Caracterização da competitividade turística dos municípios portugueses

Aprendizagem Estatística Não Supervisionada
Mestrado em Ciência de Dados - 2025/2026
Professor José Dias

Afonso Dias - 136834
João Maurício - 136811
Rafael Machado – 87341
Vasco Veríssimo - 135659

INSTITUTO UNIVERSITÁRIO DE LISBOA

Índice

Introdução

Enquadramento

Atualmente Portugal apresenta uma forte assimetria na procura e oferta turística (bastante concentrada no litoral e grandes centros urbanos). No âmbito do desafio apresentado pelo cliente (Secretaria de Estado do Turismo, Comércio e Serviços), o objetivo deste trabalho consiste em realizar uma segmentação *data-driven* dos municípios, como base para o lançamento de um programa de fundos de apoio ao desenvolvimento turístico ("Programa Turismo 2030"), sendo esta segmentação crítica para decidir que tipo de verba atribuir a diferentes tipos de municípios.

Desta forma, a análise apresentada neste trabalho pretende construir um **"Índice de Competitividade Turística Municipal"**. Este modelo pretende assim capturar as dimensões latentes do turismo e agrupar os municípios em *clusters*, garantindo informação relevante para a tomada de decisão da Secretaria de Estado.

Dados

Caracterização da base de dados

Como base de análise para este trabalho, e com o objetivo de capturar a informação potencialmente mais relevante para a caracterização do turismo em Portugal, foram recolhidos dados do Instituto Nacional de Estatística de 34 diferentes variáveis, para os 308 municípios portugueses. A análise e preparação de dados encontra-se descrita no Anexo 1. Estas variáveis, denominadas de variáveis *Input*, agregam-se num conjunto de características apresentadas abaixo:

- **Características associadas à robustez do alojamento turístico do município**: Nº de estabelecimentos, Rev Par (proveitos por quarto), Capacidade de alojamento (total e por 1000 habitantes), Nº de dormidas (total e por 100 habitantes), Estada média, Nº de hóspedes, Proveitos, Nº de estabelecimentos, Nº de quartos e Taxa líquida de ocupação cama
- **Características associadas ao investimento municipal**: Despesas em ambiente, Despesas em proteção da biodiversidade, Despesas em atividades culturais e recreativas, Despesas em atividades e equipamentos desportivos, Despesas em museus, Despesas em monumentos, centros históricos e sítios protegidos.
- **Características associadas ao dinamismo de atividades/comércio associado ao turismo**: Nº de praias de banho (incluindo fluviais), Nº de espetáculos ao vivo, Nº de estabelecimentos de restauração, Nº de estabelecimentos de atividades artísticas e

literárias, Nº de estabelecimentos de atividades desportivas, diversão e recreativas, Nº de Museus e proporção de área da Rede Natura 2000 (como indicador de potencial turismo de natureza/ruralidade)

- **Características associadas ao emprego e população com potenciais laços ao turismo:** Mediana da idade da população, População empregada em comércio a retalho, População empregada em alojamento, População empregada em restauração, População empregada em atividades desportivas, de diversão e recreativas, População empregada com ensino secundário, População empregada com até 1º ciclo, População empregada com até 3º ciclo, População empregada com ensino superior, Taxa de atração líquida de população empregada

De modo a caracterizar os resultados foram recolhidos um conjunto de outros dados para os municípios (dados *Profile*):

- População residente
- Densidade populacional
- Ganho médio mensal
- Índice de envelhecimento
- Distrito (variável categórica)
- Se concelho é litoral ou interior (variável categórica) (Distinção feita pela APA)
- Se concelho é ou não capital de distrito (variável categórica)

Identificação das dimensões de análise

Introdução às componentes principais

A identificação das dimensões contidas na base de dados foi realizada através de uma Análise de Componentes Principais (PCA) – descrita no Anexo 2, aplicada às variáveis descritas no ponto anterior, considerando apenas as variáveis *Input*. As dimensões identificadas correspondem a agregados sintéticos de informação que resumem comportamentos comuns observados entre as variáveis, permitindo reduzir a complexidade dos dados e facilitar a sua interpretação.

As componentes principais podem ser interpretadas como fatores ocultos, isto é, dimensões que estão na base de dados e resumem padrões comuns de variação entre várias variáveis. Embora não sejam diretamente observáveis, estas dimensões refletem características dos municípios que se manifestam em simultâneo nos diferentes indicadores, como por exemplo: económico, demográfico e turístico. No contexto deste trabalho, os fatores ocultos identificados correspondem a diferentes dimensões da competitividade turística nos municípios.

PC1 - Dimensão e Capacidade Económico-Turística e Cultural

Tabela 1 - Variáveis com maior contribuição para a componente: Dimensão e Capacidade Económico-Turística e Cultural

Variável	Valor de relevância para a componente
Hóspedes nos estabelecimentos	**0.960**
Espetáculos ao vivo	**0.938**
Despesas em atividades culturais e criativas	**0.925**
Proveitos totais dos estabelecimentos	**0.924**
Atividades artísticas e culturais	**0.912**
Museus	**0.898**
Dormidas totais	**0.894**
Quartos e capacidade de alojamento	**~0.88–0.86**
Estabelecimentos turísticos	**~0.85**

Esta componente reflete a atividade turística de forma global, sendo assim, composta pela concentração de infraestruturas, atividades culturais, e a sua adesão. Exemplificando, reflete a quantidade de turistas que os hotéis, alojamentos locais, concertos, teatros, feiras, entre outros, recebem por município.

Valores elevados nestas variáveis correspondem a locais com maior volume de alojamentos, também mais investimento por parte do próprio município e investimento privado, que se traduz numa presença maior de turistas. Em contraste, valores reduzidos representam municípios mais pequenos, com menos expressão turística e cultural.

Em síntese, a componente agrega, volume da atividade turística, a oferta de alojamento, e as dinâmicas culturais e criativas. E sendo a principal componente vai ter um papel fundamental a descobrir a competitividade turística de cada município.

PC2 - Intensidade Turística Relativa e a Vocação Balnear

Tabela 2 - Variáveis com maior contribuição para a componente: Intensidade Turística Relativa e a Vocação

Variável	Valor de relevância para a componente
Dormidas por 100 habitantes	0.928
Capacidade de alojamento por 1000 habitantes	0.902
Estada média	0.583
Águas balneares	0.578

Taxa líquida de ocupação cama	0.424

A segunda componente principal distingue os locais onde a atividade turística assume um peso muito elevado relativamente à população residente. Ao ter valores elevados nestas variáveis o município está fortemente especializado em turismo balnear, principalmente em praias costeiras. Estes são tipicamente localizados no litoral e nas regiões insulares, enquanto valores reduzidos caracterizam territórios com menos intensidade turística balnear.

Sendo esta componente vocacionada para praias, é expectável que esteja associado a padrões de sazonalidade, pois é um tipo de turismo presente sobretudo na época do Verão, e isso é possível de medir através da estada média dos visitantes. Resumindo, a componente agrega a intensidade turística face à população da região e a vocação balnear, que permite diferenciar municípios com turismo intenso de municípios com menos turismo na dinâmica local.

PC3 - Perfil Socioeconómico e Complementar do Turismo

Tabela 3 - Variáveis com maior contribuição para a componente: Perfil Socioeconómico e Complementar do Turismo

Variável	Valor de relevância para a componente
Emprego no comércio a retalho	**0.824**
Emprego em atividades recreativas	**0.741**
Emprego em restauração	**0.731**
Despesas ambientais – biodiversidade	**0.717**
Taxa de ocupação cama	**0.651**
Idade mediana	**−0.570**

Por fim, a terceira componente mostra o perfil socioeconómico dos municípios. É principalmente explicado pela população empregada em áreas que estão em contacto com o turismo, como a restauração, comercio e atividades recreativas. Adicionalmente é composta também por despesas na biodiversidade.

O valor de relevância para a componente com impacto negativo, “idade mediana” indica que valores mais elevados desta componente estão associados a municípios com população mais jovem, refletindo contextos socioeconómicos mais dinâmicos e maior disponibilidade de mão-de-obra ativa.

Em conjunto, esta componente permite distinguir municípios que, independentemente da sua intensidade turística, apresentam estruturas funcionais e socioeconómicas diferentes. Desta forma, distingue municípios com maior emprego em serviços ligados ao turismo, maior uso efetivo da capacidade instalada e população mais jovem, funcionando como uma dimensão de diferenciação complementar às duas outras componentes.

