Ana Teresa Lopes Lourenço

Licenciada

Estudo da estrutura amostral (variáveis e número de observações) na discriminação do local de origem da ameijoa japónica *(Ruditapes philippinarum)*

Dissertação para obtenção do Grau de Mestre em Matemática e Aplicações – Ramo Actuariado, Estatística e Investigação Operacional

Orientadora: Regina Bispo, Professora Auxiliar, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa

Co-orientadores: Maria Isabel Gomes, Professora Associada, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa

Júri:

Presidente: Prof. Doutor Filipe José Gonçalves Pereira Marques

Vogais: Prof. Doutor Miguel dos Santos Fonseca

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Setembro, 2019

Estudo da estrutura amostral (variáveis e número de observações) na discriminação do local de origem da ameijoa japónica (*Ruditapes philippinarum*)

Copyright © Ana Teresa Lopes Lourenço, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa.

A Faculdade de Ciências e Tecnologia e a Universidade Nova de Lisboa têm o direito, perpétuo e sem limites geográficos, de arquivar e publicar esta dissertação através de exemplares impressos reproduzidos em papel ou de forma digital, ou por qualquer outro meio conhecido ou que venha a ser inventado, e de a divulgar através de repositórios científicos e de admitir a sua cópia e distribuição com objectivos educacionais ou de investigação, não comerciais, desde que seja dado crédito ao autor e editor.

Agradecimentos

Começo por agradecer às professoras Regina Bispo e Maria Isabel Gomes pela orientação e paciência ao longo da dissertação.

Quero também agradecer aos meus amigos e colegas, Joana Mascarenhas e Tiago Ramalho, que pelo apoio ao longo do mestrado dando-me sempre força para alcançar os meus objetivos.

Por fim, quero agradecer aos meus pais por sempre me apoiarem incondicionalmente nas minhas decisões, sem eles nada disto seria impossível. Dedico-lhes todo o meu sucesso.

Este estudo recebeu o apoio financeiro do projeto TraSeafood(POCI- 01-0145-FEDER-029491) financiado pelo FEDER através do COMPETE2020-Programa operacional competitividade e internacionalização(POCI), e por fundos nacionais (OE) através da FCT/MCTES. Foi ainda parcialmente financiado pela Fundação para a Ciência e a Tecnologia através do projeto UID/MAT/00297/2019 (Centro de Matemática e Aplicações).

Resumo

A determinação da origem geográfica de produtos alimentares de origem marinha é um passo fundamental para controlar a sua qualidade e salvaguardar o interesse dos consumidores. Dada a importância económica da espécie alvo deste estudo, a ameijoa japónica *(Ruditapes philippinarum)*, a rastreabilidade da sua origem é particularmente importante num contexto nacional. Para este estudo foram recolhidas amostras com dimensão 30, de três regiões distintas (Ria de Aveiro, Estuário Tejo e Ria de Vigo). Os indivíduos amostrados foram analisados utilizando ferramentas bioquímicas e geoquímicas (perfil de ácidos gordos do musculo adutor e assinatura elementar da concha) perfazendo um total de 44 variáveis (18 elementos e 26 ácidos gordos). Com base nesta informação pretende-se determinar o número mínimo de variáveis e observações (diminuindo o esforço analítico e a pressão ecológica) que ainda assim garantam a discriminação do local de origem. Na diminuição/seleção do número de variáveis usou-se como ferramenta base a ACP (Análise de componentes principais ou PCA, Principal component analysis). A ACP, é um método multivariado, frequentemente utilizado, de fácil implementação que genericamente permite a redução da dimensão dos dados e, por isso, possibilita uma melhor interpretação e visualização gráfica a baixa dimensionalidade.

O estudo do efeito da diminuição do número de observações foi feito recorrendo a simulação. Como indicador do grau de separabilidade dos grupos (locais de origem) foi usado o índice Silhouette, que permite quantificar o quão bem alocada a um certo grupo se encontra uma certa observação por comparação à alocação aos restantes grupos. O estudo conclui com a apresentação do menor conjunto de variáveis e menor número de observações que ainda permitem garantir a diferenciação dos locais de origem da espécie alvo.

Palavras-chave: *Ruditapes philippinarum*, PCA, Silhouette, Bases de dados

Abstract

Determining the geographic origin of sea life food products is a fundamental step for quality control safeguarding the consumer's interest. Giving the economical importance of the target species of this study, japonica clam *(Ruditapes philippinarum)*, tracking its origin is particularly important regarding a national context. For this study were collected samples, which dimension was 30, from three distinct regions (Ria de Aveiro, Estuário Tejo e Ria de Vigo). The sampled individuals were analyzed using biochemical and geochemical tools (fat adductor muscle acids profile and elemental shell signature) completing a total of 44 variables (18 elements and 26 fat acids). Based on this information it's intended to determine the minimal number of variables and observations (reducing the analytical effort and and the ecological pressure) that guarantee the discrimination of the place of origin. In the reduction/selection of the number of variables PCA (Principal component analysis) was used as base tool. PCA is a multivariate method, frequently used, implemented with ease which genetically allows the reduction of the data size and, therefore makes possible a better graphic interpretation and visualization on a lower dimension level.

The study of the reduction of the number of observations effect was achieved recurring to simulation. As a indicator of the level of separability of the groups (place of origin) the Silhouette index was used, this allowed to quantify on how well allocated to a certain group a observation is comparing to the allocation of the remaining groups. The study concludes with a presentation of the smallest set of variables and observations which still allow assurance that we can differentiate place of origin of the target species.

Keywords: *Ruditapes philippinarum*, PCA, Silhouette, Data base

Índice

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Motivação

Estima-se que a população global aumente até os nove mil milhões, aproximadamente, ainda a meio deste século (Godfray, 2010). Consequentemente, esse aumento irá provocar um acréscimo na eficácia dos setores agrícolas e pesca, visto serem os maiores setores no fornecimento de proteína animal. Porém, principalmente no setor da pesca, a sobre-exploração está a levar à escassez dos diversos recursos, recorrendo-se cada vez mais a alternativas, como a aquacultura. Não obstante, apenas em Portugal, só no setor da pesca, a captura de bivalves, aumentou cerca de 1%, entre 2016 e 2017 (Estatísticas da Pesca 2017).

Na captura e produção de moluscos e crustáceos, a ameijoa é a espécie mais abundante (Estatísticas da Pesca 2017). Isto é explicado por vários estudos apontarem os bivalves como tendo benefícios para a saúde humana, não só por serem ricos em vitaminas e minerais(Karnjanapratum *et al.*, 2013), como por estarem associados à diminuição da incidência de doenças oncológicas e cardiovasculares, entre outras (Montilivi, 2019). No entanto, estes moluscos são organismos filtradores e, portanto, acumulam microrganismos patogénicos que possam estar presente no meio ambiente (Ricardo et al., 2017). Como muitas vezes estes animais não são bem depurados e são comidos de forma crua, estão associados a diversas ameaças à saúde do Homem.

Perante as características associadas a estes indivíduos, a rastreabilidade dos bivalves é bastante importante. Este conceito surgiu devido à necessidade de controlo de qualidade: saber de onde vem e para onde vai o produto que está a

ser tido em conta.

Neste seguimento, torna-se necessário desenvolver metodologias que permitam às autoridades competentes a identificação, tão isenta de erro quanto possível, da origem destes indivíduos. É neste contexto que surge este estudo, para o que foi considerada apenas uma espécie de bivalves, a *Ruditapes philippinarum*.

1.2 Objetivo

O principal objectivo deste estudo é diminuir a dimensionalidade da amostra (quer em termos de número de variáveis, quer em termos de número de observações) de forma a garantir a discriminação da origem (*cluster*) dos indivíduos.

Em particular pretende-se:

1. **Determinar o número mínimo de variáveis**

   A recolha indiscriminada de informação envolve maior esforço analítico e nem sempre um maior número de variáveis implica uma melhor classificação dos indivíduos. Portanto, pretende-se determinar o número mínimo de variáveis que permita a discriminação das populações correspondentes às diferentes origens dos indivíduos.

2. **Determinar o tamanho mínimo da amostra**:

   Neste âmbito o objectivo é determinar o tamanho da amostra suficiente que ainda diferencie estatisticamente as populações, e assim diminuir o esforço de amostragem e o possível impacto ambiental na espécie.

1.3 Estrutura

Esta dissertação está organizada em seis capítulos. No primeiro capítulo é feita uma introdução ao tema da dissertação. O segundo capítulo diz respeito à revisão de literatura. Aborda as diferentes técnicas utilizadas para dar resposta aos diferentes objetivos descritos.

O terceiro capítulo diz respeito à aplicação das metodologias e está dividido em três partes. A primeira é referente ao dataset, onde foram descritos os dados em estudo. A segunda parte diz respeito aos métodos usados na redução do número de variáveis (p), em particular à analise de componentes principais (da literatura anglo-saxónica, *Principal Analysis Components*, PCA) em conjunto com o índice Silhouette e o coeficiente RM. Por último, a redução do tamanho da amostra (n) onde são analisados vários critério incluindo o índice Silhouette, o

coeficiente RM, a distância Euclideana e a soma dos quadrados intra-*clusters* (da literatura anglo-saxónica, *Within Sum of Squares*, WSS).

No quarto e quinto capítulo são apresentados e discutidos os resultados deste estudo. E, no sexto e último capítulo, são apresentados as conclusões finais e possíveis trabalhos futuros.

Capítulo 2

Revisão da literatura

2.1 Análise de componentes principais

A análise de componentes principais (doravante referenciada por PCA) é, provavelmente, das técnicas mais antigas e mais conhecida em análise multivariada. Esta técnica foi inicialmente introduzida por Pearson (Pearson, 1901), "finding lines and planes of closest fit to systems of points in space" porém, foi devido ao trabalho de Hottelling (Hotelling, 1933) que se começou a utilizar esta técnica com o intuito de analisar a estrutura de covariância (ou correlação) entre variáveis graças ao avanço tecnológico, que impulsionou o avanço de capacidade computacional. Esta técnica começou a ser utilizada em várias áreas de forma abrangente, como a neurociência, engenharia, geologia, entre outras.

2.1.1 Definição

A análise de componentes principais tem como propósito a explicação da variância/covariância de um conjunto de variáveis correlacionadas a partir da combinação linear destas. Os principais objetivos são a redução da dimensionalidade, retendo o máximo de variância possível, e a representação a baixa dimensionalidade. Tal é conseguido a parir da criação de novas variáveis (m), as componentes principais (PC), forçosamente ortogonais, que não são mais que uma combinação linear das variáveis originais, onde m componentes ($m < p$) são estimadas de forma a reter o máximo de variância possível do grupo original de variáveis.

Seja X, definido por $X_{nxp} = [X_{ij}](i = 1,...,n; j = 1,...,p)$, o conjunto de dados

originais com p variáveis de dimensão n. De forma a que seja possível representar graficamente os dados e, eventualmente, observar relações que não eram antes observáveis, a análise de componentes principais é uma solução.

O primeiro passo consiste em encontrar uma função linear das p variáveis que maximize a variância explicada. Sendo t_1 um vetor de p constantes $t_{11},...,t_{1p}$, e t' a transposta de t, tem-se

$$t_1'X = t_{11}x_1 + t_{12}x_2 + ... + t_{1p}x_p \tag{2.1}$$

Esta será a primeira componente principal (PC1). O passo seguinte tem como objectivo encontrar a próxima componente principal, neste caso a segunda (PC2), que será a combinação linear das p variáveis originais $t_2'X$, não correlacionada com $t_1'X$, que apresente o máximo de variância. E assim sucessivamente, de maneira a que a combinação linear $t_k'X$ $(k \geq 2)$ não esteja correlacionada com as definidas anteriormente e que apresente o máximo de variância possível dos dados originais.

Genericamente y_{ik}=$t_k'x_i$ é a k-ésima componente principal, $k = 1,...,p$ sendo y_{ik} o *score* da observação i $(i = 1,...,n)$. No total, existe p componentes principais e n *scores*, no entanto, em regra o que se pretende das p variáveis originais é ter um conjunto de m componentes $(m < p)$, que explique o máximo da variância.

Vários métodos podem ser utilizados para definir as componentes principais, isto é, estimar as constantes dos vetores t_k. Entre eles existe a decomposição em valores próprios, o *Non-linear iterative partial least-squares* (NIPALS), o *Singular Value Decomposition* (SVD), e a técnica dos multiplicadores de Lagrange (Jolliffe, 1986), sendo esta a descrita neste estudo.

Seja a matriz de covariâncias Σ de X, a k-ésima componente principal definida por $y_{ik} = t_k'x_i$ e t_k' os vetores, com norma unitária, estimados pelos k vetores próprios de Σ (mais frequentemente denominados por *loading vetor*).Ao restringir $t_k't_k$=1, ou seja, a soma dos quadrados dos elementos de t_k igual a 1, então a variância de y_k é igual a λ_k (var(y_k)=λ_k), isto é, pelos k valores próprios da matriz Σ (Jolliffe, 1986).

Considere-se a primeira componente principal, $t_1'X$, tal que a var($t_1'X$)=$t_1'X\Sigma t_1X$, e a restrição $t_k't_k$=1. Maximizar $t_1'X\Sigma t_1X$ (a var($t_1'X$)), com essa restrição, pode ser feito recorrendo à técnica de multiplicadores de Lagrange. Derivando

$$t_1'X\Sigma t_1X - \lambda_1(t_1't_1 - 1) \tag{2.2}$$

em ordem a t_1 e igualando a zero, fica-se com

$$\Sigma t_1 - \lambda_1 t_1 = (\Sigma - \lambda_1 I_p) t_1 = 0 \tag{2.3}$$

onde λ_1 e I_p representam, respectivamente, um multiplicador de Lagrange e a matriz identidade (pxp). Da relação anterior, λ_1 representa o primeiro valor próprio e t_1 ao primeiro vetor próprio de Σ.

A primeira componente principal explica a maior parte da variância dos dados originais, em relação às restantes componentes, sendo

$$t_1' \Sigma t_1 = t_1' \lambda_1 t_1 = \lambda_1 t_1' t_1 = \lambda_1 \tag{2.4}$$

A var($t_1'X$)=$t_1'\Sigma t_1$, como acima demonstrado, é igual a λ_1, maximizando a variância da PC1 (primeira componente principal).

Para as restantes k-ésimas componentes principais de X, o mesmo se verifica, sendo λ_k o valor próprio de Σ, atendendo às restrições $t_k' t_k = 1$ (k=1,...,p) e $t_{k'}, \Sigma t_k = 0$ ($k' \neq k$) e t_k é o vetor próprio associado.

Resumidamente as componentes principais são então definidas por (Johnson e Wichern, 2007):

- **Primeira componente**: é a combinação linear $t_1'X$, onde X são os dados originais e t_1' é o vetor que maximiza var($t_1'X$), sujeita à restrição $t_1' t_1 = 1$.
- **k-ésima componente**: é a combinação linear $t_k'X$ sendo t_k' o vetor que maximiza a var($t_k'X$) sujeita à restrição $t_k' t_k = 1$ e a $t_{k'}, \Sigma t_k = 0$ ($k' \neq k$).

2.1.2 Interpretação geométrica

Algebricamente, as componentes principais são definidas pelas combinações lineares das variáveis ($x_1, x_2, ..., x_p$). Geometricamente, essas componentes representam um novo sistema de coordenadas obtido pela rotação do sistema original. Os novos eixos representam a direção que apresenta a maior variância possível (Johnson e Wichern, 2007).

Para melhor exemplificar a representação gráfica, vai-se considerar dimensão p=3, (x_1,x_2,x_3) :

O primeiro passo consiste em mover os objetos (representados pelos pontos azuis, Figura 2.1) para o centro das coordenadas. Este processo é denominado de *mean-centering*.

Depois de normalizados os dados, é traçada uma recta ao longo dos pontos de forma a explicar o máximo de variância, isto é, uma recta que explica as observações com o menor erro residual possível. Esta recta é definida pelo primeiro vetor

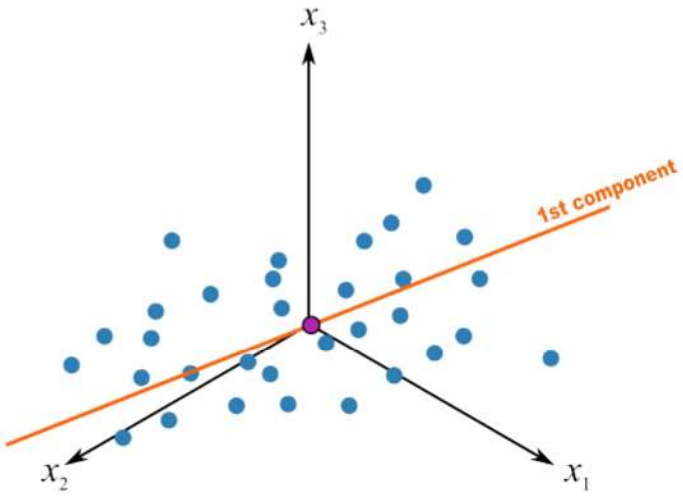


Figura 2.1: Representação gráfica dos dados X=[x_1,x_2,x_3] (pontos azuis) e a representação da primeira componente principal (recta laranja) (Kevin Dunn, 2019)

próprio pela matriz de covariância, um vetor de pesos associados a cada variável. Quando marcada a primeira recta, é possível marcar cada objecto ao longo da linha.

Considerando o objecto i, este é marcado na recta delineada a partir da projeção de 90 graus de cada ponto em relação à respectiva recta. Ao longo da recta, a distância entre a origem e o ponto marcado representa o *score* associado o objeto i.

Obtêm-se, assim, a primeira componente principal (PC1) constituída por um vetor que de coeficientes que maximiza a variância.

Fixada a primeira componente principal, adiciona-se a segunda (PC2) componente principal, perpendicular à primeira (PC1) (Figura 2.2). Começa também na origem e move-se ao longo da primeira até encontrar a direção que explique a maior variância possível. Agindo da mesma forma, os *scores* são calculados a partir da projeção de cada observação à recta que representa a segunda componente. Portanto, tal como definido para a primeira componente, a PC2 tem também associado um vetor de "direção"que melhor se ajusta aos dados, explicando grande parte da variância, e um vetor que representam a distância de cada ponto projetado perpendicularmente em relação à origem.

Este processo é repetido até o número de componentes principais corresponder com o número de variáveis em estudo, sendo que cada componente é ortogonal em relação às restantes. O conjunto de p componentes é o que melhor representa o conjunto de dados em estudo, no entanto, pode-se reduzir ao escolher um número de componentes que explique uma grande percentagem da variância.

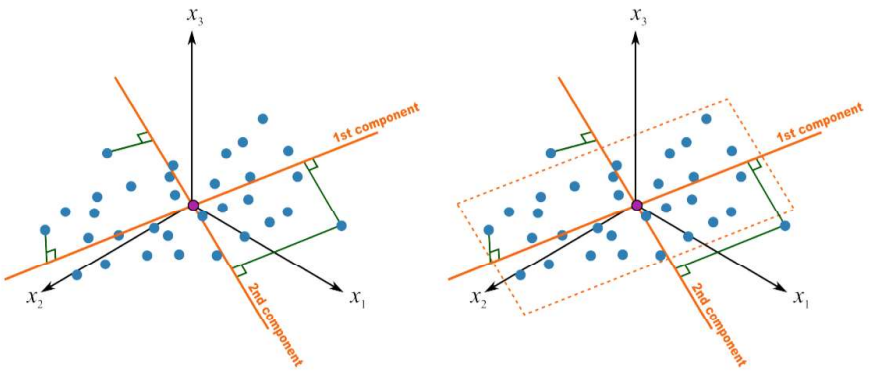


Figura 2.2: Representação geométrica da análise de componentes principais com o novo sistema de coordenadas (PC1 e PC2, rectas cor de laranja) (Kevin Dunn, 2019)

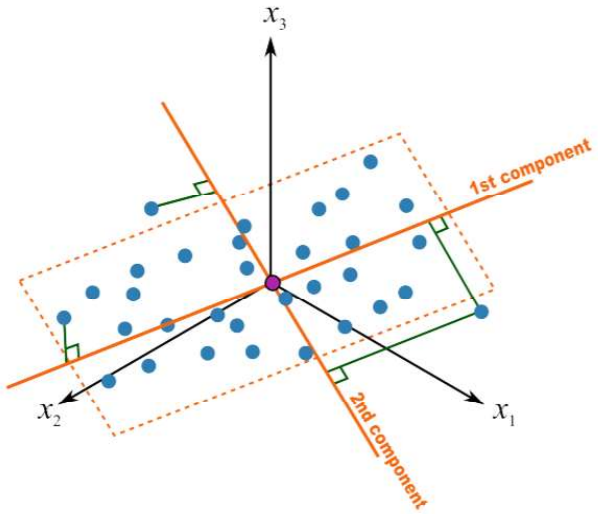


Figura 2.3: Representação geométrica da análise de componentes principais com o novo sistema de coordenadas (PC1 e PC2, rectas cor de laranja) (Kevin Dunn, 2019)

Considere-se a figura 2.3, onde estão representadas graficamente as três variáveis e as respectivas componentes principais (PC1 e PC2). Neste sistema, a primeira componente principal (PC1) tem uma orientação aproximada ao eixo de x_2. Portanto o coeficiente associado a esta variável será superior aos coeficientes associados às duas restantes variáveis (x_1 e x_3) na definição de PC1. Na verdade, variáveis que pouco contribuam para a direção da componente têm um peso próximo de zero.

Em suma, cada variável tem um peso associado que será menor ou maior conforme a sua importância na definição da componente principal. Como cada componente é definida de forma a explicar a maior percentagem de variância, assume-se que a variável associado ao menor peso tem uma menor importância

na explicação dos dados originais.

2.2 Análise de clusters

A análise de *clustering* permite-nos organizar os dados em grupos/*clusters* homogéneos, dado um conjunto de variáveis (Rubanov *et al.*, 2019). O método de *clustering* agrupa os objetos de acordo com a informação disponível, de modo a que os indivíduos do mesmo grupo sejam tão semelhantes quanto possível e sempre mais semelhantes aos elementos do mesmo grupo do que aos elementos dos restantes grupos.

Os algoritmos associados às técnicas de *clustering* baseiam-se na determinação da semelhança dos objetos com base no cálculo das distâncias entre pontos. Intuitivamente, quanto maior a distância estimada entre os pontos, menor é a semelhança entre as observações, e vice-versa (Everitt *et al.*, 2011). Para o cálculo de distâncias entre os objetos pode-se recorrer à distância Euclideana, Manhattan, Minkowski, entre outras. No entanto, a escolha do método de cálculo da distância influencia os resultados obtidos na análise de *clusters*. Para este estudo considerou-se a distância Euclideana, entre dois pontos, $\mathbf{x}=(x_1,...,x_n)$ e $\mathbf{y}=(y_1,...,y_n)$, de dimensão n, definida por:

$$d(\mathbf{x},\mathbf{y}) = \sqrt{(x_1-y_1)^2+...+(x_n-y_n)^2} = \sqrt{\sum_{j=1}^{n}(x_j-y_j)^2} \tag{2.5}$$

Existem diferentes métodos para realizar uma análise de *clusters*. Estes podem ser divididos em métodos hierárquicos e não hierárquicos.

1. **Métodos hierárquicos**: Os métodos hierárquicos incluem técnicas ou de aglomeração ou de divisão. Os métodos aglomerativos começam com um um objeto por grupo. Os objetos mais semelhantes são agrupados no mesmo grupo, sucessivamente, até todos os *clusters* se encontrarem num único *cluster*. Os métodos de divisão trabalham de forma oposta. Começa com um só grupo com todos os objetos, sucessivamente divide-os em grupos a partir da análise de semelhança entre cada um, ou seja, escolhe-se o objeto mais distante e coloca-se num grupo distinto. De seguida, avaliam-se os restantes objetos, escolhendo-se sempre a maior distância, até cada grupo ser constituído por um só objeto. Ambos os resultados obtidos de forma aglomerativa ou divisiva podem ser representados por um dendrograma (Johnson e Wichern, 2007).

2. **Métodos não hierárquicos**: Os métodos não hierárquicos são caracterizados pela necessidade de definir, inicialmente, o número de *clusters* (baseado no conhecimento inicial dos dados em estudo) e, também, pela sua flexibilidade, uma vez que os objetos podem trocar de grupo ao longo do processo, ao contrário do que acontece nos métodos hierárquicos (Johnson e Wichern, 2007).

Neste estudo, em particular, recorreu-se ao método *k-means* (Macqueen, 1967), um método de *clustering* não hierárquico.

2.2.1 K-means

O método de *clustering k-means* foi primeiramente sugerido por MacQueen (1967) como um algoritmo que atribui cada objeto, $\mathbf{x'}_i = x_1,...,x_p$, ao centróide $\bar{\mathbf{x}}' = \bar{x_1},...,\bar{x_p}$, mais próximo (ponto médio). É necessário definir inicialmente o número de *clusters* que serão gerados no final da execução do algoritmo. O algoritmo começa por escolher aleatoriamente k objetos para servirem de centro dos *clusters*. Depois, cada um dos objetos é atribuído ao *cluster* referente ao centróide mais próximo do objeto. Para avaliar a proximidade é, usualmente, utilizada a distância Euclideana, anteriormente definida, entre o objeto e o centróide. Depois dos objetos serem atribuídos a cada *cluster*, o centróide é recalculado e o processo volta-se a repetir-se, mudando, se necessário, os objetos de cada *cluster* para o grupo cujo centróide está mais próximo. Este processo iterativo é repetido até os *clusters* não se alterarem mais em cada iteração, isto é, até os grupos definidos na iteração seguinte sejam os mesmos obtidos na iteração anterior.

Como *k-means* calcula a distância entre os pontos e os centróides, o objetivo é minimizar essa distância. Como medida de homogeneidade dos grupos usa-se frequentemente a soma dos quadrados intra-*clusters* (WSS).

$$WSS(C_k) = \sum_{x_i \epsilon C_k} (x_i - \mu_k)^2 (i = 1,...,n) \tag{2.6}$$

Onde x_i é referente ao objeto pertencente ao *cluster* C_k e μ_k é o valor médio associado ao *cluster* C_k.

O algoritmo k-means (Fig. 2.3) pode ser resumido da seguinte forma (Johnson e Wichern, 2007):

1. Fixar o número de *clusters* (k) a ser criados (depende da análise do utilizador);

2. Selecionar, de forma aleatória, os objetos iniciais a partir dos quais são determinados os centróides iniciais;

3. Atribuir cada objeto ao grupo cujo centróide é mais próximo, em função da distância euclideana entre o objeto e o centróide;

4. Para cada um dos *k clusters*, atualizar o centróide de cada *cluster* calculando os novos valores médios considerando todos os pontos;

5. Minimizar a soma dos quadrados intra-*clusters* (WSS) iterativamente. Ou seja, iterar os passos 3 e 4 até que a constituição dos *cluster* pare de mudar.

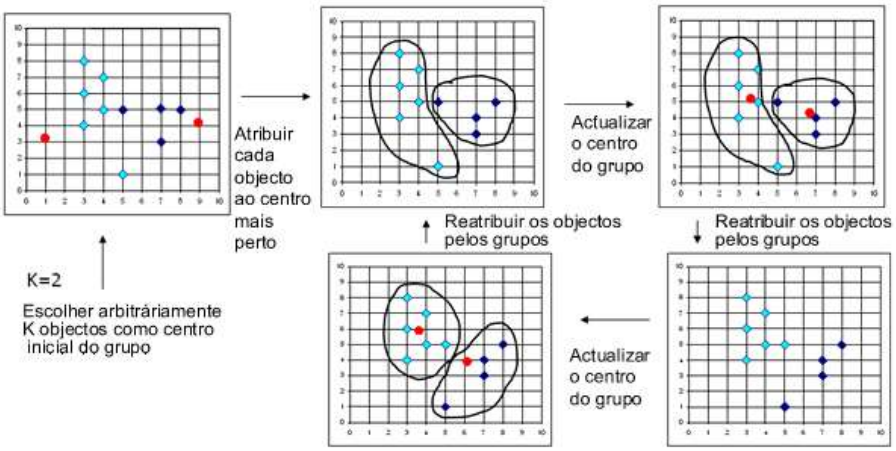


Figura 2.4: Esquema do algoritmo k-means

2.3 Índice Silhouette

No âmbito dos métodos de *clustering*, várias metodologias foram desenvolvidas para avaliar a qualidade da separação dos grupos. Em particular, para este trabalho foi utilizado o índice Silhouette.

