UNIVERSIDADE FEDERAL DO PARANÁ
SETOR DE CIÊNCIAS EXATAS
DEPARTAMENTO DE ESTATÍSTICA

CARACTERIZAÇÃO DO SISTEMA ESTUARINO-LAGUNAR DE CANANÉIA-IGUAPE – SP

CURITIBA
JUNHO 2008

REGINALDO DA SILVA
HAILTON MARCIO ARRUDA

CARACTERIZAÇÃO DO SISTEMA ESTUARINO-LAGUNAR DE CANANÉIA-IGUAPE – SP

Trabalho de graduação realizado para a disciplina de Laboratório de Estatística II do Curso de Estatística do Setor de Ciências Exatas, da Universidade Federal do Paraná.
Professor Orientador: Fernando Lucambio Perez.

CURITIBA
JUNHO 2008

Sumário

1 - Introdução 7

2 - Objetivos 8

3 – Metodologia 11

3.1 Análise de Cluster 11

3.2 - Medidas de Similaridade e Dissimilaridade 13

3.2.1 - Distância Euclidiana 13

3.2.2 - Distância Manhattan 14

3.2.3 - Distância Minkowski 15

3.3 - Gráfico da Silhueta 15

3.4 – Análise de Componentes Principais 16

4 – Resultados no Inverno 20

5 – Resultados no Verão 24

6 - Tabelas de Classificação 27

7 - Fatores Bióticos (Biológicos) 31

Conclusão 35

Anexos 36

Referências 42

Lista de Figuras

Figura 1 - Sistema estuarino-lagunar de Cananéia-Iguape 9

Figura 2 - Foraminíferos e Tecamebas 10

Figura 3 – Nova localização das amostras coletadas no inverno de 2003, no sistema estuarino-lagunar de Cananéia-Iguape. 29

Figura 4 – Nova localização das amostras coletadas no Verão de 2003, no sistema estuarino-lagunar de Cananéia-Iguape. 30

Lista de Gráficos

Gráfico 1 – Silhuetas do número de agrupamentos e as distintas distâncias – Inverno. 20

Gráfico 2 - Agrupamento das Comp. Principais (à esq.) e silhueta (à dir.) - Inverno.... 21

Gráfico 3 – Gráfico das comp. principais baseado nas variáveis Abióticas -Inverno.... 22

Gráfico 4 – Gráfico das comp. principais baseado nas variáveis Abióticas - Inverno... 22

Gráfico 5 – Silhuetas do número de agrupamentos e as devidas distâncias – Verão..... 24

Gráfico 6 - Agrupamento das comp. principais (à esq.) e silhueta (à dir.) - Verão..... 25

Gráfico 7 – Gráfico das comp. principais baseado nas variáveis Abióticas - Verão..... 26

Gráfico 8 – Gráfico das comp. crincipais baseado nas variáveis Abióticas - Verão..... 26

Gráfico 9 – Freqüências e freqüências relativas dos Clusters para variáveis bióticas - Verão. 32

Gráfico 10 –Freqüências e freqüências relativas dos Clusters para variáveis bióticas - Inverno..... 34

Lista de Tabelas

Tabela 1 – Variáveis das Comp. Principais 1 e 2 – Inverno..... 23

Tabela 2 - Variáveis das comp. principais 1 e 2 – Verão. 27

Tabela 3 - Classificação dos locais das estações dentro dos Clusters – Inverno/Verão. 27

Tabela 4 – Nova classificação das estações dentro de cada local. 28

Tabela 5 – Freqüência das observações dos fatores bióticos no Verão..... 31

Tabela 6 – Freqüência das observações dos fatores bióticos no Verão..... 31

Tabela 7 – Total de observações das variáveis bióticas no Verão..... 32

Tabela 8 – Freqüência relativa das variáveis bióticas para o Verão..... 32

Tabela 9 – Freqüência das observações dos fatores bióticos no Inverno. 32

Tabela 10 – Freqüência das observações dos fatores bióticos no Inverno. 33

Tabela 11 – Total de observações das variáveis bióticas no Inverno..... 33

Tabela 12 – Freqüência relativa das variáveis bióticas para o Inverno..... 33

1 - Introdução

Os ambientes estuarinos podem ser definidos de várias maneiras de acordo com a formação do especialista. Do ponto de vista geológico os estuários são feições efêmeras, cujo tempo de existência depende do balanço entre as taxas de sedimentação e as taxas de elevação/abaixamento do nível do mar. Em períodos de estabilidade do nível do mar, os estuários tendem a ser preenchidos pelos sedimentos trazidos pelas correntes de maré e pelos rios que deságuam no estuário. Em áreas estuarinas podem ser definidos sub-ambientes por meio de associações de foraminíferos.

Foraminíferos são microorganismos que tem sua distribuição controlada, principalmente, por fatores físicos, tais como, luz, salinidade, temperatura, etc.

Ambientes estuarinos são controlados pelas variações das influencias de origem marinha e fluvial, gerando diferentes gradientes de salinidade, temperatura, natureza de substrato, teor de carbono orgânico, PH, EH e amplitude das marés.

Em vida os foraminíferos participam ativamente da ciclagem do material orgânico e, após a morte, desde que não sofram a dissolução da suas carapaças, passam a fazer parte constituinte dos sedimentos marinhos. As carapaças dos foraminíferos geralmente permanecem bem preservadas após a morte, podendo ser utilizadas por pesquisadores para classificar estratos de antigos ambientes deposicionais, auxiliar no reconhecimento de depósitos naturais de hidrocarbonetos, permitem acompanhar a história evolutiva de ambientes costeiros, na determinação apurada das variações do nível do mar e tem sido amplamente utilizados em estudos de áreas impactadas por poluição orgânica e inorgânica.

Outro grupo de importância ambiental associado aos foraminíferos é o das Tecamebas. Tais organismos são considerados bons indicadores na detecção de ambientes deteriorados por metais pesados em ambientes poluídos.

2 - Objetivos

A pesquisadora procurou o Labest para saber se a região amostrada poderia ser dividida em quatro sub-regiões geográficas ou se pelas características físicas o agrupamento seria diferente. Além disso, uma outra questão seria saber se entre o inverno e verão as sub-regiões seriam similares ou não.

Posteriormente objetivou-se identificar os chamados indicadores biológicos, isto é, espécies de animais mais freqüentes do que outras, nas diferentes sub-regiões e estações climáticas.

Figura 1 - Sistema estuarino-lagunar de Cananéia-Iguape

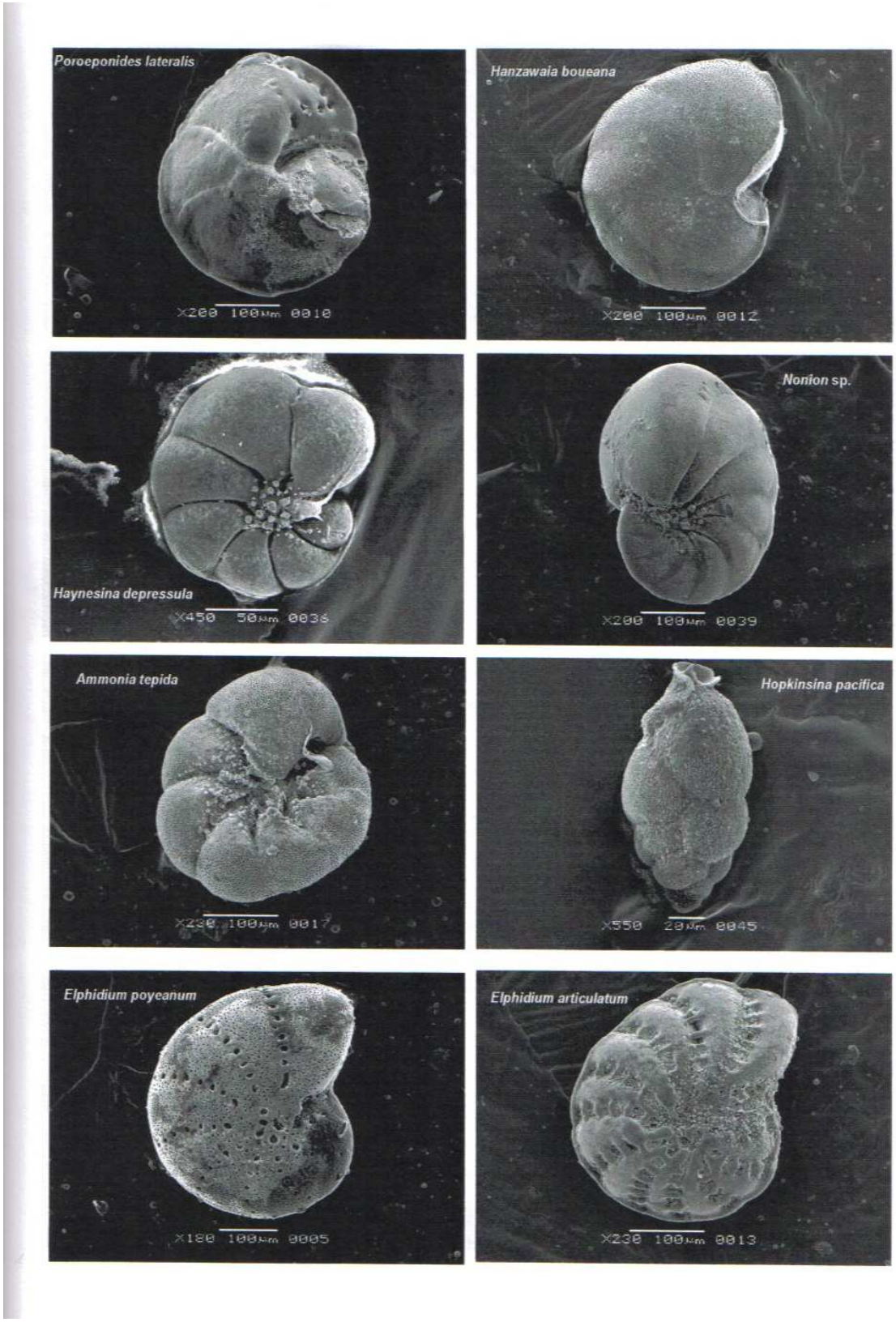

Figura 2 - Foraminíferos e Tecamebas

3 – Metodologia

3.1 Análise de Cluster

Análise de cluster, também conhecida como análise de conglomerados ou agrupamentos, é um conjunto de técnicas estatísticas cujo objetivo é agrupar objetos segundo suas características, formando grupos ou conglomerados homogêneos. A técnica classificatória multivariada da análise de agrupamentos pode ser utilizada quando se deseja explorar as similaridades entre indivíduos ou entre variáveis definindo-os em grupos, considerando simultaneamente, no primeiro caso, todas as variáveis medidas em cada indivíduo e, no segundo, todos os indivíduos nos quais foram feitas as mesmas mensurações. O agrupamento aqui não é conhecido para cada observação individual. A análise de cluster pretende fornecer uma avaliação objetiva de quantos subgrupos diferentes os dados contém.

