UFMG Cedeplar

Análise de Clusters

Simpósio Modelagem para Avaliação de Imóveis

Prefeitura de Belo Horizonte, 04/06/2019

Pedro Amaral

Ph.D. em Land Economy pela University of Cambridge

Professor do Departamento de Economia e Cedeplar/UFMG

Fellow do Center for Spatial Data Science (University of Chicago)

Análise Multivariada

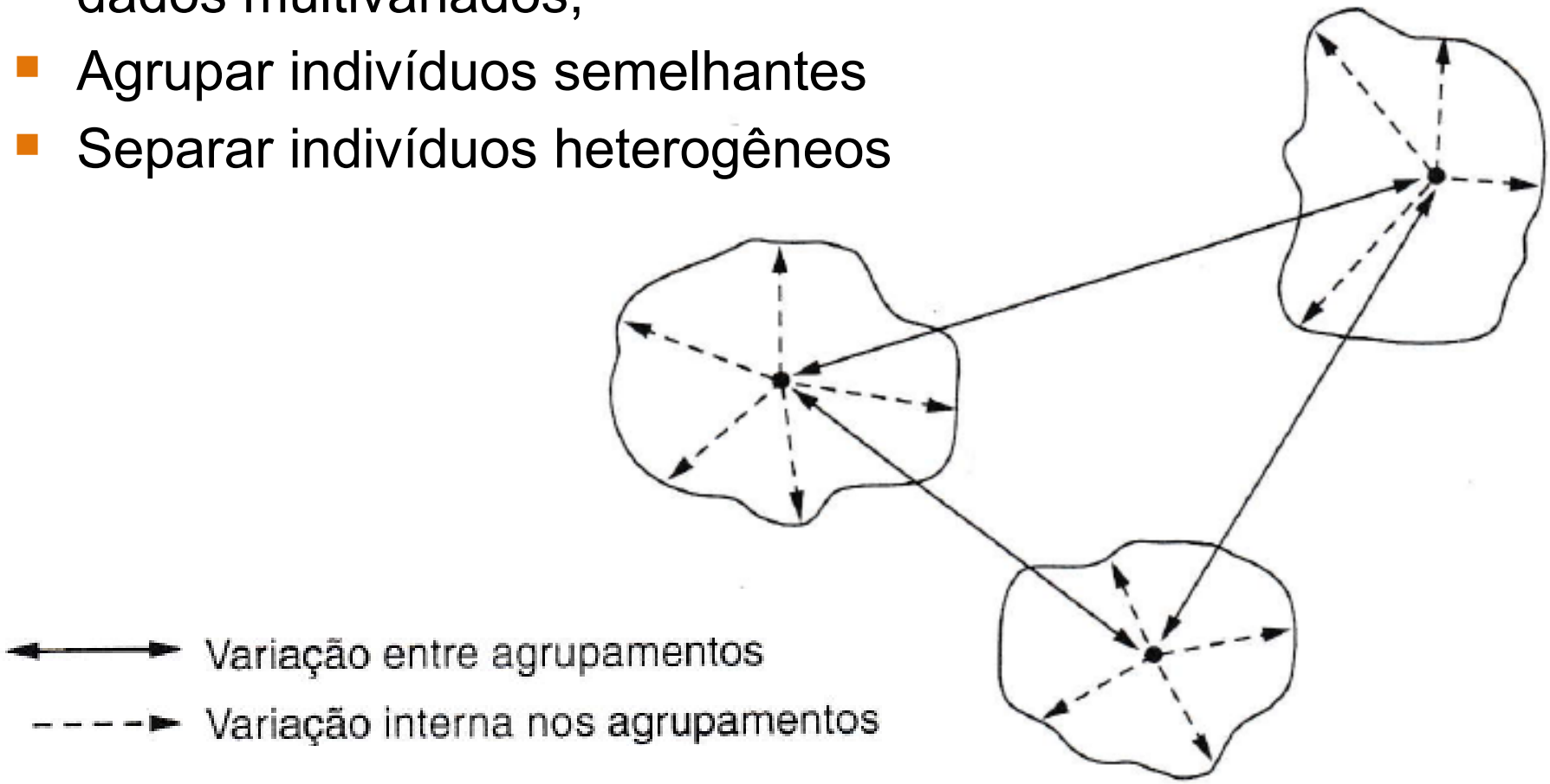
- Marriot, F. H. C. (1974):