Este índice foi primeiramente definido por Peter Rousseeuw em *"Silhouettes: a graphical aid to the interpretation and validation of cluster analysis "* (Pearce *et al.*, 1977) como técnica de interpretação e validação da coerência entre os grupos/*clusters*. Neste trabalho foi proposto uma representação gráfica do índice de silhouette $(s_{(i)}, i = 1, .., n)$, que mede o grau de separabilidade dos grupos. Este índice identifica quais os objetos se encontram "bem colocados" no grupo ou os "menos bem colocados" que estão nas margens entre grupos. Todos os índices silhouette referentes a cada objeto são representados no gráfico, permitindo uma

melhor avaliação da qualidade de agrupamento dos dados. A partir do valor médio dos índices silhouette pode-se avaliar a validade do agrupamento sendo, por isso, frequentemente usado para selecionar o número de *clusters* "ótimo".

De forma a calcular os índices silhouette ($s_{(i)}$) é necessário obter a partição dos dados (a partir da aplicação de técnicas de *clustering*) e calcular a distância entre as amostras.

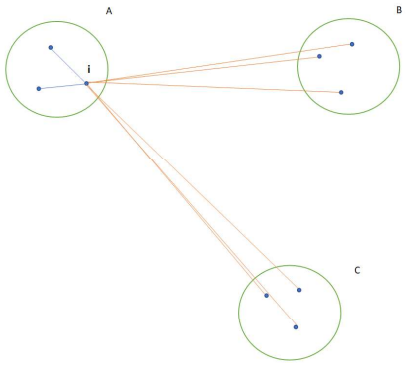


Figura 2.5: Representação de 3 grupos/*clusters* A, B e C e as suas distâncias. Retas laranjas representam a distância de um objeto i aos objetos dos restantes grupos B e C. Retas azuis representam a distância do objeto i aos restantes do mesmo grupo A

Considerando o objeto i num conjunto de dados, denominado A (Figura 2.4) de dimensão n_A, define-se $a_{(i)}$ como sendo a média das distâncias do objeto i aos restantes objetos do mesmo grupo. Note-se que quando $a_{(i)}$ tomar valores pequenos, tal indica que o objeto do mesmo grupo se encontra próximo dos restantes, ou seja, são semelhantes.

Considerando os restantes grupos, calcula-se $b_{(i)}$, que é o minimo das distâncias médias do objeto i aos restantes objetos de diferentes grupos. Primeiro calcula-se a média das distâncias do objeto i para os restantes objetos fora do grupo A. Depois de calculadas, escolhe-se o menor valor dessas médias. Portanto, para este caso, quanto maior o valor de $b_{(i)}$ melhor, indicando que os grupos estão distantes e, consequentemente, os objetos estão bem colocados. Considerando o grupo B como o grupo com menor média de distância até ao objeto i, este é considerado o *neighbouring cluster*, ou *cluster*, vizinho porque é a segunda melhor opção para o objeto i, visto que a seguir ao grupo A, B é o grupo mais próximo de i.

O índice $s_{(i)}$ é obtido conjugando os valores $a_{(i)}$ e $b_{(i)}$ da seguinte forma:

$$s_{(i)} = \begin{cases} 1 - \dfrac{a_{(i)}}{b_{(i)}}, & \text{se } a_{(i)} < b_{(i)} \\ 0, & \text{se } a_{(i)} = b_{(i)} \\ \dfrac{b_{(i)}}{a_{(i)}} - 1, & \text{se } a_{(i)} > b_{(i)} \end{cases} \tag{2.7}$$

Alternativamente,

$$s_{(i)} = \frac{b_{(i)} - a_{(i)}}{max(a_{(i)}, b_{(i)})} \tag{2.8}$$

Pelas expressões acima mencionadas, pode-se observar que

$$-1 \leqslant s_{(i)} \leqslant 1 \tag{2.9}$$

Assim, quando $s_{(i)}$ toma o valor máximo (ou seja, $s_{(i)}$=1), tal significa que as diferenças dentro do grupo onde o objeto i está inserido são muito menores em comparação à diferença entre grupos,$b_{(i)}$. Neste caso, pode-se afirmar que o objeto se encontra bem colocado.

No caso de $s(i)$=0, significa que $a_{(i)}$=$b_{(i)}$. Nestas situações não é claro se o objeto está bem colocado.

Nas piores das hipóteses, $s_{(i)}$=-1. Indica que $a_{(i)}$ é muito superior a $b_{(i)}$, ou seja, o objeto i encontra-se mais distante dos objetos do grupo onde foi posicionado e devia ser movido para o grupo mais próximo.

Em suma, o índice $s_{(i)}$ permite medir quão bem um objeto i se encontra no grupo a que pertence, independentemente do método de *clustering* utilizado. Assim, é possível comparar valores independentemente das técnicas utilizadas. Permite ainda uma interpretação e validação na análise dos resultados de *clustering*.

2.4 Coeficiente RM

Quando se aplica o método PCA sabe-se que, como resultado, as m ($m < p$) primeiras componentes principais, de um conjunto de dados nxp, são as m combinações lineares das p variáveis que maximizam a proporção de variância total explicada. Assim, consideram-se apenas os subespaços que são abrangidos por subconjuntos de variáveis do conjunto original (Cadima *et al.*, 2018), isto é, escolhe-se apenas um conjunto de variáveis do conjunto original que maximiza a proporção de variância explicada. De forma a se poder avaliar quão bem esse subconjunto representa o conjunto original, é habitual avaliar a proporção de variância retida por esse subconjunto recorrendo-se ao coeficiente RM (Equação 2.10).

$$RM = corr(X, P_k X) = \sqrt{\frac{tr(X^T P_k X)}{tr(X^T X)}} = \sqrt{\frac{tr([S^2]_{(K)} S^{-1}_{(K)})}{tr(S)}} \tag{2.10}$$

sendo:

1. X a matriz de dados original;
2. P_k a matriz de projeções ortonormais no subespaço abrangido por um dado subconjunto de k-variáveis;
3. tr o traço da matriz;
4. S a matriz de covariância (pxp);
5. K o índice do conjunto das k variáveis do subconjunto de variáveis;
6. $S_{(K)}$ uma submatriz(kxk) de S, como resultado da escolha das variáveis referentes ao índice K;
7. $[S^2]_{(K)}$ uma submatriz(kxk) de S^2, como resultado da escolha das variáveis referentes ao índice K;

O coeficiente RM mede a semelhança entre as decomposições espectrais da matriz de dados com p variáveis e a matriz de que resulta da regressão dessas variáveis em um subconjunto das mesmas. Varia entre [0,1]. Quanto maior o seu valor maior a proporção de variância explicada. Consequentemente, quanto menor for o grupo de variáveis menor será a variância explicada pelo grupo em relação ao grupo original e, portanto, menor será o valor do coeficiente.

Capítulo 3

Metodologia

Neste capítulo são descritos os métodos utilizados para alcançar os objetivos desta dissertação. É ainda de salientar que as técnicas referidas foram aplicadas com o auxilio do *software* R (R Core Team, 2014).

3.1 Dataset

O conjunto de dados alvo de análise inclui 90 observações dos indivíduos da espécie *Ruditapes philippinarum* e 44 variáveis, incluindo o local de origem (Ria de Aveiro (R), da Ria de Vigo(G) e do Estuário do Tejo (T)) existindo 30 observações de cada local. Das 44 variáveis , 18 dizem respeito à análise química de elementos da concha: Na (sódio, *el*1), Mg (magnésio, *el*2), Al (alumínio, *el*3), P (fósforo, *el*4), Mn (manganésio, *el*5), Fe (ferro, *el*6), Co (cobalto, *el*7), Ni (níquel, *el*8), Cu (cobre, *el*9), Zn (zinco, *el*10), Sr (estrôncio, *el*11), Y (ítrio, *el*12), Ba (bário, *el*13), La (lantânio, *el*14), Ce (cério, *el*15), Nd (neodímio, *el*16), Gd (gadolínio, *el*17) e U (urânio, *el*18)), perfazendo um total de 18 variáveis. Para além do perfil elementar foram também analisados ácidos gordos que caracterizam o músculo adutor da ameijoa japónica com a designação: 14:0 (*ag*1), 15:0 (*ag*2), 16:0 (*ag*3), 16:1n-9 (*ag*4), 16:1n-7 (*ag*5), 17:0 (*ag*6), 18:0 (*ag*7), 18:1n-9 (*ag*8), 18:1n-7 (*ag*9), 18:2n-6 (*ag*10), 18:3n-3 (*ag*11), 18:4n-3 (*ag*12), 20:1n-9/11 (*ag*13), 20:1n-7 (*ag*14), 20:2n-6 (*ag*15), 20:3n-6 (*ag*16), 20:4n-6 (*ag*17), 20:4n-3 (*ag*18), 20:5n-3 (*ag*19), 22:2n-9 (*ag*20), 22:2n-6 (*ag*21), 22:3n-6 (*ag*22), 22:4n-6 (*ag*23), 22:5n-6 (*ag*24), 22:5n-3 (*ag*25) e 22:6n-3 (*ag*26), somando um total de 26 variáveis.

3.2 Redução do número de variáveis

Para determinar o número mínimo de variáveis necessárias à discriminação e separação dos 3 grupos, foi seguido o seguinte algoritmo:

1. Realização do PCA, com base no dataset inicial (p=44). As duas primeiras componentes principais (PC1 e PC2) foram usadas com 2 objetivos (i) representação dos objetos a baixa dimensionalidade, ou seja, em R^2, e, (ii) determinação dos pesos factoriais associados às componentes principais (PC1 ou PC2, conforme o eixo que permite distinguir os três *clusters*) como forma de hierarquizar a importância das variáveis originais na reprodução do dataset original.

2. Cálculo do índice Silhouette.

3. Com base na ordenação das variáveis originais, feita de acordo com os respectivos pesos factoriais, foi, em cada iteração, retirado do dataset a variável com o menor peso associado, por ser essa a que, do conjunto disponível, menos contribuía para a separação dos grupos.

4. Realização do PCA e cálculo do índice Silhouette com base no dataset "reduzido". Análise da separabilidade dos grupos mediante inspeção gráfica e análise dos índices Silhouette.

5. Repetição do passo 3, até p=1.

A metodologia aplicada para reduzir o número de variáveis está representada pelo esquema da Figura 3.1.

Para além da metodologia acima descrita, foi também utilizado o coeficiente RM. Ao escolher um conjunto de variáveis, este avalia o quão semelhante é o conjunto de observações aos dados originais. Então, considerando o método descrito anteriormente, no ponto 2 e 4, foi também avaliado o coeficiente RM. Pela representação gráfica dos valores obtidos, é possível identificar qual o número e quais as variáveis (elementos e/ou ácidos gordos) necessárias e suficientes para representar os dados originais. Esta metodologia está representada na Figura 3.1.

3.3 Redução do tamanho da amostra

A determinação do número mínimo de objetos necessários à boa separação dos grupos foi feita recorrendo a métodos de simulação. Neste caso foram geradas,

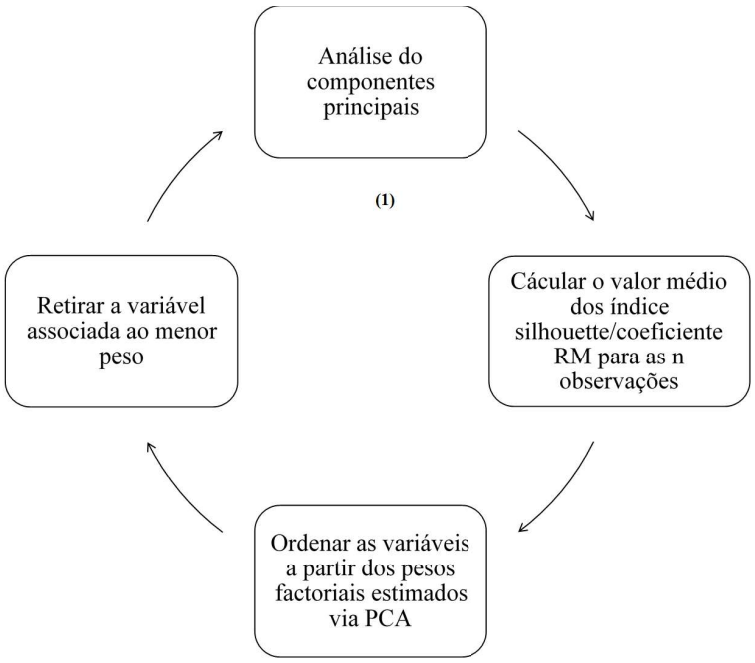


Figura 3.1: Esquema da metodologia para reduzir o número de variáveis (p). O ciclo é iniciado em (1) e repetido até p=1

a partir do dataset original, 1000 amostras (com reposição) com dimensão entre n=2 e n=29.

Com base em cada uma das amostra foram realizadas as análises PCA, *k-means* e cálculos dos índices Silhouette (média e coeficiente de variação) e coeficiente RM. Em cada iteração foram também calculados os centróides de cada grupo, as distâncias Euclideanas entre centróides (produzidos por diferentes amostras com a mesma dimensão amostral) e o coeficiente WSS como indicador do grau de variabilidade dos centróides.

Note-se que a repetição do processo (1000 repetições) permitiu aleatorizar os objetos selecionados e assim obter as distribuições empíricas de cada uma das métricas calculadas.

Seguidamente descrevem-se com maior pormenor os indicadores calculados bem como o seu uso/intenção:

1. **Índice Silhouette**: para cada uma das 1000 simulações, foi realizado uma análise de componentes principais e, como indicador de separabilidade dos grupos, foi calculado o índice Silhouette:

    - Média: Para avaliar se os grupos se encontram bem divididos, foi calculada a média dos índices Silhouettes em cada grupo para cada uma das simulações. Depois de recolhidos as mil médias dos índices Silhouette, obteve-se a distribuição empírica da média. É de esperar que este valor

com a diminuição do n, indicando qual o menor tamanho da amostra que ainda permite separar os diferentes grupos.

- Coeficiente de variação: O coeficiente de variação (CV) é a razão do desvio padrão pela média, sendo uma medida de dispersão.

  Apesar do desvio padrão ser também uma medida de dispersão dos dados, trata-se de uma medida absoluta e não uma medida relativa, portanto, para comparar variabilidades, o CV é mais indicado.

  Para cada simulação foi calculado o coeficiente de variação para os valores de índice de Silhouette obtidos. Em seguida, foi calculada a média destes valores, ou seja, se se considerar uma amostra de tamanho 6, foram realizadas 1000 simulações obtendo, assim, 1000 CV's e o valor médio dos valores. Este processo foi repetido para os diferentes tamanhos de amostra.

2. **Coeficiente RM**: O coeficiente RM foi utilizado para averiguar o quão semelhante é o conjunto de dados amostrados em relação ao original. Portanto, para cada simulação, foi calculado o RM e, depois de obtidos os 1000 valores para a mesma seleção, foi calculado o valor médio. Este processo foi repetido desde n=29 até n=2. É esperado que este valor diminua à medida que se reduz a dimensão da amostra.

3. **Distância Euclideana**: Por cada simulação, foi calculado o centróide de cada grupo. Assim, em 1000 simulações, haverá 1000 valores por cada um dos grupos (3000 centróides, 1000 por grupo). Depois de calculados os pontos médios de cada grupo foi calculada a distância Euclideana entre cada simulação referente ao mesmo grupo (por exemplo, para cada grupo, em duas simulações, foi calculada a distância entre os centróides obtidos na primeira e na segunda simulações). No final, depois de calculadas as distâncias entre centróides por grupo, é caculada a média das mesmas, ficando com três distâncias médias associadas a cada grupo,associada a cada n (Figura 3.3). Portanto, é de esperar que os centróides, à medida que a dimensão da amostra aumente, variem menos, obtendo-se, assim, distâncias entre centróides sucessivamente menores. A partir de um certo valor de n, é esperado que as distâncias estabilizem em torno de um certo valor, indicando o valor de n a partir do qual o aumento da diminuição amostra já não contribui para melhorar a distribuição dos grupos.

4. **WSS**: Por cada simulação realizada, para um determinado n, foi calculado o ponto médio de cada zona (centróide). Ao fim de 1000 simulações, por cada

grupo, há 1000 centróides associados. Depois de obtidos os pontos médios, foi aplicado o método *k-means* e obtida a soma dos quadrados intra-*clusters* de cada grupo (WSS), de forma a avaliar a variação dentro de cada grupo. Desta forma, é esperado que a variabilidade, à medida que o tamanho da amostra aumenta, vá diminuindo e, a partir de um certo n, estabilize.

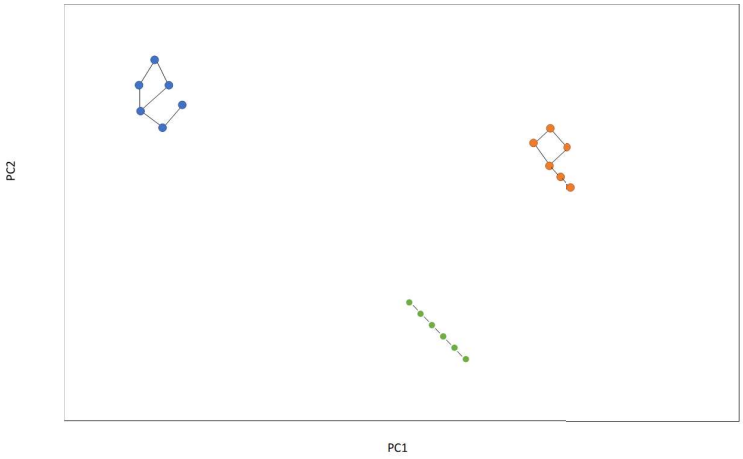


Figura 3.2: Cada ponto representa um centróide obtido numa amostra obtida por simulação. Cada traço representa a distância Euclideana calculada entre cada par de amostras (com o mesmo n). Legenda : PC1- Primeira componente principal, PC2- Segunda componente principal

Em suma, a determinação do tamanho mínimo de amostra necessária para discriminar os três grupos foi feita por simulação. Foram calculadas as médias dos índices Silhouette e o coeficiente de variação; o coeficiente RM; o WSS e a distância Euclideana entre centróides. Cada uma das análises foi sempre realizada juntamente com interpretação gráfica.

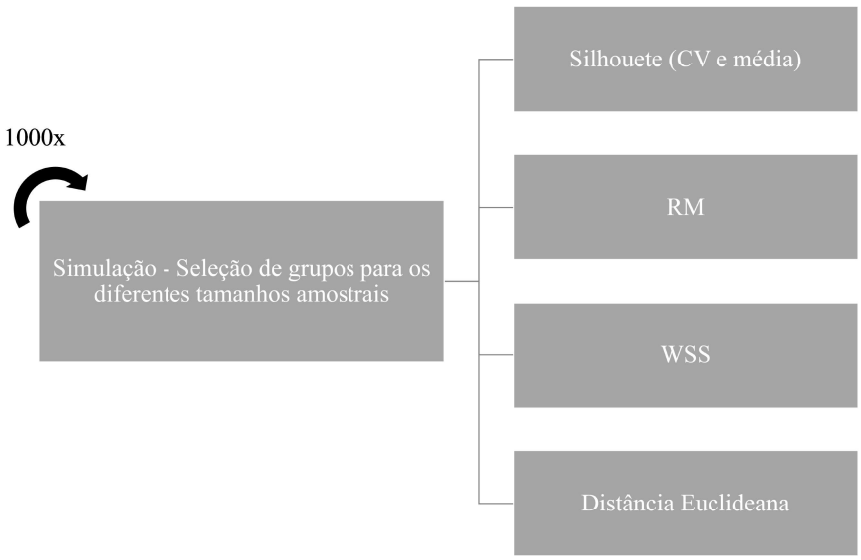


Figura 3.3: Esquema da metodologia para determinar o tamanho da amostra (n). Legenda: CV- coeficiente de variação; RM- coeficiente RM; WSS-soma dos quadrados intra-*clusters* (*within-cluster sum of squares)*

Capítulo 4

Resultados

4.1 Determinação do número mínimo de variáveis

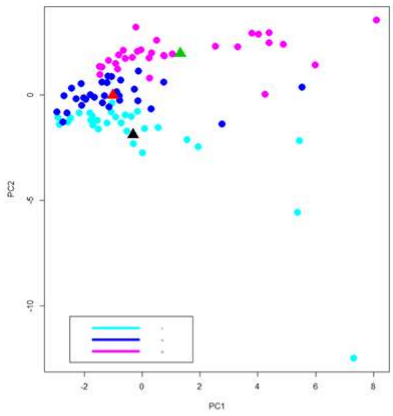
Um dos objetivos propostos nesta dissertação é a redução do número de variáveis. Num estudo preliminar aos dados, com base na análise de componentes principais (PCA), foi possível representar graficamente os dados originais (Figura 4.1). A figura mostra que usando os 18 elementos, os 26 ácidos gordos e o conjunto de ambos (44 variáveis) é possível identificar graficamente os três grupos referentes a cada origem. Seguidamente, estudou-se cada um dos conjunto de variáveis de forma a aferir a possibilidade de reduzir o número de variáveis e, em simultâneo, diferenciar as três origens.

4.1.1 Ácidos gordos

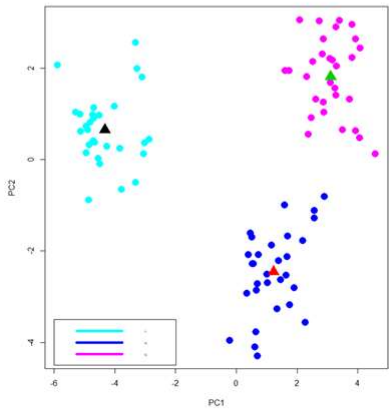
Considerando a representação gráfica da figura 4.1 B, observou-se que a partir do eixo da abcissa conseguimos dividir as observações em três grupos, portanto, a primeira componente principal (PC1) permite distinguir as três zonas T, R e G. Na tabela 4.1 apresenta-se os pesos factoriais associados a cada variável da primeira componente principal.

Da análise dos valores de t_k da tabela 4.1 verificou-se que o $ag11$ tem o menor peso factorial de 0.015. Este, por apresentar o menor peso factorial, é esperado que pouco influencie na distinção das zonas em estudo. Foi, por isso, o primeiro a ser removido da amostra.

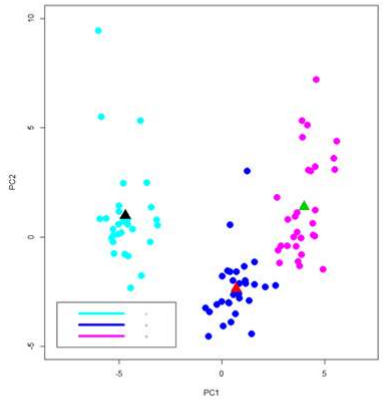
Seguindo o mesmo raciocínio, é realizada uma nova análise de componentes



A



B



C

Figura 4.1: Representação gráfica dos resultados da análise de componentes principais da amostra inicial. Legenda: A- PCA dos elementos (18 variáveis) , B- PCA dos ácidos gordos (26 variáveis), C- PCA da amostra original (44 variáveis). Cada centróide está representado por um triângulo (preto- grupo T, rosa- grupo R, verde- grupo G)

Tabela 4.1: Pesos factoriais (t_k) da PC1 associados às 26 agk ($k = 1,...,26$)

k	var	t_k
1	$ag11$	0.015
2	$ag21$	0.037
3	$ag15$	0.046
4	$ag4$	0.068
5	$ag3$	0.082
6	$ag12$	0.106
7	$ag10$	0.135
8	$ag7$	0.144
9	$ag13$	0.156
10	$ag24$	0.159
11	$ag2$	0.161
12	$ag16$	0.168
13	$ag8$	0.175
14	$ag23$	0.194
15	$ag17$	0.202
16	$ag1$	0.205
17	$ag6$	0.243
18	$ag20$	0.243
19	$ag25$	0.249
20	$ag22$	0.251
21	$ag18$	0.252
22	$ag14$	0.256
23	$ag5$	0.265
24	$ag9$	0.265
25	$ag26$	0.280
26	$ag19$	0.298

principais e, em conjunto com a visualização gráfica, é retirada a variável com menor peso factorial.

4.1.1.1 Índice Silhouette

Inicialmente, para as 26 variáveis, o índice Silhouette médio observado foi de 0.34, indicando uma boa separabilidade dos grupos. Este, à medida que o número de variáveis diminui, varia. O índice Silhouette aumenta para 0.45 para sete variáveis e para 0.61 para 2 variáveis (Figura 4.2). No entanto, apesar de para p=2 o índice médio Silhouette ser maior, a variância retida decresce para 50% da original, enquanto que, para p=7, retem aproximadamente 71%. Pela figura 4.3, pode-se também observar que, visualmente, a divisão dos grupos não é tão eficiente como para quando p=7, o que indica que o índice médio Silhouette aumentou devido aos objetos estarem mais próximos no próprio *cluster* e não por haver uma melhor

divisão de grupos. Assim, com sete ácidos gordos ($ag5$, $ag9$, $ag14$, $ag18$, $ag19$, $ag22$ e $ag26$), é possível distinguir as diferentes origens, resultando numa redução da diminuição amostral de, aproximadamente, 73%.

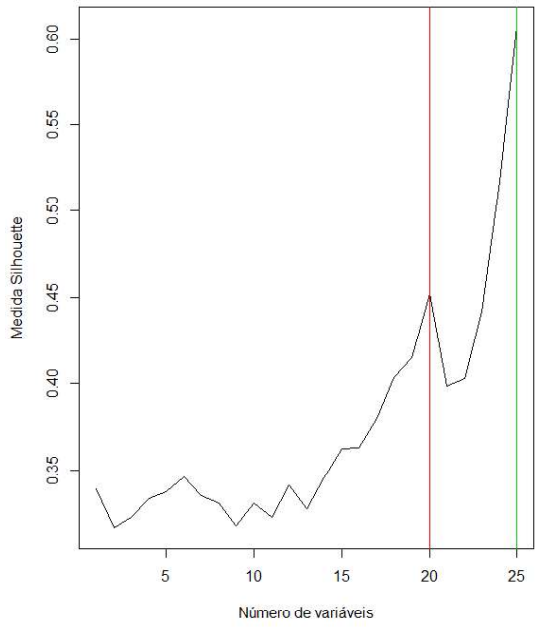


Figura 4.2: Representação gráfica dos índices Silhouette para as variáveis $ag(i)$, $i = 1,...,26$. A recta vermelha marca o índice de Silhouette associado à divisão dos grupos (para quando p=7) igual a 0.45 e a recta verde marca o índice de Silhouette associado à divisão dos grupos (para quando p=2) igual a 0.61

Pode-se assumir, então, que um maior número de variáveis nem sempre implica uma melhor representação da população, na verdade, ao serem da mesma espécie, muitas destas variáveis são comuns entre indivíduos, o que explica que um menor número de componentes possa também distinguir os grupos eficazmente.