Os objetos em cada conglomerado tendem a serem semelhantes entre si, porém diferentes dos demais objetos dos outros conglomerados.

A análise de cluster é uma ferramenta de análise exploratória de dados que tem como objetivo atribuir diferentes objetos a grupos de forma que o grau de associação entre dois objetos é máxima se eles pertencerem ao mesmo grupo e mínimo de outra forma. Se a aglomeração for bem sucedida quando representados em um gráfico, os objetos dentro dos conglomerados estarão muito próximos, e os conglomerados distintos estarão afastados.

Os algoritmos utilizados na formação dos agrupamentos são divididos em duas categorias: Não-hierárquico e Hierárquico. O método não-hierárquico caracteriza-se por dividir as observações num conjunto pré-determinado de objetos grupados. Há dois modos de fazer isso: com a análise de cluster Kmeans ou com a análise de clustermedians. A vantagem dos métodos não-hierárquicos é que em geral eles são mais simples e mais rápidos de serem operacionalizados por algum programa computacional do que os métodos

tradicionais. A desvantagem do método não-hierárquico está na necessidade que o pesquisador tem de declarar antecipadamente o número exato de clusters.

O método hierárquico começa freqüentemente com cada objeto ou observação em um grupo separado. Os dois procedimentos hierárquicos mais utilizados pelos pesquisadores são os métodos aglomerativo – em que o procedimento começa com cada objeto em um grupo separado, de forma que, em cada passo seguinte, os dois agrupamentos de objetos que são mais próximos (parecidos) são combinados para construir um novo agrupamento até que todos os objetos sejam combinados em um único agrupamento – e o divisivo, cujo procedimento de agrupamento começa com todos os objetos em um único agrupamento que é dividido em cada passo em dois agrupamentos que contêm os objetos mais distintos (EVERITT, 1980; HAIR JR. et al., 2005).

Ambos os métodos geram, como resultado gráfico, uma estrutura hierárquica em forma de árvore, chamada dendograma, que representa a formação gráfica dos clusters.

O dendograma é um meio prático e comum de representar os resultados de uma análise de cluster. Consiste de uma árvore de agrupamento hierárquico cuja altura de cada linha denota a distância entre dois objetos que estão sendo conectados.

De modo geral, os métodos de análise de cluster são de 2 tipos:

1º) Método hierárquico divisivo: Dividem o conjunto de dados em k clusters não sobrepostos, assim os objetos de um cluster estão próximos uns dos outros e objetos de diferentes clusters são dissimilares.

2º) Método hierárquico aglomerativo: Constroem um dendograma. Um método aglomerativo começa com uma situação em que cada objeto do conjunto de dados forma seu próprio cluster, e então sucessivas junções de clusters são realizadas até que apenas um grande cluster permaneça, que é o conjunto todo de dados.

3.2 - Medidas de Similaridade e Dissimilaridade

Uma questão importante refere-se ao critério a ser utilizado para se decidir até que ponto dois elementos podem ser considerados semelhantes ou não.

Dissimilaridades são números não-negativos d(i,j) que são pequenos quando i e j são próximos um do outro e se tornam grandes quando i e j são muito diferentes.

Os coeficientes de dissimilaridade mais usuais, obtidos num espaço multidimensionais, podem ser subdivididos em três categorias.

1. Os que medem a distância ou a separação angular entre pares de pontos;
2. Os que medem a correlação entre pares de valores;
3. Os que medem a associação entre pares de caracteres qualitativos;

Nesse caso, as dissimilaridades são chamadas distâncias.

Temos à disposição na linguagem de programação R as distâncias Euclidiana, Manhattan e Minkowski

3.2.1 - Distância Euclidiana

Considere o vetor x de coordenadas reais (x_1, x_2 , ..., x_p) como descritor dos objetos que investigarão os assemelhamentos. A medida mais conhecida para indicar a proximidade entre os objetos A e B é a distância euclidiana d(A,B):

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

em que:

d(A,B) = distância Euclidiana

X_i(A) = valor de abundância para a amostra i na área X;

Y_i(B) = valor de abundância para a amostra i na área Y;

n = número de amostras existentes.

A distância Euclidiana é uma das medidas de dissimilaridade entre comunidades mais utilizadas na prática (GAUCH, 1982). De acordo com BROWER e ZAR (1977), quanto menor o valor da distância Euclidiana entre duas comunidades, mais próximas elas se apresentam em termos de parâmetros quantitativos por amostra.

3.2.2 - Distância Manhattan

De uma maneira mais formal, podemos definir a distância de Manhattan entre dois pontos num espaço euclidiano com um sistema cartesiano de coordenadas fixo como a soma dos comprimentos da projeção da linha que une os pontos com os eixos das coordenadas.

Por exemplo, num plano que contém os pontos P1 e P2, respectivamente com as coordenadas (x_1,y_1) e (x_2,y_2), é definido por:

DM = $|x_1 - x_2| + |y_1 - y_2|$.

Note-se que a distância de Manhattan depende da rotação do sistema de coordenadas mas não da sua translação ou da sua reflexão em relação a um eixo coordenado.

3.2.3 - Distância Minkowski

A distancia Minkowski entre X e Y é dada por:

$$d(x,y) = \left(|x_1 - y_1|^q + |x_2 - y_2|^q + \ldots + |x_n - y_n|^q \right)^{1/q} , \text{ onde } q \in N.$$

Esta distância é a generalização das duas distâncias anteriores.

Quando q = 1, esta distância representa a distância de Manhattan e quando q = 2, a distância Euclidiana.

3.3 - Gráfico da Silhueta

Para se ter uma idéia de como os clusters resultantes estão bem separados, podemos fazer uso do Gráfico de Silhueta (ROUSSEUW, 1987).

Este gráfico nos dá uma medida de quão perto cada observação em um cluster está dos pontos nos clusters vizinhos.

O valor de silhueta s(i) do objeto i é definido como:
S(i) = b(i) – a(i) / max {a(i), b(i)}, sendo

a(i) distância media do objeto i para os objetos do seu próprio grupo.
b(i) distância media do objeto j para os objetos do seu próprio grupo.

Claramente, s(i) fica restrito entre -1 e 1. O valor de s(i) pode ser interpretado da seguinte forma:

s(i) = 1 → o objeto i está bem classificado (em a)
s(i) = 0 → o objeto i está entre dois clusters (a e b)
s(i) = -1 → o objeto i está mal classificado (mais perto de b que de a)

A silhueta do cluster A é um gráfico de todos os seus s(i), plotados em ordem crescente. Para cada observação i, uma barra é desenhada, representando sua largura de silhueta s(i). O gráfico de silhueta inteiro mostra as silhuetas de todos os clusters, um embaixo do outro. Assim, a qualidade dos clusters pode ser comparada: uma silhueta larga é melhor que uma silhueta estreita.

3.4 – Análise de Componentes Principais

Entre as várias alternativas que existem para reduzir a dimensionalidade do modelo, uma delas consiste na utilização de componentes principais. Como nos modelos de regressão, cujo propósito é a explicação da variável dependente, deve-se reter aquelas componentes principais que têm altas correlações com a variável dependente. No caso de um modelo de regressão multivariada, analisam-se as correlações das variáveis independentes com cada uma das variáveis dependentes. Existe uma tendência para os dados com componentes de grandes variâncias de melhor explicar as variáveis dependentes (MARDIA, KENT e BIBBY, 1982).

O objetivo principal da análise de componentes principais é a obtenção de um pequeno número de combinações lineares (componentes principais) de um conjunto de variáveis, que retenham o máximo possível da informação contida nas variáveis originais. Freqüentemente, um pequeno número de componentes pode ser usado, em lugar das variáveis originais, nas análises de regressões, análises de agrupamentos etc.

Os componentes são extraídos na ordem do mais explicativo para o menos explicativo. Teoricamente o número de componentes é sempre igual ao número de variáveis. Entretanto, alguns poucos componentes são responsáveis por grande parte da explicação total.

O processamento da análise de componentes principais pode ter partida na matriz de variâncias e covariâncias ou na matriz de correlação. Se você optar pela matriz de correlação, é aconselhável estabelecer o limite mínimo de 1.0 unidade para a extração dos autovalores.

Para investigar as relações entre um conjunto de p variáveis correlacionadas (X_1, X_2,... , X_p) pode ser útil transformar o conjunto de variáveis originais em um novo conjunto de variáveis não-correlacionadas chamadas componentes principais (Y_1,Y_2,... ,Y_p) de modo que Y_1 é aquela que explica a maior parcela da variabilidade total dos dados, Y_2 explica a segunda maior parcela e assim por diante, tendo propriedades especiais em termos de variâncias.

Algebricamente, as componentes principais são combinações lineares de p variáveis originais: X_1, X_2,... , X_p.

Geometricamente, as combinações lineares representam a seleção de um novo sistema de coordenadas, obtido por rotação do sistema original com X_1, X_2, ... , X_p como eixos. Os novos eixos, Y_1,Y_2, ... , Y_p, representam as direções com variabilidade máxima, permitindo uma interpretação mais simples da estrutura da matriz de covariância.

