

RAE - CEA - 9620
RELATÓRIO DE ANÁLISE ESTATÍSTICA SOBRE
O PROJETO: “IMPLICAÇÕES TAXONÔMICAS DA
DISTRIBUIÇÃO DE ALCANOS E DE
FLAVONÓIDES FOLIARES DE *HUBERIA*
D.C.(MELASTOMATACEAE).”

Dalton Francisco de Andrade

Jefferson Ferreira

São Paulo, novembro de 1996.

CENTRO DE ESTATÍSTICA APLICADA - CEA
RELATÓRIO DE ANÁLISE ESTATÍSTICA

TÍTULO : Relatório de análise estatística sobre o projeto: "Implicações taxonômicas da distribuição de alcanos e de flavonóides foliares de *Huberia* D.C. (Melastomataceae)."

PESQUISADORA : Maria Regina Martins Mimura

COLABORADORA : Maria Luiza Faria Salatino

INSTITUIÇÃO : Instituto de Biociências - USP

FINALIDADE : Publicação

RESPONSÁVEIS PELA ANÁLISE : Dalton Francisco de Andrade
Jefferson Ferreira

REFERÊNCIA DESTE TRABALHO : Andrade, D.F. e Ferreira, J. **Relatório de análise estatística sobre o projeto: "Implicações taxonômicas da distribuições de alcanos e flavonóides foliares de *Huberia* D.C. (Melastomataceae)".** São Paulo, IME-USP, 1996.

FICHA TÉCNICA :

REFERÊNCIA BIBLIOGRÁFICA :

- Bussab, W.O. , Miazaki, E.S. e Andrade, D.F. (1990). **Introdução à Análise de Agrupamentos.** IME-USP, São Paulo ; 105p.
- Everitt, B. (1980). **Cluster Analysis.** 2^a ed. Heinemann Educational Books, London; 136p.

- Gower, J.C. (1971). A general coefficient of similarity and some of its properties. **Biometrics**, 27, 857-872.
- Johnson, R.A e Wichern,D.W. (1982). **Applied Multivariate Statistical Analysis**. Prentice Hall, New Jersey; 642p.
- Mardia, K.V. , Kent, J.T. e Bibby,J. (1979). **Multivariate Analysis**. Academic Press, London; 521p.
- **Minitab for Windows - Reference Manual, Release 10** (1994) State College,PA. Minitab Inc. 1072p.
- Lopes, C.R. e Shepherd, G.P. (1991). Phylogenetic studies of some species of the genus *Coffea* - numerical analysis of Flavonoid compounds. **Revista Brasileira de Genética**, 14, 2,p. 425-435.
- Rettig, J.H, e Giannasi, D.E (1990). Foliar Flavonoids of the *Carex nigromarginata* Complex (sect. *Acrocystis*, *Cyperaceae*). **Biochemical Systematics and Ecology**, 18, 6, p. 393-397.
- **SAS Institute Inc. SAS Language Guide, Release 6.03** Edition. (1988) Cary, NC: SAS institute Inc.
- **SPSS/PC+ for Windows - Reference Manual, Release 6.0** (1993). SPSS Inc.

PROGRAMAS COMPUTACIONAIS UTILIZADOS

- Excel v 5.0
- Minitab for Windows v 10.2
- NTSYS v 1.70
- SAS for Windows v 6.08
- SPSS for Windows v 6.0
- Word for Windows v 6.0

TÉCNICAS ESTATÍSTICAS UTILIZADAS

- Análise Descritiva Unidimensional (03:010)
- Análise de Agrupamentos (06:120)

ÁREA DE APLICAÇÃO

- Botânica (14:060)

Índice

Resumo.....	6
1. Introdução	7
2. Descrição do Estudo.....	8
3. Descrição das Variáveis	8
4. Análise Exploratória dos Dados.....	9
4.1 Ceras Epicuticulares.....	9
4.2 Flavonóides	10
5. Análise de Agrupamento.....	11
5.1 Introdução	11
5.2 Medidas de Parecença	11
5.3 Métodos de Agrupamento.....	16
5.4 Avaliação do agrupamento	20
5.5 Resultados	21
6. Recursos computacionais.....	24
6.1 SAS.....	24
6.2 Minitab	25
6.3 SPSS.....	25
6.4 NTSYS - Numerical Taxonomy System of Multivariate Programs	25
7. Conclusões	26
Apêndice A - Tabelas.....	27
Apêndice B - Gráficos	31
Apêndice C - Dendogramas.....	35
Apêndice D - Programa.....	41

Resumo

Para este estudo foram coletados 25 espécimes de *Huberia* em várias regiões do Brasil e de cada espécime foram extraídos dois compostos químicos: os flavonóides foliares e as ceras epicuticulares.

O objetivo da pesquisa é verificar se esses compostos são bons marcadores taxonômicos, ou seja, produzem uma classificação auxiliar consistente, servindo para ratificar ou retificar a classificação morfológica existente.

Notamos através deste trabalho que:

- As ceras epicuticulares (através da distribuição de alcanos) discrimina os espécimes, servindo, portanto de marcador taxonômico. Os flavonóides, porém, mostraram baixo grau de discriminação dos espécimes, não servindo como marcadores taxonômicos.
- Os espécimes foram classificados, segundo análise de agrupamentos, em 4 grupos de forma consistente, ou seja, os espécimes de uma mesma espécie foram alocados no mesmo grupo.

1. Introdução

A taxonomia de vegetais é a área da Botânica responsável pela classificação das plantas baseada em suas características anatômicas, morfológicas e citogenéticas. A quimiotaxonomia, classificação com base em compostos químicos encontrados nas plantas, vem sendo utilizada para auxiliar a classificação das espécies, tornando-a mais estável e robusta, ou seja, a classificação não se altera radicalmente pela incorporação de novas informações. Os caracteres químicos servem para ratificar ou retificar a classificação tradicional, quando os caracteres morfológicos são inconsistentes ou confusos para estabelecer afinidades morfológicas. Eles porém, não são utilizados para construção de classificações sem base morfológica.

O gênero pesquisado é *Huberia* D.C. e os compostos selecionados para a análise química são as ceras epicuticulares e os flavonóides. Ele é caracterizado por possuir hábito arbóreo e arbustivo e compreende um gênero pequeno e pouco estudado com dez espécies descritas, várias delas localizadas nas manchas remanescentes da mata atlântica da costa leste brasileira.

As ceras epicuticulares funcionam como reguladoras híbridas, devido a sua característica de impermeabilidade. A sua composição pode ser complexa, apresentando alcanos, álcoois e ácidos entre outras substâncias. Os alcanos (parafinas) são amplamente utilizados neste tipo de estudo por serem de fácil obtenção e identificação.

Os flavonóides podem agir como filtros de luz ultravioleta e proteger a planta contra o ataque de microorganismos. São os compostos mais utilizados em classificação de vegetais, pois, possuem grande variedade de estruturas, podem ser separados e identificados rapidamente e são de ocorrência universal nos vegetais.

O objetivo deste estudo é verificar se estes compostos são bons marcadores taxonômicos para as espécies do gênero *Huberia*.

2. Descrição do Estudo

Foram coletadas 25 espécimes de *Huberia* D.C., sendo 22 pertencentes a cinco espécies já conhecidas e três pertencentes a três espécies novas, que ainda estão sendo descritas. As coletas foram realizadas na Bahia, Minas Gerais, Rio de Janeiro, São Paulo, Santa Catarina e Espírito Santo. Os espécimes estão distribuídos da seguinte forma:

- 6 espécimes da espécie *Huberia glazioviana* Cong. (HG1-HG6);
- 3 espécimes da espécie *Huberia minor* Cong. (HM1-HM3);
- 6 espécimes da espécie *Huberia nettoana* Brade. (HN1-HN6);
- 4 espécimes da espécie *Huberia ovalifolia* DC. (HO1-HO4);
- 3 espécimes da espécie *Huberia semiserrata* DC. (HS1-HS3);
- 1 espécime da espécie *Huberia aff ovalifolia* DC. (HA);
- 1 espécime da espécie *Huberia* sp J.B. 643 (Hsp1);
- 1 espécime da espécie *Huberia* sp J.B. 648 (Hsp2).

De cada espécime foram extraídas as ceras epicuticulares pela lavagem com clorofórmio. As parafinas foram separadas a partir da cera através de cromatografia em colunas e identificadas através de cromatografia gasosa, observando-se o comprimento de suas cadeias, que variaram de 25 a 33 átomos de carbonos, apresentando 17 tipos diferentes de parafinas. Mediu-se a concentração de cada tipo de parafina em cada espécime determinando a porcentagem de concentração das parafinas.

Para a análise dos flavonóides foram utilizados 18 espécimes. Os flavonóides foram extraídos das folhas e isolados através de técnicas cromatográficas e identificados através de espectrofotometria. Foram observados 22 tipos de flavonóides diferentes presentes nas espécimes pesquisadas.

3. Descrição das Variáveis

Ceras epicuticulares :

Tem-se 17 variáveis quantitativas, C25 a C33 (cadeias diretas) e C25R a C32R (cadeias ramificadas) representando as porcentagens de concentração dos 17 tipos de parafinas. O intervalo de variação destas variáveis é de 0 a 100% e os seus valores estão apresentados na Tabela A.1.

Flavonóides:

Tem-se 22 variáveis dicotômicas (F1 a F22) representando a presença ou ausência dos 22 tipos de flavonóides. Assume o valor 1 caso o flavonóide esteja presente no espécime e 0 caso contrário. Os valores destas variáveis estão apresentados na Tabela A.2.

