ESDRAS ADRIANO BARBOSA DOS SANTOS

ESTUDO DE RESULTADOS DO ESPECTRO MULTIFRACTAL DA RETINA HUMANA, COMO MEDIDA DE CLASSIFICAÇÃO: UMA APLICAÇÃO DE ANÁLISE DE AGRUPAMENTO

RECIFE-PE - FEV/2008

UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM BIOMETRIA E ESTATÍSTICA APLICADA

ESTUDO DE RESULTADOS DO ESPECTRO MULTIFRACTAL DA RETINA HUMANA, COMO MEDIDA DE CLASSIFICAÇÃO: UMA APLICAÇÃO DE ANÁLISE DE AGRUPAMENTO

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

Área de Concentração: Desenvolvimento de Métodos Estatísticos e Computacionais

Orientadora: Profa. Dra. Tatijana Stošić

Co-orientadores: Prof. Dr. Borko D. Stošić

Profa. Dra. Laélia P. B. Campos dos Santos

RECIFE-PE - FEV/2008.

FICHA CATALOGRÁFICA

S237e Santos, Esdras Adriano Barbosa dos
Estudo de resultados do espectro multifractal da retina humana : uma aplicação de análise de agrupamento / Esdras Adriano Barbosa dos Santos. -- 2008.
67 f. : il.

Orientadora : Tatijana Stosic
Dissertação (Mestrado em Biometria e Estatística Aplicada) – Universidade Federal Rural de Pernambuco. Departamento de Estatística e Informática.
Inclui bibliografia.

CDD 574. 018 2

1. Análise de agrupamento
2. Multifractal
3. Retina
4. Vascularização
I. Stosic, Tatijana
II. Título

UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM BIOMETRIA E ESTATÍSTICA APLICADA

ESTUDO DE RESULTADOS DO ESPECTRO MULTIFRACTAL DA RETINA HUMANA, COMO MEDIDA DE CLASSIFICAÇÃO: UMA APLICAÇÃO DE ANÁLISE DE AGRUPAMENTO

ESDRAS ADRIANO BARBOSA DOS SANTOS

Dissertação julgada adequada para obtenção do título de mestre em Biometria e Estatística Aplicada, defendida e aprovada com louvor por unanimidade em 25/02/2008 pela Comissão Examinadora.

Orientador:

Prof. Dra. Tatijana Stošić
Universidade Federal Rural de Pernambuco

Banca Examinadora:

Prof. Dr. Borko D. Stošić
Universidade Federal Rural de Pernambuco

Prof. Dr. Getúlio José Amorim do Amaral
Universidade Federal de Pernambuco

Prof. Dr. Francisco José de Azevêdo Cysneiros
Universidade Federal de Pernambuco

*À minha esposa **Laélia** e ao meu filho **Samuel**, com muito amor.*

Agradecimentos

Agradeço primeiramente a DEUS, por ter me concedido a graça de nascer do amor dos meus pais e ao longo de minha vida ter permitido chegar até aqui.

Aos meus pais Jairo Santos e Berenice Santos pelo amor, carinho e dedicação na formação de meu caráter, deixando em meu coração mais que simples ensinamentos, como também o amor a DEUS e seus preceitos, esses sempre estarão em meu coração.

Aos meus irmãos Isaías Magno pela sua sinceridade, Douglas Rafael pelo seu forte caráter e Débora Caroline pela sua maneira silenciosa de me mostrar os seus sentimentos, obrigado manos.

À minha esposa Laélia Campos e ao meu filho Samuel que são a razão da minha persistência e luta.

À minha Orientadora Tatijana Stošić pela sua paciência e motivação no decorrer da montagem deste trabalho.

Ao coordenador Prof. Eufrázio de Souza Santos pelo seu empenho em dar aos alunos da Pós-Graduação em Biometria e Estatística Aplicada as melhores condições de aprendizado e experiência na pesquisa.

Ao Prof. Gauss Moutinho Cordeiro pela maneira magistral de evocar de cada um dos seus alunos o melhor, e paralelamente ao desenvolvimento intelectual dos mesmos.

À Profa. Cláudia Regina não apenas pela sua orientação durante a graduação, mas principalmente por ter me inspirado a seguir sempre em frente, muito obrigado por tudo.

Ao Prof. Borko Stošić pela transmissão dos seus conhecimentos e principalmente pelo seu bombardeio de idéias quase que constantes.

Ao Secretário Marco Antônio dos Santos pela sua confiança e amizade durante todo o decorrer destes dois anos.

Ao meu amigo Lucas Gallindo por seu desprendimento no apoio durante todo o decorrer da nossa vida acadêmica.

Ao meu amigo Luiz Henrique por sua iniciativa de ajudar a qualquer um no que lhe é

possível fazer e pelo seu ponto de vista sempre crítico.

Ao meu amigo Luciano Sousa pela amizade incondicional, companheirismo e luta, lado a lado em muitos momentos de nossas vidas, muito obrigado Amigo.

Ao meu amigo Moacy Cabral pela sua obstinação contagiante em sempre estar aprendendo novas teorias.

À amiga Rosângela do Nascimento pelos seus conselhos e experiência passados durante nossa convivência.

Ao meu amigo Cândido Dantas que me ensinou mais do que simples conhecimentos filosóficos, mostrando-me humanidade.

Aos amigos e companheiros de estudo durante minha graduação em Estatística, vocês marcaram minha vida de uma forma que nunca vão ser esquecidos.

Aos amigos e companheiros de estudo do curso de mestrado em Biometria, pelos momentos de estudo e descontração que passamos juntos.

Aos professores e funcionários do Departamento de Estatística e Informática pela convivência agradável durante esses dois anos.

Ao Rev. Núzio Daniel pelas palavras de sabedoria ministradas ao meu coração.

A CAPES e ao CNPq pelo apoio financeiro e ao Programa de Pós-Graduação em Biometria e Estatística Aplicada pelo suporte logístico e intelectual.

A todos que de alguma forma deixaram em meu coração um pouco de si, cada um de vocês faz agora parte de minha vida, muito obrigado.

Resumo

A análise de imagens é freqüentemente praticada por oftalmologistas para diagnóstico de doenças. A inspeção da vascularização da retina pode revelar inícios de doenças como a retinopatia diabética. Desta forma, existem vários esforços para o desenvolvimento de métodos mais eficazes no diagnóstico destas doenças. A identificação de anormalidades requer uma trabalhosa inspeção de uma grande quantidade das imagens por especialistas. Assim sendo, há necessidade de desenvolvimento de softwares para o auxílio dos oftalmologistas na busca de uma diagnose mais rápida e mais precisa. O uso da dimensão fractal na busca de diferenciação entre retinas com e sem patologias é mais um dos ramos de pesquisa realizados nesta área. Recentemente, foi mostrado que a retina humana não é um fractal simples, mas um multifractal, caracterizado pelo espectro multifractal não trivial. Neste trabalho, foram aplicados métodos de agrupamento nos resultados da análise multifractal para verificar a sensibilidade desta análise na diferenciação entre casos patológicos e casos normais da retina humana. As variáveis usadas são os elementos de espectro multifractal $f(\alpha)$ e dimensões generalizadas $D(q)$, das quais foram escolhidos três conjuntos distintos. Os métodos de agrupamento usados para análise foram o método de Ward, K-médias, PAM e Fuzzy c-means. Como medida para a validação dos grupos obtidos, foi usada a correlação cofenética para o método de Ward e gráficos de silhueta e silhueta média para os métodos K-médias, PAM e Fuzzy c-means. Os resultados mostraram que, para imagens esqueletonizadas, 70-80% das retinas patológicas (dependendo do método e do conjunto de variáveis usadas) foram agrupadas corretamente, enquanto que para as imagens segmentadas originais, o agrupamento não apresentou resultados tão satisfatórios. Este fato indica que a largura dos vasos apresenta menor influência para as conclusões da análise atual, em comparação com o comprimento dos vasos e suas ramificações. Diante disso, é possível concluir que a análise multifractal, aliada ao pré-processamento adequado das imagens e a escolha adequada das variáveis, pode ser utilizada para detecção de casos patológicos ou para o pré-diagnóstico.

Palavras-chave: Análise de Agrupamento, Multifractal, Retina, Vascularização.

Abstract

Image analysis is frequently used by ophthalmologists as part of the diagnostic procedure. Inspection of the vascular structure of the retina may reveal early stages of pathologies such as diabetic retinopathy, and there have been various efforts to develop more efficient methods for diagnosing such diseases. Currently, identification of abnormalities requires a laborious inspection of a large number of images from the part of specialists, and there is a necessity of automating this process to provide auxiliary diagnostic tools of high speed and precision. One of the lines of research conducted in the direction of differentiating between healthy and pathological retinal images uses the concept of fractal dimension. Recently it was shown that the vascular structure of the human retina is not a simple fractal, but rather a multifractal, characterized by a non trivial multifractal spectrum. In this work, multivariate clustering methods are applied to the results of the multifractal analysis, in order to establish the sensitivity of this analysis, and its ability to differentiate between the normal and pathological cases of the human retina. The variables used for this purpose are the elements of the multifractal spectrum $f(\alpha)$ and the generalized dimension $D(q)$, from which three distinct sets of variables were chosen. The clustering methods used for this study are the Ward method, K-means, PAM and Fuzzy C-means. As a measure of validation of the obtained groups the cophenetic correlation was used for the Ward method, and the silhouette graphs for K-means, PAM and Fuzzy C-means. The results show that for the skeletonized images 70-80% of the pathological images were correctly classified (depending on the method and the variables used), while for the original segmented images clustering produces worse results. This fact indicates that the width of the vessels exerts less influence on the conclusions of the current analysis in comparison with the length distribution and the ramification structure. Thus, we may conclude that the multifractal analysis, with adequate pre-processing of the images and choice of variables, can be used for detection of pathological cases, as part of the pre-diagnostic procedure.

Keywords: Clustering Methods, Multifractal, Retina, Vascular Structure.

Lista de Figuras

Lista de Tabelas

Sumário

1 Introdução

A imagem da retina representa uma estrutura complexa composta de vários elementos como vasos sangüíneos, fóvea e nervo óptico. Os vasos sangüíneos se apresentam na imagem como uma rede ramificada que cresce de um ponto (nervo óptico) e segue por toda a retina. As alterações vasculares da retina são associadas a doenças como diabetes, hipertensão arterial, arteriosclerose, etc., que podem comprometer a visão do paciente e em casos graves levar a cegueira. Algumas destas doenças como retinopatia diabética não causam sintomas iniciais. Quando os pacientes apresentam uma baixa na visão é sinal que a doença já está avançada com danos irreversíveis aos vasos sangüíneos. Isso torna importante a detecção precoce de alterações vasculares na retina, que pode ser feita examinando o fundo de olho por meio de fundus câmera ou angiografia.

Os avanços dos sistemas computacionais facilitam o desenvolvimento e implementação de métodos (baseados em uso de modelos matemáticos, físicos e estatísticos) que ajudam a analisar as imagens médicas, com o objetivo de aumentar a qualidade do processo diagnóstico e tratamento de doenças. Para aplicar os métodos analíticos e numéricos de análise sobre os vasos sangüíneos da retina, a estrutura dos vasos deve ser extraída da imagem (segmentada manualmente utilizando ferramentas gráficas, ou automaticamente usando processamento digital de imagens). Após segmentação, a imagem do sistema de vasos pode ser tratada como um objeto geométrico em busca de propriedades relevantes para identificação de casos patológicos.

Durante a década passada foram feitas várias tentativas da utilização da dimensão fractal como quantificador das propriedades geométricas do sistema dos vasos sangüíneos da retina humana. Os resultados ainda não são conclusivos principalmente porque não existe um método eficiente de segmentação automática dos vasos, a partir de imagens obtidas pelo fundus câmera ou angiografia (MASTERS, 2004). Nessa linha de pesquisa, tem-se recentemente a publicação de um trabalho realizado por Stosic e Stosic (2006), no qual foi aplicada a análise multifractal aos vasos da retina para casos patológicos e não patológicos, com o objetivo de estabelecer se estes objetos representam fractais regulares,

ou devem ser tratados como multifractais. Os resultados de Stosic e Stosic (2006) mostraram o comportamento multifractal em ambos os casos. As diferenças entre as dimensões fractais generalizadas apresentadas por Stosic e Stosic (2006), bem como a forma das curvas do espectro multifractal, indicam possibilidade de uso dessa análise na detecção de casos patológicos.

Nesse contexto, o objetivo dessa dissertação foi aplicar a análise de agrupamento sobre os resultados da análise multifractal encontrada, com o intuito de verificar a viabilidade do uso desta análise para diferenciar casos patológicos de casos normais.

No Capitulo 2 do presente trabalho encontra-se a revisão de literatura sobre fractais e multifractais, bem como análise multifractal cujos resultados de Stosic e Stosic (2006) foram usados como variáveis para aplicação dos métodos de agrupamento. Também é apresentada uma breve revisão referente aos métodos de agrupamentos da análise multivariada que serão usados neste estudo. No Capitulo 3 são descritos os materiais e métodos referentes aos bancos de dados usados, bem como a aplicação dos métodos de Ward, K-médias, PAM e Fuzzy c-means ao referido banco. No Capítulo 4 encontram-se os resultados encontrados após aplicação dos métodos e a discussão dos mesmos. O trabalho é finalizado no Capitulo 5 com a apresentação das conclusões e perspectivas futuras.

2 Revisão de Literatura

2.1 O Olho

O olho é um órgão que permite detectar a luz e transformar essa percepção em impulsos elétricos. O globo ocular tem este nome por ter a forma de um globo, que por sua vez fica acondicionado dentro de uma cavidade óssea chamada órbita, e protegido pelas pálpebras. O globo ocular possui em seu exterior seis músculos que são responsáveis pelos movimentos oculares, e também três camadas concêntricas aderidas entre si com a função de visão, nutrição e proteção. A camada externa é constituída pela córnea e esclera, servindo para proteção. A camada média é dividida em duas partes: a anterior contendo a íris e o corpo ciliar e a posterior contendo a coróide. A camada interna é constituída pela retina que é a parte nervosa, sendo composta de células nervosas que leva a imagem através do nervo óptico para que o cérebro a interprete (GUYTON, 1993).

2.1.1 A Retina

A retina é a parte do olho sensível a luz, sendo composta de células nervosas fotorreceptoras: os cones, responsáveis pela visão colorida e os bastonetes, responsáveis pela visão no escuro. Quando os cones e os bastonetes são excitados, os sinais são transmitidos ao longo dos neurônios, na própria retina, e finalmente chegam às fibras do nervo óptico e córtex cerebral. As camadas da retina estão dispostas da seguinte forma: camada pigmentada, camada de cones e bastonetes, membrana limitante externa, camada nuclear externa, contendo os corpos celulares dos cones e bastonetes, camada plexiforme externa, camada nuclear interna, camada plexiforme interna, camada ganglionar, camada de fibras do nervo óptico e finalmente a membrana limitante interna. Após a luz atravessar o sistema de lentes do olho, a mesma penetra na retina por sua superfície interna atravessando as células ganglionares e as camadas de cones e bastonetes localizados em toda

superfície externa da retina. O suprimento sangüíneo que nutre as camadas internas da retina é derivado da artéria retiniana central, que penetra no globo ocular juntamente com o nervo óptico, e depois divide-se para irrigar toda a superfície interna da retina. Assim, em um grau bastante alto, a retina possui seu próprio suprimento sangüíneo, independente das outras estruturas do olho. Entretanto, a superfície externa da retina é aderente à coróide, que é um tecido muito vascularizado entre a retina e a esclerótica. As camadas externas da retina, incluindo os segmentos externos dos bastonetes e cones, dependem em grande parte da difusão, a partir dos vasos coróides, para nutrição e principalmente para o oxigênio (GUYTON, 1993).

2.2 Fractais & Multifractais

2.2.1 Fractais

O conceito de geometria fractal foi introduzido por Benoit Mandelbrot (MANDELBROT, 1982) para descrever os sistemas naturais formados pelos processos estocásticos que são longe do equilíbrio. Como exemplos desses sistemas podem-se citar as árvores ramificadas, linhas costeiras, nuvens, polímeros, estruturas cardiopulmonares (rede arterial, árvore traqueobronquial), etc. (MANDELBROT, 1982; VICSEK, 1993; FEDER, 1988; BASSINGTHWAIGHTE et al., 1994). A diferença existente entre a geometria fractal e a geometria euclidiana é que fractais possuem dimensão não inteira (fracionária) e propriedade de auto-similaridade (pedaços de objeto que se assemelham ao objeto todo).

Os exemplos citados representam fractais estocásticos e possuem a propriedade de auto-similaridade em sentido estatístico dentro de um intervalo de escala $s \leq \ell \leq S$ onde s e S representam os limites da escala (s é proporcional à distância entre as partículas do sistema e S é proporcional à dimensão linear do sistema). Dentro desse intervalo de escalas, o volume da região com dimensão linear R é dado por

$$V(R) \propto R^{d_f} \tag{2.1}$$

onde $d_f < D$ é a dimensão fractal do sistema e D é a dimensão euclidiana do espaço, no qual o fractal está incorporado (VICSEK, 1993).

