

MÉTODOS DE APRENDIZAGEM NÃO SUPERVISIONADA

Turma 2 - Grupo 2



REALIZADO POR:

Carolina Morgado - 123794

Diogo Nobre - 123441

Inês Silva - 123407

Marcos Mestre - 123436

Maria Pereira - 123393

JANEIRO 2025

Índice

1. Introdução.....	3
2. Descrição dos Dados.....	3
3. Tratamento e Compreensão dos dados.....	4
4. Análise dos Componentes Principais (PCA).....	5
5. <i>Clustering</i>.....	7
5.1. <i>Clustering</i> Hierárquico.....	8
5.2. <i>Clustering</i> Partitivo.....	9
5.3. <i>Clustering</i> Probabilístico.....	9
6. Discussão dos Resultados e Conclusão.....	12

1. Introdução

No domínio da unidade curricular Métodos de Aprendizagem Não Supervisionada, foi nos proposto a realização de um projeto com o objetivo de consolidar e aplicar os conhecimentos obtidos ao longo do trimestre.

O setor imobiliário é um dos pilares da economia, englobando atividades cruciais que vão desde a construção e venda de imóveis até à gestão de propriedades. Este mercado é influenciado por diversos fatores, como a localização das casas, a sua infraestrutura, a oferta e a procura. A base de dados disponibilizada apresenta dados relativos às características intrínsecas de diversos imóveis. Traçamos como o nosso principal objetivo agrupar as casas consoante as suas características intrínsecas e tentar encontrar padrões que nos permitam compreender melhor como essas características se relacionam entre si. Idealmente, pretendemos identificar grupos homogêneos que representem diferentes tipos de imóveis, por exemplo imóveis de luxo (de grandes dimensões, em excelente condição e com boa vista) ou imóveis urbanos compactos (com uma área menor, com menos divisões e menos espaço de armazenamento).

Este relatório irá conter todos os procedimentos necessários para atingir o objetivo definido anteriormente, começando por uma descrição detalhada da base de dados e o seu respetivo tratamento. De seguida, será feita a análise de componentes principais e, posteriormente, a realização de diversos modelos de *clustering*. Finalmente, será efetuada uma breve conclusão com as ideias principais retiradas ao longo do projeto.

2. Descrição dos Dados

O *dataset* disponibilizado ao nosso grupo (T2Gr02) inclui dados relativos a diversas características de imóveis, contendo inicialmente 4152 observações e 13 variáveis, sendo 11 quantitativas e 2 qualitativas ordinais (Vista e Condição). As variáveis que foram utilizadas para realizar a análise dos componentes principais, denominadas variáveis de *input*, foram todas as variáveis exceto o Preço e o ID.

3. Tratamento e Compreensão dos dados

Posteriormente à visualização do *dataset*, verificámos que tínhamos de resolver algumas incongruências e limpar os dados que nos foram fornecidos.

Observamos que tanto na variável Número de Casas de Banho, como na variável Número de Andares existiam valores decimais, o que não faria sentido, portanto, procedemos ao arredondamento dos mesmos. Optamos por substituir sempre pelo valor inteiro a seguir, por exemplo, quando o Número de Casas de Banho era 1.25, 1.50 e 1.75 arredondamos para 2 casas de banho e quando o Número de Andares era 2.5, optamos por substituir por 3 andares.

Foi também notado pelo grupo que existiam observações com a variável Preço igual a 0 e ainda que existiam observações em que o Ano de Construção era superior ao Ano de Renovação, revelando incongruências. Como estes imóveis representavam apenas 7.39% dos dados, decidimos eliminá-los, por isso o nosso *dataset* final conta com 3845 observações.

Após termos finalizado o tratamento e a limpeza dos dados, fomos estudar as correlações entre as diversas variáveis de *input*. Para isso, realizamos uma matriz de correlações e o seu respetivo gráfico, para facilitar a visualização e compreensão (Imagem 1).

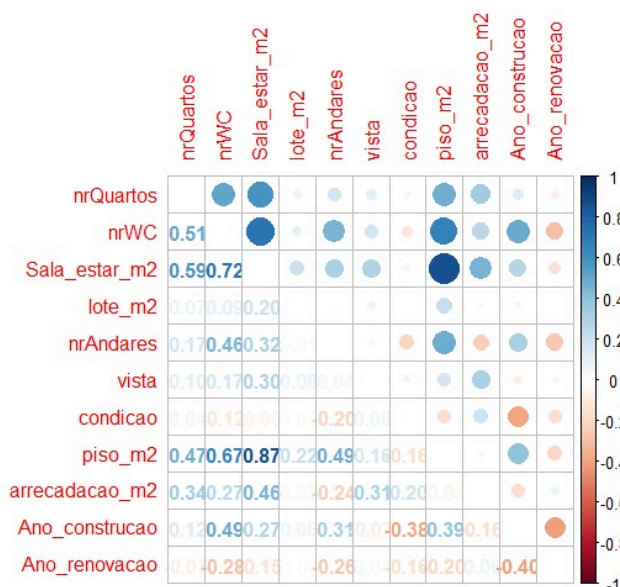


Imagem 1 - Gráfico de correlações entre as variáveis de *input*

Visualizando o gráfico concluímos que, no geral, as variáveis não são muito correlacionadas entre si, com exceção da relação entre a Área do Piso e a Área da Sala de Estar ($r = 0,87$, correlação forte), e entre a Área da Sala de Estar e o Número de Casas de Banho ($r = 0.72$, correlação moderada).

