

Análise de Agrupamentos algoritmos e aplicações

José Francisco Moreira Pessanha

Universidade do Estado do Rio de Janeiro (Uerj)

Centro de Pesquisas de Energia Elétrica (Eletrobras Cepel)

francisc@cepel.br

Programa

Conceitos fundamentais da análise de agrupamentos

Algoritmos de análise de agrupamentos:

- Kmeans
- Métodos de encadeamento
- Método de Ward
- Fuzzy c-means
- Mapa auto organizável

Implementação em ambiente R

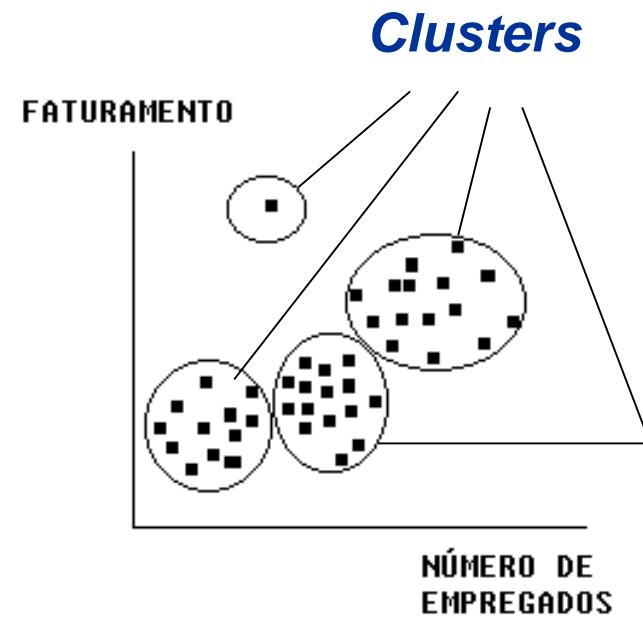
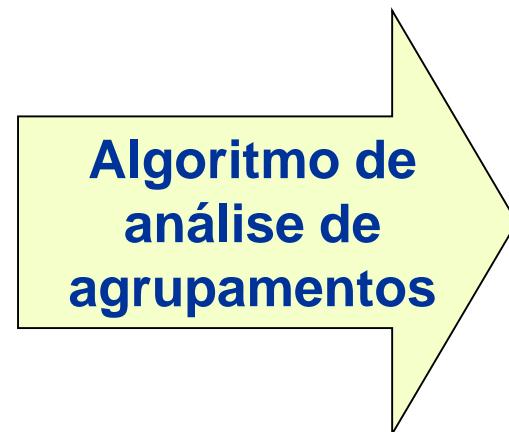
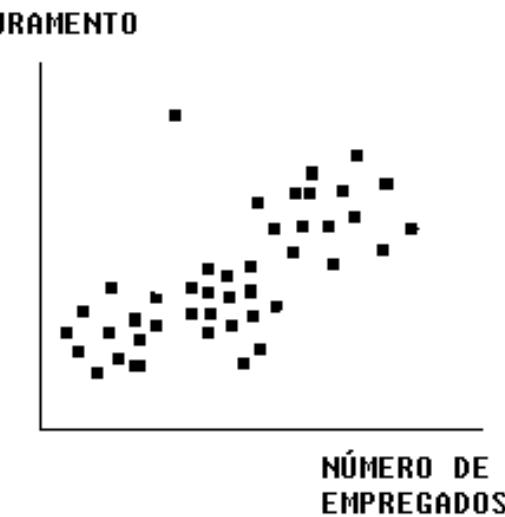


Aplicações em problemas reais do setor elétrico

Análise de agrupamentos (Clusters Analysis)

Técnica estatística multivariada, cujo objetivo consiste em organizar um conjunto de objetos em um determinado nº de subconjuntos mutuamente exclusivos (*clusters*), de tal forma que os objetos em um mesmo *cluster* sejam semelhantes entre si, porém diferentes dos objetos nos outros *clusters*.

Cada ponto representa uma empresa



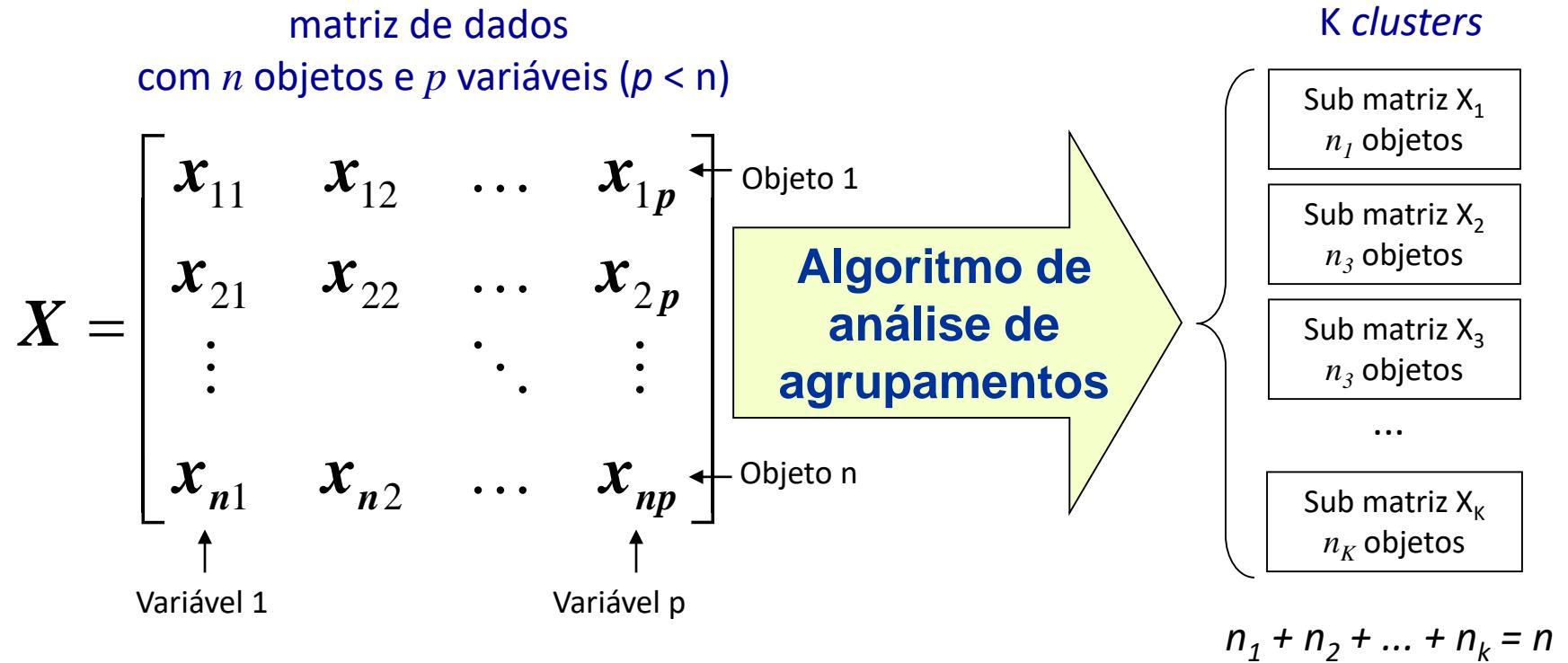
Agrupamentos (Clusters)

Clusters são grupos de objetos semelhantes.

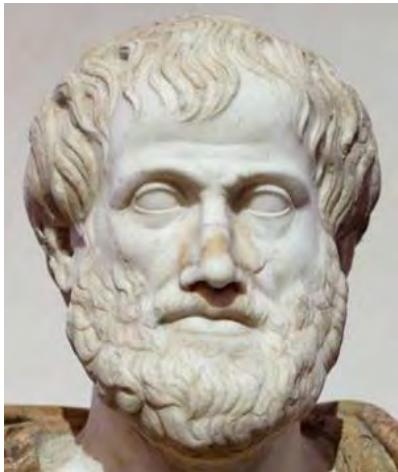
Os *clusters* devem ser bem separados.

Os *clusters* não são conhecidos previamente e são identificados a partir da análise das semelhanças/diferenças entre os objetos de um conjunto de dados.

Análise de agrupamentos

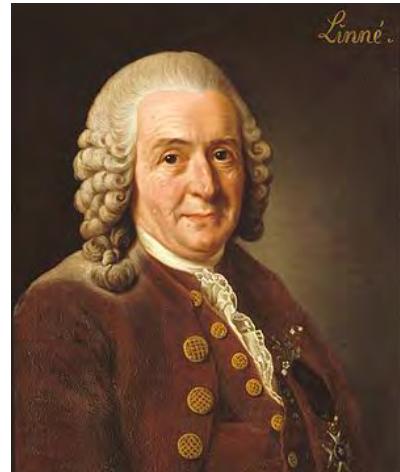


Problema de análise de agrupamentos: Determinar uma partição dos n objetos (linhas) em K clusters, alocando cada objeto (linha) em apenas um cluster, de forma que objetos semelhantes sejam reunidos em um mesmo cluster e objetos diferentes sejam alocados em clusters distintos.



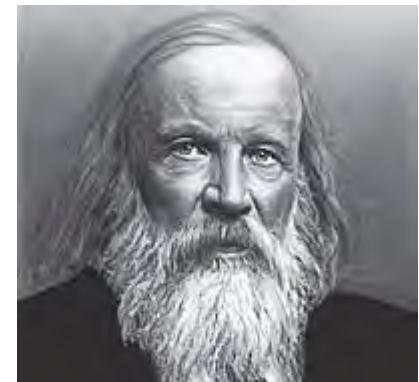
Aristoteles
384 AC – 322 AC

Aristóteles dividiu os seres vivos conhecidos à época em dois Reinos: o dos Animais, móveis, e o das Plantas, imóveis



Lineu
1707 - 1778

Classificação de Lineu
(*Systema Naturae*, 1758)



Mendeleev
1834 - 1907
Tabela Periódica

Periodic Table of the Elements																										
 hydrogen										 poor metals																
 Li	 Be	 Na	 Mg	 K	 Ca	 Sc	 Ti	 V	 Cr	 Mn	 Fe	 Co	 Ni	 Cu	Zn	Ga	Ge	As	Se	Br	Kr					
 Rb	 Sr	 Y	 Zr	 Nb	 Mo	 Tc	 Ru	 Rh	 Pd	 Ag	 Cd	 In	 Sn	 Sb	Te	I	Xe	Rn								
 Cs	 Ba	 La	 Hf	 Ta	 W	 Re	 Os	 Ir	 Pt	 Au	 Hg	 Tl	 Pb	 Bi	Po	At	Rn									
 Fr	 Ra	 Ac	 Un	 Un	 Un	 Un	 Un	 Un	 Un	 Un	 Un	 Un	 Un	 Un	Un	Un	Lu									
 Ce	 Pr	 Nd	 Pm	 Sm	 Eu	 Gd	 Dy	 Ho	 Er	 Tm	 Yb	 Lu	 Th	 Pa	 U	 Np	 Pu	 Am	 Cm	 Bk	 Cf	 Es	 Fm	 Md	 No	 Lr

103 elementos classificados em 8 clusters

Periodic Table of the Elements

The periodic table is color-coded into eight clusters:

- hydrogen** (H, green)
- alkali metals** (Li, Na, K, Rb, Cs, Fr, orange)
- alkali earth metals** (Be, Mg, Ca, Sr, Ba, Ra, light blue)
- transition metals** (Sc, Ti, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, La, Hf, Ta, W, Re, Os, Ir, Pt, Au, Hg, orange)
- poor metals** (B, Al, Ga, In, Ti, Pb, Bi, Po, grey)
- nonmetals** (C, Si, Ge, Sn, Sb, Br, I, grey)
- noble gases** (He, Ne, Ar, Kr, Xe, Rn, orange)
- rare earth metals** (Ce, Pr, Nd, Pm, Sm, Eu, Gd, Tb, Dy, Ho, Er, Tm, Yb, Lu, Th, Pa, U, Np, Pu, Am, Cm, Bk, Cf, Es, Fm, Md, No, Lr, grey)

58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb	71 Lu
90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No	103 Lr

Classificação supervisionada x Classificação não supervisionada

Classificação supervisionada

Modelos para classificação de objetos em classes pré-estabelecidas, por exemplo, clientes adimplentes e inadimplentes, negativo e positivo.

Os modelos são ajustados com base em uma amostra de dados classificados.

Análise de discriminante
Modelos Logit e Probit

Classificação não supervisionada

Algoritmos que particionam um conjunto de objetos em classes (clusters) não conhecidas a priori.

Os algoritmos criam as classes a partir de uma amostra de dados não classificados.

Análise de agrupamentos

Classificação não supervisionada

Os clusters não são conhecidos a priori: pouco ou nada se conhece sobre a estrutura dos grupos, nem o número de grupos é conhecido.

Dados não classificados: a pertinência de um objeto às categorias é desconhecida.

Agrupamento natural dos próprios dados: Busca-se uma estrutura de grupos que se ajuste aos dados.

Característica da análise de agrupamentos

- É descritiva, os métodos são exploratórios e a idéia é gerar hipóteses.
- Cria grupos indiferentemente a existência de alguma estrutura nos dados.
- Variedade de vias e critérios para definição dos grupos o que implica na possibilidade de obter soluções diferentes variando-se um ou mais critérios.
- A solução da análise de agrupamentos não é generalizável por que ela é totalmente dependente tanto das variáveis usadas como dos critérios em que ela se baseia.

O que podemos fazer com a análise de agrupamentos?

