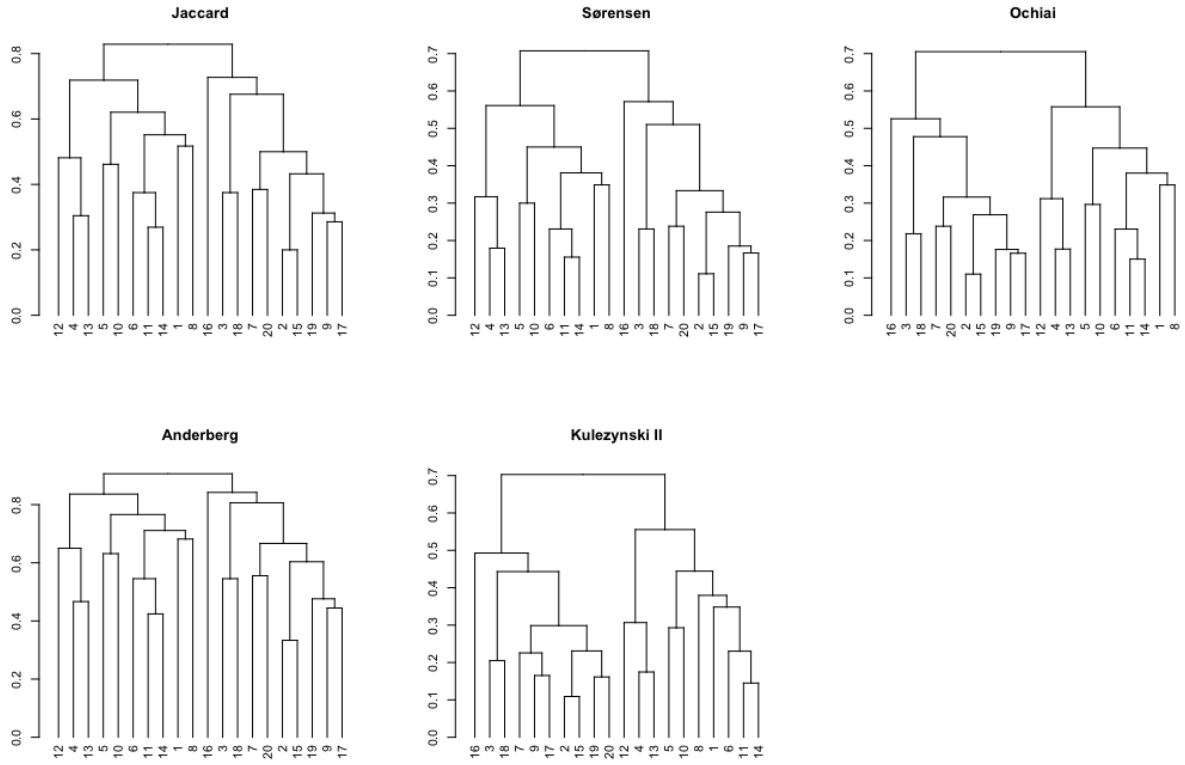
Figura 3.7: Dendograma dos indivíduos com o tratamento MALDI para os métodos: Jaccard, Sørenen, Ochiai, Anderberg, Kulezynski II



Para esse resultado, podemos ver que todos as medidas de distância obtiveram os mesmos agrupamentos do Dendograma 3 (figura 3.4), na seção 3.1.2, contendo os indivíduos: 7, 14, 15, 13, 11, 12 em um grupo e os indivíduos: 8, 20, 9, 19, no outro grupo, como obtivemos o mesmo resultado, então mantem-se a mesma análise descritiva das características do questionário e dos íons.

SALDI

Figura 3.8: Dendograma dos indivíduos com o tratamento SALDI para os métodos: Jaccard, Sørenen, Ochiai, Anderberg, Kulezynski II

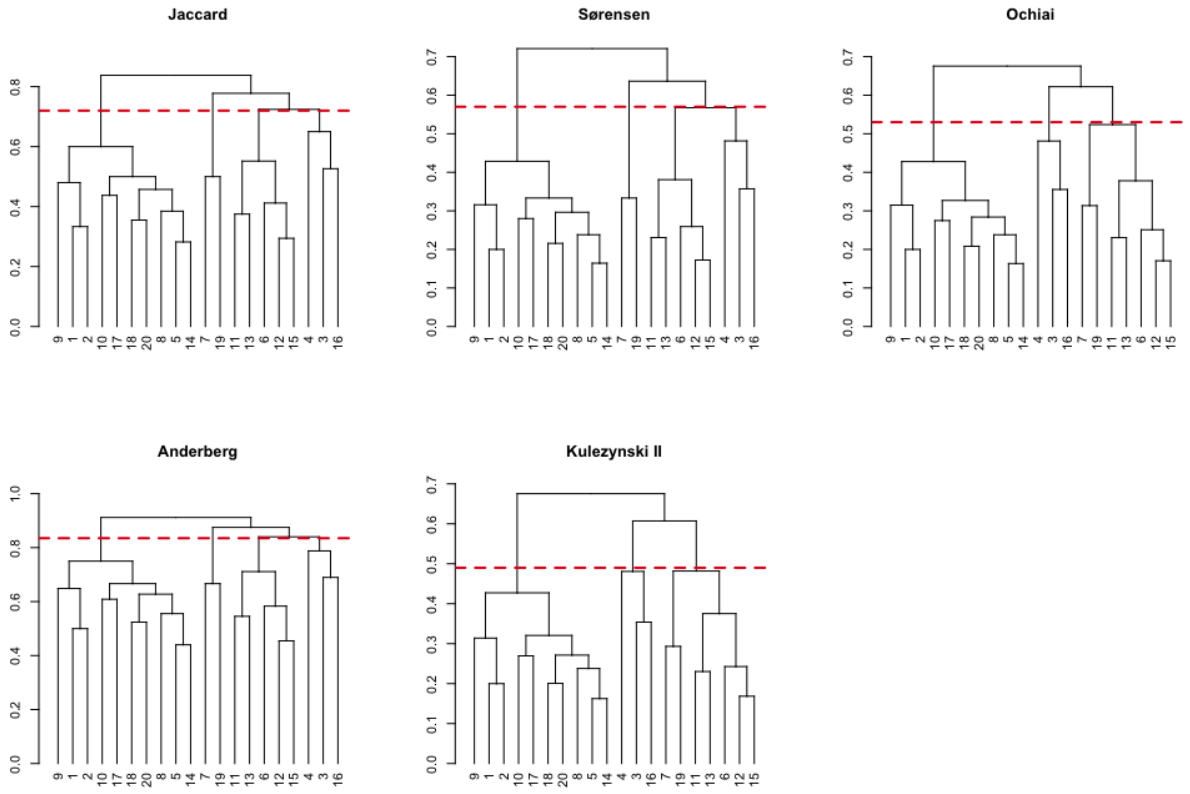


No tratamento SALDI, não obtemos nenhuma semelhança de resultados como procedimentos feitos anteriormente, um grupo que aparece em todas as medidas é o agrupamento dos indivíduos: 12, 4, 13. Podemos observar também que na maioria das vezes o individuo 16 é formado primeiramente e fica sozinho até o final das iterações.