Chamou-nos à atenção o facto que também existiam correlações negativas, sendo as mais fortes (mas mesmo assim, fracas) entre o Ano de Construção e o Ano de Renovação ($r = -0.4$) e entre a Condição e o Ano de Construção ($r = -0.38$).

4. Análise dos Componentes Principais (PCA)

Uma vez feita a compreensão dos dados e da sua matriz de correlações, procedeu-se à análise dos componentes principais (PCA), utilizando as variáveis de *input* escolhidas. Em primeiro lugar, foram realizados diversos testes para verificar a adequação da aplicação da técnica de PCA. Segundo os critérios de adequação abordados ao longo da unidade curricular, foi verificado que:

- O número de observações é superior a 300, indicando uma condição favorável para a realização de PCA.
- Analisando a matriz e o gráfico de correlações, verificou-se que mais de metade das correlações (36 em 55: 65%) têm o valor do coeficiente r menor, em módulo, do que 0.3. Este resultado indica que algumas variáveis podem não estar suficientemente correlacionadas entre si, podendo vir a limitar a eficiência dos PCs em captar padrões subjacentes nos dados.
- De seguida, realizou-se o Teste de *Bartlett*, que testa se a matriz de correlações é igual a uma matriz de identidade. O teste interpreta-se tendo em conta o *p-value* obtido, que tem de ser inferior ao nível de significância escolhido (sendo 0.05 o valor padrão). O resultado do nosso Teste de *Bartlett* foi *p-value* = 0, logo, rejeitou-se a hipótese nula de que a matriz de correlações se trata de uma matriz identidade. Desta forma, concluiu-se que, segundo este teste, é adequado prosseguir com a aplicação de PCA.
- Calculou-se ainda a medida de adequação amostral de *Kaiser-Meyer-Olkin* (KMO), sendo um teste que retorna valores entre 0 e 1. Quanto mais perto de 1 o resultado obtido for, mais adequados são os dados para realização de PCA. No nosso caso, o

KMO retornou um valor de 0.58, o que indica uma adequação medíocre dos dados para a realização de PCA.

Apesar do valor de KMO não ser ideal, e da presença de coeficientes de correlação inferiores a 0.3 em mais de metade dos pares de variáveis, decidiu-se prosseguir com a implementação de PCA, pois esse era um dos principais objetivos do trabalho.

O primeiro passo para criar os Componentes Principais foi a *standardização* dos dados, pois a sua realização permite assegurar que as diferentes unidades de medida não influenciam os resultados. Posteriormente, foi estabelecido um número inicial de PCs, que foi igual ao número de variáveis de *input* que estamos a utilizar, ou seja, 11 PCs.

Após esta extração inicial, tornou-se fulcral determinar o número ideal de PCs que devem ser mantidos para prosseguir com a análise. Para tomar esta decisão tivemos em conta diversos fatores:

- **Critério de *Kaiser*:** Este critério indica que apenas os componentes com valores próprios (*eigenvalues*) superiores a 1 devem ser mantidos. Seguindo este critério, devemos retirar os primeiros 4 componentes principais.
- **Variância total extraída:** É recomendado que os PCs selecionados contenham pelo menos 60% da variância total dos dados. No nosso caso, com os primeiros 3 componentes este critério estava cumprido (61%) e com 4 PCs esta percentagem aumenta para 70%.
- **Critério do cotovelo (*Scree Plot*):** Ao analisar o gráfico *Scree Plot* (Imagem 2), observou-se que o "cotovelo" do gráfico (o ponto em que a taxa de diminuição dos *eigenvalues* desacelera) ocorre por volta do 3º e 4º PC, estando de acordo com os métodos realizados previamente.

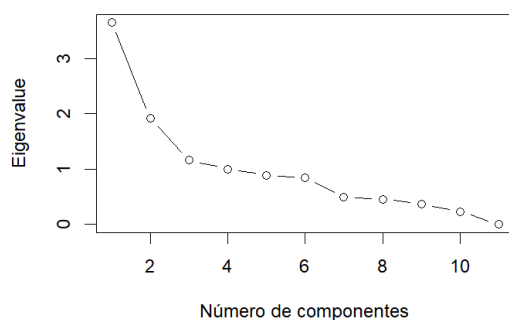


Imagem 2 - *Scree Plot* - Método do cotovelo

Após a realização destes 3 testes com resultados unânimes, optamos por manter os primeiros 4 componentes principais.

Após a extração dos PCs, fomos verificar a variância que ficou explicada para cada variável de *input* utilizada. Notou-se que a variável Vista apresenta o valor de comunalidade mais baixo (0.304), estando subrepresentada nos PCs escolhidos, sendo a única que tem um valor abaixo dos 0.5.