**Formar uma taxonomia das observações
Classificação empírica dos objetos.**

**Simplificar os dados (compressão de dados)
Descrever de forma compacta um razoável volume de dados
por meio de elementos típicos.**

Identificar relações entre objetos

- **Revelar similaridades e diferenças entre objetos.**
- **Identificar observações aberrantes.**

Conceitos fundamentais da análise de agrupamentos



François Morellet

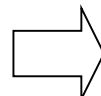
6 répartitions aléatoires de 4 carrés noirs et blancs d'après les chiffres pairs et impairs du nombre Pi, 1958.



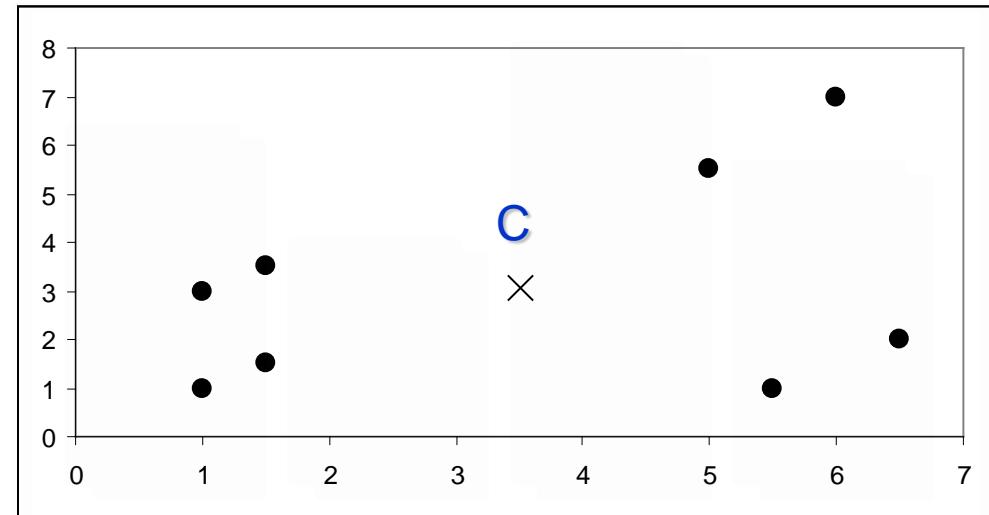
Centroide: Centro de gravidade

Matriz de dados

objetos	atributo X	atributo Y
#1	1	1
#2	1,5	1,5
#3	5	5,5
#4	6	7
#5	1	3
#6	1,5	3,5
#7	5,5	1
#8	6,5	2



Nuvem de dados com centro de gravidade C



As coordenadas do centro de gravidade C são as médias em cada atributo.

O centro de gravidade é o vetor de médias:

$$\text{Abscissa de } C = (1 + 1,5 + 5 + 6 + 1 + 1,5 + 5,5 + 6,5) / 8 = 3,5$$

$$\text{Ordenada de } C = (1 + 1,5 + 5,5 + 7 + 3 + 3,5 + 1 + 2) / 8 = 3,06$$

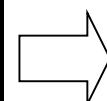
← C = (3,5 ; 3,06)



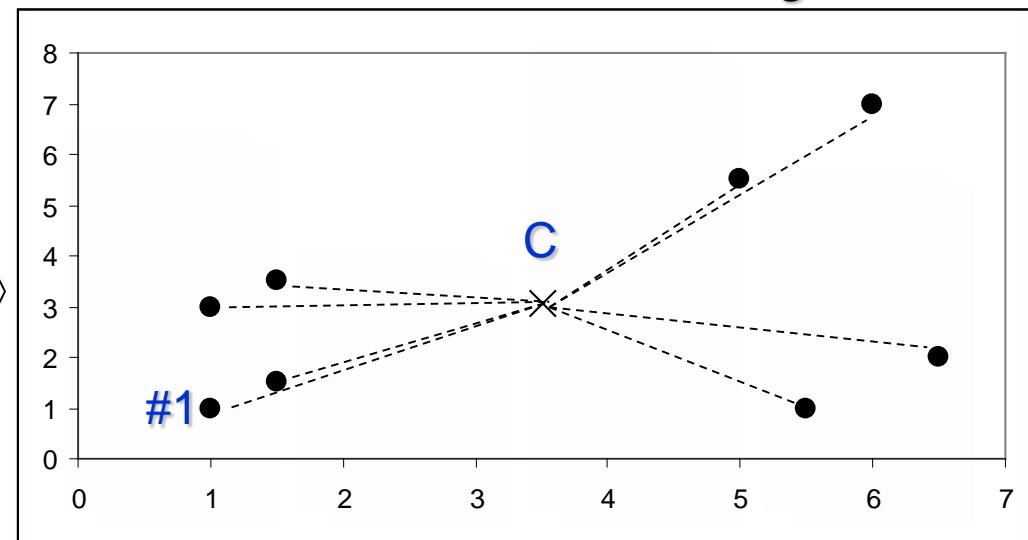
Inércia total: Medida de dispersão da nuvem de dados

Matriz de dados

objetos	atributo X	atributo Y
#1	1	1
#2	1,5	1,5
#3	5	5,5
#4	6	7
#5	1	3
#6	1,5	3,5
#7	5,5	1
#8	6,5	2
Centro de gravidade	3,50	3,06



Nuvem de dados com centro de gravidade C



Quadrado da distância entre o objeto #1 e o centro de gravidade C:

$$\| \#1 - C \|^2 = (1 - 3,5)^2 + (1 - 3,06)^2 = 10,1$$

A inércia total é a soma dos quadrados das distâncias entre cada objeto e o centro de gravidade da nuvem de pontos (ponto C):

$$\text{Inércia total} = \| \#1 - C \|^2 + \| \#2 - C \|^2 + \dots + \| \#8 - C \|^2 = \sum_{i=1}^8 \| \#i - C \|^2 = 75,72$$

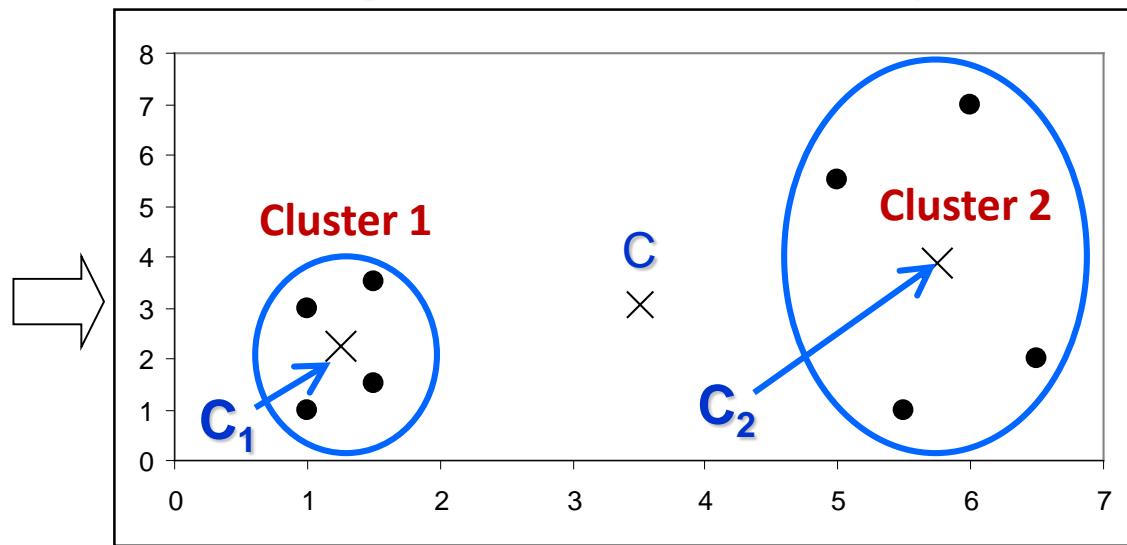
Inércia total = Total Squares Sum (TSS)

Centroides dos *clusters*

Matriz de dados

objetos	atributo X	atributo Y
#1	1	1
#2	1,5	1,5
#3	5	5,5
#4	6	7
#5	1	3
#6	1,5	3,5
#7	5,5	1
#8	6,5	2

Centros de gravidade dos *clusters* (centróides)



O centro de gravidade de um *cluster* é o vetor de médias dos seus objetos:

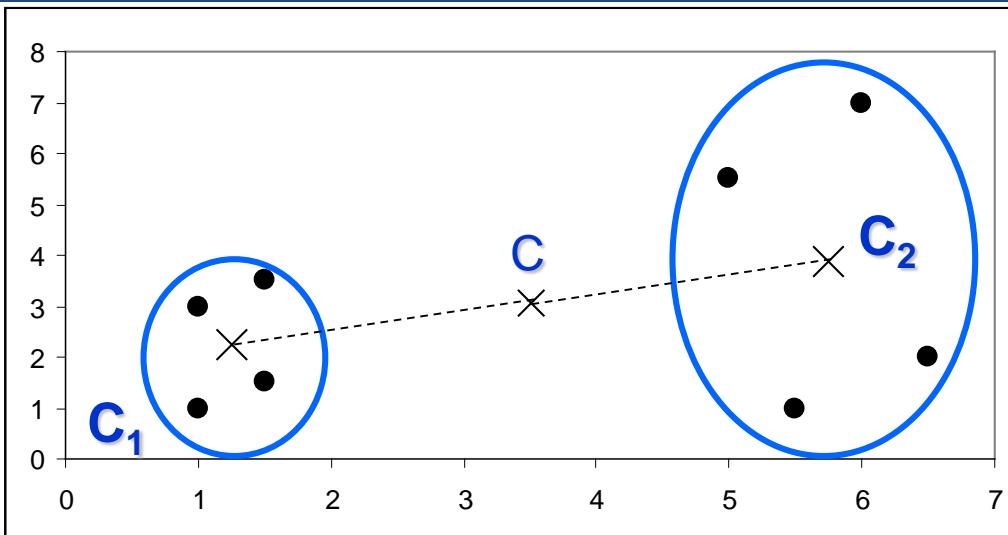
$$\text{Abscissa do centróide } C1 = (1 + 1,5 + 1 + 1,5) / 4 = 1,25$$

$$\text{Ordenada do centróide } C1 = (1 + 1,5 + 3 + 3,5) / 4 = 2,25$$

$$\text{Abscissa do centróide } C2 = (5 + 6 + 5,5 + 6,5) / 4 = 5,75$$

$$\text{Ordenada do centróide } C2 = (5,5 + 7 + 1 + 2) / 4 = 3,88$$

Inércia entre *clusters*: medida de separação entre os *clusters*



É a soma dos quadrados das distâncias dos centroides dos *clusters* ao centro de gravidade da nuvem de dados, ponderadas pelos respectivos tamanhos dos *clusters*.

$$4x||C_1 - C||^2 + 3x||C_2 - C||^2$$

Centro de gravidade	abscissa (X)	ordenada (Y)
cluster 1 (C_1)	1,25	2,25
cluster 2 (C_2)	5,75	3,88
nuvem de pontos (C)	3,50	3,06

Contribuição do *cluster 1* para a inércia entre *cluster* = $4x||C_1 - C||^2$

$$4x[(1,25 - 3,50)^2 + (2,25 - 3,06)^2] = 22,89$$

Contribuição do *cluster 2* para a inércia entre *cluster* = $4x||C_2 - C||^2$

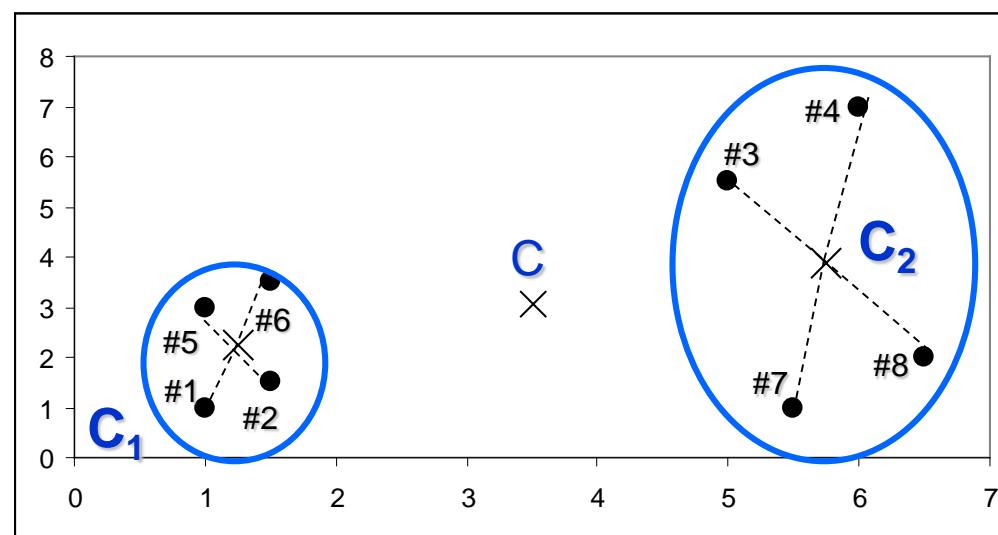
$$4x[(5,75 - 3,50)^2 + (3,88 - 3,06)^2] = 22,89$$

Inércia entre *clusters* = $22,89 + 22,89 = 45,78$

Inércia entre *clusters* = Between Squares Sum (BSS)



Inércia dentro dos *clusters*: medida de homogeneidade dos *clusters*



É a soma dos quadrados das distâncias dos objetos aos centroides dos *clusters* onde estão alocados.