Para o grupo observados em todas as medidas (12, 4, 13) percebemos que pelas características obtidas pelo questionário, o único aspecto que têm em comum é que não são fumante e que não tomaram banho antes do experimento.

LDI

Figura 3.9: Dendograma dos indivíduos com o tratamento LDI para os métodos: Jaccard,Sørenen, Ochiai,Anderberg,Kulezynski II



Assim como o SALDI, quando não usamos tratamento químico para analisar as digitais com as medidas para dados binário obtemos alguns grupos que se destacam. Para todas as medidas podemos perceber a formação do grupo com os indivíduos 7,19. E outro grupo que também é bastante evidente são das pessoas 4, 3, 16.

Usando a presença/ausência dos íons nas digitais em comparação as características observadas no dia do experimento, o grupo de indivíduos 4, 3, 16, não teve todas as características coletadas no questionário iguais.

Já para os indivíduos 7 e 19, ambos usam medicamentos, tem contato frequente com produto de limpeza, e usam cosméticos, porém para as outras características mostram-se diferentes entre-si.

Assim como esperávamos, o LDI, nos trouxe resultados que não foi aderente aos aspectos registrados no questionário.

Capítulo 4

Conclusões

Diante de tantas expectativas para as conclusões dos tratamentos químicos MALDI, SALDI e LDI, vimos que obtivemos diferentes resoluções para cada um. Olhando os resultados vemos que alguns grupos foram bem coerentes com as características observadas no questionário, porém para algumas metodologias como os métodos hierárquicos para o tratamento SALDI e LDI ou então os métodos não hierárquicos para o tratamento LDI, não obtivemos resultados responsivos, isso nos leva a análise de algumas questões.

Em primeiro lugar podemos indicar que existem diferenças entre os tratamentos, a captação de íons e a intensidade em que foram registrados diferem para cada um dos tratamentos.

Outra questão a ser abordada, é que nos diferentes métodos multivariados aplicados no banco de dados, também foram obtidos, em alguns casos, resultados diferentes para um mesmo tratamento, isso pode mostrar que há diversos métodos estatísticos que podem ser adequados, ou não, para um determinado banco de dados.

E uma terceira perspetiva de toda essa análise, teve como pressuposto dois fatores, na analise descritivas observamos e pontuamos uma variância da intensidade muito grande, além disso em muitos trabalho que envolvem a espectrometria de massa, é analisado apenas a presença do íon, descartando então a intensidade em que aparece na digital, levando em consideração esses fatores foi proposto a análise de agrupamento para dados com a presença ou ausência dos íons para cada indivíduo. Essa proposta foi realizada a fim de analisar os pressupostos assumidos anteriormente.

Em ultimo lugar, devemos levar em consideração que a validação dos dendogramas e agrupamentos obtidos foram realizadas com base nas respostas obtidas pelo

questionário. Contudo, devemos ponderar que os resultados obtidos podem estar revelando outras características que não foram obtidas nas respostas ao questionário. Por isso, citamos nas abordagens analisadas a ocorrência dos íons para que em um trabalho futuro seja possível identificá-los e relacioná-los a alguma substância que talvez esteve presente para todos os indivíduos de um *cluster* que foi analisado e obtido pelos procedimentos realizados neste trabalho. Ainda sobre essa relação entre as características e os agrupamentos observamos que em nenhum dos dendogramas encontrados os indivíduos fumantes ficam em um mesmo *cluster*. Isso pode indicar que o fato de fumar pode não ser tão relevante para a construção dos agrupamentos, até mesmo porque os íons que correspondem ao cigarro não foram identificados no experimento. Em contrapartida, percebeu-se a partir dos agrupamentos realizados que os indivíduos que permaneceram em um mesmo *cluster* obtiveram respostas parecidas quando perguntados sobre banho, uso de cosméticos e produtos de limpeza. Levando em consideração que estas atividades possibilitam um contato direto com a pele, pode-se verificar que estas informações podem ter sido relevantes para o resultado final dos agrupamentos.