Seja X'= [X_1, X_2, ..., X_p] um vetor aleatório p-dimensional com vetor de médias μ, matriz de covariância Σ e autovalores: $\lambda_1 \geq \lambda_2 \geq$, ..., λ_p.
Considere as combinações lineares:

Y = C' X onde:

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_p \end{bmatrix} \quad e \quad C = \begin{bmatrix} c_{11} & c_{12} & --- & c_{1p} \\ c_{21} & c_{22} & --- & c_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ c_{p1} & c_{p2} & --- & c_{pp} \end{bmatrix}$$

com:

$$E(Y_j) = E(c'_j X) = c'_j E(X) = c'_j \mu$$

$$V(Y_j) = V(c'_j X) = c'_j V(X) c_j = c'_j \Sigma c_j$$

$$Cov(Y_i, Y_j) = V(c'_i X, c'_j X) = c'_i \Sigma c'_j$$

$$i \neq j = 1, 2, \ldots, p.$$

O método de eliminação de variáveis explicativas pelo uso de componentes principais não é o único existente, por ser comum a aplicação dos métodos de regressão que envolve a análise do coeficiente de determinação. No entanto, o método das componentes principais permite uma redução significativa no número de variáveis, fundamentalmente quando se tem um significado adequado para a componente retida, a qual pode ser tratada como a nova variável explicativa.

A aplicação desse método é adequado principalmente nos casos envolvendo um número muito grande de variáveis explicativas em que as componentes principais têm uma interpretação significativa para o pesquisador. A substituição das variáveis explicativas originais pelas componentes principais retidas proporciona um modelo com uma redução substancial no número de variáveis explicativas.

As componentes principais aplicam-se a análise de cluster para mostrar graficamente os agrupamentos obtidos.

Uma decisão a ser tomada diz respeito ao número de componentes principais que deve ser retido na análise. Se esse número é muito pequeno pode ser haver uma redução exagerada da dimensionalidade e muita informação pode ser perdida. Se o número é grande, pode-se não atender aos objetivos de redução. Na verdade, essa redução depende das correlações e das variâncias das variáveis originais.

Critérios para determinar o número de componentes principais a serem retidas na análise:

Critério de Kaiser (1958), o qual sugere considerar apenas os componentes com autovalor superior a 1, o que significa que o componente contabiliza mais variância do que uma variável.

Critério da proporção: observa-se a proporção de variância acumulada e um nível de corte é estabelecido, representando o total da variância contabilizado pelos componentes selecionados; e

Scree test: através de uma análise gráfica, consideram-se apenas os componentes situados antes de um certo intervalo, se houver.

Cada estação ou unidade experimental foi amostrada em dois diferentes momentos: Março de 2003 (Verão) e Julho de 2003 (Inverno).

4 – Resultados no Inverno

Iniciamos a análise considerando as 81 unidades amostrais e a informação das características ambientais disponíveis: Temperatura, salinidade, granulo, areia muito grossa, areia grossa, areia média, areia fina, areia muito fina, profundidade e argila. Mais detalhadamente: Razão C/N, carbonato de cálcio, silte grosso, silte fino, razão C/S, carbono orgânico, nitrogênio total e enxofre total. Identificamos primeiro o número de agrupamentos, para isso realizamos o gráfico 1.

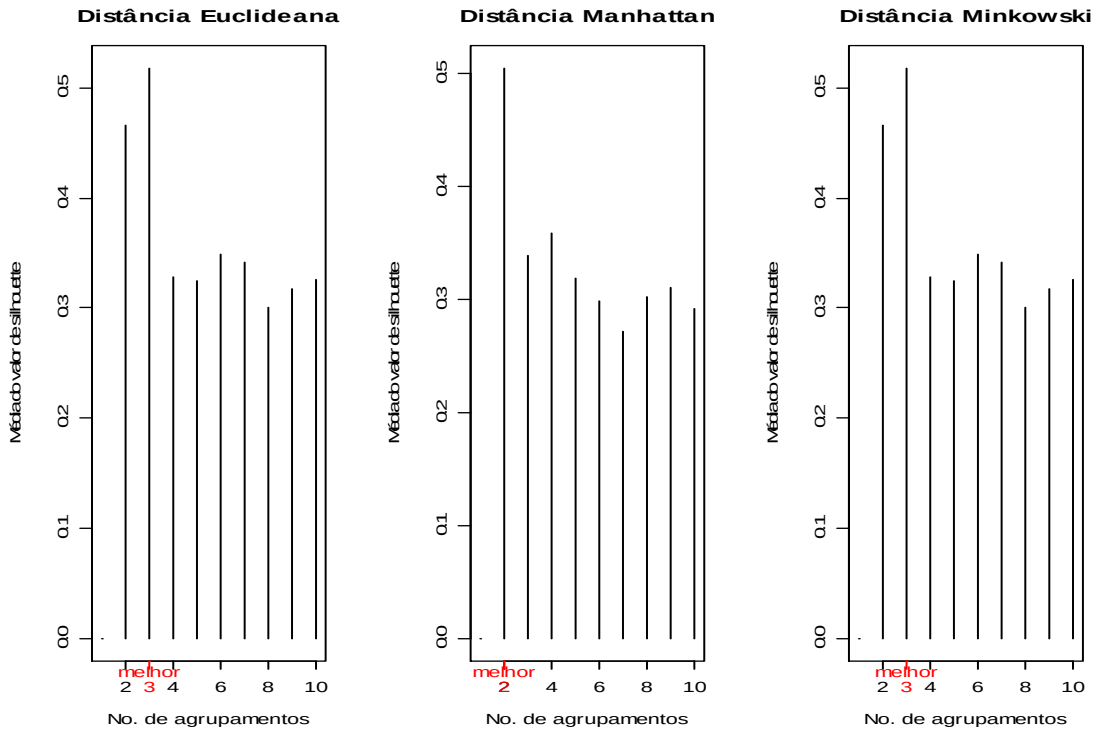

Gráfico 1 – Silhuetas do número de agrupamentos e as distintas distâncias – Inverno.

No gráfico 2 temos duas figuras: a primeira refere-se ao agrupamento obtido pelas componentes principais, e a segunda refere-se à silhueta dos clusters formados.

Percebemos que em duas distâncias o número de agrupamentos obtido foi 3 e somente na distância Manhattan o valor máximo da silhueta é obtido quando escolhemos somente 2 cluster. Por esta razão, decidimos realizar a análise de agrupamentos para agrupar as estações em 3 grupos, utilizando a distância euclidiana.

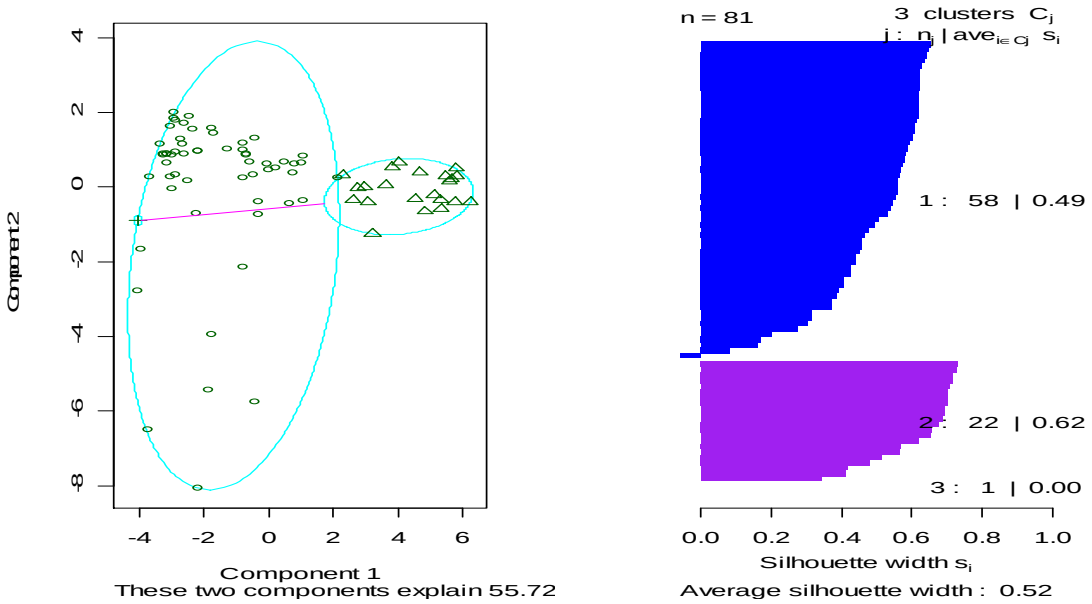

Gráfico 2 - Agrupamento das Comp. Principais (à esq.) e silhueta (à dir.) - Inverno.

Considerando 2 clusters, que são o número de regiões, notamos que os clusters estão bem definidos, o que pode ser verificado pelo valor de cada silhueta obtida, bem como pelo valor da média geral das silhuetas (0,52).

No cluster 1 há ocorrência de valor negativo de silhueta, indicando que tais amostras não estão bem classificadas.

Podemos observar que uma das unidades amostrais não se assemelha a nenhuma outra, formando assim um cluster de tamanho 1. Isso não tem sentido prático, por isso, selecionamos 2 clusters e observamos que pertencia ao cluster 3, passando a fazer parte do cluster, mantendo-se as outras unidades amostrais nos mesmos agrupamentos mostrados no gráfico 2.

O método de análise de componentes principais foi aplicado para o conjunto de dados no Inverno envolvendo todas as variáveis Abióticas explicativas do Sistema Estuarino-Lagunar de Cananéia Iguape. A análise foi realizada com uso do software R e o método utilizado proporcionou a redução de todas as variáveis explicativas para apenas 2 componentes principais, componente1 e componente 2.

Sendo que as variáveis Abióticas da componente 1 são Lama, Areia e Argila, as quais representam a maior variabilidade dos dados. Também na Componente 2 temos as variáreis Abióticas identificadas pela sua maior representatividade, as quais são Areia Muito Grossa, Areia Grossa e Granulometria.