"If the results disagree with informed opinion, do not admit a simple logical interpretation, and do not show up clearly in a graphical presentation, they are probably wrong. There is no magic about numerical methods, and many ways in which they can break down. They are a valuable aid to the interpretation of data, not sausage machines automatically transforming bodies of number into packets of scientific fact".

Métodos de Análise Multivariada se aproximam muito mais das **artes** que das ciências.

Análise de Clusters

- Análise de clusters é a arte de encontrar grupos em dados multivariados;
- Agrupar indivíduos semelhantes
- Separar indivíduos heterogêneos



Análise de Clusters

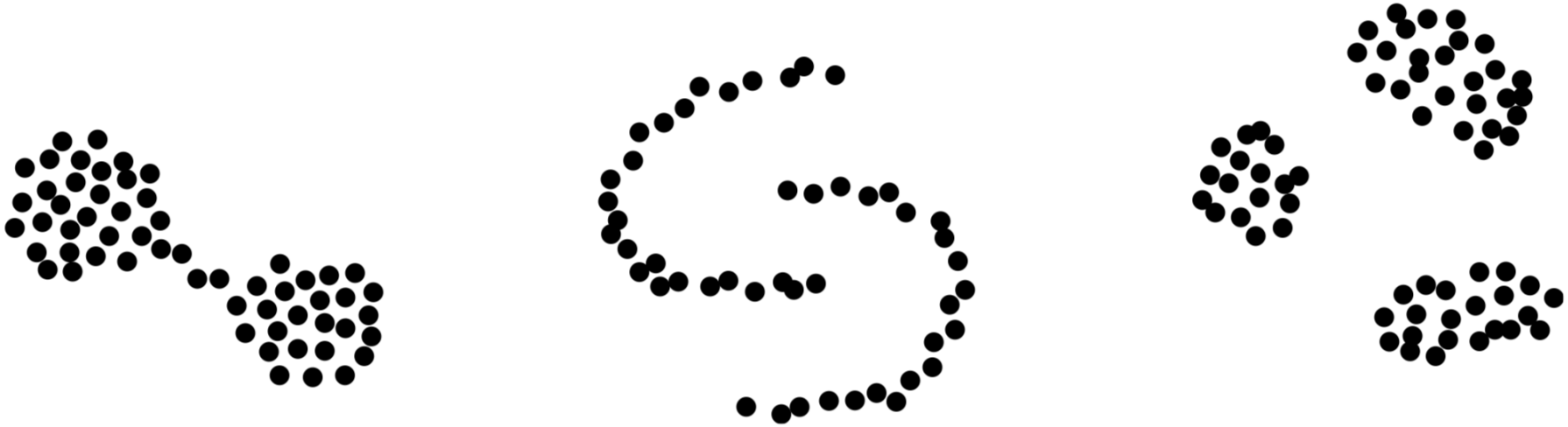
- Pode ser usada tanto para identificar grupos em conjunto de dados quanto para dividir conjunto de dados de maneira “justa”.

- Objetivos:
  - Classificar objetos (taxonomia);
  - Reduzir número de objetos;
  - Identificação de relações

- GI-GO
  - Classificação livros segundo tema (fantasia, romance, etc.) ou cor da capa (livro vermelho, azul, etc.)