De seguida foram ainda realizadas rotações dos componentes principais obtidos, com o objetivo de lhes dar um maior significado e os tornar mais fáceis de interpretar. Para isso, utilizou-se o método de rotação *Varimax*, que aumenta os maiores *loadings* e diminui os menores *loadings* dentro de cada componente principal. Com base nesta rotação, foram atribuídos nomes às novas variáveis criadas, que devem espelhar as variáveis de *input* utilizadas que mais nele estão representadas. Os nomes atribuídos a cada um dos PCs foram:

- **PC1 - Índice de Estrutura e Idade** (relaciona principalmente as variáveis *nrWC*, *nrAndares*, *piso_m2* e *Ano_construção*).
- **PC2 - Índice de Espaço Interior e Arrumos** (relaciona principalmente as variáveis *nrQuartos*, *sala_estar_m2*, *arrecadação_m2*).
- **PC3 - Índice de Condição e Renovação** (relaciona inversamente as variáveis *condição* e *Ano_renovação*).
- **PC4 - Índice do Terreno** (é composto principalmente pela variável *lote_m2*).

No Anexo 1, encontra-se o *pairs plot*, que permite visualizar as relações entre os PCs, a pares. Embora os PCs sejam sempre não correlacionados linearmente entre si, a visualização dos seus *scatterplots* pode sugerir possíveis relações mais complexas e interessantes entre os componentes. Este tipo de gráfico tem também utilidade para visualização dos *outliers*, pois são pontos que se destacam significativamente dos outros em qualquer uma das combinações de PCs. No nosso caso, não se destacam relações significativas entre componentes principais e pode ser notada a presença de alguns *outliers*.

5. Clustering

Para realizar *clusters*, existe sempre a necessidade de calcular a matriz de distâncias para cada ponto, sendo que a forma de calcular estas distâncias pode variar. No nosso

trabalho usamos a distância euclidiana (a distância *default* nas funções do R utilizadas) em todos os nossos métodos de *clustering*.

Foi decidido pelo grupo utilizar os três tipos de *clustering* abordados na Unidade Curricular: *Clustering* Hierárquico, *Clustering* Partitivo e *Clustering* Probabilístico.

5.1. *Clustering* Hierárquico

Os *clusters* hierárquicos formam estruturas não sobrepostas (cada observação só pode pertencer a um único *cluster*) e, idealmente, separadas entre si. Para a criação dos *clusters* as observações são agrupadas progressivamente em forma de árvore hierárquica, sendo possível visualizá-la através de um dendrograma. Utilizando este tipo de *clustering*, não é necessário definir previamente um número de *clusters*. Dentro do *clustering* hierárquico optamos por utilizar diferentes métodos, nomeadamente o Algoritmo de *Ward*, o método *Single* e o método *Complete*.

Na Imagem 3 encontra-se o dendrograma obtido através do Algoritmo de *Ward*.

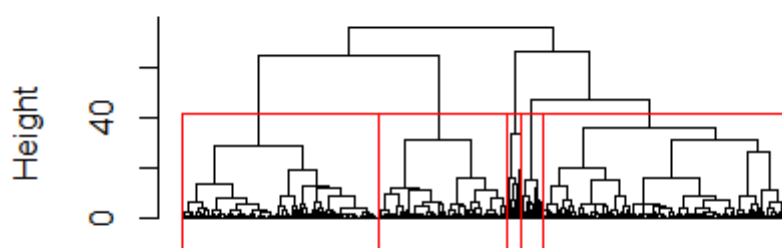


Imagem 3 - Dendrograma obtido através do Algoritmo de *Ward* com 5 *clusters*

Pela análise do dendrograma optou-se por realizar 5 *clusters*. Após esta decisão, realizou-se um gráfico com a função *silhouette* (Anexo 2) e obtivemos o valor 0.33, que nos indica uma estrutura fraca, tendo em conta que, valores próximos de 0, sugerem que os pontos estão localizados entre *clusters*. O ideal seria obter valores próximos de 1, pois isso significa que os pontos estariam bem classificados no seu *cluster*.

Realizamos também os métodos *Complete* e *Single*, mas os dendrogramas obtidos demonstraram *clusters* desproporcionais e uma visualização pouco clara (Anexo 3).

5.2. Clustering Partitivo

Ao contrário dos métodos de *clustering* anteriores, este consiste em repartir as observações num número pré-definido de *clusters*, satisfazendo critérios de otimização. Neste tipo de *clustering*, cada observação também pertence a um único *cluster* e os resultados obtidos irão depender do valor de K escolhido.

Para escolher qual seria o melhor número de *clusters*, utilizamos uma função que gera um gráfico *screeplot* (Imagem 4) que nos permite, através do método do cotovelo, concluir que o valor ótimo de K a utilizar são 5 *clusters*. Após realizarmos o gráfico *silhouette*, concluímos que a estrutura dos *clusters* com este método também era fraca, pois revelou um valor de silhueta de 0.37 (Anexo 4).

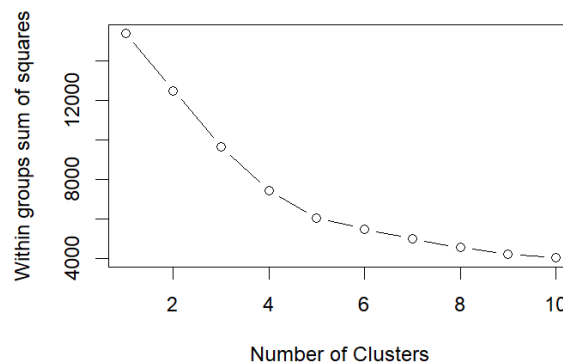


Imagem 4 - Screeplot do algoritmo *K-means*

Para concluir os métodos de partição, utilizamos também o método PAM (*Partition Around Medoids*), onde o número de *clusters* foi definido *à priori*, tendo sido escolhidos 5 *clusters*. Este método apresentou também um baixo valor de silhueta (igual a 0.36 - Anexo 5), o que revelou que, à semelhança dos métodos prévios, o PAM também apresenta uma estrutura fraca.