A análise dos gráficos para p=26 (inicial) e p=7 (final) (Figura 4.4) corrobora as conclusões acima referidas. No primeiro caso, como acima mencionado, consegue-se observar uma boa divisão das três zonas. Quando reduzidos para p=7, também se observa as três diferentes zonas. No entanto, apesar do maior índice de Silhouette, os grupos parecem mais próximos mas, ao mesmo tempo, os dados encontram-se mais próximos dentro do seu próprio *cluster*. Isto indica que, tendo em conta a definição do índice (equação 2.8, capítulo 2), o valor subiu devido a uma diminuição de $a_{(i)}$, a distância média Euclideana do objeto i aos restantes objetos do próprio *cluster*. Isso pode-se observar na figura 4.5, em que o índice Silhouette de cada objeto, na maioria, aumenta quando reduzido o valor de p e,

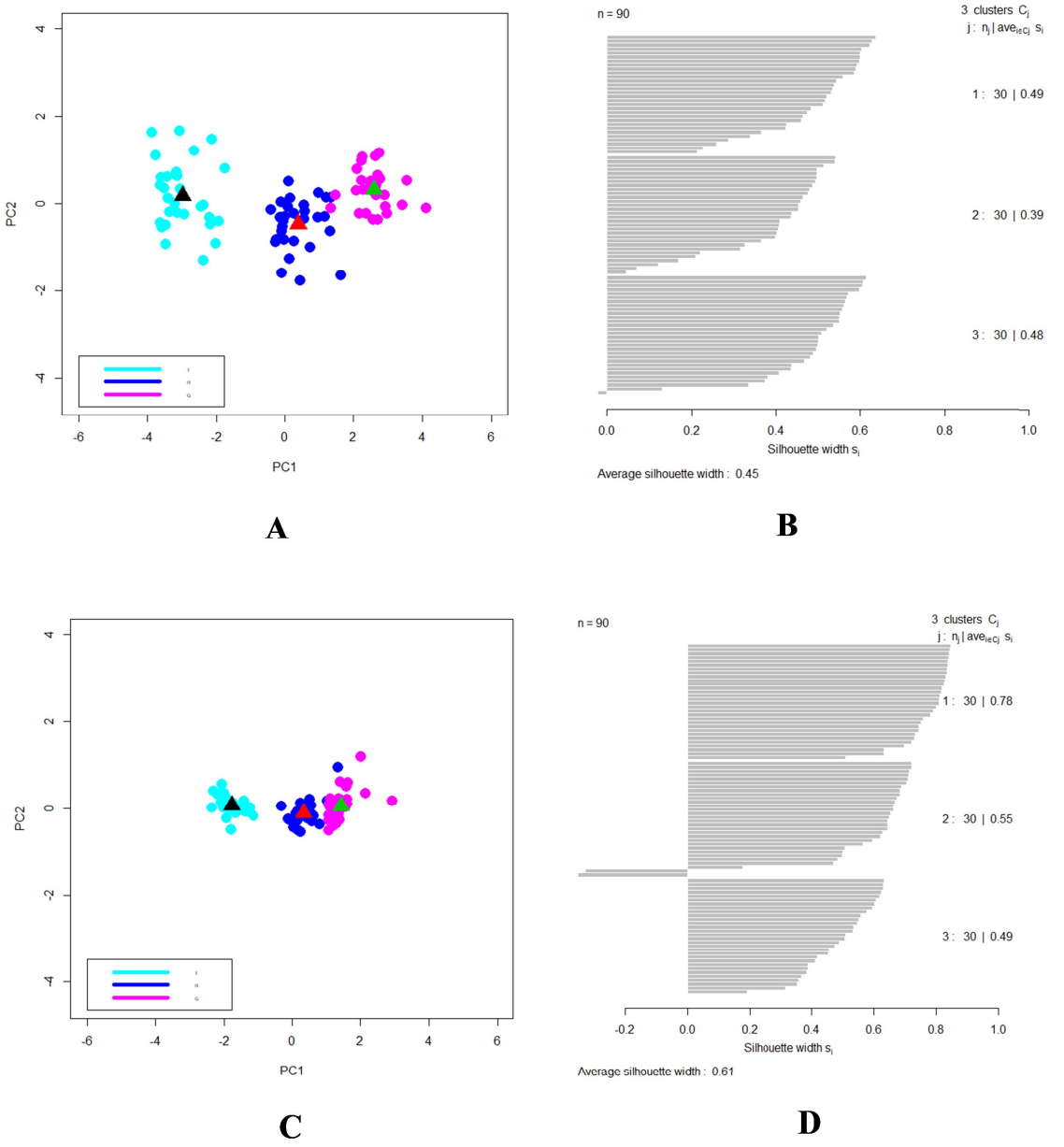


Figura 4.3: Representação gráfica dos índices Silhouette para as variáveis $ag(i)$, $i = 1,...,26$ e dos resultados da análise de componentes principais da amostra inicial. Legenda: A- PCA dos ácidos gordos (7 variáveis), B- Silhouette dos ácidos gordos (7 variáveis, índice médio Silhouette= 0.45), C- PCA dos ácidos gordos (2 variáveis), D- Silhouette dos ácidos gordos (2 variáveis, índice médio Silhouette= 0.61). Cada centróide está representado por um triângulo (preto- grupo T, rosa- grupo R, verde- grupo G)

consequentemente, o valor médio do grupo também aumenta. Porém, é possível distinguir graficamente as três origens em estudo da ameijoa japónica apenas com sete variáveis (ácidos gordos). Portanto, é possível afirmar que é exequível diminuir o número de variáveis originais e ainda identificar as diferentes zonas.

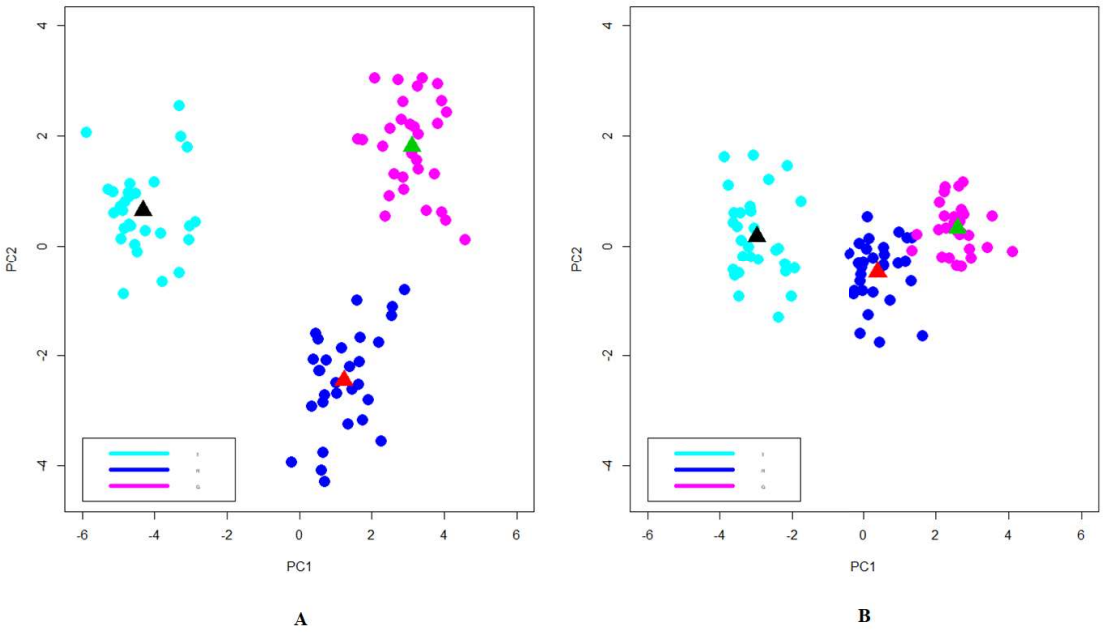


Figura 4.4: Representação gráfica dos resultados da análise de componentes principais da amostra representada apenas com as variáveis agk, $k = 1,...26$. Legenda: A- PCA da amostra representada pelos ácidos gordos (26 variáveis) , B- PCA da amostra representada pelos ácidos gordos (7 variáveis). Cada centróide está representado por um triângulo (preto- grupo T, rosa- grupo R, verde- grupo G)

4.1.2 Elementos

De forma a avaliar a importância dos elementos na distinção dos grupos, procedeu-se da mesma forma que no processo de análise dos ácidos gordos, ou seja, selecionou-se apenas as variáveis referentes aos elementos, um total de 18 variáveis. E repetiu-se o mesmo processo.

Pela análise do gráfico que representa as duas primeiras componentes principais (Figura 4.1 A), resultantes da análise de componentes principais dos dados referentes aos elementos, pode-se observar que a divisão dos grupos se realiza ao longo do eixo das ordenadas, contrariamente ao que acontece no caso dos ácidos gordos. Assumiu-se, desta forma, a segunda componente principal (PC2) como a que mais contribui para a separação dos grupos T, G e R.

A tabela 4.2 apresenta os pesos factoriais associados às variáveis elk, resultantes da primeira análise PCA. Aplicada a metodologia descrita no terceiro capítulo,

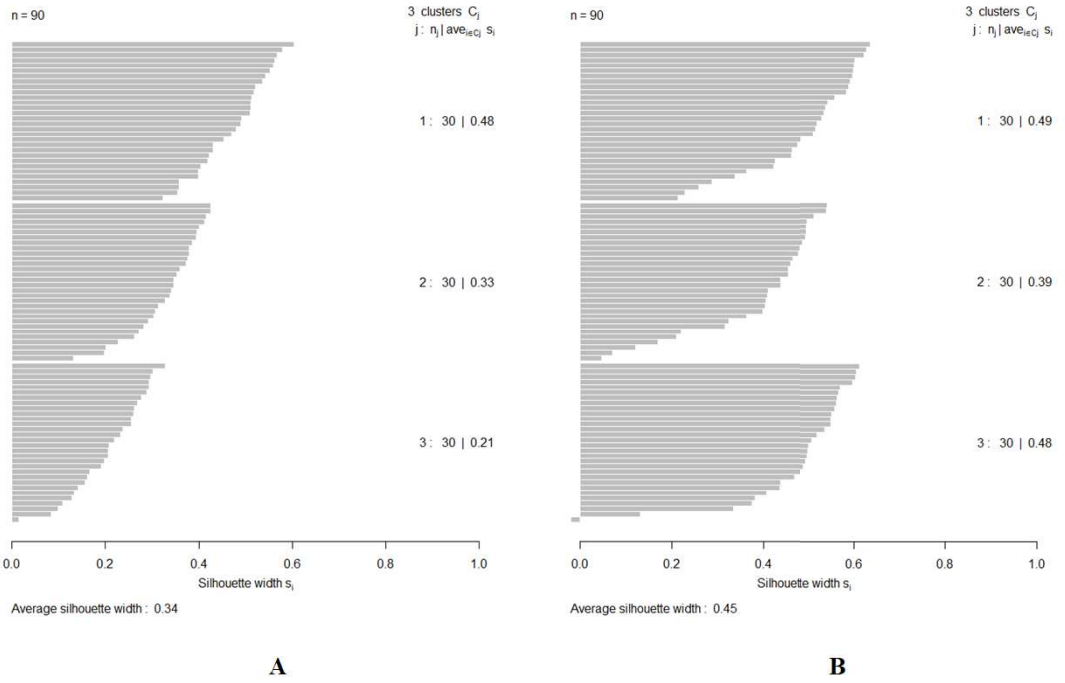


Figura 4.5: Representação gráfica dos índices Silhouette associados a cada objeto i, resultante da análise de componentes principais da amostra representada pelas variáveis agk, $k = 1,...26$. Legenda:A- Silhouette dos ácidos gordos (26 variáveis, índice médio Silhouette= 0.34), B- Silhouette dos ácidos gordos (7 variáveis, índice médio Silhouette= 0.45)

foi retirada o $el15$, por apresentar um menor peso factorial (0.004). Ao realizar um novo PCA, os pesos vão variando (o que significa que os valores na Tabela 4.2 diferem),sendo as variáveis retiradas de acordo como esses pesos.

Novamente, o objetivo é conseguir encontrar um número mínimo de elementos que permita distinguir diferentes populações da espécie em estudo. Nomeadamente, é necessário ter uma métrica que permita avaliar se os grupos se encontram ou não, bem discriminados.

4.1.2.1 Silhouette

Como esperado, pela representação gráfica, o valor médio dos índices Silhouette para 18 elementos é baixo (aproximadamente, 0.12). Este valor varia à medida que se calcula o valor médio dos índices Silhouette para cada PCA, e, a partir de um certo valor, apresenta um máximo indicando a melhor divisão dos *clusters*. Este pico acontece quando o número de variáveis p é reduzido para 6 ($el1$, $el6$, $el7$, $el10$, $el11$ e $el17$) (Figura 4.6). No entanto, este índice pouco aumenta. Com um valor máximo de 0.15, a divisão dos grupos pouco melhorou.

Tabela 4.2: Pesos factoriais (t_k) da PC2 associados às 18 variáveis elk ($k = 1,...,18$)

k	var	t_k
1	$el15$	0.004
2	$el14$	0.006
3	$el8$	0.008
4	$el17$	0.036
5	$el13$	0.058
6	$el16$	0.065
7	$el4$	0.095
8	$el7$	0.189
9	$el5$	0.197
10	$el3$	0.232
11	$el18$	0.241
12	$el12$	0.268
13	$el1$	0.308
14	$el9$	0.308
15	$el6$	0.323
16	$el10$	0.354
17	$el11$	0.355
18	$el2$	0.268

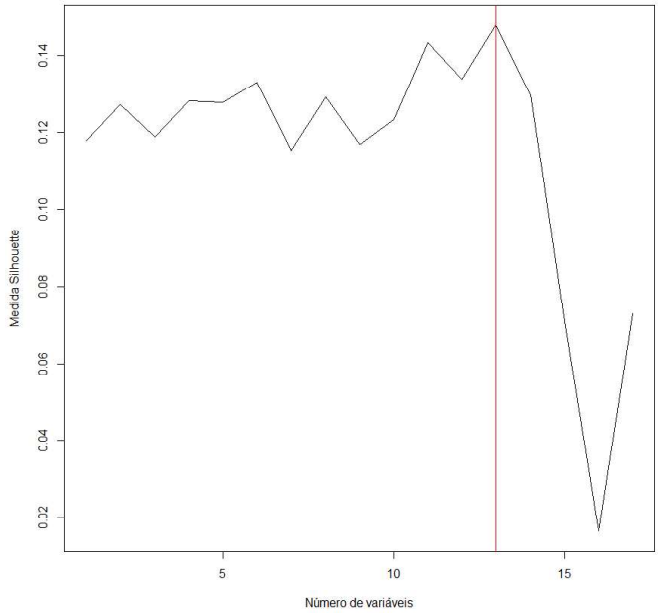


Figura 4.6: Representação gráfica dos índices Silhouette para as variáveis $el(i)$, $i = 1,...,18$. A recta vermelha marca o índice de Silhouette associado à melhor divisão dos grupos (para quando p=6) igual a 0.15

Depois de analisados os gráficos, antes e depois da redução (figura 4.7) pôde-se inferir que, a representação das zonas em estudo, recorrendo apenas aos elementos não permite distinguir os grupos. Para além da visualização gráfica, pela representação dos índices Silhouette de cada objecto (Figura 4.8) observa-se que, para além da zona do Estuário do Tejo, que tanto pelos ácidos gordos como pelos elementos, é o grupo mais distante dos restantes, as Ria de Aveiro e Vigo pouco se distinguem, apresentando um valor médio Silhouette negativo ou muito baixo, por grupo (0.12 e -0.08 respectivamente, para p=6).

Em suma, só os elementos, que são responsáveis pela assinatura elementar da concha da ameijoa japónica, não permitiram uma boa distinção da origem da espécie em estudo. No entanto, o mesmo não aconteceu para os ácidos gordos portanto, ou (i) obtém-se uma melhor divisão de *clusters* apenas com as variáveis dos ácidos gordos para p=7, ou (ii) com um conjunto de variáveis de elementos e ácidos gordos, consegue-se um melhor índice Silhouette, ou seja, uma melhor caracterização das populações.

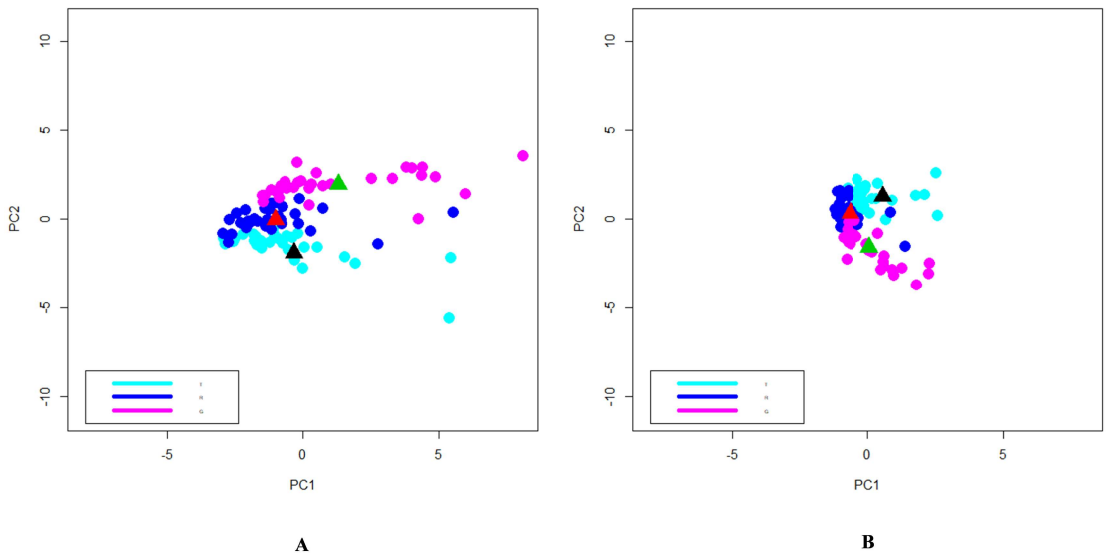


Figura 4.7: Representação gráfica dos resultados da análise de componentes principais da amostra representada apenas com as variáveis elk, k = 1,...18. Legenda: A- PCA da amostra representada pelos elementos (18 variáveis) , B- PCA da amostra representada pelos elementos (6 variáveis). Cada centróide está representado por um triângulo (preto- grupo T, rosa- grupo R, verde- grupo G)

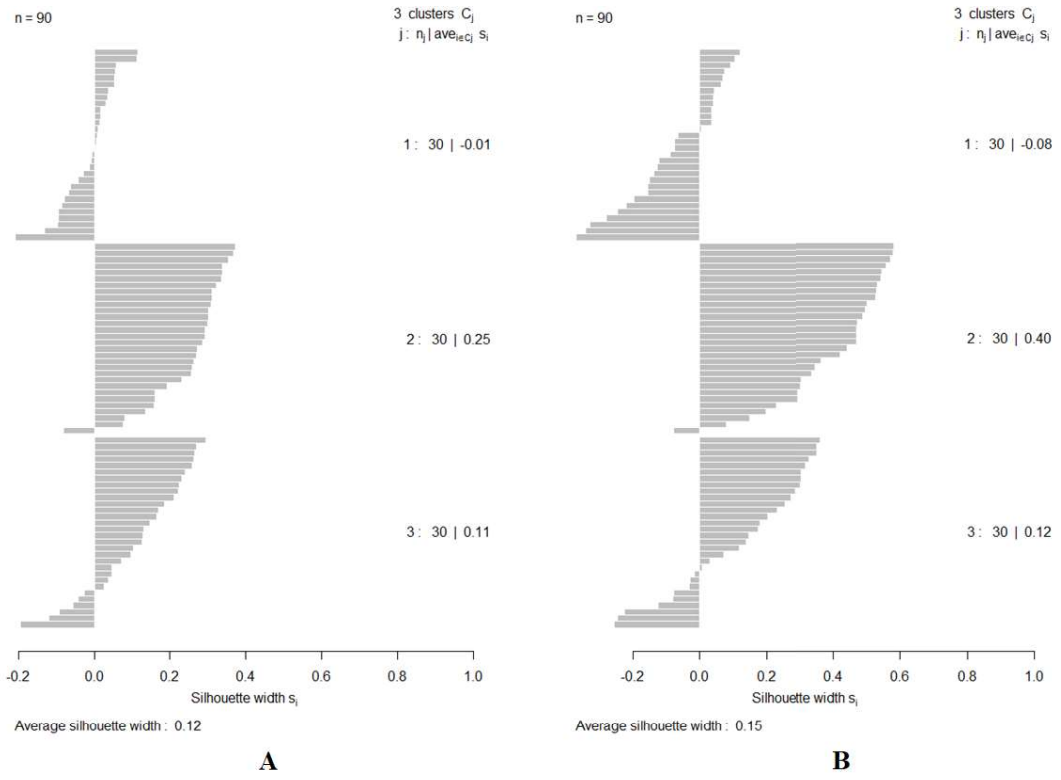


Figura 4.8: Representação gráfica dos índices Silhouette associados a cada objeto i, resultante da análise de componentes principais da amostra representada pelas variáveis elk, $k = 1,...18$. Legenda:A- Silhouette dos elementos (18 variáveis, índice médio Silhouette= 0.12), B- Silhouette dos elementos (6 variáveis, índice médio Silhouette= 0.15)

4.1.3 Todas as variáveis

Inicialmente, quando observados os resultados da análise de componentes principais para todas as variáveis (Figura 4.1), foi possível, com as 44 variáveis (elementos e ácidos gordos), identificar as diferentes origens de forma clara. Porém, pela visualização gráfica (Figura 4.1), a divisão dos grupos quando caracterizados apenas pelos ácidos gordos é maís visível do que quando se tem em conta a amostra original indicando que, possivelmente, ao analisar os dados em conjunto com os elementos pode piorar a distinção dos grupos.

Recorrendo à metodologia e raciocínio descrito anteriormente, tal como para os ácidos gordos, é a primeira componente principal que distingue os três *clusters*. Assim, tal como antes, foi analisada a combinação linear obtido por PCA. Na primeira análise pode-se observar que o $el7$ apresenta o menor peso factorial de 0.003 (Tabela 4.3). O processo repetiu-se de forma a se obter uma medida de separação dos grupos para cada análise.

Tabela 4.3: Pesos factoriais (t_k) da PC1 associados às 44 variáveis *elk* e *agk*

k	var	t_k	k	var	t_k
1	$el13$	0.005	27	$ag10$	0.156
2	$el17$	0.006	28	$ag8$	0.163
3	$el14$	0.007	29	$ag23$	0.173
4	$el4$	0.014	30	$ag24$	0.177
5	$ag15$	0.017	31	$el2$	0.179
6	$el15$	0.019	32	$ag17$	0.193
7	$el16$	0.022	33	$el11$	0.198
8	$el8$	0.024	34	$ag22$	0.202
9	$ag11$	0.039	35	$ag18$	0.203
10	$ag21$	0.045	36	$ag20$	0.210
11	$ag4$	0.060	37	$ag1$	0.214
12	$el7$	0.062	38	$ag6$	0.219
13	$el9$	0.071	39	$ag14$	0.223
14	$el3$	0.076	40	$ag25$	0.224
15	$ag3$	0.085	41	$ag5$	0.239
16	$el10$	0.095	42	$ag9$	0.241
17	$ag7$	0.101	43	$ag26$	0.245
18	$el18$	0.108	44	$ag19$	0.266
19	$ag16$	0.113			
20	$el1$	0.124			
21	$ag2$	0.134			
22	$el12$	0.137			
23	$ag12$	0.137			
24	$el5$	0.147			
25	$ag13$	0.149			
26	$el6$	0.155			

4.1.3.1 Índice Silhouette

Com as 44 variáveis iniciais, a média dos índice Silhouette é 0.26 (Figura 4.9). Este valor decresce só com os elementos para 0.12 mas, em comparação com análise inicial dos ácidos gordos, é atingido um menor valor, o que corrobora a análise anterior referente aos gráficos apresentados pela figura 4.1.

Na figura 4.9 representam-se os índices médios calculados para cada PCA em função do p. É observável que, em média, os índices vão aumentando até p=2. Mas tal como para o estudo dos ácidos gordos, essa amostra não é representativa da original, visto só explicar 50% da variância da amostra original. Considerando o pico anterior, para p=7, a variância explicada é de 71%, aproximadamente. Isto indica que depois de retiradas 37 variáveis pelos critérios acima descritos, a divisão dos grupos é a máxima para os objetos em estudo. Portanto, apenas são necessárias sete variáveis para identificar a origem da espécie.

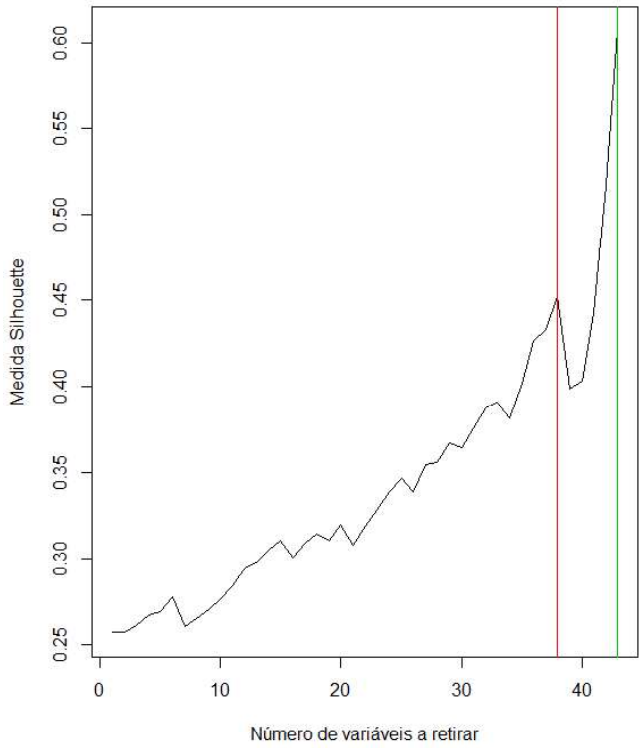


Figura 4.9: Representação gráfica dos índices Silhouette para as variáveis $el(i)$ e $ag(k)$, $i = 1,...,18$ e $k = 1,...,26$. A recta vermelha marca o índice de Silhouette associado à melhor divisão dos grupos (para quando p=7) igual a 0.45

As sete variáveis necessárias para a determinação da origem geográfica da ameijoa dizem respeito apenas a ácidos gordos, mais respectivamente, aos sete ácidos gordos obtidos anteriormente ($ag5$, $ag9$, $ag14$, $ag18$, $ag19$, $ag22$ e $ag26$).

Portanto, foi possível reduzir o número de variáveis de 44 elementos e ácidos gordos para apenas 7 ácidos gordos e ainda distinguir as populações da *Ruditapes philippinarum* das diferentes zonas de origem em estudo Ria de Aveiro e Vigo e estuário do Tejo (Figura 4.10).

Portanto, foram obtidos os seguintes resultados:

1. Redução do número de elementos de 18 para 6: $el1$, $el6$, $el7$, $el10$, $el11$ e $el17$. No entanto, o valor médio dos índices Silhouette associado às observações é baixo (0.15) indicando que pouco distingue a origem de cada população em estudo;

2. Redução do número do conjunto de elementos e ácidos gordos de 44 para 7: $ag5$, $ag9$, $ag14$, $ag18$, $ag19$, $ag22$ e $ag26$, com um valor médio dos índices Silhouette de 0.45, indicando uma boa divisão das diferentes zonas.