Apêndice A

MALDI - Dendogramas

A.1 Indivíduos: 14 e 15

Figura A.1: Método Single

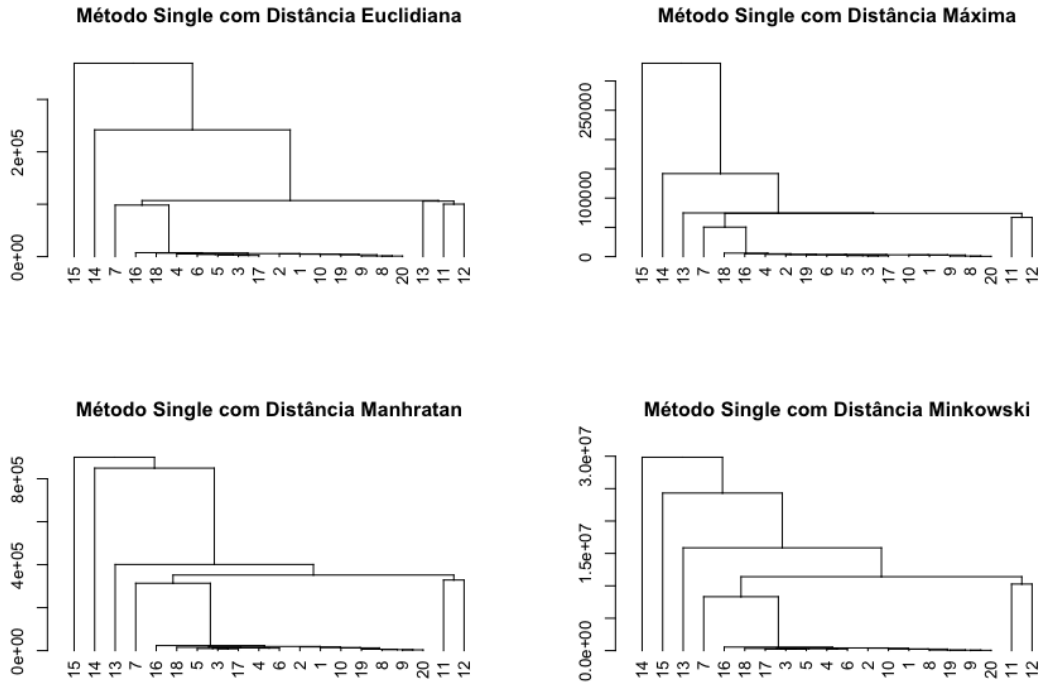


Figura A.2: Método Average