5.3. Clustering Probabilístico

Por último, procedeu-se à realização de *clustering* Probabilístico. O seu algoritmo é mais complexo e possui uma convergência lenta. Neste processo, a formação dos *clusters* baseia-se nas probabilidades atribuídas a cada observação de pertencer a diferentes *clusters*, permitindo que uma mesma observação tenha associação a mais do que um *cluster*.

Para auxiliar na escolha do melhor número de *clusters* e no Modelo de Mistura Gaussiana que deve ser utilizado, recorreu-se ao *Bayesian Information Criterion* (BIC). Após a visualização do gráfico BIC abaixo (Imagem 5), concluiu-se que iríamos realizar os *clusters* seguindo o modelo VVV com 7 *clusters*. Idealmente, o modelo escolhido seria o VVV com 9 *clusters*, pois deve ser escolhido o modelo com um número de BIC maior. Porém, esse número é quase o dobro do número de *clusters* selecionados pelos métodos anteriores, que agrupam todos os dados em 5 *clusters*. Tendo em consideração que a diferença do BIC entre 7 e 9 *clusters* é mínima, optou-se por realizar a análise com 7 *clusters*.

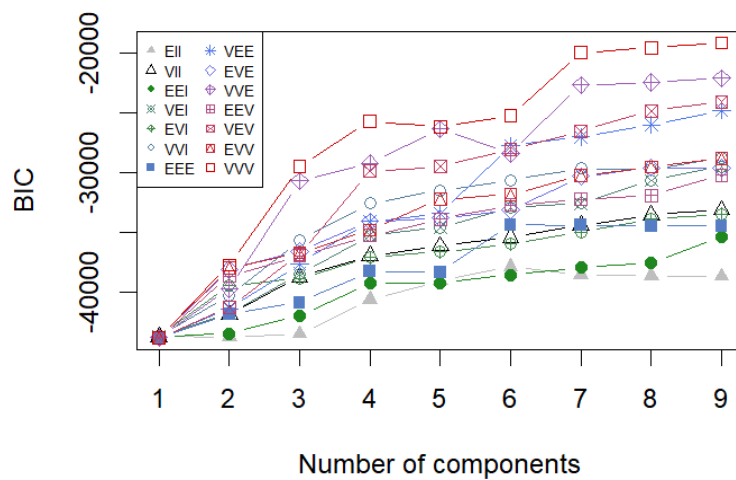


Imagem 5 - Número de *clusters* associados ao valor de BIC para cada modelo GMM

De seguida foi realizado o gráfico *classification* para o modelo de *clustering* probabilístico que utilizamos e este pode ser visualizado na Imagem 6.

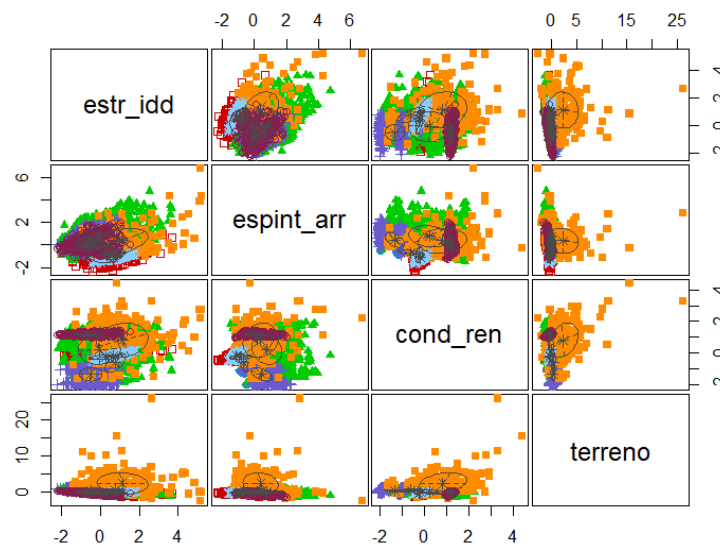


Imagem 6 - Gráfico *classification* para o modelo probabilístico VVV

A interpretação do gráfico permite perceber quais são as observações que estão atribuídas a cada *cluster* mostrando através de diferentes cores e formas qual é o grupo a que cada ponto pertence. Pode-se ainda visualizar que em cada um dos gráficos existem desenhadas elipses pretas (por vezes muito sobrepostas), que correspondem a cada *cluster*. Neste tipo de gráfico, a cada observação é atribuído um único número de *cluster*, que é aquele que tem o maior valor de probabilidade calculado pelo modelo. Este gráfico simplifica a incerteza inerente aos modelos de *clustering* probabilísticos, pois mostra apenas a atribuição definitiva de cada imóvel a um único *cluster*.