Contribuição do *cluster* 1 para a inércia dentro dos *clusters*

$$\| \#1 - C_1 \|^2 + \| \#2 - C_1 \|^2 + \| \#5 - C_1 \|^2 + \| \#6 - C_1 \|^2 = 4,5$$

cluster 1	atributo X	atributo Y
#1		1
#2	1,5	1,5
#5		3
#6	1,5	3,5
Centro de gravidade C1	1,25	2,25

Contribuição do *cluster* 2 para a inércia dentro dos *clusters*

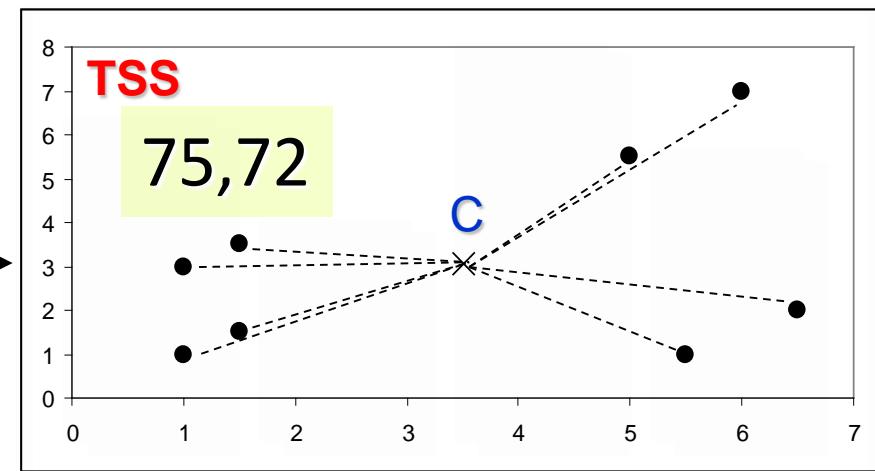
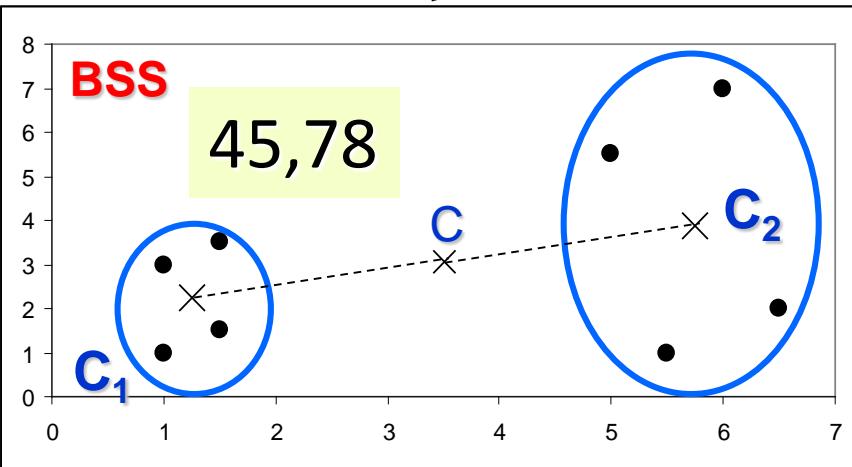
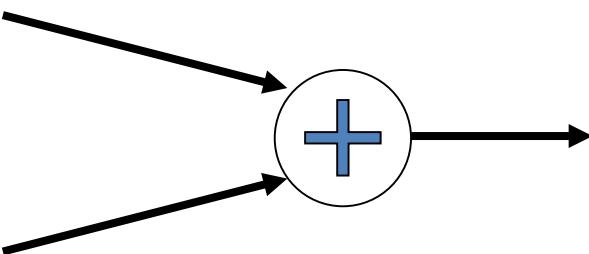
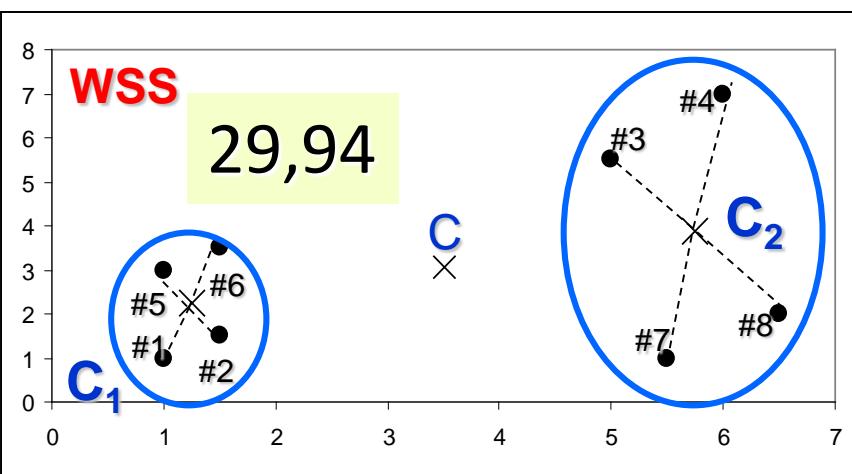
$$\| \#3 - C_2 \|^2 + \| \#4 - C_2 \|^2 + \| \#7 - C_2 \|^2 + \| \#8 - C_2 \|^2 = 25,44$$

cluster 2	atributo X	atributo Y
#3	5	5,5
#4	6	7
#7	5,5	1
#8	6,5	2
Cntrro de gravidade C2	5,75	3,88

$$\text{Inércia dentro dos } clusters = 4,5 + 25,44 = 29,94$$

Inércia dentro dos *clusters* = Within Squares Sum (WSS)

Decomposição da Inércia Total



$$\text{TSS} = \text{BSS} + \text{WSS}$$

Decomposição da Inércia Total: Teorema de Huygens

Quadrado da distância entre o objeto x_i e o centro de gravidade da nuvem de dados C

$$\sum_{i=1}^n \|x_i - C\|^2$$

Inércia total TSS

Quadrado da distância entre o centroide do cluster h (C_h) e o centro de gravidade da nuvem de dados C

$$\sum_{h=1}^K n_h \|C_h - C\|^2$$

Inércia entre os clusters BSS

Quadrado da distância entre o objeto x_j no cluster h e o centróide do cluster h

$$\sum_{h=1}^K \sum_{j=1}^{n_h} \|x_{j,h} - C_h\|^2$$

Inércia dentro dos clusters WSS

$K = \text{nº de clusters}$

$n = \text{nº de objetos}$

$n_h = \text{nº de objetos no cluster } h$



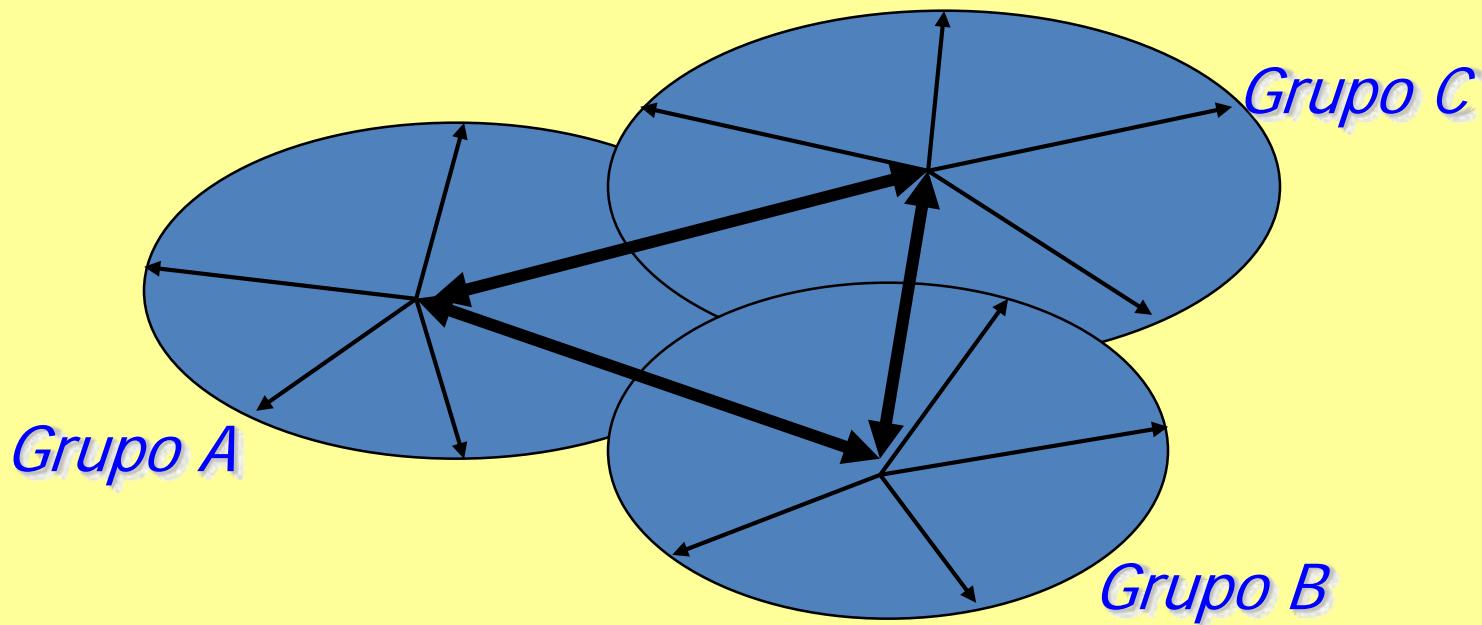
Huygens
1629 - 1695

BSS e WSS



Variação entre clusters BSS = Maximizar

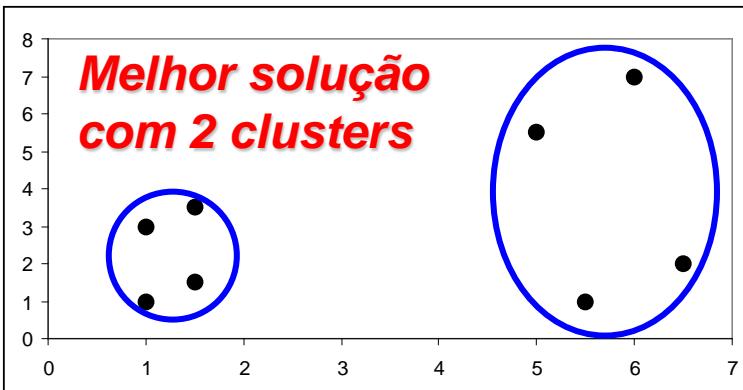
Variação dentro dos clusters WSS = Minimizar





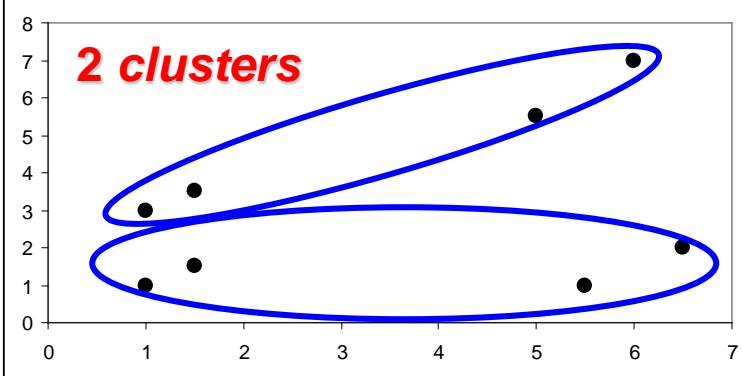
Qual a melhor forma de organizar n objetos em K clusters ?

A melhor alternativa forma *clusters* homogêneos e bem separados



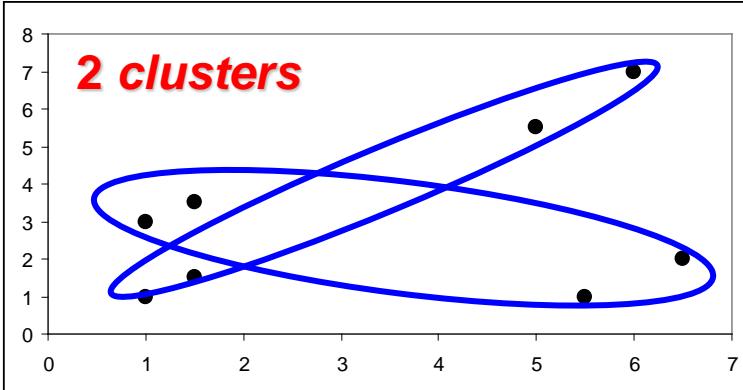
$$\begin{aligned}WSS &= 29,94 \\BSS &= 45,78 \\TSS &= 75,72\end{aligned}$$

A inércia total *TSS* permanece constante e independe da forma como os objetos são agrupados



$$\begin{aligned}WSS &= 52,81 \\BSS &= 22,91 \\TSS &= 75,72\end{aligned}$$

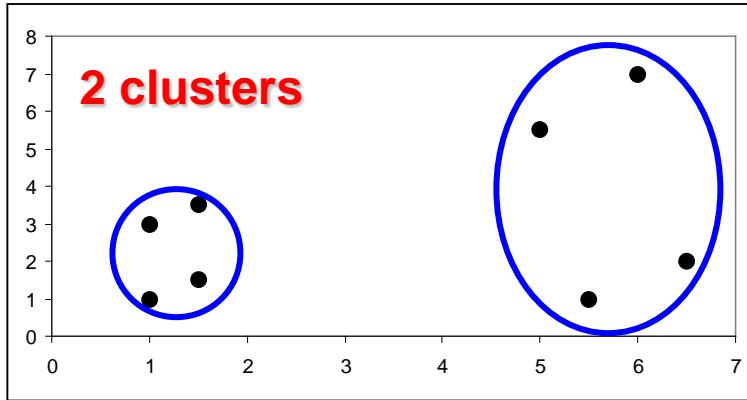
Quanto menor a inércia dentro dos *clusters* *WSS*, maior a homogeneidade interna dos *clusters*



$$\begin{aligned}WSS &= 71,81 \\BSS &= 3,91 \\TSS &= 75,72\end{aligned}$$

Quanto maior a inércia entre clusters *BSS*, maior é a separação dos clusters

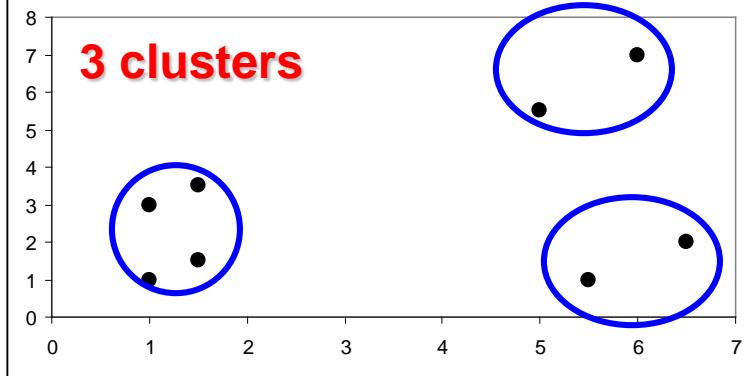
Quantos *clusters* ?