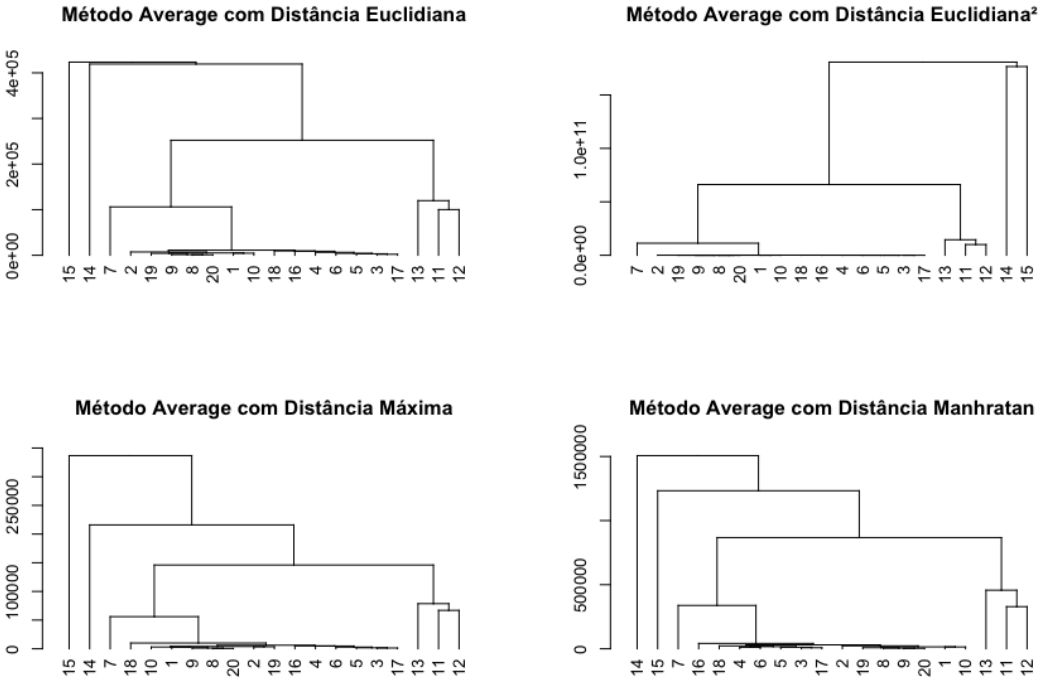


Figura A.3: Método MCquitty

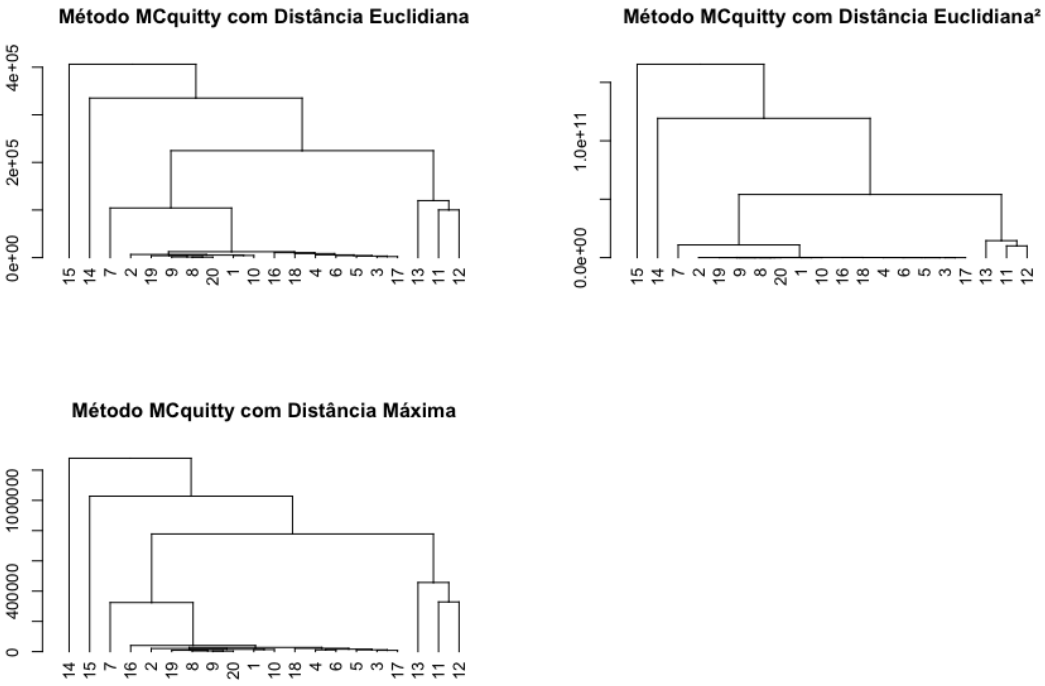
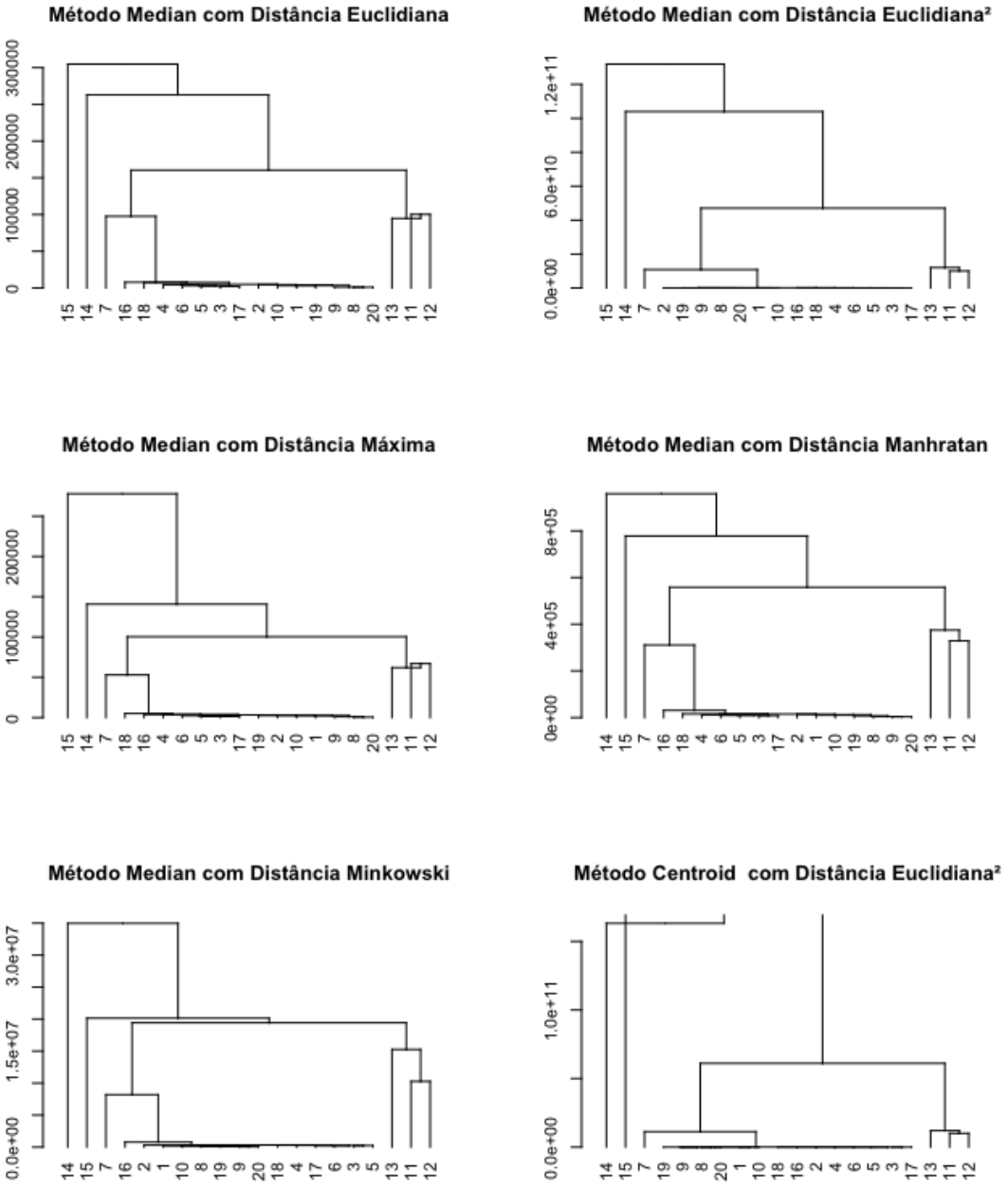
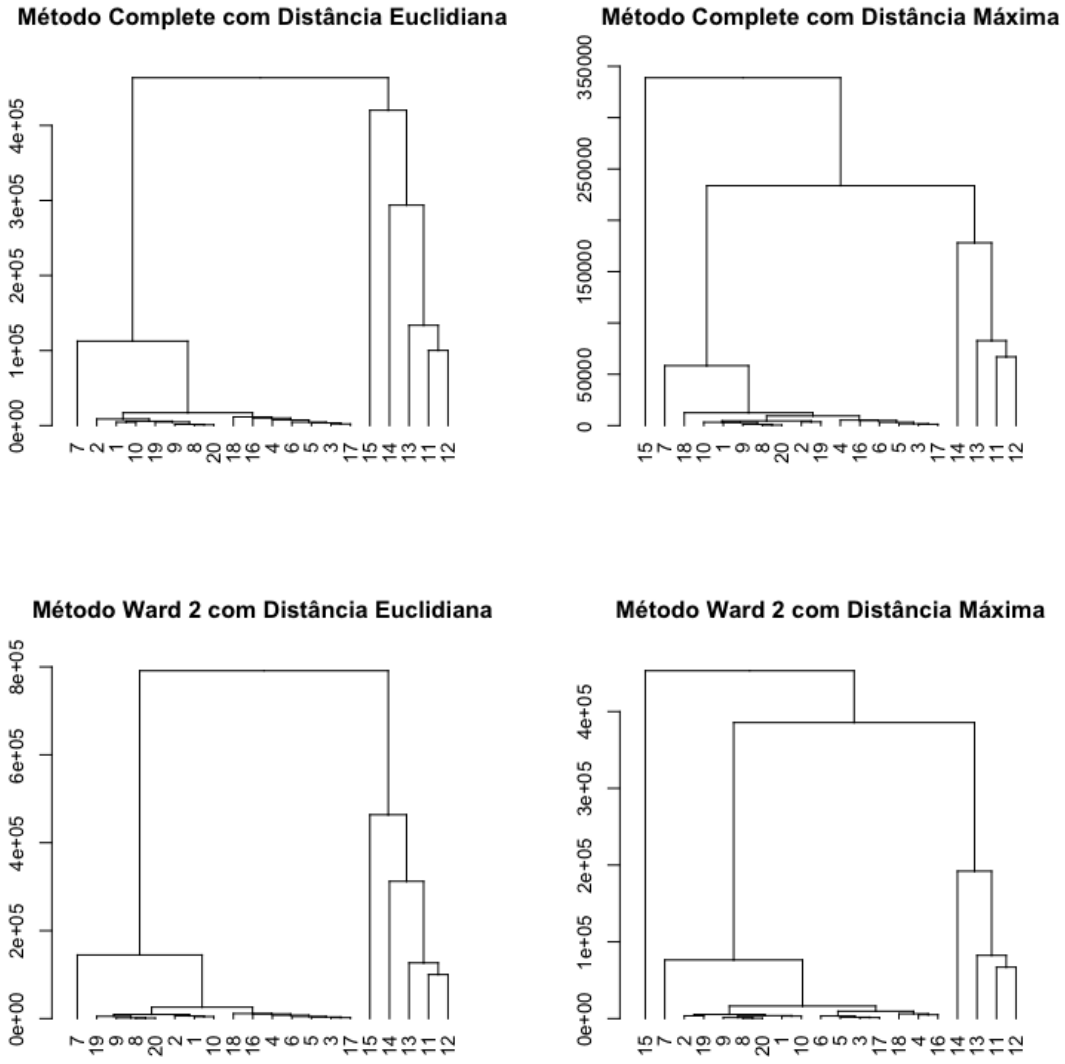