Para visualizar a incerteza, foi gerado outro gráfico com a função *uncertainty* (Imagem 7). O gráfico *uncertainty* mostra visualmente a confiança que o modelo tem na associação de cada observação a um *cluster*. À semelhança do gráfico anterior, também se podem notar elipses pretas desenhadas em cada um dos gráficos, que representam os *clusters*. Porém, ao contrário do gráfico de classificação, no de incerteza a forma de cada observação é sempre circular, e muda apenas a sua cor de preenchimento e tamanho do círculo. Na nossa interpretação, quanto maior o tamanho do círculo, menor a confiança do modelo em posicionar a observação no respetivo *cluster*.

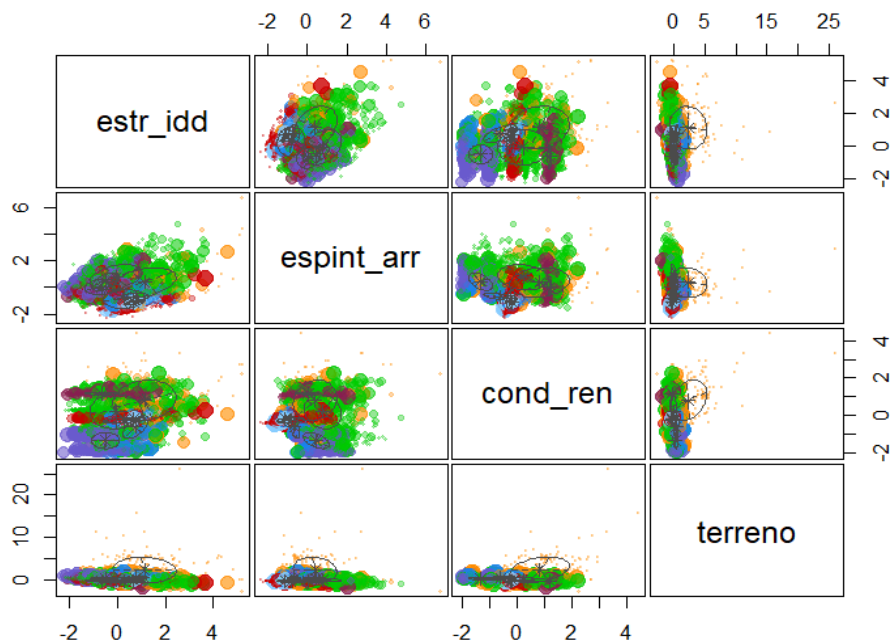


Imagem 6 - Gráfico *uncertainty* para o modelo probabilístico VVV

Em concordância com os métodos abordados previamente, interpretando visualmente os gráficos, esta estrutura também se revela fraca, porque não se conseguem distinguir *clusters* bem definidos no gráfico de classificação e a quantidade de círculos com maior diâmetro no gráfico de *uncertainty* é notória.

6. Discussão dos Resultados e Conclusão

Como referido anteriormente, foram utilizados 6 métodos de *clustering* ao longo do projeto. Dentro dos métodos hierárquicos foram utilizados 3: *Ward*, *Single* e *Complete*. Dos métodos partitivos foram utilizados 2: *K-means* e PAM. Dos métodos de *clustering* probabilísticos foi escolhido o modelo VVV dentro dos Modelos de Mistura Gaussianos (GMM). Na discussão de resultados vamos abordar e comparar todos os métodos exceto o *Single* e o *Complete*, por terem, aparentemente, uma estrutura muito fraca logo na etapa da visualização do dendrograma.

Começamos por tentar perceber que tipos de casas é que os *clusters* de cada método tentavam agrupar. Realizamos gráficos de barras para cada *cluster*, de modo a podermos visualizar a média dos componentes principais nele incorporados. Começando pelo Método *Ward*, na Imagem 6 podem ser visualizados os 5 gráficos de barras criados, cada um para cada *cluster*.

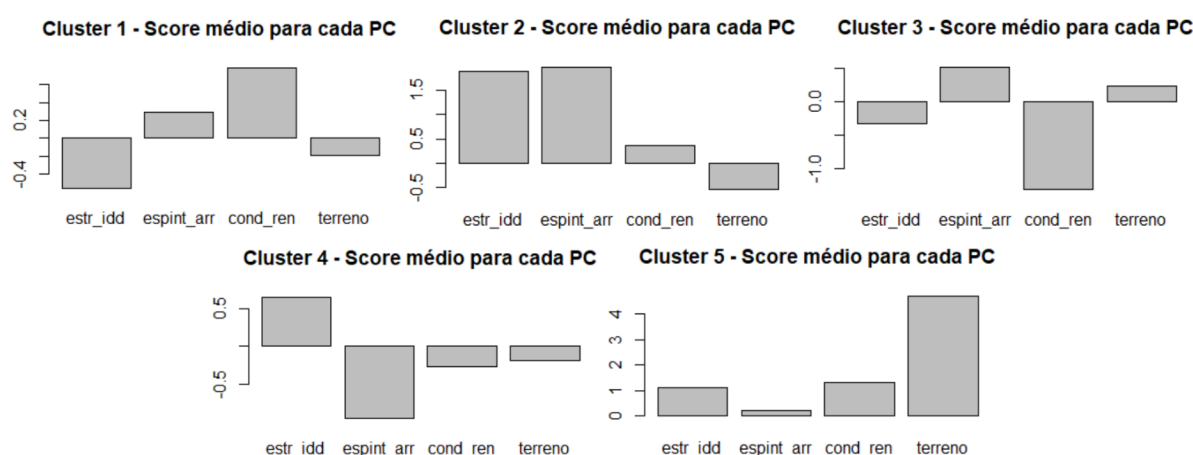


Imagem 6 - Modelo *Cluster Ward*: média do valor dos PCs em cada *cluster*.