Por outro lado, é possível construir os fractais determinísticos, objetos geométricos que possuem a propriedade de auto-similaridade em todas as escalas. O processo de construção desses fractais consiste em um procedimento iterativo onde os segmentes do objeto

(e.g. triângulos, quadrados) são substituídos por uma estrutura característica (gerador) para cada tipo de fractal (VICSEK, 1993). Para fractais determinísticos, o procedimento

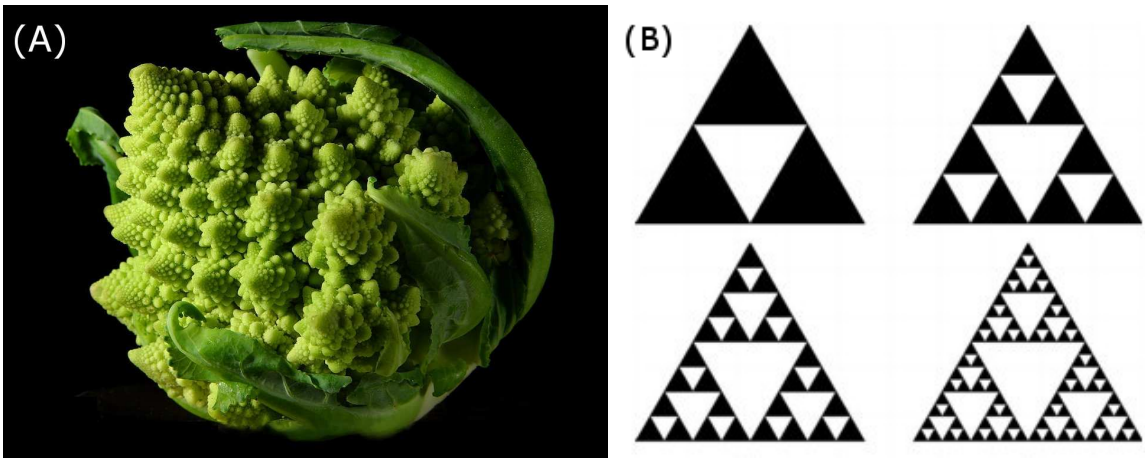

Figura 2.1: Exemplos de fractais: (A) fractal estocástico. (B) fractal determinístico Sierpinski Gasket.

de cálculo da dimensão fractal é a seguinte: se $N(\ell)$ é o número de unidades da estrutura em escala ℓ, a diminuição da escala b vezes resulta em um novo número de unidades da estrutura

$$N\left(\frac{\ell}{b}\right) = N(\ell)\, b^{d_f} \tag{2.2}$$

A dimensão fractal d_f é dada pela expressão

$$d_f = \frac{\log \frac{N\left(\frac{\ell}{b}\right)}{N(\ell)}}{\log b} \tag{2.3}$$

válida para todas as escalas. No caso de Sierpinski Gasket exposto na figura (2.1b), a diminuição da escala duas vezes resulta do aumento de triângulos três vezes e a dimensão fractal calculada pela equação (2.3) tem valor

$$d_f = \frac{\log 3}{\log 2} = 1,585 \tag{2.4}$$

Para fractais estocásticos, existem vários métodos para o cálculo de dimensão fractal, como contagem de caixas (“box counting”), método massa-raio (“mass-radius method”), e método de correlação densidade-densidade (“density-density correlation function method”). O método contagem de caixas consiste em cobrir a estrutura com uma grade de caixas com arestas de tamanho r e contar o número $N(r)$ de caixas que contém pelo menos uma partícula do sistema. O número de caixas $N(r)$ depende de r segundo a relação $N(r) \propto r^{-d_f}$. Repetindo este valor para vários valores de r, a dimensão fractal pode ser calculada usando o coeficiente angular da reta obtida pela regressão do gráfico $\log N(r)$

versus $\log r$ (VICSEK, 1993). Outro método muito usado é massa-raio, onde se constrói uma seqüência de circunferências de raios crescentes, centralizadas no centro da massa do sistema. O número de partículas (*massa*) M dentro da circunferência de raio r, depende de tamanho r segundo a relação $M \propto r^{-d_f}$. A dimensão fractal d_f é calculada como o coeficiente angular da reta de regressão do gráfico $\log M$ versus $\log r$. O método da função de correlação consiste em calcular a função de correlação

$$c(r) = \frac{1}{N} \sum_{r'} \rho(r + r')\rho(r) \tag{2.5}$$

onde N é o número de partículas do sistema e $\rho(r)$ densidade local ($\rho(r) = 1$ se existe uma partícula em posição r, e $\rho(r) = 0$ caso contrário). Para objetos fractais, $c(r)$ depende de r segundo a relação $c(r) \sim r^{d_f - D}$, onde d_f é a dimensão fractal do sistema e D a dimensão euclidiana. A dimensão fractal é calculada usando o coeficiente angular da reta de regressão do gráfico $\log c(r)$ versus $\log r$.

2.2.2 Multifractais

Em contraste com fractais simples (ou monofractais), os multifractais são caracterizados por uma hierarquia de expoentes (VICSEK, 1993; FEDER, 1988). Mais precisamente, multifractais podem ser vistos como um entrelace de simples fractais. A palavra "hierarquia" aqui se refere aos diferentes membros deste entrelace, os quais, têm dimensões fractais distintas. Se essa propriedade é ignorada e o objeto multifractal é tratado como um fractal simples (monofractal), os métodos tradicionais do cálculo de dimensão fractal resultam em um valor intermediário. A dimensão de capacidade (calculada usando o método contagem de caixas) tem valor maior do que a dimensão de correlação, calculada usando o método de função e correlação. Assim, para testar se um objeto geométrico deve ser tratado como um multifractal, o primeiro passo é calcular a dimensão de capacidade e a dimensão de correlação, e se essas dimensões forem distintas, deve-se realizar a análise multifractal usando o cálculo de espetro multifractal para descrever as propriedades geométricas do objeto. Um dos exemplos mais investigados de multifractalidade é a distribuição de probabilidade de crescimento durante o processo de **"Diffusion Limited Aggregation" (DLA)** (HAYAKAWA et al., 1987; NITTMANN et al., 1987). O modelo DLA foi introduzido por Witten e Sander (1981) para descrever processos como a deposição dos íons nos eletrodos e descargas elétricas (raios) (VICSEK, 1993).

Para a montagem do DLA, coloca-se primeiro uma partícula (semente) na grade. Uma partícula é lançada longe da semente e se movimenta aleatoriamente (caminho aleatório,

ou processo de difusão) até a mesma se posicionar junto da partícula semente, e fica incluída como parte do cluster. Uma nova partícula é lançada longe da semente até ficar absorvida pelo cluster, e assim por diante. Esse processo é repetido até o cluster atingir um tamanho pré-definido, resultando em uma estrutura ramificada mostrada na figura 2.2

Figura 2.2: Exemplo de um DLA - Diffusion Limited Aggregation em 3 dimensões

Análise Multifractal

No caso de um multifractal geométrico, analisa-se o número de partículas dentro de uma região (VICSEK et al., 1990). O procedimento do cálculo da dimensão multifractal generalizada consiste em cobrir a estrutura analisada com caixas de aresta de tamanho ℓ, variando posteriormente os valores de ℓ, e registrando os valores de M_i dentro de i-ésima caixa, sendo M_o o número total de partículas do sistema. A dimensão generalizada D_q para distribuição de massa é definida por:

$$\sum_i \left(\frac{M_i}{M_0}\right)^q \sim \left(\frac{\ell}{L}\right)^{(q-1)D_q} \tag{2.6}$$

onde q é uma variável contínua que torna possível enfatizar as propriedades fractais em diferentes escalas. As dimensões generalizadas D_0 , D_1 e D_2 representam a dimensão de capacidade, dimensão de informação e dimensão de correlação, respectivamente. Finalmente, $D_{-\infty}$ e D_{∞} representam os limites de espetro de dimensões generalizadas, onde a

medida de interesse é a mais diluída e mais densa, respectivamente. Para monofractais, todas as dimensões generalizadas são iguais, dando um único valor de dimensão fractal.

Para monofractais, o espetro $D(q)$ é constante, enquanto para multifractais $D(q)$ representa uma função monótona decrescente. A aplicação direta da equação 2.6 é difícil porque para $q < 0$ as caixas que contêm um pequeno número de partículas têm uma grande contribuição na soma do lado esquerdo de 2.6. Para evitar este problema, usa-se o método Sand Box generalizado (VICSEK et al., 1990; TÉL et al., 1989). Esse método foi usado com sucesso para demonstrar a multifractalidade geométrica do DLA (VICSEK et al., 1990). O procedimento consiste em uma seleção aleatória de N pontos que pertencem à estrutura com total de M_0 pontos e contando para cada um desses pontos escolhidos o número de partículas $M_i(R)$ que pertencem à estrutura dentro das caixas de dimensão linear crescente R, centralizadas nas partículas escolhidas. A quantidade $M_i(R)/M_0$ pode ser entendida como uma estimação empírica da densidade de probabilidade espacial de encontrar a partícula pertencente à estrutura na posição correspondente ao ponto escolhido (que aumenta com a densidade de uma região analisada, sendo maior ou igual a zero, e sua soma tem valor unitário para o conjunto de caixas não sobrepostas de dimensão linear R que cobrem completamente a imagem). O lado esquerdo da equação 2.6 pode ser interpretado como a média da quantidade $(M_i(R)/M)^{q-1}$ de acordo com a distribuição espacial $(M_i(R)/M_0)$. Como no método atual, os centros de caixas de tamanho R são escolhidos aleatoriamente, a média pode ser calculada para o conjunto escolhido, e pela equação 2.6 temos

$$\left\langle \left(\frac{M(R)}{M_0} \right)^{q-1} \right\rangle \sim \left(\frac{R}{L} \right)^{(q-1)D_q} \tag{2.7}$$

A equação 2.7 representa a síntese de um método generalizado de Sand Box (VICSEK et al., 1990; TÉL et al., 1989), amplamente aceito na literatura como análise de multifractalidade geométrica. A vantagem desse método é que as caixas são centralizadas na estrutura, então, por construção, não se encontram caixas com um número de partículas muito pequeno (ou nulo).

Para um tratamento alternativo da multifractalidade usa-se o espectro $f(\alpha)$ (VICSEK, 1993; FEDER, 1988; HALSEY et al., 1986) onde

$$N(\alpha) = L^{-f(\alpha)} \tag{2.8}$$

representa o número de caixas $N(\alpha)$ para as quais a probabilidade P_i de encontrar uma

partícula dentro de i-ésima região é regida pela lei de escala

$$P_i = L^{\alpha_i} \tag{2.9}$$

Sendo $f(\alpha)$ entendido como a dimensão fractal da união de regiões com singularidade entre α e $\alpha + d\alpha$, α variando entre $[-\infty, \infty]$. A relação entre a função $D(q)$ e o espectro $f(\alpha)$ é feita via transformação de Legendre

$$f(\alpha(q)) = q\alpha(q) - \tau(\alpha) \tag{2.10}$$

em que

$$\alpha(q) = \frac{d\tau(\alpha)}{dq} \tag{2.11}$$

e

$$\tau(\alpha) \equiv (q-1)D_q \tag{2.12}$$

é o expoente de correlação de massa da q-ésima ordem. No caso de monofractal, a dimensão fractal não depende de q $\left(D_q \equiv D\right)$, e usando as equações 2.11e 2.12 tem-se $f(\alpha) = D$ e o espetro $f(\alpha)$ consiste um único ponto, onde $f(\alpha)$ é igual a dimensão fractal do sistema. Estruturas multifractais são caracterizadas por um espetro $f(\alpha)$ não trivial.

2.3 Análise Multivariada

Na pesquisa cientifica aplicada, muitas vezes é essencial a análise de um conjunto de várias medidas feitas sobre um mesmo indivíduo da amostra. Por exemplo, para o cálculo do índice de massa corpórea (IMC), duas variáveis são necessárias, o peso e a altura. Este simples exemplo mostra a necessidade da coleta e da análise de mais de uma variável presente na amostra. Outro exemplo é o conjunto de dados cujas medidas são o tamanho e largura das pétalas da planta Iris[1] tomada de duas espécies (ANDERSON, 2003, p. 1).

Para esse tipo de análise são necessárias técnicas que possam lidar com todas as relações existentes entre as variáveis. O ramo da estatística para este tipo de aplicação é chamado de estatística multivariada e também denominado de análise multivariada. Uma definição dada para esta área da estatística é a seguinte: a análise multivariada é um conjunto de técnicas utilizadas em situações que várias variáveis são medidas simultaneamente, para cada elemento amostral (MINGOTI, 2005; HAIR et al., 2005; MARDIA et al., 1979). Alguns exemplos de técnicas multivariadas para análise de dados são:

[1] planta da família das Iridáceas

1. Análise de regressão múltipla: método que tem por objetivo examinar as relações entre uma variável resposta e um conjunto de variáveis explicativas, sendo esta relação linear ou não-linear. Esta relação linear é enquadrada no conjunto de modelos lineares generalizados (HAIR et al., 2005).

2. Análise multivariada de dados: generalização da ANOVA, que objetiva analisar a variância dos dados onde há mais de uma variável independente.

3. Análise de componentes principais: busca determinar um conjunto reduzido e significante de variáveis que expliquem o conjunto de dados (MINGOTI, 2005).

4. Análise fatorial: busca descrever a variabilidade original do conjunto de dados em termos de um número reduzido, porém significante de variáveis aleatórias, chamadas de fatores comuns e que estão relacionadas aos dados através de um modelo linear (MINGOTI, 2005).

5. Análise de discriminante linear: dado o conhecimento *a priori* das características dos agrupamentos existentes em uma população, esta técnica desenvolve uma função capaz de classificar novos elementos a algum dos grupos com perfil semelhante a esta (MINGOTI, 2005).

6. Análise de correlação canônica: técnica que tenta estabelecer se há ou não uma relação linear entre dois conjuntos de variáveis (covariáveis e respostas) (MINGOTI, 2005).

7. Análise de agrupamento: técnica que objetiva a classificação de indivíduos em grupos, através de características presentes nestes indivíduos (MINGOTI, 2005). Este trabalho utilizará a análise de agrupamento para avaliar o banco de dados descrito no Capítulo 3.

Outra área em que existe a necessidade do uso de múltiplas variáveis é a de **Reconhecimento de padrões**. Esta área é um ramo da inteligência artificial direcionada a classificação ou descrição de observações. O reconhecimento de padrões permite classificar informações (padrões) baseado, ou no conhecimento *a priori*, ou em informações estatísticas extraídas dos padrões. O reconhecimento de padrões pode ser dividido em duas classes de métodos. Os métodos de classificação supervisionados (e.g. análise de discriminante), no qual os objetos são identificados como membros de uma classe predefinida e os métodos de classificação não-supervisionados (e.g. análise de agrupamentos), onde indivíduos são assinalados em uma classe não definida (JAIN et al., 2000).

2.4 Análise de Agrupamento

A Análise de agrupamento é o termo usado para nomear uma série de técnicas que têm por finalidade dividir os elementos de uma amostra, ou população, em grupos de forma que os elementos pertencentes a um mesmo grupo sejam semelhantes entre si com respeito às variáveis (características) que neles foram medidas. Enquanto os grupos sejam o mais heterogênios quanto possível (MINGOTI, 2005).

Para atingir este objetivo, algumas questões precisam ser levantadas e discutidas. Segundo Hair et al. (2005) estas questões são três: (i) Como é definida a similaridade; (ii) Como é formado o agrupamento; (iii) e Quantos grupos podem ser formados. Porém, Barroso e Artes (2003, pg 10-14) aporta uma divisão detalhada destas questões, sendo expostas a seguir.

1. **Escolha do critério de presença**[2] - etapa em que as variáveis devem ser definidas e todas as transformações realizadas, além da escolha do critério que será utilizado para a determinação dos grupos. No presente trabalho, a proximidade dos pontos é usada como medida de semelhança entre os objetos da amostra.

2. **Definição do número de grupos** - etapa em que o número de grupos deve ser definido dado um conhecimento prévio dos dados.

3. **Formação dos grupos** - etapa em que se deve definir o algoritmo que será utilizado na identificação dos grupos.

4. **Validação dos agrupamentos** - etapa onde deve ser garantido o fato de que as variáveis têm comportamento diferenciado nos diversos grupos, supondo que cada grupo seja uma amostra aleatória de alguma subpopulação, aplicando técnicas inferenciais para compará-las.

5. **Interpretação dos grupos** - ao final do processo de formação de grupos é importante caracterizar os grupos formados. O uso de estatísticas descritivas é recomendado nesta fase da análise.

2.4.1 Medidas de Similaridade e Dissimilaridade

Uma questão importante refere-se ao critério que deve ser adotado para decidir o quão dois elementos do conjunto de dados podem ser considerados como semelhantes ou dis-

[2]Esta é uma fase investigativa do conjunto de dados.

tintos. Para responder esta questão é necessário considerar medidas de semelhança entre os elementos amostrais. Assim, considerando que cada elemento amostral possui informações de p variáveis armazenadas em um vetor, a comparação de diferentes elementos amostrais poderá ser feita através de medidas matemáticas (métricas), que possibilitem a comparação destes vetores (MINGOTI, 2005). Estas medidas podem ser chamadas de coeficiente de parecença segundo Bussab et al. (1990). As medidas podem ser classificadas em qualitativas e quantitativas. As medidas qualitativas são atributos, características ou propriedades categóricas que particularizam ou descrevem o objeto. Estas podem descrever diferenças, indicar presença ou ausência de uma característica ou propriedade. Por exemplo, a variável sexo só pode assumir as características masculina ou feminina. Já as variáveis quantitativas, como seu nome já expressa, são incógnitas que assumem valores numéricos (HAIR et al., 2005). As medidas de comparação podem ser divididas em duas categorias: medidas de similaridade (quanto maior o valor, mais semelhantes são os objetos) e dissimilaridade (quanto maior o valor, mais distintos são os objetos) (BUSSAB et al., 1990). As distâncias são as mais usadas no estudo de dados constituídos de variáveis quantitativas. Uma medida de distância $d(i.j)$ representa uma distância entre os elementos i e j se as condições abaixo forem satisfeitas:

a) $d(i,j) \geq 0 \quad \forall i,j; \quad i \neq j$;

b) $d(i,i) = 0$;

c) $d(i,k) = d(k,i)$;

c) $d(i,k) \leq d(i,m) + d(m,k)$.

Para se definir medidas de distâncias é necessário tomar um conjunto de dados constituído de n elementos amostrais, onde para cada um destes foram medidas p variáveis. Então, para cada elemento j, $i = \{1,2,\ldots n\}$, é definido o vetor de medidas como

$$X_j = (X_{1j}, X_{2j}, \ldots, X_{nj},) \tag{2.13}$$

em que $X_{i,j}$ é o valor observado da variável i medida no elemento j.

Entre as medidas de dissimilaridade mais comuns estão a distância Euclidiana e a distância de Manhattan e entre as medidas de similaridade tem-se a Correlação de Pearson. Estas são definidas da seguinte forma:

- A distância euclidiana entre um par de elementos amostrais l e k sendo $l \neq k$ é

definida por:

$$d(l,k) = \left[(X_l - X_k)' (X_l - X_k) \right]^{1/2} = \left[\sum_{i=1}^{p} (X_{il} - X_{ik})^2 \right]^{1/2} \quad (2.14)$$

- A distância Manhattan (City-block distance) entre elementos amostrais l e k sendo $l \neq k$ é definida por:

$$d(l,k) = \sum_{i=1}^{p} w_i |X_{il} - X_{ik}| \quad (2.15)$$

em que W_i são os pesos de ponderação para as variáveis.

- A correlação entre dois elementos amostrais l e k sendo $l \neq k$ é definida por:

$$Cor(l,k) = \frac{\sum_{i=1}^{p} \left(X_{il} - \overline{X}_l\right)\left(X_{ik} - \overline{X}_k\right)}{\sqrt{\sum_{i=1}^{p} \left(X_{il} - \overline{X}_l\right)} \cdot \sqrt{\sum_{i=1}^{p} \left(X_{jk} - \overline{X}_k\right)}} \quad (2.16)$$

Existem várias outras medidas que podem ser usadas na análise de agrupamento e algumas destas são mostradas em Mingoti (2005), Mardia et al. (1979), Barroso e Artes (2003) ou Hair et al. (2005).