4. Análise Exploratória dos Dados

Para a análise dos dados foram consideradas as 22 variáveis referentes aos flavonóides e apenas 3 variáveis referentes as ceras epicuticulares, pois as demais eram iguais a zero ou apresentaram valores muito próximos para todos os espécimes, não servindo para discriminá-los. As variáveis utilizadas foram C29, C31 e C33, referentes as cadeias de comprimento 29,31 e 33 respectivamente.

4.1 Ceras Epicuticulares

A análise das parafinas foi realizada para cada espécime conforme a sua espécie.

Através do Gráfico B.1, que nos mostra a porcentagem de concentração dos três tipos de parafinas considerados, verificamos que para a espécie *Huberia glazioviana* o espécime HG2 apresenta maior concentração de cadeias de comprimento 29 diferindo dos demais que apresentam maior concentração de cadeias de comprimento 31. Os espécimes HG4,HG5 e HG6 apresentam concentrações de cadeias de comprimento 33 próximas as de comprimento 29. HG1 e HG3 diferem destes espécimes apresentando concentração de cadeias de comprimento 29 maior

que as de comprimento 33. Esta espécie é a que apresenta maior variabilidade entre os espécimes.

Para as espécies *Huberia minor* e *Huberia nettoana* (Gráficos B.2 e B.3) verificamos que todos os espécimes apresentam o mesmo comportamento: maior concentração de cadeias de comprimento 29, seguida das cadeias de comprimento 31 e 33, assim como o espécime HG2 da espécie *Huberia glazioviana*.

Para a espécie *Huberia ovalifolia* (Gráfico B.4) todos os espécimes mostram maior concentração de cadeias de comprimento 31, seguida das cadeias de comprimento 33 e 29, comportamento semelhante a alguns espécimes de *Huberia glazioviana* (HG4, HG5 e HG6).

Para a espécie *Huberia semiserrata* (Gráfico B.5) verificamos que HS1 e HS2 apresentam o mesmo comportamento das espécies *Huberia minor* e *Huberia nettoana*. O espécime HS3 difere destes, apresentando comportamento semelhante a alguns espécimes de *Huberia glazioviana* (HG4, HG5 e HG6) porém com concentração menor de cadeias de comprimento 31.

Para as três novas espécies (Gráfico B.6) verificamos que a espécie *Huberia aff ovalifolia* apresenta o mesmo comportamento que a espécie *Huberia ovalifolia* (Gráfico B.4). O espécime Hsp1 (*Huberia sp 643*) difere de todos os espécimes observados, sendo o único a apresentar maior concentração de cadeias de comprimento 33. O espécime Hsp2 (*Huberia sp 648*) é semelhante aos espécimes das espécies *Huberia minor*, *Huberia nettoana* e *Huberia semiserrata* (HS1 e HS2).

Verificamos, através dos Gráficos B.1 a B.7, que as espécies *Huberia minor*, *Huberia nettoana*, *Huberia semiserrata* e *Huberia sp 648* são semelhantes em relação a composição da cera epicuticular. *Huberia ovalifolia*, *Huberia aff ovalifolia* e *Huberia glazioviana* também são semelhantes, sendo que a espécie *Huberia glazioviana* apresenta espécimes muito diferentes. A espécie *Huberia sp 643* é diferente de todas as outras espécies.

4.2 Flavonóides

A análise deste item deve ser cuidadosa devido ao pequeno número de observações por espécie.

Na Tabela A.3 observamos as porcentagens de presença dos 22 tipos de flavonóide em cada espécie. Verificamos que, em grande parte dos flavonóides, a porcentagem de presença é 0 ou 100%, o que indica que o flavonóide estudado está ausente ou presente em todos os espécimes da espécie. Verificamos também que não há um flavonóide que esteja presente em todas as espécies. Observamos que o flavonóide F2 está presente em 94% dos espécimes, sendo altamente frequente em todas as espécies. Os flavonóides F1, F3, F10, F14, F15, F19 e F22 estão presentes em apenas um espécime (6% do total), sendo praticamente inexistentes em todas as espécies. A espécie *Huberia minor* foi a que apresentou maior homogeneidade entre os espécimes estudados. Pouco se pode observar das espécies *Huberia aff ovalifolia*, *Huberia sp 643*, *Huberia sp 648* e *Huberia glazioviana* pois apresentam um número muito reduzido de observações (2 de *Huberia glazioviana* e 1 das demais).

5. Análise de Agrupamento

5.1 Introdução

A Análise de Agrupamento ou “Cluster Analysis” (Bussab *et al*, 1990 e Everitt, 1980) engloba uma variedade de técnicas e algoritmos cujo objetivo é, através de diversas variáveis medidas, juntar objetos em grupos que sejam homogêneos internamente e heterogêneos entre si. Este tipo de análise é amplamente utilizado na área de taxonomia (Rettig e Giannasi, 1990 e Lopes e Sheperd, 1991).

Neste trabalho abordaremos os critérios de “parecença” entre objetos e descreveremos os algoritmos mais conhecidos para produção de agrupamentos.

5.2 Medidas de Parecença

O termo parecença é utilizado aqui para representar dois conceitos: similaridade e dissimilaridade.

A dissimilaridade é medida através da distância entre os elementos, ou seja, quanto maior o seu valor menos parecidos são os elementos.

A similaridade é obtida através de coeficientes que medem a proximidade dos elementos, ou seja, quanto maior o seu valor mais parecidos serão os elementos.

A maioria dos algoritmos de Análise de Agrupamento trabalham com medidas de dissimilaridade (distância), sendo portanto, importante definir claramente a medida utilizada.

Para definir a medida de parença devemos considerar o tipo das variáveis do estudo, como veremos a seguir.

5.2.1 Variáveis quantitativas

Como há vários tipos de medidas de parença (Bussab *et al*, 1990) para variáveis quantitativas, enfatizaremos aquelas que são derivadas da distância euclidiana, por serem mais comumente usadas.

- Distância Euclidiana : é a distância geométrica entre dois pontos, sendo a medida mais conhecida para indicar a proximidade entre dois objetos e é dada por:

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

onde p é o número total de variáveis que estão sendo comparadas;

A, B são os objetos que estão sendo comparados;

$x_i(A)$, $x_i(B)$ são os valores assumidos pelos objetos A e B na i-ésima variável.

- Distância Euclideana Média : é a distância euclideana ponderada pelo número de variáveis envolvidas. Produz resultados análogos a distância euclideana, apresentando a vantagem de poder ser utilizada na ausência de dados para algumas variáveis. É dada por:

$$d(A,B) = \left[\frac{1}{p} \sum_{i=1}^p (x_i(A) - x_i(B))^2 \right]^{1/2}$$

- Distância Euclideana Padronizada : é a distância euclideana utilizando variáveis padronizadas (subtrai-se a média e divide-se pelo respectivo desvio padrão). Esta medida é importante quando a ordem de grandeza das variáveis é diferente.

- Distância de Mahalanobis : é a distância euclidiana ponderada pela variabilidade de cada uma das variáveis e as correlações entre elas. É de difícil interpretação.

5.2.2 Variáveis categorizadas dicotômicas

São muitos os tipos de medidas (coeficientes) existentes. Apresentaremos as mais utilizadas. Considerando dois espécimes, podemos construir a tabela abaixo:

		Espécime A		
		1	0	Total
Espécime B	1	a	b	a+b
	0	c	d	c+d
Total		a+c	b+d	p

onde a é o número de flavonóides que estão presentes nos dois espécimes;
 b é o número de flavonóides presentes no espécime B e ausentes em A;
 c é o número de flavonóides presentes no espécime A e ausentes em B;
 d é o número de flavonóides que estão ausentes nos dois espécimes e
 p é o número total de flavonóides.

A partir desta tabela podemos definir as seguintes medidas:

- Distância euclidiana média ou Distância binária de Sokal: indica a proporção de atributos não coincidentes nos dois espécimes, ou seja, é uma medida de dissimilaridade. Seu intervalo de variação é entre 0 e 1, sendo que o valor 0 indica que as espécimes são idênticas. É dada por:

$$d(A,B) = \left(\frac{b + c}{a + b + c + d} \right)^{1/2}$$

- Coeficiente de concordâncias simples: é uma medida de similaridade que indica a proporção de coincidências. Seu intervalo de variação é entre 0 e 1 e é dada por:

$$s(A,B) = \frac{a + d}{a + b + c + d}$$

- Coeficiente de concordâncias positivas ou coeficiente de Russel e Rao: também é uma medida de similaridade que varia de 0 a 1. Indica a proporção de atributos presentes nos dois espécimes.

$$s(A,B) = \frac{a}{a + b + c + d}$$

- Coeficiente de Jaccard: é um coeficiente muito utilizado que indica a proporção de coincidências nos dois espécimes, desconsiderando os atributos que estão ausentes nos dois espécimes, ou seja, são considerados somente os atributos que aparecem em ao menos um dos espécimes

$$s(A,B) = \frac{a}{a + b + c}$$

5.2.3 Variáveis categorizadas com vários níveis

Quando as variáveis categorizadas apresentam mais de dois níveis, elas podem ser transformadas em variáveis binárias (dicotômicas) fictícias e as mesmas medidas acima podem ser empregadas. Para isso, transformamos cada nível da variável em uma nova variável binária, que indicará se o atributo está presente neste nível. Por exemplo, se uma variável possui três níveis, esta será transformada em três variáveis binárias, sendo que 1 na primeira variável binária indica que o atributo está no primeiro nível da variável original e assim por diante.