2 clusters

$$\begin{aligned} \text{WSS} &= 29,94 \\ \text{BSS} &= 45,78 \\ \text{TSS} &= 75,72 \end{aligned}$$

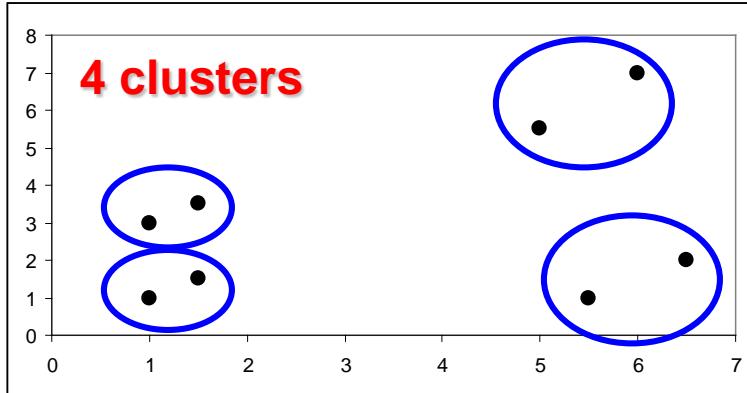
A inércia total *TSS* permanece constante e independe do nº de *clusters*



3 clusters

$$\begin{aligned} \text{WSS} &= 7,13 \\ \text{BSS} &= 68,59 \\ \text{TSS} &= 75,72 \end{aligned}$$

A inércia dentro dos *clusters* *WSS* diminui com a inclusão de mais *clusters*



4 clusters

$$\begin{aligned} \text{WSS} &= 3,13 \\ \text{BSS} &= 72,59 \\ \text{TSS} &= 75,72 \end{aligned}$$

A inércia entre *clusters* *BSS* aumenta com a inclusão de mais *clusters*

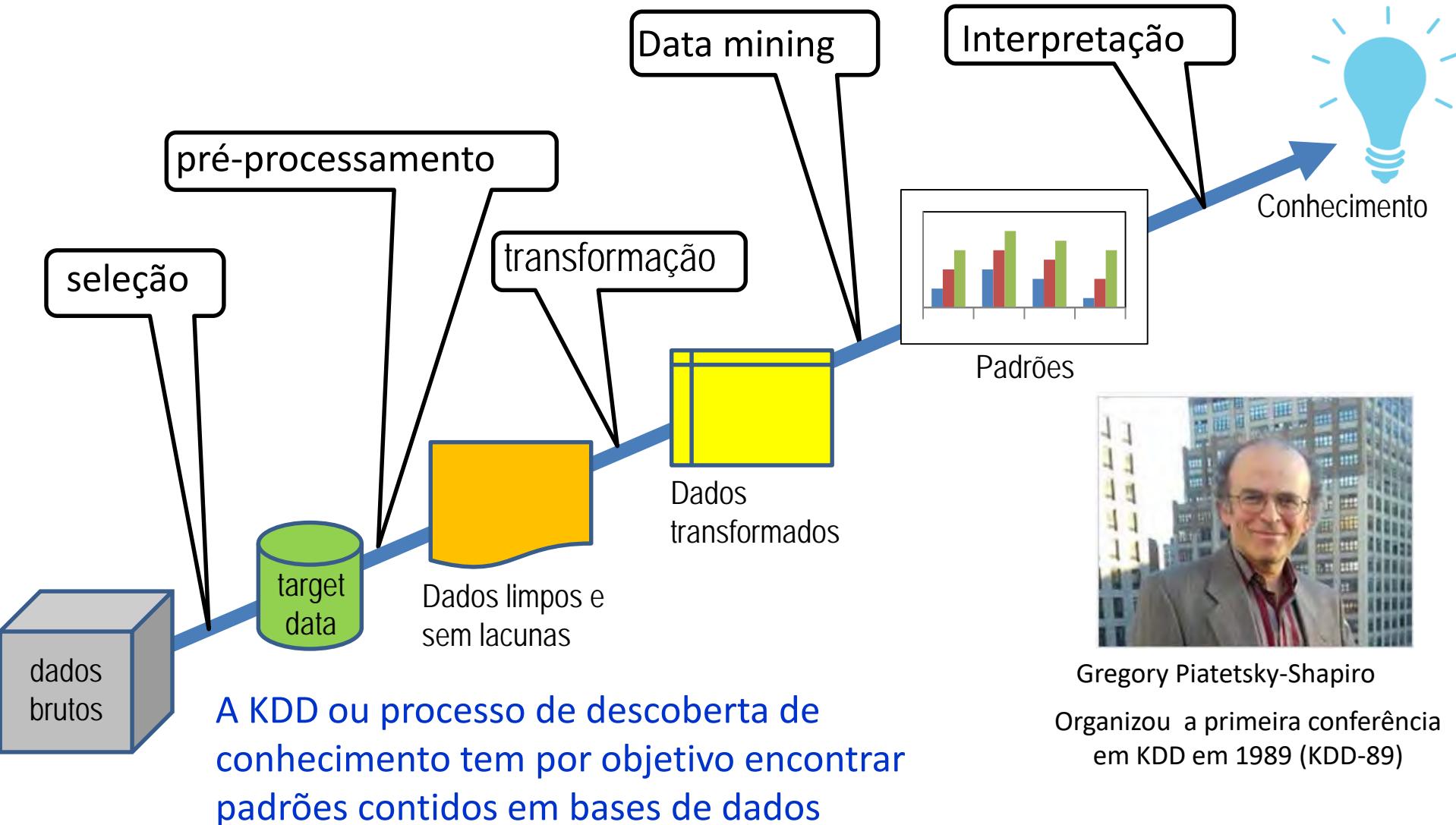


Etapas da análise de agrupamentos

- 1) Seleção dos objetos a serem agrupados.**
- 2) Definição de um conjunto de atributos que caracterizam os objetos.**
- 3) Seleção de uma medida de dissimilaridade entre os objetos.**
- 4) Seleção de um algoritmo de agregação dos objetos.**
- 5) Definição do nº de *clusters*.**
- 6) Interpretação e validação dos *clusters* obtidos.**

Extração do conhecimento – KDD Knowledge Discovery in Databases

O objetivo do BI é permitir uma fácil interpretação do grande volume de dados. Identificando novas oportunidades e implementando uma estratégia efetiva baseada nos dados, também pode promover negócios com vantagem competitiva no mercado e estabilidade a longo prazo.





Passo 2 - Critérios para seleção das variáveis

- Use variáveis que caracterizem os objetos e que estejam relacionadas com os objetivos da pesquisa.
- Procure diminuir o número de variáveis (Princípio da parcimônia), de forma que sua seleção conte cole tanto a relevância como o poder de discriminação face o problema em estudo.
- Variáveis que assumem praticamente o mesmo valor para todos os objetos são pouco discriminatórias e podem ser descartadas.
- Variáveis com grande poder de discriminação, porém irrelevantes frente ao problema, podem mascarar os grupos e levar a resultados equivocados. Estas também devem ser descartadas.
- Use técnicas estatísticas para a redução da dimensionalidade da matriz de dados (análise de componentes principais e a análise fatorial).



Passo 2 - Transformação das variáveis

- Quando as variáveis se apresentam definidas em diferentes escalas de medida a utilização dos valores originais pode implicar no estabelecimento de uma ponderação implícita nas variáveis.
- Sem uma padronização prévia, a classificação dos objetos vai refletir sobretudo o peso das variáveis com os maiores valores e maiores variâncias.
- As variáveis devem ser padronizadas de forma a evitar problemas nos resultados causados pelo uso de escalas diferentes.
- A padronização mais comum é o Z escore, porém não pode ser tomada como solução ideal para todos os casos.
- A padronização pode reduzir as diferenças entre os indivíduos anulando os agrupamentos naturais que possam existir nos dados.
- Só a experiência e o conhecimento no assunto em estudo poderão ajudar a encontrar a solução mais correta para cada caso.



Passo 3 - Medidas de dissimilaridade

Clusters são grupos de objetos semelhantes.

Para formar *clusters* é necessário ter uma medida do grau de similaridade entre os objetos ou de diferença (dissimilaridade) entre eles.

Há várias medidas possíveis e a mais adequada para um caso depende dos tipos de atributos que caracterizam os objetos.

Quatro tipos de medidas (Aldenderfer & Blashfield, 1985)

- Medidas de associação
- Medidas de correlação
- Medidas de distância
- Medidas probabilísticas



Passo 3 - Medidas de distância

Medida de dissimilaridade entre dois objetos.

Distâncias pequenas indicam objetos próximos, ou seja, semelhantes em um conjunto de atributos.

Distâncias grandes indicam objetos distantes, logo diferentes.

Dados três objetos x , y e z a distância entre eles deve verificar as seguintes propriedades:

Simetria: $d(x,y) = d(y,x) \geq 0$

Diferenciabilidade de não idênticos: $d(x,y) \neq 0 \Rightarrow x \neq y$

Indiferenciabilidade de idênticos: $d(x,y) = 0 \Rightarrow x = y$

Desigualdade triangular: $d(x,y) \leq d(x,z) + d(z,y)$

Passo 3 - Medidas de distância

Dados dois objetos i e j caracterizados por p variáveis

$$X_i^T = \begin{pmatrix} x_{i1} & \dots & x_{ip} \end{pmatrix}$$

$$X_j^T = \begin{pmatrix} x_{j1} & \dots & x_{jp} \end{pmatrix}$$

Tipos de medidas de distância

Distância Euclidiana

$$d_{ij} = \sqrt{\sum_{v=1}^p (x_{iv} - x_{jv})^2} = \sqrt{(X_i - X_j)^T (X_i - X_j)}$$

Quadrado da distância Euclidiana

$$d_{ij}^2 = \sum_{v=1}^p (x_{iv} - x_{jv})^2 = (X_i - X_j)^T (X_i - X_j)$$

Distância absoluta ou
City-Block

$$d_{ij} = \sum_{v=1}^p |x_{iv} - x_{jv}|$$

Distância de Minkowski

$$d_{ij} = \left(\sum_{v=1}^p |x_{iv} - x_{jv}|^r \right)^{\frac{1}{r}}$$

Σ = matriz de covariância

Distância de Mahalanobis ou
generalizada

$$d_{ij} = (X_i - X_j)^T \Sigma^{-1} (X_i - X_j)$$

Distância de Chebishev

$$d_{ij} = \max_v |x_{iv} - x_{jv}|$$

Passo 3 - Medidas de distância

Matriz de dados

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \ddots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

← Objeto 1

← Objeto n

↑ Variável 1 ↑ Variável p

Métrica de dissimilaridade



Matriz de distâncias

$$D = \begin{bmatrix} 0 & d_{12} & \dots & d_{1n} \\ d_{21} & 0 & \dots & d_{2n} \\ \vdots & \ddots & \ddots & \vdots \\ d_{n1} & d_{n2} & \dots & 0 \end{bmatrix}$$

Variedade de métodos de análise de agrupamentos

- Métodos estatísticos
 - Métodos não hierárquicos
(k-Means , Nuvens dinâmicas)
 - Métodos hierárquicos
(Métodos de encadeamento, Método de Ward)
- Métodos baseados em lógica fuzzy: *Fuzzy Clustering Method - FCM*
- Métodos baseados em redes neurais: *Self-Organizing Map - SOM*
- Métodos baseados em densidade: DBSCAN

Passo 4 – Seleção do algoritmo de análise de agrupamentos

Lista de pacotes e funções do R para análise de agrupamentos

<http://cran.r-project.org/web/views/Cluster.html>

Hierarchical Clustering:

- Functions `hclust()` from package `stats` and `agnes()` from `cluster` are the primary functions for agglomerative hierarchical clustering, function `diana()` can be used for divisive hierarchical clustering. Faster alternatives to `hclust()` are provided by the packages `fastcluster` and `flashClust`.
- Function `dendrogram()` from `stats` and associated methods can be used for improved visualization for cluster dendograms.
- The `dendextend` package provides functions for easy visualization (coloring labels and branches, etc.), manipulation (rotating, pruning, etc.) and comparison of dendograms (tanglegrams with heuristics for optimal branch rotations, and tree correlation measures with bootstrap and permutation tests for significance).
- Package `dynamicTreeCut` contains methods for detection of clusters in hierarchical clustering dendograms.
- `hybridHclust` implements hybrid hierarchical clustering via mutual clusters.
- Package `isopam` uses an algorithm which is based on the classification of ordination scores from isometric feature mapping. The classification is performed either as a hierarchical, divisive method or as non-hierarchical partitioning.
- `pvclust` is a package for assessing the uncertainty in hierarchical cluster analysis. It provides approximately unbiased p-values as well as bootstrap p-values.
- Package `sparcl` provides clustering for a set of n observations when $p \gg n$. It adaptively chooses a set of variables to use in clustering the observations. Sparse K-means clustering and sparse hierarchical clustering are implemented.