É relevante apontar ainda que a primeira e segunda componentes têm um maior "poder agregador" da informação recolhida das variáveis *input* utilizadas na análise, sendo a terceira componente utilizada com uma ótica complementar às anteriores.

Visão geral das componentes anteriores:

As três componentes estão relacionadas com a utilização e estrutura da oferta turística, mas em níveis diferentes:

- **Dimensão e Capacidade Económico-Turística e Cultura:** mostra o volume do turismo.
- **Intensidade Turística Relativa e a Vocação Balnear:** mostra o quão intenso é o turismo em relação ao número de habitantes.
- **Perfil Socioeconómico e Complementar do Turismo:** mostra como características complementares se relacionam com o turismo no município.

O elemento comum às três componentes é a forma como a oferta turística é utilizada no território.

Relação entre componentes

Com o objetivo de aprofundar a interpretação das dimensões, os componentes principais foram representados de forma gráfica, evidenciando semelhanças e contrastes entre os territórios.

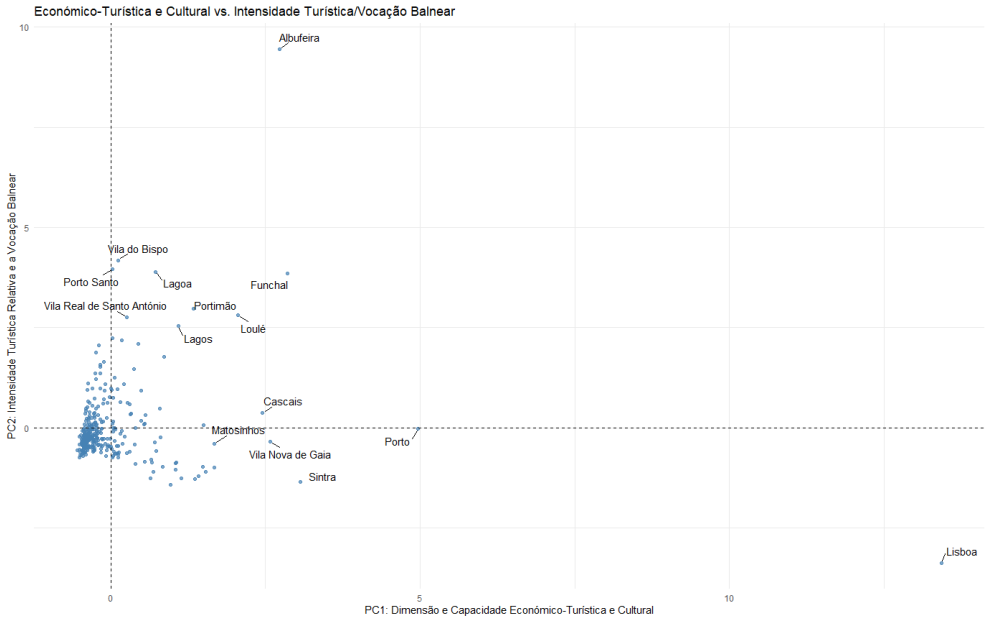

Figura 1 - Gráfico PC1 vs PC2

O gráfico da figura 1 mostra no eixo horizontal (PC1) a dimensão do turismo, ou seja, onde existe mais atividade turística no geral. No eixo vertical (PC2) está a importância do turismo balnear, isto é, até que ponto o turismo de praia é relevante para cada município.

A distribuição gráfica confirma uma forte assimetria territorial: enquanto a maioria dos municípios se agrupa na origem com relevância turística mínima, apenas uma minoria se destaca pela sua elevada capacidade económica ou especialização balnear.

Os pontos que estão mais afastados da origem, são dos municípios de Lisboa e Albufeira, mas por motivos diferentes. Lisboa, é o mais afastado da origem no eixo horizontal, com valores baixos no eixo vertical, o que significa que tem muita capacidade para receber turismo em termos globais. Aqui o índice do turismo assume características ligadas a outros fatores, podendo estes ser por exemplo: urbanos, culturais, históricos entre outros. Por outro lado, Albufeira aparece mais afastado no sentido do eixo vertical, o que nos indica, que é um destino bastante baseado em turismo balnear.

Existem ainda municípios como Funchal, Lagos, Portimão ou Loulé, que também são de turismo balnear, mas neste caso mostra que é possível ter turismo relevante e alguma vocação balnear em simultâneo, sem atingir níveis extremos como o observado em Albufeira. Isto ajuda-nos a perceber o tamanho do turismo no município, e alguma da sua natureza. É ainda possível concluir que no geral os municípios portugueses ocupam posições muito diferentes da capital e das zonas litorais.

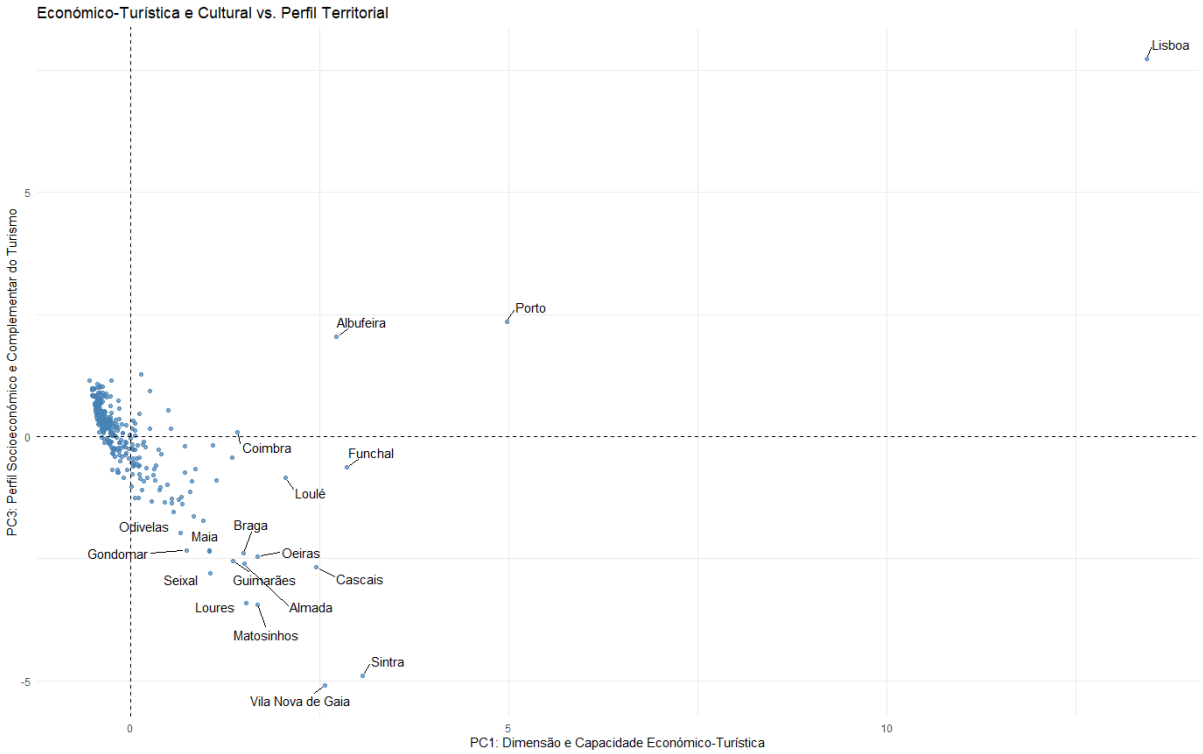

Figura 2 - Gráfico PC1 vs PC3

Este gráfico compara o tamanho do turismo em cada município com a forma como o turismo está organizado e funciona no território. No eixo horizontal (PC1) a dimensão do turismo, ou seja, onde existe mais atividade turística no geral, como no gráfico anterior. No eixo vertical (PC3) está o perfil socioeconómico e estrutural do turismo, que reflete aspetos ligados ao turismo, mas mais circundantes, como empregos e características da população.

Novamente a maioria dos municípios estão concentrados perto da origem, indicado turismo de pequena dimensão. Lisboa destaca-se claramente, mais afastado da origem, que mostra grande dimensão turística e um perfil estrutural muito próprio. Neste município existe muito emprego e serviços complementares que elevam a oferta turística.