5.2.4 Variáveis mistas (quantitativas e categorizadas)

A seguir são dadas algumas sugestões (Bussab *et al*, 1990) para analisar estas variáveis conjuntamente:

A) Transformação em variáveis binárias

Uma maneira de se resolver este problema é transformar todas as variáveis quantitativas em variáveis binárias e utilizar os coeficientes descritos para este tipo de variável. O grande inconveniente deste método é a perda de informações decorrentes da transformação.

B) Transformação das variáveis quantitativas de modo que o intervalo de variação fique entre 0 e 1.

Devemos transformar as variáveis quantitativas, por exemplo, dividindo as variáveis referentes as porcentagens por 100 e assim o intervalo de variação será de 0 e 1. Em seguida calcular a distância euclidiana considerando todas as variáveis.

C) Coeficiente combinado de semelhança.

Constrói-se coeficientes para as variáveis quantitativas e para as variáveis categorizadas e a partir destes constrói-se um único coeficiente ponderado como é dado a seguir:

$$C(A,B) = w_1 C_{cat}(A,B) + w_2 C_{quanti}(A,B)$$

onde: w_1 é o peso das variáveis categorizadas e

w_2 é o peso das variáveis quantitativas.

Quando utilizamos este coeficiente devemos verificar que:

- Os coeficientes de parecença devem ter o mesmo sentido (similaridade ou dissimilaridade).

Caso os coeficientes não tenham o mesmo sentido, devemos transformá-los convenientemente. Como a maioria dos algoritmos de agrupamentos utilizam as medidas de dissimilaridade, é conveniente, então, transformar todas as medidas em medidas de dissimilaridade.

- Os coeficientes devem ter os mesmos intervalos de variação.

Se isto não ocorrer podemos padronizar os coeficientes (subtrair a média e dividir pelo desvio padrão).

- A escolha dos pesos deve ser adequada.

Em geral utiliza-se o número total de variáveis quantitativas e categorizadas para se definir o peso de cada coeficiente. O peso das variáveis quantitativas seria o número de variáveis quantitativas sobre o total de variáveis envolvidas. Analogamente temos o peso para as variáveis categorizadas.

D) Coeficiente de Gower

Gower (1971) propõe um coeficiente de parecença que é uma forma mais elaborada do coeficiente combinado, pois permite trabalhar com observações perdidas para algumas variáveis.

Inicialmente, define-se um coeficiente de parecença para cada variável (S_i), variando entre 0 e 1. Utiliza-se uma variável indicadora I_i , que assume valor 1 se os dois espécimes podem ser comparados em relação a variável i (se ambos tiveram esta variável medida) e 0 caso contrário. Assim o coeficiente será calculado considerando-se somente as variáveis medidas para os dois espécimes. A similaridade entre eles é dada por:

$$S(A,B) = \frac{\sum_{i=1}^p I_i(A,B) \cdot S_i(A,B)}{\sum_{i=1}^p I_i(A,B)}$$

Observação: Este coeficiente será indefinido quando os dois espécimes não puderem ser comparados segundo nenhuma variável.

5.3 Métodos de Agrupamento

A idéia principal que envolve a análise de agrupamento é a coesão interna e o isolamento externo entre os grupos. Como já foi dito, a maioria dos métodos utiliza medidas de dissimilaridade.

As técnicas de agrupamento podem ser classificadas em diversos tipos. Citaremos aqui as técnicas hierárquicas e as técnicas de partição.

5.3.1 Técnicas hierárquicas

Técnicas hierárquicas são aquelas nas quais os objetos são classificados em grupos em diferentes etapas, produzindo uma árvore de classificação. Os quatro métodos mais utilizados são:

5.3.1.1 Método da Centróide

A distância entre os grupos é definida pela distância entre os centros dos grupos (Figura 1). Devemos então fundir os grupos com a menor distância.

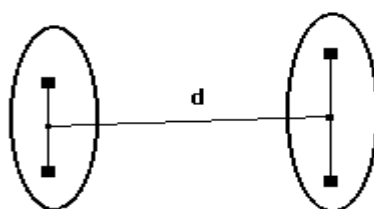


Figura 1 - Método da Centróide

5.3.1.2 Método das Médias das Distâncias

Neste método a distância entre os grupos é definida pela média das distâncias de cada elemento de um grupo e cada elemento do outro grupo (Figura 2). Este método também é conhecido como UPGMA (Unweighted Pair-Group Method using arithmetic Average). Deve-se fundir os grupos com a menor distância média.

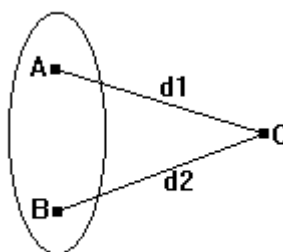


Figura 2 -Método das Médias das Distâncias

5.3.1.3 Método da Ligação Simples ou do Vizinho mais Próximo

A distância entre os grupos é definida pela distância entre os elementos mais próximos (mais parecidos) de cada grupo (Figura 3). Funde-se os grupos com a menor distância. Uma crítica que se faz a este método é que ele tende a agrupar objetos distintos, não diferenciando grupos próximos porém diferentes.

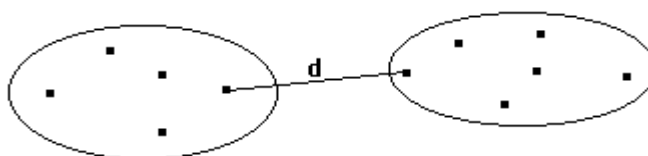


Figura 3 - Método da Ligação Simples

5.3.1.4 Método da Ligação Completa ou do Vizinho mais Longe

A distância entre os grupos é definida pela distância entre os elementos mais distantes (menos parecidos) de cada grupo (Figura 4).

Neste método também devemos fundir os grupos com a menor distância.

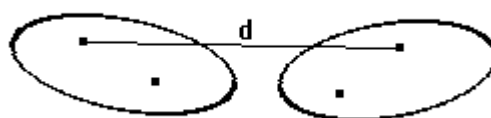


Figura 4 - Método da Ligação Completa

5.3.2 Técnicas de partição

As técnicas de partição procuram separar diretamente os espécimes em grupos. Estes métodos pressupõe o conhecimento do número de grupos que se deseja obter e permite trabalhar com um número grande de objetos. O método mais utilizado para esta técnica é o método das k-médias.

Neste método também devemos ter homogeneidade dentro dos grupos e heterogeneidade entre os grupos. O critério mais utilizado para se verificar isto é o da soma de quadrados residual, baseado na Análise de Variância.

Faz-se uma partição dos espécimes em k grupos. Calcula-se a soma de quadrados residual dentro de cada grupo (utilizando o quadrado da distância euclideana de cada elemento ao centro do grupo) e em seguida obtém-se a soma de quadrados total somando-se todas as somas de quadrados de cada grupo. Quanto menor for este valor mais homogêneos são os grupos e melhor será a partição.

As partições são realizadas de modo iterativo, ou seja, são feitas várias partições até encontrarmos a “melhor”. O processo ocorre da seguinte forma:

- Escolhemos os k centros dos grupos para começarmos o processo. Os centros podem ser escolhidos segundo vários critérios, como por exemplo, os k primeiros elementos, k elementos sorteados aleatoriamente, os k mais distantes, etc...

- Após alocarmos estes k elementos, devemos designar os demais aos grupos. Cada elemento será colocado no grupo mais próximo, definido pela proximidade ao centro deste grupo.

- Estando os elementos separados nos k grupos, devemos calcular a soma de quadrados residual.

- Como a partição é arbitrária devemos encontrar uma que diminua esta soma. Para isso, realocamos o primeiro objeto no grupo que lhe for mais próximo e verificamos se há diminuição significativa na soma de quadrados residual. Caso haja diminuição, o elemento passa ao correspondente grupo. Repete-se este passo para cada elemento.

O procedimento é repetido até encontrarmos a “melhor” partição ou até um número definido de iterações.

Neste método podemos realizar uma análise de variância (ANOVA) para identificarmos quais são as variáveis mais importantes na análise, ou seja, aquelas que melhor discriminam os espécimes. Para isso, calculamos as somas de quadrados dentro dos grupos e entre os grupos, considerando-se cada variável.

Existe um teste de hipóteses que nos permite escolher o número de grupos da partição. A ideia deste teste é verificar o ganho na soma de quadrados residual ao se passar de uma partição com k grupos para uma com $k+1$ grupos. Pode-se, portanto obter partições para várias quantidades de grupos e definir qual a melhor. Decide-se por $k+1$ grupos quando a variação na soma de quadrados residual ao se passar de k para $k+1$ grupos for significativamente grande. A estatística do teste é dado por:

$$R(k+1/k) = [SQRes(k)/SQRes(k+1) - 1](n - k - 1)$$

onde $SQRes(k)$ é a soma de quadrados residual do k -ésimo grupo e n é o número total de elementos.

Supondo que as variáveis (quantitativas) têm distribuição normal, a estatística $R(k+1,k)$ tem distribuição F com p e $n-k-1$ graus de liberdade, onde p é o número de variáveis envolvidas. Desta forma, valores de R maiores que um determinado valor crítico indicam ganho na passagem de k para $k+1$ grupos.

O método das k -médias fornece indicações mais precisas quando ao número de grupos a ser formado e permite que, através de análises de variância, identifiquemos quais variáveis contribuem mais para a discriminação dos elementos e as que não contribuem.