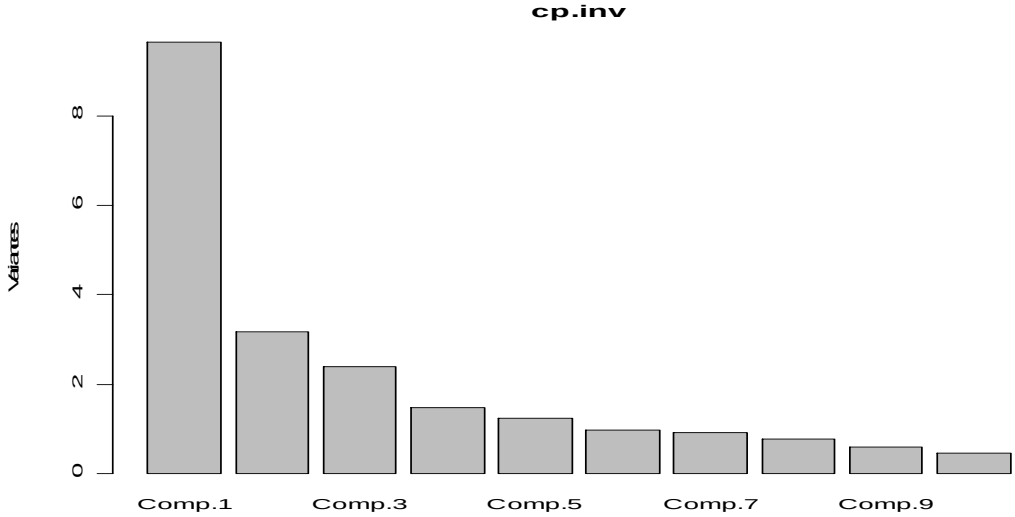

Gráfico 3 – Gráfico das comp. principais baseado nas variáveis Abióticas -Inverno.

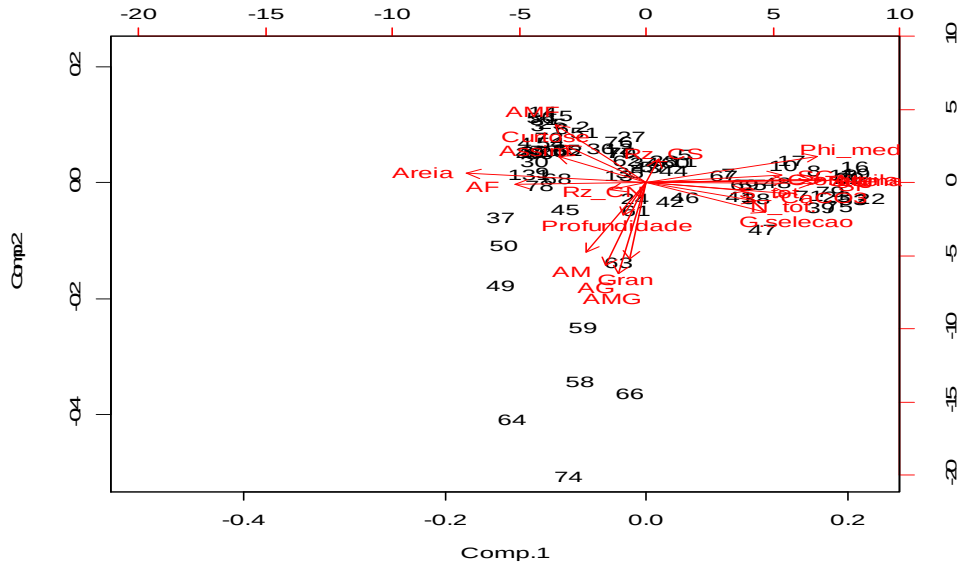

Gráfico 4 – Gráfico das comp. principais baseado nas variáveis Abióticas - Inverno.

Loadings:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7
Profundidade		-0.173	0.110		0.675	0.185	0.225
CaCO3	0.250		0.118	0.191	0.233	-0.145	
Gran		**-0.406**		-0.318		0.461	
AMG		**-0.485**		-0.251		0.167	
AG		**-0.442**			-0.121	-0.425	
AM	-0.105	-0.369	-0.176	0.151		-0.465	
AF	-0.230		-0.118	0.364	-0.160	0.188	-0.110
AMF	-0.159	0.310	0.273	-0.290			
SG	0.238		-0.248	-0.224			
SF	0.294		-0.144				0.152
Areia	**-0.315**						
Silte	0.291		-0.205	-0.136			
Argila	**0.309**				0.124		
Lama	**0.317**						
Phi_medio	0.303	0.143			0.115		
G.selecao	0.212	-0.158	0.275		0.179	-0.102	-0.430
Assim.	-0.153	0.143	0.281	-0.199	0.127	-0.180	-0.520
Curtose	-0.142	0.201		-0.244			0.571
C.org.	0.249		0.244		-0.325		
N_tot.	0.193		0.243		-0.418	0.215	
S_tot.	0.176		0.424	0.240	-0.223		0.139
Rz_CN			-0.197	0.433		0.399	-0.180
Rz_CS		0.125	-0.453	-0.322	-0.123		-0.193

Tabela 1 – Variáveis das Comp. Principais 1 e 2 – Inverno.

Pode-se identificar na Tabela 1 todas as variáveis abióticas que formam a Componente Principal 1, destacando-se Areia, Argila e Lama entre as que explicam melhor a variabilidade. Similarmente, identificam-se as variáveis Grânulo, Areia Muito Grossa e Areia Grossa entre as que explicam melhor a variabilidade da Componente Principal 2.

5 – Resultados no Verão

Procedimento similar foi realizado nas observações, nas mesmas estações, mas no Verão. Observamos no gráfico 5 que o número de clusters apropriado é 2. Com base nestes dados, decidimos realizar a análise de agrupamentos, como no caso anterior, para agrupar as estações em 2 grupos, utilizando a distância euclidiana.

Podemos identificar as devidas distâncias Euclidiana, Máxima e Minkowski no gráfico 5.

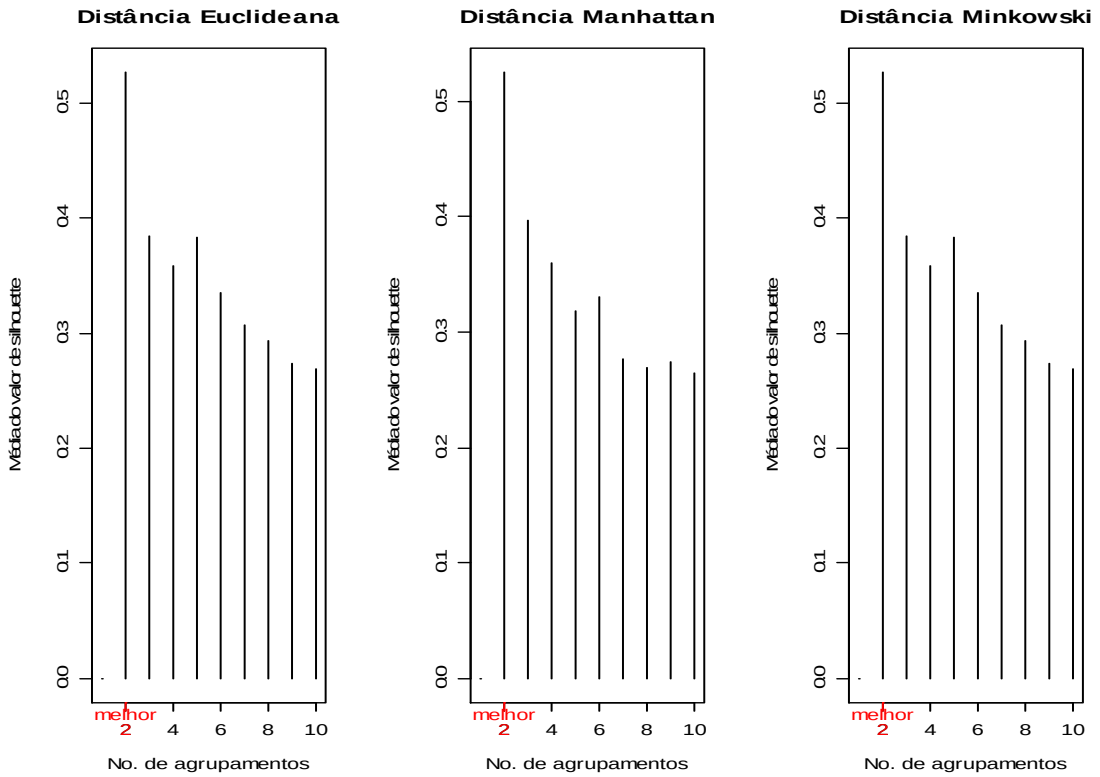

Gráfico 5 – Silhuetas do número de agrupamentos e as devidas distâncias – Verão.

Considerando 2 clusters, que são o número de regiões, pode-se observar que os clusters estão bem definidos, o que pode ser verificado pelo valor de cada silhueta obtida, bem como pelo valor da média geral das silhuetas (0,53).

No cluster 1 há ocorrência de valor negativo de silhueta, indicando que tais amostras não estão bem classificadas.