O Porto aparece também afastado da origem no eixo horizontal, mas mais moderado no eixo vertical, o que sugere que, embora tenha dimensão turística relevante, o seu perfil estrutural é menor, ou seja, tem menos oferta de restaurantes, menos capacidade de alojamento, e menos atividades.

Alguns municípios metropolitanos, como Vila Nova de Gaia, Sintra ou Matosinhos, surgem com valores negativos, no eixo vertical, significando que apesar de também ter alguma dimensão turística, o turismo não é o foco central desses locais, sendo vocacionado para outras atividades económicas.

Este gráfico ajuda-nos a compreender que muito turismo não significa que funcione da mesma forma em municípios próximos, porque eles podem apresentar infraestruturas diferentes, o que se reflete na economia local. Ou seja, não importa apenas o volume de turistas, mas também o tipo de atividades, do emprego gerado e do contexto social.

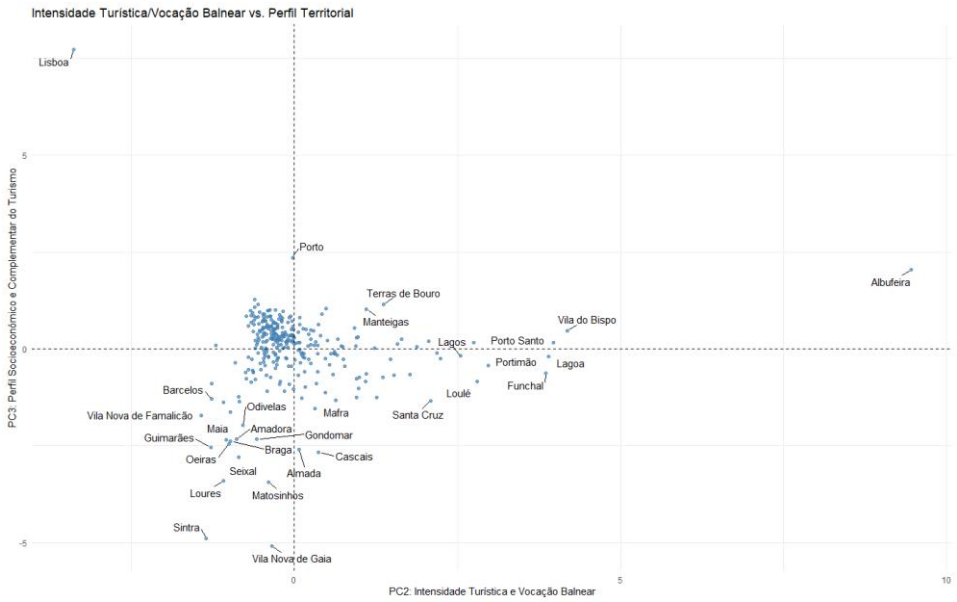

Figura 3 - Gráfico PC2 vs PC3

Este gráfico compara a intensidade do turismo balnear com a forma como o turismo está organizado e integrado na economia local. No eixo horizontal (PC2) como explicado no primeiro gráfico, mostra a importância balnear. No eixo vertical (PC3) como explicado no gráfico anterior, está o perfil socioeconómico e estrutural do turismo.

Neste gráfico a maioria dos municípios continuam perto da origem, mas estão mais dispersos. Indica que para a maioria o turismo balnear é pouco relevante, e quando existe turismo é pouco impactante na estrutura e na economia regional.

Ainda assim existem municípios que se destacam, como novamente Albufeira, com valores muito elevados no eixo horizontal, por se tratar de uma zona muito procurada pelas praias. No entanto, tem um posicionamento baixo, no eixo vertical, o que pode indicar pouca estrutura dada a procura turística elevada, sendo distinto de outros grandes centros urbanos como Lisboa, que surge com valores elevados no eixo vertical. No caso de Lisboa, os valores elevados de PC3 confirmam que o turismo tem forte impacto estrutural e socioeconómico, mas sem representatividade balnear, estando mais ligado a atividades urbanas, culturais e artísticas. É aqui que se percebe evidentemente as distintas formas de fazer turismo, com Lisboa e Albufeira representando comportamentos bastante distintos. Tal como observamos no gráfico anterior, a análise conjunta destes componentes, reforça que municípios com intensidade turística podem apresentar perfis estruturais distintos.

Representação geográfica das componentes

Após a análise conjunta das dimensões através dos gráficos, foi feita a representação geográfica das componentes. Estes mapas não introduzem novas dimensões de análise, mas permitem confirmar o que observamos anteriormente, e facilitam a leitura.

Mapa PC1 - Dimensão e Capacidade Económico-Turística e Cultural

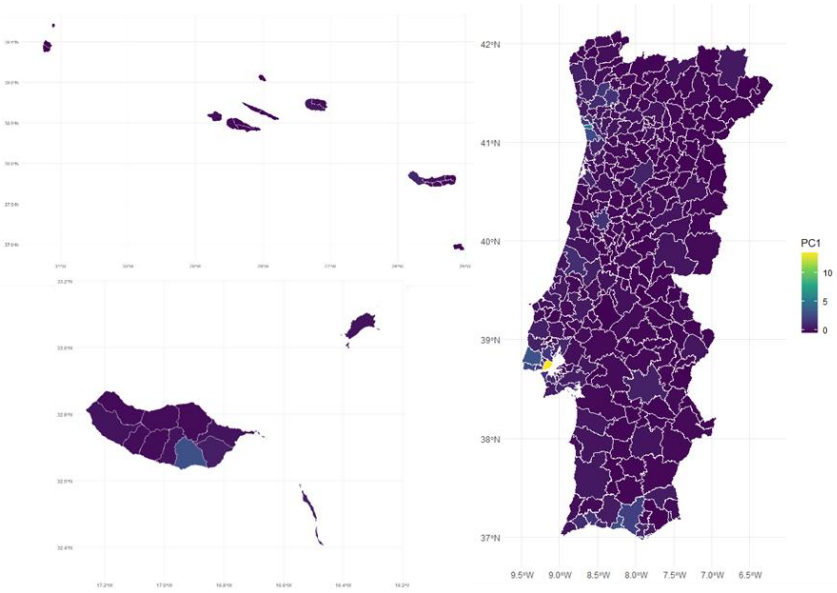

Figura 4 - Mapa Scores PC1

A representação geográfica da primeira componente confirma a forte concentração económica, turística e cultural, num número limitado de municípios, enquanto a maioria do território apresenta valores baixos.

Mapa PC2 - Intensidade Turística Relativa e a Vocação Balnear

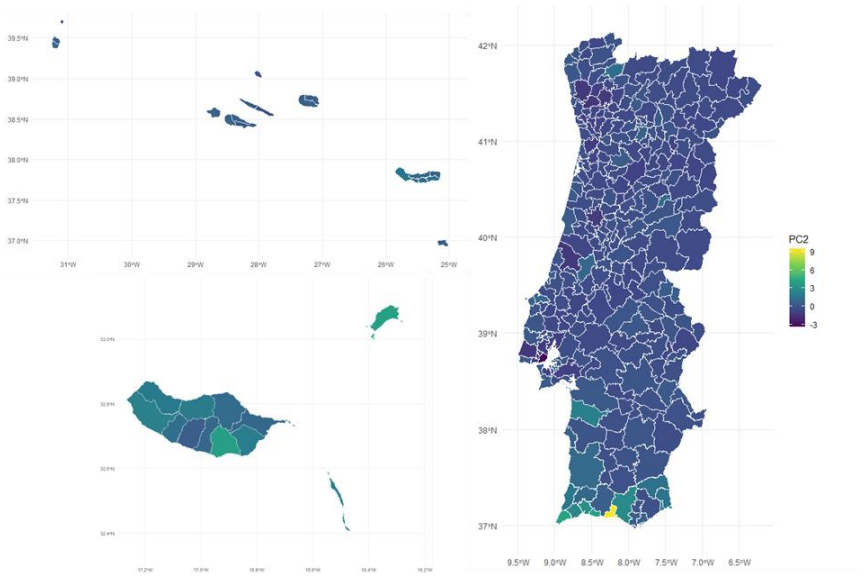

Figura 5 - Mapa Scores PC2

A distribuição geográfica desta componente destaca valores mais elevados sobretudo em zonas litorais, e nas regiões autónomas, confirmando a associação desta dimensão à intensidade turística balnear, em contraste com os valores reduzidos do interior.