Partitioning Clustering:

- Function `kmeans()` from package `stats` provides several algorithms for computing partitions with respect to Euclidean distance.
- Function `pam()` from package `cluster` implements partitioning around medoids and can work with arbitrary distances. Function `clara()` is a wrapper to `pam()` for larger data sets. Silhouette plots and spanning ellipses can be used for visualization.
- Package `apcluster` implements Frey's and Dueck's Affinity Propagation clustering. The algorithms in the package are analogous to the Matlab code published by Frey and Dueck.
- Package `bayesclust` allows to test and search for clusters in a hierarchical Bayes model.
- Package `clusterSim` allows to search for the optimal clustering procedure for a given dataset.
- Package `flexclust` provides k-centroid cluster algorithms for arbitrary distance measures, hard competitive learning, neural gas and QT clustering. Neighborhood graphs and image plots of partitions are available for visualization. Some of this functionality is also provided by package `clust`.
- Package `kernlab` provides a weighted kernel version of the k-means algorithm by `kkmeans` and spectral clustering by `specc`.
- Packages `kml` and `kml3d` provide k-means clustering specifically for longitudinal (joint) data.
- Package `skmeans` allows spherical k-Means Clustering, i.e. k-means clustering with cosine similarity. It features several methods, including a genetic and a simple fixed-point algorithm and an interface to the CLUTO vcluster program for clustering high-dimensional datasets.
- Package `trimcluster` provides trimmed k-means clustering.

Passo 5 – Definição do número de agrupamentos

Critérios:

- 3 grupos (pequeno, médio, grande)
- Dendrograma gerada por métodos hierárquicos
- Proporção da BSS na TSS
- Pseudo-F
- Medidas de validação e estabilidade

Análise de agrupamentos

Métodos não hierárquicos

Procuram diretamente uma partição de N objetos em um número pré-definido de k clusters que satisfaçam duas premissas básicas: coesão interna e isolamento dos clusters.

O número de agrupamentos deve ser especificado a priori

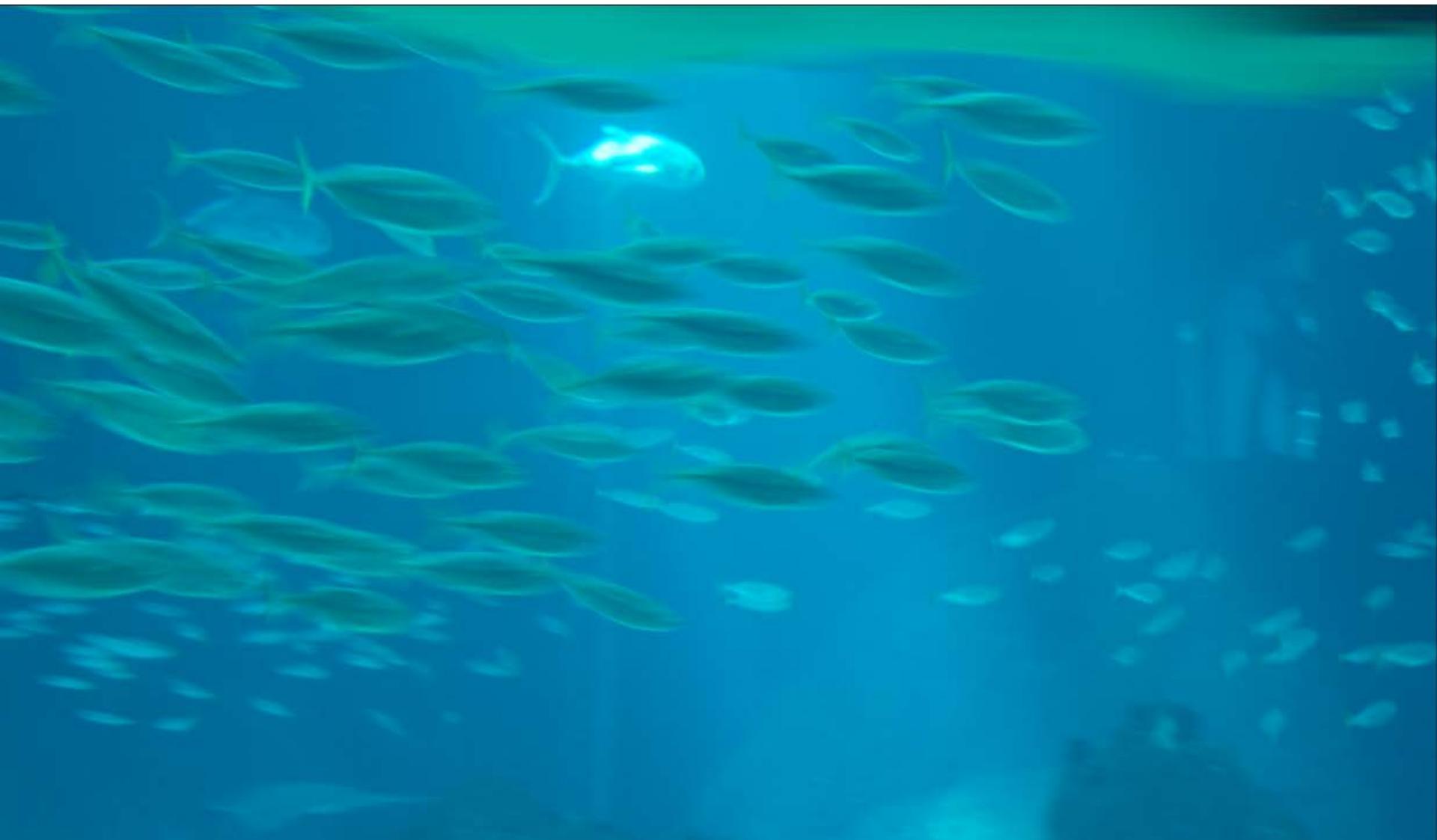
Métodos hierárquicos

Particionam um conjunto com N objetos seqüencialmente em 1,2,3,4 até N clusters, obtendo no final uma estrutura em árvore, semelhante as classificações zoológicas (espécies, gêneros, famílias, ordem, etc.).

Métodos não hierárquicos: k-Means

Métodos hierárquicos: métodos de encadeamento

Métodos não hierárquicos: K Means



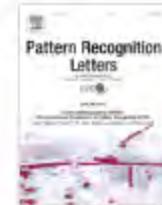
K Means



Pattern Recognition Letters

Volume 31, Issue 8, 1 June 2010, Pages 651–666

Award winning papers from the 19th International Conference on Pattern
Recognition (ICPR)

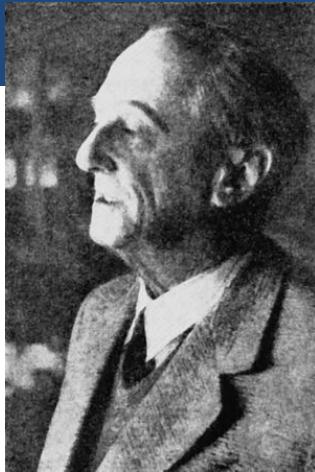


19th International Conference in Pattern Recognition (ICPR)

Data clustering: 50 years beyond K-means

Anil K. Jain

One of the most popular and simple clustering algorithms, K-means, was first published in 1955. In spite of the fact that K-means was proposed over 50 years ago and thousands of clustering algorithms have been published since then, K-means is still widely used.



K Means

BULLETIN DE L'ACADEMIE
POLONAISE DES SCIENCES
CL. III — VOL. IV, No. 12, 1956

MATHÉMATIQUE

Hugo Steinhaus
1887 - 1972

Sur la division des corps matériels en parties¹
par

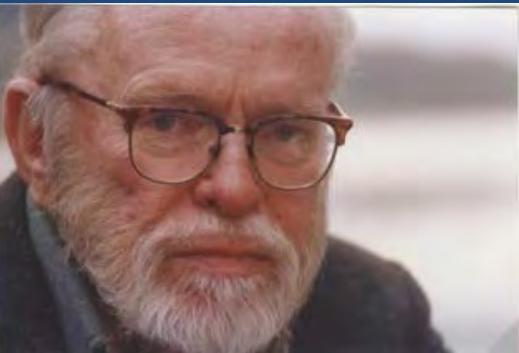
H. STEINHAUS
Présenté le 19 Octobre 1956

Le PROBLÈME de cette Note est la division d'un corps en n parties K_i ($i = 1, 2, \dots, n$) et le choix de n points A_i de manière à rendre aussi petite que possible la somme

$$(1) \quad S(K, A) = \sum_{i=1}^n I(K_i, A_i) \quad (K \equiv \{K_i\}, \quad A \equiv \{A_i\}),$$

où $I(Q, P)$ désigne, en général, le *moment d'inertie* d'un corps quelconque Q par rapport à un point quelconque P .

Steinhaus, H. (1956). «Sur la division des corps matériels en parties». *Bull. Acad. Polon. Sci.* 4 (12): 801–804.



K Means

Usou o termo K Means pela primeira vez

James B. MacQueen
1929-2014

MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press.

SOME METHODS FOR CLASSIFICATION AND ANALYSIS OF MULTIVARIATE OBSERVATIONS

J. MACQUEEN
UNIVERSITY OF CALIFORNIA, LOS ANGELES

1. Introduction

The main purpose of this paper is to describe a process for partitioning an N -dimensional population into k sets on the basis of a sample. The process, which is called 'k-means,' appears to give partitions which are reasonably efficient in the sense of within-class variance.

K Means

Divide um conjunto de n objetos x_i , ($i=1,n$) em k clusters, de tal forma que a heterogeneidade interna (WSS) dos clusters seja minimizada.

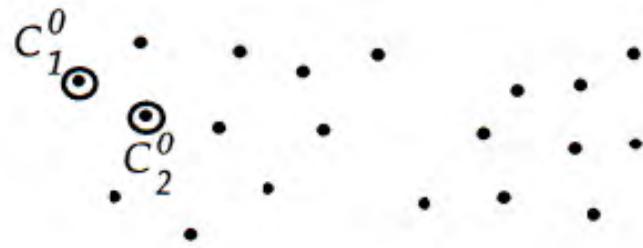
O número de clusters k é um dado de entrada.

Dado um conjunto de k centróides, cada objeto x_i , $i=1,n$, é alocado ao cluster com o centróide c_j mais próximo.

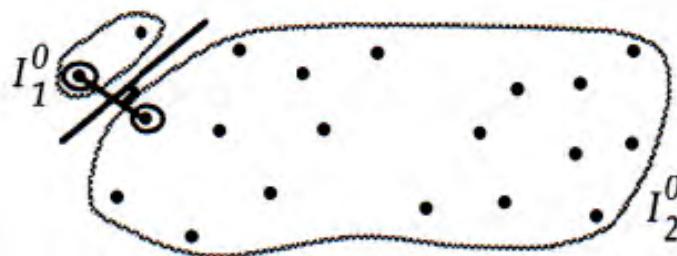
Problema: Identifique os k centróides c_j , $j=1,k$ e as alocações dos objetos aos clusters que minimizam a inércia dentro dos clusters (WSS).

$$WSS = \sum_{j=1}^k WSS(j) = \sum_{j=1}^k \left(\sum_{x_i \in \text{cluster } j} \|x_i - c_j\|^2 \right)$$

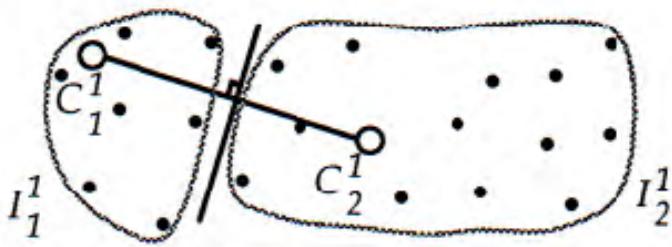
Algoritmo K Means



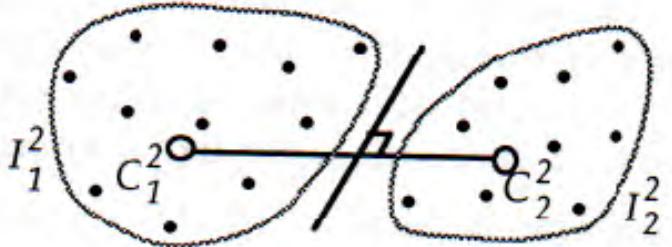
Tirage au hasard
des centres
 C_1^0 et C_2^0



Constitution des classes
 I_1^0 et I_2^0



Nouveaux centres
 C_1^1 et C_2^1
et nouvelles classes
 I_1^1 et I_2^1



Nouveaux centres
 C_1^2 et C_2^2
et nouvelles classes
 I_1^2 et I_2^2

Algoritmo K Means

Passo 1 Especifique K centroides iniciais c_j , $j=1, K$ (escolhidos ou selecionados aleatoriamente a partir da amostra).

Passo 2 Aloque cada objeto ao centroide mais próximo.

Passo 3 Atualize os centroides, i.e., recalcule o vetor de médias em cada cluster

Passo 4 Retorne ao passo 2 enquanto o critério de parada não for alcançado. Caso contrário, pare a execução do algoritmo.

A partição em K clusters corresponde à ultima alocação realizada.