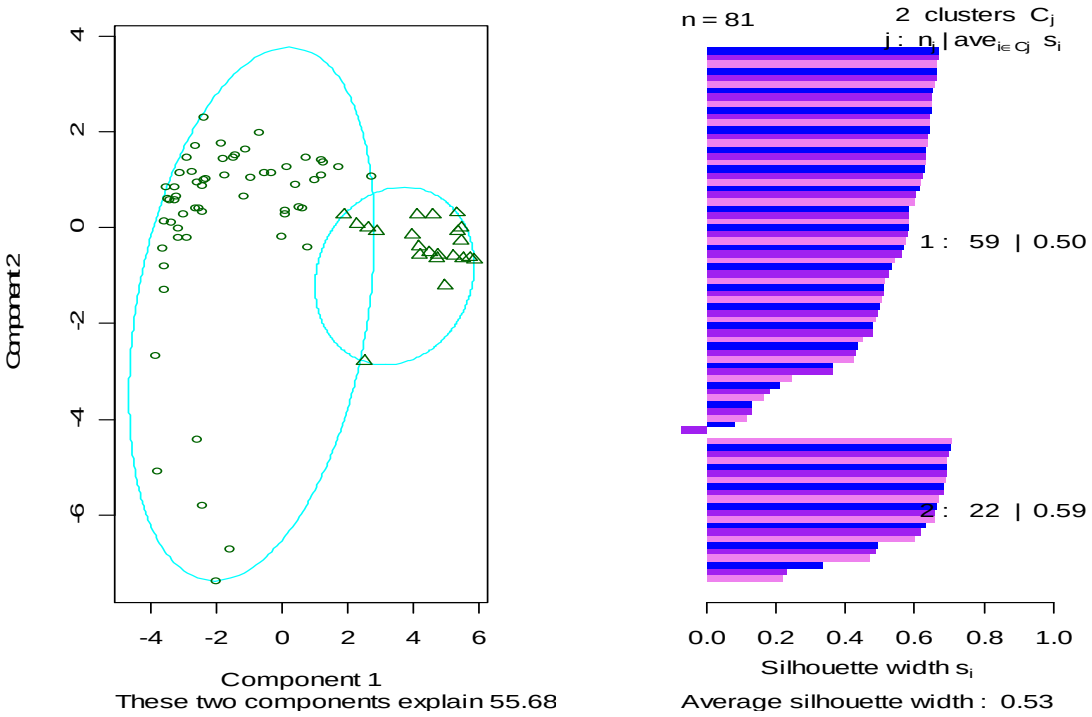

Gráfico 6 - Agrupamento das comp. principais (à esq.) e silhueta (à dir.) - Verão.

O método de análise de componentes principais foi aplicado para o conjunto de dados no Inverno envolvendo todas as variáveis Abióticas explicativas do Sistema Estuarino-Lagunar de Cananéia Iguape. A análise foi realizada com uso do software R e o método utilizado proporcionou a redução de todas as variáveis explicativas para apenas 2 componentes principais.

Sendo que as variáveis Abióticas da Comp.1 são Lama, Areia e Argila, as quais representam a maior variabilidade dos dados. Também na Comp.2 temos as variáreis Abióticas identificadas pela sua maior representatividade, as quais são Areia Muito Grossa, Areia Grossa e Areia Média.

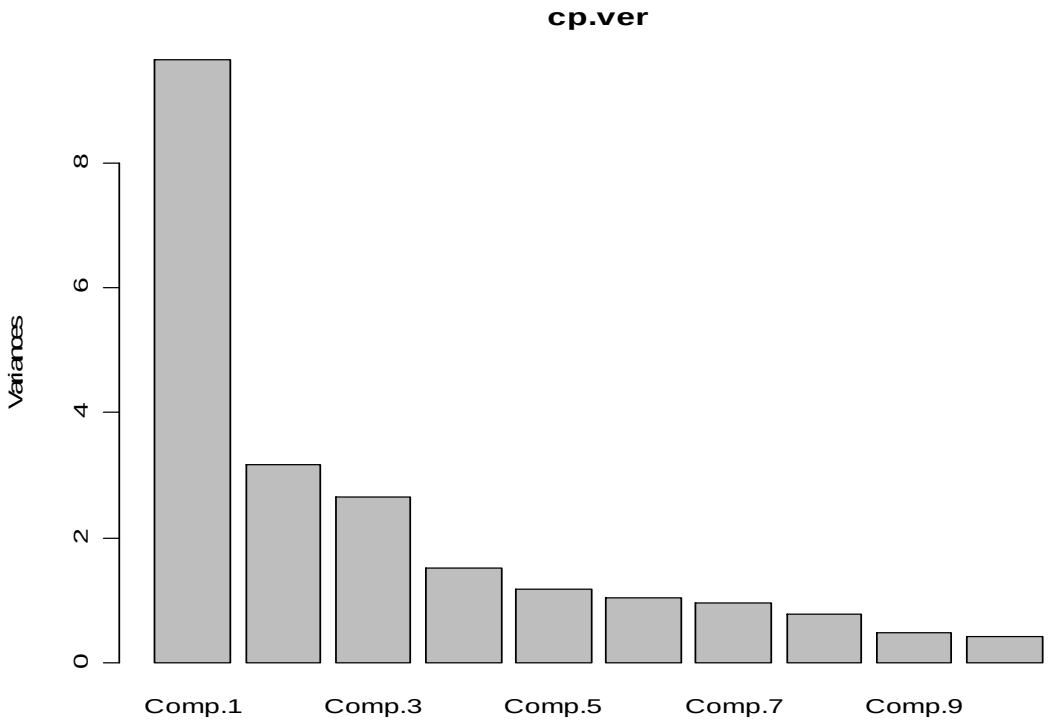

Gráfico 7 – Gráfico das comp. principais baseado nas variáveis Abióticas - Verão.

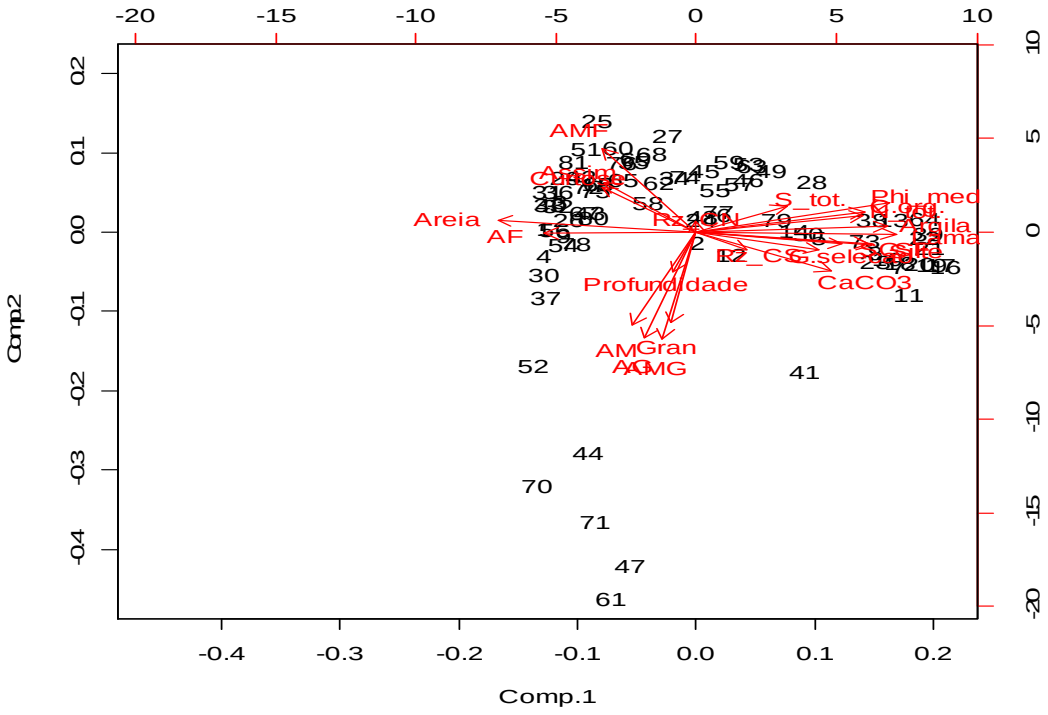

Gráfico 8 – Gráfico das comp. crincipais baseado nas variáveis Abióticas - Verão.

Loadings:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7
Profundidade		-0.167	-0.174	-0.278	0.115	0.716	0.245
CaCO3	0.214	-0.160			0.154	0.246	0.381
Gran		-0.378	-0.228	-0.195	0.282	-0.109	-0.386
AMG		**-0.444**	-0.219	-0.166	0.190		-0.283
AG		**-0.442**			-0.107	-0.220	0.177
AM	-0.102	**-0.387**		0.138	-0.433		0.234
AF	-0.243		0.197	0.248		0.202	-0.171
AMF	-0.149	0.347	-0.208	-0.208	0.234	-0.180	0.156
SG	0.233		0.169	-0.327		-0.188	
SF	0.275		0.161	-0.160			-0.178
Areia	**-0.317**						
Silte	0.279		0.180	-0.251			-0.100
Argila	**0.305**			0.109			
Lama	**0.318**						
Phi_medio	0.308	0.121					
G.selecao	0.196		-0.350			-0.197	0.335
Assim.	-0.146	0.207	-0.273	-0.302		-0.296	0.244
Curtose	-0.152	0.188		-0.368		0.211	-0.239
C.org.	0.268		-0.198	0.144			-0.230
N_tot.	0.267		-0.179	0.196			-0.157
S_tot.	0.144	0.110	-0.436	0.274	-0.137		-0.168
Rz_CN			-0.118	-0.355	-0.735		-0.138
Rz_CS			0.461	-0.127	0.126	-0.227	0.127

Tabela 2 - Variáveis das comp. principais 1 e 2 – Verão.

Pode-se identificar na Tabela 1 todas as variáveis abióticas que formam a Componente Principal 1, destacando-se Areia, Argila e Lama entre as que explicam melhor a variabilidade. Similarmente, identificam-se as variáveis Areia Muito Grossa, Areia Grossa e Areia Média entre as que explicam melhor a variabilidade da componente principal 2.

6 - Tabelas de Classificação

Tabela 3 - Classificação dos locais das estações dentro dos Clusters – Inverno/Verão.

Inverno		Cluster	
	Local	1	2
Local 1	Baia do Trapandé	13	2
Local 2	Mar de Cananéia	12	6
Local 3	Mar de Cubatão	13	4
Local 4	Mar Pequeno	20	11
		58	**23**

Verão		Cluster	
	Local	1	2
Local 1	Baia do Trapandé	13	2
Local 2	Mar de Cananéia	14	4
Local 3	Mar de Cubatão	16	1
Local 4	Mar Pequeno	16	15
		59	**22**

Tabela 4 – Nova classificação das estações dentro de cada local.