Figura A.4: Método Median e Centroid



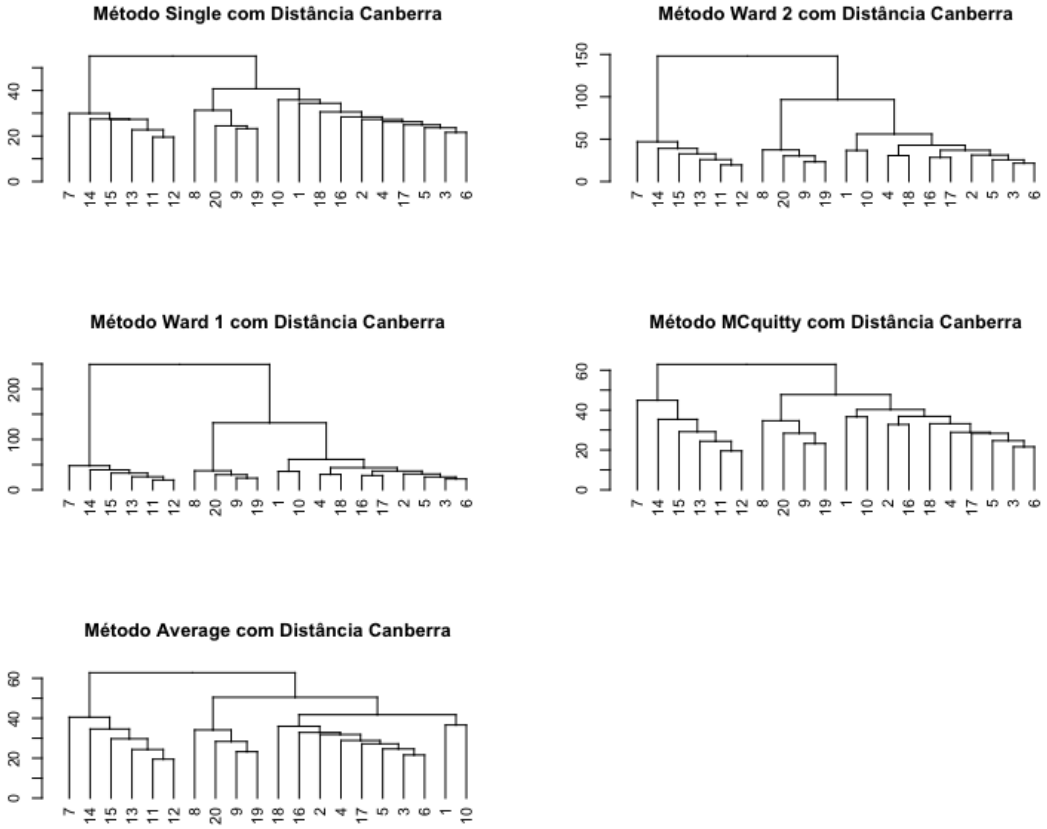
A.2 Indivíduos: 15 e 11 a 14

Figura A.5: Método Complete e Ward 2



A.3 Indivíduos: 7 11 12 13 14 15

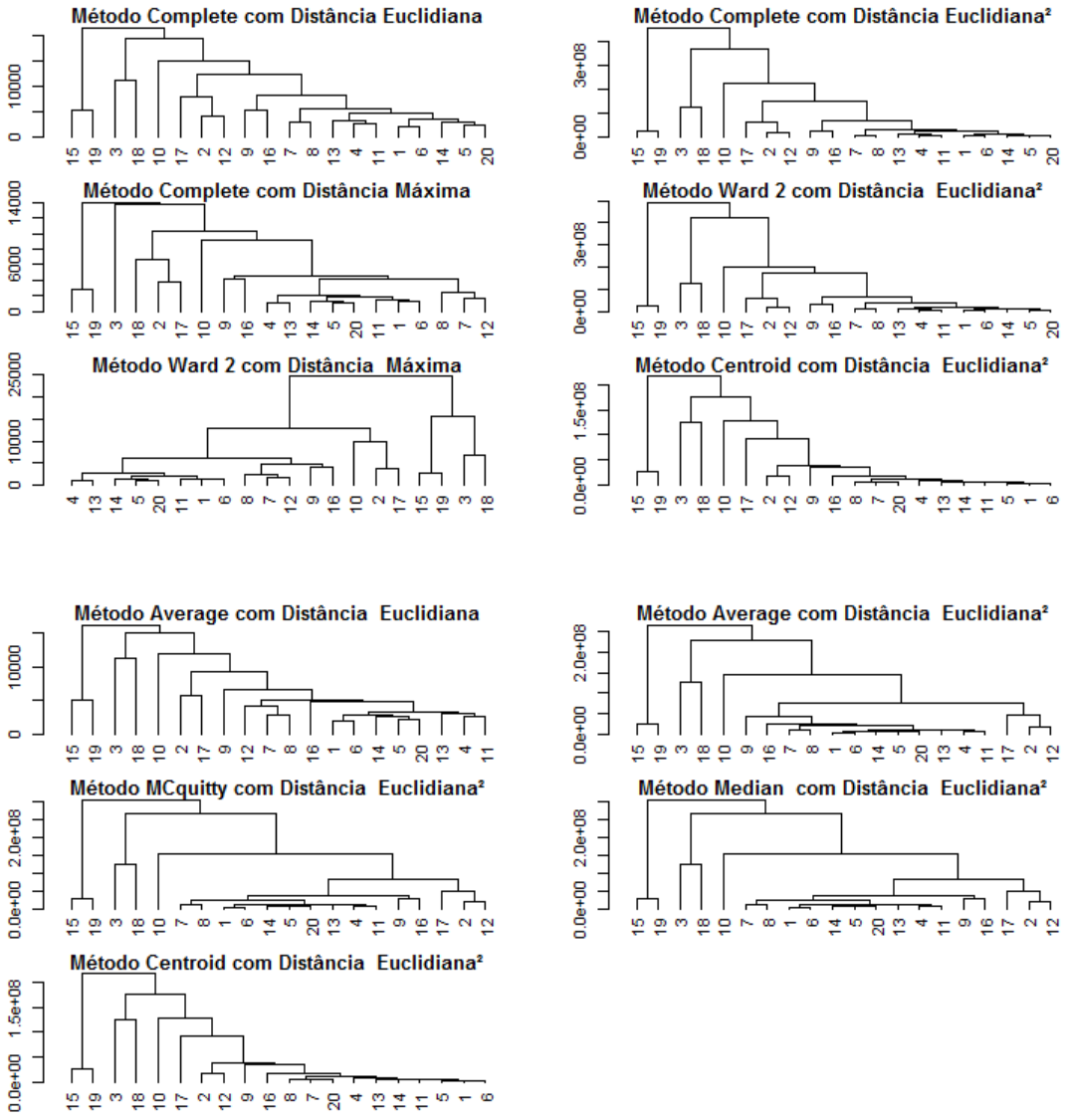
Figura A.6: Distância Canberra



Apêndice B

SALDI - Dendogramas

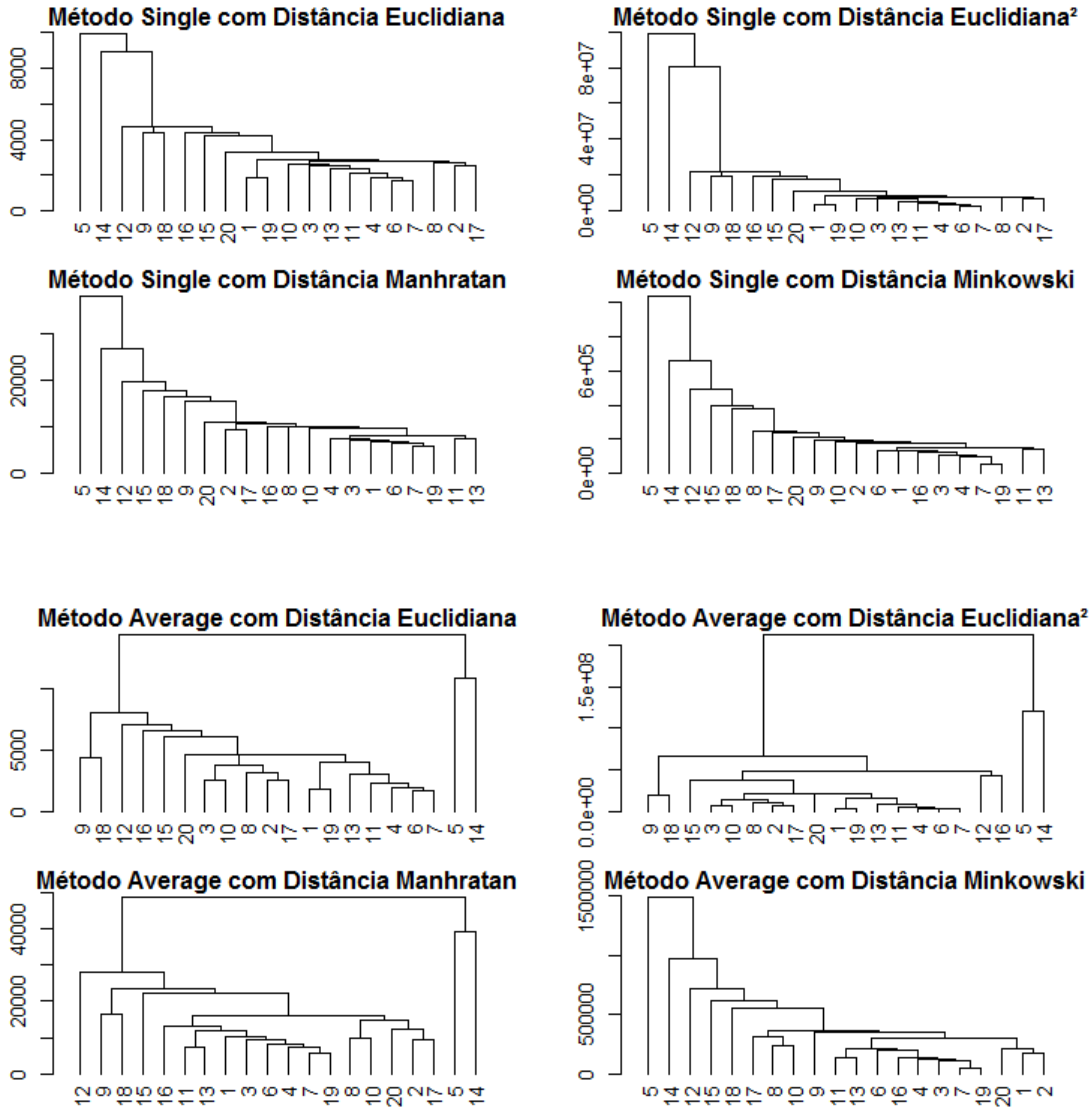