Critérios de parada:

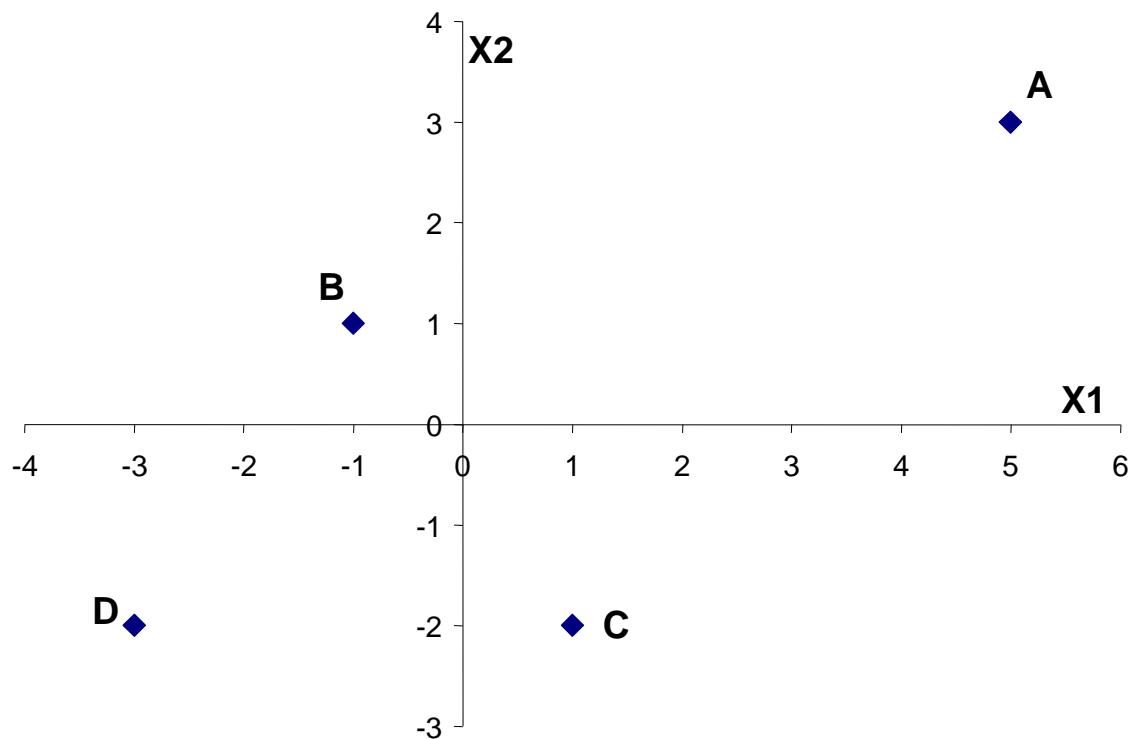
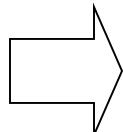
Número máximo de iterações

Estabilidade dos centroides

Exemplo 1 - K Means

Dado o conjunto de 4 objetos ($n=4$), use o algoritmo k-Means para identificar 2 clusters ($k=2$)

Objetos	Coordenadas	
	X1	X2
A	5	3
B	-1	1
C	1	-2
D	-3	-2





Exemplo 1 - K Means

Centroides iniciais (seleção aleatória)

Centróide	Coordenadas dos centróides dos clusters	
	X1	X2
C1	2	2
C2	-1	-2

Lista de objetos

Objetos	Coordenadas	
	X1	X2
A	5	3
B	-1	1
C	1	-2
D	-3	-2

Alocação dos objetos aos clusters

$$\|A - C1\|^2 = (5-2)^2 + (3-2)^2 = 10$$

$$\|A - C2\|^2 = (5+1)^2 + (3+2)^2 = 61$$

A Cluster com centroide C1

$$\|B - C1\|^2 = (-1-2)^2 + (1-2)^2 = 10$$

$$\|B - C2\|^2 = (-1+1)^2 + (1+2)^2 = 9$$

B Cluster com centroide C2

$$\|C - C1\|^2 = (1-2)^2 + (-2-2)^2 = 5$$

$$\|C - C2\|^2 = (1+1)^2 + (-2+2)^2 = 4$$

C Cluster com centroide C2

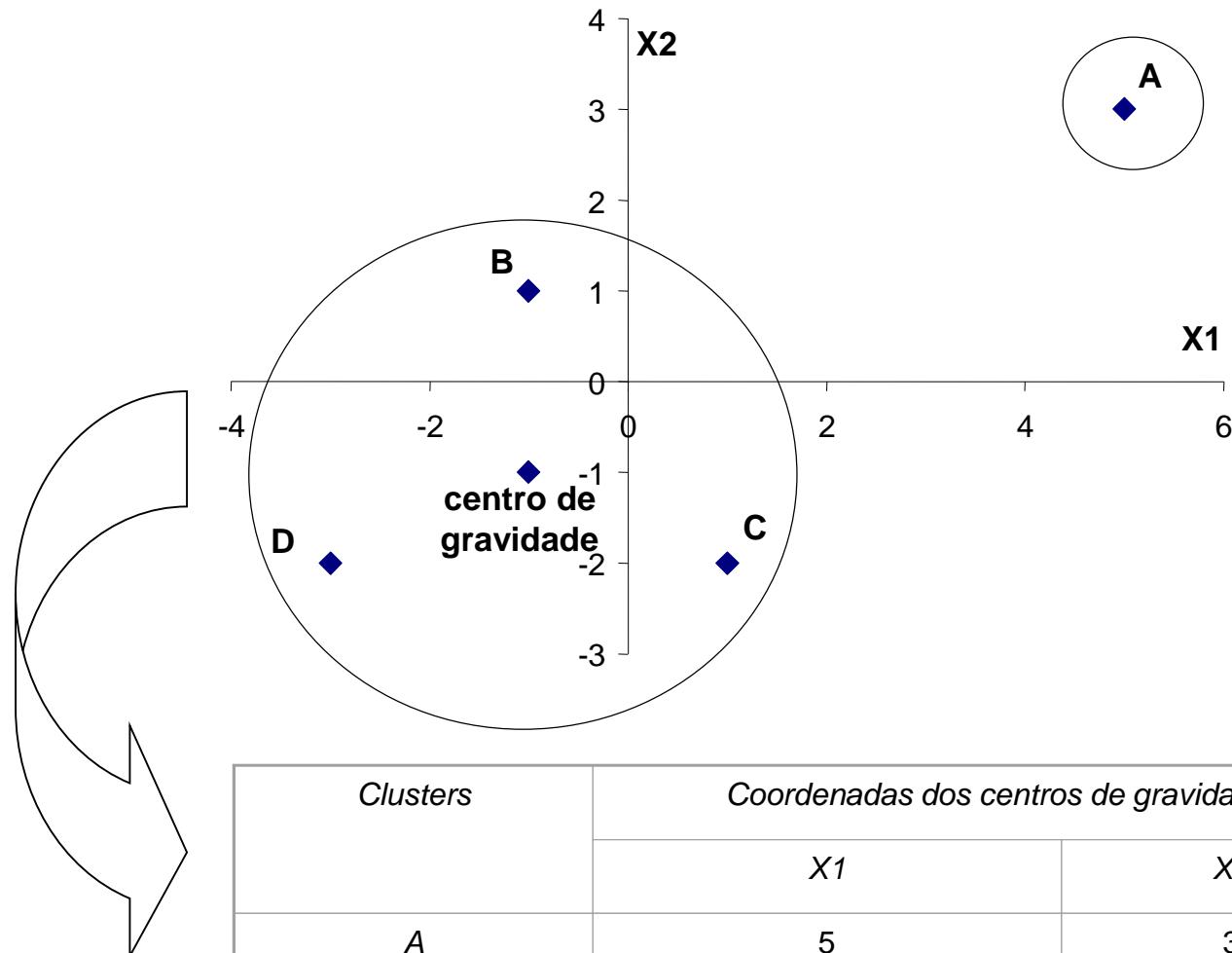
$$\|D - C1\|^2 = (-3-2)^2 + (-2-2)^2 = 39$$

$$\|D - C2\|^2 = (-3+1)^2 + (-2+2)^2 = 4$$

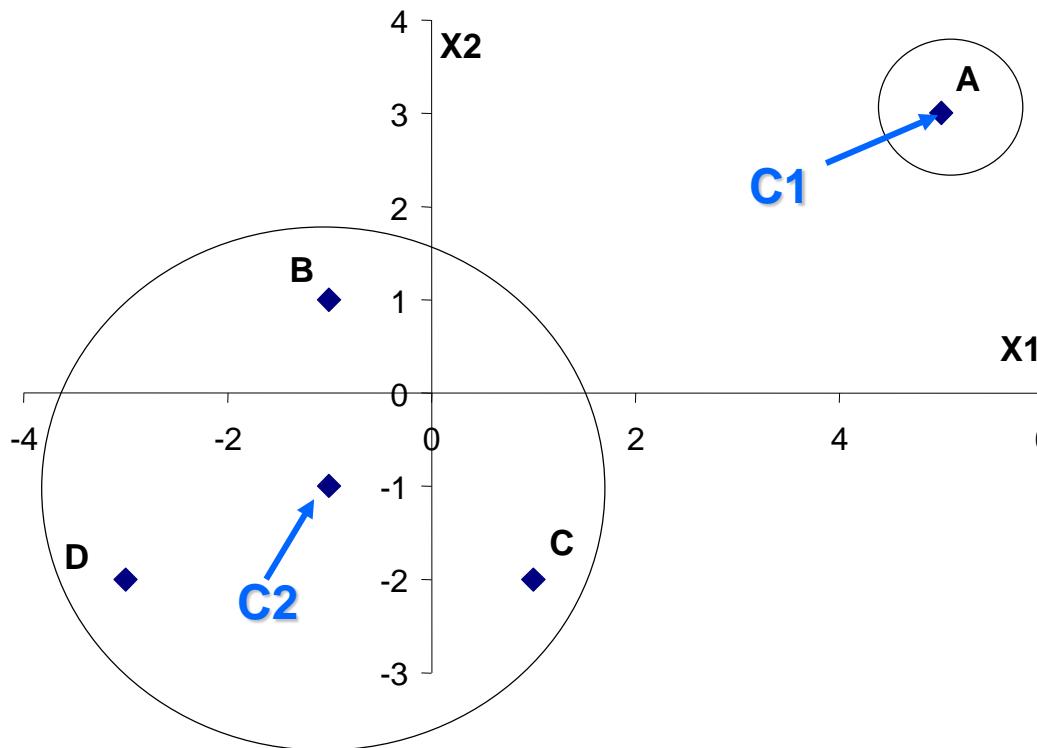
D Cluster com centroide C2

Exemplo 1 - K Means

Atualiza os
centróides



Exemplo 1 - K Means



Alocação dos objetos aos clusters

$$\|A - C_1\|^2 = (5-5)^2 + (3-3)^2 = \mathbf{0}$$

$$\|A - C_2\|^2 = (5+1)^2 + (3+1)^2 = 61$$

A \rightarrow Cluster com centróide C1

$$\|B - C_1\|^2 = (-1-5)^2 + (1-3)^2 = 40$$

$$\|B - C_2\|^2 = (-1+1)^2 + (1+1)^2 = \mathbf{4}$$

B \rightarrow Cluster com centróide C2

$$\|C - C_1\|^2 = (1-5)^2 + (-2-3)^2 = 39$$

$$\|C - C_2\|^2 = (1+1)^2 + (-2+1)^2 = \mathbf{5}$$

C \rightarrow Cluster com centróide C2

$$\|D - C_1\|^2 = (-3-5)^2 + (-2-3)^2 = 41$$

$$\|D - C_2\|^2 = (-3+1)^2 + (-2+1)^2 = \mathbf{5}$$

D \rightarrow Cluster com centróide C2

**Não houve realocação de objetos,
portanto, o algoritmo convergiu e dois
clusters foram identificados: A e B,C,D**



Vantagens e Desvantagens

Vantagens

- Tendem a maximizar a dispersão entre os centros de gravidade dos clusters (clusters bem separados).
- Simplicidade de cálculo, calcula somente as distâncias entre os objetos e os centros de gravidade dos clusters.

Desvantagens

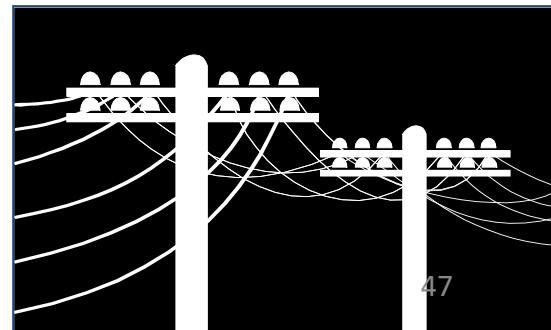
- A solução é dependente dos conjuntos de sementes iniciais, principalmente se a seleção das sementes é aleatória.
- Não há garantias de um agrupamento ótimo dos objetos.

Exemplo 2 - Classificação de distribuidoras de energia elétrica

Para fins de regulação econômica da distribuição de energia elétrica é interessante segmentar as empresas distribuidoras em clusters de forma a permitir análises comparativas entre empresas semelhantes.

Considere uma matriz de dados formada por 59 concessionárias de distribuição que atuam no setor elétrico brasileiro, cada uma descrita por 9 variáveis (Fonte: Aneel):

- Mercado atendido (MWh)
- Número de consumidores
- Tamanho da rede de distribuição (km)
- Densidade de consumidores (consumidores/km de rede)
- Consumo por unidades consumidora – CPC (MWh/consumidor)
- Índice de complexidade no combate às perdas não técnicas
- Composição do mercado por classe de tendão BT, MT e AT.