2.5 Algoritmos de Agrupamento

Os algoritmos utilizados na construção de grupos são geralmente classificados em duas famílias de técnicas: Hierárquicas e Não Hierárquicas. As hierárquicas são técnicas que muitas vezes são usadas como uma espécie de análise exploratória, com o intuito de identificar indícios de possíveis agrupamentos presentes na amostra. Enquanto para as técnicas não hierárquicas é necessária a definição prévia do número de grupos para sua aplicação (MINGOTI, 2005).

2.5.1 Algoritmos Hierárquicos

Os métodos hierárquicos são divididos em aglomerativos e divisivos. Os aglomerativos partem do pressuposto de que existem na amostra n grupos de indivíduos, ou seja, cada elemento amostral forma um grupo distinto. Em cada passo do algoritmo, os elementos mais similares são agregados formando grupos, até que todos os elementos formem um único grupo. Os métodos hierárquicos divisivos trabalham na direção oposta, onde todos os elementos são alocados em um único grupo inicial. Este grupo é dividido em

dois subgrupos de modo que exista grande semelhança entre os objetos dos mesmos subgrupos e também uma grande dissimilaridade entre elementos de subgrupos distintos. A cada passo do algoritmo, os elementos são subdivididos em outros subgrupos dissimilares até que haja tantos subgrupos quantos elementos amostrais. A variância interna no início do processo aglomerativo é nula pois cada grupo é representado por um elemento e ao final de todas as etapas do algoritmo tem-se a variância máxima, porque todos os elementos estão alocados em um grupo. Sendo o contrário para o processo divisivo. Em cada estágio do procedimento de agrupamento, os grupos são comparados através de uma medida de similaridade ou dissimilaridade previamente escolhida (HAIR et al., 2005). Como os métodos de agrupamento aglomerativos e divisivos se comportam de maneira análoga, mas oposta, será considerado nesse estudo apenas os métodos aglomerativos. Existem vários métodos de agrupamento hierárquicos. Os mais comuns são: o método de Ligação Simples (Single Linkage), Ligação Completa (Complet Linkage), Ligação Média (Average Linkage), Centróide (Centroid Linkage) e Ward, entre outros apresentados em Mingoti (2005), Mardia et al. (1979), Barroso e Artes (2003), Hair et al. (2005). Nesta dissertação, dos métodos supracitados, apenas será detalhado o método de Ward.

Método de WARD

Também conhecido como método de “mínima variância”, o método de Ward como todo método hierárquico aglomerativo parte do pressuposto de que cada elemento da amostra seja um grupo original e a cada passo os conglomerados que minimizam a função objetivo sejam agrupados até a formação de um único grupo. Assim, pode-se observar em alguns métodos hierárquicos, que a cada passo de execução do algoritmo, a similaridade entre os elementos do grupo decresce. Isto é, a variação entre os grupos diminui e a variação entre os elementos aumenta. No ano de 1963, Ward propôs um método de agrupamento baseado nesta mudança de variação da similaridade interna e externa dos grupos formados a cada passo. Assim, a medida de homogeneidade é a soma de quadrados total de uma análise de variância (MINGOTI, 2005). Os princípios básicos do método são os seguintes:

(a) Inicialmente cada elemento é considerado um conglomerado;

(b) Em cada passo a soma dos quadrados da distância euclidiana de cada elemento amostral é correspondente ao vetor das médias do conglomerado, isto é

$$SS_i = \sum_{j=1}^{n_i} \left(X_{ij} - \overline{X}_{i.}\right)' \left(X_{ij} - \overline{X}_{i.}\right) = \sum_{j=1}^{n_i} \left[\sum_{k=1}^{p} \left(X_{ikj} - \overline{X}_{ki.}\right)^2\right]^{1/2} \tag{2.17}$$

em que n_i é o número de elementos da amostra no agrupamento g_i, X_{ij} é o vetor do j-*ésimo* elemento do i-*ésimo* grupo, $\overline{X}_i$ é o centróide do i-*ésimo* grupo g_i e SS_i representa a soma de quadrados do conglomerado g_i. Em cada passo t a soma total dos quadrados é dada por:

$$SSR = \sum_{l=1}^{n_g} = SS_i \tag{2.18}$$

em que n_g é o número de grupos existentes no passo t.

A distância entre os grupos g_i e g_j é definida como:

$$d(g_i, g_k) = \left[\frac{n_i \cdot n_j}{n_i + n_j}\right] \left(\overline{X}_{i.} - \overline{X}_{j.}\right)' \left(\overline{X}_{i.} - \overline{X}_{j.}\right) = \left[\sum_{k=1}^{p} \left(\overline{X}_{ik.} - \overline{X}_{jk.}\right)^2\right]^{1/2} \tag{2.19}$$

que é a soma de quadrados entre os grupos g_i e g_j. Em cada passo do algoritmo de agrupamento, os dois conglomerados que minimizam a distância (2.19) são combinados. Dada a distância (2.19) é possível demonstrar que esta é a diferença entre a soma total dos quadrados, antes e depois de combinar os conglomerados g_i e g_j. Logo, o método de Ward combina os grupos que minimizam SSR.

2.5.2 Algoritmos Não Hierárquicos

Os procedimentos não hierárquicos atuam de modo diferente dos métodos hierárquicos. Estes métodos não envolvem um processo de construção em árvore. Em vez disso, designam os objetos da amostra a agrupamentos assim que o número de conglomerados seja designado. Assim, por exemplo, a solução de um agrupamento com seis grupos não é apenas a combinação de dois grupos a partir da solução de sete agregados, mas sim busca a melhor solução com exatos seis grupos. Resumidamente, estes procedimentos primeiramente selecionam uma semente de agrupamento como centro inicial deste, e os objetos restantes que estejam dentro de uma distância previamente especificada são alocados ao grupo representado por cada um dos centros. Em seguida, outra semente é escolhida e esta alocação aos grupos continua até que um critério de parada seja alcançado (HAIR et al., 2005). Entre os métodos de agrupamento não hierárquicos os mais comuns são: K-médias (K-means), PAM (Partitional Around Medoids) (KAUFMAN; ROUSSEEUW, 1990) e Fuzzy (KAUFMAN; ROUSSEEUW, 1990), este último seguindo a lógica fuzzy de conjuntos.

Método de K-médias

O método de K-médias (HARTIGAN, 1975, p. 84) é muito utilizado em problemas práticos. Basicamente, cada objeto amostral é alocado àquele grupo cujo centróide (vetor de médias amostral para o grupo) é o mais próximo do vetor de valores observados para o respectivo elemento. Sua descrição formal é dada a seguir.

Seja $A(i,j)$ o valor da *j-ésima* variável para o *i-ésimo* objeto em que $(1 \leq i \leq n)$ e $(1 \leq j \leq p)$. Os valores da variável são apropriados de modo que a distância euclidiana possa ser usada como medida de dissimilaridade. Dada uma partição $(P(n,k))$ que divide os objetos em k, a saber $l = \{1,2,3,\ldots k\}$, cada um dos n objetos deve ser alocado a apenas um dos k grupos. Seja $B(l,j) = \{c_1,\ldots c_p\}$ a média da j-ésima variável sobre os objetos do grupo l. Seja $N(l)$ o número de objetos pertencentes ao grupo l. A distância entre cada objeto i e o grupo l é dada por

$$d(i,l) = \sum_{j=1}^{p} \left[(A(i,j) - B(l,j))^2 \right]^{1/2} \tag{2.20}$$

O erro referente a partição é

$$e\,[P(n,k)] = \sum_{i=1}^{n} d\,(i,l(i))^2 \tag{2.21}$$

em que $l(i)$ é o grupo que contêm o objeto i. O procedimento geral consiste em buscar uma partição que minimize $e[(P(n,k))]$ pelo movimento de um objeto i de um grupo para o outro. A busca se encerra quando o movimento do objeto não muda o valor de $e[(P(n,k))]$.

Os passos de aplicação do algoritmo são:

1. Assumindo uma partição inicial em k clusters, calcule $B(l,j)$, o valor de $e[(P(n,k))]$ inicial.

   $$e\,[P(n,k)] = \sum_{i=1}^{n} d\,(i,l(i))^2 \tag{2.22}$$

   em que $d\,(i,l(i))$ denota a distância euclidiana entre o objeto i e a média do grupo que o contém i.

2. Para o objeto 1 calcule:

   $$A = \frac{N(l)\cdot(d(1,l))^2}{N(l)+1} - \frac{N\,[l(1)]\cdot(d(1,l(1)))^2}{N\,[l(1)]-1} \tag{2.23}$$

Este é o acréscimo no erro da transferência do objeto 1 do cluster onde o mesmo está alocado para o cluster l. Se o mínimo de A para algum $l \neq l(1)$ for negativo, transfere-se o objeto 1 do grupo $l(1)$ para o grupo l. Ajusta-se o valor da média de $l(1)$ e l e adiciona o valor do incremento A ao erro $e[(P(n,k))]$.

3. Repita o passo 2 para os demais objetos.

4. Se o movimento de um objeto de um cluster para outro não provocar mudança em $e[(P(n,k))]$, pare. Caso contrário, repita o passo 2.

O algoritmo implementado no software (R Development Core Team, 2007) está exposto no trabalho de Hartigan e Won (1979).

Método PAM (*Partitional Around Medoids*)

O método PAM apresentado por Kaufman e Rousseeuw (1990, p. 102) é usado para agrupar objetos para os quais foram medidas p variáveis de escala no mínimo intervalar. Este método busca por k *elementos representativos* chamados de *Medóides* entre os objetos do conjunto de dados. Após encontrar os k elementos, os k grupos são construídos pela atribuição dos elementos restantes ao agrupamento representado pelo medóide mais próximo. O algoritmo continua escolhendo os k elementos até que a soma total das dissimilaridades dos objetos representativos seja a mínima. O algoritmo consiste de duas partes. A primeira parte é chamada de BUILD, esta busca objetos um a um até que k objetos representativos sejam escolhidos. E a segunda, chamada de SWAP, busca refinar o agrupamento formado no BUILD. Esta escolha é finalizada quando a soma total das dissimilaridades entre os objetos representativos e os restantes dos elementos da amostra alcança um valor mínimo. O primeiro objeto é aquele mais central, para o qual a soma das distâncias naquele passo seja a menor possível. Na seqüência, em cada passo outro objeto é escolhido, o qual, decresce a soma das dissimilaridades. Para encontrar este objeto deve-se executar os seguintes passos:

1. Considere um objeto i ainda não selecionado.

2. Considere então outro objeto j. Calcule a distância euclidiana (D_j) entre o elemento j e o elemento previamente selecionado mais similar. Calcule a distância $d(j,i)$ entre j e o objeto i. Após isso calcule $D_j - d(j,i)$.

3. Se esta diferença for positiva, o objeto j contribui para a decisão de selecionar o objeto i. E assim:

$$C_{ji} = max(D_j - d(j,i), 0) \tag{2.24}$$

4. Calcule o total do ganho obtido pela seleção de i

$$\sum_j C_{ji} \tag{2.25}$$

5. Escolha o objeto ainda não selecionado que

$$\underset{i}{\text{minimize}} \sum_j C_{ji} \tag{2.26}$$

Este processo continua até que k objetos sejam encontrados. A segunda parte chamada SWAP tem o objetivo de tentar refinar o agrupamento primariamente estabelecido pelo BUILD. Este processo realiza todas as possíveis intercomparações entre pares de elementos (i,h), onde i foi selecionado como elemento representativo e h não. Este determina qual será o impacto, sobre o agrupamento formado, ao selecionar h em lugar de i. Para calcular o efeito do SWAP sobre o agrupamento deve-se seguir dois passos:

1. Considere um objeto j e calcule a contribuição C_{jih} ao SWAP:

   a. Se j é mais distante de i e h do que de outro elemento representativo, tem-se $C_{jih} = 0$.

   b. Se j não é tão distante de i do que de algum outro elemento representativo selecionado no BUILD $(d(j,i) = D_j)$ serão consideradas duas situações:

      i. j é mais próximo de h e de i do que dos outros elementos representativos, ou seja $d(j,i) < E_i$, onde E_i é a distância entre j e todos os objetos representativos exceto i. Nesta condição a contribuição é dada por:

$$C_{jih} = d(j,h) - d(j,i) \tag{2.27}$$

      ii. j está no mínimo tão distante de h quanto do segundo objeto representativo, ou seja, $d(j,h) \geq E_j$. Neste caso, a contribuição do objeto j ao SWAP é

$$C_{jih} = E_j - D_j \tag{2.28}$$

      OBS: Pode-se observar que no item anterior a contribuição C_{jih} pode ter valores positivos ou negativos dependendo da posição dos objetos j,h e

i, sendo positiva apenas quando j está mais próximo de i do que h. Isso indicando que o SWAP não é favorável do ponto de vista do objeto j. Por outro lado, no caso ii a contribuição será sempre positiva. Pois não será vantajoso trocar i por h, quando este último está mais próximo de j que de outro objeto representativo.

iii. j está mais distante de i que de pelo menos um dos outros objetos representativos, mas próximo de h que de algum objeto representativo. Neste caso a contribuição de j para o SWAP é

$$C_{jih} = d(j,h) - D_j \tag{2.29}$$

2. Calcula-se o valor:

$$T_{ih} = \sum_j C_{jih} \tag{2.30}$$

3. Seleciona-se o par (i,h) o qual

$$\underset{i,h}{\text{minimize}}\, T_{ih} \tag{2.31}$$

4. Caso o mínimo T_{ih} seja negativo o SWAP é executado novamente, isto é, o algoritmo retorna ao passo 1. Entretanto, se o mínimo de T_{ih} for positivo ou zero o algoritmo pára.

Pode-se notar que todos os pares são considerados e que o algoritmo não depende da ordem das variáveis dos objetos na entrada do programa.

O algoritmo Fuzzy

Nos métodos de partição até aqui citados cada elemento pode apenas pertencer a um grupo, ou seja, se for atribuído um valor para indicar sua presença ou ausência em um grupo seu valor seria 0 se pertencesse e 1 se não. Portanto, métodos de partição clássicos são algumas vezes chamados de métodos de formação de grupos fechados. Já o método fuzzy de agrupamento (KAUFMAN; ROUSSEEUW, 1990), baseado na lógica fuzzy de conjuntos, permite a escolha através do coeficiente de agrupamento, com variação entre 0 e 1, onde a qual grupo o objeto deve ser mais bem agrupado.

Seu objetivo principal é minimizar a função:

$$C = \sum_{\upsilon=1}^{k} \frac{\sum_{i,j=1}^{n} u_{i\upsilon}^2 u_{j\upsilon}^2 d(i,j)}{2\sum_{j=1}^{n} u_{j\upsilon}^2} \tag{2.32}$$

em que $d(i,j)$ representa a distância entre os objetos i e j, enquanto $u_{i\upsilon}$ é o coeficiente de pertinência do objeto i ao cluster υ. É possível notar alguns detalhes nesta função. Para começar a mesma contém apenas a $d(i,j)$ e $u_{i\upsilon}$, sendo este último o valor desconhecido que deve ser encontrado. Tem-se ainda que a soma presente no numerador seja feita para todos os pares de elementos da amostra. Nesta soma, cada par aparece duas vezes porque os pares (i,j) e (j,i) ocorrem e é por isso que a fração é dividida por 2. E a soma externa é feita para todos os clusters υ. Finalmente, é possível notar com estas considerações que a função (2.32) é um tipo de dispersão. O coeficiente de pertinência possui as seguintes restrições:

$$u_{i\upsilon} \geq 0 \quad \text{para} \quad i = \{1,\ldots,n\};\ \upsilon = \{1,\ldots,k\} \tag{2.33}$$

$$\sum_{\upsilon} u_{i\upsilon} = 1 \quad \text{para} \quad i = \{1,\ldots,n\} \tag{2.34}$$

Expressando que o coeficiente não pode ser negativo e que a soma de todos os coeficientes deve ser igual a 1 logo o $u_{i\upsilon}$ assemelha-se a uma probabilidade do objeto pertencer ao grupo υ.

O método fuzzy oferece uma vantagem sobre os outros métodos de agrupamento. Este apresenta para cada elemento uma probabilidade de pertencer a cada um dos grupos cujo número é arbitrariamente determinado. Estes valores podem ser agrupados em uma tabela que será denominada **Tabela de probabilidades de pertinência**.

O processo de maximização está bem detalhado no livro “Finding Groups in Data: an introdution to Cluster Analysis” escrito por Kaufman e Rousseeuw.

Uma variação deste método é o *Fuzzy c-means* proposto por Bezdek (1981), citado por (MINGOTI, 2005) e consiste em dado um número c de grupos predefinidos, o mesmo busca minimizar a função objetivo 2.35

$$J = \sum_{i=1}^{c} \sum_{j=1}^{n} \left(u_{ij}\right)^m d\left(X_j, V_i\right) \tag{2.35}$$

em que V_i é o centróide ponderado do conglomerado i, para $1 \leq i \leq c$, $m > 1$ é o parâmetro Fuzzy; u_{ij} é a probabilidade de que o elemento X_j pertença ao conglomerado; $d\left(X_j, V_i\right)$ é a distância euclidiana entre o objeto X_j e o centróide V_i.

A partição é dada através da maximização da função 2.35 com a atualização dos valores u_{ij} e dos centróides V_i por

$$u_{ij} = \left[\sum_{k=1}^{c} \left(\frac{\left\| x_i - c_j \right\|}{\| x_i - c_k \|} \right)^{\frac{2}{m-1}} \right]^{-1} \tag{2.36}$$

em que

$$V_i = \frac{\sum_{j=1}^{n} \left(u_{ij} \right) X_j}{\sum_{j=1}^{n} u_{ij}} \tag{2.37}$$

para todo $i = \{1, 2, \ldots, c\}$ e $j = \{1, 2, \ldots, n\}$. Para encontrar a solução final, devem-se ter os valores de V_i e u_{ij} iniciais. Nos algoritmos disponíveis, os valores de u_{ij} são gerados por uma distribuição uniforme no intervalo $[0, 1]$; os valores dos centróides vão se modificando a cada interação e o algoritmo é interrompido quando o número de iterações é alcançado ou quando o programa é incapaz de minimizar o valor da função objetivo por um valor ε. Ao contrário do método K-médias, que fornece como resultado uma partição na qual cada elemento pertence a um único cluster, no método *Fuzzy*, para cada elemento amostral, estima-se a probabilidade de que o mesmo elemento pertença a cada um dos c clusters formados. Assim, é possível identificar os elementos amostrais que estão na interface, ou seja, que se assemelham a mais de um dos c grupos. Passos do Algoritmo:

1. Inicialize com os valores da matriz $U = [U_{ij}]$, $U^{(0)}$
2. Em cada passo k calcule os valores de $V_j^{(k)}$ com $U^{(k)}$

$$V_i^{(k)} = \frac{\sum_{j=1}^{n} \left(u_{ij}^{(k)} \right) X_j}{\sum_{j=1}^{n} u_{ij}^{(k)}}$$

3. Calcule os valores de U^k, U^{k+1}, $J^{(k)}$ e $J^{(k+1)}$

$$u_{ij} = \left[\sum_{k=1}^{c} \left(\frac{\left\| X_i - c_j \right\|}{\| X_i - c_k \|} \right)^{\frac{2}{m-1}} \right]^{-1}$$

4. Se $\left\| J^{(k+1)} - J^{(k)} \right\| < \varepsilon$ for satisfeito pare, senão retorne ao passo 2.

2.5.3 Apresentação Gráfica e Verificação da Qualidade do Agrupamento

A construção de gráficos permite ao pesquisador uma melhor visualização dos dados sob investigação. Nos métodos de agrupamento podem-se usar alguns tipos de gráficos para visualização e verificação da qualidade do agrupamento.