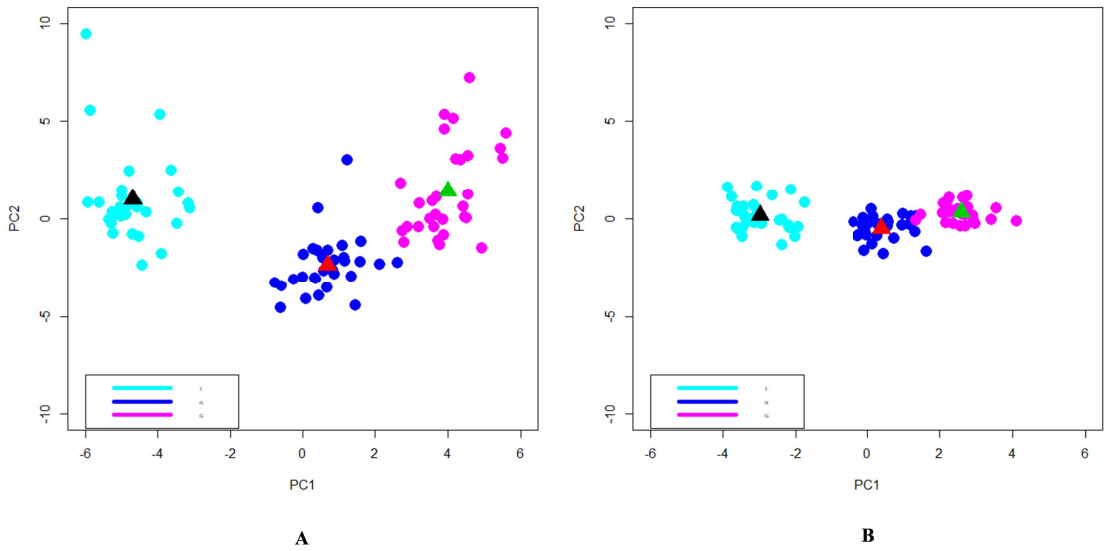


Figura 4.10: Representação gráfica dos resultados da análise de componentes principais da amostra representada com as variáveis em estudo ($el(i)$ e $ag(k)$, $i = 1,...,18$ e $k = 1,...,26$). Legenda: A- PCA da amostra com as variáveis em estudo (44 variáveis) , B- PCA da amostra com as variáveis em estudo (7 variáveis). Cada centróide está representado por um triângulo (preto- grupo T, rosa- grupo R, verde- grupo G)

4.1.4 Coeficiente RM *vs.* Indíce Silhouette

O índice Silhouette foi utilizado para avaliar qual o melhor conjunto de variáveis que permite distinguir a origem geográfica da espécie em estudo. Permite avaliar o quão bem colocado o objeto está no grupo em que se encontra e, a partir da média dos valores associados a cada objeto, a qualidade da divisão dos grupos no geral.

Complementarmente ao método aplicado anteriormente, usou-se a análise a partir do coeficiente RM. Este indica a quantidade de variância explicada por um conjunto de dados. Portanto, em contraste com o índice Silhouette, este não avalia a divisão de *clusters* mas sim o quão bem um certo subconjunto de dados representa os dados originais.

Foi tido em conta o conjunto de variáveis original, 26 ácidos gordos e 18 elementos. Por cada grupo gerado a partir do PCA, foi calculado o coeficiente RM. Evidentemente, por cada variável que é retirada, a variância explicada é menor, afastando-se dos dados originais.

Pode-se observar, pela análise da figura 4.11, que com 25 variáveis consegue-se obter um índice Silhouette perto de 0.32 e um coeficiente RM de 0.95 (Tabela 4.4).

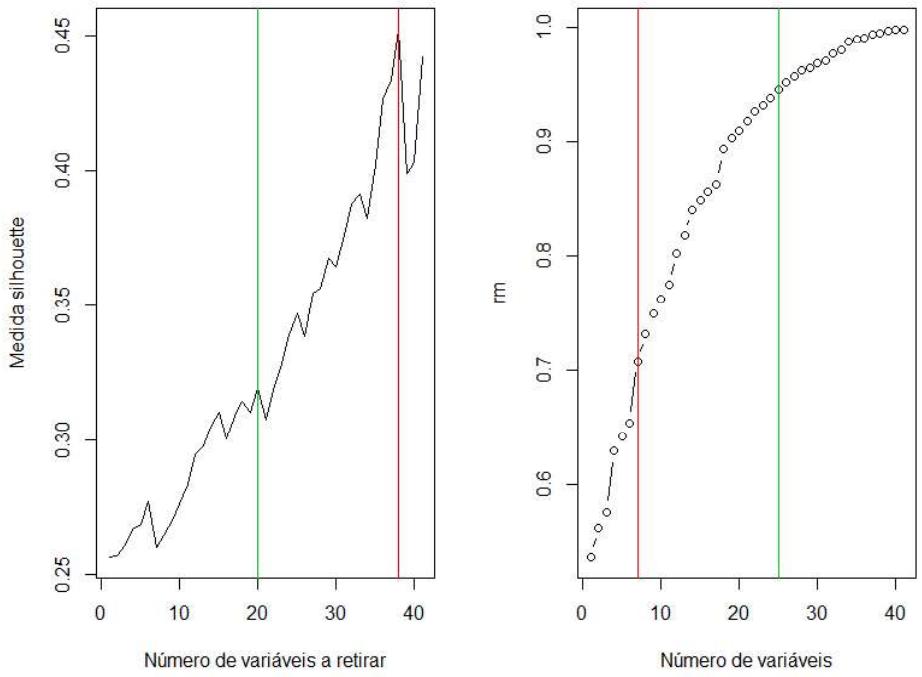


Figura 4.11: Comparação da escolha do número de variáveis pelo índice Silhouette (medida Silhouette) e pelo coeficiente RM (Rm)

Tabela 4.4: Comparação dos resultados com o coeficiente RM e o índice Silhouette

Conjunto de dados	**Coeficiente RM**	**Índice Silhouette**
$ag1$, $ag2$, $ag5$, $ag6$, $ag7$, $ag8$, $ag9$, $ag13$, $ag14$, $ag16$, $ag17$, $ag18$, $ag19$, $ag20$, $ag22$, $ag23$, $ag24$, $ag25$, $ag26$, $el1$, $el2$, $el6$, $el18$, $el11$	0.945	0.319
$ag5$, $ag9$, $ag14$, $ag18$, $ag19$, $ag22$ e $ag26$	0.707	0.452

Enquanto que, com 7 variáveis consegue-se um índice Silhouette superior de 0.45 e um coeficiente RM de 0.71 (Tabela 4.4). Portanto, pelo coeficiente RM é escolhido o conjunto de dados com 25 variáveis e pelo índice Silhouette é escolhido o conjunto de 7 variáveis.

O fator de decisão para o conjunto de dados a escolher tem em conta o objetivo do estudo realizado. A finalidade deste trabalho é reduzir o número de variáveis e, ainda assim, conseguir distinguir as diferentes populações. Com 25 variáveis observou-se uma boa divisão dos grupos (Figura 4.12) e um coeficiente RM alto. No entanto, com 7 variáveis consegue-se observar, pela representação gráfica do PCA (Figura 4.12), uma boa divisão dos três grupos associada ao maior

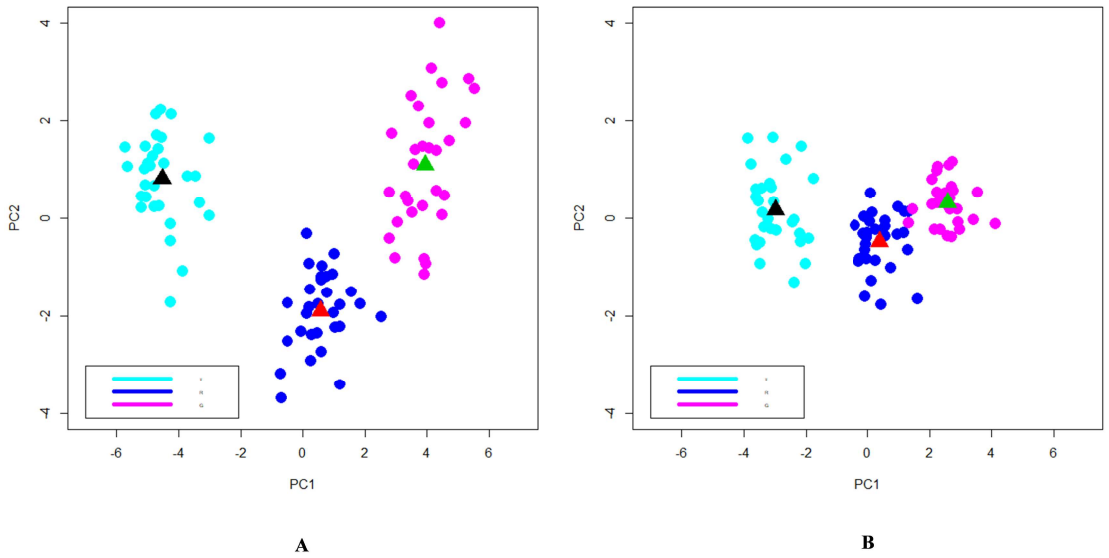


Figura 4.12: Representação gráfica dos resultados da análise de componentes principais da amostra representada com as variáveis em estudo ($el(i)$ e $ag(k)$, $i = 1,...,18$ e $k = 1,...,26$). Legenda: A- PCA da amostra com as variáveis em estudo (25 variáveis) , B- PCA da amostra com as variáveis em estudo (7 variáveis).Cada centróide está representado por um triângulo (preto- grupo T, rosa- grupo R, verde- grupo G)

valor do índice Silhouette e ainda obter um coeficiente RM de 0.71, uma diferença menor que 0.3 dos dados originais. Portanto, indo ao encontro do principal objetivo, tendo em conta o índice Silhouette (que avalia a qualidade da divisão de *clusters*) e ainda a representação gráfica que mostra uma boa separação das diferentes origens, conclui-se que o subconjunto de 7 variáveis apresenta uma solução adequada para o problema em questão.

Finalizando este capítulo, conclui-se que é possível diminuir, de forma significativa, o número de variáveis sem prejudicar a divisão das populações em estudo da espécie *Ruditapes philippinarum* originárias da Ria de Aveiro, Ria de Vigo ou Estuário do Tejo.

4.2 Determinação do número médio de observações amostrais

A redução do tamanho da amostra (n) é um dos desafios desta dissertação. Como mencionado, há três áreas em estudo que dizem respeito ao habitat das populações da ameijoa japónica. Por cada uma das áreas, foram amostrados 30 indivíduos, somando um total de n=90 observações.

Para reduzir o tamanho da amostra foram tidos em conta os seguintes pontos:

- Para poder reduzir a amostra tem que avaliar se, independentemente do objecto que se retira, os resultados não se alteram;
- Tem que se conseguir visualizar a distinção dos grupos (Ria de Aveiro, Estuário do Tejo e Ria de Vigo);
- Por último, é necessário avaliar se a amostra reduzida representa bem os dados originais

Ao longo deste capítulo serão abordados os três pontos acima mencionados recorrendo às metodologias descritas no capítulo 3, que permitiram alcançar o objectivo mencionado.

4.2.1 Simulação

Como dito anteriormente é necessário avaliar a separação dos grupos varia conforme o(s) objecto(s) que se retira(m) da amostra. Foram estudadas duas maneiras diferentes de o fazer:

1. Índice Silhouette: este permite avaliar a distinção entre os grupos. Tem como premissa que a divisão dos grupos não se altera se o valor médio dos índices de Silhouette não varia consoante o(s) objeto(s) retirado(s).
2. Coeficiente RM: indica a quantidade de variância retida numa amostra dos dados originais. Tal como para o caso dos índices Silhouette, ao longo das mil simulações, é esperado que este valor não se altere consoante o objeto retirado;

De seguida, são descritos os resultados obtidos pelos passos acima mencionados.

4.2.1.1 Índice Silhouette

Neste secção apresentam-se os resultados depois de retiradas diferentes amostras de forma aleatória, em mil simulações, sendo representados os respectivos índices Silhouette num histograma. Desta forma, é possível observar a distribuição dos índices calculados e avaliar se é indiferente a observação que é retirada.

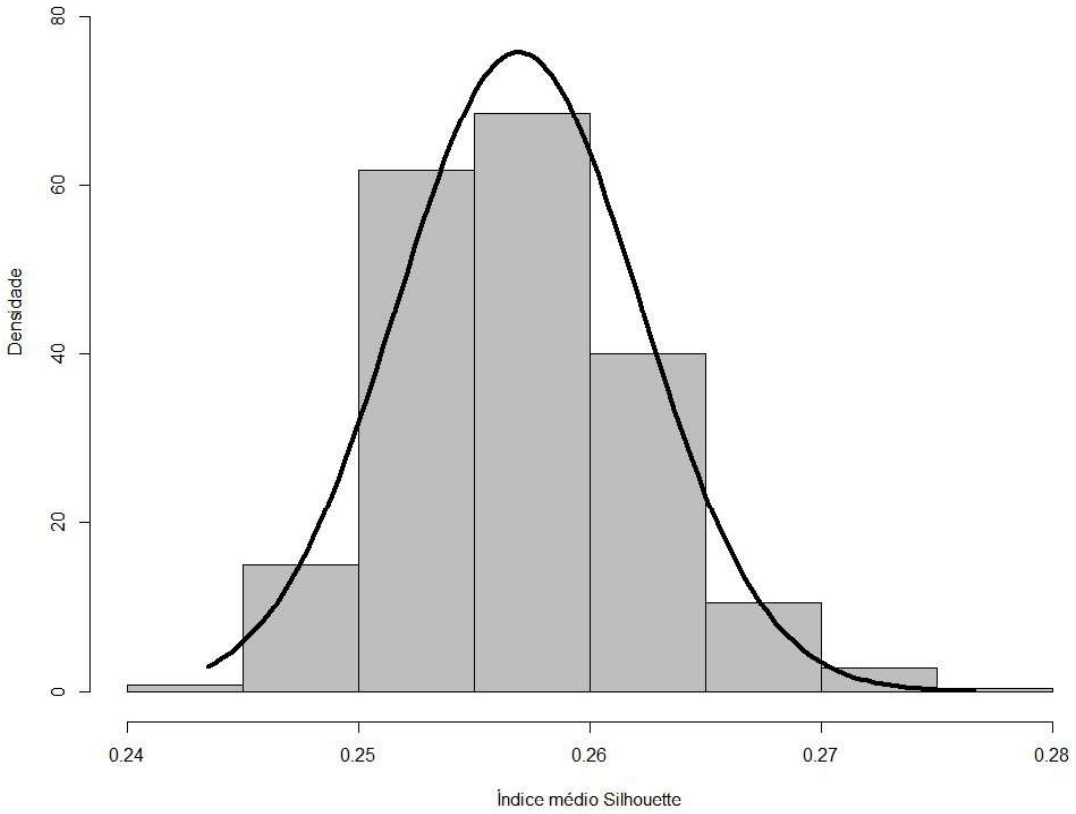


Figura 4.13: Histograma dos índices médios Silhouette resultantes das 1000 simulações para quando a amostra é reduzida para n=81 (com as variáveis iniciais), com a sobreposição da curva da distribuição normal com a média e desvio padrão dos índices médios Silhouette

Pela análise da figura 4.13 pode-se depreender que não é indiferente o objecto que é retirado. Caso fosse observado uma distribuição uniforme, para um determinado n, ao fim de mil simulações, tal indicaria que, qualquer que fosse a observação retirada para qualquer amostra de tamanho n, o índice Silhouette seria sempre o mesmo e, portanto, não afetava a separação dos grupos. No entanto, como se observa para n=81 (Figura 4.13), os índices de Silhouette variam conforme a observação retirada, apresentando uma distribuição aproximada à normal.

Contudo, observando os valores dos índices da figura 4.13, estes pouco variam. Na verdade, estes valores apresentam uma amplitude de 0.04 entre eles (entre 0.24 e 0.28), continuando a mostrar em qualquer caso uma boa divisão das zonas em estudo. Como é calculado um valor médio dos índices Silhouette, este é sensível ao objeto que é retirado. No "melhor" dos casos (quando o índice tem o valor máximo de 0.28), os objetos retirados são as que se encontram na fronteira entre os três *clusters*, melhorando a separação dos grupos e, consequentemente, o valor médio de $s_{(i)}$. O contrário também se verifica para o "pior" dos casos, quando $s_{(i)}$ diminui para 0.24.

Resumidamente, apesar dos índices Silhouette variarem em função do objeto retirado, estes valores pouco diferem indicando, dessa maneira, pouca sensibilidade ao elemento retirado.

4.2.1.2 Coeficiente RM

Como descrito no capítulo 3, sempre que era retirada uma observação dos dados originais e nova amostra era comparada com a original considerando o coeficiente RM.

Para cada n, é esperado que haja apenas um coeficiente RM associado, ou seja, que a variância explicada por cada amostra seja constante indicando ser irrelevante que não importa qual a observação que é retirada. Recorrendo à representação gráfica por histograma, é avaliada a distribuição dos coeficientes obtidos. Verificou-se que os dados referentes aos coeficientes RM apresentam uma distribuição aproximadamente uniforme. Apesar do coeficiente para o mesmo n (para alguns dos casos) variar, é uma diferença máxima de 0.0002, considerando-se por isso constante. É ainda de salientar que, a título exemplificativo, quando comparados os valores para n=29 e n=2, como seria de esperar, os valores decresceram, neste caso, de 0.85 e de 0.67, respectivamente.

Concluindo, é corroborado que, para cada tamanho de n (a variar de 2 a 29), é irrelevante qual a amostra retirada para o estudo.

4.2.2 Índice Silhouette

Garantir a separação dos grupos é o principal objectivo deste estudo diminuindo p e n sendo que, para avaliar essa separação, tem sido utilizado o índice Silhouette. Por cada análise de componentes principais, o índice Silhouette é calculado para cada objecto e a respectiva média desses valores.

À medida que é reduzido o tamanho da amostra (considerando agora as 7 variáveis definidas anteriormente (ver capítulo 4), é esperado que os grupos cada

vez mais apresentem índices Silhouette significativamente mais pequenos indicando a partir de que valor não se pode reduzir mais sob pena de penalizar a discriminação dos grupos.

Analisando a figura 4.14 é possível observar que, para cada subamostra, os valores médios pouco diferem do índice Silhouette médio para n=30 (0.45). No entanto, ao seguir o raciocínio acima descrito, podia-se reduzir a amostra até, pelo menos, n=6. Todavia, com apenas 6 observações por grupo, a amostra original não é fielmente representada, isto é, a proporção da variância retida decresce para 0.745. Note-se que o comportamento observado na figura 4.14, onde é representada a distribuição (empírica) da média, é o esperado à luz do Teorema Limite Central, isto é, o valor médio é igual (ou muito aproximado) ao da população (0.45) e a variância decresce proporcionalmente com o aumento de n.

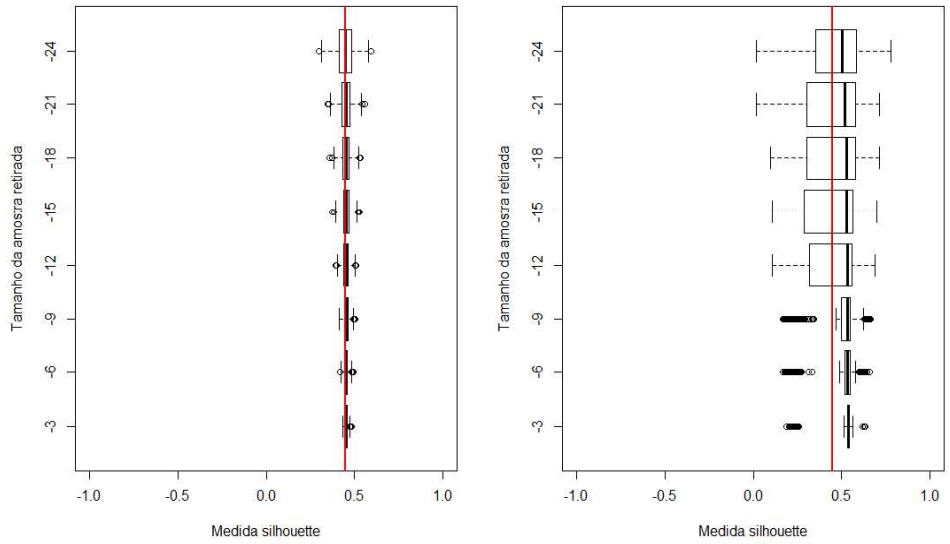


Figura 4.14: Legenda da esquerda para a direita:A-*Boxplot* da média dos índices de Silhouette, das 1000 simulações, para cada redução da amostra:-3,-6,-9,-12,-15,-18,-21,-24. A linha vermelha indica o valor médio do índice de Silhouette para a amostra de p=7 e n=90.B-*Boxplot* dos índices de Silhouette, das 1000 simulações, para cada redução da amostra:-3,-6,-9,-12,-15,-18,-21,-24. A linha vermelha indica o valor médio do índice de Silhouette para a amostra de p=7 e n=90.

Na figura 4.14 B, foram representadas as distribuições de todos os índices de Silhouette (ao invés das médias), para cada n (Figura 4.14 B). Contudo, nenhuma conclusão foi retirada. Para além da presença de bastantes *outliers*, não existe uma variação do n que permita identificar qual o tamanho da amostra mínima adequada. Isto deve-se ao facto de serem retiradas observações de forma aleatória.

Por cada simulação foram retiradas n observações, e foi calculado o índice Silhouette. Poucos foram os casos em que este tomou valores negativo. Apenas quando os objetos da fronteira entre *clusters* foram selecionados para representar os três grupos (o pior dos casos) é que se verificou uma diferença relativamente ao dados originais. Mas, como é visível na figura 4.14, esses casos são raros e mascarados pela maioria dos restantes acontecimentos possíveis.

4.2.3 WSS, coeficiente RM, distância Euclideana e índice Silhouette

A redução do tamanho da amostra depende principalmente da variância explicada pelo subconjunto e da adequada separação dos grupos (T, R e G). Como antes comprovado, avaliar apenas a separação das zonas em estudo, como forma de obter n mínimo adequado, não é suficiente.

A tabela 4.13 resume os indicadores calculados para n entre 2 e 29.

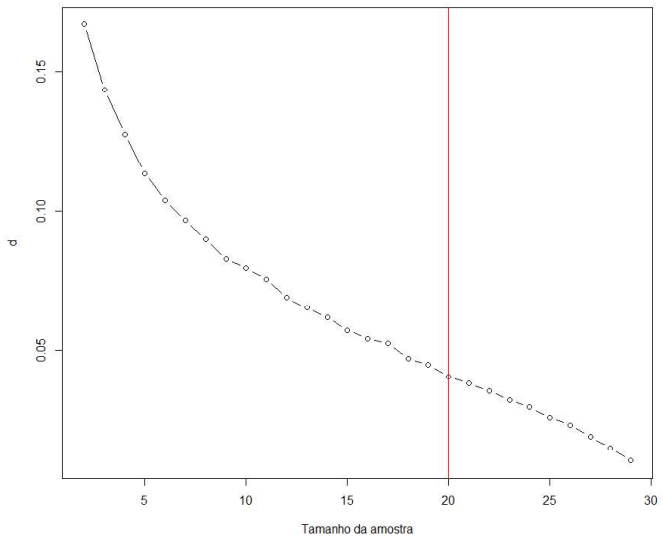


Figura 4.15: Representação gráfica dos valores d (distância média Euclideana entre os centróides obtidos em cada simulação) das 1000 simulações para determinar o tamanho da amostra necessário para distinguir as três zonas em estudo

4.2.3.1 Distância Euclideana

Por cada simulação, o centróide de cada grupo vai mudando. Evidentemente, à medida que o tamanho da amostra diminui o ponto médio do grupo varia mais. Isto implica que a distância entre os sucessivos centróides (para o mesmo n) aumente.

Com esta análise pretendeu-se observar a partir de que n, o valor médio das distâncias euclideanas (d) dos três grupos estabiliza. Quando representados graficamente os valores médios, em função de n, observa-se como esperado, para n=2 a distância média é superior a 0.8 e, quando n aumenta, a média das distâncias diminui para menos que 0.2, aproximando-se do valor 0 (Figura 4.15). Pode-se observar que, embora não sendo atingido um valor constante, os decréscimos são sucessivamente menores sendo, por isso, cada vez menos relevante o aumento do n.

4.2.3.2 WSS

Tal como para a distância Euclideana, para este caso também foi tido em conta os centróides. É conhecido de antemão que se tem 3 grupos e, evidentemente, considera-se três *clusters* quando aplicado o método *k-means*, para os conjuntos de centróides.

Verificou-se que quanto maior a diminuição da amostra, menor é a variação do valor médio ao longo das 1000 simulações, ou seja, os centróides encontram-se menos dispersos (Figura 4.17) para um maior valor de n. Tal como para a distância, também o WSS tem uma diminuição acentuada (Figura 4.16), variando entre 272.479 e 0.288, para n=2 e n=29, respectivamente. Como de esperado, este valor diminui conforme o tamanho da amostra aumenta. Como é observado na figura 4.16, a partir de n=15, os decréscimos de WSS são praticamente constantes. Isto indica que a caracterização do grupo, pelo seu centróide, já não se altera marcadamente com o aumento do n na vizinhança daqueles valores.

Tal como mostrado anteriormente, o índice Silhouette médio pouco ou nada varia para n entre 2 e 29 (entre 0.437 e 0.452; Tabela 4.13). Contudo, pela observação da figura 4.18, o índice médio Silhouette, varia até estabilizar a partir de aproximadamente n=21 atingindo o valor mínimo de 0.452.

O coeficiente de variação do índice Silhouette médio mostra uma diminuição acentuada para $n \leqslant 10$ e parece estabilizar a partir de n=11 (Figura 4.18). No entanto, estes valores também pouco diferirem entre si (Tabela 4.13) .

Quanto ao coeficiente RM, este compara a variância retida para os diferentes n's, com p=7, em comparação à amostra original (n=30 e p=7). Portanto, à medida que a amostra aumenta, esta será uma melhor aproximação dos dados originais e, consequentemente, será associada a um maior coeficiente RM (Figura 4.19).