Pela interpretação da imagem acima pode ser visualizado que, por exemplo, no *cluster* 1, o método de *Ward* colocou alojamentos com um Índice de Estrutura e Idade menor e um Índice de Condição e Renovação maior, o que significa que são casas com menos casas de banho e andares, com menor área de piso e ano de construção, com uma pior condição e com um ano de renovação mais recente (dentro do PC *cond_ren* as variáveis são inversamente relacionadas). Pela mesma lógica, por exemplo, no *cluster* 5 o algoritmo colocou as casas com um maior Índice de Terreno, logo, com um lote de maior área.

Estes gráficos foram também realizados para os restantes algoritmos e podem ser consultados nos Anexos 6 (*K-means*) e 7 (PAM).

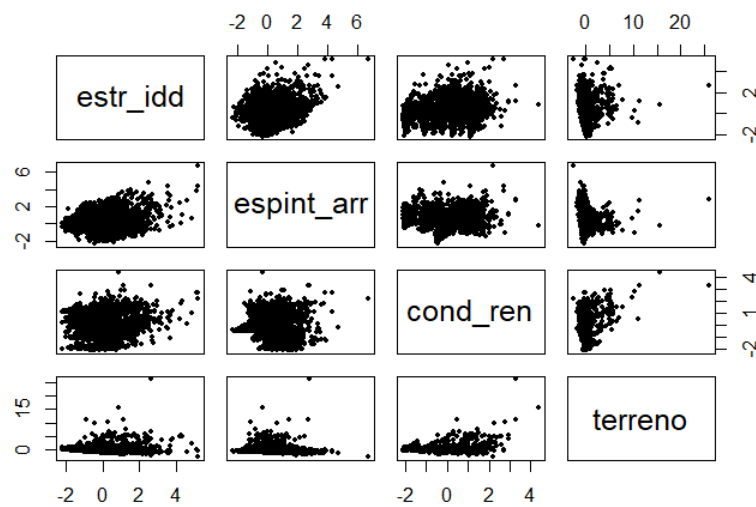
Após esta análise, foi relevante para nós perceber se os modelos concordavam entre si na atribuição dos *clusters*. Para isso realizamos as tabelas encontradas na Imagem 7 que fazem a contagem do número de observações colocadas em cada *cluster*. Num caso ideal, a tabela devia ter apenas números na diagonal principal (assinalada a amarelo) e o resto das entradas devia ser igual a 0, pois isso significaria que ambos os modelos tinham posto o mesmo número de observações em cada *cluster*. Seguindo este raciocínio, concluímos que os modelos mais concordantes entre si são o *Ward* e o PAM, pois são estes os que têm um maior número de observações na diagonal principal. Nesta análise não foi utilizado o modelo probabilístico por ter um número de *clusters* diferente.

	■ <i>Ward</i>					■ <i>K-means</i>					■ PAM				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
1	109	1	193	179	1092	0	19	860	1	3	919	364	181	109	1
2	0	0	0	136	1	0	0	0	0	73	0	136	0	1	0
3	744	0	9	63	0	38	12	100	1208	1	0	67	741	5	3
4	30	1	1155	34	11	0	397	1	15	0	1	28	94	1107	1
5	0	71	2	1	13	892	169	55	0	1	10	2	0	2	73

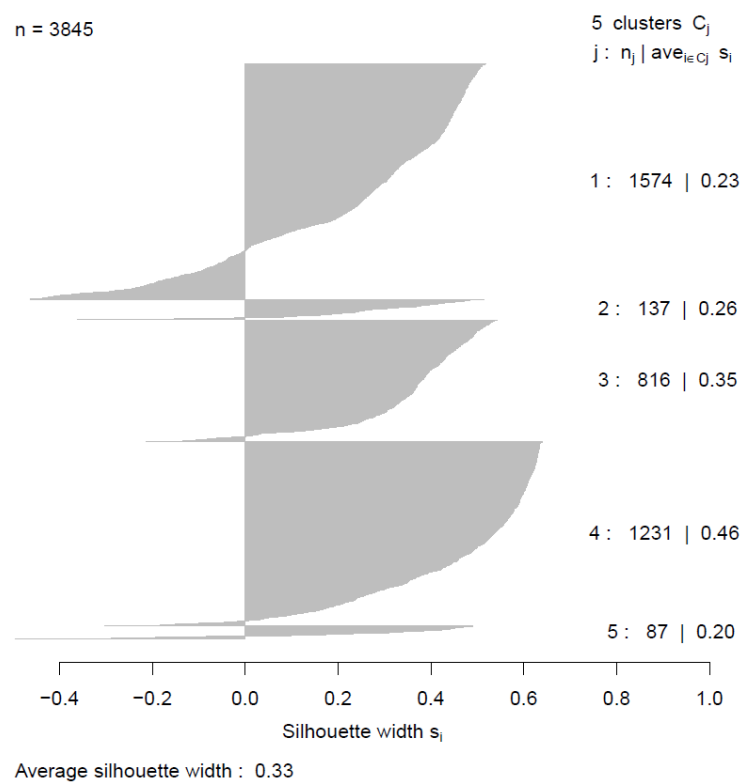
Imagem 7 - Tabelas de concordância dos modelos *Ward*, *K-means* e PAM