5.4 Avaliação do agrupamento

Para avaliação do agrupamento podemos utilizar o coeficiente de correlação cofenética que é dado pela correlação entre a matriz de similaridade original e a produzida pelo agrupamento. Altos valores deste coeficiente indicam que o agrupamento foi satisfatório, mantendo as diferenças originais entre os elementos.

Existem outros índices e teste utilizados para avaliação do agrupamento. Para maiores detalhes ver Bussab *et al* (1990).

5.5 Resultados

5.5.1 Métodos hierárquicos

Primeiramente realizamos a análise para as variáveis quantitativas (C29, C31 e C33) e categorizadas (F1 a F22) separadamente.

Para as variáveis quantitativas (distribuição de alcanos) utilizou-se a distância euclidiana como medida de “parecença” e todos os métodos de agrupamentos (Gráfico C.1). Notamos, inicialmente, a existência de quatro possíveis grupos, que são:

Grupo 1: Todos os espécimes das espécies *Huberia minor* (HM) e *Huberia nettoana* (HN), HS1 e HS2 da espécie *Huberia semiserrata* (HS), HG2 da espécie *Huberia glazioviana* (HG) e a espécie *Huberia sp 648* (Hsp2).

Grupo 2: A espécie *Huberia sp 643* (Hsp1).

Grupo 3: Todos os espécimes das espécies *Huberia ovalifolia* (HO), um espécime de *Huberia glazioviana* (HG4) e a espécie *Huberia aff ovalifolia* (HA).

Grupo 4: Os espécimes de *Huberia glazioviana* (exceto HG2 e HG4) e o espécime HS3 da espécie *Huberia semiserrata* (HS).

De maneira geral, observamos a presença do grupo 1 bem definido em todos os métodos e que os espécimes de *Huberia ovalifolia* e *Huberia aff ovalifolia* estão sempre presentes no mesmo grupo.

Para as variáveis categorizadas (Flavonóides) utilizou-se o coeficiente de Jaccard e todos os métodos descritos (Gráfico C.2). Observamos que os grupos não estão bem definidos e a classificação é confusa. Isso ocorre, possivelmente, porque os flavonóides não discriminam bem os espécimes.

Consideremos agora as variáveis mistas (quantitativas e categorizadas). Para esta análise empregamos duas propostas:

- Transformamos as variáveis quantitativas de modo que o intervalo de variação ficou entre 0 e 1.

A transformação realizada foi:

$$Z_i = \frac{X_i - X_{\min}}{X_{\max} - X_{\min}}$$

onde x_i é a i -ésima observação da variável;

$x_{\min}$ é o menor valor observado;

$x_{\max}$ é o maior valor observado.

Como as variáveis categorizadas já estão no intervalo [0,1], utilizamos a distância euclideana como medida de “parecença”. Atráves dos dendrogramas (Gráfico C.3) verificamos que os resultados obtidos são parecidos com aqueles obtidos na análise com as variáveis quantitativas, porém não tão claros quanto os primeiros. Este fato pode estar relacionado à transformação, que altera o grau de discriminação das variáveis quantitativas e, deste modo, a medida de “parecença” sofre maior influência das variáveis categorizadas.

- Contruimos um coeficiente combinado de semelhança. Para as variáveis quantitativas utilizamos a distância euclideana (dissimilaridade) e para as categorizadas utilizamos o coeficiente de Jaccard (similaridade). Como as medidas não tem o mesmo sentido, transformamos o coeficiente de Jaccard em um medida de dissimilaridade (troca de sinal). Em seguida rescalonamos as medidas para o intervalo [0,1] para que tivessem o mesmo intervalo de variação. Para a determinação do peso utilizamos duas propostas:

- ponderar os coeficientes pelo número de variáveis envolvidas (3 quantitativas e 22 categorizadas) e
- ponderar os coeficientes dando a mesma importância para os dois conjuntos de variáveis.

Os coeficientes para as propostas acima são:

$$C(A,B) = \frac{22}{25} \cdot C_{\text{cat}}(A,B) + \frac{3}{25} \cdot C_{\text{quanti}}(A,B) \text{ e}$$

$$C(A,B) = \frac{1}{2} \cdot C_{\text{cat}}(A,B) + \frac{1}{2} \cdot C_{\text{quanti}}(A,B),$$

respectivamente.

Atráves dos dendrogramas (Gráfico C.4 e C.5) notamos que:

Os espécimes de *Huberia ovalifolia* e *Huberia aff ovalifolia* estão sempre no mesmo grupo, assim como os espécimes de *Huberia nettoana*, *Huberia minor*, *Huberia semiserrata* e *Huberia sp* (Hsp2) formam outro grupo.

Os espécimes de *Huberia glazioviana* estão sempre juntos, porém algumas vezes estão isolados, outras vezes formando grupo com os espécimes de *Huberia ovalifolia* e *Huberia aff ovalifolia*.

Os espécimes *Huberia sp* (Hsp1) e HS3 (*Huberia semiserrata*) quase sempre estão isolados ou não se agrupam de maneira consistente, ou seja, ora entram em um grupo, ora entram em outro.

5.5.2 Método das k-médias

Neste método só podemos utilizar as variáveis quantitativas.

Nos métodos hierárquicos, notamos que os espécimes podem ser agrupados em até 5 grupos. Aplicamos o método das k-médias separando os espécimes em 2, 3, 4 e 5 grupos e calculamos a estatística R para detectarmos o melhor número de grupos para a partição. Através da Tabela 5.1 verificamos que o ganho é significativo quando aumentamos o número de grupos, exceto quando passamos de 4 para 5 grupos, o que nos indica que a melhor partição é a de 4 grupos.

Tabela 5.1 - Estatística R

R	Distribuição	Nível descritivo
R(3/2)	F(3,15)	0.0004
R(4/3)	F(3,14)	0.0044
R(5/4)	F(3,13)	0.0723

Através da análise de variância (Tabela 5.2) considerando-se 4 grupos, verificamos que as três variáveis, C29, C31 e C33, discriminam muito bem os espécimes, ou seja, estas variáveis servem para classificá-los.

Tabela 5.2 - ANOVA para 4 grupos

Fonte de Variação	Nível descritivo
C29	0.0000
C31	0.0000
C33	0.0000

Os grupos formados são apresentados na tabela a seguir.

Tabela 5.3 - Grupos formados pelo método das k-médias

Grupo 1	Grupo 2	Grupo 3	Grupo 4
HG3	HA	HM1	Hsp1
HG6	HO1	HM2	
	HO2	HM3	
	HO3	HN1	
	HO4	HN2	
	HS3	HN5	
		HS1	
		HS2	
		Hsp2	

Como podemos observar os grupos formados são muito parecidos com os grupos que aparecem na análise com métodos hierárquicos, mostrando que a distribuição de alcanos classifica bem os espécimes.

6. Recursos computacionais

Existem vários programas disponíveis para Análise de Agrupamentos, dos quais discutiremos o SAS, Minitab, SPSS e o NTSYS.

6.1 SAS

Possui 2 procedimentos para aplicação de análise de agrupamentos:

- CLUSTER : possui 11 métodos hierárquicos e utiliza, de modo geral, a distância euclideana como medida de parecença.

- FASTCLUS: realiza o método das k-médias utilizando a distância euclideana.

Os procedimentos assumem que todas as variáveis são quantitativas, ou seja, não há medidas nem tratamento especial para as variáveis categorizadas. Para a construção do dendrograma utiliza-se o procedimento TREE.

6.2 Minitab

Este pacote é de simples manipulação. Apresenta 3 opções para análise: CLUSTER-OBSERVAÇÃO, CLUSTER- VARIÁVEIS e CLUSTER- KMÉDIA.

Dispõe de 6 métodos hierárquicos e 6 medidas de “parecença” e assim como o SAS, assume que todas as variáveis são quantitativas.

No método das k-médias, não há opção para obtermos a ANOVA.

6.3 SPSS

O SPSS trabalha com variáveis quantitativas e categorizadas e possui dois procedimentos para Análise de Agrupamentos:

- CLUSTER (ou Hierarchical Cluster): dispõe de 7 métodos hierárquicos, 8 medidas de “parecença” para variáveis quantitativas , 7 distâncias e 18 coeficientes de similaridade para variáveis categorizadas.

- QUICK CLUSTER (ou kmeans Cluster): realiza o método das k-médias.

Este pacote, assim como os demais, não permite que se obtenha o coeficiente combinado de semelhança. A maneira encontrada para calculá-lo foi a seguinte:

- Obter as matrizes de coeficientes para as variáveis quantitativas e para as categorizadas e armazenar cada uma em um arquivo de saída.

- Utilizar uma planilha eletrônica (Excel, por exemplo) para fazer a operação com as matrizes. Após a operação, armazenar a matriz de coeficientes combinados em um arquivo.

- Utilizar a matriz de coeficientes (o arquivo deve estar no formato SPSS) como entrada do procedimento CLUSTER.

O programa elaborado para a análise encontra-se no Apêndice D.

6.4 NTSYS - Numerical Taxonomy System of Multivariate Programs

Este pacote foi desenvolvido especificamente para pesquisa em taxonomia numérica. O programa dispõe de 8 métodos hierárquicos, 16 coeficientes para variáveis categorizadas e 9 medidas para variáveis quantitativas e não possui o método das k-médias.

Assim como no SPSS, podemos calcular o coeficiente combinado de semelhança de modo indireto: calculamos os coeficientes para cada tipo de variável, armazenamos as matrizes em arquivos de saída, manipulamos estas matrizes em uma planilha eletrônica e entramos nos procedimentos de agrupamento com a matriz de coeficientes combinados.