Dendograma

Na análise de agrupamento hierárquica, após o término do procedimento, pode-se construir um gráfico chamado de dendograma ou dendrograma. Este gráfico tem a forma de uma árvore, onde no eixo vertical tem-se a medida de similaridade ou dissimilaridade e no eixo horizontal apresentam-se os elementos da amostra numa ordem conveniente relacionada ao histórico do agrupamento. As linhas verticais, partindo dos elementos amostrais agrupados, têm a altura correspondente ao nível em que os elementos foram considerados semelhantes, isto é, a distância do agrupamento ou o nível de similaridade (BUSSAB et al., 1990; BARROSO; ARTES, 2003; HAIR et al., 2005).

Gráfico de Silhueta

Uma dificuldade na aplicação do método de agrupamento é decidir o número de clusters e como distinguir uma má alocação de um elemento a um grupo. Uma maneira de solucionar estes problemas é através do cálculo da silhueta de cada elemento. A silhueta é um coeficiente que mede o quão bem alocado cada elemento está ao seu grupo comparado aos outros clusters formados. Esta medida é calculada em termos da média da distância euclidiana entre o elemento i e todos os elementos do grupo de i, comparando com as médias das distâncias entre i e os elementos do grupo vizinho.

A silhueta é construída da seguinte maneira.

Considere um objeto i presente na amostra pertencente a um cluster A e calcule

$$a(i) = \frac{\sum_{(j \in A) \;\; j \neq i} d_{(i,j)}}{n_A - 1} \tag{2.38}$$

que representa a média de dissimilaridade entre todos os elementos do grupo A.

Agora considerando um cluster C diferente de A, pode-se definir

$$d_{i,C} = \frac{\sum_{(j \in C)} d_{(i,j)}}{n_C} \tag{2.39}$$

onde esta representa a média da dissimilaridade entre i e todos os elementos de C. Calcule esta medida para todos os agrupamentos $C \neq A$ cluster objetos e escolha a menor desta, a qual é definida como $b(i)$.

$$b(i) = mim\left(d_{i,C}\right); \quad C \neq A \tag{2.40}$$

O cluster B para o qual $d_{i,C} = b(i)$ é chamado de vizinhança o objeto i. Logo, pode-se definir o valor da silhueta de $s(i)$ como

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \tag{2.41}$$

É notório que o valor de $s(i)$ está no intervalo $[-1, 1]$ e o mesmo pode ser interpretado da seguinte forma:

- para $s(i)$ próximo de 1 é muito pequeno em relação à $b(i)$. Em outras palavras, o objeto i está muito próximo dos objetos do seu grupo em comparação com o seu vizinho.
- para $s(i)$ em torno de zero, $a(i)$ e $b(i)$ são aproximadamente iguais indicando que o mesmo pode ser um objeto intermediário entre A e B.
- para $s(i)$ próximo de -1, o valor de $b(i)$ é muito menor do que o valor de $a(i)$. Em outras palavras, o objeto i está muito próximo do seu vizinho do que do grupo ao qual ele foi assinalado, ou seja, i está erroneamente alocado a A.

Assim, para a construção do gráfico de silhueta, os objetos devem ser divididos em grupos de acordo com o resultado do método de agrupamento. Em cada grupo, os elementos são ordenados em ordem decrescente seguindo o valor da silhueta. Cada objeto é representado por uma barra horizontal, cujo o comprimento é o valor da silhueta. Desta forma, todos os elementos são expostos em um único diagrama (figura 2.3) onde a qualidade do agrupamento pode ser analisada.

A silhueta é uma boa ferramenta para a verificação do número de clusters. A média das silhuetas $\overline{s}(k)$ é definida por:

$$\overline{s}(k) = \sum_{i=1}^{n} s(i) \tag{2.42}$$

e pode ser usada para selecionar o melhor valor do número de grupos (k) pela escolha do valor de $\overline{s}(k)$ quando $\overline{s}(k)$ é máximo. Executa-se o método de agrupamento escolhido para

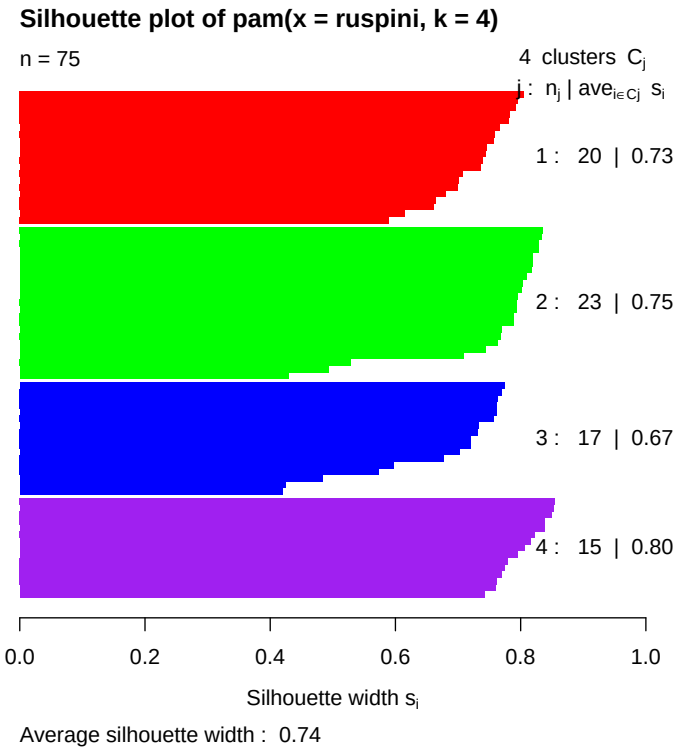

Figura 2.3: Exemplo de um gráfico de silhueta encontrado no modo de ajuda do software (R Development Core Team, 2007).

todos os possíveis valores de (k) a saber $k = \{2,3,\ldots,n-1\}$ onde tem-se max$(\bar{s}(k))$ como o número de grupos para amostra. A medida $SM = \bar{s}(k)$ é um coeficiente de qualidade da perda de dimensionalidade da estrutura do agrupamento. Kaufman e Rousseeuw (1990) apresenta uma proposta de interpretação apresentada na Tabela 2.1.

Tabela 2.1: Proposta de Interpretação para o valor da silhueta média $(\bar{s}(k))$ (KAUFMAN; ROUSSEEUW, 1990, p. 88)

Intervalo de SM	Interpretação
$0,71-1,00$	Existe uma forte estrutura de agrupamento formada.
$0,51-0,70$	Existe uma razoável estrutura para o agrupamento.
$0,26-0,50$	Existe uma fraca estrutura para o agrupamento, e poderia ser artificial. Para uma melhor conclusão é aconselhável a aplicação de métodos de agrupamentos adicionais.
$\leq 0,25$	Foi encontrada uma estrutura não substancial de agrupamento.

Correlação Cofenética

Na Bioestatística, a **correlação cofenética** ou coeficiente de correlação cofenética é uma medida que quantifica o quão fielmente um dendograma preserva a distância original entre um par de objetos. A correlação é uma medida de validação utilizada principalmente

para os métodos de agrupamento hierárquicos. Estas medidas assemelham-se a correlação de Pearson entre a matriz de distâncias originais (O) e a matriz de distâncias baseadas no dendrograma, esta última chamada de **Matriz Cofenética** (C). Esta é definida da seguinte maneira:

$$cor_{cof} = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(c_{ij}-\overline{c})(o_{ij}-\overline{o})}{\sqrt{[\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(c_{ij}-\overline{c})][\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(o_{ij}-\overline{o})]}} \tag{2.43}$$

Valores desta medida que são próximos de 1, e segundo Bussab et al. (1990), maiores que $0,8$, indicam uma pequena distorção provocada pelo método de agrupamento. Existem outras maneiras de verificar a qualidade de métodos de agrupamento tanto hierárquicos quanto não hierárquicos. Isto é feito através do gráfico de silhueta dos agrupamentos formados.

3 Materiais e Métodos

3.1 Dados

3.1.1 Projeto STARE (STructured Analysis of the Retina Project)

O projeto STARE foi concebido e iniciado em 1975 por Dr. Michael Goldbaum, da Universidade de Califórnia, San Diego, e foi financiado continuamente pelos Institutos Nacionais de Saúde dos EUA, (National Institutes of Health (U.S.A.)) desde 1986. Durante este tempo, mais de 30 pesquisadores contribuíram ao projeto, com os conhecimentos que vão da medicina à engenharia. As imagens e os dados clínicos foram fornecidos pelo Shiley Eye Center na universidade de Califórnia e pelo Veterans Administration Medical Center também em San Diego.

3.1.2 Bancos

Do projeto STARE foram coletadas 20 imagens sendo 10 destas pertencentes a indivíduos com alguma patologia e 10 pertencentes a indivíduos com retinas normais. Estas imagens foram segmentadas[1] por dois observadores chamados aqui de **AH** e **VK**. Após coletar estas imagens do bando STARE, as imagens foram esqueletonizadas[2] por Stosic e Stosic (2006). Posteriormente, foram calculadas as dimensões generalizadas e o espectro multifractal de todas as imagens, tanto as segmentadas manualmente, quanto as imagens esqueletonizadas, para os dois observadores. Assim, a estrutura dos bando ficou dividida em 20 imagens segmentadas manualmente por **AH** e **VK**, totalizando 40 e estas 20 imagens foram esqueletonizadas somando ao estudo mais 40. Logo, tem-se um conjunto de 80 imagens de retinas. Na figura 3.1 tem-se o exemplo de duas imagens, uma patológica e outra não patológica.

[1] Este termo significa em marcar os pixels na imagem que pertencem a vasos da retina, com auxílio de algum programa gráfico.

[2] Esqueletonizar significa tomar apenas a ramificação da árvore vascular da retina sem levar em conta o volume.

As imagens patológicas foram nomeadas de $P_1 \ldots P_{10}$ e as imagens de pacientes com retinas normais foram nomeadas de $N_1 \ldots N_{10}$. Este banco de dados não contém informação de quais patologias estão presentes e assim é possível que haja patologias em estados diversos ou até patologias que não afetem a vascularização da retina.

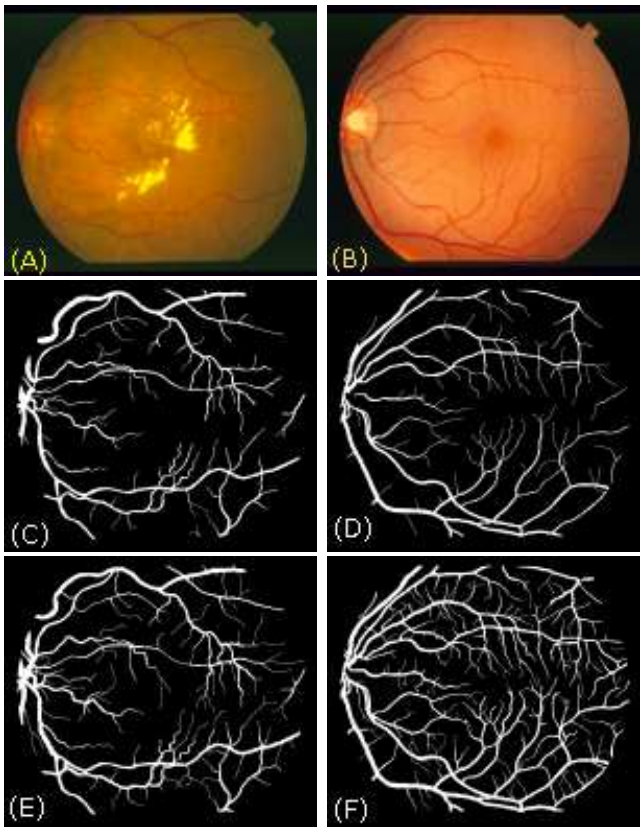

Figura 3.1: Exemplo de imagens do banco STARE: (A) Imagem original e uma retina patológica; (B) Imagem original de uma retina sadia; (C) Imagem patológica segmentada pelo observador AH; (D) Imagem sadia segmentada pelo observador AH; (E) Imagem patológica segmentada pelo observador VK; (D) Imagem sadia segmentada pelo observador VK.

Os bancos de dados consistem de valores das dimensões multifractais generalizadas e de elementos do espectro multifractal $f(\alpha)$, ambos resultados de análise realizada por Stosic e Stosic (2006). Destes, foram escolhidas como variáveis os pares de valores de α e $f(\alpha)$ do espectro, para $\tau(q = -3,0,3)$ e $\tau(q = -2,0,2)$ constituinte dos bancos nomeados de Banco 1 e 2, respectivamente. As dimensões generalizadas (multifractais) D_0, D_1 e D_2 são constituintes do Banco 3, sendo estes apresentados nas Tabelas 3.1 a 3.5. Para cada observador existem dois tipos de imagens (esqueletonizadas e segmentadas manualmente) e 3 bancos, constituindo um total de 12 matrizes de dados analisadas.

Tabela 3.1: Banco 1: valores de α e $f(\alpha)$ para $q = (-3, 0, 3)$ para as imagens esqueletonizadas

Imagens	Observador AH						Observador VK					
	$\alpha(-3)$	$f(\alpha(-3))$	$\alpha(0)$	$f(\alpha(0))$	$\alpha(3)$	$f(\alpha(3))$	$\alpha(-3)$	$f(\alpha(-3))$	$\alpha(0)$	$f(\alpha(0))$	$\alpha(3)$	$\alpha(f(3))$
P1	1,70	1,33	1,56	1,54	1,48	1,44	1,73	1,40	1,60	1,59	1,55	1,52
P2	1,69	1,28	1,54	1,52	1,45	1,40	1,75	1,32	1,59	1,57	1,48	1,42
P3	1,62	1,33	1,51	1,50	1,46	1,43	1,73	1,42	1,62	1,61	1,58	1,56
P4	1,62	1,35	1,52	1,51	1,47	1,44	1,70	1,45	1,61	1,60	1,56	1,53
P5	1,69	1,36	1,57	1,55	1,51	1,49	1,77	1,51	1,67	1,66	1,63	1,61
P6	1,65	1,40	1,55	1,54	1,49	1,45	1,81	1,47	1,68	1,66	1,61	1,57
P7	1,68	1,40	1,58	1,56	1,51	1,47	2,01	1,32	1,70	1,68	1,65	1,65
P8	1,46	1,54	1,49	1,49	1,44	1,39	1,76	1,46	1,61	1,58	1,48	1,41
P9	1,54	1,32	1,46	1,45	1,41	1,39	1,66	1,41	1,57	1,56	1,53	1,50
P10	1,63	1,33	1,52	1,50	1,45	1,42	1,76	1,46	1,64	1,62	1,56	1,53
N1	1,68	1,45	1,60	1,59	1,56	1,54	1,75	1,52	1,67	1,66	1,65	1,65
N2	1,66	1,38	1,56	1,55	1,51	1,49	1,81	1,46	1,68	1,67	1,64	1,61
N3	1,76	1,32	1,60	1,59	1,55	1,53	1,81	1,45	1,69	1,68	1,67	1,67
N4	1,80	1,39	1,65	1,64	1,59	1,56	1,84	1,48	1,71	1,70	1,67	1,66
N5	1,73	1,45	1,63	1,61	1,55	1,51	1,78	1,44	1,66	1,65	1,61	1,59
N6	1,72	1,42	1,61	1,59	1,52	1,47	1,83	1,44	1,69	1,67	1,63	1,60
N7	1,69	1,43	1,59	1,58	1,54	1,51	1,79	1,45	1,67	1,66	1,65	1,64
N8	1,71	1,43	1,61	1,60	1,57	1,55	1,76	1,48	1,66	1,65	1,63	1,62
N9	1,65	1,42	1,57	1,56	1,55	1,54	1,78	1,48	1,67	1,66	1,65	1,64
N10	1,79	1,41	1,65	1,63	1,59	1,56	1,85	1,44	1,70	1,69	1,68	1,67

Tabela 3.2: Banco 1: valores de α e $f(\alpha)$ para $q = (-3, 0, 3)$ para as imagens segmentadas manualmente

Imagens	Observador AH						Observador VK					
	$\alpha(-3)$	$f(\alpha(-3))$	$\alpha(0)$	$f(\alpha(0))$	$\alpha(3)$	$f(\alpha(3))$	$\alpha(-3)$	$f(\alpha(-3))$	$\alpha(0)$	$f(\alpha(0))$	$\alpha(3)$	$\alpha(f(3))$
P1	2,02	0,96	1,59	1,54	1,40	1,31	2,06	1,07	1,65	1,58	1,41	1,31
P2	1,98	1,04	1,60	1,55	1,39	1,30	2,00	1,21	1,65	1,57	1,37	1,26
P3	1,92	1,03	1,56	1,51	1,40	1,34	1,94	1,24	1,66	1,59	1,43	1,34
P4	1,76	1,21	1,55	1,52	1,40	1,31	1,97	1,19	1,65	1,57	1,34	1,21
P5	1,89	1,20	1,62	1,59	1,49	1,44	2,12	1,19	1,74	1,68	1,53	1,44
P6	1,92	1,12	1,60	1,54	1,39	1,30	2,18	1,06	1,72	1,67	1,51	1,41
P7	1,93	1,16	1,62	1,56	1,42	1,35	2,11	1,13	1,72	1,68	1,57	1,52
P8	1,87	1,09	1,56	1,52	1,39	1,30	2,05	1,24	1,68	1,60	1,41	1,30
P9	1,73	1,08	1,48	1,44	1,32	1,24	1,96	1,14	1,62	1,55	1,39	1,30
P10	1,96	1,11	1,62	1,57	1,42	1,34	2,18	1,05	1,71	1,64	1,47	1,38
N1	1,95	1,18	1,63	1,58	1,44	1,37	1,99	1,22	1,69	1,66	1,57	1,53
N2	1,96	1,08	1,60	1,55	1,44	1,38	2,02	1,20	1,71	1,67	1,57	1,50
N3	2,07	1,10	1,65	1,58	1,40	1,32	2,03	1,23	1,71	1,66	1,52	1,44
N4	2,00	1,27	1,71	1,65	1,47	1,38	2,06	1,27	1,76	1,71	1,59	1,52
N5	2,08	1,17	1,71	1,64	1,49	1,41	2,10	1,19	1,73	1,68	1,57	1,51
N6	1,99	1,18	1,65	1,60	1,45	1,38	2,11	1,15	1,74	1,69	1,57	1,51
N7	1,90	1,23	1,63	1,58	1,46	1,38	2,12	1,11	1,71	1,66	1,54	1,48
N8	1,99	1,07	1,63	1,59	1,46	1,38	2,15	1,02	1,70	1,66	1,55	1,50
N9	1,92	1,15	1,63	1,59	1,50	1,45	2,16	1,02	1,72	1,68	1,59	1,54
N10	1,94	1,23	1,67	1,63	1,54	1,49	2,09	1,16	1,74	1,70	1,59	1,53