Mapa PC3 - Perfil Socioeconómico e Complementar do Turismo

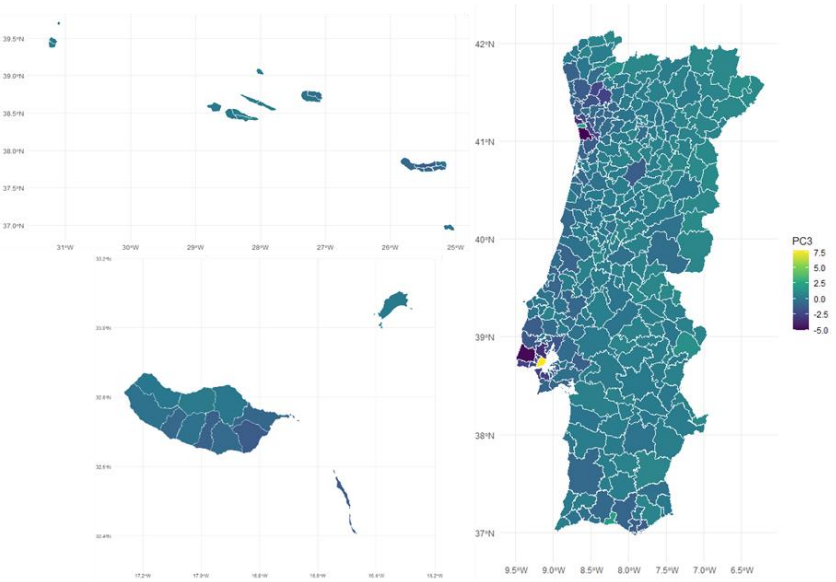

Figura 6 - Mapa Scores PC3

Por fim, o mapa da terceira componente demostra contrastes entre os municípios, com valores concentrados em grandes centros urbanos, como Lisboa e Porto, e valores baixos nos municípios envolventes, refletindo a diferença clara na estrutura socioeconómica e funcional do território.

Identificação da heterogeneidade da base de dados

Uma vez que a análise em componentes principais permite reduzir a complexidade da informação e identificar os principais eixos de diferenciação entre municípios esta não evidencia, por si, a existência de grupos distintos de municípios. Deste modo, através de uma análise de *clustering* é possível obter estes grupos. Foram realizados diferentes métodos de *clustering* (explorados no Anexo 2), sendo a melhor solução alcançada através do K-means, com a qual foram identificados 4 *cluster*s distintos: 1 – Competitivo Balnear, 2 – Lisboa, 3 – Competitivo Citadino e 4 – Menos competitivo. Sendo que os *cluster*s agrupam 21, 1, 27 e 259 municípios, respetivamente.

Tabela 4 - Perfil médio por cluster e nº de municípios

Cluster	**Perfil médio por *cluster***			**Nº de municípios**
	PC1	PC2	PC3	
1 – Competitivo Balnear	0,57	2,74	-0,2	21
2 – Lisboa	13,42	-3,38	7,74	1
3 – Competitivo Citadino	1,14	-0,63	-2,15	27
4 – Menos Competitivo	-0,22	-0,14	0,21	259

O *cluster* do Competitivo Balnear encontra-se principalmente localizado na Região do Algarve e no Arquipélago da Madeira, contudo ainda está presente no Distrito de Setúbal e nos Açores. O *cluster* de Lisboa, como o nome indica corresponde apenas ao município de Lisboa. O *cluster* Competitivo Citadino localiza-se principalmente nas Áreas Metropolitanas de Lisboa e do Porto, estando ainda presente no Algarve, Leiria e Viseu. Por fim, o *cluster* Menos Competitivo, uma vez que corresponde a maior parte dos municípios irá, consequentemente, corresponder também à maioria do território continental, estando também representado na maioria dos Açores e menos concentrado no Arquipélago da Madeira.

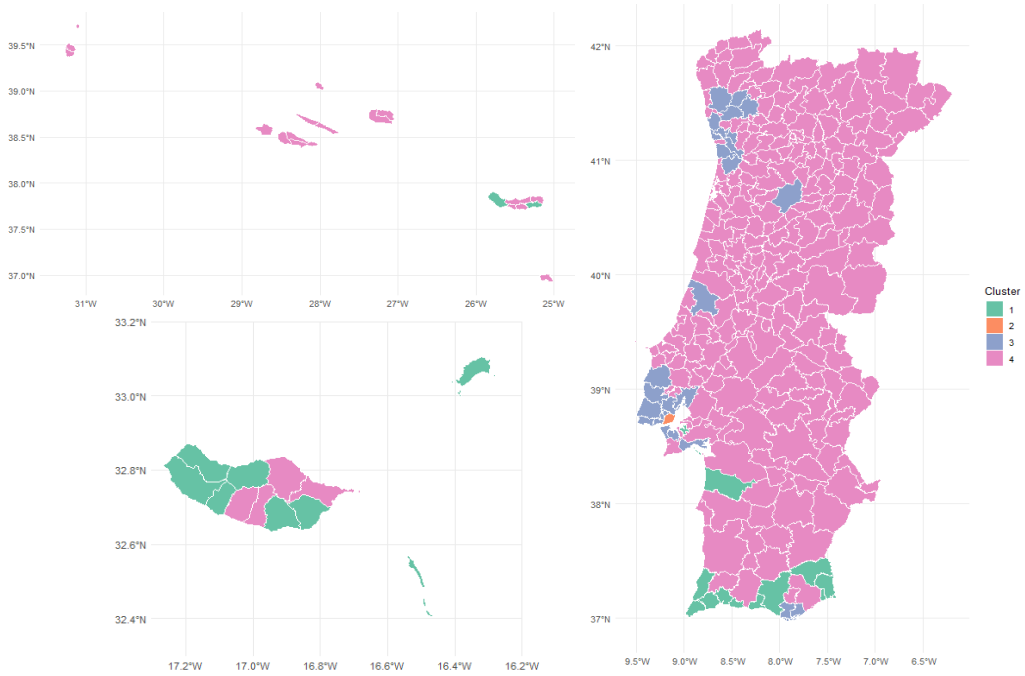

Figura 7 - Mapa de clusters de competitividade turística

Estes *cluster*s podem ainda ser caracterizados pelos valores das variáveis *Profile*, sendo estas divididas entre variáveis numéricas e categóricas.

Olhando para as tabelas com as variáveis numéricas é possível concluir que o *cluster* 1 apresenta uma baixa densidade populacional e reduzida população, sendo esta também envelhecida, com ganho médio mensal pouco acima dos 1100€. O *cluster* de Lisboa (2) apresenta uma densidade populacional muito elevada tal como a população total, contudo esta é consideravelmente envelhecida, apresentando ainda o maior ganho médio mensal, acima de 1800€. Já o *cluster* Competitivo Citadino (3) apresenta uma densidade populacional e população total elevadas, sendo esta a população menos envelhecida do país, com um ganho mensal considerável de 1300€. Por fim, o *cluster* 4 apresenta uma densidade populacional muito baixa e uma população total muito reduzida e envelhecida, apresentando ainda o menor ganho médio mensal com cerca de 1100€.

Tabela 5 - População e densidade populacional por cluster

Cluster	**População**			**Densidade populacional**		
	Min	Max	Média	Min	Max	Média
1	2517	105782	30046	17.0	1403.9	270.1
2	545796	545796	545796	5572.4	5572.5	5572.4
3	44614	385606	151683	197.9	7363.6	1530.4
4	384	231800	19579	4.3	5867.1	152.6

Tabela 6 - Índice de envelhecimento e ganho médio mensal

Cluster	**Índice envelhecimento**			**Ganho médio mensal**		
	Min	Max	Média	Min	Max	Média
1	109.1	338.9	194.5	938.8	1335.2	1121.2
2	169.5	169.5	169.5	1856.2	1856.3	1856.4
3	117.3	195.5	163.4	1085.5	2019.9	1303.3
4	65.6	720.1	281.6	902.8	2075.0	1105.3

Passando para as variáveis categóricas verifica-se que o *cluster* Competitivo Balnear (1) é representado totalmente por municípios que não são capitais de distrito e que são quase totalmente municípios litorais. Para o *cluster* 2 que corresponde a Lisboa é capital de distrito e que, nesta situação, se localiza no interior. No *cluster* 3 verifica-se que cerca de 20% dos municípios são capitais de distrito e que se encontram divididos quase igualmente entre o interior e o litoral. Por fim, para o *cluster* Menos Competitivo (4), é onde se encontram as restantes capitais de distrito que dentro do *cluster* correspondem a menos de 5% e os municípios do interior são mais de 80%.