7. Conclusões

Verificamos que a distribuição de alcanos discrimina os espécimes, servindo, portanto de marcador taxonômico. Os flavonóides, porém, mostraram baixo grau de discriminação dos espécimes, não servindo como marcadores taxonômicos.

Observamos 4 grupos na Análise de Agrupamentos:

- Os espécimes de *Huberia ovalifolia* e *Huberia aff ovalifolia*.
- Os espécimes de *Huberia nettoana*, *Huberia minor*, *Huberia semiserrata* e *Huberia sp* J.B. 648.
- Os espécimes de *Huberia glazioviana*.
- O espécime de *Huberia sp* J.B. 643.

Um espécime de *Huberia semiserrata* apresenta uma classificação indefinida, sendo mais semelhante aos espécimes de *Huberia ovalifolia* e *Huberia glazioviana*.

Apêndice A

Tabelas

Apêndice B

Gráficos

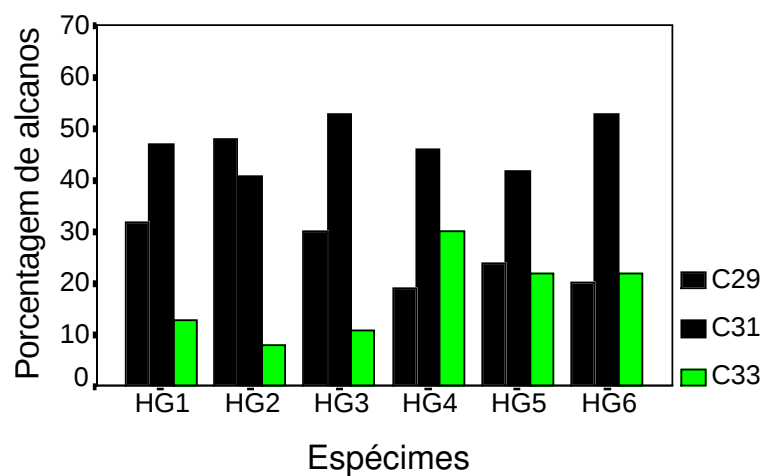
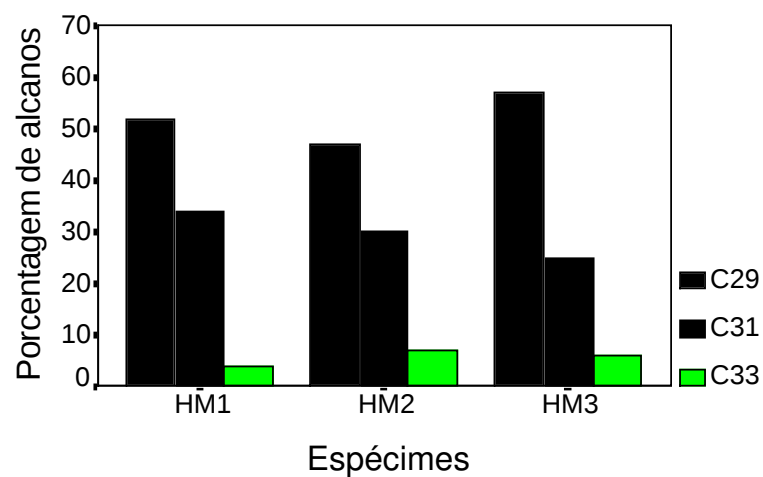
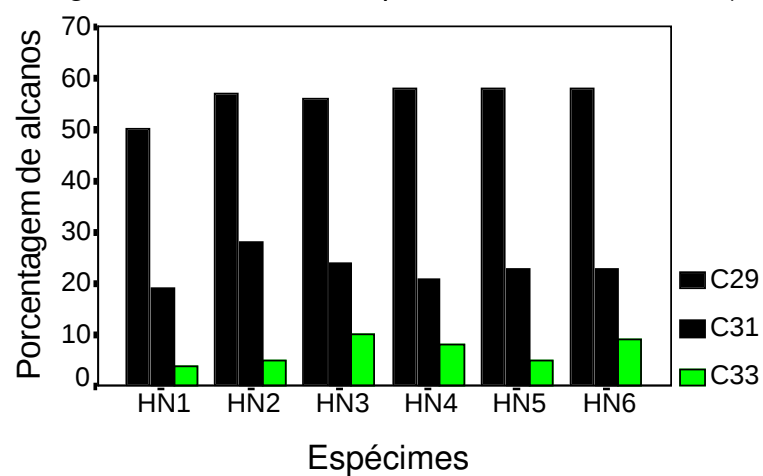
Gráfico B.1 - Porcentagem de alcanos na espécie *Huberia glazioviana* (HG)Gráfico B.2 - Porcentagem de alcanos na espécie *Huberia minor* (HM)Gráfico B.3 - Porcentagem de alcanos na espécie *Huberia nettoana* (HN)

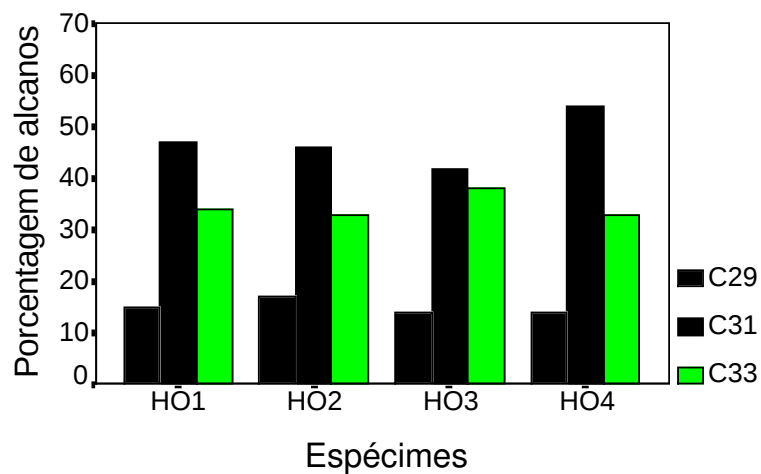
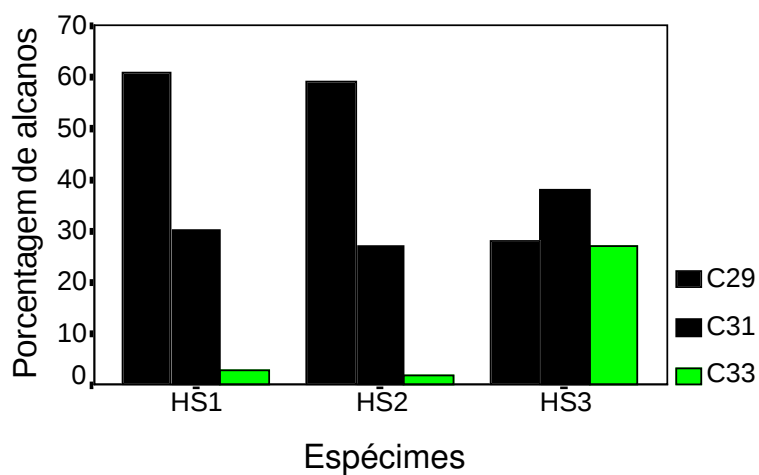
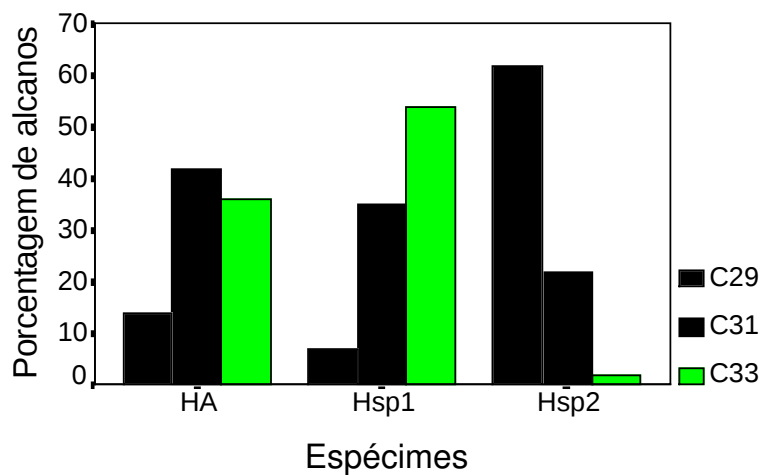
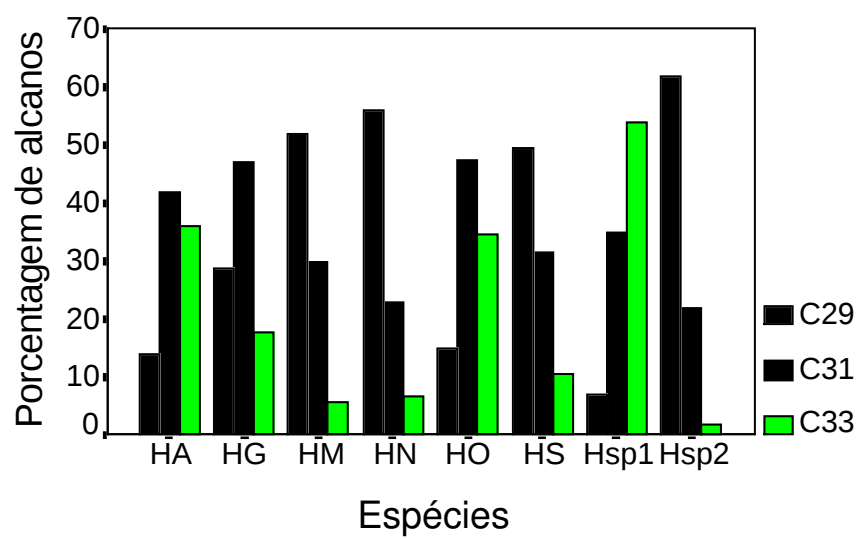
Gráfico B.4 - Porcentagem de alcanos na espécie *Huberia ovalifolia* (HO)Gráfico B.5 - Porcentagem de alcanos na espécie *Huberia semiserrata* (HS)Gráfico B.6 - Porcentagem de alcanos nas espécies *Huberia aff ovalifolia* (HA), *Huberia sp* 643 (Hsp1) e *Huberia sp* 648 (Hsp2).