Após ter sido analisada a concordância entre *clusters*, tiramos conclusões definitivas sobre os *clusters* formados pelos diferentes métodos. Para qualquer um dos métodos utilizados a conclusão foi semelhante, logo, concluímos que apesar das nossas tentativas de criação de *clusters* com diversos métodos, todos aqueles que foram utilizados resultaram numa estrutura fraca e com separações mal definidas. Este resultado indica que os dados não apresentaram uma segmentação clara, o que nos impediu de alcançar o objetivo definido de forma eficaz.

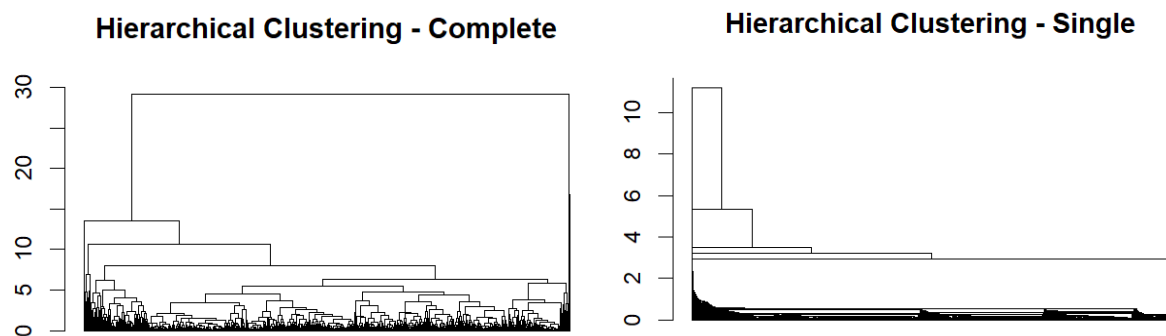
7. Anexos



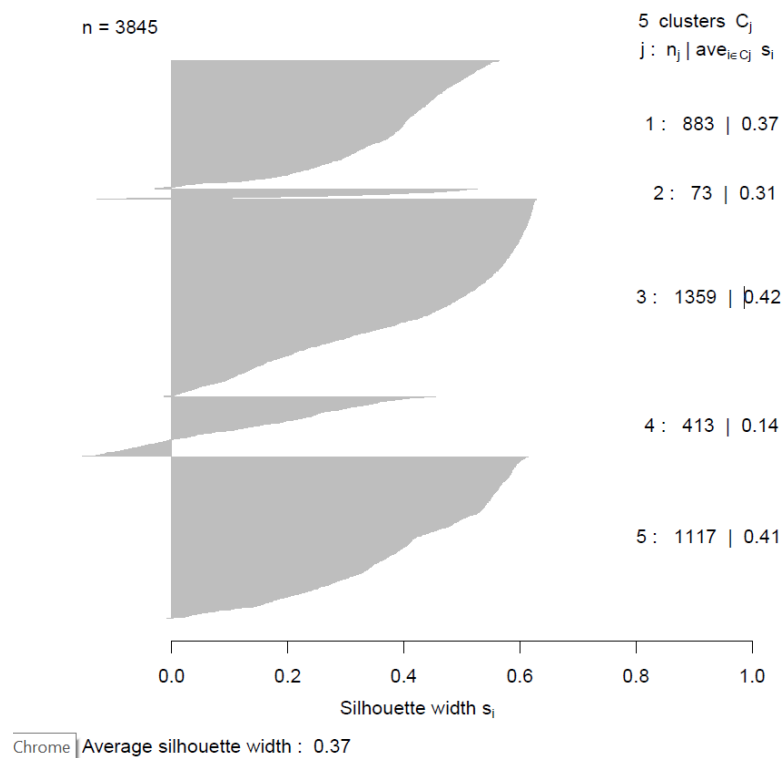
ANEXO 1 - *Pairs plot* entre os componentes principais criados