Tabela 3.3: Banco 2: valores de α e $f(\alpha)$ para $q = (-2, 0, 2)$ para as imagens esqueletonizadas

Imagens	Observador AH						Observador VK					
	$\alpha(-2)$	$f(\alpha(-2))$	$\alpha(0)$	$f(\alpha(0))$	$\alpha(2)$	$f(\alpha(2))$	$\alpha(-2)$	$f(\alpha(-2))$	$\alpha(0)$	$f(\alpha(0))$	$\alpha(2)$	$f(\alpha(2))$
P1	1,65	1,45	1,56	1,54	1,51	1,49	1,68	1,51	1,60	1,59	1,56	1,56
P2	1,64	1,42	1,54	1,52	1,48	1,46	1,71	1,44	1,59	1,57	1,52	1,50
P3	1,59	1,42	1,51	1,50	1,47	1,46	1,69	1,53	1,62	1,61	1,59	1,58
P4	1,58	1,44	1,52	1,51	1,48	1,47	1,67	1,53	1,61	1,60	1,57	1,57
P5	1,65	1,47	1,57	1,55	1,53	1,52	1,74	1,60	1,67	1,66	1,64	1,64
P6	1,62	1,47	1,55	1,54	1,50	1,49	1,77	1,56	1,68	1,66	1,62	1,61
P7	1,65	1,49	1,58	1,56	1,53	1,52	1,95	1,46	1,70	1,68	1,66	1,66
P8	1,47	1,52	1,49	1,49	1,46	1,45	1,75	1,48	1,61	1,58	1,52	1,50
P9	1,51	1,39	1,46	1,45	1,42	1,42	1,63	1,50	1,57	1,56	1,54	1,53
P10	1,60	1,42	1,52	1,50	1,47	1,46	1,74	1,52	1,64	1,62	1,58	1,57
N1	1,65	1,54	1,60	1,59	1,57	1,56	1,71	1,61	1,67	1,66	1,65	1,65
N2	1,62	1,48	1,56	1,55	1,53	1,52	1,76	1,59	1,68	1,67	1,65	1,64
N3	1,70	1,47	1,60	1,59	1,56	1,55	1,75	1,60	1,69	1,68	1,67	1,67
N4	1,74	1,54	1,65	1,64	1,61	1,60	1,79	1,61	1,71	1,70	1,68	1,68
N5	1,70	1,54	1,63	1,61	1,57	1,56	1,74	1,56	1,66	1,65	1,62	1,62
N6	1,68	1,51	1,61	1,59	1,54	1,53	1,77	1,58	1,69	1,67	1,65	1,64
N7	1,66	1,51	1,59	1,58	1,55	1,54	1,74	1,58	1,67	1,66	1,65	1,65
N8	1,67	1,53	1,61	1,60	1,58	1,57	1,72	1,59	1,66	1,65	1,64	1,63
N9	1,62	1,51	1,57	1,56	1,55	1,55	1,73	1,59	1,67	1,66	1,65	1,65
N10	1,73	1,54	1,65	1,63	1,61	1,60	1,78	1,60	1,70	1,69	1,68	1,68

Tabela 3.4: Banco 2: valores de α e $f(\alpha)$ para $q = (-2, 0, 2)$ para as imagens segmentadas manualmente

Imagens	Observador AH						Observador VK					
	$\alpha(-2)$	$f(\alpha(-2))$	$\alpha(0)$	$f(\alpha(0))$	$\alpha(2)$	$f(\alpha(2))$	$\alpha(-2)$	$f(\alpha(-2))$	$\alpha(0)$	$f(\alpha(0))$	$\alpha(2)$	$f(\alpha(2))$
P1	1,92	1,21	1,59	1,54	1,44	1,41	1,98	1,28	1,65	1,58	1,45	1,42
P2	1,88	1,27	1,60	1,55	1,43	1,40	1,96	1,31	1,65	1,57	1,41	1,37
P3	1,83	1,24	1,56	1,51	1,42	1,40	1,88	1,39	1,66	1,59	1,46	1,43
P4	1,69	1,38	1,55	1,52	1,44	1,41	1,91	1,35	1,65	1,57	1,40	1,36
P5	1,81	1,40	1,62	1,59	1,52	1,50	2,03	1,40	1,74	1,68	1,56	1,53
P6	1,85	1,30	1,60	1,54	1,43	1,40	2,08	1,30	1,72	1,67	1,55	1,52
P7	1,86	1,34	1,62	1,56	1,45	1,43	2,00	1,40	1,72	1,68	1,60	1,58
P8	1,79	1,27	1,56	1,52	1,43	1,40	2,02	1,31	1,68	1,60	1,46	1,43
P9	1,66	1,26	1,48	1,44	1,36	1,33	1,89	1,30	1,62	1,55	1,43	1,40
P10	1,88	1,31	1,62	1,57	1,46	1,43	2,08	1,29	1,71	1,64	1,51	1,48
N1	1,88	1,34	1,63	1,58	1,47	1,44	1,91	1,44	1,69	1,66	1,59	1,58
N2	1,87	1,29	1,60	1,55	1,47	1,44	1,92	1,45	1,71	1,67	1,59	1,57
N3	2,00	1,25	1,65	1,58	1,44	1,41	1,94	1,44	1,71	1,66	1,56	1,53
N4	1,93	1,43	1,71	1,65	1,52	1,48	1,97	1,50	1,76	1,71	1,63	1,60
N5	1,99	1,37	1,71	1,64	1,52	1,49	2,01	1,41	1,73	1,68	1,59	1,57
N6	1,92	1,35	1,65	1,60	1,49	1,46	2,01	1,41	1,74	1,69	1,59	1,57
N7	1,83	1,39	1,63	1,58	1,49	1,46	2,02	1,35	1,71	1,66	1,57	1,54
N8	1,89	1,32	1,63	1,59	1,49	1,47	2,03	1,31	1,70	1,66	1,58	1,56
N9	1,82	1,38	1,63	1,59	1,53	1,51	2,02	1,35	1,72	1,68	1,61	1,59
N10	1,86	1,44	1,67	1,63	1,56	1,55	1,98	1,44	1,74	1,70	1,62	1,60

Tabela 3.5: Banco 3: valores de D_0, D_1 e D_2 para as imagens esqueletonizadas e segmentadas manualmente

Imagens	Observador AH						Observador VK					
	Esqueletonizadas			Seg. Manualmente			Esqueletonizadas			Seg. Manualmente		
	D_0	D_1	D_2	D_0	D_1	D_2	D_0	D_1	D_2	D_0	D_1	D_2
P1	1,54	1,53	1,52	1,59	1,58	1,57	1,54	1,49	1,46	1,58	1,52	1,49
P2	1,52	1,51	1,49	1,57	1,55	1,53	1,55	1,50	1,46	1,57	1,50	1,45
P3	1,50	1,49	1,48	1,61	1,60	1,59	1,51	1,47	1,44	1,59	1,54	1,50
P4	1,51	1,50	1,49	1,60	1,59	1,58	1,52	1,49	1,46	1,57	1,50	1,45
P5	1,55	1,54	1,53	1,66	1,65	1,65	1,59	1,56	1,54	1,68	1,63	1,59
P6	1,54	1,52	1,51	1,66	1,64	1,63	1,54	1,49	1,46	1,67	1,62	1,58
P7	1,56	1,55	1,54	1,68	1,67	1,66	1,56	1,52	1,48	1,68	1,64	1,62
P8	1,49	1,48	1,47	1,58	1,56	1,54	1,52	1,48	1,45	1,60	1,54	1,50
P9	1,45	1,44	1,43	1,56	1,55	1,54	1,44	1,41	1,38	1,55	1,50	1,46
P10	1,50	1,49	1,48	1,62	1,60	1,59	1,57	1,52	1,49	1,64	1,58	1,54
N1	1,59	1,58	1,58	1,66	1,66	1,66	1,58	1,53	1,50	1,66	1,63	1,61
N2	1,55	1,54	1,53	1,67	1,66	1,65	1,55	1,51	1,49	1,67	1,64	1,61
N3	1,59	1,57	1,57	1,68	1,68	1,67	1,58	1,52	1,48	1,66	1,62	1,59
N4	1,64	1,63	1,62	1,70	1,69	1,69	1,65	1,59	1,55	1,71	1,68	1,65
N5	1,61	1,60	1,59	1,65	1,64	1,63	1,64	1,59	1,55	1,68	1,64	1,62
N6	1,59	1,57	1,56	1,67	1,66	1,65	1,60	1,55	1,51	1,69	1,64	1,62
N7	1,58	1,57	1,56	1,66	1,66	1,65	1,58	1,54	1,51	1,66	1,62	1,59
N8	1,60	1,59	1,58	1,65	1,65	1,64	1,59	1,55	1,52	1,66	1,62	1,59
N9	1,56	1,56	1,55	1,66	1,66	1,65	1,59	1,56	1,54	1,68	1,64	1,62
N10	1,63	1,62	1,61	1,69	1,69	1,68	1,63	1,60	1,58	1,70	1,66	1,64

3.2 Métodos Estatísticos

Com o objetivo de verificar a viabilidade de diferenciação das retinas, através do uso da dimensão multifractal como medida de parecença entre objetos, serão usados os conjuntos de dados citados na seção (3.1.2). A análise de agrupamento será usada como ferramenta para quantificar a diferenciação entre as imagens patológicas e não patológicas.

3.2.1 Análise de Agrupamento Aplicada ao Banco de Dados

Neste trabalho foram utilizados o método de agrupamento hierárquico **WARD** e os métodos não hierárquicos **K-médias**, **PAM** e **Fuzzy c-means**. O método hierárquico foi primeiramente aplicado como uma forma descritiva dos agrupamentos e a aplicação dos métodos não hierárquicos como uma forma confirmatória dos agrupamentos. Os métodos de **WARD** e **K-médias** foram escolhidos por seu grande uso na literatura, enquanto os demais foram selecionados para que seus resultados fossem comparados ao método K-médias. Em particular, foi escolhido o método Fuzzy por sua versatilidade no processo

de alocação dos grupos, pois é baseado na lógica fuzzy de conjuntos, e assim formando grupos de uma maneira diferenciada. Além dos motivos citados para escolha dos métodos, outro foi a disponibilidade destes no software (R Development Core Team, 2007), o qual é gratuito e de código fonte aberto. Este programa estatístico dá ao operador uma boa liberdade na reprogramação de suas rotinas para o cálculo e ainda apresentação dos resultados de modo elegante. A medida numérica de dissimilaridade entre um par de elementos amostrais escolhida foi a distância euclidiana, definida na seção 2.4.1. Como os bancos coletados do STARE estão divididos em imagens de retinas patológicas e não patológicas, o número de grupos foi dividido em dois para todos os métodos de agrupamento.

3.2.2 Apresentação Gráfica e Validação

Gráfico de Silhueta

O gráfico de silhueta é um procedimento usado para análise de qualidade dos agrupamentos obtidos, apresentados por vários autores como (BARROSO; ARTES, 2003; KAUFMAN; ROUSSEEUW, 1990; ROUSSEEUW, 1987). Neste procedimento cada objeto é representado por um valor da silhueta o qual mostra se este está bem alocado, mal alocado ou representa um cluster individual.

Dendrogramas

Para ser observada a seqüência de formação dos agrupamentos, foi construído o dendrograma para o método Ward de agrupamento.

Correlação Cofenética

A correlação é uma medida de validação utilizada principalmente para os métodos de agrupamento hierárquicos. Estas medidas assemelham-se a correlação de Pearson entre a matriz de distâncias originais e a matriz de distâncias baseadas no dendrograma, esta última chamada de **Matriz Cofenética**. Valores próximos de 1, segundo Bussab et al. (1990) maiores que $0,8$, já indicam uma pequena distorção provocada pelo método de agrupamento. Existem outras maneiras de verificar a qualidade de métodos de agrupamento tanto hierárquicos quanto não hierárquicos. Isto é feito através do gráfico de silhueta dos agrupamentos formados.

4 Resultados e Discussão

Neste item serão apresentados os resultados da aplicação de métodos de análise de agrupamento, que teve como finalidade verificar a sensibilidade da análise multifractal para classificação de imagens de retinas. Como exposto na seção 2.4, estes métodos têm o objetivo de separar os elementos amostrais em grupos distintos. Esta separação é baseada em características medidas nos indivíduos da amostra estudada. Como os métodos multivariados são divididos em hierárquicos e não hierárquicos, os resultados deste estudo também estão divididos desta forma.

4.1 Métodos Hierárquicos

Esse estudo foi iniciado pela aplicação dos métodos hierárquicos por dois motivos. Primeiro por sua facilidade de execução; e segundo pela facilidade de apresentação gráfica do resultado. A medida de dissimilaridade usada foi a distância euclidiana. Os dendrogramas expostos a seguir apresentam os grupos formados após a aplicação do método de Ward aos dados de cada um dos bancos. Aqui será convencionado como bom resultado a presença de apenas imagens de um tipo alocadas em um único grupo [1].

No caso das imagens esqueletonizadas segmentadas pelo observador VK para todos os bancos analisados, 70% de retinas patológicas (P1,P2,P3,P4,P8,P9,P10) formam um grupo e 30% (P5,P6,P7) se agrupam junto às retinas normais, (Figura 4.1 B,D,F). No caso das imagens segmentadas manualmente pelo mesmo observador, o grupo de retinas patológicas contém 60% do total, sendo que a retina (P10) encontrada no grupo anterior foi deslocada para o grupo de retinas normais (Figura 4.2 B,D,F). No caso do observador AH, os resultados não são tão consistentes (Figura 4.1 & Figura 4.2 A,C,E), pois não se tem a homogeneidade encontrada tanto para as imagens esqueletonizadas quanto para as imagens segmentadas manualmente, onde para o banco 3 tem-se um grupo constituído

[1] Os resultados serão informados através de porcentagens de alocação, como por exemplo, se 10 imagens patológicas são encontradas em um grupo, tem-se 100% de alocação.

apenas de imagens de um tipo.

Os resultados citados no parágrafo anterior mostram que as imagens segmentadas pelo observador VK produzem resultados melhores e mais regulares do que as imagens segmentadas pelo observado AH. Isto pode ser explicado pelo fato de que a segmentação manual feita pelo mesmo resulta em imagens mais detalhadas. E em ambos os casos, as imagens esqueletonizadas mostraram-se melhores para aplicação do método de agrupamento com a formação de grupos homogêneos[2]. Isso evidencia que as alterações em comprimento e ramificação dos vasos são mais significantes para detecção de casos patológicos do que a largura dos vasos.

Para avaliar os resultados apresentados pelos dendrogramas pode-se usar a correlação cofenética, a qual mede o grau de distorção provocado pela aplicação do dendrograma nos resultados. Os seus valores são apresentados na Tabela 4.1. Segundo Bussab et al. (1990) o valor ideal desta correlação é subjetivo e aconselha aos leitores um patamar em torno de $0,8$, o qual já indica uma pequena distorção provocada pelo dendrograma. Logo, observando a Tabela 4.1, na qual estão exibidas as correlações, pode-se observar que o observador VK em todos os bancos possui uma correlação maior quando comparado com o observador AH. Focalizando agora o observador VK individualmente, percebe-se que o mesmo possui valores maiores de correlação para imagens segmentadas manualmente. Isso mostra que o resumo realizado pelo dendrograma é mais sensível para a segmentação manual com o diâmetro dos vasos incluído no cálculo da análise da dimensão multifractal.

Tabela 4.1: Correlações cofenéticas para os agrupamentos formados pelo método hierárquico

Bancos	Banco 1		Banco 2		Banco 3	
Observador	AH	VK	AH	VK	AH	VK
Esqueletonizadas	0,63	0,71	0,57	0,75	0,69	0,87
Seg. Manualmente	0,58	0,86	0,58	0,85	0,59	0,9

Observando estes resultados como uma avaliação inicial percebe-se que os dados podem ser usados como variáveis de classificação. Mas, para confirmar esta afirmação, serão usados os métodos de classificação não hierárquicos.

[2]Aqui este termo indica grupos com apenas um tipo de imagem.

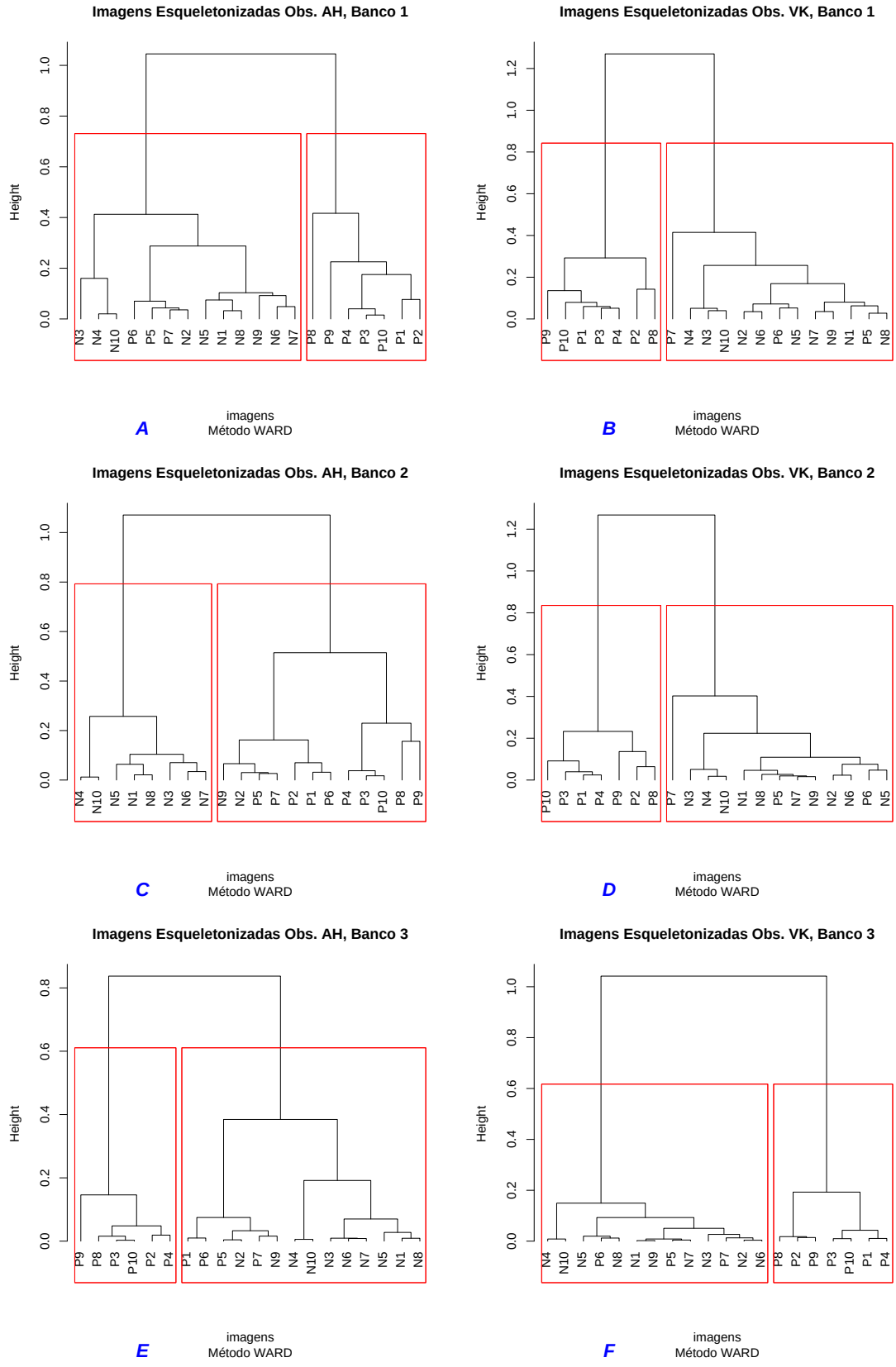

Figura 4.1: Dendrogramas expondo a hierarquia dos agrupamentos formados após a aplicação do método de Ward às imagens esqueletonizadas, para cada um dos bancos e observadores.