B.1 Indivíduos: [15,19] e [3,18]

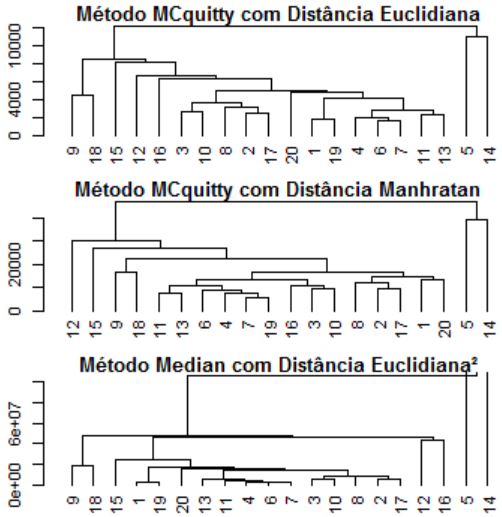


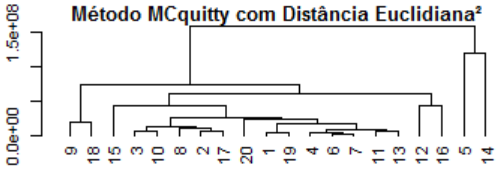
Apêndice C

LDI - Dendogramas

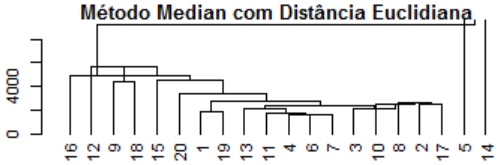
C.1 Indivíduos: 5 e 14



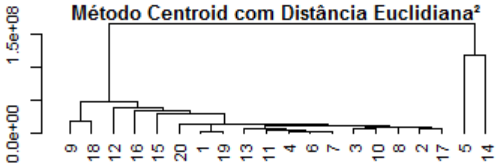

Método MCquitty com Distância Euclidiana
0
4000
10000
9 18 15 12 16 3 10 8 2 17 20 1 19 4 6 7 11 13 5 14