Tabela 7 - # de municípios capital de distrito e interior/litoral por cluster

Cluster	**Capital de distrito**		**Interior/litoral**	
	Capital	Não capital	Interior	Litoral
1	0	21	2	19
2	1	0	0	1
3	5	22	15	12
4	12	247	210	49

Conclusão

Tendo como base o desafio proposto inicialmente, vemos que a solução de 4 *clusters* permite fazer uma diferenciação de municípios que permita segmentar a atribuição de fundos de apoio ao desenvolvimento turístico ("Programa Turismo 2030").

Começando pelo *cluster* de Lisboa, dado ser o *cluster* que representa unicamente a capital, vemos que poderá ser um dos municípios onde a necessidade de atribuição de fundos de apoio é claramente menos necessária, devido a ter já atualmente uma capacidade de alojamento forte e um volume de turistas elevados, bem como a estrutura complementar necessária para suportar o turismo de massas que visita a capital. A prioridade de fundos aqui deve ser mínima em expansão, focando-se antes na gestão da sustentabilidade e mitigação das externalidades negativas da elevada densidade turística.

No caso do *cluster* Competitivo Balnear, onde a preponderância do turismo tem um impacto relevante no dia-a-dia dos habitantes, e de certo modo transversal ao município, pode ser relevante explorar fundos que se foquem em alternativas ao turismo de massas (apoiando produtos de nicho como turismo de natureza ou desportivo), tentando reduzir a sazonalidade elevada mais associada a estes municípios, e garantindo uma maior diversificação das fontes de rendimento associadas ao setor.

Enquanto no caso do *cluster* Competitivo Balnear faz sentido uma aposta na redução da sazonalidade, no caso do *cluster* Competitivo Citadino, a procura turística é potencialmente mais uniforme ao longo do ano, mas pode representar tendências que não representam os padrões tradicionais do turismo (por exemplo estadias em âmbito de trabalho, deslocações ocasionais dentro do país, etc.), uma vez que estes municípios representam grandes centros urbanos e áreas envolventes. Ainda assim, o objetivo para este *cluster* deve ser a conversão de "visitas de um dia" em estadias de longa duração. Investimentos em infraestruturas culturais e reforço da capacidade de alojamento podem ser relevantes.

Por fim, relativamente ao *cluster* Menos Competitivo, que abrange a vasta maioria dos municípios e a maior extensão do território nacional, especialmente nas regiões do interior, a abordagem deve ser profundamente estratégica. Os fundos para este *cluster* devem funcionar como uma ferramenta essencial de coesão territorial e social. Em vez de grandes infraestruturas, o investimento deve priorizar projetos de pequena escala que valorizem o património endógeno e o potencial para o turismo de natureza, visando a criação de microeconomias que ajudem a fixar a população e a combater a desertificação destas regiões.

Bibliografia

Agência Portuguesa do Ambiente. (2025). *Programa COSMO*. Obtido de Programa de Monitorização da Faixa Costeira de Portugal Continental (COSMO): https://cosmo.apambiente.pt/geovisualizador

Forest-GIS. (2025). *Shapefiles e dados GIS de Portugal*. Obtido de Forest GIS: https://forest-gis.com/shapefiles-de-portugal/

Instituto Nacional de Estatística. (2025). *Base de dados de indicadores*. Obtido de Instituto Nacional de Estatística: https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_base_dados&contexto=bd&selTab=tab2

Anexos

Anexo 1 – Preparação dos dados para análise

Os dados recolhidos como base para análise foram recolhidos do site do Instituto Nacional de Estatística, sob o formato de XLS, com um conjunto de dados restantes para a criação de variáveis *profile* encontrando-se no formato CSV.

De forma a facilitar a importação e tratamento de dados foram criados inicialmente pastas para dividir os ficheiros correspondentes às variáveis *profile* e *input*, divididos ainda em subpastas conforme o formato do ficheiro. Dado o elevado nº de ficheiros em formato XLS, foi criada a função processar_ine_xls no R de modo a simplificar o processo, sendo o ficheiro *localização_cap_dist.csv* importado "manualmente", resultando em listas com as tabelas de cada um dos ficheiros de dados originais.

Após a importação dos ficheiros, foram realizados um conjunto de processos de modo a manter apenas as colunas relevantes, convertendo ainda as colunas necessárias para formato numérico e realizando a soma de colunas para obtermos variáveis mais relevantes.

```
193 ##################
194 # LIMPEZA DE DADOS
195 ##################
196
197 # INPUT - Converter colunas para numéricas
198 lista_dados_input <- lapply(lista_dados_input, function(df) {
199   cols_para_converter <- setdiff(names(df), c("codigo_municipio", "localizacao"))
200   df[cols_para_converter] <- lapply(df[cols_para_converter], function(x) as.numeric
    (as.character(x)))
201   return(df)
202 })
203
204 # INPUT - somar colunas de populaçao por niveis de educacao
205 lista_dados_input[[19]] <- lista_dados_input[[19]] %>%
206
207   mutate(
208
209     `Até 1º ciclo` = rowSums(across(2:3), na.rm = TRUE),
210     `Até 3º ciclo` = rowSums(across(4:5), na.rm = TRUE),
211     `Ensino Superior` = rowSums(across(7:11), na.rm = TRUE)
212   )
213
214 # INPUT - alterar nomes das colunas
215 names(lista_dados_input[[3]])[2] <- "Total"
216 names(lista_dados_input[[9]])[2] <- "Total"
```

Figura 8 - Exemplo de passos da limpeza/preparação

Após a seleção das colunas/variáveis relevantes, foram agregadas as colunas em 2 tabelas para utilização subsequente: dados_profile e dados_input. De forma a garantir a correta identificação da informação correspondente a cada coluna das tabelas foram revistos os nomes das colunas, e, no caso da tabela dados_profile, foi mantido o prefixo com o nome do

ficheiro original. Estas tabelas correspondem assim a 36 e 9 colunas, respetivamente, sendo as 2 primeiras colunas de cada tabela correspondentes ao nome e código do município, resultando assim num total de 34 variáveis *input* e 7 *profile*.

Foi ainda criada a tabela df_municipios, com o objetivo de manter a informação relevante de cada município para a criação dos mapas nas fases posteriores de análise. Esta tabela inclui o nome, código de município, e 2 outros códigos relevantes para o cruzamento de dados com o ficheiro GEOJSON para a construção dos mapas.

Data	
dados_input	308 obs. of 36 variables
dados_profile	308 obs. of 9 variables

Figura 9 - Tabelas finais de dados

Após a criação das tabelas base para a análise dos dados, foi validada a existência de resultados omissos. Ao selecionar as colunas da tabela dados_input com valores omissos e utilizar a biblioteca mice, foi observado que os valores omissos se encontravam apenas em 5 variáveis, sendo os omissos comuns em 3 das variáveis a 4 municípios: Mondim de Basto, Vizela, Moimenta da Beira e Murça. Dado tratar-se de uma amostra reduzida de observações omissas, optou-se pela imputação simples dos dados através da média.

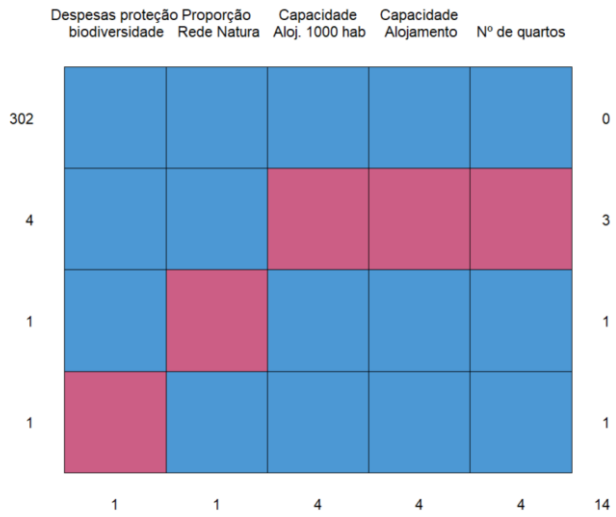

Figura 10 - Distribuição de valores omissos (inclui apenas as variáveis com valores omissos)

Anexo 2 – Análise de componentes principais

Adequabilidade dos dados à extração de componentes

Uma análise exploratória inicial, através de matrizes de dispersão e da matriz de correlações, revelou elevados níveis de correlação entre múltiplas variáveis, sugerindo redundância informacional e a adequação da aplicação de uma Análise de Componentes Principais (PCA).