Análise de Clusters

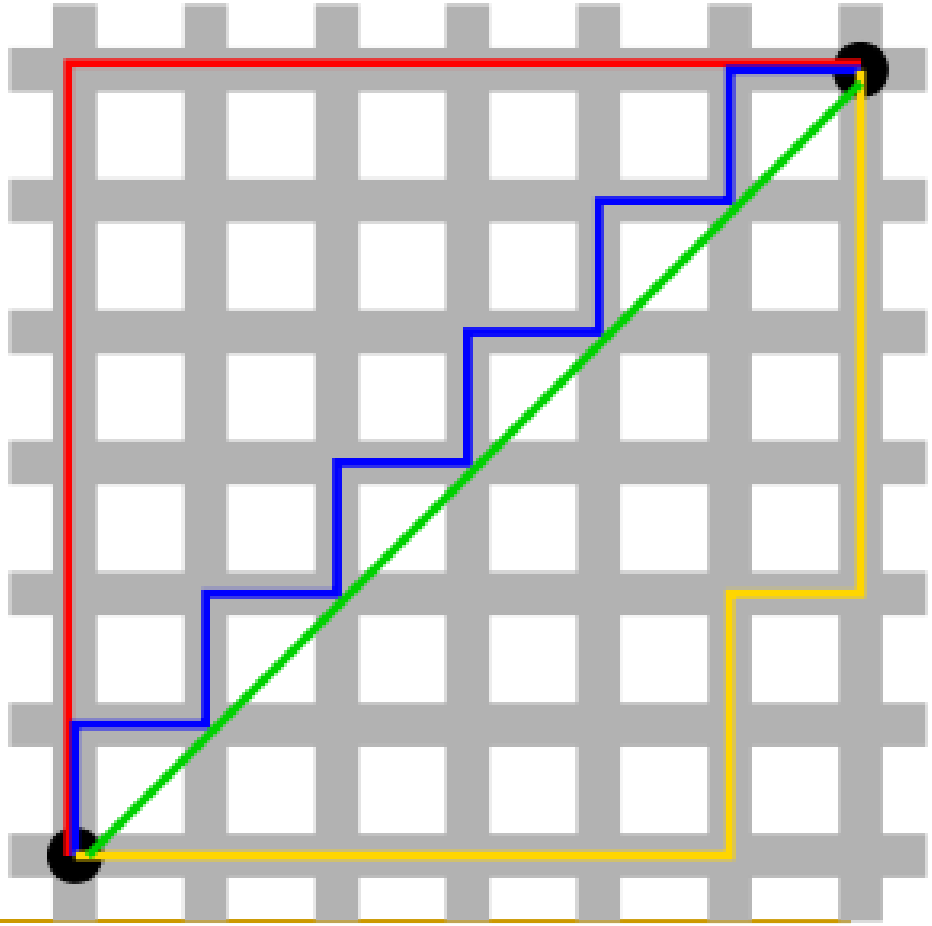
- Três questões principais:
  - Como medir semelhança entre objetos?
  - Como agrupar objetos semelhantes?
  - Como avaliar semelhança dos grupos?



Medidas de (dis)similaridade

- Medidas de associação.
- Medidas de correlação;
- Medidas de distância;
  - Euclidiana
  - Problema:
    - Multicolinearidade:

$$d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \cdots + (x_{ip} - x_{jp})^2}$$



Outros dados

- Outras possibilidades:
  - Dados categóricos não binários
    - Scores
  - Dados de proporções
    - Ignorar
    - Considerar em log
    - Considerar como ordinal
  - Dados mistos
    - Categorizar todas variáveis
    - Procedimento de Gower

Como agrupar objetos semelhantes?

- Métodos hierárquicos;
  - Resultado não ótimo para número específico de grupos
- Métodos de partição (não-hierárquicos);
  - K-means
    - Envolve o calculo da média (centroide) de cada cluster
  - K-median
    - Ao invés de usar a média do cluster, baseia-se na mediana do grupo, ou objeto central / representativo.