Método MCquitty com Distância Euclidiana²
0.0e+00
1.5e+08
9 18 15 3 10 8 2 17 20 1 19 4 6 7 11 13 12 16 5 14

Método MCquitty com Distância Manhratan
0
20000
12 15 9 18 11 13 6 4 7 19 16 3 10 8 2 17 1 20 5 14


Método Median com Distância Euclidiana
0
4000
16 12 9 18 15 20 1 19 13 11 4 6 7 3 10 8 2 17 5 14

Método Median com Distância Euclidiana²
0e+00
6e+07
9 18 15 1 19 20 13 11 4 6 7 3 10 8 2 17 12 16 5 14


Método Centroid com Distância Euclidiana²
0.0e+00
1.5e+08
9 18 12 16 15 20 1 19 13 11 4 6 7 3 10 8 2 17 5 14

Apêndice D

Programação R - Dados Binários

Programação em R para medidas de similaridades para dados binários

```
# Autor: Jose Luis Vicente Villardon
# Dpto. de Estadistica
# Universidad de Salamanca
# Revisado: Noviembre/2013
BinaryDistances<-function(x,y=NULL,coefficient="Simple_Matching",
transformation="sqrt(1-S)") {
if (!is.matrix(x)) stop("Input must be a matrix")
#if (!CheckBinaryMatrix(x))
stop("Input must be a binary matrix (with 0 or 1 values)")
coefficients = c("Kulezynski", "Russell_and_Rao", "Jaccard",
 "Simple_Matching", "Anderberg", "Rogers_and_Tanimoto",
  "Sorensen_Dice_and_Czekanowski", "Sneath_and_Sokal", "Hamman",
  "Kulezynski2", "Anderberg2", "Ochiai", "S13","Pearson_phi",
  "Yule","Sorensen", "Dice")
if (is.numeric(coefficient)) coefficient=coefficients[coefficient]
if (is.null(y)) y=x
a = y %*% t(x)
b = y %*% t(1 - x)
c = (1 - y) %*% t(x)
d = (1 - y) %*% t(1 - x)
switch(coefficient, Kulezynski.1 = {
sim = a/(b + c)
}, Russell_and_Rao = {
```

```
sim = a/(a + b + c+d)
}, Jaccard = {
sim = a/(a + b + c)
}, Simple_Matching = {
sim = (a + d)/(a + b + c + d)
}, Anderberg = {
sim = a/(a + 2 * (b + c))
}, Rogers_and_Tanimoto = {
sim = (a + d)/(a + 2 * (b + c) + d)
}, Sorensen_Dice_and_Czekanowski = {
sim = a/(a + 0.5 * (b + c))
}, Sneath_and_Sokal = {
sim = (a + d)/(a + 0.5 * (b + c) + d)
}, Hamman = {
sim = (a - (b + c) + d)/(a + b + c + d)
}, Kulezynski.2 = {
sim = 0.5 * ((a/(a + b)) + (a/(a + c)))
}, Anderberg2 = {
sim = 0.25 * (a/(a + b) + a/(a + c) + d/(c + d) + d/(b + d))
}, Ochiai = {
sim = a/sqrt((a + b) * (a + c))
}, S13 = {
sim = (a * d)/sqrt((a + b) * (a + c) * (d + b) * (d + c))
}, Pearson_phi = {
sim = (a * d - b * c)/sqrt((a + b) * (a + c) * (d + b) * (d + c))
}, Yule = {
sim = (a * d - b * c)/(a * d + b * c)
}, Sorensen = {
sim = (2*a)/(2* a  + b + c)
}, Dice = {
sim = (2*a)/(2* a  + b + c)
})
```

```
transformations= c("Identity", "1-S", "sqrt(1-S)", "-log(s)",
"1/S-1", "sqrt(2(1-S))", "1-(S+1)/2", "1-abs(S)", "1/(S+1)")
if (is.numeric(transformation)) transformation=transformations[transformation]

switch(transformation, `Identity` = {
dis=sim
}, `Identity` = {
dis=sim
}, `1-S` = {
dis=1-sim
}, `sqrt(1-S)` = {
dis = sqrt(1 - sim)
}, `-log(s)` = {
dis=-1*log(sim)
}, `1/S-1` = {
dis=1/sim -1
}, `sqrt(2(1-S))` = {
dis== sqrt(2*(1 - sim))
}, `1-(S+1)/2` = {
dis=1-(sim+1)/2
}, `1-abs(S)` = {
dis=1-abs(sim)
}, `1/(S+1)` = {
dis=1/(sim)+1
})

return(dis)
}
```

Apêndice E

Questionário

1. Qual a sua idade?
2. Possui dieta específica? (ex: vegetariano, vegano)
3. É fumante? Se sim, quantos cigarros por dia?
4. Faz uso frequente de algum medicamento? Se sim, qual ou quais?
5. Utilizou algum medicamento nas últimas 24 horas? Se sim, qual ou quais?
6. Tomou banho pela manhã (antes de doar suas impressões)
7. Utilizou cosméticos como xampu, sabonetes, cremes, protetor solar e/ou maquiagens nas horas que antecederam o exame? Se sim, quais?
8. Possui o hábito de lavar a louça e/ou realizar a faxina em casa sem o uso de luvas?
9. Ingeriu café (bebida) pela manhã (antes de doar suas impressões)?

Referências Bibliográficas

Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. In: *International Conference on Database Theory. Springer Berlin Heidelberg*, pages 420–434. Springer.

Allen, D. A. N. & Goldstein, G. A. (2013). *Cluster analysis in neuropsychological research*. Springer.

Andeberg, M. R. (1973). Cluster analysis for applications academic press. *New York*.

Archer, N. E., Charles, Y., Elliott, J. A., & Jickells, S. (2005). Changes in the lipid composition of latent fingerprint residue with time after deposition on a surface. *Forensic Science International*, 154(2):224–239.

Baker, F. B. (1974). Stability of two hierarchical grouping techniques case i: Sensitivity to data errors. *Journal of the American Statistical Association*, 69(346):440–445.

Battaglia, O. R., Paola, B., & Fazio, C. (2016). A new approach to investigate students. behavior by using cluster analysis as an unsupervised methodology in the field of education. *Applied Mathematics*, 7(15):1649.

Carlini-Garcia, L. (1998). Estudo da estrutura ge nética populacional através de marcadores moleculares.