Exemplo 2 - Leitura e padronização dos dados

Leitura do arquivo de dados

```
setwd("c:/curso_R/")
dados = read.csv2("distribuidoras.csv",sep=",",dec=".",
header=TRUE)
empresas=dados[,2:10]
rownames(empresas)=dados[,1]
```

Padronização dos dados

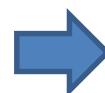
```
dados_padron=scale(empresas,center=TRUE,scale=TRUE)
```



$$\frac{x_{i,k} - \bar{X}_k}{S_k}$$

Em cada objeto i (empresa), o valor da variável k é subtraído da média da variável k e o resultado é dividido pelo desvio padrão da variável k

Para ver as médias e desvios padrão considerados no cálculo de `dados_padron`, digite `dados_padron` no R



```
attr(,"scaled:center")
      AT          MT          BT          Rede        Densidade       CPC
1.115508e-01 3.390610e-01 5.493898e-01 5.031273e+04 2.786153e+01 5.325424e+00
Consumidores    Mercado Complexidade
1.095152e+06 5.806429e+06 1.470068e-01
attr(,"scaled:scale")
      AT          MT          BT          Rede        Densidade       CPC
1.125434e-01 1.099551e-01 1.319779e-01 7.752209e+04 2.022090e+01 2.443612e+00
Consumidores    Mercado Complexidade
1.506989e+06 8.836134e+06 1.130079e-01
```

Médias das variáveis

Desvios padrão das variáveis

Exemplo 2 - Execução do K Means

```
# Execução do K Means para partição em 3 grupos de empresas (P,M,G)
saída_k_means=kmeans(dados_padron,3,nstart=10)
```

```
# Lista de objetos em saída_k_means
```

```
names(saída_k_means)
```

```
[1] "cluster"    "centers"    "totss"      "withinss"    "tot.withinss"  
[6] "betweenss"  "size"       "iter"       "ifault"
```

Exemplo 2 - Resultados

```
names(saida_k_means)
```

```
[1] "cluster"    "centers"     "totss"       "withinss"     "tot.withinss"  
[6] "betweenss"  "size"        "iter"        "ifault"
```

```
# Lista de alocação das empresas nos clusters
```

```
saida_k_means$cluster
```

```
> saida_k_means$cluster  
AES SUL          AMPLA      BANDEIRANTE      CEB          CEEE          CELESC          CEMIG          COPEL  
2                 3           3               1           1               3           3  
CPFL - Paulista CPFL - Piratininga ELEKTRO      ELETROPAULO      ESCELSA          LIGHT          RGE          AME  
3                 3           3               3           3               3           3  
CAIUAI          CEAL          CELG          CELPA          CELPE          CELTINS          CEMAR          CEMAT  
1                 1           1               1           3               1           1  
CEPISA          CLFSC         CNEE          COELBA          COELCE          COSERN          CSPE          EBO  
1                 1           1               3           3               1           2  
EDEVAP          EEB           EMG          ENERSUL          EPB            ESE          SULGIPE          BOA VISTA  
1                 2           1               1           1               3           1  
CFLO             CHESP         CJE           CLFM          COCEL          COOPERALIANÇA      CPEE          DEMEI  
2                 1           2               1           2               2           1  
DME-PC          EFLJC         EFLUL         ELETROACRE      ELETROCAR        ELFSM          ENF           HIDROPAN  
2                 1           2               1           1               1           1  
IENERGIA        MUX-Energia UHENPAL         1
```

```
# Total de elementos em cada cluster  
table(saida_k_means$cluster)
```

```
> table(saida_k_means$cluster)  
1   2   3  
28 13 18
```

Exemplo 2- Resultados

Lista de empresas no cluster 1

```
empresas1 = which(saida_k_means$cluster==1)
```

CEB 4	CEEE 5	CAIUÁ 17	CEAL 18	CELPÁ 20	CELTINS 22	CEMAR 23	CEMAT 24	CEPISA 25	CLFSC 26	CNEE 27	COSERN 30	EDEVPA 33	EMG 35	ENERSUL 36
EPB 37	SULGIPE 39	BOA VISTA 40	CHESP 42	CLFM 44	CPEE 47	DEMEI 48	EFLJC 50	ELETROACRE 52	ELETROCAR 53	ELFSM 54	ENF 55	UHENPAL 59		

Lista de empresas no cluster 2

```
empresas2 = which(saida_k_means$cluster==2)
```

AES SUL 1	RGE 15	CSPE 31	EEB 34	CFLO 41	CJE 43	COCEL 45	COOPERALIANÇA 46	DME-PC 49	EFLUL 51	HIDROPAN 56
IENERGIA 57	MUX-Energia 58									

Lista de empresas no cluster 3

```
empresas3 = which(saida_k_means$cluster==3)
```

AMPLA 2	BANDEIRANTE 3	CELESC 6	CEMIG 7	COPEL 8	CPFL - Paulista 9	CPFL - Piratininga 10	ELEKTRO 11
ELETROPAULO 12	ESCELSA 13	LIGHT 14	AME 16	CELG 19	CELPE 21	COELBA 28	COELCE 29
EBO 32	ESE 38						

Exemplo 2 - Resultados

Para ver os dados das empresas do cluster 1 faça
empresas[empresas1,]

	AT	MT	BT	Rede	Densidade	CPC	Consumidores	Mercado	Complexidade
CEB	0.0985	0.3052	0.5963	17064.54	48.45	6.30	826730	5210863	0.1315
CEEE	0.1278	0.3319	0.5403	71892.26	20.00	5.06	1438072	7277929	0.1877
CAIUA	0.0028	0.3534	0.6438	8344.33	24.69	4.81	206018	991036	0.0549
CEAL	0.1632	0.2924	0.5444	32079.00	25.93	2.95	831711	2453674	0.3175
CELPA	0.0989	0.3007	0.6003	92616.60	18.00	3.44	1666664	5734325	0.4581
CELTINS	0.0121	0.2581	0.7298	63436.47	6.59	2.95	417952	1232501	0.1875
CEMAR	0.0453	0.1958	0.7590	89929.45	18.77	2.12	1687939	3571718	0.3681
CEMAT	0.1200	0.3179	0.5622	106111.71	9.35	5.61	992365	5570387	0.1465
CEPISA	0.0708	0.2140	0.7152	48288.15	18.48	2.12	892390	1894082	0.2830
CLFSC	0.0487	0.3454	0.6059	9055.61	19.52	4.96	176767	876898	0.0543
CNEE	0.0129	0.3082	0.6789	3090.61	31.60	4.97	97669	485456	0.0399
COSERN	0.1961	0.2746	0.5293	43272.17	24.86	3.84	1075590	4130636	0.1734
EDEVP	0.0298	0.3042	0.6660	7375.96	21.21	4.60	156463	720177	0.0472
EMG	0.1591	0.2578	0.5831	25338.27	14.65	3.49	371253	1295720	0.0626
ENERSUL	0.0462	0.3449	0.6089	74503.98	10.53	4.39	784816	3441581	0.1212
EPB	0.2251	0.2193	0.5556	65373.65	16.21	2.89	1059575	3061728	0.2633
SULGIPE	0.1684	0.2824	0.5492	6905.84	16.86	2.21	116426	257335	0.2309
BOA VISTA	0.0000	0.3514	0.6486	2764.27	28.59	6.35	79029	501789	0.1625
CHESP	0.0000	0.1993	0.8007	3072.47	9.47	2.83	29110	82267	0.1068
CLFM	0.0000	0.3803	0.6197	1639.00	24.11	4.91	39522	193924	0.0341
CPEE	0.0000	0.3758	0.6242	2580.00	19.33	5.55	49883	276728	0.0268
DEMEI	0.0000	0.2508	0.7492	406.32	65.90	3.80	26777	101827	0.0382
EFLJC	0.0000	0.2730	0.7270	49.60	51.29	4.09	2544	10395	0.0028
ELETROACRE	0.0000	0.2169	0.7831	10914.95	17.43	3.19	190279	607916	0.2858
ELETROCAR	0.0000	0.3254	0.6746	2324.46	13.85	4.44	32185	142877	0.0521
ELFSM	0.0000	0.2423	0.7577	7404.27	11.54	4.42	85410	377751	0.0991
ENF	0.0000	0.2681	0.7319	1955.64	46.70	3.46	91336	316428	0.1134
UHENPAL	0.0000	0.2332	0.7668	1683.00	8.26	4.13	13909	57419	0.1129

Exemplo 2 - Resultados

```
# Mercado total atendido e extensão total das redes em cada cluster
total_mercado=rep(0,3)
total_clientes=rep(0,3)
total_rede=rep(0,3)
for (i in 1:3){
  if (i==1) {
    nomes=paste("cluster ",i,sep="")
  } else {
    nomes=c(nomes,paste("cluster ",i,sep ""))
  }
  selec=which(saida_k_means$cluster==i)
  total_rede[i]=sum(empresas[selec,4])
  total_clientes[i]=sum(empresas[selec,7])
  total_mercado[i]=sum(empresas[selec,8])
}
```

Exemplo 2 - Resultados

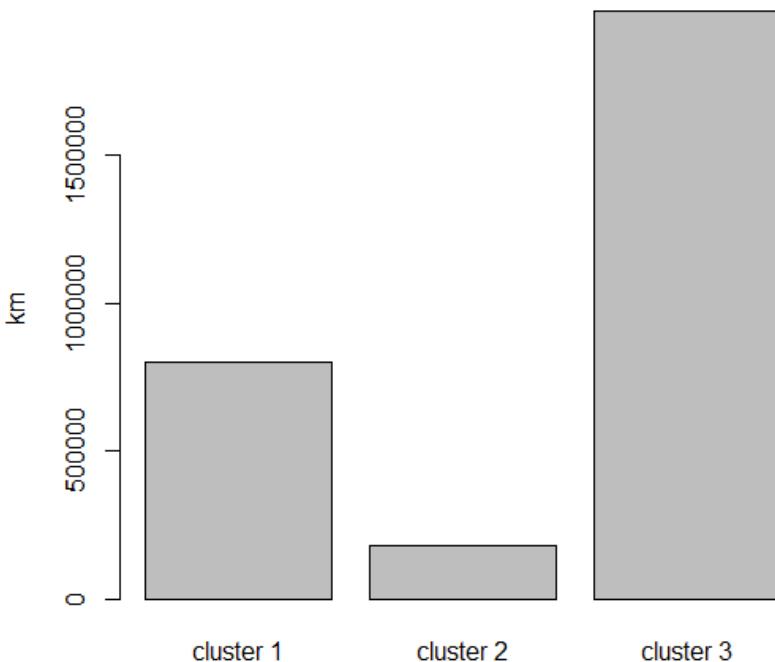
Gráficos da extensão total das redes em cada cluster

```
names(total_rede)=nomes
```

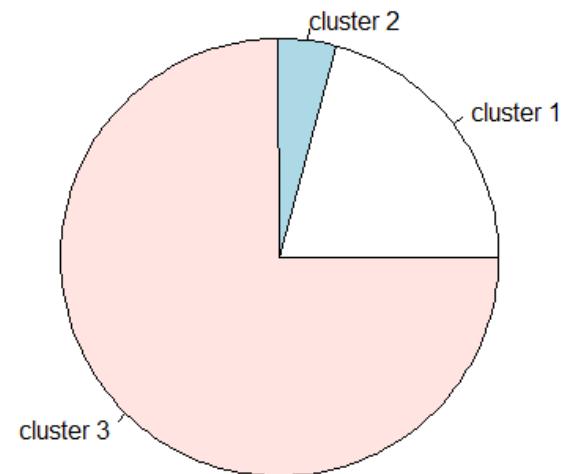
```
barplot(total_rede,ylab="km",main="Extensão total da rede")
```

```
pie(total_clientes,main="Extensão total da rede")
```

Extensão total da rede



Extensão total da rede



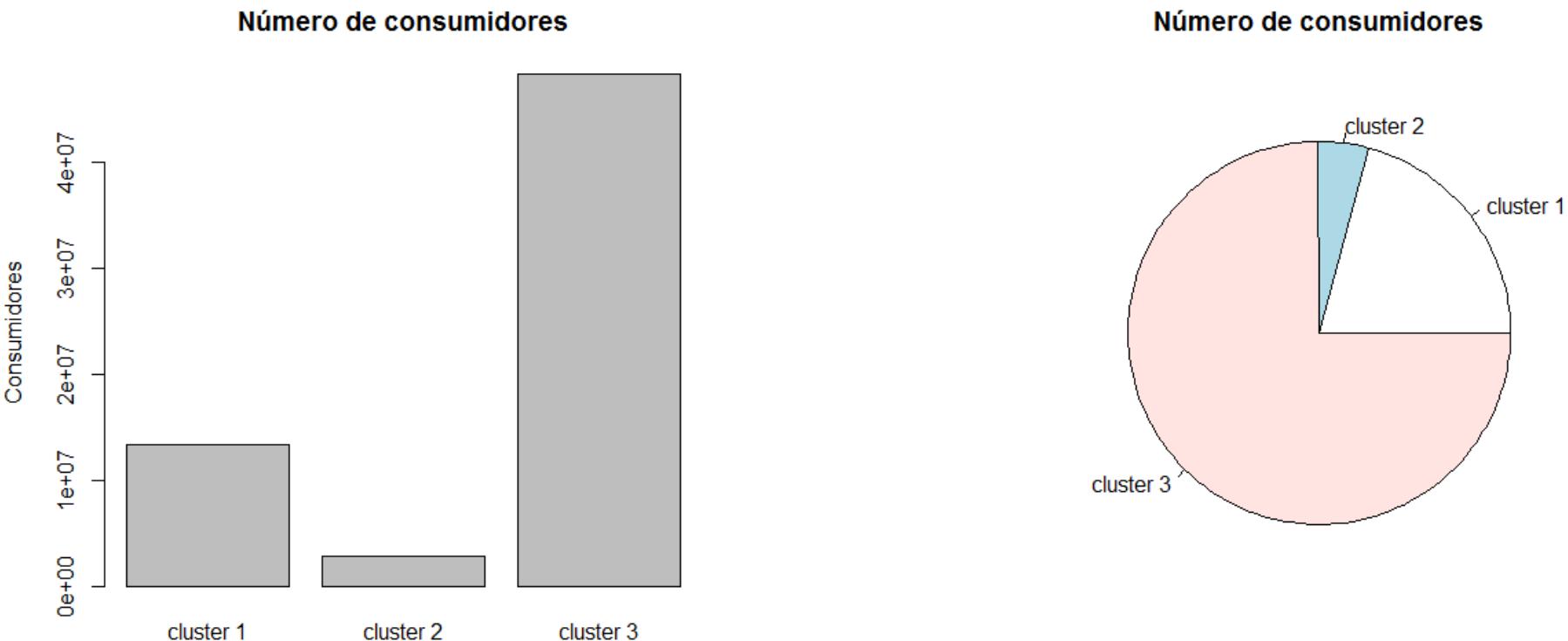
Exemplo 2 - Resultados

Gráficos do total de unidades consumidoras em cada cluster

```
names(total_clientes)=nomes
```

```
barplot(total_clientes,ylab="Cosumidores",main="Número de consumidores")
```

```
pie(total_clientes,main="Número de consumidores")
```