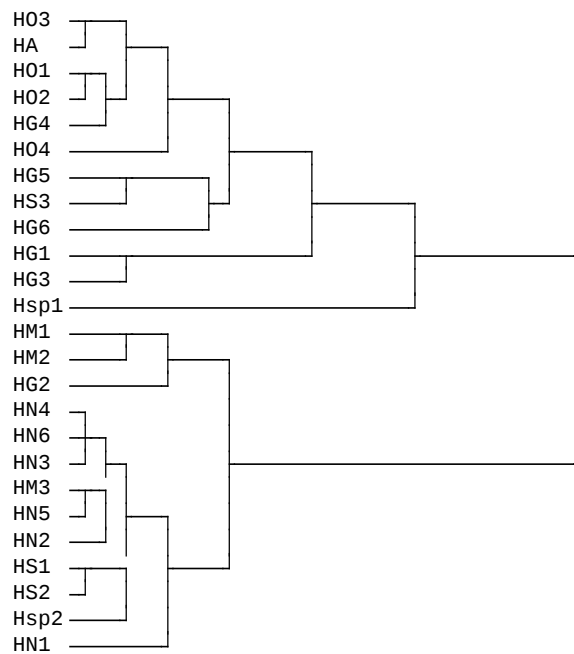
Gráfico B.7 - Porcentagem média de alcanos por espécie



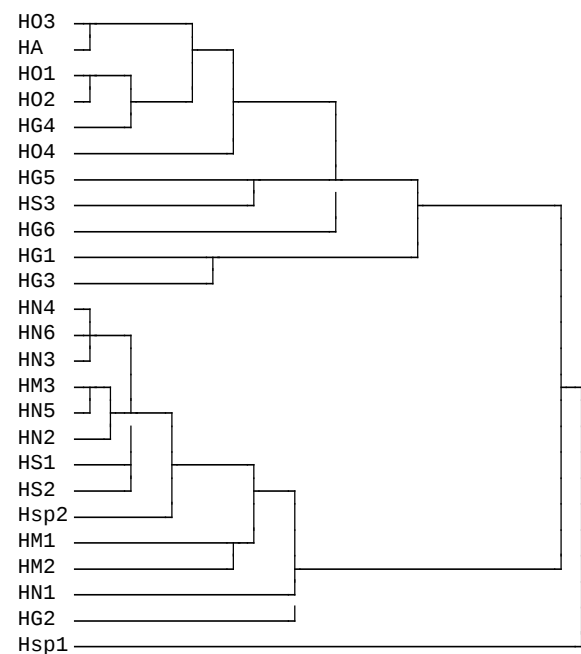
Apêndice C

Dendogramas

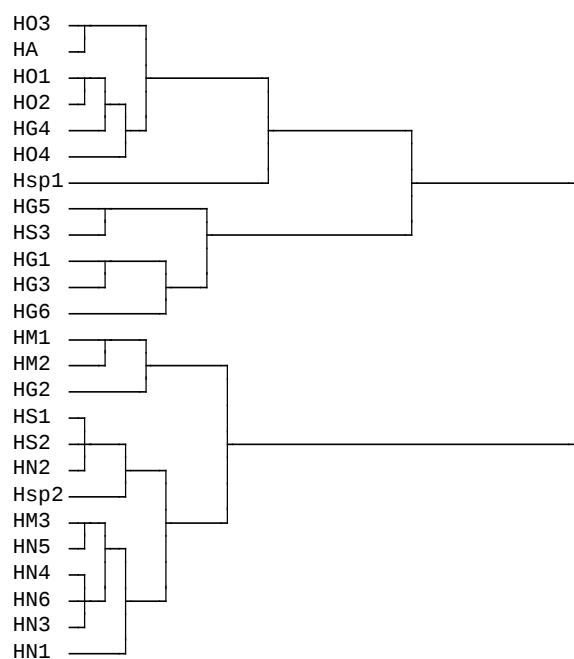
Gráfico C.1 - Dendograma para variáveis quantitativas.



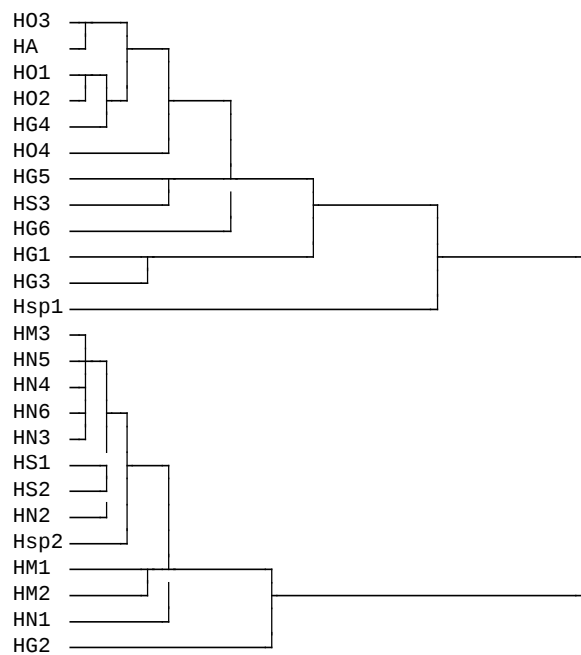
Média das Distâncias



Ligação Simples

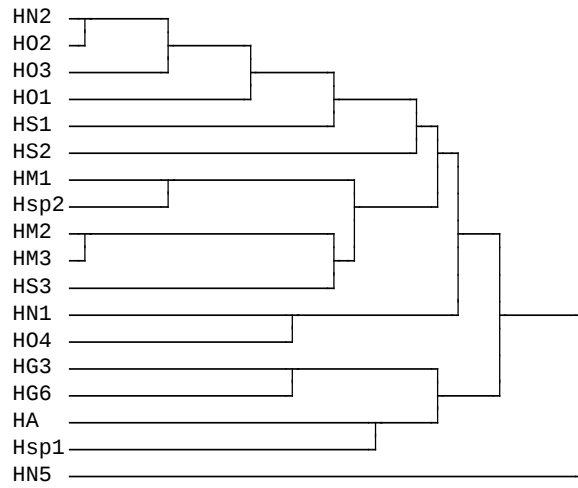


Ligação Completa

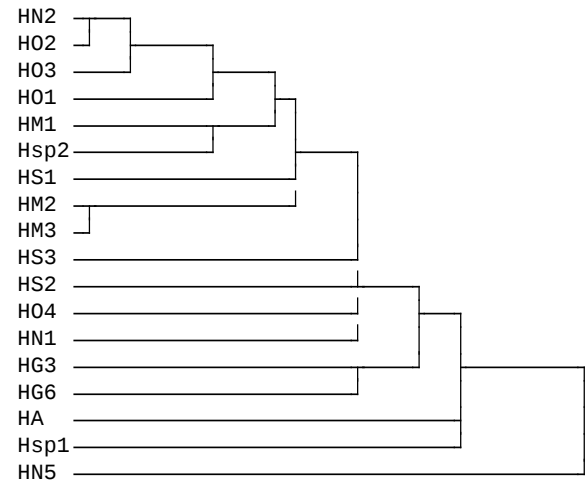


Centróide

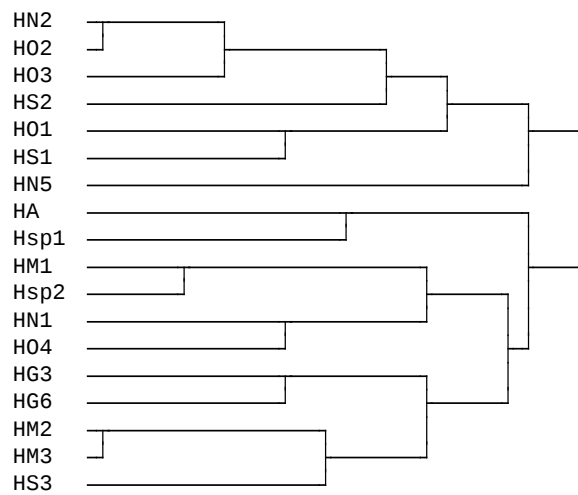
Gráfico C.2 - Dendrograma para variáveis Categorizadas.