Assim, considerando o gráfico da figura 4.15 referente ás distâncias médias Euclideanas considera-se a estabilização das distâncias a partir de n=20. Considerando os resultados de WSS (Figura 4.16), este indicador parece estabilizar antes

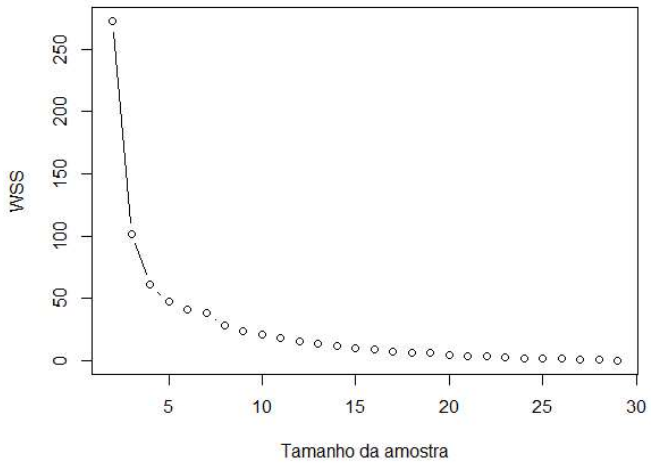
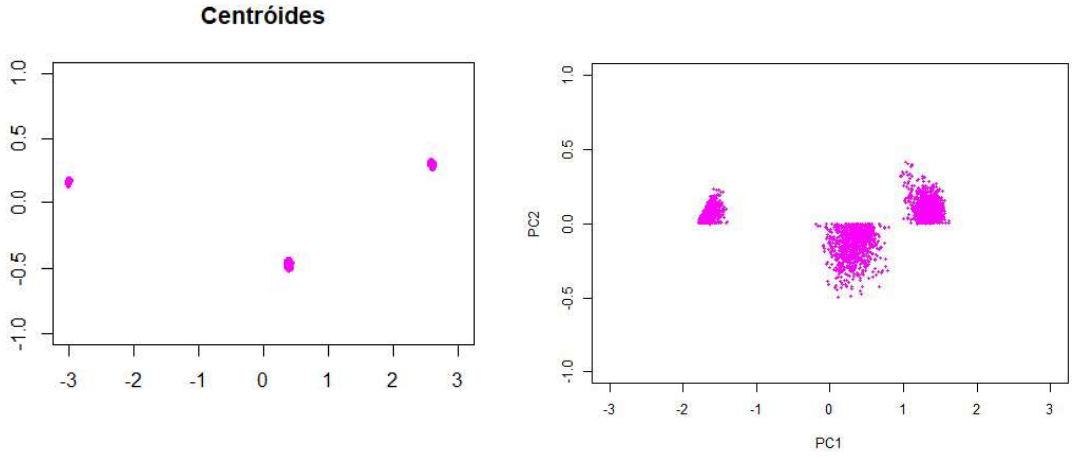


Figura 4.16: Representação gráfica dos valores WSS (*withinsumofsquares* para os centróides obtidos em cada simulação) das 1000 simulações para determinar o tamanho da amostra necessário para distinguir as três zonas em estudo



a - Amostra com n=29 e p=7

b - Amostra com n=2 e p=7

Figura 4.17: Representação gráfica dos centróides obtidos para cada grupo nas 1000 simulações

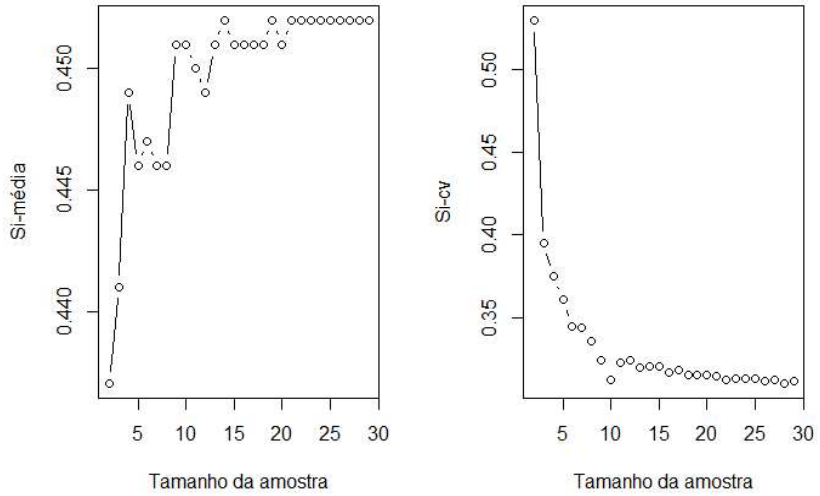


Figura 4.18: Representação gráfica dos índices médios Silhouette e do coeficiente de variação dos mesmos das 1000 simulações para determinar o tamanho da amostra necessário para distinguir as três zonas em estudo

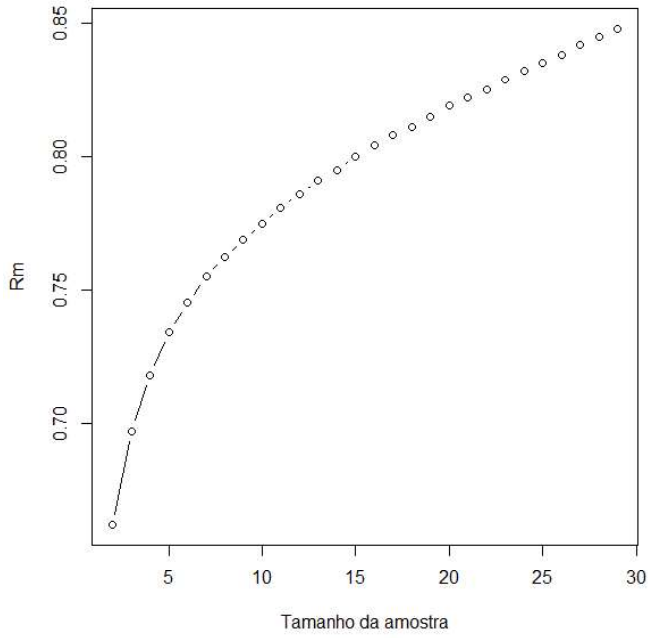


Figura 4.19: Representação gráfica dos coeficientes RM obtidos nas 1000 simulações para determinar o tamanho da amostra necessária para distinguir as três zonas em estudo

(n=15). Comparando os resultados, para n=15 e n=20 (Tabela 4.13):

1. Como já mencionado, o WSS para n=15 apresenta um valor de 10.026 enquanto que, para n=20 este toma o valor de 4.522;

2. O coeficiente de variação (CV) dos $s_{(i)}$ é idêntico: para n=20 obteve-se 0.315 e para n=15 tem-se o valor de 0.320;

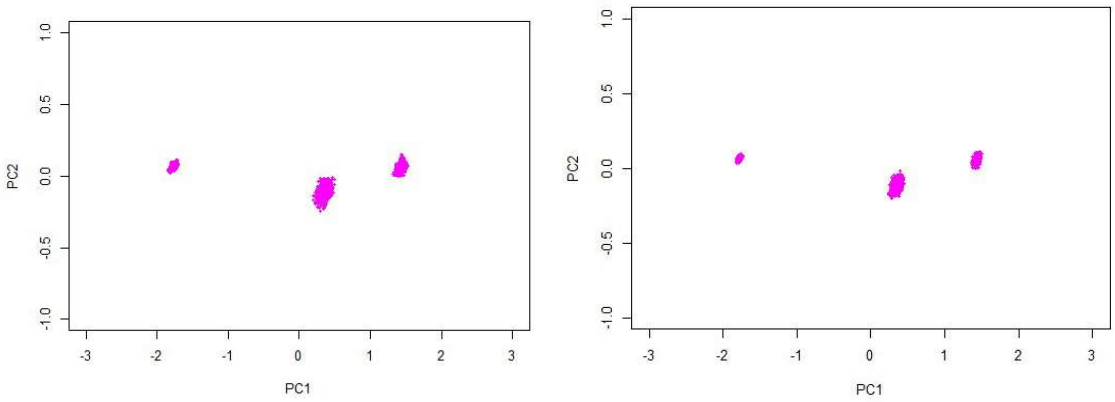
3. O valor médio de $s_{(i)}$ para n=15 e n=20 é igual a 0.451. Indica, portanto, que, independentemente do tamanho escolhido, a qualidade de separação dos grupos é a mesma;

4. O coeficiente RM para n=15 e n=20 é 0.800 e 0.819, respectivamente. Assim, ambos os subconjuntos explicam, aproximadamente, 80% dos dados originais;

5. A distância média Euclideana é de 0.120 e 0.080 para n=15 e n=20, pela respectiva ordem.

Os gráficos da figura 4.20 ilustram a variação dos centróides (por simulação) para n=15 e n=20. Quando comparados os centróides das zonas T, R e G das simulações para n=15 e n=20 (Figura 4.20) consegue-se observar uma maior dispersão dos pontos, como esperado, para n=15, o que corrobora os resultados obtidos para WSS e d. Assim, pela visualização gráfica dos centróides, verifica-se uma dispersão ainda considerada destes valores para n=15 (Figura 4.20). Atendendo os valores de WSS e d, que avaliam a dispersão dos centróides, verifica-se uma diferença considerável, principalmente para os valores de WSS, onde n=15 apresenta mais que o dobro em relação a n=20 (10.026 e 4.522, respectivamente).

Portanto, os resultados resumidos na tabela 4.13 em conjunto com as figuras 4.20 e 4.13, apontam para n=20 (ou na vizinhança de 20) como um tamanho da amostra suficiente para distinguir as populações das Ria de Aveiro e Vigo, e do Estuário do Tejo da espécie *R. philippinarum*, com um índice Silhouette médio de 0.45 e um coeficiente RM de 0.819 explicando mais de 80% da variância original.



a -Amostra com n=15 e p=7

b -Amostra com n=20 e p=7

Figura 4.20: Representação gráfica dos centróides obtidos para cada grupo nas 1000 simulações

Tabela 4.5: Coeficientes $s_{(i)}$ (médio e cv), RM, d e WSS obtidos, em função do n (entre 2 e 29), por simulação. Legenda: n= tamanho da amostra, Si.média= Média dos valores médios do índices Silhouette obtido em cada uma das 1000 simulações, Si.cv= Média dos coeficientes de variação do índices Silhouette obtido em cada uma das 1000 simulações, RM= Média do coeficiente RM obtido em cada uma das 1000 simulações para cada n, d=Média das distâncias euclideanas das 1000 simulações para cada n, WSS= Média do WSS das 1000 simulações para cada n

n	**Si**		**Rm**	**D**	**WSS**
	média	**cv**			
2	0.437	0.530	0.662	0.841	272.479
3	0.441	0.395	0.697	0.396	101.776
4	0,449	0,375	0.718	0.299	60.513
5	0.446	0.361	0.734	0.268	47.043
6	0.447	0.345	0.745	0.241	40.659
7	0.446	0.344	0.755	0.220	37.721
8	0.446	0.336	0.762	0.201	27.921
9	0.451	0.324	0.769	0.184	23.999
10	0.451	0.312	0.775	0.172	20.643
11	0.450	0.322	0.781	0.161	18.359
12	0.449	0.324	0.786	0.161	15.415
13	0.451	0.319	0.791	0.139	13.429
14	0.452	0.320	0.795	0.125	11.659
15	0.451	0.320	0.800	0.120	10.026
16	0.451	0.316	0.804	0.111	8.631
17	0.451	0.318	0.808	0.105	6.866
18	0.451	0.315	0.811	0.095	6.321
19	0.452	0.315	0.815	0.087	6.321
20	0.451	0.315	0.819	0.080	4.522
21	0.452	0.314	0.822	0.075	3.619
22	0.452	0.312	0.825	0.068	3.166
23	0.452	0.313	0.829	0.061	2.560
24	0.452	0.313	0.832	0.057	2.161
25	0.452	0.313	0.835	0.049	1.666
26	0.452	0.311	0.838	0.043	1.309
27	0.452	0.312	0.842	0.037	0.892
28	0.452	0.310	0.845	0.030	0.615
29	0.452	0.311	0.848	0.020	0.288

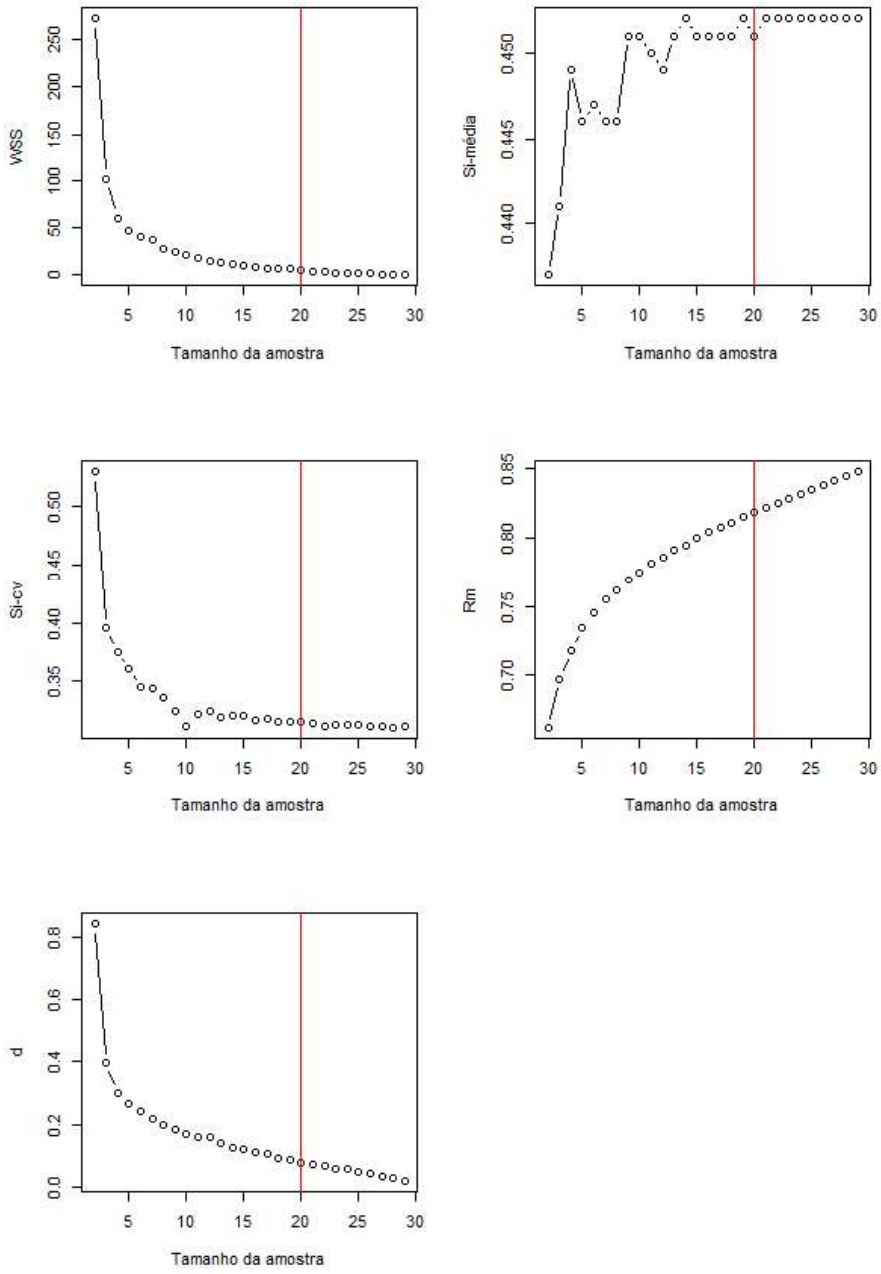


Figura 4.21: Representação gráfica dos diferentes valores obtidos nas simulações para determinar o tamanho da amostra necessário para distinguir as três zonas em estudo. Legenda: Si-média= Média dos valores médios do índices Silhouette obtido em cada uma das 1000 simulações, Si-cv= Média dos coeficientes de variação do índices Silhouette obtido em cada uma das 1000 simulações, RM= Média do coeficiente RM obtido em cada uma das 1000 simulações, d= Média das distâncias euclideanas das 1000 simulações para cada n, WSS= Média do WSS das 1000 simulações

Capítulo 5

Conclusões

5.1 Conclusões gerais

Nesta dissertação foi abordado o tema de alteração da estrutura de um conjunto de dados organizados em *clusters*/grupos de forma a (1) minimizar o esforço de amostragem quer em termos de quantidades avaliadas (variáveis) quer relativamente ao número de observações (dimensão amostral) e (2) manter a capacidade descriminante dos grupos do conjunto de dados original. Em concreto, neste estudo foi trabalhado um conjunto de dados referente à espécie a *Ruditapes philippinarum*. Conhecida por ameijoa japónica, esta é uma espécie com valor económico e social importante a nível nacional, o que despoletou o interesse e necessidade de criação técnicas que permitam a identificação de habitat de diferentes populações da mesma e, desta forma, proteger o interesse dos consumidores.

Vários estudos foram realizados para a rastreabilidade de bivalves a partir da ácidos gordos presentes no músculo adutor dos elementos presentes nas conchas. Em particular, este estudo visou identificar quais os elementos e/ou ácidos gordos (variáveis) são necessários e suficientes para garantir a discriminação da origem da espécie (Ria de Aveiro, Ria de Vigo e Estuário do Tejo). Utilizaram-se a análise de componentes principais (PCA) e a análise k-means em conjunto com o cálculo dos índices: silhouette, coeficiente RM, WSS e distância Euclideana entre centróides.

Para a redução do número de variáveis partiu-se da interpretação dos pesos factoriais da análise de componentes principais. As duas primeiras componentes principais permitiram visualizar os diferentes grupos, em R^2. A partir dos índices Silhouette, foi possível reduzir as 44 variáveis (18 elementos e 26 ácidos gordos)

em 84% (7 variáveis), aproximadamente. As 7 variáveis incluem ácidos gordos (em particular, 16:1n -7; 18:1n -7; 20:1n -7; 20:4n -3; 20:5n -3; 22:3n -6 e 22:6n).

Estes resultados foram obtidos a partir da análise dos índices Silhouette. Quando observado a quantidade de variância retida por cada redução de variável, a partir do coeficiente RM, os resultados diferem. Como esperado, quando considerado o coeficiente RM mínimo de 0.95, aproximadamente, o número de variáveis aumenta para 25 (43%). No entanto, o objectivo é distinguir a separação dos diferentes grupos. Com apenas 7 variáveis, essa distinção é possível e ainda apresenta um coeficiente RM de 0.7.

Na redução do tamanho da amostra inicial recorreu-se à simulação. A aplicação da análise de componentes principais foi também usada para observar gráficamente os diferentes grupos e inspecionar visualmente a sua separação. Ao contrário do alcançado para a redução do número de variáveis, apenas com o índice Silhouette e coeficiente RM, não foi possível definir uma dimensão amostral adequada. No entanto, a partir do resultado das 1000 simulações, em que foi obtido o centróide de cada grupo e calculado o valor médio de WSS e da distância Euclideana, entre cada objecto por grupo, foi possível definir um valor para n. Em conjunto com a visualização gráfica dos centróides e dos resultados obtidos com WSS, distância Euclideana, índice Silhouette e coeficiente RM, concluiu-se que a dimensão n=20 (ou na sua vizinhança) será necessário e suficiente para simultaneamente reproduzir adequadamente a informação inicial e garantir a separação dos grupos.

5.2 Trabalho futuro

Seria interessante aplicar diferentes técnicas, que apresentam alternativas de visualização gráfica ao PCA. A aplicação da técnica de análise de componentes principais, apesar de ser a mais utilizada, não é a mais eficaz para o objetivo de visualização de grupos. Porém, a partir da aplicação desta foi possível alcançar os objetivos propostos e observar resultados interessantes e importantes para este tema.

De algumas dessas técnicas, são de salientar duas:

- **t-SNE**: Desenvolvida por Geoffrey Hinton e Laurens van der Maaten em 2008 (Laurens et al., 2008). A técnica t-*Distributed Stochastic Neighbor Embedding* (t-SNE) é também utilizada para redução de dimensão dos dados, permitindo uma visualização de dados por grupo muito eficaz.

- **UMAP**: Desenvolvida por Leland McInnes, John Healy, James Melville o ano passado, 2018 (McInnes et al., 2018), com principal aplicação em aprendizagem automática. A *Uniform Manifold Approximation and Projection*, tal como PCA e t-SNE, é utilizada para a redução de dimensão dos dados. Este algoritmo apresenta resultados competitivos com os obtidos a partir da aplicação do t-SNE, exibindo melhores visualizações gráficas. Tal como o t-SNE, este algoritmo une os objetos de forma a criar grupos distintos e, de forma a distinguir melhor os grupos, este também separa melhor os grupos entre si, afastando-os.

Capítulo 6

Referências

- Bodoy, A., Maître-Allain, T. e Riva, A., (1980). Croissance comparée de la palourde européenne Ruditapes decussatus et de la palourde japonaise Ruditapes philippinarum dans un écosystème artificiel méditerranéen. Vie Mar. 2 39- 51 39–52;
- Cadima, J., Cerdeira, J.O., Silva, P.D., Minhoto, M., (2018). The subselect R package 1–35;
- Chainho, P. et al., (2010). Long-Term Trends in Intertidal and Subtidal Benthic Communities in Response to Water Quality Improvement Measures. *Estuaries and Coasts*, 33(6), pp.1314–1326;
- Estatísticas da Pesca, (2017);
- Everitt, S.,B. e Landau, Sabine e M. Leese, M e Stahl, Daniel.,(2011). *Cluster Analysis*. 10.1002/9780470977811.ch8.
- Gaspar, M.B., (2010). Distribuição, abundância e estrutura demográfica da amêijoajaponesa (*Ruditapes philippinarum*) no Rio Tejo. Relatório do IPIMAR, 6 pp;
- Godfray, H.C.J., (2010). Food Security: The Challenge of Feeding 9 Billion People Food Security: The Challenge of Feeding 9 Billion People ;
- Hotelling, H., (1933). Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 24(6), 417-441;

- Johnson, Richard A. e Wichern, D.W., (2007). Applied multivariate statistical analysis, 6th ed;
- Jolliffe, I.T., (1986). Principal Component Analysis, Second Edition;
- Karnjanapratum, S., Benjakul, S., Kishimura, H., Tsai, Y., (2013). Chemical compositions and nutritional value of Asian hard clam (Meretrix lusoria) from the coast of Andaman Sea. Food Chem. 141, 4138–4145;
- Kevin D., (2019). PID: Process Improvement using Data 323–414;
- Laurens van der van der, M. e Geoffrey, H., (2008). Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research :2579-2605, 2008
- Macqueen, J., (1967). Some methods for classification and analysis of multivariate observations 233, 281–297;
- McInnes, L., Healy, J., Melville, J., (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.
- Melià, P. e Gatto, M., (2005). A stochastic bioeconomic model for the management of clam farming. Ecol. Model., 184(1): 163-174;
- Montilivi, C., (2019). Challenging the links between seafood and human health in the context of global change 96, 29–42;
- Pearce, J.G., Shaar, Z., Crosbie, R.E., (1977). Scattering of energetic ions by solids — a simulation. Simulation 29, 97–104;
- Pearson, K., (1901). On lines and planes of closest fit to systems of points in space. 559–572;
- R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/;
- Ricardo, F., Pimentel, T., Maciel, E., Moreira, A.S.P., Rosário Domingues, M., Calado, R., (2017). Fatty acid dynamics of the adductor muscle of live cockles (*Cerastoderma edule*) during their shelf-life and its relevance for traceability of geographic origin. Food Control 77, 192–198;
- Rubanov, P., Vasylieva, T., Lyeonov, S., Pokhylko, S., (2019). Cluster analysis of development of alternative finance models depending on the regional affiliation of countries 90–107;

• Scarlato, O. A., (1981). Bivalves of temperate waters of the northwestern part of the Pacific Ocean. Leningrad, Russia: Nauka Press. 408 pp

Anexo I

Apêndice A- Gráficos complementares no estudo da redução do número de variáveis

I.1 Ácidos gordos

Representação gráfica dos resultados da análise de componentes principais da amostra representada apenas com as variáveis agk, $k = 1,...26$. Cada um dos grupos representados por pontos: (azul claro)- Estuário do Tejo (T), (azul escuro)- Ria de Aveiro (R), (rosa)- Ria de Vigo (G). Cada centróide está representado por um triângulo (preto- grupo T, rosa- grupo R, verde- grupo G).

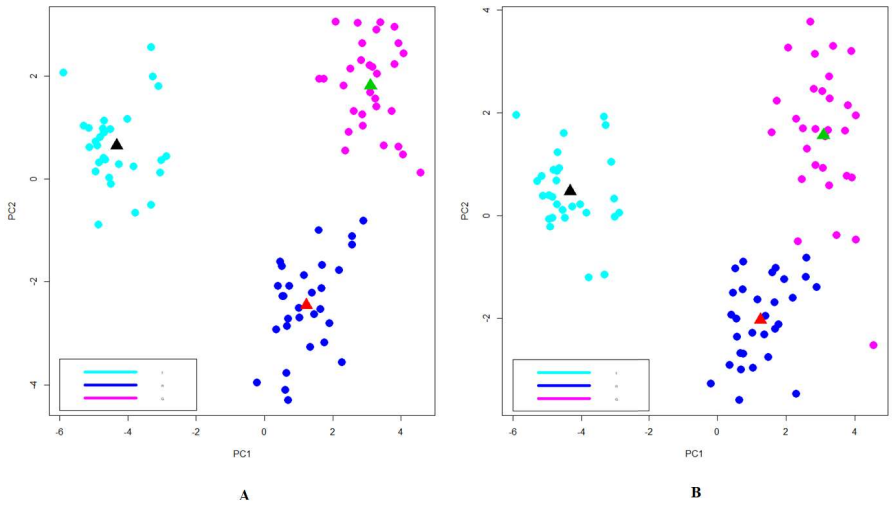


Figura I.1: Legenda: A- PCA da amostra representada pelos ácidos gordos (26 variáveis), B- PCA da amostra representada pelos ácidos gordos (25 variáveis, sem $ag11$)

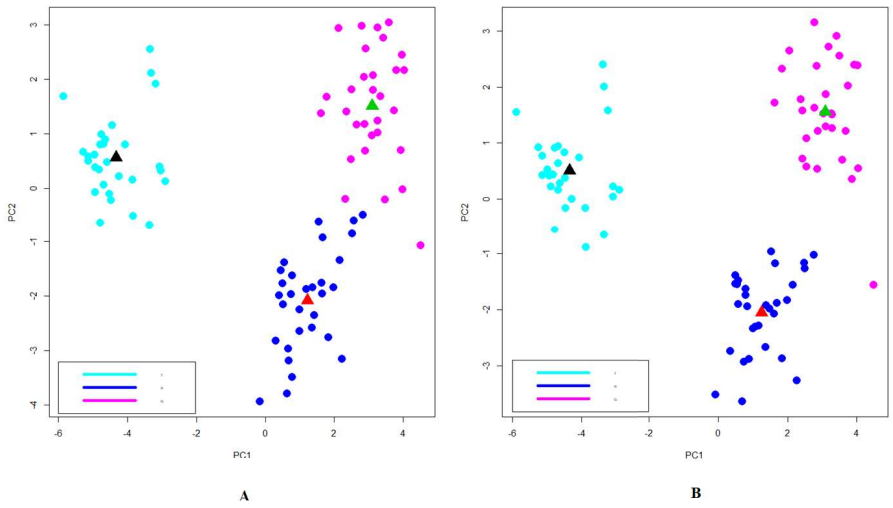


Figura I.2: Legenda: A- PCA da amostra representada pelos ácidos gordos (24 variáveis, sem $ag21$), B- PCA da amostra representada pelos ácidos gordos (23 variáveis, sem $ag15$)

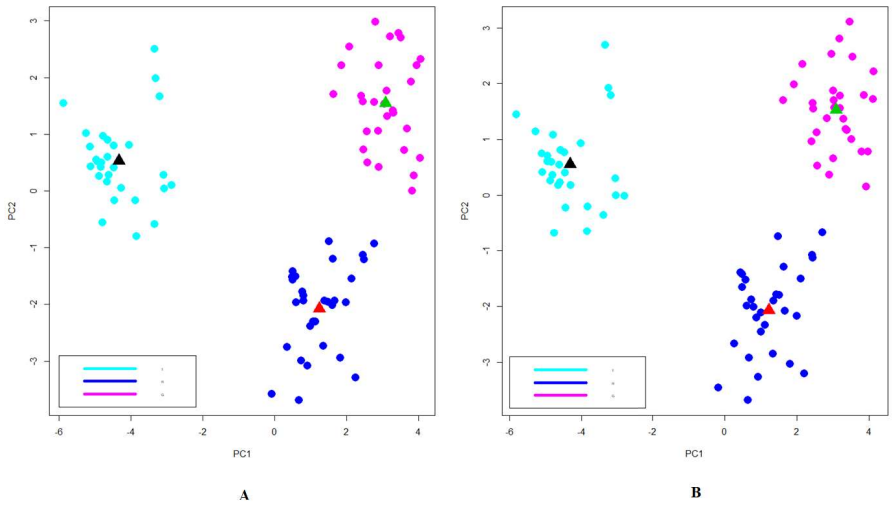


Figura I.3: Legenda: A- PCA da amostra representada pelos ácidos gordos (22 variáveis, sem $ag4$), B- PCA da amostra representada pelos ácidos gordos (21 variáveis, sem $ag3$)