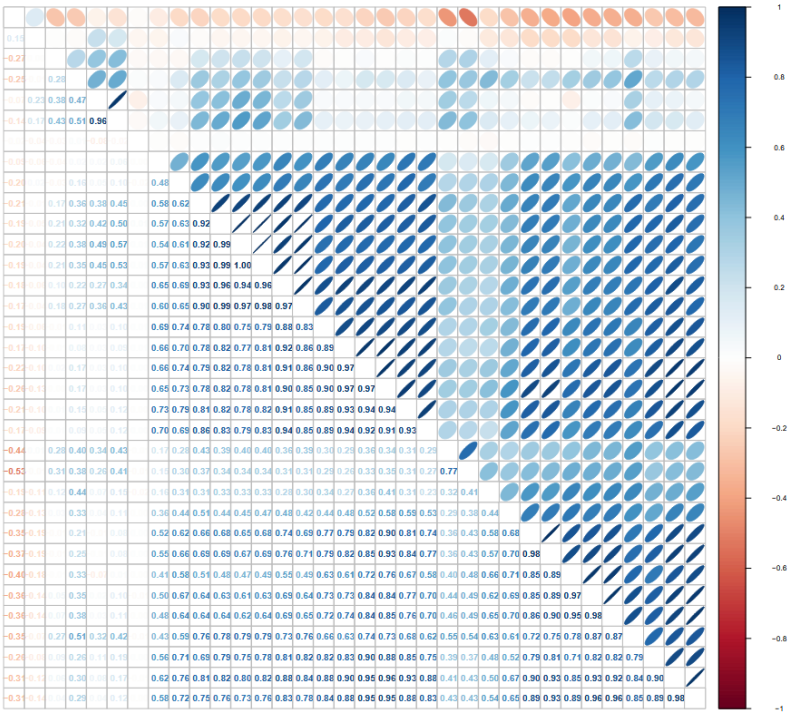

Figura 11 - Estrutura de correlações das variáveis input

A adequação da base de dados à aplicação da Análise de Componentes Principais foi avaliada recorrendo a testes estatísticos clássicos. Inicialmente foi aplicado o teste de Bartlett, que rejeita claramente a hipótese nula de que a matriz de correlação é uma matriz identidade (p-value ≈ 0), indicando a existência de correlações estatisticamente significativas entre as variáveis, justificando a aplicação da PCA.

Adicionalmente, a medida de Kaiser-Meyer-Olkin (KMO) apresenta um valor global de aproximadamente 0,91, o que corresponde a um nível considerado excelente segundo os critérios da literatura, confirmando que a estrutura de correlações é adequada para uma redução dimensional.

Seleção do número de componentes principais

Após a estimação inicial do modelo com o número máximo de componentes (igual ao número de variáveis), procedeu-se à análise dos valores próprios (eigenvalues) associados a cada componente.

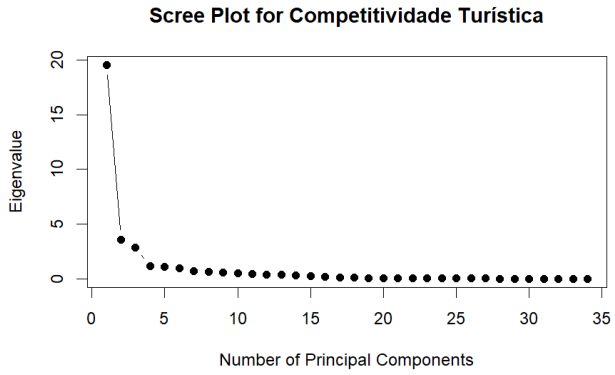

Figura 12 – Scree plot

A análise visual do gráfico de declive (*scree plot*) indicou uma inflexão acentuada ("cotovelo") na terceira componente, sugerindo que a retenção de 3 componentes capturaria a maior parte da variância estrutural relevante, sendo as restantes maioritariamente variância residual ou ruído.

No critério de Kaiser (retenção de componentes com valor próprio superior a 1), observando os valores próprios da extração inicial, 5 componentes apresentavam valores superiores a 1. Embora este critério indique a possibilidade de reter um número superior de componentes, este critério tende a sobrestimar a dimensionalidade em contextos com um número elevado de variáveis, como é o caso do presente estudo.

Adicionalmente, a análise da variância explicada acumulada mostra que a retenção de três componentes permite captar uma proporção substancial da variância total da base de dados (76,4%), assegurando simultaneamente um elevado grau de síntese e interpretabilidade.

Tabela 8 - Resumo de análise de componentes principais

Estatística	PC1	PC2	PC3	PC4	PC5	PC6	PC7	…
SS Loadings (eigenvalue) *(CRITÉRIO DE KAISER)*	19.59	3.54	2.84	1.16	1.05	0.97	0.70	…
Proporção da variância	57,6%	10,4%	8,4%	3,4%	3,1%	2,9%	2,1%	…
Variância acumulada	57,6%	68,0%	76,4%	79,8%	82,9%	85,8%	87,9%	…

Optou-se pela solução de 3 componentes pois, após testes exploratórios, esta solução ofereceu o melhor compromisso entre variância explicada e interpretação semântica dos eixos para o contexto turístico.

Rotação e Interpretação

Após a seleção das três componentes principais, foi aplicada uma rotação ortogonal do tipo Varimax. Este método maximiza a variância dos *loadings* (pesos) dentro de cada componente, polarizando as correlações (tornando os valores altos mais altos e os baixos mais próximos de

zero), o que facilitou a atribuição de nomes às novas dimensões latentes (PC1, PC2 e PC3) utilizadas nas fases subsequentes.

Qualidade da representação e cálculo dos scores

Após a rotação, avaliou-se a qualidade da representação das variáveis originais através das comunalidades. Verificou-se que a maioria das variáveis apresenta valores elevados (superiores a 0.5), o que indica que as três componentes retidas conseguem recuperar uma parte substancial da variância de cada indicador individual. Variáveis com comunalidades baixas indicariam uma fraca explicação pelo modelo, mas a solução final demonstrou-se robusta.

Como passo final da Análise em Componentes Principais, procedeu-se à estimação dos *scores* fatoriais para cada um dos 308 municípios nas três dimensões retidas. Estes *scores*, que constituem as novas variáveis quantitativas para a fase subsequente de *Cluster Analysis*, foram calculados de forma padronizada (média 0 e desvio-padrão 1), o que facilita a interpretação da posição relativa de cada concelho face ao panorama nacional. Nesta escala, o valor 0 representa a média exata do país, enquanto valores positivos ou negativos indicam, respetivamente, uma performance acima ou abaixo dessa média, sendo a magnitude medida em unidades de desvio-padrão. Os *scores* das componentes principais, juntamente com as variáveis *input*, são depois exportados para o ficheiro PCA_Turismo.xlsx.

Anexo 3 – Clustering de observações

Após a estimação dos scores das componentes, passamos à fase de *clustering*, na qual foram aplicados diferentes modelos, de forma a analisar qual apresentava uma estrutura que melhor representasse os dados obtidos.

Começamos pelo cálculo da matriz de distâncias usando a distância euclidiana, com base nos *scores* das componentes principais selecionadas, resultando no objeto demodist. Esta matriz representa assim a proximidade de cada município, sendo depois disso usada como *input* para métodos de *clustering*.

Clustering hierárquico

Aplicou-se seguidamente *clustering* hierárquico utilizando 3 métodos diferentes de ligação: complete, single e Ward.D2. O método Ward.D2 foi escolhido para prosseguir com a análise por produzir clusters menos suscetíveis a outliers, dadas as *scores* extremadas de alguns municípios.

Depois de aplicado o *clustering* hierárquico foi realizado o teste de corte em três diferentes níveis k=3, 4 e 5 e foram analisados o número de observações por cluster, o perfil médio de

cada cluster nas componentes principais e a silhueta de cada observação, para avaliar a qualidade do agrupamento.