Tabela de Classificação Inverno		
Estação	Local Original	Nova Classificação
3	1	1
6	1	1
8	1	1
11	1	1
12	1	1
14	1	2
17	1	2
18	1	1
20	1	1
21	1	2
24	1	1
29	1	1
31	1	1
33	1	1
39	1	2
43	1	2
47	1	2
54	1	2
62	1	2
70	1	2
71	1	2
74	1	2
79	1	1
87	1	1
91	1	1
95	1	1
99	1	1
103	1	1
108	1	1
111	1	1
120	1	1
124	2	1
130	2	1
132	2	1
137	2	1
142	2	2
147	2	2
148	2	2
153	2	1
158	3	1
164	3	1
165	3	1
170	3	1
175	3	1
180	3	1
183	3	1
191	3	1
194	3	2
196	3	2
199	3	1
203	2	1
206	4	1
212	4	1
217	4	2
218	4	1
221	4	2
223	4	1
225	4	1
228	4	2
230	4	1
234	4	1
236	4	1
239	4	2
240	4	1
244	4	1
249	4	1
250	4	1
254	4	1
255	3	1
260	3	1
267	3	1
268	3	1
271	2	2
279	2	1
281	2	1

Tabela de Classificação Verão		
Estação	Local Original	Nova Classificação
3	1	1
6	1	1
8	1	1
11	1	1
12	1	2
14	1	1
17	1	2
18	1	2
20	1	1
21	1	2
24	1	2
29	1	1
31	1	2
33	1	2
39	1	1
43	1	2
47	1	2
54	1	2
62	1	2
70	1	2
71	1	2
74	1	2
79	1	2
87	1	1
91	1	1
95	1	1
99	1	1
103	1	1
108	1	1
111	1	1
120	2	1
124	2	1
130	2	1
132	2	1
137	2	1
142	2	1
147	2	1
148	2	2
153	2	2
158	3	1
164	3	2
165	3	1
170	3	1
175	3	1
180	3	1
183	3	1
191	3	1
194	3	1
196	3	1
199	3	2
203	2	1
206	4	1
212	4	1
217	4	1
218	4	1
221	4	1
223	4	1
225	4	1
228	4	1
230	4	1
234	4	1
236	4	1
239	4	1
240	4	2
244	4	1
249	4	1
250	4	1
254	4	1
255	3	1
260	3	1
267	3	1
268	3	1
271	2	2
279	2	1
281	2	1

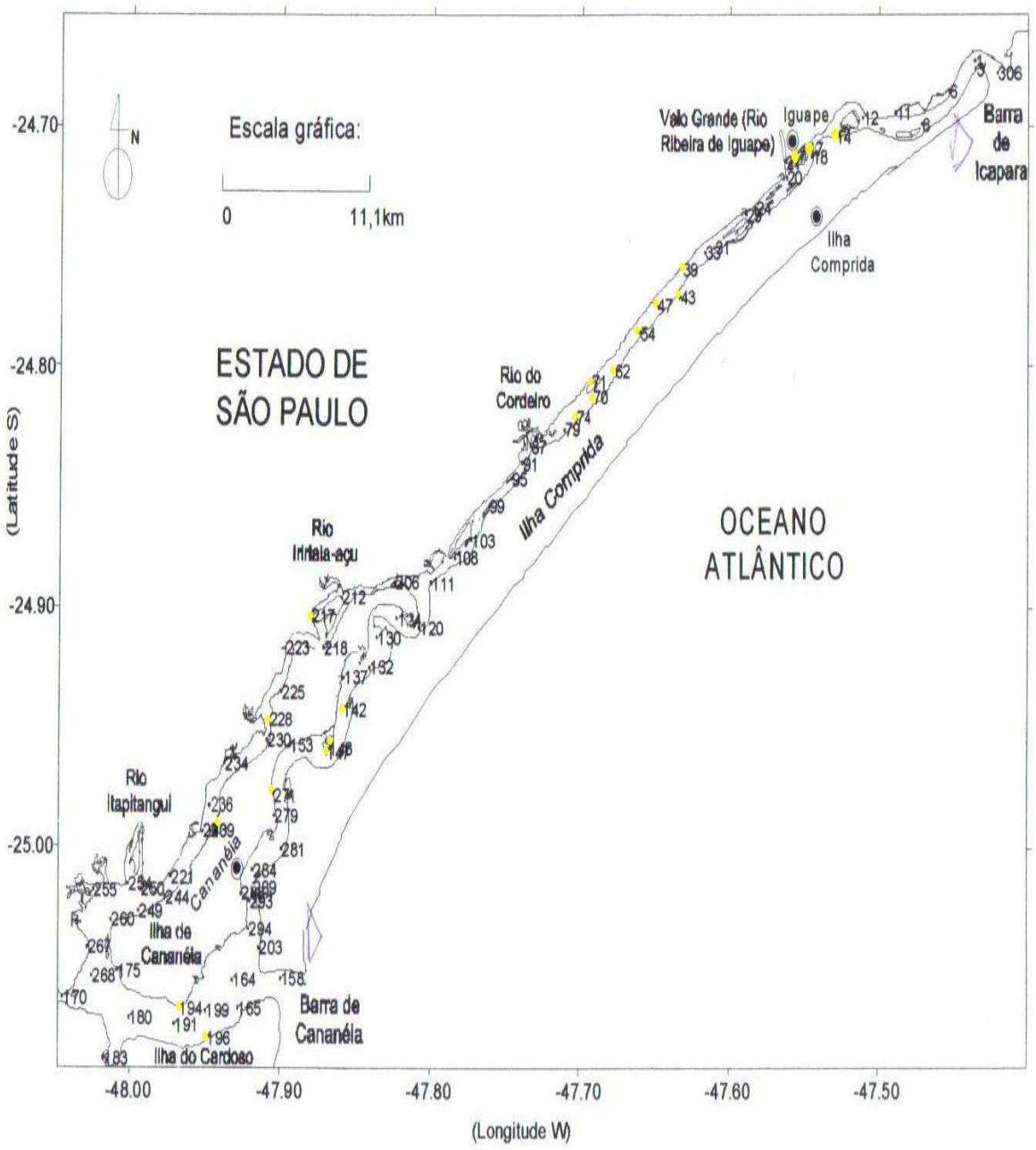

Figura 3 – Nova localização das amostras coletadas no inverno de 2003, no sistema estuarino-lagunar de Cananéia-Iguape.

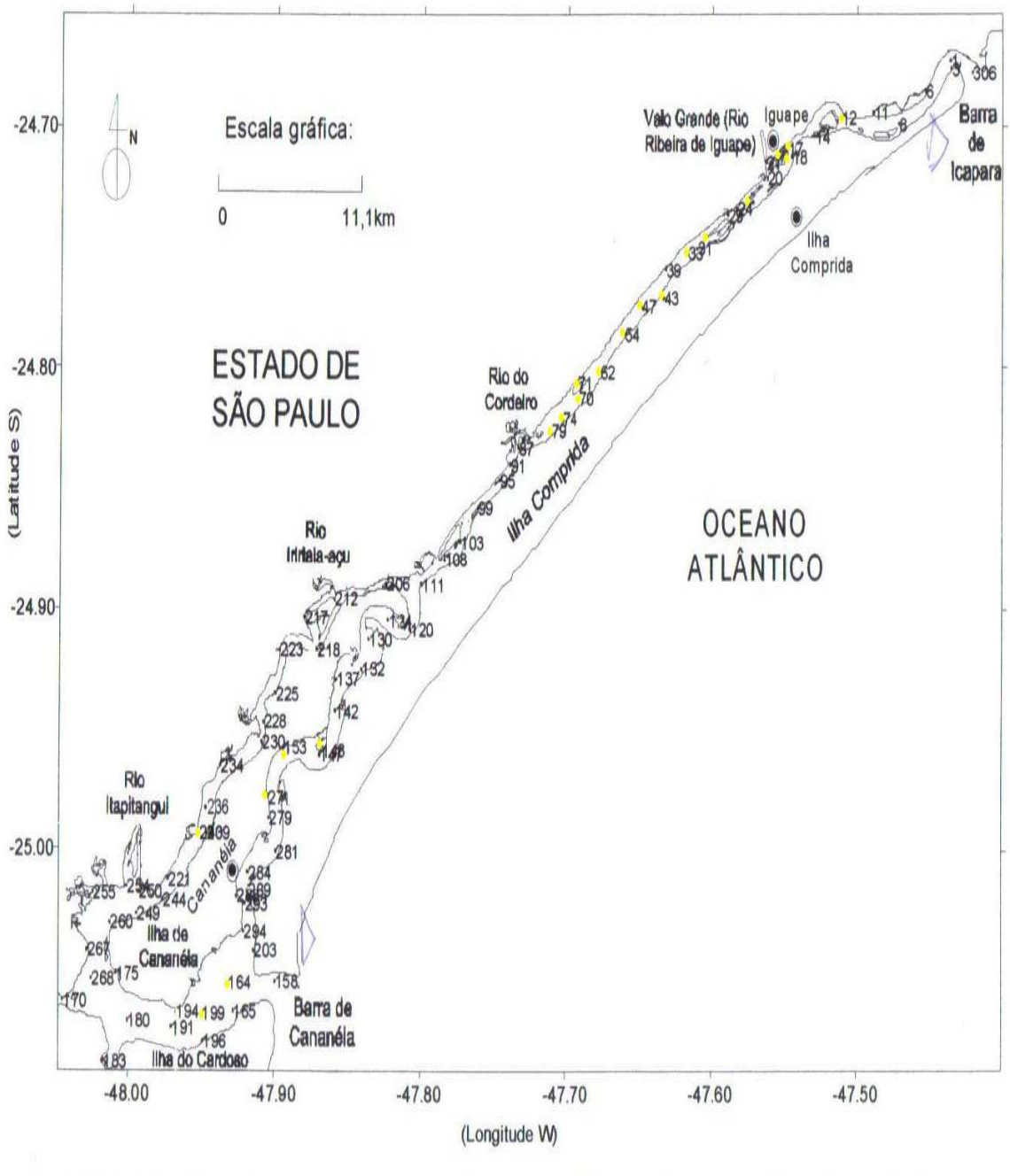

Figura 4 – Nova localização das amostras coletadas no Verão de 2003, no sistema estuarino-lagunar de Cananéia-Iguape.