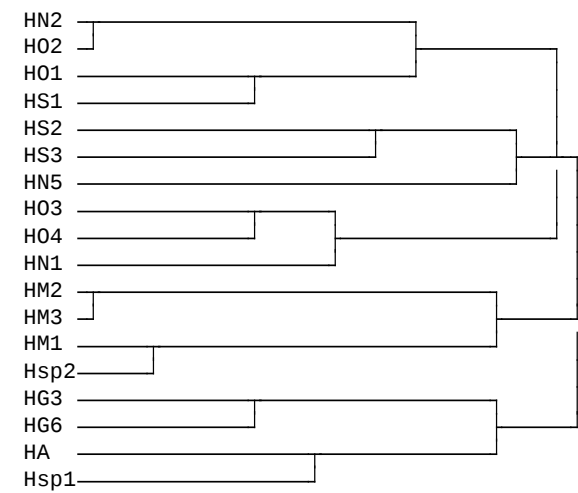
Média das Distâncias



Ligação Simples

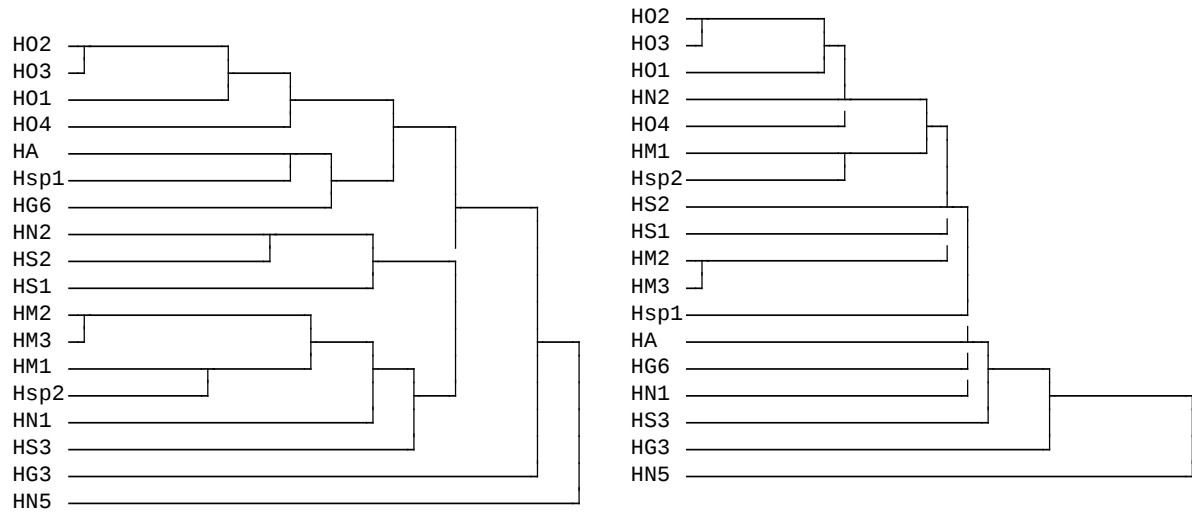


Ligação Completa



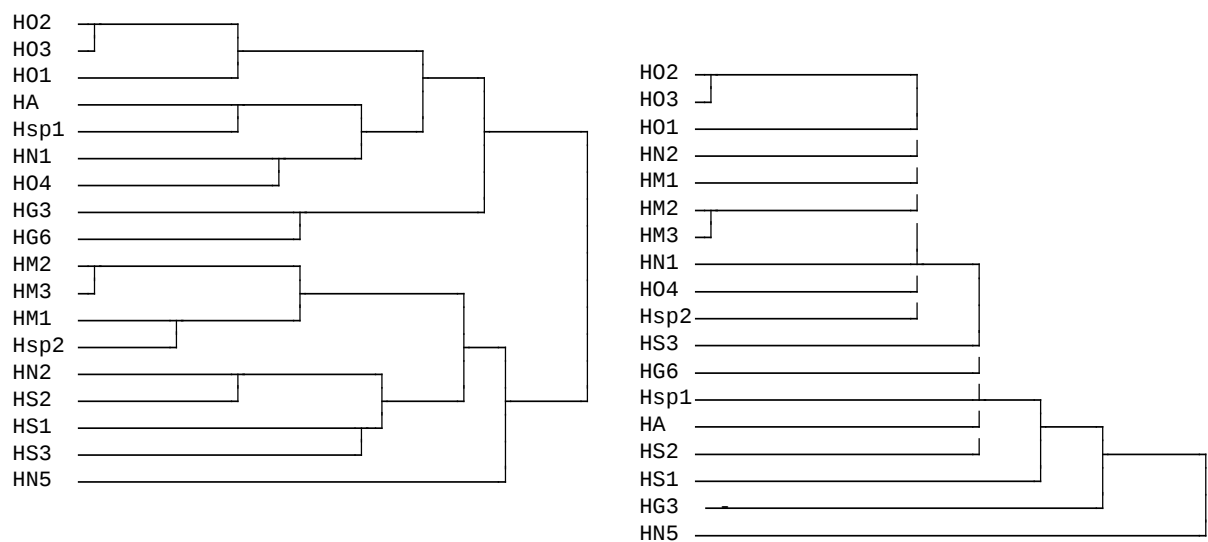
Centróide

Gráfico C.3 - Dendrograma para variáveis mistas utilizando transformação.



Média das Distâncias

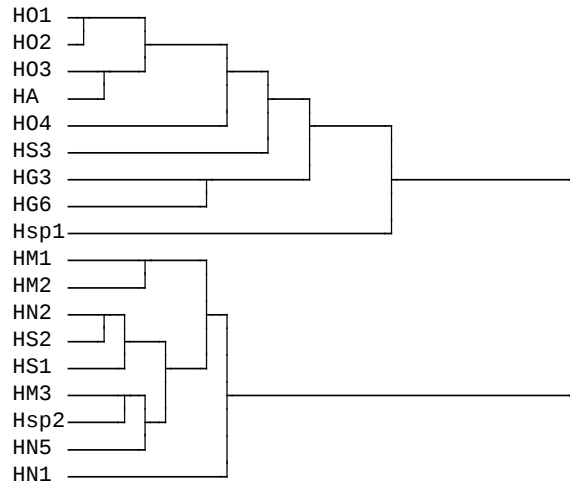
Ligação Simples



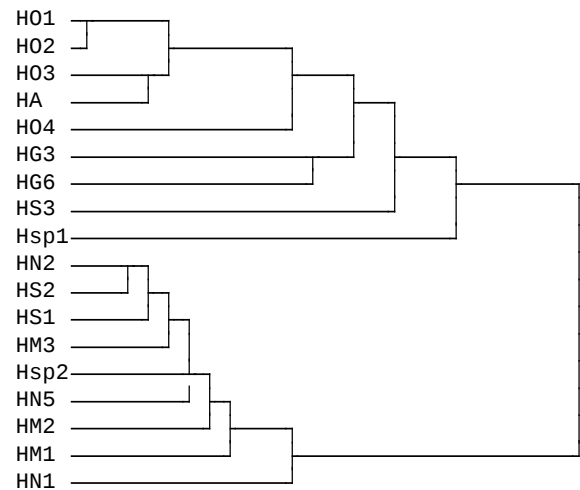
Ligação Completa

Centróide

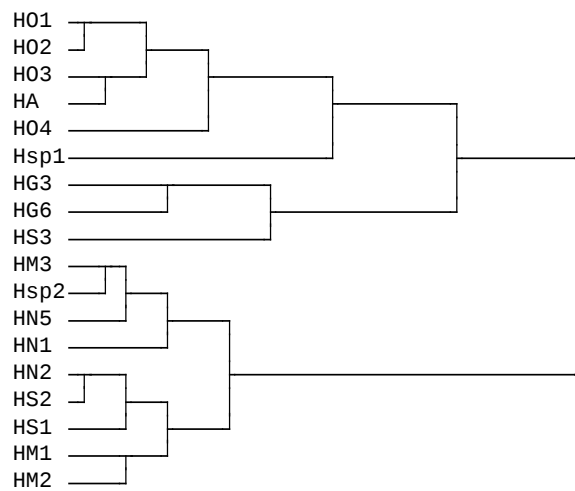
Gráfico C.4 - Dendrograma para variáveis mistas utilizando coeficiente combinado
($w_1=3/25$ e $w_2=22/25$)