K-means

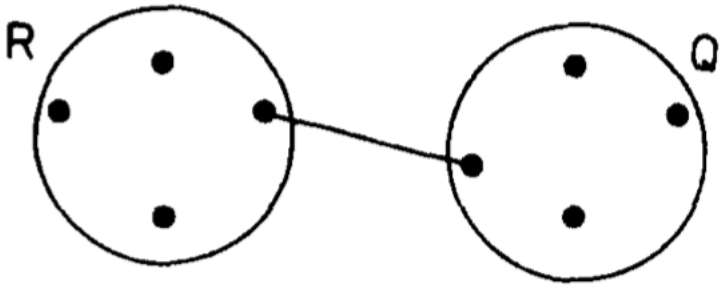
- Algumas características desse método são:
  - Tendência a formar grupos esféricos;
  - O número de grupos é o mesmo durante todo o processo;
  - Inadequado para descobrir grupos com formas não convexas ou de tamanhos muito diferentes;
  - Sensibilidade a ruídos, uma vez que um elemento com um valor extremamente alto pode distorcer a distribuição dos dados;
  - Resultado pode ser diferente para ordenações diferentes dos dados

Partitioning around medoids

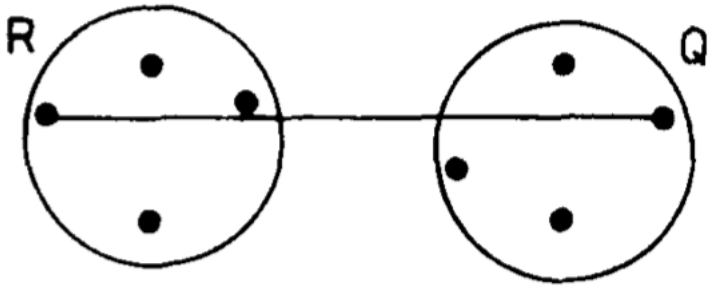
- Algumas características desse método são:
  - Independente da ordem, os resultados serão os mesmos; Tendência a encontrar grupos esféricos;
  - Mais robusto do que o k-means na presença de ruídos porque o medóide é menos influenciado pelos ruídos do que a média;
  - Processamento mais custoso que o k-means;
  - Não aplicável à grandes bases de dados, pois o custo de processamento é alto;
    - Possibilidade do uso do CLARA – Clustering Large Applications

Como avaliar semelhança dos grupos?

- Proximidade entre grupos derivada da matriz de dissimilaridade
  - Nearest neighbour distance

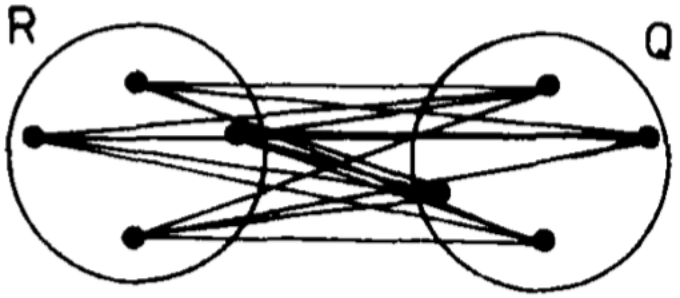


  - Furthest neig



Como avaliar semelhança dos grupos?

- Proximidade entre grupos derivada da matriz de dissimilaridade
  - Group average



- Proximidade entre grupos baseada em medidas-resumo para dados contínuos
  - Média (ou centróide)
  - Mediana

Como avaliar semelhança dos grupos?

- Existem inúmeros métodos. O mais comum é a medida de silhueta.
- Silhueta é uma medida s(i) E [-1,1]
  - Quando s(i) é próximo de 1, a heterogeneidade do cluster do objeto i é menor que sua separação, então o objeto é considerado bem classificado.
  - Média de silhuetas acima de 0,5 seria indicativo de agrupamento bem definido. Abaixo de 0,2 indicaria ausência de estrutura clara de grupos.

Novamente….

- Três questões principais:
  - Como medir semelhança entre objetos?
  - Como agrupar objetos semelhantes?
  - Como avaliar semelhança dos grupos?

- Não existe necessariamente uma resposta correta
  - Depende de tipo de dados e objetivo específico
  - É possível testar diferentes abordagens sem qualquer problema, diferentemente de testes estatísticos tradicionais (pré-teste)

Obrigado.

Pedro Amaral – pedroamaral@cedeplar.ufmg.br