7 - Fatores Bióticos (Biológicos)

Muitos fatores bióticos podem influenciar na formação dos grupos a serem estudados. Observemos agora as variáveis bióticas para cada novo cluster. Os fatores bióticos foram representados pelos grupos de espécies identificadas como:

FCH - Foraminíferos Calcários Hialinos.
FCP – Foraminíferos Calcários Porcelanáceos.
ALLG – Espécie Blysmasphaera Brasiliensis
FA – Foraminíferos Aglutinantes.
TEC – Tecamebas.

Primeiro construiremos as tabelas de freqüências para os fatores bióticos no Verão.

Tabela 5 – Freqüência das observações dos fatores bióticos no Verão.

	FCH	FCP	Allg	FA	TEC
Cluster 1	18550	2484	6	6458	660
Cluster 2	2535	1152	0	2140	926

Observamos a ausência de allg, por isso o desconsideramos para análise dos fatores através da tabela de freqüências. Passamos desta forma a considerar a tabela de 5 variáveis para apenas 4 variáveis, a qual é representada pela tabela 6.

Tabela 6 – Freqüência das observações dos fatores bióticos no Verão.

	FCH	FCP	FA	TEC
Cluster 1	18550	2484	6458	660
Cluster 2	2535	1152	2140	926

Uma vez em posse desta tabela calculamos a freqüência relativa em relação ao total de fatores bióticos, mostrado na tabela 8.

Tabela 7 – Total de observações das variáveis bióticas no Verão.

FCH	FCP	FA	TEC
21085	3636	8598	1586

Tabela 8 – Freqüência relativa das variáveis bióticas para o Verão

	FCH	FCP	FA	TEC
Cluster 1	88,0%	68,3%	75,1%	41,6%
Cluster 2	12,0%	31,7%	24,9%	58,4%

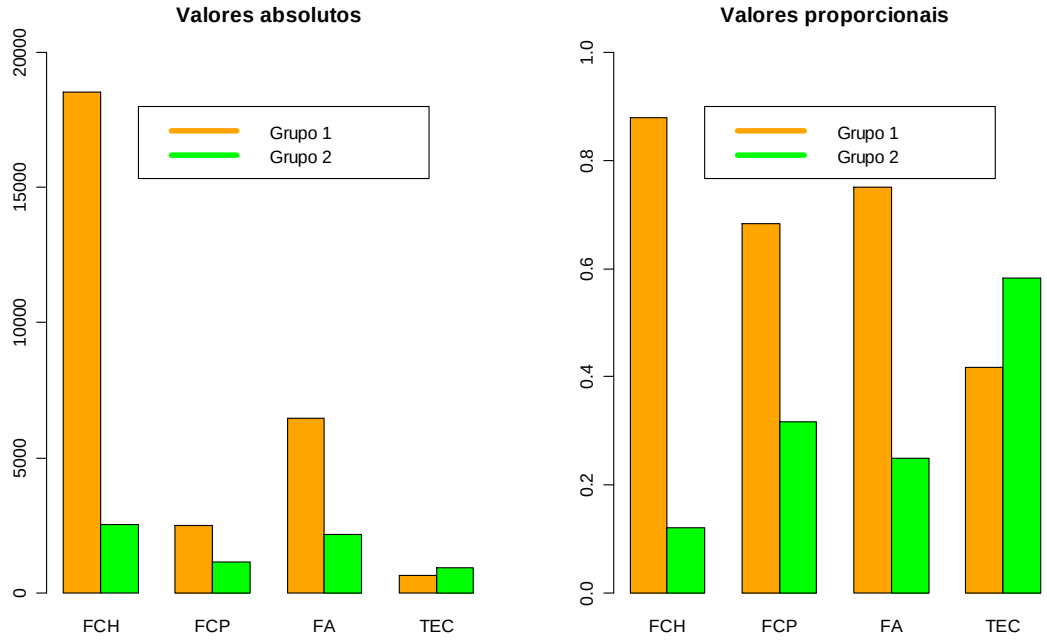

Gráfico 9 – Freqüências e freqüências relativas dos Clusters para variáveis bióticas - Verão.

Analisando o gráfico 9 percebe-se claramente que o grupo 2 é formado especificamente pela espécie Tecamebas. Analogamente, temos as espécies FCH, FCP e FA formando o grupo 1.

Na seqüência construiremos as tabelas de freqüências para os fatores bióticos no Inverno.

Tabela 9 – Freqüência das observações dos fatores bióticos no Inverno.

	FCH	FCP	Allg	FA	TEC
Cluster 1	11056	401	12	9086	466
Cluster 2	2197	604	9	2381	1030

Da mesma forma observamos a ausência significativa de allg, por isso o desconsideramos também para análise dos fatores através da tabela de freqüências. Passamos desta forma a considerar a tabela de 5 variáveis para apenas 4 variáveis, a qual é representada pela tabela 10.

Tabela 10 – Freqüência das observações dos fatores bióticos no Inverno.

	FCH	FCP	FA	TEC
Cluster 1	11056	401	9086	466
Cluster 2	2197	604	2381	1030

Uma vez em posse desta tabela calculamos a freqüência relativa em relação ao total de fatores bióticos, mostrado na tabela 12.

Tabela 11 – Total de observações das variáveis bióticas no Inverno.

FCH	FCP	FA	TEC
13253	1005	11467	1496

Tabela 12 – Freqüência relativa das variáveis bióticas para o Inverno

	FCH	FCP	FA	TEC
Cluster 1	83,4%	39,9%	79,2%	31,1%
Cluster 2	16,6%	60,1%	20,8%	68,9%

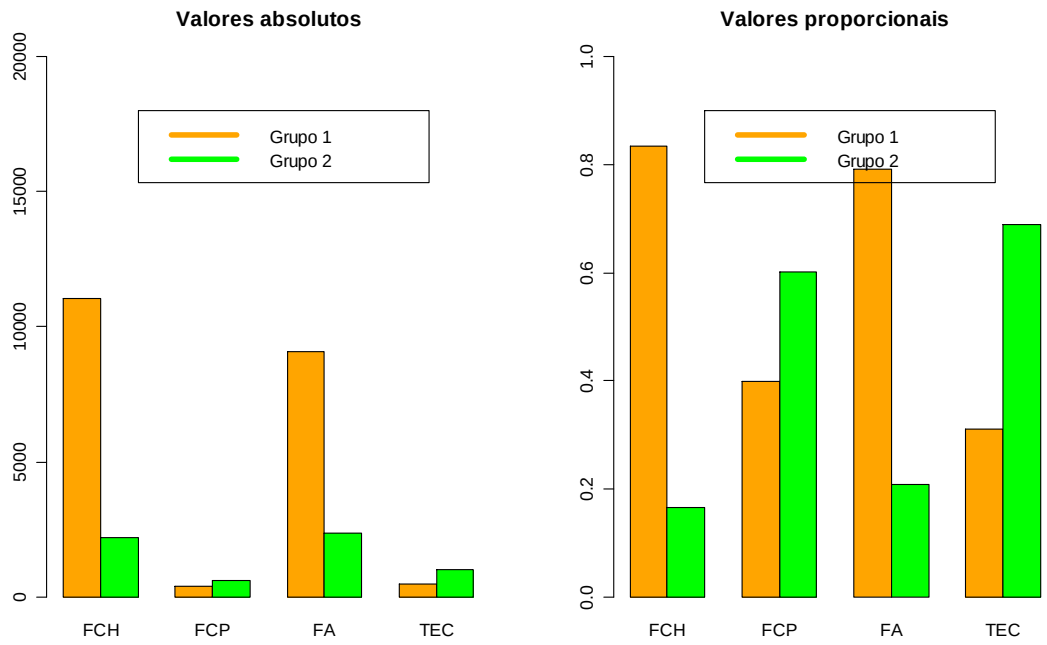

Gráfico 10 –Freqüências e freqüências relativas dos Clusters para variáveis bióticas - Inverno.

Da mesma forma analisando o gráfico 10 percebe-se claramente que o grupo 2 é formado predominantemente pelas espécies FCP e Tecamebas. Analogamente, temos as espécies FCH e FA formando o grupo 1.

Conclusão

Foram identificadas que ao invés de quatro sub-regiões, para cada estação climática apresentam-se duas sub-regiões mostradas nos gráficos 2 e 6. Pode-se notar uma grande semelhança entre estas novas regiões no inverno e no verão.

Identificamos que as Tecamebas constituem o indicador biológico do grupo 2 e os FCH (Foraminíferos Calcários Hialinos), FCP (Foraminíferos Calcários Porcelanáceos) e FA (Foraminíferos Aglutinantes) os indicadores biológicos do grupo 1 no verão.

No inverno as Tecamebas e FCP (Foraminíferos Calcários Porcelanáceos) identificam o grupo 2 e os FCH (Foraminíferos Calcários Hialinos) e FA (Foraminíferos Aglutinantes) o grupo 1.

