# Análise de Clusters

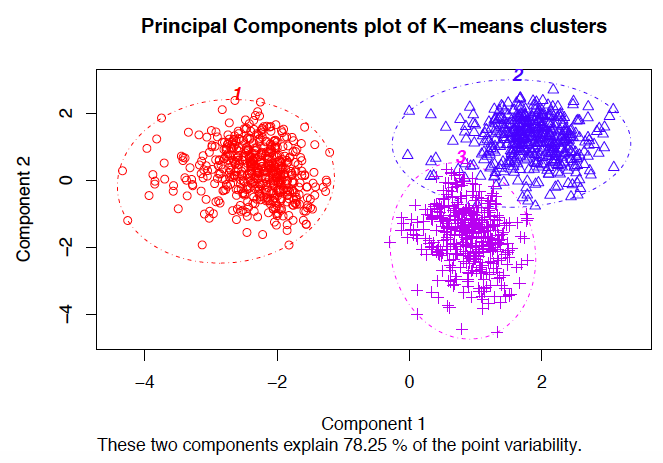
Out 10,2012

Fonte: <https://analise-estatistica.pt/2012/10/analise-de-clusters.html>

A análise de clusters é um procedimento estatístico multivariado que serve para identificar grupos homogéneos nos dados, com base em variáveis ou em casos.

A análise de clusters, ou análise de agrupamentos, permite assim classificar objetos e pessoas com base na observação das semelhanças e das dissemelhanças: dado um conjunto de *n* indivíduos, sobre os quais existe informação de *p* variáveis, o método agrupa os indivíduos em função da informação existente, de modo que os indivíduos de um grupo sejam tão semelhantes entre si quanto possível e tão diferentes dos restantes grupos quanto possível.

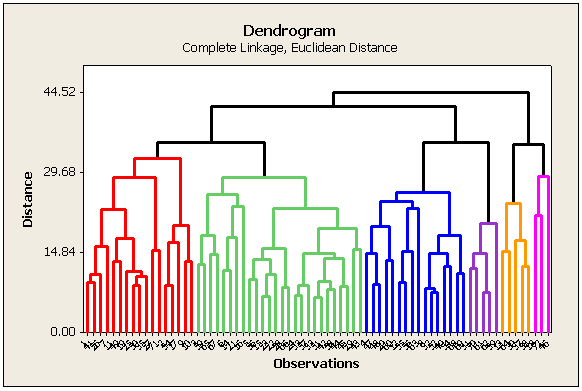
A análise de clusters pretende organizar um conjunto de casos em grupos homogéneos, de tal modo que os indivíduos pertencentes a um grupo são o mais semelhante possível entre si e diferentes dos restantes.



## Quando recorrer à análise de clusters

A análise de clusters é um bom procedimento quando se suspeita, na fase de exploração dos dados, que a amostra não é homogénea, isto é:

* Quando numa análise univariada se observa mais de uma moda (cf. histograma, diagrama de caule e folhas ou gráficos circulares);
* Quando numa análise bivariada se observa, pelo diagrama de dispersão, que os dados estão dispostos em grupos;
* Quando numa análise multivariada o dendograma aponta para a formação de grupos dos variáveis ou de casos.



A formação de grupos pode ter implicações teóricas e práticas importantes, como no caso da medicina onde, por exemplo, a classificação correta das perturbações mentais pode ajudar na descoberta das suas causas e dos seus tratamentos; assim como na pesquisa de mercado, onde a classificação correta dos clientes com iguais preferências permite identificar nichos de mercado.

Neste procedimento de análise, a escolha das variáveis a incluir na análise de clusters é crucial, porque a inclusão ou a exclusão de uma determinada variável pode significar resultados e conclusões bem diferentes. A escolha das variáveis determina quais as caraterísticas que irão identificar os grupos. Se considerarmos, por exemplo, apenas a cor e o sabor do vinho, estamos a deixar de fora outras caraterísticas como a cor, o cheiro, a acidez e a concentração de glucose, por exemplo.

## Análise de clusters de variáveis

A análise de clusters aplica-se tanto a variáveis quantitativas, discretas ou contínuas, como à contagem de dados (frequências), ou ainda a dados binários.

A análise de clusters de variáveis é semelhante à análise fatorial, porque ambos os procedimentos identificam grupos de variáveis relacionadas entre si. Nesta situação, a opção pela análise de clusters é mais rara. A análise fatorial é preferível porque é um modelo mais teórico, enquanto a análise de clusters é um procedimento mais ad hoc.

## A análise de clusters de casos

A análise de clusters de casos é semelhante à análise discriminante, uma vez que procura classificar um conjunto de dados iniciais em grupos ou categorias, usando os valores observados das variáveis que se referem ao fenómeno em estudo, não se conhecendo nem o número de grupos, nem o número de membros dos grupos.

## As cinco etapas da análise de clusters:

1. A seleção dos casos a serem agrupados;
2. A definição de um conjunto de variáveis a partir das quais será obtida a informação necessária ao agrupamento dos casos;
3. A seleção de uma medida de semelhança ou de distância entre cada par de casos;
4. A escolha de um critério de agregação ou de um critério de desagregação dos casos;
5. A validação dos resultados encontrados.

Os métodos de agrupamento das variáveis podem ser hierárquicos ou não hierárquicos. A formação de clusters com base nos métodos hierárquicos é feita com base no agrupamento nos pares de casos mais próximos de acordo com a medida de distância escolhida. O algoritmo continua passo a passo, juntando pares de casos, pares de clusters, ou um objecto com um cluster, até que todos os dados estejam num só cluster. [Saber mais sobre os métodos hierárquicos de agrupamento](https://analise-estatistica.pt/inicio/2017/02/metodos-hierarquicos-da-analise-clusters.html).

Fonte: Pestana, M. H. & Gageiro, J. N. (2005). Análise de Dados para Ciências Sociais: A Complementaridade do SPSS. Lisboa, Edições Sílabo.