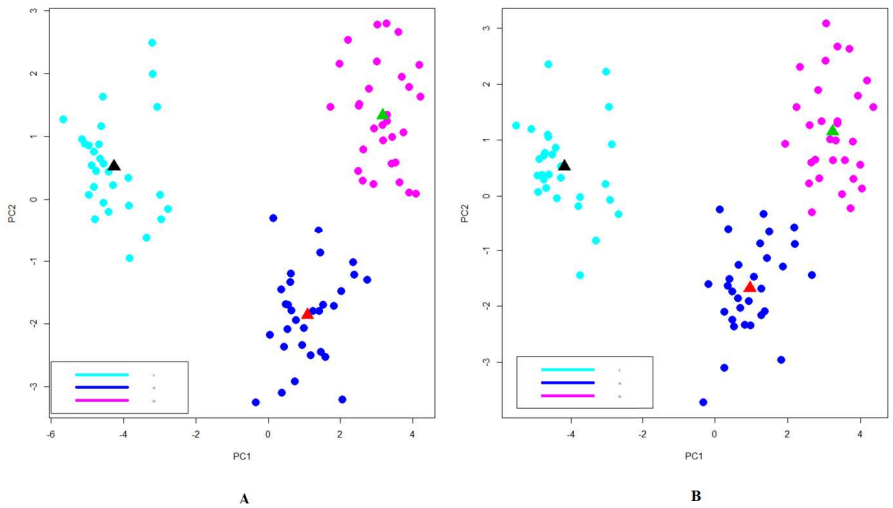


Figura I.4: Legenda: A- PCA da amostra representada pelos ácidos gordos (20 variáveis, sem $ag12$), B- PCA da amostra representada pelos ácidos gordos (19 variáveis, sem $ag10$)

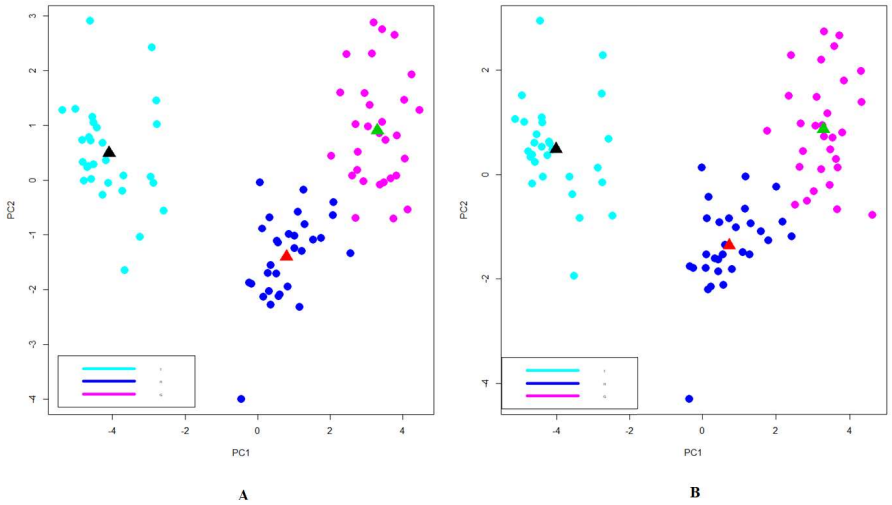


Figura I.5: Legenda: A- PCA da amostra representada pelos ácidos gordos (18 variáveis, sem $ag = 24$), B- PCA da amostra representada pelos ácidos gordos (17 variáveis, sem $ag13$)

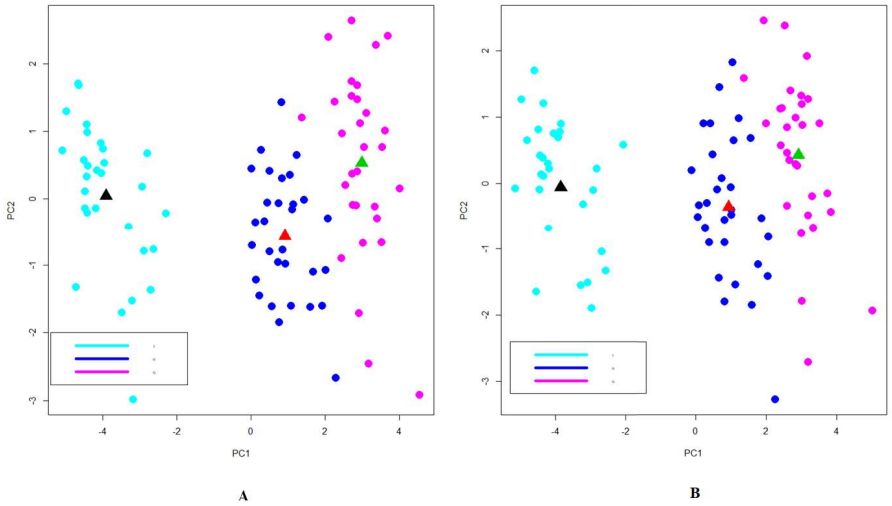


Figura I.6: Legenda: A- PCA da amostra representada pelos ácidos gordos (16 variáveis, sem $ag = 7$), B- PCA da amostra representada pelos ácidos gordos (15 variáveis, sem $ag8$)

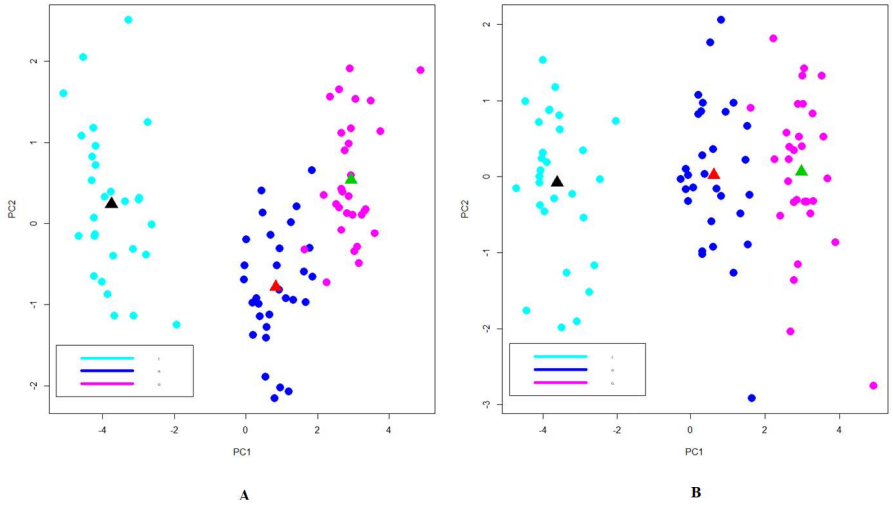


Figura I.7: Legenda: A- PCA da amostra representada pelos ácidos gordos (14 variáveis, sem $ag = 16$), B- PCA da amostra representada pelos ácidos gordos (13 variáveis, sem $ag2$)

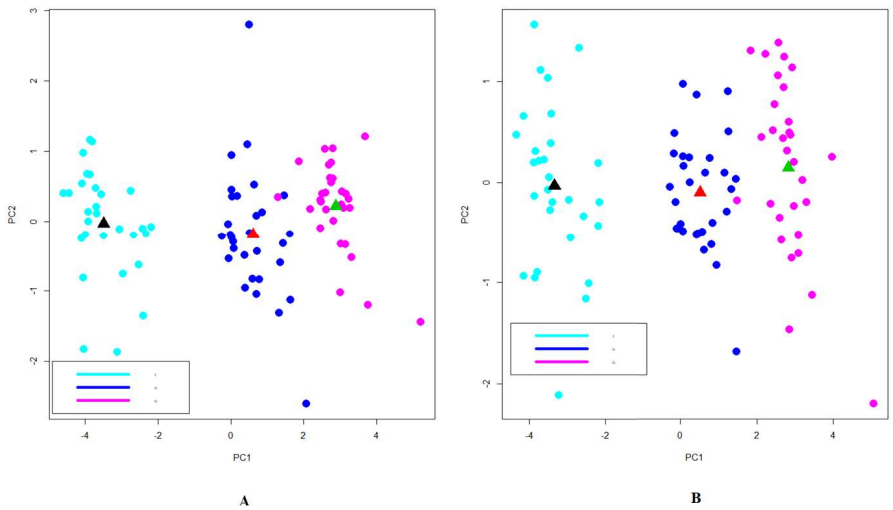


Figura I.8: Legenda: A- PCA da amostra representada pelos ácidos gordos (12 variáveis, sem $ag = 17$), B- PCA da amostra representada pelos ácidos gordos (11 variáveis, sem $ag1$)

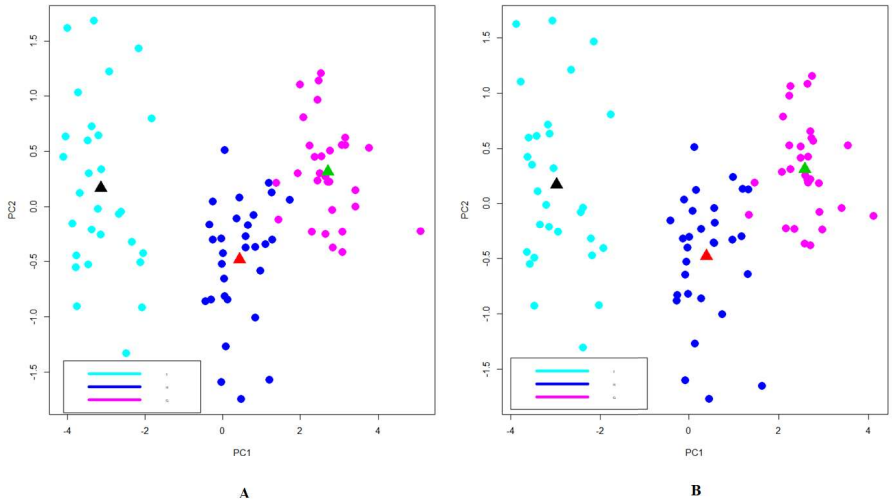


Figura I.9: Legenda: A- PCA da amostra representada pelos ácidos gordos (10 variáveis, sem $ag = 23$), B- PCA da amostra representada pelos ácidos gordos (9 variáveis, sem $ag6$)

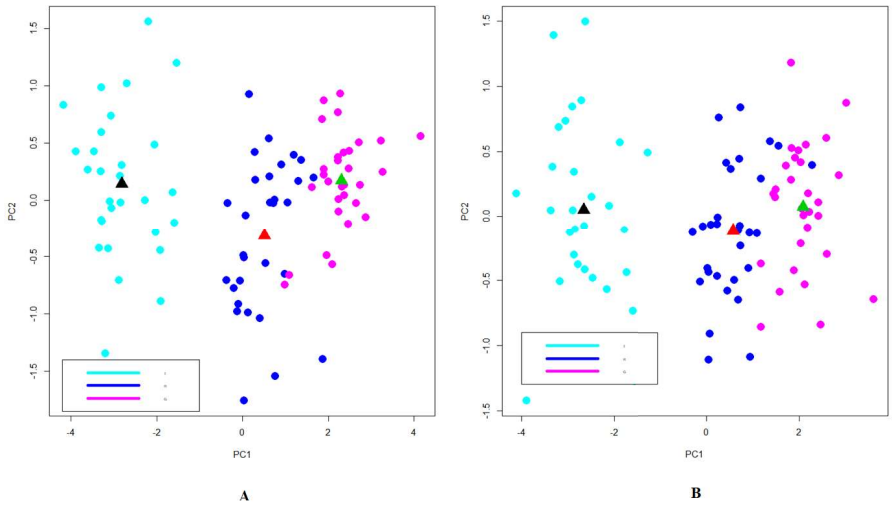


Figura I.10: Legenda: A- PCA da amostra representada pelos ácidos gordos (8 variáveis, sem $ag = 25$), B- PCA da amostra representada pelos ácidos gordos (7 variáveis, sem $ag21$)

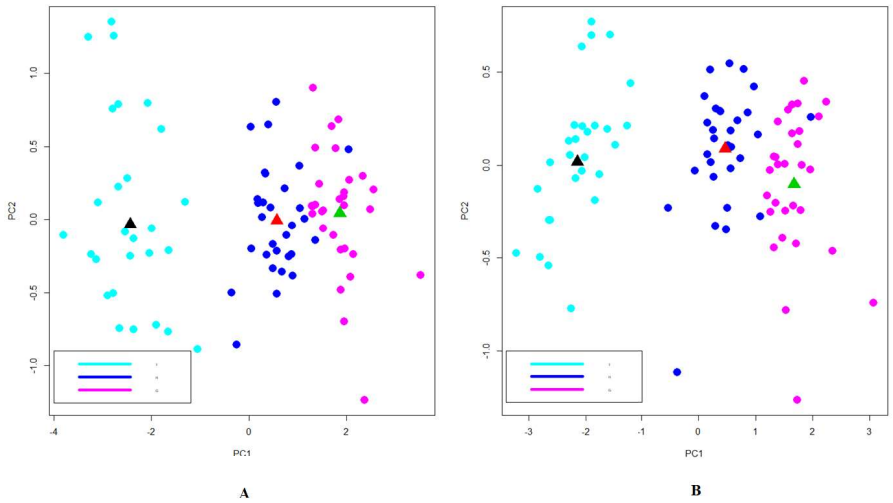


Figura I.11: Legenda: A- PCA da amostra representada pelos ácidos gordos (6 variáveis, sem $ag = 22$), B- PCA da amostra representada pelos ácidos gordos (5 variáveis, sem $ag18$)

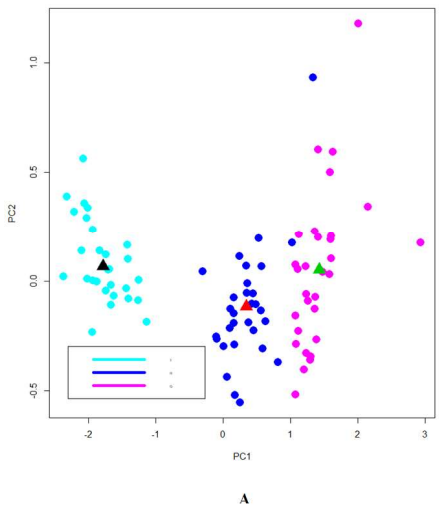


Figura I.12: Legenda: A- PCA da amostra representada pelos ácidos gordos (4 variáveis, sem $ag = 14$), B- PCA da amostra representada pelos ácidos gordos (3 variáveis, sem $ag5$)

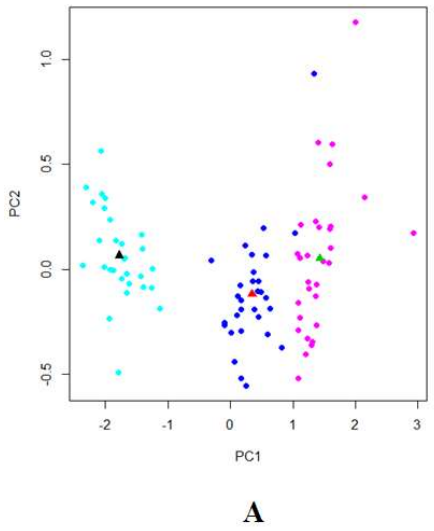


Figura I.13: Legenda: A- PCA da amostra representada pelos ácidos gordos (2 variáveis, sem $ag = 9$)

I.2 Elementos

Representação gráfica dos resultados da análise de componentes principais da amostra representada apenas com as variáveis elk, $k = 1,...18$. Cada um dos grupos representados por pontos: (azul claro)- Estuário do Tejo (T), (azul escuro)- Ria de Aveiro (R), (rosa)- Ria de Vigo (G). Cada centróide está representado por um triângulo (preto- grupo T, rosa- grupo R, verde- grupo G).

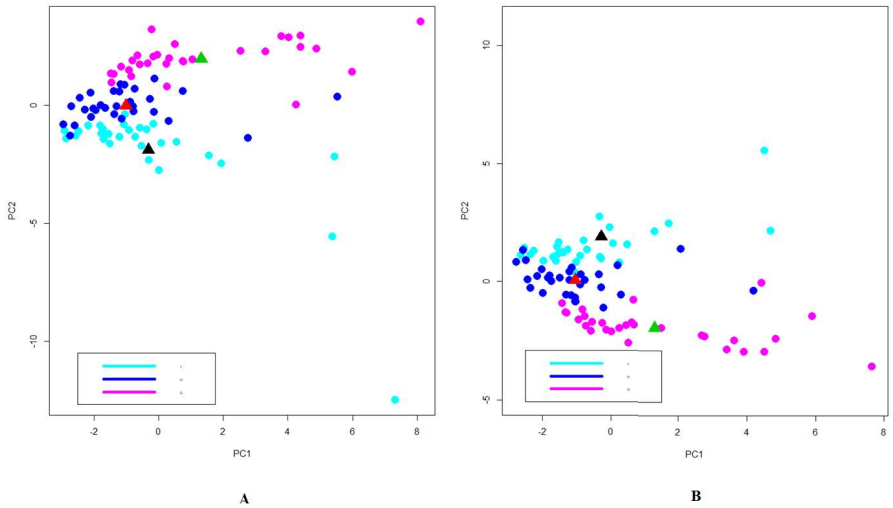


Figura I.14: Legenda: A- PCA da amostra representada pelos elementos (18 variáveis), B- PCA da amostra representada pelos ácidos gordos (17 variáveis, sem $el15$)

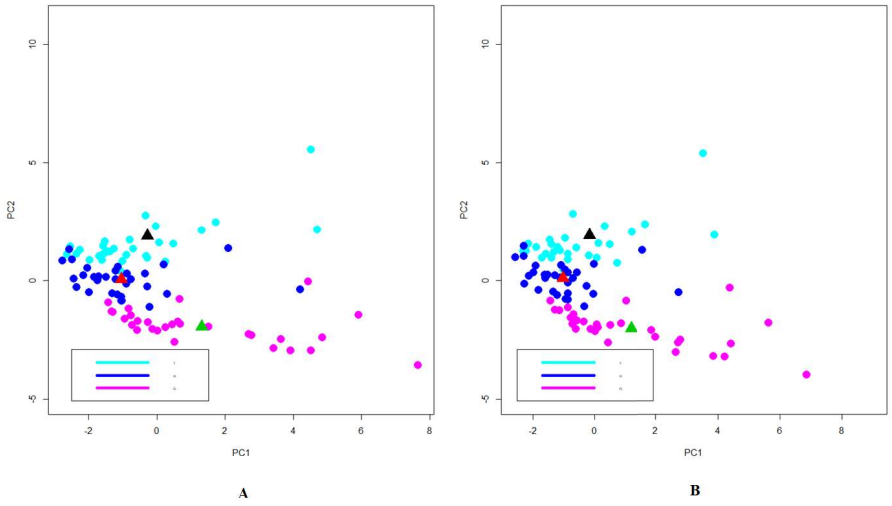


Figura I.15: Legenda: A- PCA da amostra representada pelos elementos (16 variáveis, sem *el*8 variáveis), B- PCA da amostra representada pelos ácidos gordos (15 variáveis, sem *el*14)

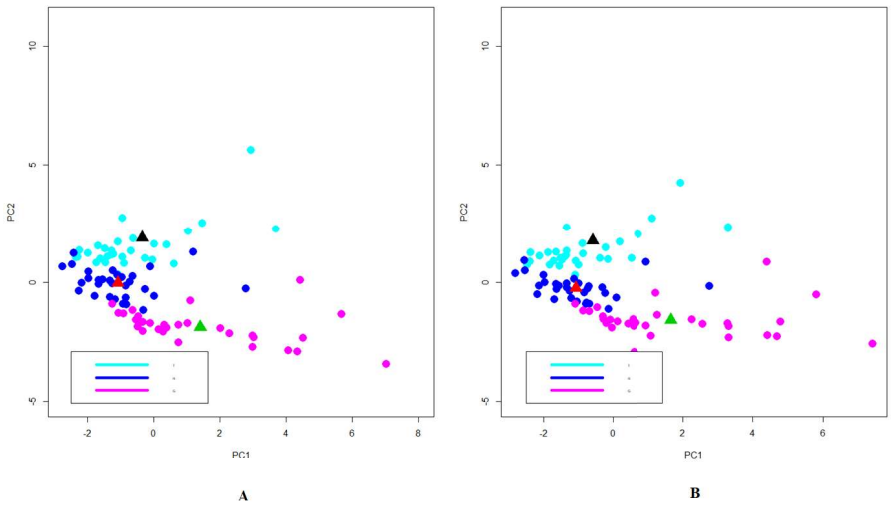


Figura I.16: Legenda: A- PCA da amostra representada pelos elementos (14 variáveis, sem *el*13 variáveis), B- PCA da amostra representada pelos ácidos gordos (13 variáveis, sem *e*3)

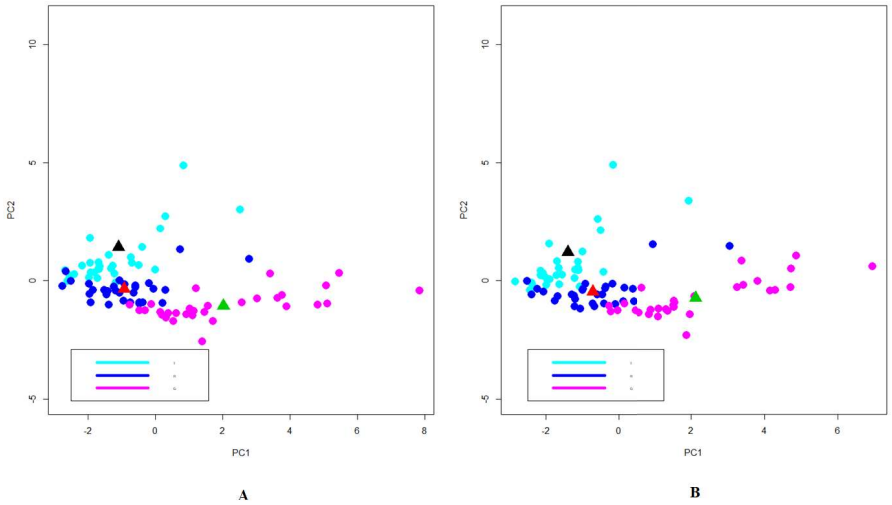


Figura I.17: Legenda: A- PCA da amostra representada pelos elementos (12 variáveis, sem *el*5 variáveis), B- PCA da amostra representada pelos ácidos gordos (11 variáveis, sem *el*4)

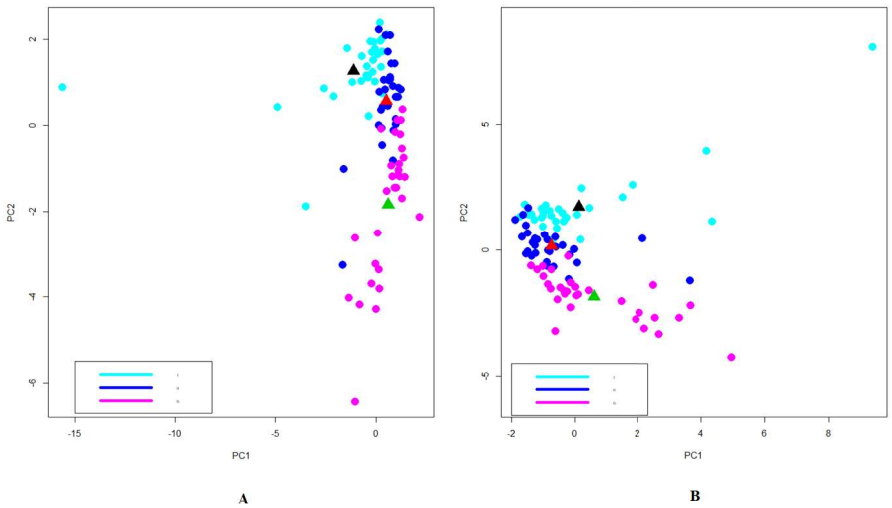


Figura I.18: Legenda: A- PCA da amostra representada pelos elementos (10 variáveis, sem *el*2 variáveis), B- PCA da amostra representada pelos ácidos gordos (9 variáveis, sem *el*9)

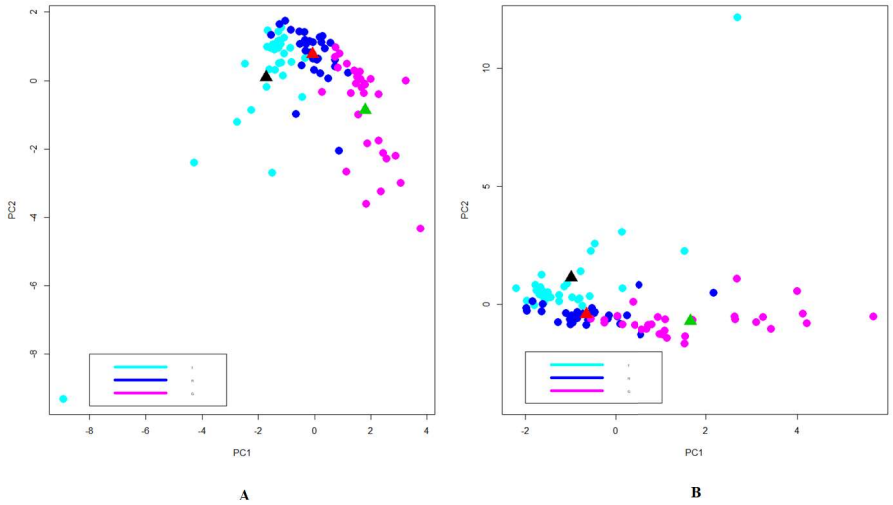


Figura I.19: Legenda: A- PCA da amostra representada pelos elementos (8 variáveis, sem *el*6 variáveis), B- PCA da amostra representada pelos ácidos gordos (7 variáveis, sem *el*2)

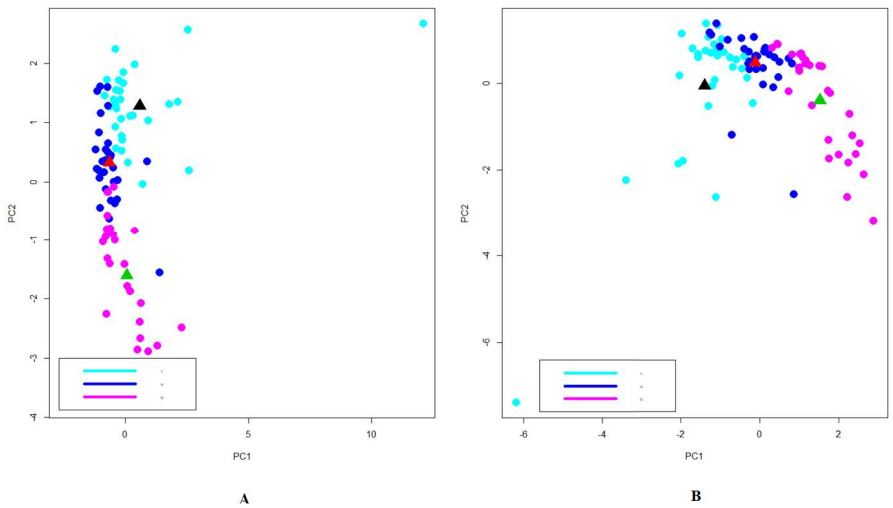


Figura I.20: Legenda: A- PCA da amostra representada pelos elementos (6 variáveis, sem *el*8 variáveis), B- PCA da amostra representada pelos ácidos gordos (5 variáveis, sem *el*7)

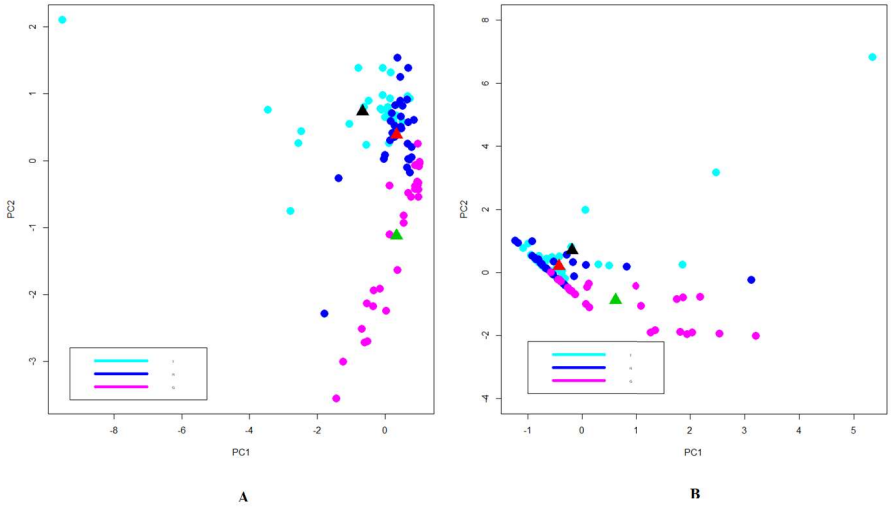


Figura I.21: Legenda: A- PCA da amostra representada pelos elementos (4 variáveis, sem *el*1 variáveis), B- PCA da amostra representada pelos ácidos gordos (3 variáveis, sem *el*6)

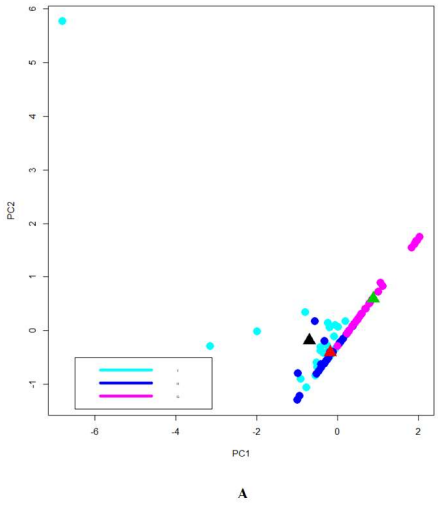


Figura I.22: Legenda: A- PCA da amostra representada pelos elementos (2 variáveis, sem *el*17 variáveis)

I.3 Todas as variáveis

Representação gráfica dos resultados da análise de componentes principais da amostra representada com as variáveis em estudo ($el(i)$ e $ag(k)$, $i = 1,...,18$ e $k = 1,...,26$). Cada um dos grupos representados por pontos: (azul claro)- Estuário do Tejo (T), (azul escuro)- Ria de Aveiro (R), (rosa)- Ria de Vigo (G). Cada centróide está representado por um triângulo (preto- grupo T, rosa- grupo R, verde- grupo G).