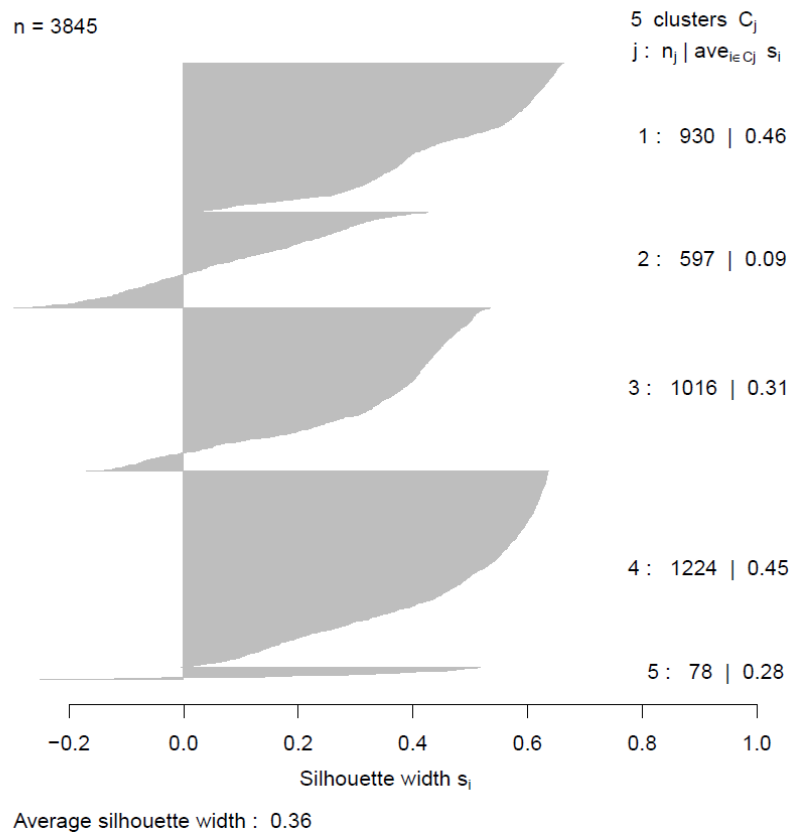
ANEXO 2 - Modelo *Ward*: Observações por *cluster* e o respetivo valor médio da silhueta



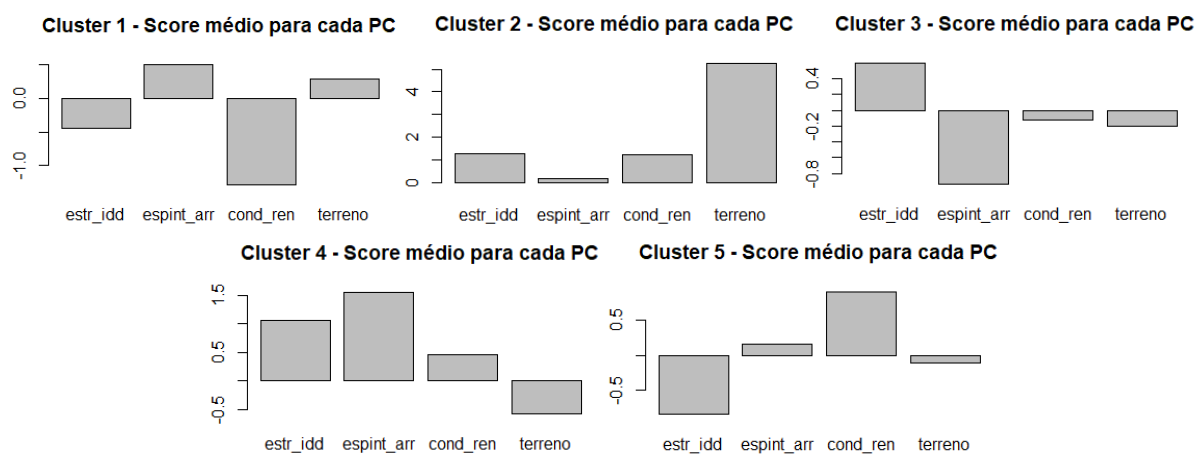
ANEXO 3 - Dendrograma obtido com o Método *Complete* (esquerda) e *Single* (direita)



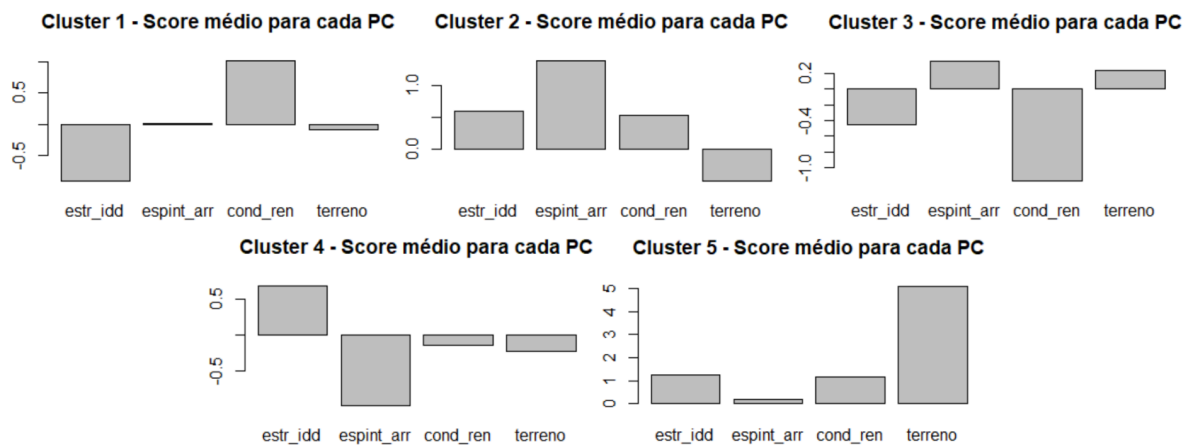
ANEXO 4 - Modelo *K-means*: Observações por *cluster* e o respetivo valor médio da silhueta



ANEXO 5 - Modelo PAM: Observações por *cluster* e o respetivo valor médio da silhueta



ANEXO 6 - Modelo *Cluster K-means*: distribuições dos *clusters* pelos PCs



ANEXO 7 - Modelo *Cluster* PAM: distribuições dos *clusters* pelos PCs