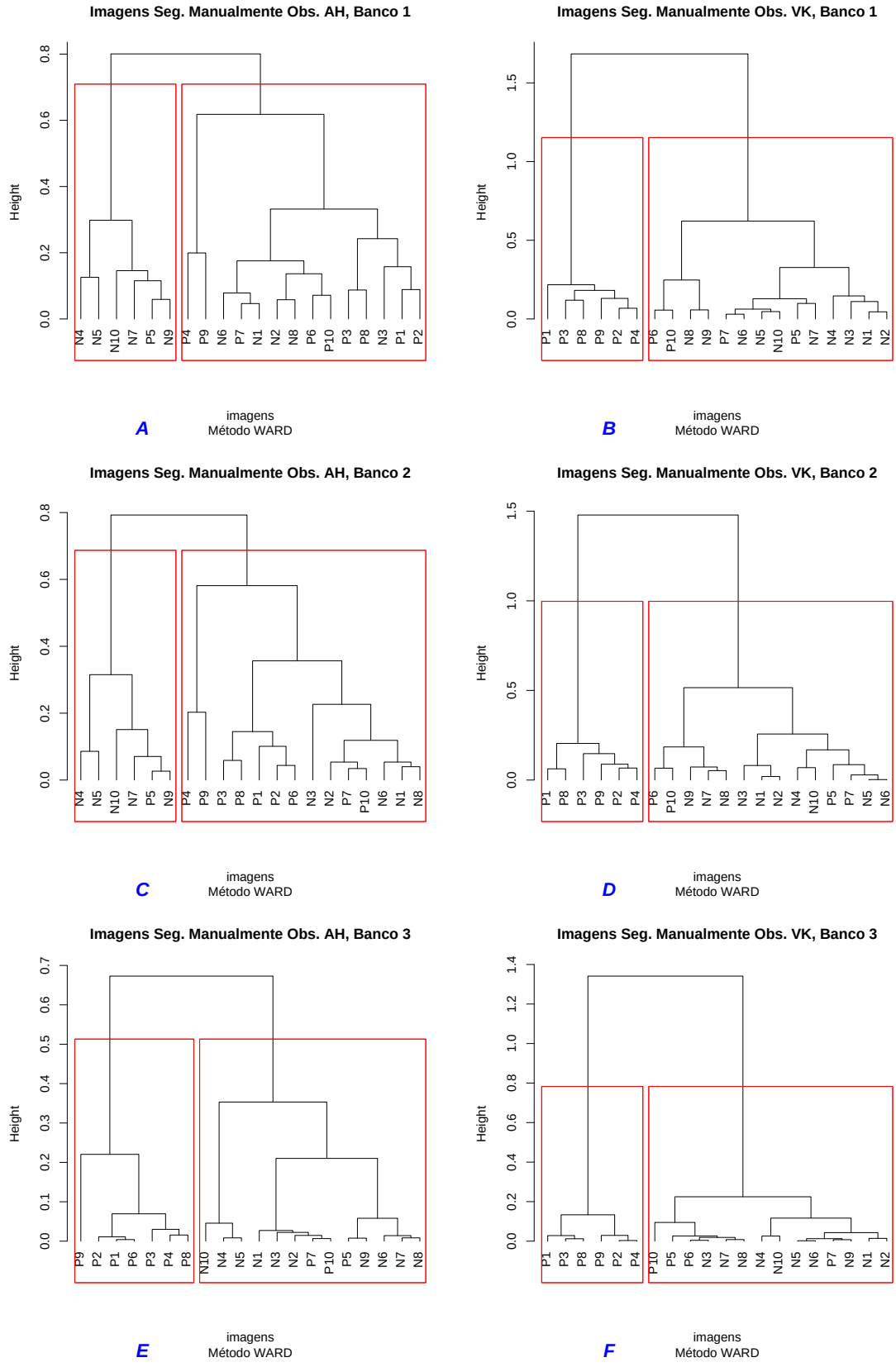

Figura 4.2: Dendrogramas expondo a hierarquia dos agrupamentos formados após a aplicação do método de Ward às imagens segmentadas, para cada um dos bancos e observadores.

4.2 Métodos Não Hierárquicos

Nesta fase do estudo, alguns métodos particionais de agrupamento foram aplicados como **K-médias**, **PAM**, e **Fuzzy c-means**. Primeiramente, foi aplicado o método K-médias, passando posteriormente para o método PAM e concluindo com o Fuzzy c-means, verificando cada um os resultados e validando os resultados através de algumas medidas como a silhueta e silhueta média.

4.2.1 K-médias

Os resultados do método k-médias estão apresentados nas Tabelas 4.2, 4.3 e 4.4. Semelhantes aos resultados exibidos pelo método de Ward, para as imagens esqueletonizadas, as retinas patológicas agrupam-se com 80% (P1,P2,P3,P4,P6,P8,P9,P10) para o observador AH e com 70% (P1,P2,P3,P4,P8,P9,P10) para o observador VK em todos os bancos analisados. Em todos os casos as retinas normais agrupam-se corretamente. No caso das imagens segmentadas manualmente pelo observador AK, os dois grupos representam a mistura de retinas normais e patológicas, enquanto para o observador VK as retinas patológicas se agrupam com 60% (P1,P2,P3,P4,P8,P9) em todos os bancos analisados. Novamente, as imagens esqueletonizadas mostraram-se melhores para detecção de casos patológicos.

Tabela 4.2: Grupos formados pelo algoritmo K-médias referentes ao Banco 1

Observador AH				Observador VK			
Esqueletonizadas		Seg. Manualmente		Esqueletonizadas		Seg. Manualmente	
Grupo 1	Grupo 2	Grupo 1	Grupo 2	Grupo 1	Grupo 2	Grupo 1	Grupo 2
P1	P5	P5	P1	P1	P5	P5	P1
P2	P7	N1	P2	P2	P6	P6	P2
P3	N1	N4	P3	P3	P7	P7	P3
P4	N2	N5	P4	P4	N1	P10	P4
P6	N3	N6	P6	P8	N2	N1	P8
P8	N4	N7	P7	P9	N3	N2	P9
P9	N5	N9	P8	P10	N4	N3	
P10	N6	N10	P9		N5	N4	
	N7		P10		N6	N5	
	N8		N2		N7	N6	
	N9		N3		N8	N7	
	N10		N8		N9	N8	
					N10	N9	
						N10	

Tabela 4.3: Grupos formados pelo algoritmo K-médias referentes ao Banco 2

Observador AH				Observador VK			
Esqueletonizadas		Seg. Manualmente		Esqueletonizadas		Seg. Manualmente	
Grupo 1	Grupo 2	Grupo 1	Grupo 2	Grupo 1	Grupo 2	Grupo 1	Grupo 2
P1	P5	P1	P5	P1	P5	P5	P1
P2	P7	P2	N1	P2	P6	P6	P2
P3	N1	P3	N4	P3	P7	P7	P3
P4	N2	P4	N5	P4	N1	P10	P4
P6	N3	P6	N6	P8	N2	N1	P8
P8	N4	P7	N7	P9	N3	N2	P9
P9	N5	P8	N8	P10	N4	N3	
P10	N6	P9	N9		N5	N4	
	N7	P10	N10		N6	N5	
	N8	N2			N7	N6	
	N9	N3			N8	N7	
	N10				N9	N8	
					N10	N9	
						N10	

Tabela 4.4: Grupos formados pelo algoritmo K-médias referentes ao Banco 3

Observador AH				Observador VK			
Esqueletonizadas		Seg. Manualmente		Esqueletonizadas		Seg. Manualmente	
Grupo 1	Grupo 2	Grupo 1	Grupo 2	Grupo 1	Grupo 2	Grupo 1	Grupo 2
P5	P1	P1	P5	P1	P5	P1	P5
P7	P2	P2	N1	P2	P6	P2	P6
N1	P3	P3	N4	P3	P7	P3	P7
N2	P4	P4	N5	P4	N1	P4	P10
N3	P6	P6	N6	P8	N2	P8	N1
N4	P8	P7	N7	P9	N3	P9	N2
N5	P9	P8	N8	P10	N4		N3
N6	P10	P9	N9		N5		N4
N7		P10	N10		N6		N5
N8		N2			N7		N6
N9		N3			N8		N7
N10					N9		N8
					N10		N9
							N10

Para se obter mais informações a respeito da alocação dos elementos aos grupos formados usam-se os gráficos de silhueta apresentados nas figuras 4.3 a 4.6. Os gráficos expostos no lado esquerdo das figuras exibem as silhuetas dos agrupamentos formados pelo método K-médias. Enquanto no lado direito têm-se os gráficos de silhueta dos elementos agrupados arbitrariamente em dois grupos contendo apenas patológicos e normais. Os valores de $s(i)$ baixos e negativos evidenciam que o objeto está erroneamente alocado a este grupo. Além disso, valores próximos de zero indicam que o objeto pode ser considerado como um elemento intermediário segundo Kaufman e Rousseeuw (1990, pg 86).

Os valores da silhueta média são maiores para as imagens esqueletonizadas do que para as imagens segmentadas manualmente pelo observador AH em todos os bancos. Este resultado indica que as imagens esqueletonizadas têm uma melhor alocação geral, confirmando que estas imagens são mais apropriadas para aplicação do método de agrupamento. No caso do observador VK não existe uma grande diferença entre os valores das silhuetas médias para as imagens esqueletonizadas e segmentadas manualmente, da mesma forma que foi mostrado pelo método de Ward, o método K-médias evidencia a regularidade nas imagens segmentadas por este observador.

Tabela 4.5: Silhueta média dos agrupamentos formados pelo método K-médias, para todos os Bancos

Observadores	Tipos	Banco 1		Banco 2		Banco 3	
AH	Esqueletonizadas	0,42	0,35	0,48	0,48	0,56	0,48
	Seg. Manualmente	0,30	0,19	0,32	0,21	0,51	0,32
VK	Esqueletonizadas	0,53	0,33	0,59	0,37	0,74	0,40
	Seg. Manualmente	0,54	0,28	0,56	0,32	0,76	0,40

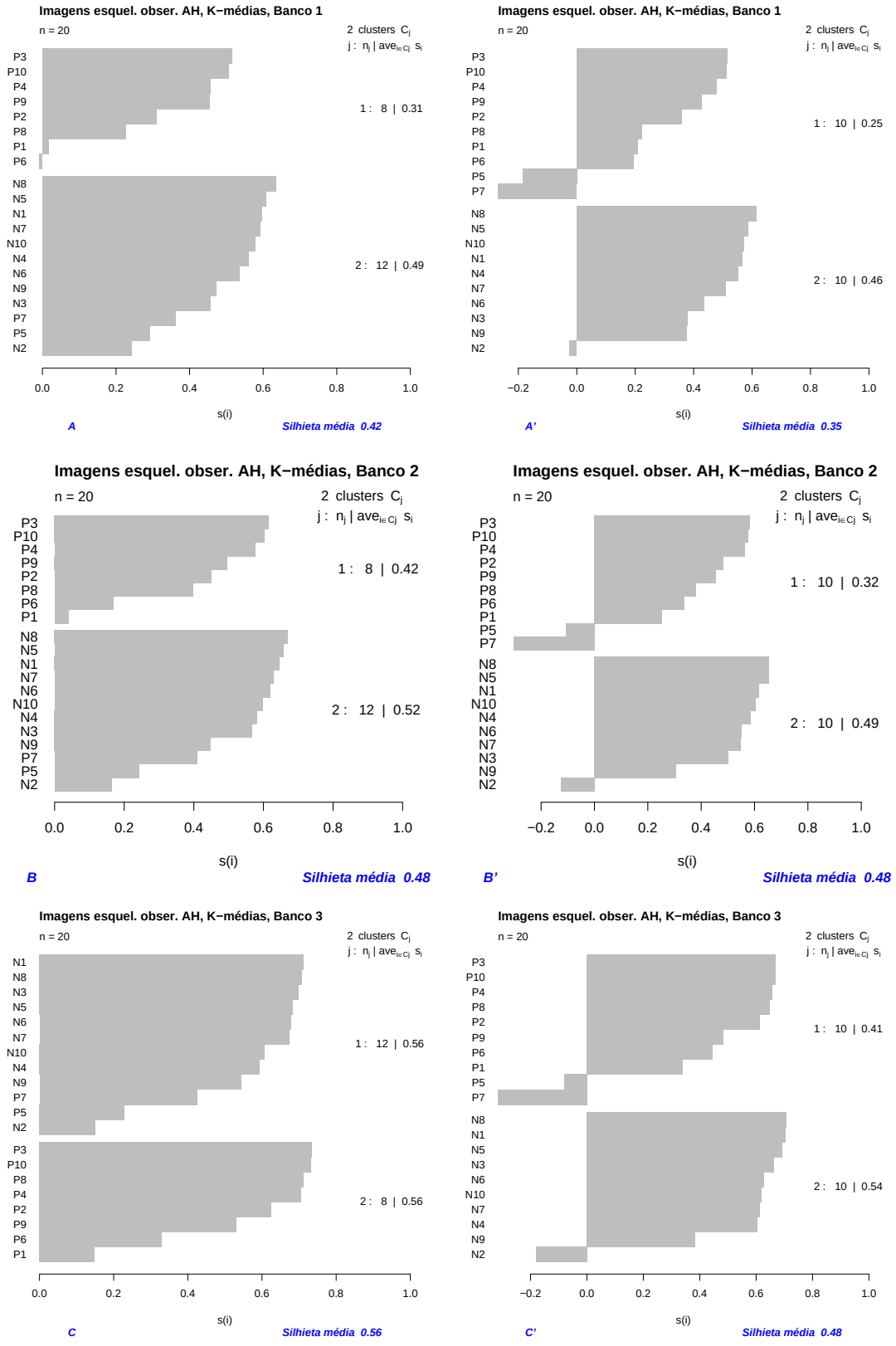

Figura 4.3: Gráficos de silhuetas das imagens esqueletonizadas, observador AH, dos grupos formados pelo algoritmo K-médias.

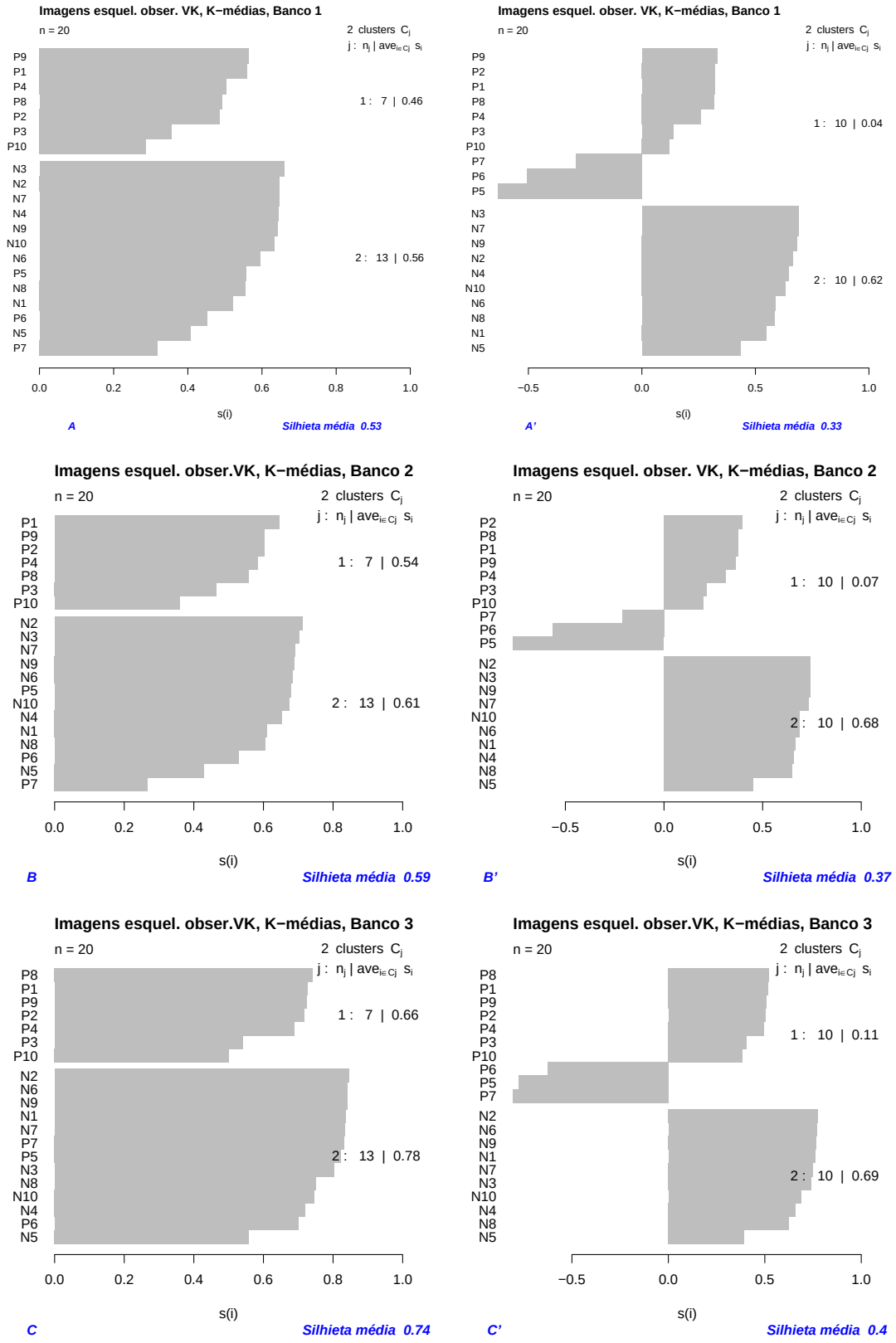

Figura 4.4: Gráficos de silhuetas das imagens esqueletonizadas, observador VK, dos grupos formados pelo algoritmo K-médias.