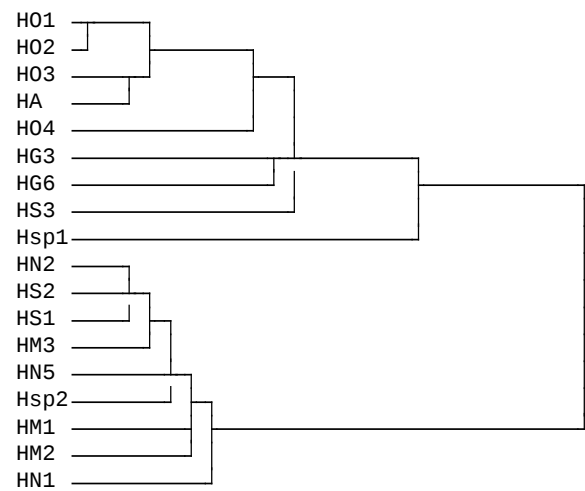
Média das Distâncias



Ligação Simples

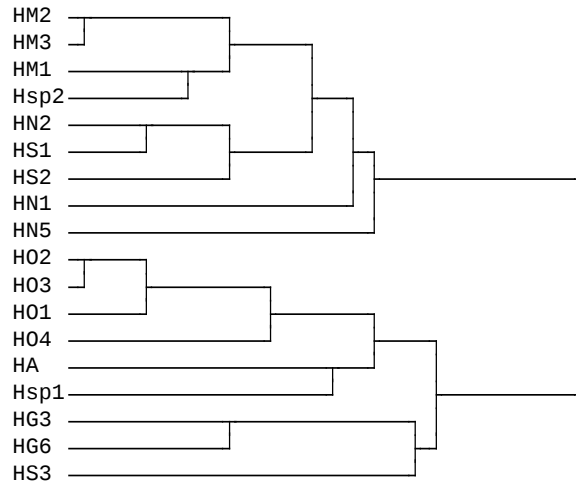


Ligação Completa

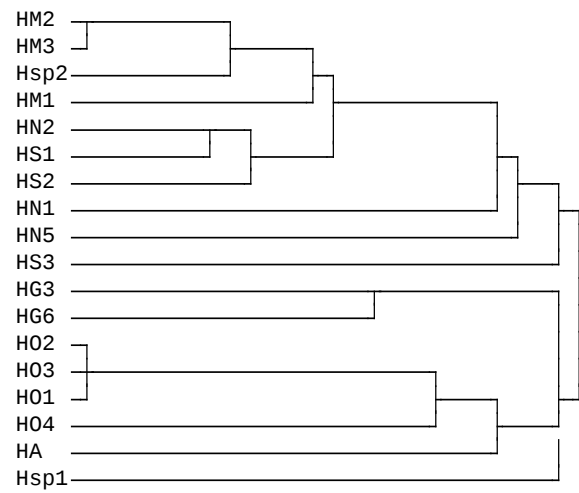


Centróide

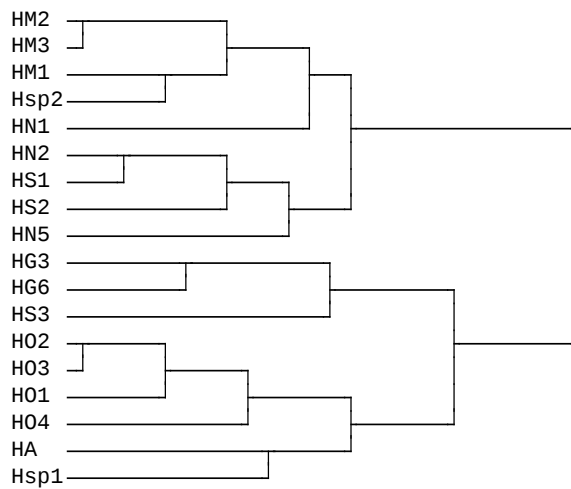
Gráfico C.5 - Dendrograma para variáveis mistas utilizando coeficiente combinado
($w_1=w_2=1/2$)



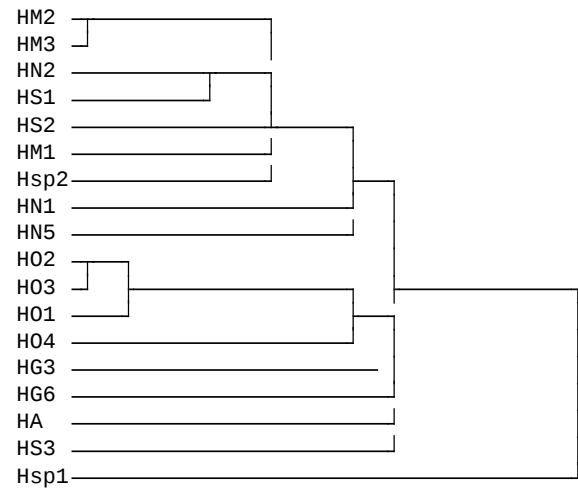
Média das Distâncias



Ligação Simples



Ligação Completa



Centróide

Apêndice D

Programa

O programa a seguir foi elaborado no pacote SPSS para Análise de Agrupamentos utilizando o coeficiente combinado. Primeiramente, calculamos as medidas de “parecença” através do procedimento PROXIMITIES e armazenamos as matrizes em arquivos de saída (formato SPSS). Com as matrizes, calculamos o coeficiente combinado e armazenamos em um arquivo (deve estar no formato SPSS) que foi a entrada do procedimento CLUSTER. O programa é:

PROXIMITIES

```
f1 f3 f4 f5 f6 f7 f8 f9 f10 f11 f12 f13 f14 f15 f16 f17 f18 f19 f2 f20 f21 f22
/MATRIX OUT ('C:\JACCARD.DAT')
/ID=especime
/VIEW=CASE
/MEASURE= REVERSE RESCALE JACCARD (1,0) . {* medida de dissimilaridade*}
```

PROXIMITIES

```
c29 c31 c33
/MATRIX OUT ('C:\EUCLID.DAT')
/ID=especime
/VIEW=CASE
/MEASURE= RESCALE EUCLID {*medida de dissimilaridade no intervalo [0,1]*}
/STANDARDIZE= VARIABLE Z .
```

CLUSTER

```
/MATRIX IN ('C:\COMBINADO.DAT')
/METHOD BAVERAGE SINGLE COMPLETE CENTROID
/PRINT NONE
/PLOT DENDROGRAM .
```

Observação: Os procedimentos PROXIMITIES e CLUSTER foram executados separadamente.

Tabela A.1 - Porcentagem de alcanos nas ceras epicuticulares de espécimes de *Huberia* D.C.

Espécime	Local	Comprimento da cadeia																
		C25	C25R	C26	C26R	C27	C27R	C28	C28R	C29	C29R	C30	C30R	C31	C31R	C32	C32R	C33
HG1	RJ	0	0	0	0	3	0	3	0	32	0	2	0	47	0	0	0	13
HG2	RJ	0	0	0	0	0	0	0	0	48	0	2	0	41	0	1	0	8
HG3	RJ	0	0	0	0	1	0	1	0	40	0	2	0	53	0	2	0	11
HG4	RJ	0	0	0	0	1	0	1	0	19	0	1	0	46	0	2	0	30
HG5	RJ	0	0	0	0	4	1	1	0	24	1	1	0	42	0	2	0	22
HG6	MG	0	0	0	0	2	0	0	0	20	0	1	0	53	0	2	0	22
HM1	RJ	0	0	0	0	2	0	3	0	52	3	2	0	34	0	0	0	4
HM2	RJ	0	0	0	0	3	0	8	0	47	0	5	0	30	0	0	0	7
HM3	RJ	0	0	0	0	8	0	3	0	57	0	1	0	25	0	0	0	6
HN1	MG	0	0	0	0	10	2	4	0	50	4	2	0	19	0	0	0	4
HN2	MG	0	0	0	0	6	0	2	0	57	0	2	0	28	0	0	0	5
HN3	MG	0	0	0	0	7	0	2	0	56	0	0	5	24	0	0	0	10
HN4	RJ	0	0	0	0	9	0	2	0	58	0	1	0	21	0	1	0	8
HN5	RJ	0	0	0	0	8	0	3	0	58	0	2	0	23	0	1	0	5
HN6	RJ	0	0	0	0	6	0	2	0	58	0	1	0	23	0	1	0	9
HO1	SP	0	0	0	0	0	0	0	0	15	0	2	0	47	0	2	0	34
HO2	ES	0	0	0	0	0	0	0	0	17	1	0	0	46	0	3	0	33
HO3	BA	0	0	0	0	0	0	1	0	14	0	2	0	42	0	3	0	38
HO4	RJ	0	0	0	0	0	0	3	0	14	0	1	0	54	0	4	0	11
HS1	SP	0	0	0	0	3	0	1	0	61	0	2	0	30	0	0	0	3
HS2	SP	0	0	0	0	3	0	2	0	59	4	3	0	27	0	0	0	2
HS3	SC	0	0	0	0	2	0	1	0	28	0	1	0	38	0	3	0	27
HA	RJ	0	0	0	0	1	0	3	0	14	0	1	0	42	0	3	0	36
Hsp1	BA	0	0	0	0	0	0	0	0	7	0	0	0	35	0	0	0	54
Hsp2	ES	0	0	0	0	8	0	3	0	62	0	2	1	22	0	0	0	2

onde R indica cadeia ramificada.

Tabela A.2 - Perfil flavonoídico dos espécimes de *Huberia* D.C. (1=presença, 0=ausência)

Espécime	Flavonóides																					
	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15	F16	F17	F18	F19	F20	F21	F22
HG3	1	1	0	1	1	0	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0	0
HG6	0	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
HM1	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0
HM2	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0
HM3	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0
HN1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
HN2	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0
HN5	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	1	0	1	0	0	0	0
HO1	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	1	1	0	0	0
HO2	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
HO3	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
HO4	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
HS1	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1	1	1	0	0	0	0
HS2	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0
HS3	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0	0
HA	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Hsp1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
Hsp2	0	1	0	1	0	0	0	0	1	0	0	1	1	0	0	0	0	1	0	0	0	0

Tabela A.3 - Proporção de flavonóides em cada espécie

Espécie	n	Flavonóides																					
		F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15	F16	F17	F18	F19	F20	F21	F22
HG	2	50	100	0	100	50	0	0	100	0	0	0	0	0	0	0	0	50	50	0	100	0	0
HM	3	0	100	0	0	0	0	0	0	0	0	0	33	67	0	0	0	67	67	0	100	0	0
HN	3	0	67	33	0	0	0	33	0	0	0	0	0	0	33	33	67	0	100	0	33	0	33
HO	4	0	100	0	0	25	0	75	0	0	0	25	0	0	0	0	0	0	100	25	50	0	0
HS	3	0	67	0	0	0	0	67	0	33	33	33	0	0	0	0	100	67	33	0	67	0	0
HA	1	0	100	0	100	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Hsp1	1	0	100	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	100	0	0	100	0
Hsp2	1	0	100	0	0	0	0	0	0	100	0	0	100	100	0	0	0	0	100	0	100	0	0
Total	18	6	94	6	24	12	6	35	12	12	6	12	12	18	6	6	29	29	76	6	65	6	6