Carvalho, A. X. Y., Albuquerque, P. H. M., de Almeida Junior, G. R., Guimarães, R. D., & Laureto, C. R. (2009). Clusterização hierárquica espacial com atributos binários. Technical report, Texto para Discussão, Instituto de Pesquisa Econômica Aplicada (IPEA).

Cavalcante, D. M. C. & Vasconcelos, P. H. (2018). Análise de impressões digitais utilizando modelos mistos. *Departamento de Estatística - Universidade de Brasília*.

Chemello, E. (2006). Ciência forense: impressões digitais. *Química Virtual*, pages 1–11.

Clifford, H. T., Stephenson, W., et al. (1975). *An introduction to numerical classification*, volume 240. Academic Press New York.

Dikshitulu, Y., Prasad, L., Pal, J., & Rao, C. (1986). Aging studies on fingerprint residues using thin-layer and high performance liquid chromatography. *Forensic science international*, 31(4):261–266.

Doni, M. V. (2004). Análise de cluster: métodos hierárquicos e de particionamento. *Universidade Presbiteriana Mackenzie*.

Dunn, G. & Everitt, B. S. (1980). *An introduction to mathematical taxonomy*. New York:Cambridge University Press.

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). Hierarchical clustering. *Cluster Analysis, 5th Edition*, pages 71–110.

Farris, J. S. (1969). On the cophenetic correlation coefficient. *Systematic Zoology*, 18(3):279–285.

Gan, G., Ma, C., & Wu, J. (2007). *Data clustering: theory, algorithms, and applications*, volume 20. Siam.

Girod, A., Ramotowski, R., & Weyermann, C. (2012). Composition of fingermark residue: a qualitative and quantitative review. *Forensic science international*, 223(1):10–24.

Hair, Black, B. A. T. (2009). *Análise Multivariada de dados*. Bookman Editora.

Hubert, L. (1974). Approximate evaluation techniques for the single-link and complete-link hierarchical clustering procedures. *Journal of the American Statistical Association*, 69(347):698–704.

Jarccard, P. (1908). Nouvelles recherches sur la distribution floral. *Bull. Soc. Vard. Sci. Nat*, 44:223–270.

Kaplan-Sandquist, K., LeBeau, M. A., & Miller, M. L. (2014). Chemical analysis of pharmaceuticals and explosives in fingermarks using matrix-assisted laser desorption ionization/time-of-flight mass spectrometry. *Forensic science international*, 235:68–77.

Kaufman, L. & Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons.

Kuiper, F. K. & Fisher, L. (1975). 391: A monte carlo comparison of six clustering procedures. *Biometrics*, pages 777–783.

Kurczynski, T. (1970). Generalized distance and discrete variables. *Biometrics*, pages 525–534.

Lance, G. N. & Williams, W. T. (1966). Computer programs for hierarchical polythetic classification ("similarity analyses"). *The Computer Journal*, 9(1):60–64.

Lattin, J., Carroll, J. D., & Green, P. E. (2011). Análise de dados multivariados. *São Paulo: Cengage Learning*, 475.

Legendre, P. & Legendre, L. F. (1983). *Numerical ecology*. New York: Elsevier Scientific.

Legendre, P. & Legendre, L. F. (2012). *Numerical ecology*, volume 24. Elsevier.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

Malhotra, N. K. (2006). Questionnaire design and scale development. *The handbook of marketing research: Uses, misuses, and future advances*, pages 176–202.

Malhotra, N. K. (2012). *Pesquisa de marketing: uma orientação aplicada*. Bookman Editora.

Marques, M. D. (2017). Análise crítica da aderência das taxonomias industriais ã realidade da indústria de transformação brasileira.

Metz, J. (2006). *Interpretação de clusters gerados por algoritmos de clustering hierárquico. 2006. 126f.* PhD thesis, Tese de Doutorado. Instituto de Ciências Matemáticas e de Computação.Universidade de São Paulo.

Milligan, G. W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*, 45(3):325–342.

Mingoti, S. A. (2005). *Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada*. Editora UFMG.

Mountford, M. (1962). An index of similarity and its application to classification problems. *Progress in soil zoology*, pages 43–50.

Mountfort, K. A., Bronstein, H., Archer, N., & Jickells, S. M. (2007). Identification of oxidation products of squalene in solution and in latent fingerprints by esi-ms and lc/apci-ms. *Analytical Chemistry*, 79(7):2650–2657.

Pereira, J. R. G. et al. (1993). Um estudo sobre alguns métodos hierárquicos para análise de agrupamentos.

Rohlf, F. J. (1970). Adaptive hierarchical clustering schemes. 19(1):58–82.

Romesburg, H. (2004). Cluster analysis for researchers. lulu. *City*.

Saracli, S., Doğan, N., & Doğan, İ. (2013). Comparison of hierarchical cluster analysis methods by cophenetic correlation. *Journal of Inequalities and Applications*, 2013(1):203.

Sokal, R. & Sneath, P. (1963). Principles of numerical taxonomy. w.h. freeman and company. *San Francisco, CA*.

Sokal, Robert R, R. F. J. (1962). The comparison of dendrograms by objective methods. *Taxon*, 11(2):33–40.

Souza, A. M. & Vicini, L. (2005). Análise multivariada da teoria à prática. *Santa Maria: Departamento de Estatística UFSM*.

Stewart, J., Miller, M., Audo, C., & Stewart, G. (2012). Using cluster analysis to identify patterns in students, responses to contextually different conceptual problems. *Physical Review Special Topics-Physics Education Research*, 8(2):020112.

Tochetto, D. et al. (1999). Identificação humana.

Tryon, R. C. (1939). *Cluster analysis: Correlation profile and orthometric (factor) analysis for the isolation of unities in mind and personality*. Edwards brother, Incorporated, lithoprinters and publishers.

Wanner, K. T. & Höfner, G. (2007). *Mass spectrometry in medicinal chemistry: applications in drug discovery*. John Wiley & Sons.

Wolstenholme, R., Bradshaw, R., Clench, M. R., & Francese, S. (2009). Study of latent fingermarks by matrix-assisted laser desorption/ionisation mass spectrometry imaging of endogenous lipids. *Rapid communications in mass spectrometry*, 23(19):3031–3039.

Xu, S., Li, Y., Zou, H., Qiu, J., Guo, Z., & Guo, B. (2003). Carbon nanotubes as assisted matrix for laser desorption/ionization time-of-flight mass spectrometry. *Analytical Chemistry*, 75(22):6191–6195.