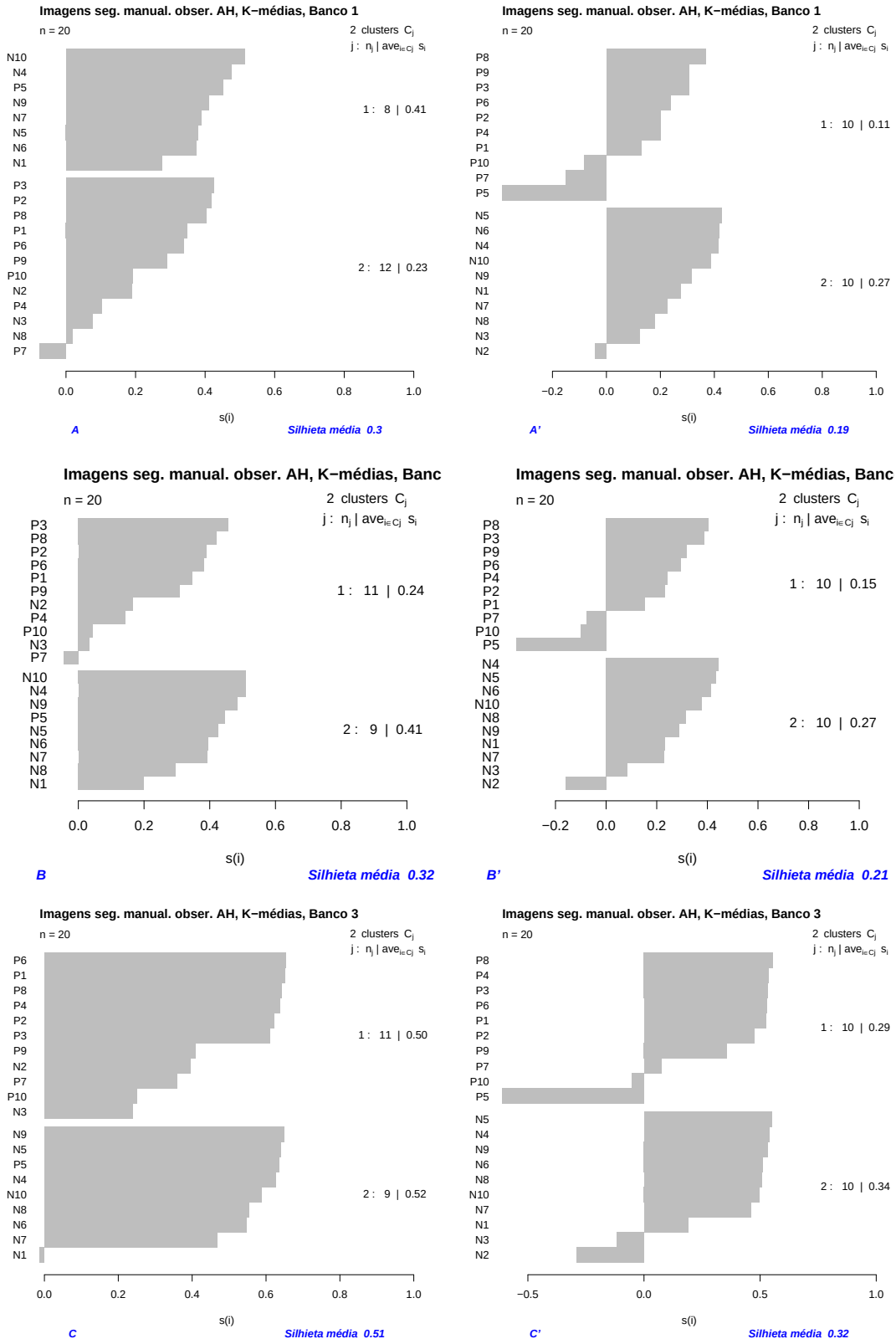

Figura 4.5: Gráficos de silhuetas das imagens segmentadas manualmente, observador AH, dos grupos formados pelo algoritmo K-médias.

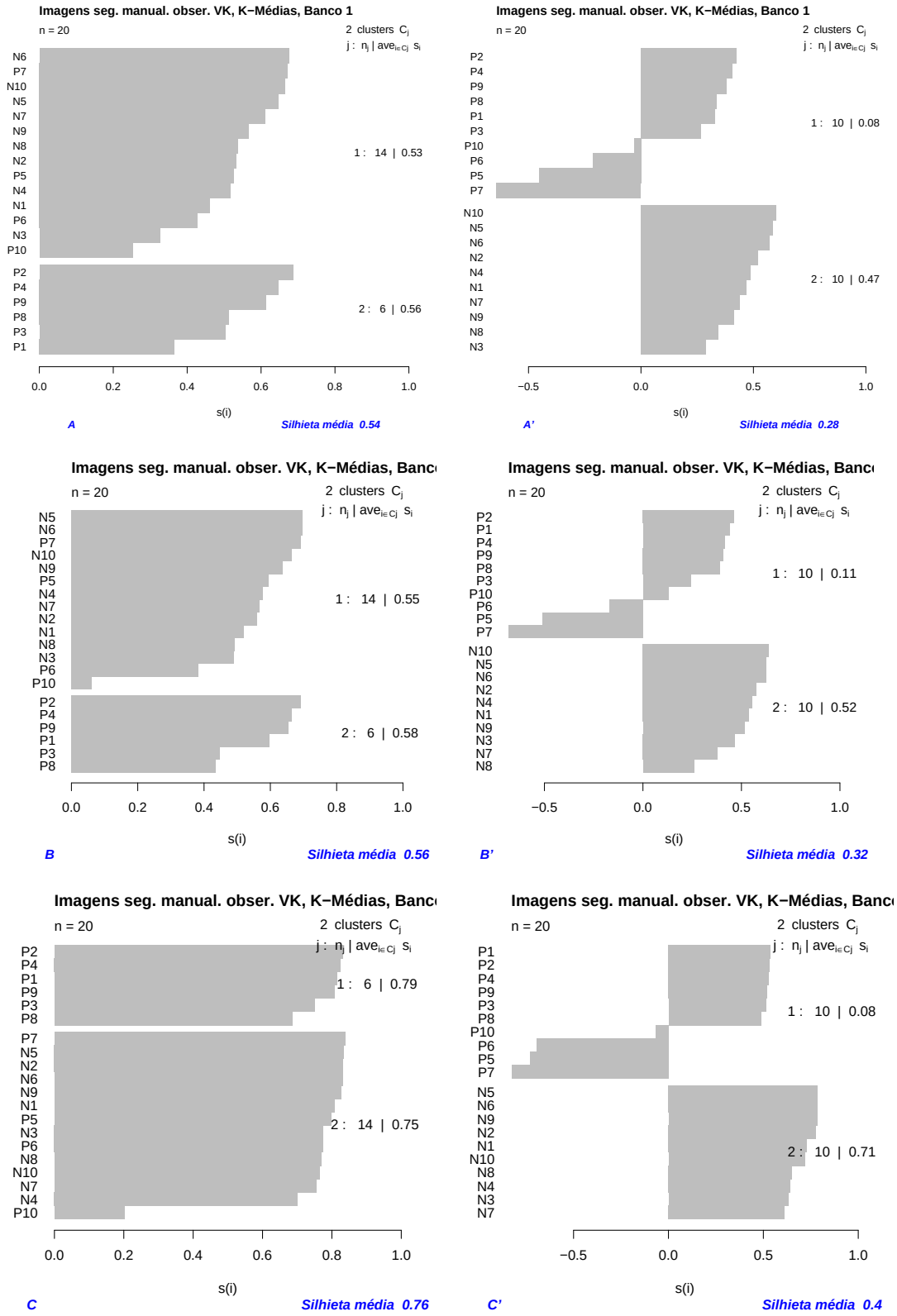

Figura 4.6: Gráficos de silhuetas das imagens segmentadas manualmente, observador VK, dos grupos formados pelo algoritmo K-médias.

4.2.2 PAM e Fuzzy c-means

Os grupos formados por estes métodos de agrupamento, são apresentados nas Tabelas de 4.6 a 4.11. No caso das retinas esqueletonizadas os resultados são idênticos ao método K-médias com exceção do método PAM para banco 2 observador AH onde a imagem (P1) foi alocada ao grupo onde estão alocadas as imagens normais. No caso das imagens segmentadas manualmente surgem outras diferenças em relação ao K-médias. No método PAM banco 1 observador AH, foi obtido um grupo de retinas patológicas com 70% do total, contra uma mistura encontrada para o método K-médias. Já no método FUZZY foram encontradas diferenças para o banco 2, observador VK no grupo onde se encontram apenas imagens patológicas, com acréscimo da imagem P10.

Estas afirmações podem ser confirmadas através da análise dos valores das probabilidades de pertinência expostos nas Tabelas 4.12 a 4.14. Nestas Tabelas, pode-se observar que as maiores probabilidades de pertinência de cada um dos elementos são dadas aos grupos que estão apresentados nas Tabelas 4.9, 4.10 e 4.11, confirmando assim os resuldados dos agrupamentos formados.

Tabela 4.6: Grupos formados pelo algoritmo PAM referentes ao Banco 1

Observador AH				Observador VK			
Esqueletonizadas		Seg. Manualmente		Esqueletonizadas		Seg. Manualmente	
Grupo 1	Grupo 2	Grupo 1	Grupo 2	Grupo 1	Grupo 2	Grupo 1	Grupo 2
P1	P5	P1	P5	P1	P5	P1	P5
P2	P7	P2	P7	P2	P6	P2	P6
P3	N1	P3	P10	P3	P7	P3	P7
P4	N2	P4	N1	P4	N1	P4	P10
P6	N3	P6	N2	P8	N2	P8	N1
P8	N4	P8	N3	P9	N3	P9	N2
P9	N5	P9	N4	P10	N4		N3
P10	N6		N5		N5		N4
	N7		N6		N6		N5
	N8		N7		N7		N6
	N9		N8		N8		N7
	N10		N9		N9		N8
			N10		N10		N9
							N10

Tabela 4.7: Grupos formados pelo algoritmo PAM referentes ao Banco 2

Observador AH				Observador VK			
Esqueletonizadas		Seg. Manualmente		Esqueletonizadas		Seg. Manualmente	
Grupo 1	Grupo 2	Grupo 1	Grupo 2	Grupo 1	Grupo 2	Grupo 1	Grupo 2
P1	P2	P1	P4	P1	P5	P1	P5
P5	P3	P2	P5	P2	P6	P2	P6
P7	P4	P3	N4	P3	P7	P3	P7
N1	P6	P6	N7	P4	N1	P4	P10
N2	P8	P7	N9	P8	N2	P8	N1
N3	P9	P8	N10	P9	N3	P9	N2
N4	P10	P9		P10	N4		N3
N5		P10			N5		N4
N6		N1			N6		N5
N7		N2			N7		N6
N8		N3			N8		N7
N9		N5			N9		N8
N10		N6			N10		N9
		N8					N10

Tabela 4.8: Grupos formados pelo algoritmo PAM referentes ao Banco 3

Observador AH				Observador VK			
Esqueletonizadas		Seg. Manualmente		Esqueletonizadas		Seg. Manualmente	
Grupo 1	Grupo 2	Grupo 1	Grupo 2	Grupo 1	Grupo 2	Grupo 1	Grupo 2
P1	P5	P1	P5	P1	P5	P1	P5
P2	P7	P2	N1	P2	P6	P2	P6
P3	N1	P3	N4	P3	P7	P3	P7
P4	N2	P4	N5	P4	N1	P4	P10
P6	N3	P6	N6	P8	N2	P8	N1
P8	N4	P7	N7	P9	N3	P9	N2
P9	N5	P8	N8	P10	N4		N3
P10	N6	P9	N9		N5		N4
	N7	P10	N10		N6		N5
	N8	N2			N7		N6
	N9	N3			N8		N7
	N10				N9		N8
					N10		N9
							N10

Tabela 4.9: Grupos formados pelo algoritmo FUZZY referentes ao Banco 1

Observador AH				Observador VK			
Esqueletonizadas		Seg. Manualmente		Esqueletonizadas		Seg. Manualmente	
Grupo 1	Grupo 2	Grupo 1	Grupo 2	Grupo 1	Grupo 2	Grupo 1	Grupo 2
P5	P1	P1	P5	P5	P1	P1	P5
P7	P2	P2	P7	P6	P2	P2	P6
N1	P3	P3	N1	P7	P3	P3	P7
N2	P4	P4	N4	N1	P4	P4	P10
N3	P6	P6	N5	N2	P8	P8	N1
N4	P8	P8	N6	N3	P9	P9	N2
N5	P9	P9	N7	N4	P10		N3
N6	P10	P10	N8	N5			N4
N7		N2	N9	N6			N5
N8		N3	N10	N7			N6
N9				N8			N7
N10				N9			N8
				N10			N9
							N10

Tabela 4.10: Grupos formados pelo algoritmo FUZZY referentes ao Banco 2

Observador AH				Observador VK			
Esqueletonizadas		Seg. Manualmente		Esqueletonizadas		Seg. Manualmente	
Grupo 1	Grupo 2	Grupo 1	Grupo 2	Grupo 1	Grupo 2	Grupo 1	Grupo 2
P5	P1	P5	P1	P5	P1	P5	P1
P7	P2	N1	P2	P6	P2	P6	P2
N1	P3	N4	P3	P7	P3	P7	P3
N2	P4	N5	P4	N1	P4	N1	P4
N3	P6	N6	P6	N2	P8	N2	P8
N4	P8	N7	P7	N3	P9	N3	P9
N5	P9	N8	P8	N4	P10	N4	P10
N6	P10	N9	P9	N5		N5	
N7		N10	P10	N6		N6	
N8			N2	N7		N7	
N9			N3	N8		N8	
N10				N9		N9	
				N10		N10	

Tabela 4.11: Grupos formados pelo algoritmo FUZZY referentes ao Banco 3

Observador AH				Observador VK			
Esqueletonizadas		Seg. Manualmente		Esqueletonizadas		Seg. Manualmente	
Grupo 1	Grupo 2	Grupo 1	Grupo 2	Grupo 1	Grupo 2	Grupo 1	Grupo 2
P5	P1	P5	P1	P1	P5	P5	P1
P7	P2	N1	P2	P2	P6	P6	P2
N1	P3	N4	P3	P3	P7	P7	P3
N2	P4	N5	P4	P4	N1	P10	P4
N3	P6	N6	P6	P8	N2	N1	P8
N4	P8	N7	P7	P9	N3	N2	P9
N5	P9	N8	P8	P10	N4	N3	
N6	P10	N9	P9		N5	N4	
N7		N10	P10		N6	N5	
N8			N2		N7	N6	
N9			N3		N8	N7	
N10					N9	N8	
					N10	N9	
						N10	

Tabela 4.12: Probabilidade de pertinência para o Banco 1

Retinas	Observador AH						Observador VK					
	Esqueletonizadas			Seg. Manualmente			Esqueletonizadas			Seg. Manualmente		
	1	2	Escolha	1	2	Escolha	1	2	Escolha	1	2	Escolha
P1	0,34	0,66	2,00	0,76	0,24	1,00	0,02	0,98	2,00	0,81	0,19	1,00
P2	0,20	0,80	2,00	0,88	0,12	1,00	0,15	0,85	2,00	0,98	0,02	1,00
P3	0,01	0,99	2,00	0,91	0,09	1,00	0,17	0,83	2,00	0,89	0,11	1,00
P4	0,01	0,99	2,00	0,59	0,41	1,00	0,08	0,92	2,00	0,93	0,07	1,00
P5	0,68	0,32	1,00	0,14	0,86	2,00	0,91	0,09	1,00	0,08	0,92	2,00
P6	0,28	0,72	2,00	0,91	0,09	1,00	0,86	0,14	1,00	0,21	0,79	2,00
P7	0,76	0,24	1,00	0,45	0,55	2,00	0,69	0,31	1,00	0,01	0,99	2,00
P8	0,33	0,67	2,00	0,91	0,09	1,00	0,14	0,86	2,00	0,92	0,08	1,00
P9	0,14	0,86	2,00	0,69	0,31	1,00	0,08	0,92	2,00	0,95	0,05	1,00
P10	0,01	0,99	2,00	0,72	0,28	1,00	0,20	0,80	2,00	0,34	0,66	2,00
N1	0,94	0,06	1,00	0,10	0,90	2,00	0,87	0,13	1,00	0,18	0,82	2,00
N2	0,61	0,39	1,00	0,70	0,30	1,00	0,99	0,01	1,00	0,11	0,89	2,00
N3	0,82	0,18	1,00	0,54	0,46	1,00	0,96	0,04	1,00	0,25	0,75	2,00
N4	0,86	0,14	1,00	0,17	0,83	2,00	0,94	0,06	1,00	0,15	0,85	2,00
N5	0,94	0,06	1,00	0,23	0,77	2,00	0,84	0,16	1,00	0,03	0,97	2,00
N6	0,93	0,07	1,00	0,07	0,93	2,00	0,96	0,04	1,00	0,01	0,99	2,00
N7	0,97	0,03	1,00	0,15	0,85	2,00	0,99	0,01	1,00	0,03	0,97	2,00
N8	0,96	0,04	1,00	0,49	0,51	2,00	0,92	0,08	1,00	0,13	0,87	2,00
N9	0,84	0,16	1,00	0,16	0,84	2,00	0,97	0,03	1,00	0,12	0,88	2,00
N10	0,88	0,12	1,00	0,17	0,83	2,00	0,93	0,07	1,00	0,03	0,97	2,00

Assim, é possível perceber que nos métodos PAM e Fuzzy as imagens esqueletonizadas possuem resultados que apresentam melhores alocações do que imagens segmentadas manualmente. E ainda, as imagens segmentadas pelo observador AH possuem