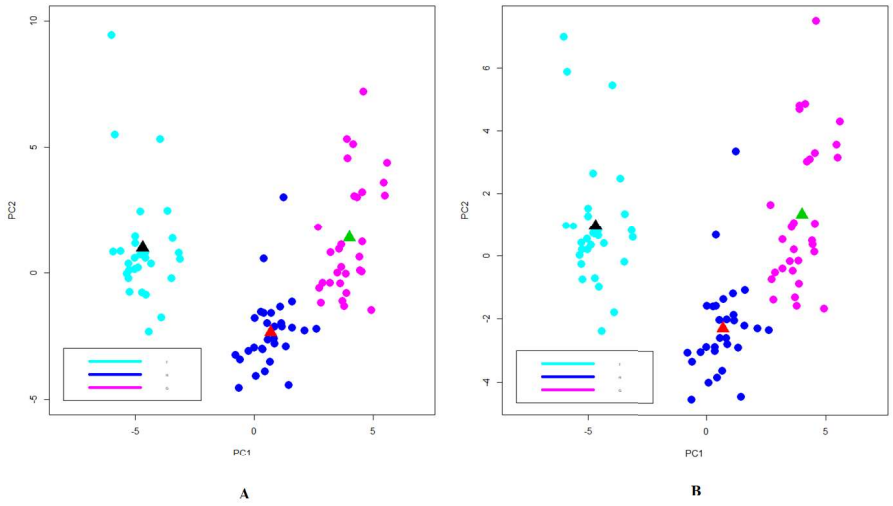


Figura I.23: Legenda: A- PCA da amostra com as variáveis em estudo (44 variáveis) , B- PCA da amostra com as variáveis em estudo (43 variáveis, sem)

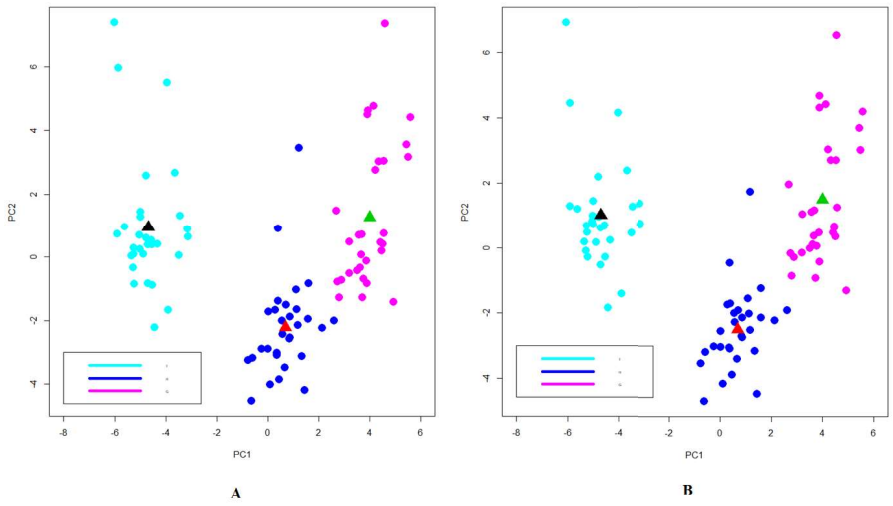


Figura I.24: A- PCA da amostra com as variáveis em estudo (42 variáveis, sem) , B- PCA da amostra com as variáveis em estudo (41 variáveis, sem $el17$)

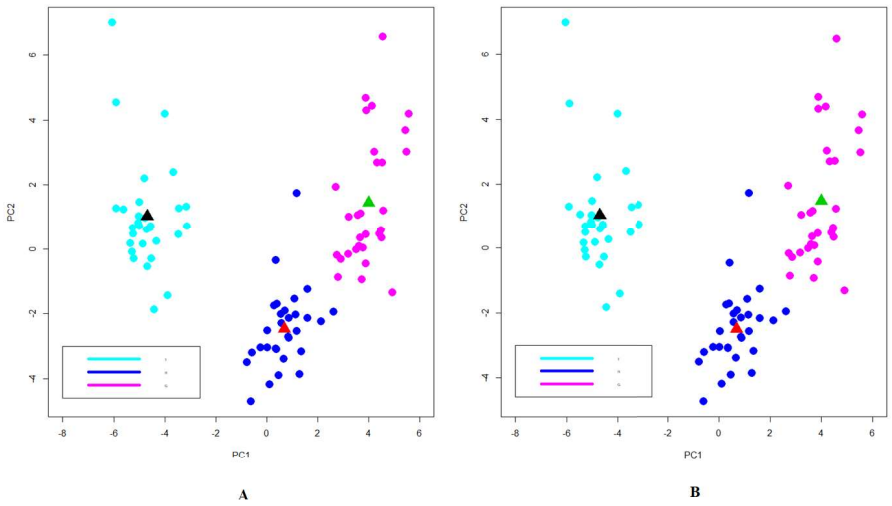


Figura I.25: A- PCA da amostra com as variáveis em estudo (40 variáveis, sem *ag*21) , B- PCA da amostra com as variáveis em estudo (39 variáveis, sem *el*16)

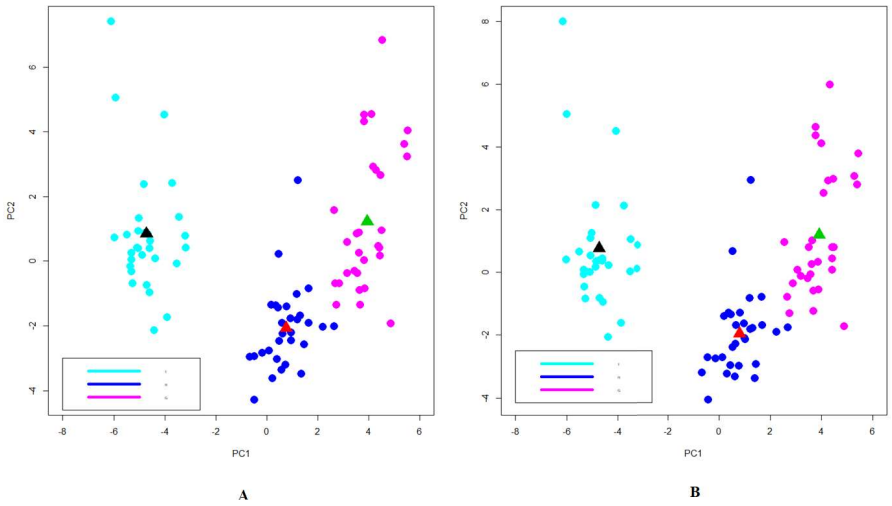


Figura I.26: A- PCA da amostra com as variáveis em estudo (38 variáveis, sem *el*8) , B- PCA da amostra com as variáveis em estudo (37 variáveis, sem *el*13)

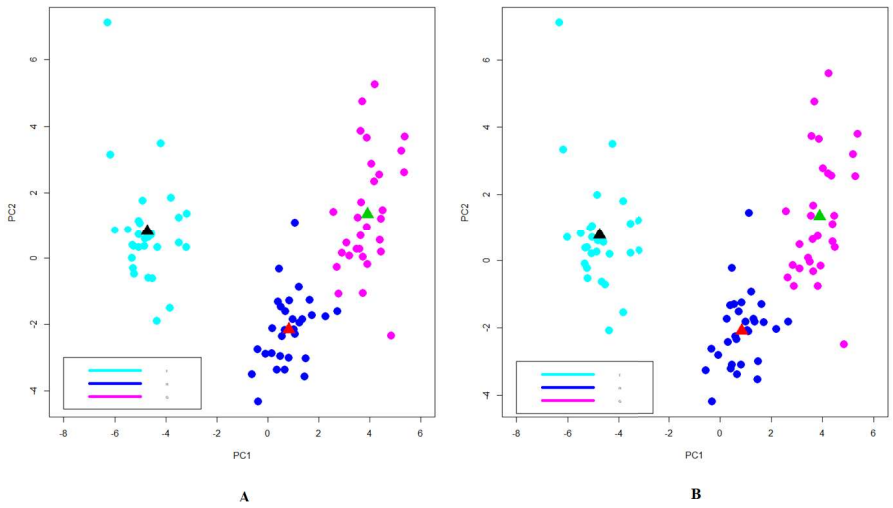


Figura I.27: A- PCA da amostra com as variáveis em estudo (36 variáveis, sem *ag*11) , B- PCA da amostra com as variáveis em estudo (35 variáveis, sem *el*14)

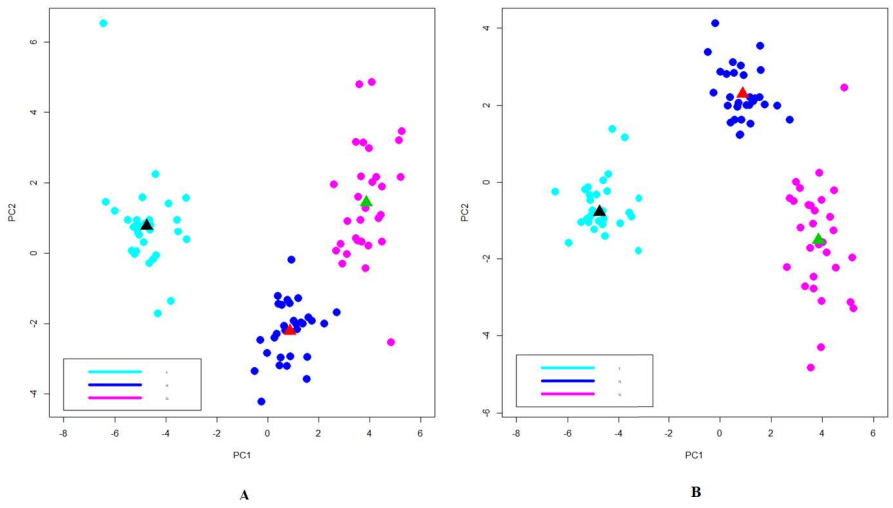


Figura I.28: A- PCA da amostra com as variáveis em estudo (34 variáveis, sem *el*17) , B- PCA da amostra com as variáveis em estudo (33 variáveis, sem *ag*15)

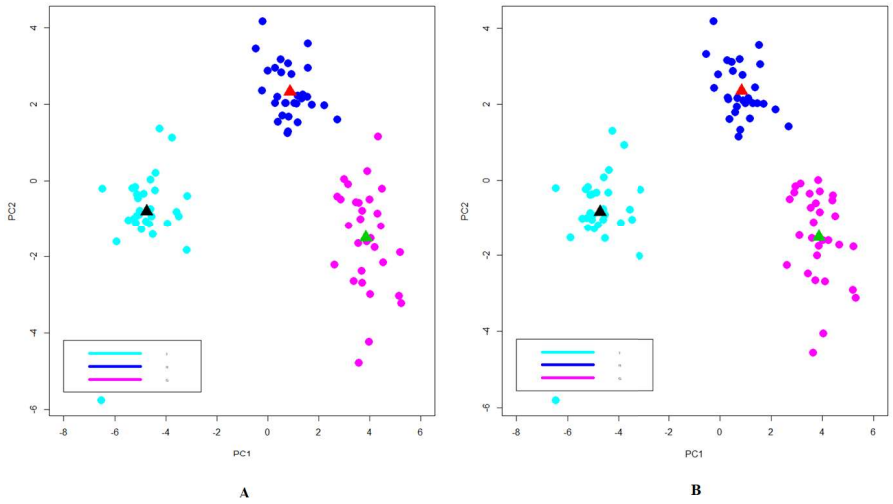


Figura I.29: A- PCA da amostra com as variáveis em estudo (32 variáveis, sem *el*14) , B- PCA da amostra com as variáveis em estudo (31 variáveis, sem *el*15)

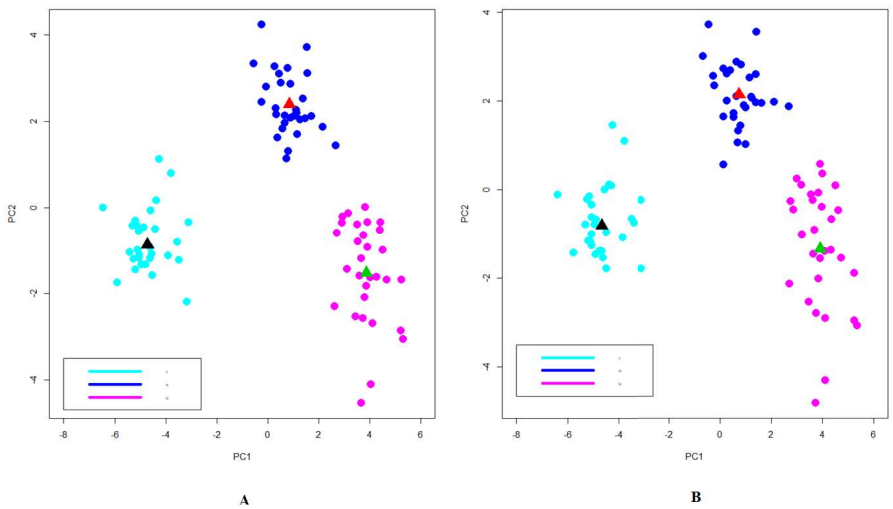


Figura I.30: A- PCA da amostra com as variáveis em estudo (30 variáveis, sem *ag*4) , B- PCA da amostra com as variáveis em estudo (29 variáveis, sem *ag*3)

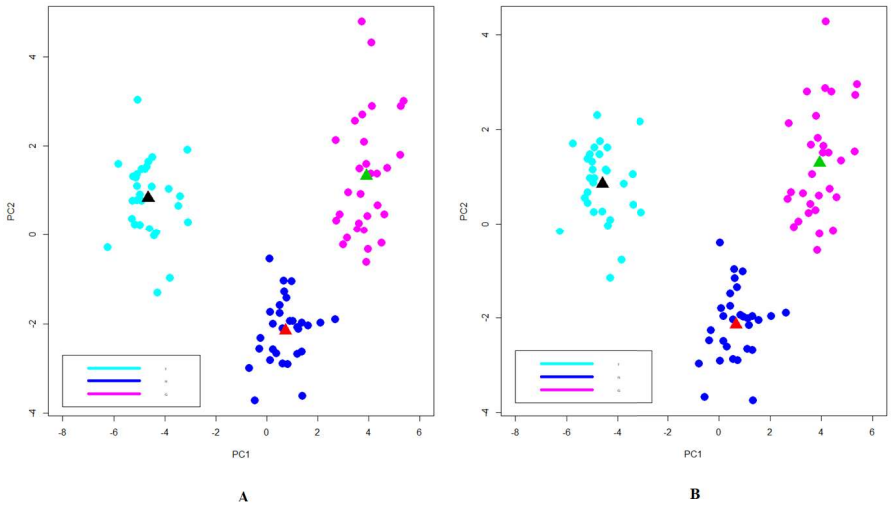


Figura I.31: A- PCA da amostra com as variáveis em estudo (28 variáveis, sem $el19$) , B- PCA da amostra com as variáveis em estudo (27 variáveis, sem $el12$)

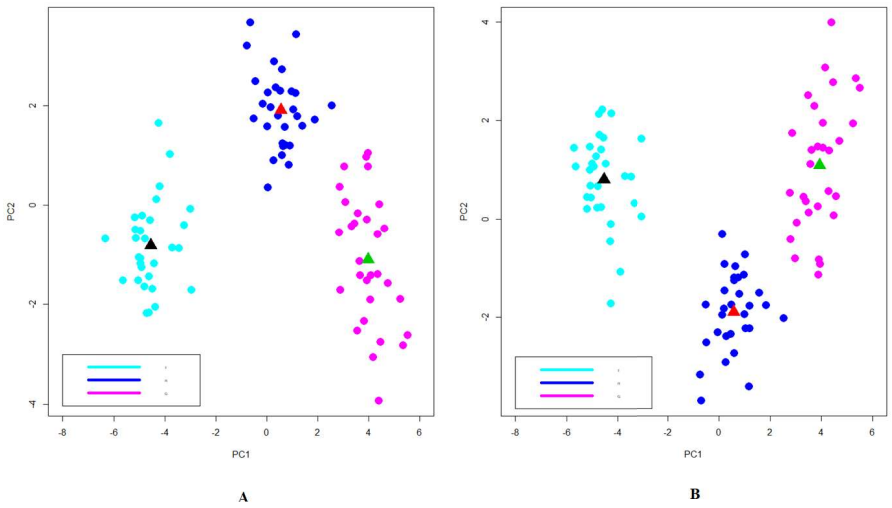


Figura I.32: A- PCA da amostra com as variáveis em estudo (26 variáveis, sem $el10$) , B- PCA da amostra com as variáveis em estudo (25 variáveis, sem $el15$)

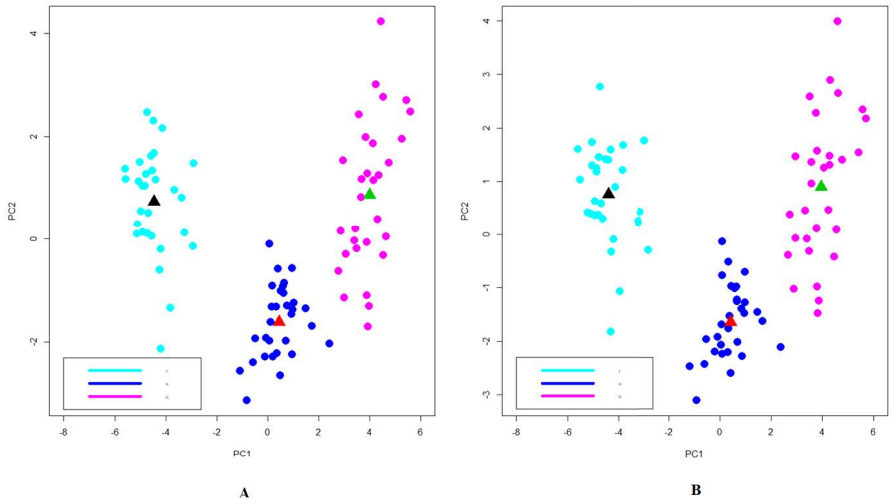


Figura I.33: A- PCA da amostra com as variáveis em estudo (24 variáveis, sem $ag24$) , B- PCA da amostra com as variáveis em estudo (23 variáveis, sem $el16$)

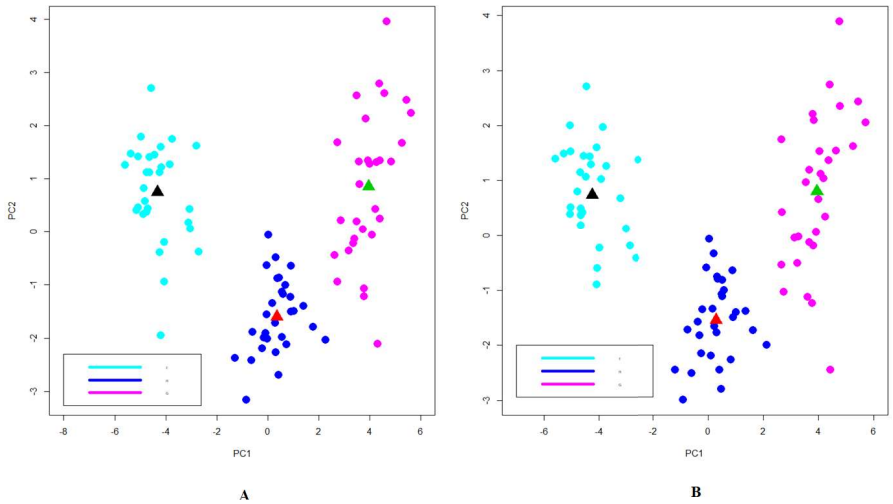


Figura I.34: A- PCA da amostra com as variáveis em estudo (22 variáveis, sem $ag13$) , B- PCA da amostra com as variáveis em estudo (21 variáveis, sem $ag8$)

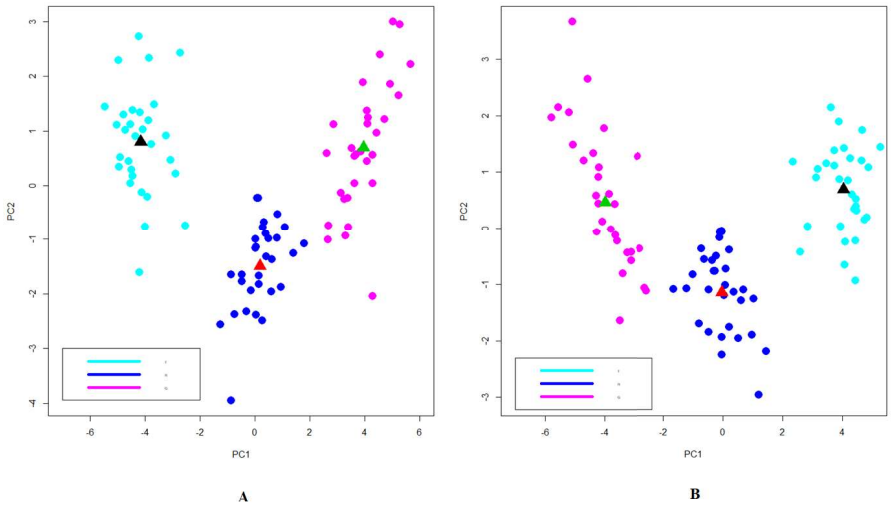


Figura I.35: A- PCA da amostra com as variáveis em estudo (20 variáveis, sem $ag17$) , B- PCA da amostra com as variáveis em estudo (19 variáveis, sem $ag1$)

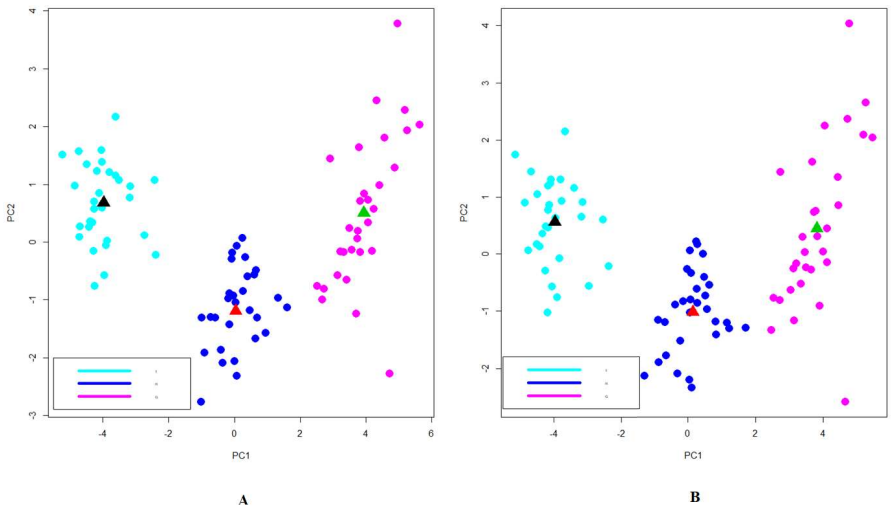


Figura I.36: A- PCA da amostra com as variáveis em estudo (18 variáveis, sem $ag2$) , B- PCA da amostra com as variáveis em estudo (17 variáveis, sem $ag7$)

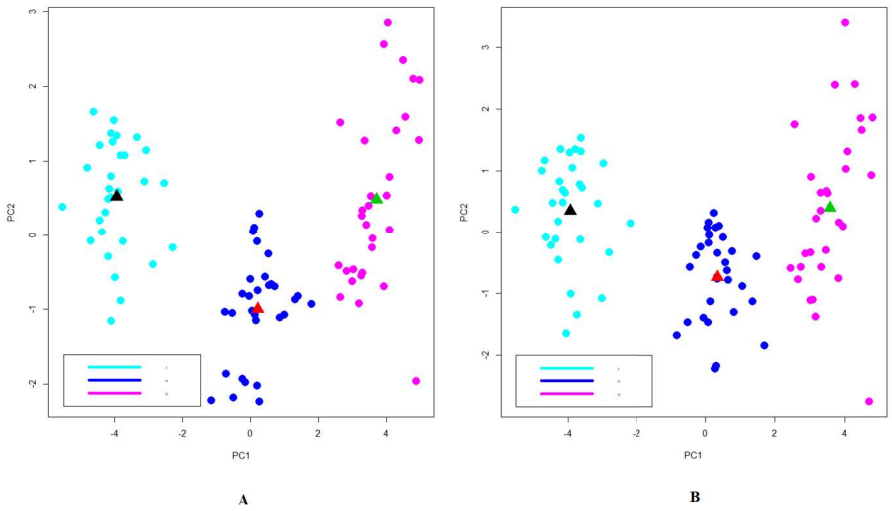


Figura I.37: A- PCA da amostra com as variáveis em estudo (16 variáveis, sem *el*18) , B- PCA da amostra com as variáveis em estudo (15 variáveis, sem *ag*16)