Tabela 9 - Resultados clustering hierárquico

K Clusters	K = 3	K = 4	K = 5
Nº de elementos por grupo	Grupo 1: 266 Grupo 2: 41 Grupo 3: 1	Grupo 1: 266 Grupo 2: 14 Grupo 3: 1 Grupo 4: 27	Grupo 1: 206 Grupo 2: 60 Grupo 3: 14 Grupo 4: 1 Grupo 5: 27
Perfil médio por cluster *(PC1, PC2, PC3)*	Grupo 1: -0,18; -0,19; 0,16 Grupo 2: 0,83; 1,30; -1,24 Grupo 3: 13,42; -3,38; 7,74	Grupo 1: -0,17; -0,18; 0,16 Grupo 2: 1,56; -0,72; -2,95 Grupo 3: 13,42; -3,38; 7,74 Grupo 4: 0,45; 2,35; -0,35	Grupo 1: -0,32; -0,13; 0,38 Grupo 2: 0,30; -0,38; -0,61 Grupo 3: 1,56; -0,72; -2,95 Grupo 4: 13,42; -3,38; 7,74 Grupo 5: 0,45; 2,35; -0,35
Silhueta	0,6	0,6	0,43

Observou-se que K = 3 e k=4 apresentaram uma silhueta de 0,6, indicando *clusters* bem definidos e homogéneos. Valores maiores de K, como K = 5 não melhoraram significativamente a separação dos grupos, sendo que K = 5 apresentou uma silhueta menor (0,43), sugerindo *clusters* menos coerentes.

Clustering partitivo

Após isso aplicamos também uma abordagem de *clustering* partitivo através da aplicação do método K-means. Para determinar o número apropriado de *cluster*s para K-means, aplicou-se o método do cotovelo. O gráfico resultante mostra que, a partir de K=4, a redução na soma total de quadrados intra-*cluster* (WSS) deixa de ser significativa, indicando o ponto de equilíbrio entre a complexidade do modelo e a qualidade do agrupamento.

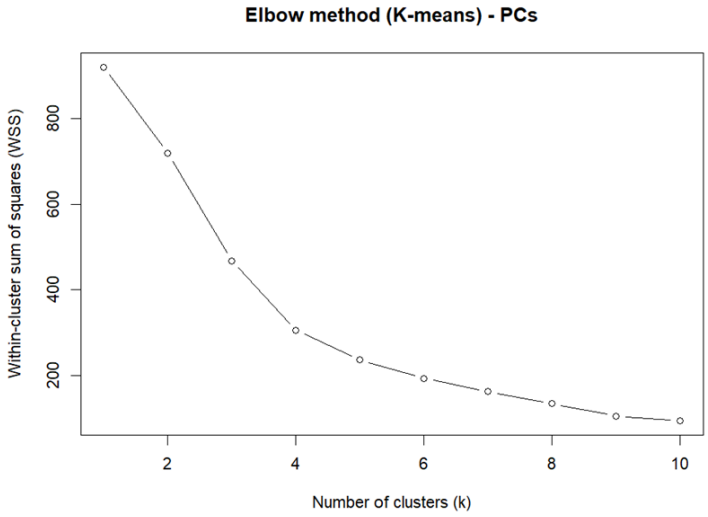

Figura 13 - Método do cotovelo para a aplicação de K-means

Com base nos resultados do método do cotovelo foi realizado o *clustering* K-means para K = 4 e K = 5, sendo K=4 o resultado com a silhueta mais alta (0,61).

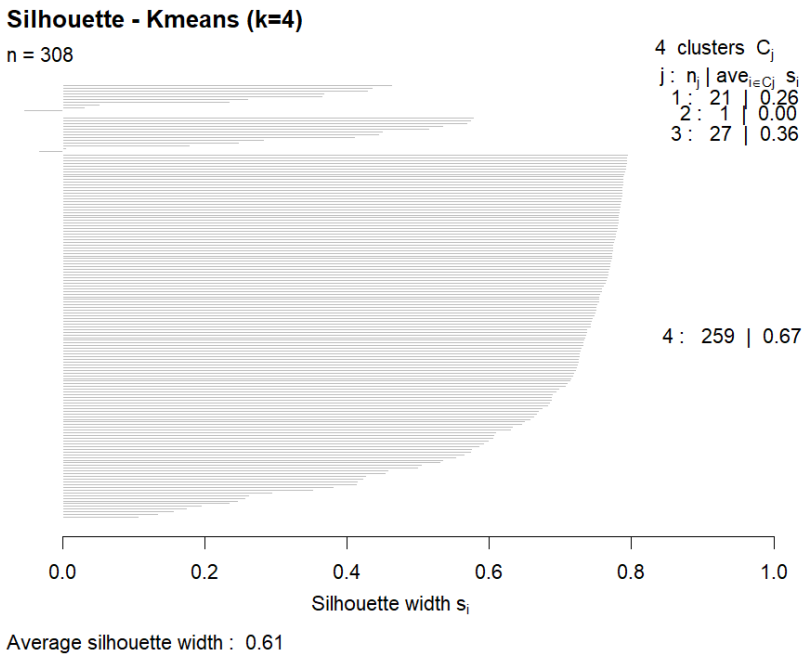

Figura 14 - Silhueta para K-means com k=4

Seguidamente foi ainda aplicado o método PAM, com K = 4. Esta técnica é mais robusta a outliers que o K-means, pois utiliza os medoides em vez das médias dos clusters (centroides). Apesar disso, os resultados de silhueta alcançados com o método PAM apresentaram uma silhueta mais baixa de 0,38, sugerindo agrupamentos menos coerentes. Como é possível ver nos resultados da tabela abaixo, uma das principais diferenças surge no facto de, na abordagem PAM, não termos um cluster isolado para Lisboa, como ocorre nas outras

soluções. Assim, para este conjunto de dados, K-means ou o *clustering* hierárquico fornecem uma solução de clusters mais adequada que o PAM.

Tabela 10 - Comparação de resultados de soluções partitivas e hierárquicas

Clusters	**PAM** **K = 4**	**K-MEANS** **K = 4**	**H CLUST** **K = 4**
Nº de elementos por grupo	Grupo 1: 176 Grupo 2: 86 Grupo 3: 21 Grupo 4: 35	Grupo 1: 21 Grupo 2: 1 Grupo 3: 27 Grupo 4: 259	Grupo 1: 266 Grupo 2: 14 Grupo 3: 1 Grupo 4: 27
Perfil médio por cluster (*PC1, PC2, PC3*)	Grupo 1: -0,27; -0,25; 0,51 Grupo 2: 0,10; -0,01; -0,41 Grupo 3: 1,30; -0,85; -2,42 Grupo 4: 0,49; 2,48; -0,12	Grupo 1: 0,57; 2,74; -0,20 Grupo 2: 13,42; -3,38; 7,74 Grupo 3: 1,14; -0,63; -2,15 Grupo 4: -0,22; -0,14; 0,21	Grupo 1: -0,17; -0,18; 0,16 Grupo 2: 1,56; -0,72; -2,95 Grupo 3: 13,42; -3,38; 7,74 Grupo 4: 0,45; 2,35; -0,35
Silhueta	0,38	0,61	0,6

Clustering probabilístico

Por fim, para complementar e validar as abordagens de particionamento rígido anteriores, aplicou-se um modelo de mistura finita gaussiana (*Gaussian Mixture Models* - GMM). Ao contrário do K-Means, que assume clusters esféricos, o GMM ajusta-se à geometria dos dados.

O algoritmo Mclust selecionou automaticamente o modelo VEV (Ellipsoidal, Equal shape) como o que melhor se ajusta aos dados (BIC = -976.69). Isto indica que os clusters turísticos portugueses assumem formas elipsoidais (alongadas) e não esféricas, confirmando que métodos como o K-Means podem estar a simplificar excessivamente a realidade territorial.

A aplicação deste modelo obedeceu a uma estratégia de validação comparativa, tendo-se testado duas configurações distintas quanto ao número de clusters:

- **K=4 (Cenário de Comparabilidade):** Fixou-se o número de grupos em 4 para permitir uma comparação direta com a solução de K-Means que apresentou melhor desempenho nas fases anteriores. O objetivo foi validar se a estrutura de macro-segmentos (Elite vs. Baixa Intensidade) se mantinha estável quando relaxada a restrição de esfericidade dos clusters.
- **K=6 (Cenário de Granularidade):** Testou-se uma solução com maior número de grupos, (que foi dada como ótima pelo modelo) para investigar a heterogeneidade latente. Pretendia-se verificar se os grandes grupos de "baixa intensidade" poderiam ser fragmentados em nichos mais específicos (e.g., distinguir turismo rural incipiente

de ausência total de turismo), avaliando se o ganho em detalhe compensava a penalização da complexidade do modelo (BIC).