Tabela 4.13: Probabilidade de pertinência para o Banco 2

Retinas	Observador AH						Observador VK					
	Esqueletonizadas			Seg. Manualmente			Esqueletonizadas			Seg. Manualmente		
	1	2	Escolha	1	2	Escolha	1	2	Escolha	1	2	Escolha
P1	0,36	0,64	2,00	0,21	0,79	2,00	0,02	0,98	2,00	0,04	0,96	2,00
P2	0,11	0,89	2,00	0,10	0,90	2,00	0,09	0,91	2,00	0,03	0,97	2,00
P3	0,01	0,99	2,00	0,06	0,94	2,00	0,13	0,87	2,00	0,17	0,83	2,00
P4	0,00	1,00	2,00	0,37	0,63	2,00	0,07	0,93	2,00	0,06	0,94	2,00
P5	0,64	0,36	1,00	0,83	0,17	1,00	0,98	0,02	1,00	0,94	0,06	1,00
P6	0,25	0,75	2,00	0,03	0,97	2,00	0,91	0,09	1,00	0,72	0,28	1,00
P7	0,78	0,22	1,00	0,48	0,52	2,00	0,66	0,34	1,00	1,00	0,00	1,00
P8	0,19	0,81	2,00	0,07	0,93	2,00	0,10	0,90	2,00	0,12	0,88	2,00
P9	0,15	0,85	2,00	0,29	0,71	2,00	0,09	0,91	2,00	0,06	0,94	2,00
P10	0,02	0,98	2,00	0,41	0,59	2,00	0,17	0,83	2,00	0,50	0,50	2,00
N1	0,97	0,03	1,00	0,80	0,20	1,00	0,93	0,07	1,00	0,88	0,12	1,00
N2	0,56	0,44	1,00	0,27	0,73	2,00	1,00	0,00	1,00	0,90	0,10	1,00
N3	0,93	0,07	1,00	0,47	0,53	2,00	0,98	0,02	1,00	0,88	0,12	1,00
N4	0,88	0,12	1,00	0,85	0,15	1,00	0,93	0,07	1,00	0,89	0,11	1,00
N5	0,96	0,04	1,00	0,79	0,21	1,00	0,84	0,16	1,00	0,99	0,01	1,00
N6	0,97	0,03	1,00	0,90	0,10	1,00	0,99	0,01	1,00	0,99	0,01	1,00
N7	0,98	0,02	1,00	0,88	0,12	1,00	0,99	0,01	1,00	0,92	0,08	1,00
N8	0,98	0,02	1,00	0,86	0,14	1,00	0,94	0,06	1,00	0,83	0,17	1,00
N9	0,81	0,19	1,00	0,87	0,13	1,00	0,98	0,02	1,00	0,95	0,05	1,00
N10	0,89	0,11	1,00	0,83	0,17	1,00	0,95	0,05	1,00	0,96	0,04	1,00

Tabela 4.14: Probabilidade de pertinência para o Banco 3

Retinas	Observador AH						Observador VK					
	Esqueletonizadas			Seg. Manualmente			Esqueletonizadas			Seg. Manualmente		
	1	2	Escolha	1	2	Escolha	1	2	Escolha	1	2	Escolha
P1	0,32	0,68	2,00	0,00	1,00	2,00	0,99	0,01	1,00	0,00	1,00	2,00
P2	0,03	0,97	2,00	0,02	0,98	2,00	0,95	0,05	1,00	0,02	0,98	2,00
P3	0,01	0,99	2,00	0,05	0,95	2,00	0,87	0,13	1,00	0,03	0,97	2,00
P4	0,00	1,00	2,00	0,01	0,99	2,00	0,97	0,03	1,00	0,02	0,98	2,00
P5	0,62	0,38	1,00	0,98	0,02	1,00	0,01	0,99	2,00	0,99	0,01	1,00
P6	0,21	0,79	2,00	0,00	1,00	2,00	0,06	0,94	2,00	0,97	0,03	1,00
P7	0,78	0,22	1,00	0,24	0,76	2,00	0,01	0,99	2,00	1,00	0,00	1,00
P8	0,02	0,98	2,00	0,02	0,98	2,00	0,97	0,03	1,00	0,05	0,95	2,00
P9	0,14	0,86	2,00	0,22	0,78	2,00	0,96	0,04	1,00	0,02	0,98	2,00
P10	0,01	0,99	2,00	0,32	0,68	2,00	0,84	0,16	1,00	0,62	0,38	1,00
N1	1,00	0,00	1,00	0,57	0,43	1,00	0,00	1,00	2,00	0,99	0,01	1,00
N2	0,57	0,43	1,00	0,19	0,81	2,00	0,00	1,00	2,00	1,00	0,00	1,00
N3	0,99	0,01	1,00	0,32	0,68	2,00	0,02	0,98	2,00	0,97	0,03	1,00
N4	0,89	0,11	1,00	0,91	0,09	1,00	0,06	0,94	2,00	0,93	0,07	1,00
N5	0,97	0,03	1,00	0,93	0,07	1,00	0,14	0,86	2,00	1,00	0,00	1,00
N6	0,98	0,02	1,00	0,94	0,06	1,00	0,00	1,00	2,00	0,99	0,01	1,00
N7	0,98	0,02	1,00	0,87	0,13	1,00	0,01	0,99	2,00	0,96	0,04	1,00
N8	0,99	0,01	1,00	0,93	0,07	1,00	0,04	0,96	2,00	0,97	0,03	1,00
N9	0,88	0,12	1,00	0,99	0,01	1,00	0,00	1,00	2,00	0,99	0,01	1,00
N10	0,90	0,10	1,00	0,89	0,11	1,00	0,05	0,95	2,00	0,96	0,04	1,00

silhuetas médias menores do que as imagens segmentadas pelo observador VK, para as imagens segmentadas manualmente, para todos os bancos.

Para verificar a qualidade dos grupos formados podem-se usar as silhuetas médias apresentadas nas Tabelas 4.15 e 4.16. Nestas tabelas, as silhuetas médias das imagens esqueletonizadas são maiores do que as das imagens segmentadas manualmente como também as silhuetas médias do observador AH são menores que as silhuetas das imagens segmentadas pelo observador VK, ratificado assim as afirmações delineadas no parágrafo anterior.

Tabela 4.15: Silhueta média dos agrupamentos formados pelo método PAM para todos os Bancos

Observadores	Tipos	Banco 1		Banco 2		Banco 3	
AH	Esqueletonizada	0,42	0,35	0,48	0,48	0,56	0,48
	Seg. Manualmente	0,32	0,19	0,26	0,21	0,51	0,32
VK	Esqueletonizada	0,53	0,33	0,59	0,37	0,74	0,40
	Seg. Manualmente	0,54	0,28	0,56	0,32	0,76	0,40

Tabela 4.16: Silhueta média dos agrupamentos formados pelo método Fuzzy para todos os Bancos

Observadores	Tipos	Banco 1		Banco 2		Banco 3	
AH	Esqueletonizada	0,42	0,35	0,48	0,48	0,56	0,48
	Seg. Manualmente	0,28	0,19	0,32	0,21	0,51	0,32
VK	Esqueletonizada	0,53	0,33	0,59	0,37	0,74	0,40
	Seg. Manualmente	0,54	0,28	0,54	0,32	0,76	0,40

Observando todos os agrupamentos formados pelos métodos hierárquicos e não hierárquicos, nota-se que quase sempre as mesmas imagens deslocam-se do grupo imagens patológicas para o grupo de imagens normais. Usando esta quantidade para as imagens que se deslocaram como medida de erro, pode-se fazer uma análise mais profunda dos resultados obtidos. Estas imagens estão expostas nas Tabelas 4.17 e 4.18. Assim, para as imagens esqueletonizadas, observam-se que as imagens patológicas (P5,P7; 20%) são deslocadas do seu grupo para o grupo de imagens normais em todos os métodos exceto no método de Ward no Banco 2 para o observador AH. Já para o observador VK, as imagens deslocadas para todos os métodos de agrupamento são as (P5,P6,P7; 30%). Analisando a Tabela 4.18, observa-se que para o observador AH apenas a imagem (P5; 10%) é deslocada para todos os bancos, com exceção do Banco 2 para o método PAM. E do mesmo modo que ocorreu para o observador VK para as imagens esqueletonizadas, tem-se uma repetição das (P5,P6,P7,P10; 10%) para todos os métodos, exceto para o método Fuzzy

c-means no Banco 2. Este comportamento indica que as imagens (P5,P7) e talvez as (P6 e P10) sejam imagens de retinas com a presença de um tipo de patologia que não altere a árvore vascular da retina. Considerando esta alocação errônea como uma margem de erro, e usando como medida de comparação, as imagens esqueletonizadas produzem melhores resultados, bem como o observador VK é mais constante em suas segmentações.

Tabela 4.17: Imagens esqueletonizadas alocadas em grupo oposto ao tipo ao qual fazem parte

	Ward			K-Médias			PAM			Fuzzy-cmeans		
Bancos	1	2	3	1	2	3	1	2	3	1	2	3
AH	P5 P6 P7	N7 N9	P1 P5 P6 P7	P5 P7	P5 P7	P5 P7	P5 P7	P1 P5 P7	P5 P7	P5 P7	P5 P7	P5 P7
VK	P5 P6 P7	P5 P6 P7	P5 P6 P7	P5 P6 P7	P5 P6 P7	P5 P6 P7	P5 P6 P7	P5 P6 P7	P5 P6 P7	P5 P6 P7	P5 P6 P7	P5 P6 P7

Tabela 4.18: Imagens segmentadas manualmente alocadas em grupo oposto ao tipo ao qual fazem parte

Métodos	Ward			K-Médias			PAM			Fuzzy-cmeans		
Bancos	1	2	3	1	2	3	1	2	3	1	2	3
AH	P5 N1 N2 N3 N6 N8	P5 N1 N2 N3 N6 N8	P5 P7 P10	P5 N2 N3 N8	P5 N2 N3	P5 N2 N3	P5 P7 P10	N1 N2 N3 N5 N6 N8 P4 P5	P5 N2 N3	P5 P7 N2 N3	P5 N2 N3	P5 N2 N3
VK	P5 P6 P7 P10	P5 P6 P7 P10	P5 P6 P7 P10	P5 P6 P7 P10	P5 P6 P7 P10	P5 P6 P7 P10	P5 P6 P7 P10	P5 P6 P7 P10	P5 P6 P7 P10	P5 P6 P7 P10	P5 P6 P7	P5 P6 P7 P10

5 Conclusões

A inspeção do sistema vascular da retina humana é extremamente importante para detecção de doenças como retinopatia diabética e oclusão de vasos causada pela hipertensão e arteriosclerose. Estas doenças se manifestam como alterações em vasos sangüíneos da retina e têm como conseqüência a diminuição da visão do paciente e em estados avançados podem levar a cegueira. Para prevenir danos irreversíveis é necessária a detecção precoce dessas doenças. Isso pode ser feito através de exames periódicos incluindo a inspeção da imagem da retina obtida pelo fundus câmera, angiografia, juntamente com outros aparelhos de captação de imagens que estão em constante evolução. Um método automático composto de segmentação dos vasos, e posterior análise de vasos segmentados usando modelos matemáticos físicos e estatísticos, pode ser usado na detecção dessas doenças, mas ainda representa um desafio para ciência. Durante a década passada foram feitas várias tentativas com o uso da dimensão fractal para descrever e quantificar as propriedades geométricas do sistema vascular da retina humana (MASTERS, 2004). Os resultados ainda não são conclusivos principalmente porque não existe um método eficiente e preciso de segmentação automática de vasos a partir da imagem obtida pelo fundus câmera, angiografia e outros aparelhos. Na maioria dos estudos foram usadas imagens segmentadas manualmente que incluem fatores subjetivos como o nível dos detalhes de segmentação e que depende do treinamento do observador.

Recentemente, foi mostrado que o sistema vascular da retina possui uma complexidade maior do que um fractal simples representando um multifractal geométrico caracterizado pela hierarquia de expoentes e espectro multifractal não trivial (STOSIC; STOSIC, 2006). Os resultados da análise multifractal de retinas patológicas e retinas normais indicam que essa pode ser usada para detectar casos patológicos. Para analisar essa possibilidade, aplicam-se métodos de agrupamento nos resultados da análise multifractal das imagens de retinas segmentadas manualmente e também das imagens esqueletonizadas. As variáveis usadas no agrupamento foram as dimensões fractais generalizadas e os elementos de espectro multifractal, dos quais foram escolhidos três conjuntos distintos. Os métodos de agrupamento usados foram Ward, K-médias, PAM e Fuzzy c-means. Para

avaliar os grupos formados foram usados a correlação cofenética e o gráfico da silhueta. Os resultados obtidos após a aplicação dos quatro métodos de agrupamento usando como variáveis três conjuntos de dados extraídos dos resultados da análise multifractal das imagens segmentadas manualmente pelo dois observadores e também das mesmas imagens esqueletonizadas (sendo 10 imagens de retinas normais e 10 de retinas patológicas) mostraram que:

a) As imagens esqueletonizadas são mais apropriadas para identificação de casos patológicos que as imagens segmentadas manualmente. Para imagens esqueletonizadas, 70-80% das retinas patológicas (dependendo do método e conjunto de variáveis usadas) foram agrupadas corretamente, enquanto que as imagens segmentadas manualmente os resultados não foram consistentes. Este fato indica que o comprimento de vasos e suas ramificações são fatores mais relevantes para as conclusões da análise atual do que a largura dos vasos.

b) O fato do deslocamento das retinas patológicas P5, P6 e P7 para o grupo de retinas normais, observado em todos os casos de agrupamento (usando as imagens esqueletonizadas), pode significar que essas retinas não possuem patologias que causam alterações nos vasos. O banco de dados STARE de onde se originaram as imagens usadas no estudo atual, não disponibiliza informações sobre o tipo das patologias.

c) A diferença nos resultados de agrupamento para as retinas segmentadas pelo dois observadores, indica a necessidade de desenvolvimento de algoritmos eficientes de segmentação automática que deverão provavelmente eliminar os fatores subjetivos (Influência de Julgadores) presentes na segmentação manual. Além disso, a segmentação automática possibilitará o uso de amostras de tamanho maior, aumentando a consistência das análises estatísticas.

Finalmente, é possível concluir que a análise multifractal (com pré-processamento adequado das imagens e escolha das variáveis) pode ser utilizada para detecção de casos patológicos da retina humana.

Uma seqüência natural desse trabalho seria analisar uma patologia específica (e.g. retinopatia diabética) usando amostras de tamanho maior e outros métodos de agrupamento (e.g. Análise de Discriminante).

Referências Bibliográficas

ANDERSON, T. W. **An Introduction to Multivariate Statistical Analysis**. 3. ed. New York: John Wiley and Sons, 2003. 742 p.

BARROSO, L. P.; ARTES, R. **Análise Multivariada**. Lavras: UFLA, 2003. 156 p.

BASSINGTHWAIGHTE, J. B.; LIEBOVITCH, L.; WEST, B. J. **Fractal Physiology**. New York: Oxford University Press, 1994. 384 p.

BEZDEK, J. C. **Pattern Recognition with Fuzzy Objective Function Algorithms**. Norwell, MA, USA: Kluwer Academic Publishers, 1981. ISBN 0306406713.

BUSSAB, W. de O.; MIAZAKI, E. S.; ANDRADE, D. F. de. Introdução a análise de agrupamentos. In: **simpósio nacional de probabilidade e estística (SINAPE)**. São Paulo: ABE, 1990. p. 105.

FEDER, J. **Fractals**. New York: Plenum, 1988. 283 p.

GUYTON, A. C. **Neurociência Básica**. Rio de Janeiro: Editora Guanabara Koogan S.A., 1993. 345 p.

HAIR, J. F. et al. **Análise Multivariada de Dados**. 5. ed. Porto Alegre: Bookman, 2005. 137 p.

HALSEY, T. C. et al. Fractal measures and their singularities: The characterization of strange sets. **Physical Review A**, American Physical Society, United States, v. 33, n. 2, p. 1141–1151, Feb 1986.

HARTIGAN, J. A. **Clustering Algorithms**. New York: John Wiley & Sons, 1975. 137 p.

HARTIGAN, J. A.; WON, M. A. A k-means clustering algorithm. **Journal of the Royal Statistical Society**, Royal Statistical Society, England, v. 28, p. 100–108, 1979.

HAYAKAWA, Y.; SATO, S.; MATSUSHITA, M. Scaling structure of the growth-probability distribution in diffusion-limited aggregation processes. **Physical Review. A**, American Physical Society, United States, v. 36, n. 4, p. 1963–1966, Aug 1987.

JAIN, A. K.; DUIN, R. P. W.; MAO, J. Statistical pattern recognition: A review. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 22, n. 1, p. 4–37, 2000.

KAUFMAN, L.; ROUSSEEUW, P. J. **Finding Groups in Data: an introduction to Cluster Analysis**. New York: John Wiley & Sons, 1990.

MANDELBROT, B. B. **The Fractal Geometry of Nature**. San Francisco: Freeman, 1982. 468 p.

MARDIA, K. V.; KENT, J. T.; M.BIBBY, J. **Multivariate Analysis**. 3. ed. London: Academic Press, 1979. 521 p.

MASTERS, B. R. Fractal analysis of the vascular tree in the human retina. **Annual Review of Biomedical Engineering**, Annual Reviews, United States, v. 6, p. 427–452, abril 2004.

MINGOTI, S. A. **Análise de Agrupamento Através de métodos de Estatística Multivariada: uma abordagem aplicada**. Belo Horizonte: Editora UFMG, 2005.

NITTMANN, J. et al. Experimental evidence for multifractality. **Physical review letters**, American Physical Society, United States, v. 58, n. 6, p. 619, Feb 1987.

R Development Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2007. ISBN 3-900051-07-0. Disponível em: <http://www.R-project.org>.

ROUSSEEUW, P. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics**, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, v. 20, n. 1, p. 53–65, 1987. ISSN 0377-0427.

STOSIC, T.; STOSIC, B. Multfractal analysis of human retinal vessels. **IEEE Transactions on Medical Imaging**, United States, v. 25, p. 1101–1107, 2006.

TÉL, T.; FULLOP, A.; VICSEK, T. Determination of fractal dimensions for geometrical multifractals. **Physica A Statistical Mechanics and its Applications**, Netherlands, v. 159, p. 155–166, ago. 1989.

VICSEK, T. **Fractal Growth Phenomena**. 2. ed. Singapore: World Scientific, 1993. 488 p.

VICSEK, T.; FAMILY, F.; MEAKIN, P. Multifractal geometry of diffusion-limited aggregates. **Europhysics Letters**, France, v. 12, p. 217–222, jun. 1990.

WARD, J. H. Hierarchical grouping to optimize an objective function. **Journal of the American Statistical Association**, v. 58, p. 236–244, 1963.

WITTEN, T. A.; SANDER, L. M. Diffusion-limited aggregation, a kinetic critical phenomenon. **Physical review letters**, American Physical Society, United States, v. 47, n. 19, p. 1400–1403, Nov 1981.