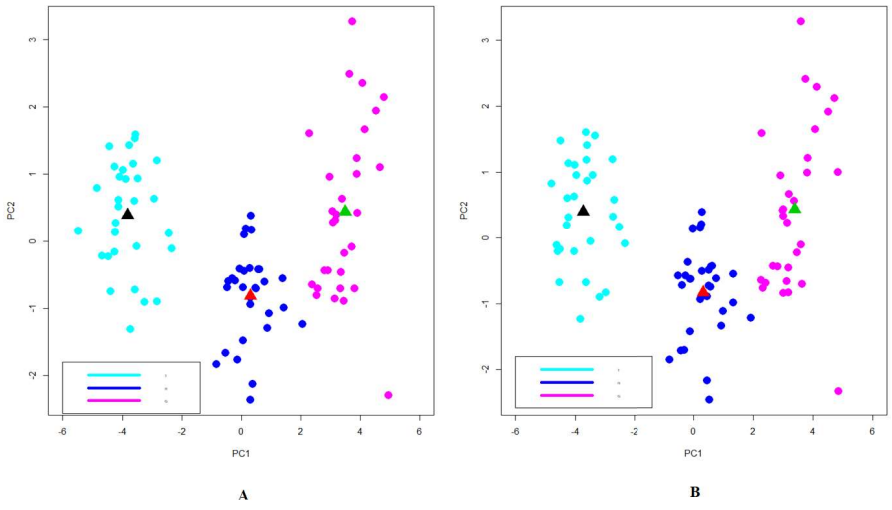


Figura I.38: A- PCA da amostra com as variáveis em estudo (14 variáveis, sem *ag*23) , B- PCA da amostra com as variáveis em estudo (13 variáveis, sem *el*12)

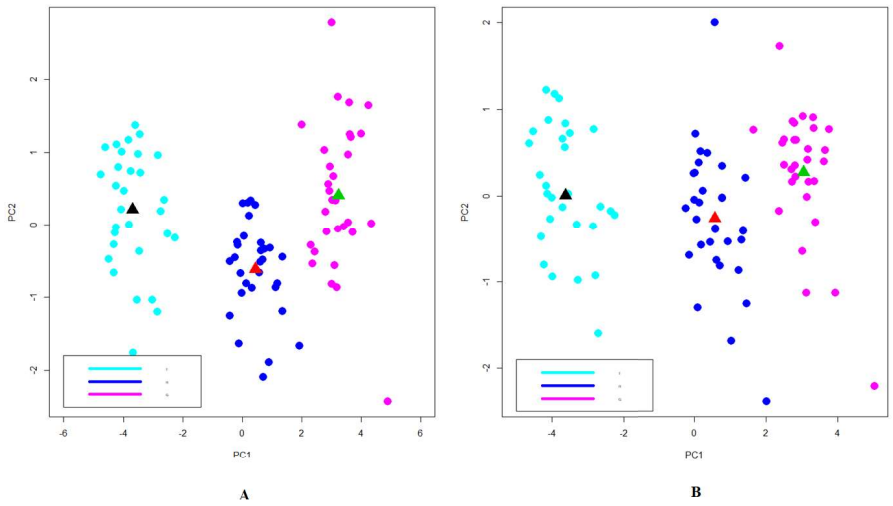


Figura I.39: A- PCA da amostra com as variáveis em estudo (12 variáveis, sem $el1$) , B- PCA da amostra com as variáveis em estudo (11 variáveis, sem $el12$)

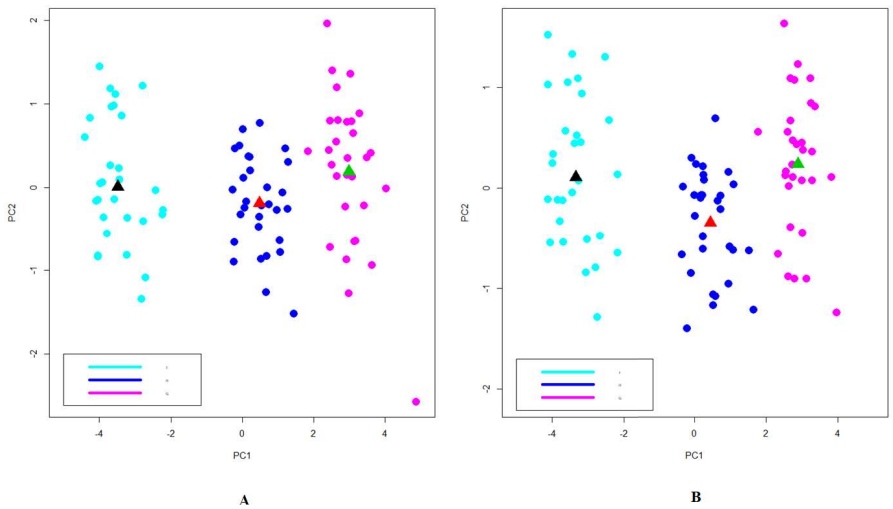


Figura I.40: A- PCA da amostra com as variáveis em estudo (10 variáveis, sem $ag6$) , B- PCA da amostra com as variáveis em estudo (9 variáveis, sem $ag20$)

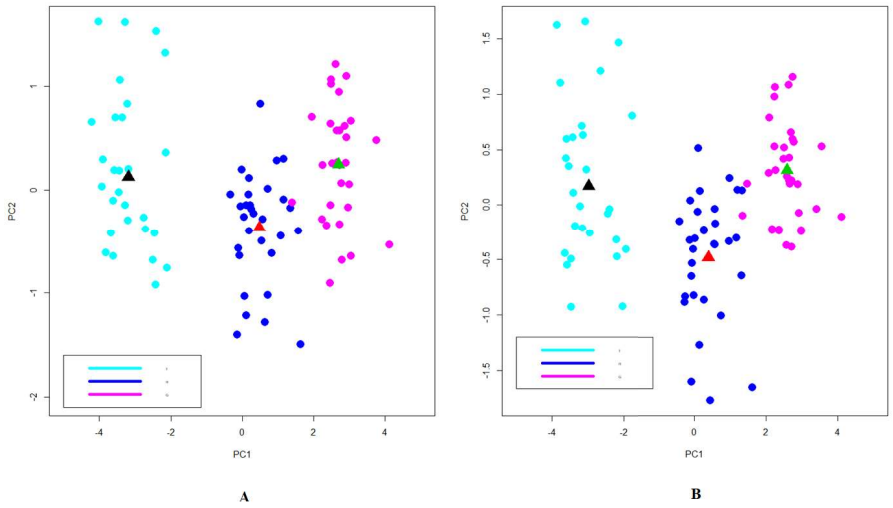


Figura I.41: A- PCA da amostra com as variáveis em estudo (8 variáveis, sem $el11$) , B- PCA da amostra com as variáveis em estudo (7 variáveis, sem $ag25$)

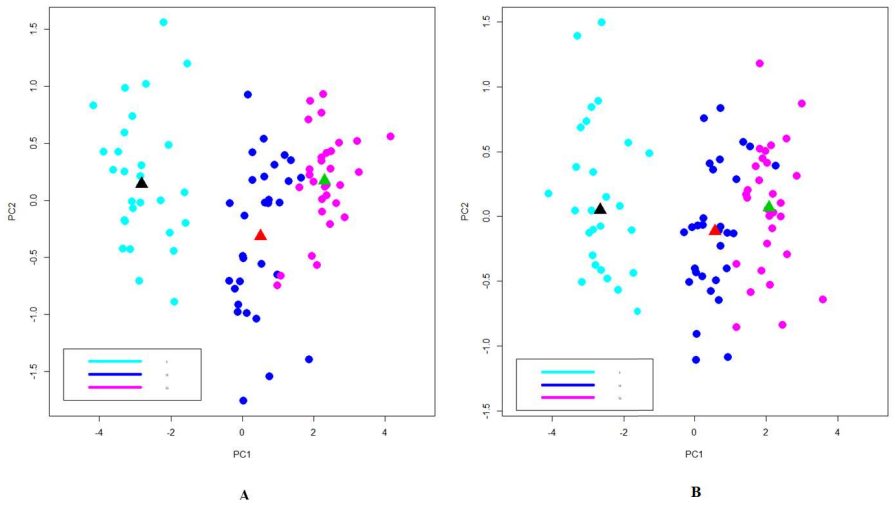


Figura I.42: A- PCA da amostra com as variáveis em estudo (6 variáveis, sem $ag22$) , B- PCA da amostra com as variáveis em estudo (5 variáveis, sem $ag18$)

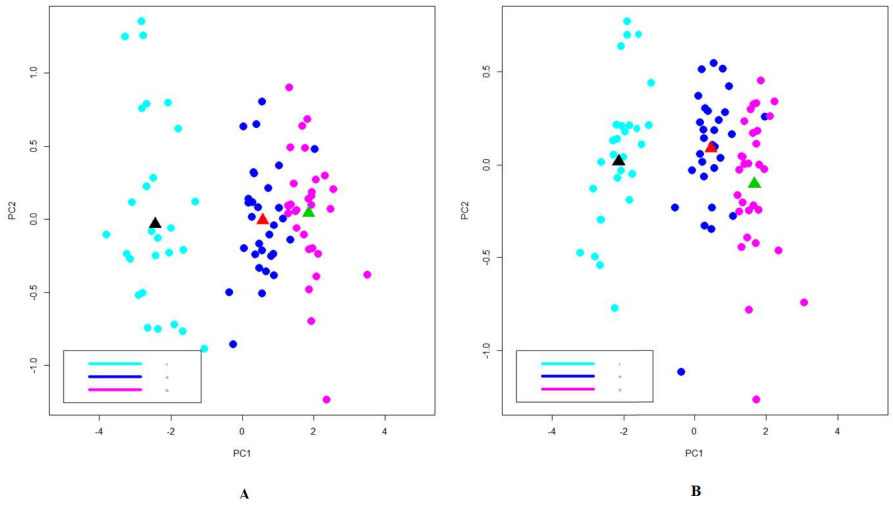


Figura I.43: A- PCA da amostra com as variáveis em estudo (4 variáveis, sem *ag*14) , B- PCA da amostra com as variáveis em estudo (3 variáveis, sem *ag*5)

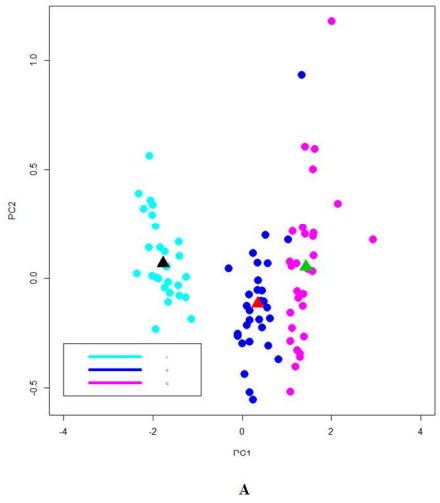


Figura I.44: A- PCA da amostra com as variáveis em estudo (2 variáveis, sem *ag*9)

ANEXO II

APÊNDICE B- GRÁFICOS COMPLEMENTARES NO ESTUDO DA REDUÇÃO DO TAMANHO DA AMOSTRA

II.1 Histogramas do coeficiente RM para diferentes n

Distribuição dos valores de coeficiente RM, para cada n.

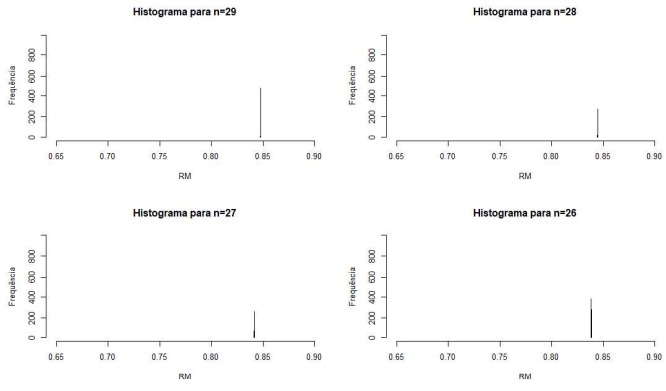


Figura II.1: Histograma do coeficiente RM para cada n

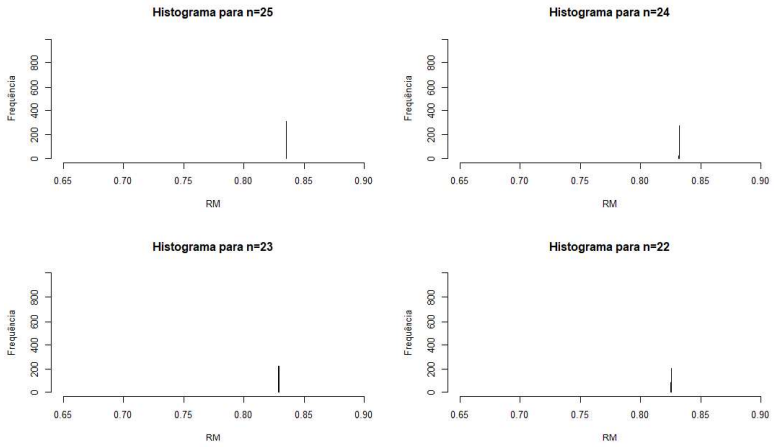


Figura II.2: Histograma do coeficiente RM para cada n

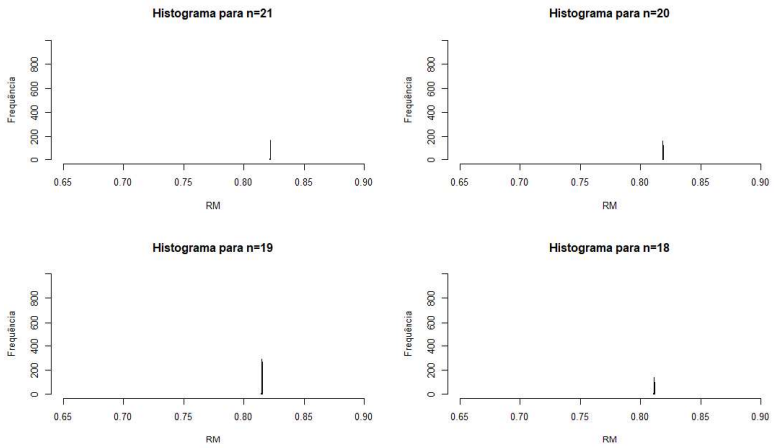


Figura II.3: Histograma do coeficiente RM para cada n

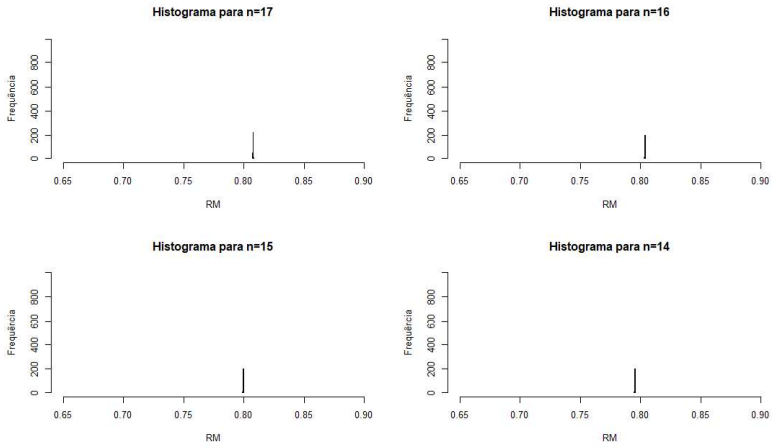


Figura II.4: Histograma do coeficiente RM para cada n

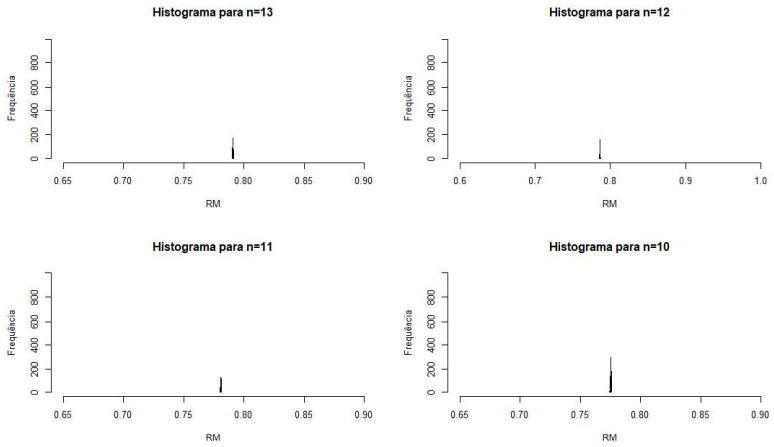


Figura II.5: Histograma do coeficiente RM para cada n

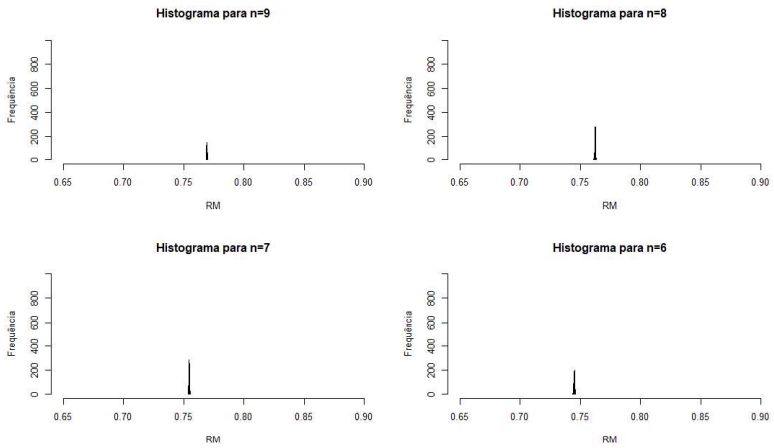


Figura II.6: Histograma do coeficiente RM para cada n

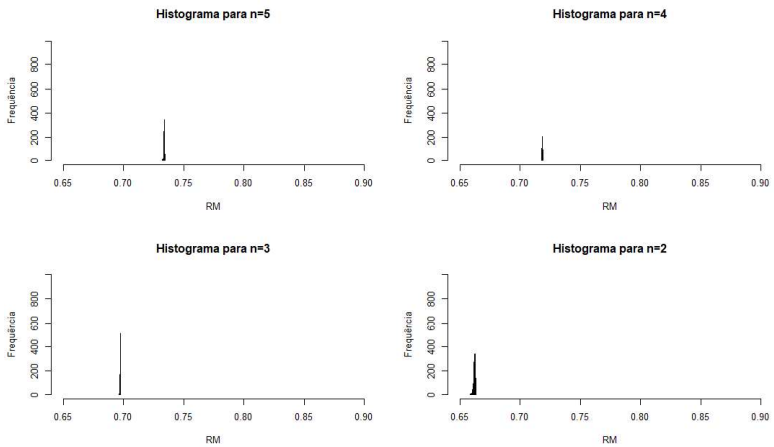
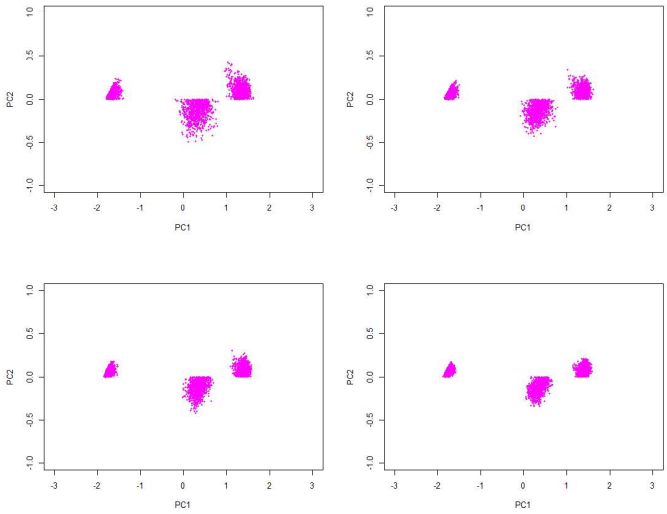


Figura II.7: Histograma do coeficiente RM para cada n

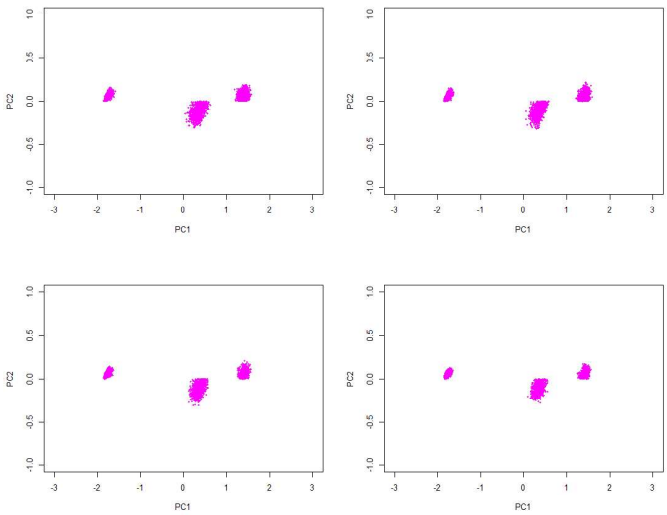
II.2 Gráficos dos centróides para diferentes n

Representação gráfica dos centróides obtidos em cada simulação para o mesmo n.



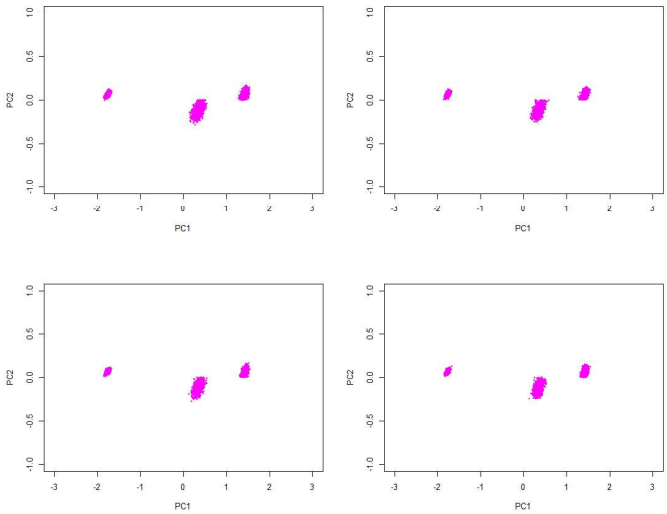
a - Centróides para n=2,..,5

Figura II.8: Representação gráfica dos centóides obtidos em cada simulação



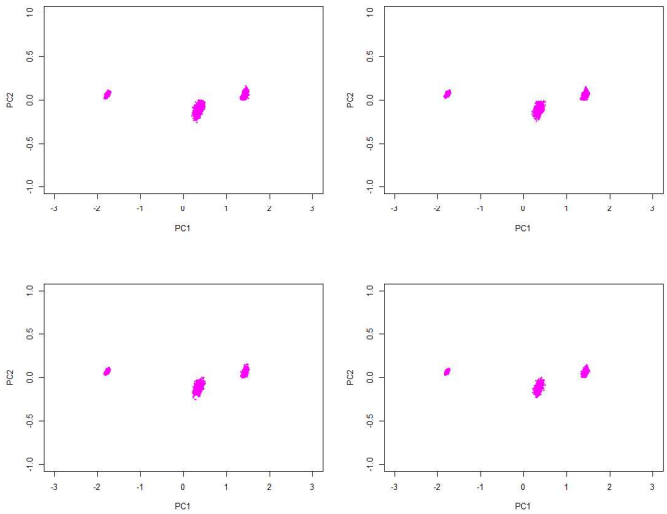
a - Centróides para n=6,..,9

Figura II.9: Representação gráfica dos centóides obtidos em cada simulação



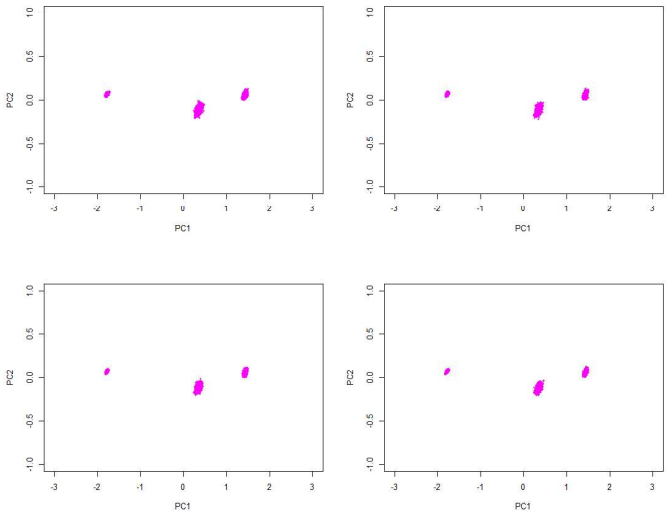
a - Centróides para n=10,..,13

Figura II.10: Representação gráfica dos centóides obtidos em cada simulação



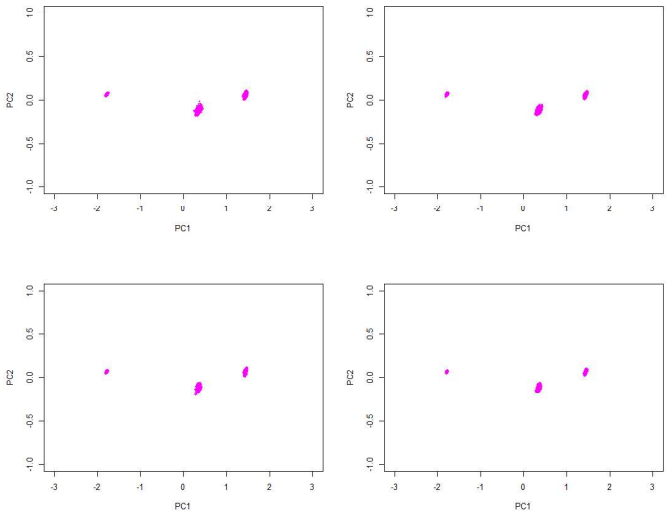
a - Centróides para n=14,..,17

Figura II.11: Representação gráfica dos centóides obtidos em cada simulação



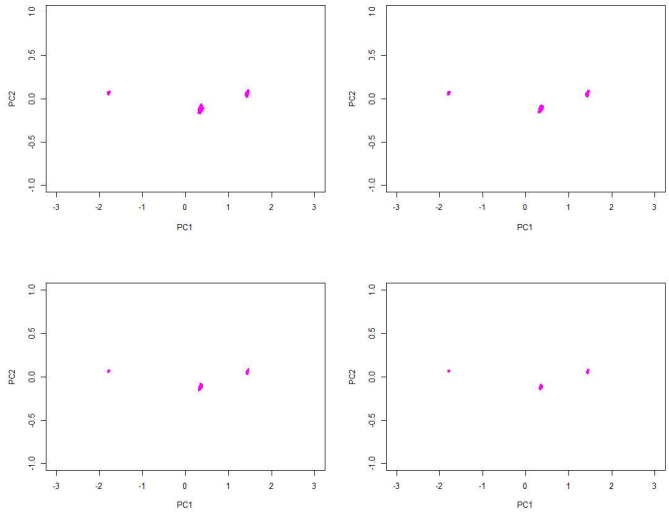
a - Centróides para n=18,..,21

Figura II.12: Representação gráfica dos centóides obtidos em cada simulação



a - Centróides para n=22,..,25

Figura II.13: Representação gráfica dos centóides obtidos em cada simulação



a - Centróides para n=26,..,29

Figura II.14: Representação gráfica dos centóides obtidos em cada simulação