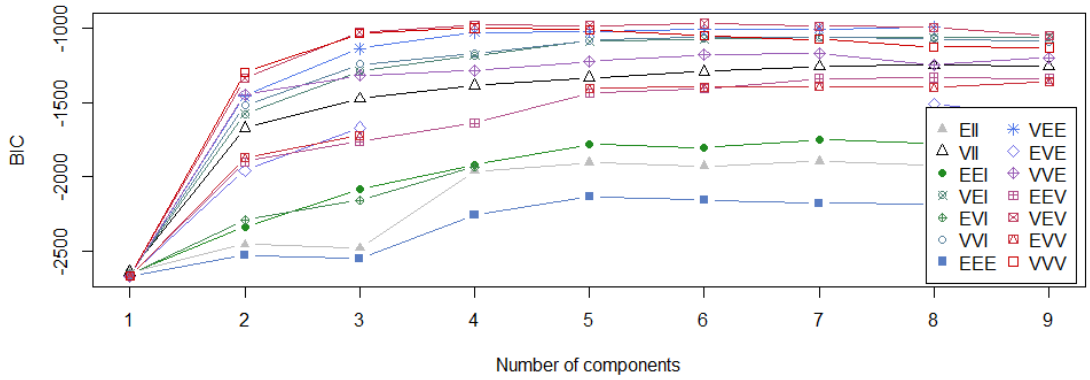

Figura 15 - Decisão do nº de clusters ótimos para GMM (k=6 VEV)

Uma das principais vantagens desta análise foi a identificação da incerteza de classificação. O gráfico de incerteza gerado permitiu visualizar os municípios que se encontram em "zonas de transição" entre perfis turísticos, onde a probabilidade de pertença não é decisiva (ex: 50/50).

Abaixo apresentam-se os resultados visuais e numéricos para a solução de 4 clusters (K=4):

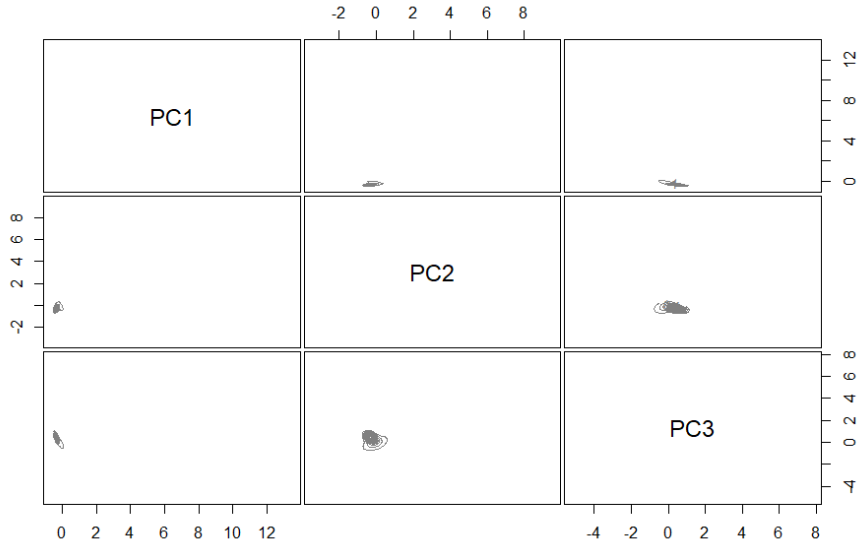

Figura 16 - Gráfico de Densidade (GMM)

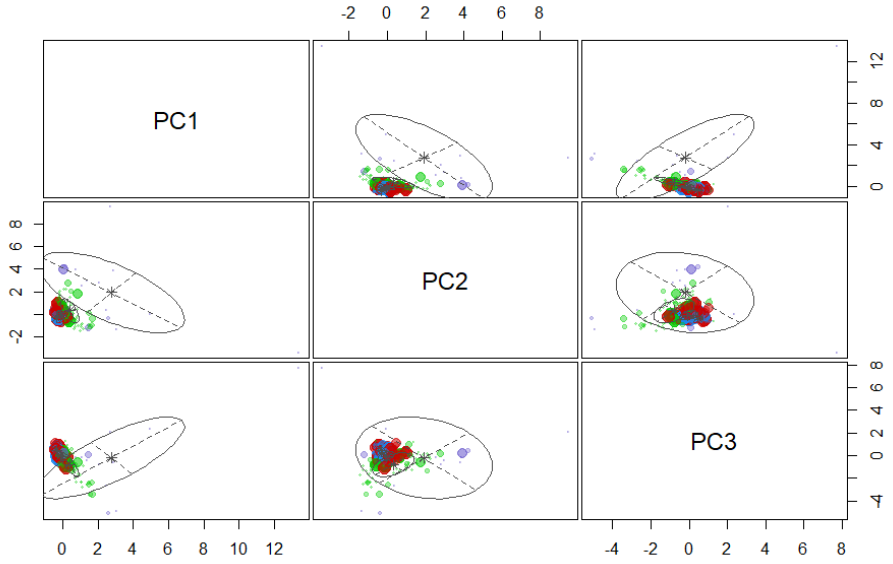

Figura 17 - Gráfico de Incerteza (GMM)

A análise dos perfis médios (Tabela 4) confirma a estrutura detetada anteriormente:

Tabela 11 - Resultados Cluster Probablistico

Parâmetro	**Grupo 1**	**Grupo 2**	**Grupo 3**	**Grupo 4**
Dimensão (nº municípios)	**147**	**80**	**67**	**14**
Médias de Componentes				
PC1	-0,38	-0,15	0,38	**2,75**
PC2	-0,31	-0,07	0,31	**1,92**
PC3	0,45	-0,06	-0,80	-0,21

- Grupo 4 (Elite Turística): Um grupo reduzido de 14 municípios (vs. 21 no K-Means) que apresenta scores médios extremamente elevados no PC1 (2.75) e PC2 (1.92). O GMM foi mais seletivo que o K-Means, isolando apenas os *outliers* mais fortes.
- Grupo 1 (Baixa Intensidade): O maior grupo (147 municípios), caracterizado por valores negativos em ambas as dimensões principais, representando a "média" do interior ou zonas sem vocação turística marcada.
- Grupos 2 e 3 (Grupos intermédios): Os Grupos 2 e 3 representam zonas de transição. O Grupo 3 (67 municípios), por exemplo, já apresenta uma média positiva no PC1 (0,38) e PC2 (0,31), sugerindo concelhos com alguma vitalidade turística emergente, mas ainda distantes da escala do Grupo 4.

Embora o modelo probabilístico (GMM) tenha demonstrado robustez teórica ao identificar a geometria elipsoidal dos dados, a análise comparativa revelou uma diferença crucial na estruturação dos grupos de topo.

Enquanto o GMM agregou os outliers num único cluster de 14 municípios (misturando o caso extremo de Lisboa com outros polos turísticos), a solução de K-Means com K=4 ofereceu uma granularidade superior no topo da pirâmide, isolando Lisboa num cluster unitário e identificando um grupo distinto de 21 municípios de alta performance ("Elite Turística").

Face a isto, optou-se pela solução de **K-Means (K=4)** para a caracterização final no corpo do relatório, fundamentada em dois critérios:

- **Utilidade para *Benchmarking*:** Ao isolar o caso único de Lisboa (Cluster 2), o K-Means permite focar a análise de *benchmarking* no Cluster 1 (21 municípios). Isto é mais útil para o contexto, pois fornece aos decisores um grupo de referência realista de casos de maior competitividade turística replicáveis, separando-os da realidade incomparável da capital, Lisboa**.**
- **Parcimónia e Clareza:** A solução K-Means apresenta um valor de Silhueta global de 0,61, o que garante estatisticamente que os grupos estão bem separados e são coesos. Esta métrica oferece uma validação direta, garantindo a parcimónia necessária para a definição de políticas públicas, ao evitar a complexidade e a ambiguidade das probabilidades de pertença (zonas de incerteza) geradas pelo GMM.