Anexos

Análise dos dados – Fatores abióticos - Inverno

```
dados=read.table('c:\\temp\\Inverno.txt',sep=';',h=T)

names(dados)=c("Estacao","Local","FCH","FCP","allg","FA","TEC","Profun
didade","CaCO3","Gran","AMG","AG","AM","AF","AMF","SG","SF","Areia","S
ilte","Argila","Lama","Phi_medio","Phi_medio2","G.selecao","G.selecao2
","Assim.","Curtose","Shepard","C.org.","N_tot.","S_tot.","Rz_CN","Rz_
CS")

require(cluster)

#png(file='silhouette.png',width=1260)
par(mfrow=c(1,3))

# Distância euclidean

asw=numeric(10)

dados.dist=dist(dados[,c(8:22,24,26:27,29:33)],method='euclidean')
     ## Note that "k=1" won't work!
     for (k in 2:10)
       asw[k]=pam(dados.dist, k) $ silinfo $ avg.width
     k.best=which.max(asw)
     cat("Número ótimo de agrupamentos - silhouette:", k.best, "\n")

     plot(1:10, asw, type= "h", main = "Distância Euclideana",
          xlab= "No. de agrupamentos", ylab = "Média do valor de
silhouette")
     axis(1, k.best, paste("melhor",k.best,sep="\n"), col = "red",
col.axis = "red")
asw

# Distância manhattan
     ##
     asw=numeric(10)

dados.dist=dist(dados[,c(8:22,24,26:27,29:33)],method='manhattan')
     ## Note that "k=1" won't work!
     for (k in 2:10)
       asw[k]=pam(dados.dist, k) $ silinfo $ avg.width
     k.best=which.max(asw)
     cat("Número ótimo de agrupamentos - silhouette:", k.best, "\n")

     plot(1:10, asw, type= "h", main = "Distância Manhattan",
          xlab= "No. de agrupamentos", ylab = "Média do valor de
silhouette")
     axis(1, k.best, paste("melhor",k.best,sep="\n"), col = "red",
col.axis = "red")
asw

# Distância minkowski
     ##
     asw=numeric(10)

dados.dist=dist(dados[,c(8:22,24,26:27,29:33)],method='minkowski')
```

```
     ## Note that "k=1" won't work!
     for (k in 2:10)
       asw[k]=pam(dados.dist, k) $ silinfo $ avg.width
     k.best=which.max(asw)
     cat("Número ótimo de agrupamentos - silhouette:", k.best, "\n")

     plot(1:10, asw, type= "h", main = "Distância Minkowski",
          xlab= "No. de agrupamentos", ylab = "Média do valor de
silhouette")
     axis(1, k.best, paste("melhor",k.best,sep="\n"), col = "red",
col.axis = "red")
asw

dev.off()

# Resultado

cluster1m=pam(dados[,c(8:22,24,26:27,29:33)],3,metric='minkowski')
#png(file='cluster.png',width=1260)
par(mfrow=c(1,2))
clusplot(cluster1m,main='')
si.cluster1m=silhouette(cluster1m)
plot(si.cluster1m,col=c("blue","purple","violet"),main='')
#dev.off()
#
cluster1m$clustering
table(dados$Local,cluster1m$clustering)
table(cluster1m$clustering,dados[,1])
```

Análise de Componentes Principais – Dados no Inverno

```
cp.inv <- princomp(dados[,c(8:22,24,26:27,29:33)])
princomp(dados[,c(8:22,24,26:27,29:33)], cor = TRUE)
summary(cp.inv <- princomp(dados[,c(8:22,24,26:27,29:33)], cor =
TRUE))
loadings(cp.inv)
plot(cp.inv) # shows a screeplot.
biplot(cp.inv)
```

Tabela de freqüências – Variáveis bióticas – Inverno

```
tabela=xtabs(cbind(dados$FCH,dados$FCP,dados$allg,dados$FA,dados$TEC)~
cluster1m$clustering)
tabela

tabela1=xtabs(cbind(dados$FCH,dados$FCP,dados$FA,dados$TEC)~cluster1m$
clustering)
tabela1

# Totais
tabela2=apply(xtabs(cbind(dados$FCH,dados$FCP,dados$FA,dados$TEC)~clus
ter1m$clustering),2,sum)
tabela2
#
ttabela2=rbind(tabela2,tabela2)
ttabela2

## Tabela ajustada
```

```
tabela3 <-  rbind(tabela1[1,]+tabela1[3,],tabela1[2,])
tabela3 <- as.matrix(tabela3)
ttabs=tabela3/ttabela2
ttabs <- as.matrix(ttabs)
ttabs

par(mfrow=c(1,2))
barplot(tabela3,beside=T,names=c('FCH','FCP','FA','TEC'),ylim=c(0,2000
0),col=c('orange','green'),main='Valores absolutos')
legend(3,18000,legend=c('Grupo 1','Grupo
2'),lwd=5,col=c('orange','green'))
#
barplot(ttabs,beside=T,names=c('FCH','FCP','FA','TEC'),ylim=c(0,1),col
=c('orange','green'),main='Valores proporcionais')
legend(3,0.9,legend=c('Grupo 1','Grupo
2'),lwd=5,col=c('orange','green'))
```

 # Análise dos dados – Fatores abióticos - Verão

```
dados=read.csv('c:\\temp\\Verao.txt',h=T,sep=';')

names(dados)=c("Estacao","Local","FCH","FCP","allg","FA","TEC","Profun
didade","CaCO3","Gran","AMG","AG","AM","AF","AMF","SG","SF","Areia","S
ilte","Argila","Lama","Phi_medio","Phi_medio2","G.selecao","G.selecao2
","Assim.","Curtose","Shepard","C.org.","N_tot.","S_tot.","Rz_CN","Rz_
CS")

require(cluster)

#png(file='silhouette.png',width=1260)
par(mfrow=c(1,3))

# Distância euclidean
     ##
     asw=numeric(10)

dados.dist=dist(dados[,c(8:22,24,26:27,29:33)],method='euclidean')
     ## Note that "k=1" won't work!
     for (k in 2:10)
       asw[k]=pam(dados.dist, k) $ silinfo $ avg.width
     k.best=which.max(asw)
     cat("Número ótimo de agrupamentos - silhouette:", k.best, "\n")

     plot(1:10, asw, type= "h", main = "Distância Euclideana",
          xlab= "No. de agrupamentos", ylab = "Média do valor de
silhouette")
     axis(1, k.best, paste("melhor",k.best,sep="\n"), col = "red",
col.axis = "red")
asw

     asw=numeric(10)

dados.dist=dist(dados[,c(8:22,24,26:27,29:33)],method='manhattan')
     ## Note that "k=1" won't work!
     for (k in 2:10)
       asw[k]=pam(dados.dist, k) $ silinfo $ avg.width
     k.best=which.max(asw)
     cat("Número ótimo de agrupamentos - silhouette:", k.best, "\n")

     plot(1:10, asw, type= "h", main = "Distância Manhattan",
          xlab= "No. de agrupamentos", ylab = "Média do valor de
silhouette")
     axis(1, k.best, paste("melhor",k.best,sep="\n"), col = "red",
col.axis = "red")
asw

# Distância minkowski
     ##
     asw=numeric(10)

dados.dist=dist(dados[,c(8:22,24,26:27,29:33)],method='minkowski')
     ## Note that "k=1" won't work!
     for (k in 2:10)
       asw[k]=pam(dados.dist, k) $ silinfo $ avg.width
     k.best=which.max(asw)
     cat("Número ótimo de agrupamentos - silhouette:", k.best, "\n")

     plot(1:10, asw, type= "h", main = "Distância Minkowski",
```

```
          xlab= "No. de agrupamentos", ylab = "Média do valor de
silhouette")
     axis(1, k.best, paste("melhor",k.best,sep="\n"), col = "red",
col.axis = "red")
asw

dev.off()

# Resultado

cluster1m=pam(dados[,c(8:22,24,26:27,29:33)],2,metric='minkowski')
#png(file='cluster.png',width=1260)
par(mfrow=c(1,2))
clusplot(cluster1m,main='')
si.cluster1m=silhouette(cluster1m)
plot(si.cluster1m,col=c("blue","purple","violet"),main='')
#dev.off()

cluster1m$clustering
table(dados$Local,cluster1m$clustering)
table(cluster1m$clustering,dados[,1])
```

Análise de Componentes Principais – Dados no Verão

```
cp.ver <- princomp(dados[,c(8:22,24,26:27,29:33)])
princomp(dados[,c(8:22,24,26:27,29:33)], cor = TRUE)
summary(cp.ver <- princomp(dados[,c(8:22,24,26:27,29:33)], cor =
TRUE))
loadings(cp.ver)
plot(cp.ver) # shows a screeplot.
biplot(cp.ver)
```

Tabela de frequencias – Variáveis bióticas - Verão

```
# Tabela de freqüências

tabela=xtabs(cbind(dados$FCH,dados$FCP,dados$allg,dados$FA,dados$TEC)~
cluster1m$clustering)
tabela

# Observamos a ausência de allg, por isso o desconsideramos.

tabela1=xtabs(cbind(dados$FCH,dados$FCP,dados$FA,dados$TEC)~cluster1m$
clustering)
tabela1

# Totais

tabela2=apply(xtabs(cbind(dados$FCH,dados$FCP,dados$FA,dados$TEC)~clus
ter1m$clustering),2,sum)
tabela2

ttabela2=rbind(tabela2,tabela2)
ttabs=tabela1/ttabela2

par(mfrow=c(1,2))
barplot(tabela1,beside=T,names=c('FCH','FCP','FA','TEC'),ylim=c(0,2000
0),col=c('orange','green'),main='Valores absolutos')
legend(3,18000,legend=c('Grupo 1','Grupo
2'),lwd=5,col=c('orange','green'))
```

```
barplot(ttabs,beside=T,names=c('FCH','FCP','FA','TEC'),ylim=c(0,1),col
=c('orange','green'),main='Valores proporcionais')
legend(3,0.9,legend=c('Grupo 1','Grupo
2'),lwd=5,col=c('orange','green'))
```

Referências

DA SILVA, J. L. P., OLIVEIRA, M. F. **"Nova Proposta para Classificação das Agências de Correio"**, UFPR. (2007).

JAWORSKI, K. S. **"Caracterização do Sistema estuarino-lagunar de Cananéia-Iguape – SP.** USP. (2006).

R Development Core Team. R version 2.5.1: A language and environment for statistical computing. **The R Foundation for Statistical Computing**, Viena, Austria. ISBN 3-900051-07-0. (2007). URL http://www.R-project.org.

MARQUES, J. M. **"The principal components in the reduction of variables in a multiple regression model".** Revista FAE. (2005).

WANGENHEIM, A. von, Prof. Dr. **"Análise de Agrupamentos".** UFSC. (2006).

ROUSSEEUW, P. J. **"Silhouetes: A Graphical and to the interpretation and validation of cluster analysis.** J. Comput Appl. Math., 20, 53-65. (1987).

Manual e Ajuda para o add-in Metrixus - **Funções Quantitativas para Mercado de Capitais** - Élin Duxus - Brasil - 2002 – 2008. www.duxus.com.br

48ª Reunião RBRAS e 10ª SEAGRO, p.37-40, p.59-61 – **Universidade Federal de Lavras** – Departamento de Ciências Exatas – MG – (2003).