Exemplo 2 - Resultados

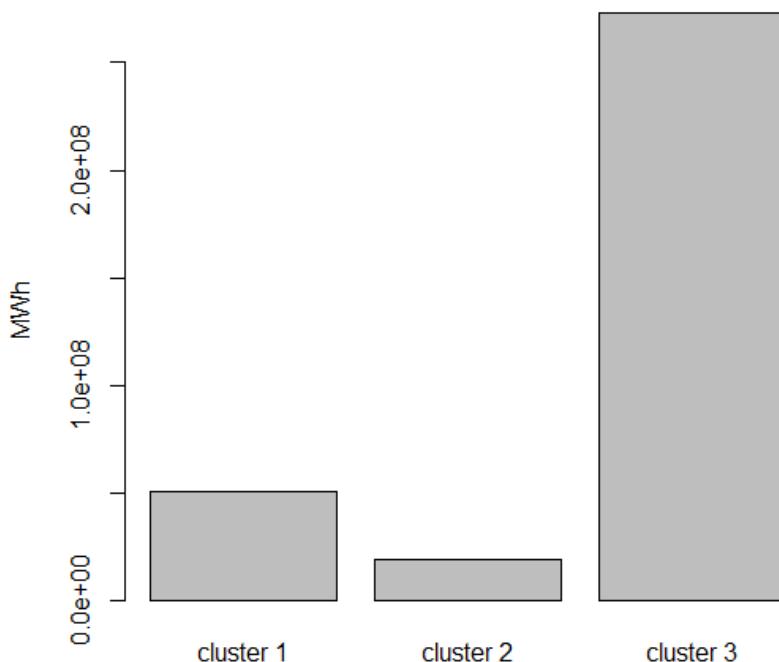
Gráficos do mercado total atendido em cada cluster

```
names(total_mercado)=nomes
```

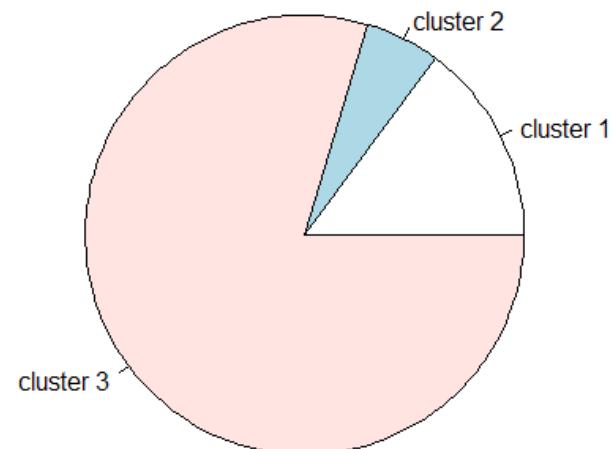
```
barplot(total_mercado,ylab="MWh",main="Total do mercado atendido")
```

```
pie(total_mercado,main="Total do mercado atendido")
```

Total do mercado atendido



Total do mercado atendido



Exemplo 2 - Resultados

```
names(saida_k_means)
```

```
[1] "cluster"    "centers"
```

```
[6] "betweenss"  "size"
```

```
"totss"        "withinss"      "tot.withinss"  
"iter"         "ifault"
```

```
# Inércia total
```

```
saida_k_means$totss
```

Inércia total = 522

```
# Inércia dentro dos agrupamentos (WSSi, i =1:k clusters)
```

```
saida_k_means$withinss
```

Inércias dentro dos grupos = 86,15 59,89 167,07

```
# Inércia total dentro os agrupamentos (WSS)
```

```
saida_k_means$tot.withinss
```

Inércia intra cluster = 313,10 (60%)

```
# Inércia total entre os agrupamentos (BSS)
```

```
saida_k_means$betweenss
```

Inércia inter cluster = 208,90 (40%)

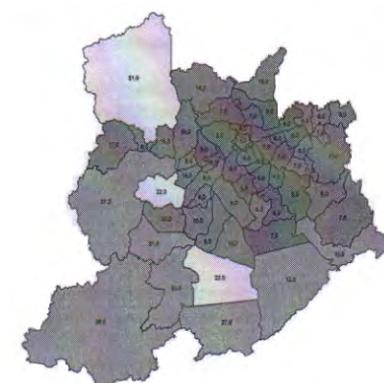
Exemplo 3 - Definição das metas de continuidade

A continuidade do fornecimento é uma das dimensões da qualidade da energia elétrica.

A continuidade do fornecimento de energia elétrica é avaliada por dois indicadores: DEC (duração média da interrupção do fornecimento) e FEC (freqüência média de interrupções do fornecimento).

As metas de continuidade do fornecimento de energia elétrica são padrões de qualidade definidos por uma análise comparativa dos desempenhos apurados nos conjuntos de unidades consumidoras.

Conjuntos de unidades consumidoras
são subdivisões de uma área de concessão



Exemplo 3 – Regulação por comparação (*Yardstick competition*)

- Baseia-se no princípio de que áreas com semelhantes condições de atendimento devem ter desempenhos compatíveis.
- Condições de atendimento representadas pelos atributos nº de unidades consumidoras, consumo anual (MWh), extensão da rede aérea primária (km), potência instalada (kVA), área (km^2) e tipo do sistema (isolado ou interligado)
- A identificação das áreas semelhantes pode ser efetuada por meio de uma técnica de *cluster analysis*, por exemplo, o *K Means*.
- Áreas classificadas em um mesmo *cluster* devem receber os mesmos padrões de qualidade. As áreas classificadas em um mesmo *cluster* devem “competir” com o respectivo padrão de qualidade.
- Evita o problema de assimetria de informação e estabelece metas diferentes para cada empresa
- O ponto sensível da metodologia é a confiabilidade dos dados analisados



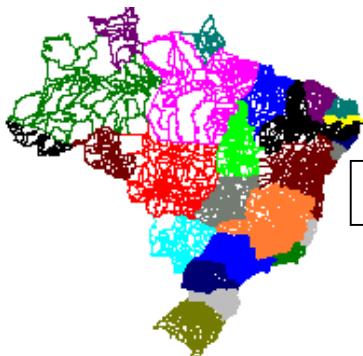
Exemplo 3 - K Means na definição das metas de continuidade do fornecimento de energia elétrica

A Resolução ANEEL 024/2000: introduziu a análise comparativa (yardstick competition) dos desempenhos dos conjuntos de unidades consumidoras, como meio de definição das metas dos indicadores DEC e FEC.

A idéia é estabelecer metas diferenciadas que reconhecem a diversidade física e econômica das regiões atendidas pelas distribuidoras.

A análise comparativa e a definição das metas de continuidade está implementada no ANABENCH, um sistema computacional desenvolvido pelo CEPEL.

Cerca de 6000 conjuntos de unidades consumidoras informados pelo sistema GESTTOR/ANEEL



Atributos dos conjuntos:

- km de rede aérea primária – ERAP
- área do conjunto – AREA
- potência instalada – PNI
- número de consumidores - NUC
- consumo médio - CMM
- Isolado ou interligado

Formação dos clusters de conjuntos semelhantes pelo K-Means

Clusters

Definição das metas de DEC e FEC em cada cluster

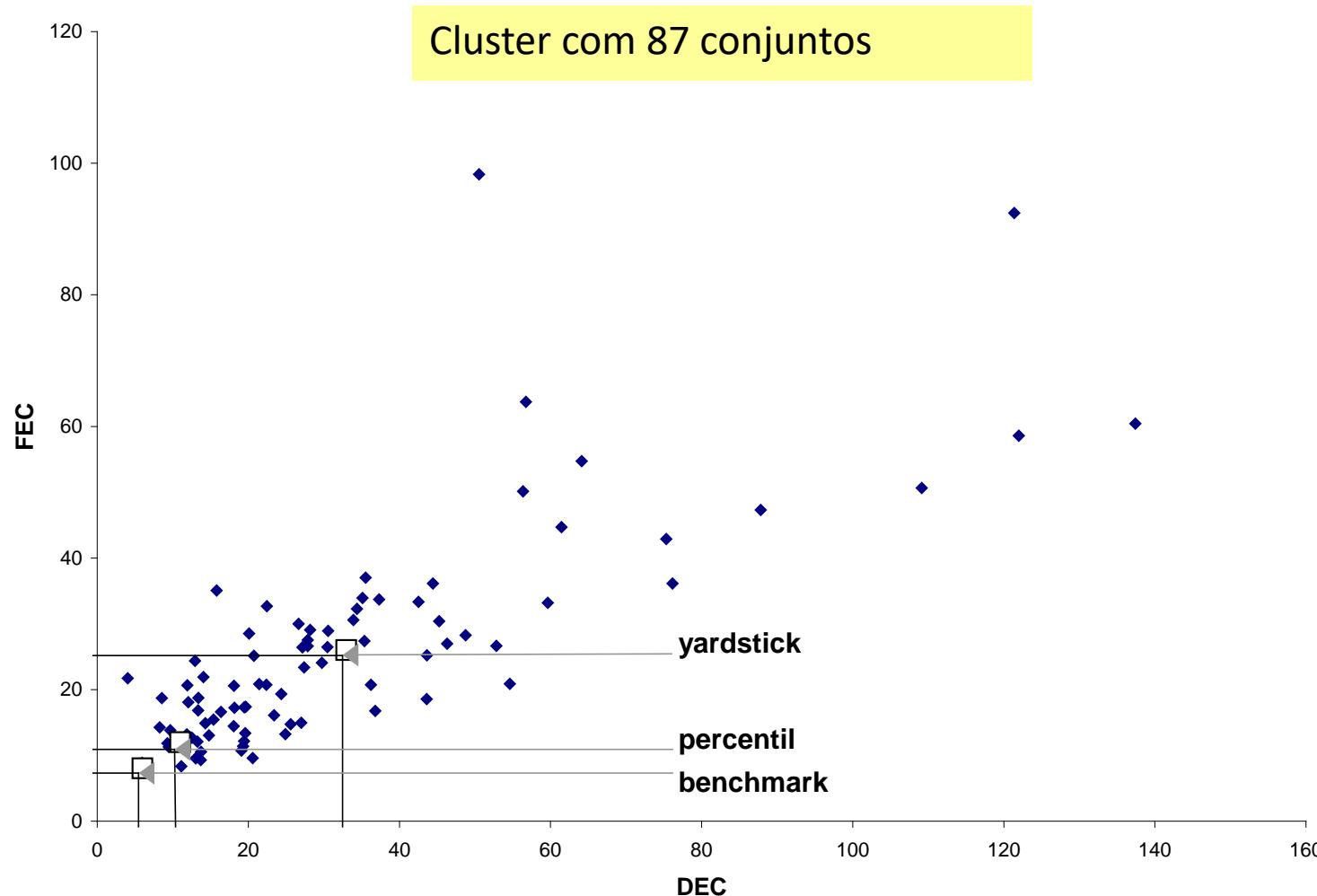
Metas

Exemplo 3 - Definição das metas de continuidade

Resolução ANEEL nº 501/2000
Metas de continuidade dos conjuntos da Light

Conjunto de unidades consumidoras	DEC			FEC		
	2001	2002	2003	2001	2002	2003
BAIA DE GUARATIBA	22	21	20	19	18	17
BANGU	9	9	9	10	10	10
BARRA	6	6	6	7	7	7
BARRA DO PIRAI	10	10	10	13	12	12
BARRA DO PIRAI NAO URBANO	16	16	16	15	14	14
BARRA MANSA	15	14	14	11	11	11
BARRA MANSA NAO URBANO	13	12	12	11	11	11
BELFORD ROXO	18	17	16	14	13	13
BOTAFOGO	6	6	6	6	6	6
CAMPO GRANDE	10	10	10	8	8	8
CASCADURA	14	13	13	11	10	10
CATUMBI FAVELAS	20	20	20	22	21	21
CAVA	12	12	12	10	10	10
CAXIAS	12	11	11	8	8	8
CENTRO	5	5	5	3	3	3
COMPLEXO DA MARE FAVELAS	16	16	16	13	13	13
COPACABANA	5	5	5	3	3	3
CURICICA	16	15	15	15	15	14
DEMOCRATICOS	12	12	12	11	11	11
FAZENDA BOTAFOGO FAVELAS	22	22	21	20	19	19
FLAMENGO	5	5	5	3	3	3
FLORESTA DA TIJUCA	18	16	15	17	15	14
FREGUESIA	12	12	11	13	12	12
GAVEA	7	7	7	8	8	8
GRAJAU FAVELAS	21	21	20	26	25	24
ILHA DE PAQUETA	11	11	11	10	10	10
ILHA DO GOVERNADOR	9	9	9	9	9	9
IRAJA	10	10	10	9	9	9
ITAGUAI	20	19	18	22	21	20
ITAGUAI NAO URBANO	18	17	17	18	17	17

Exemplo 3 - Definição das metas de continuidade



Três critérios para definição das metas de continuidade :

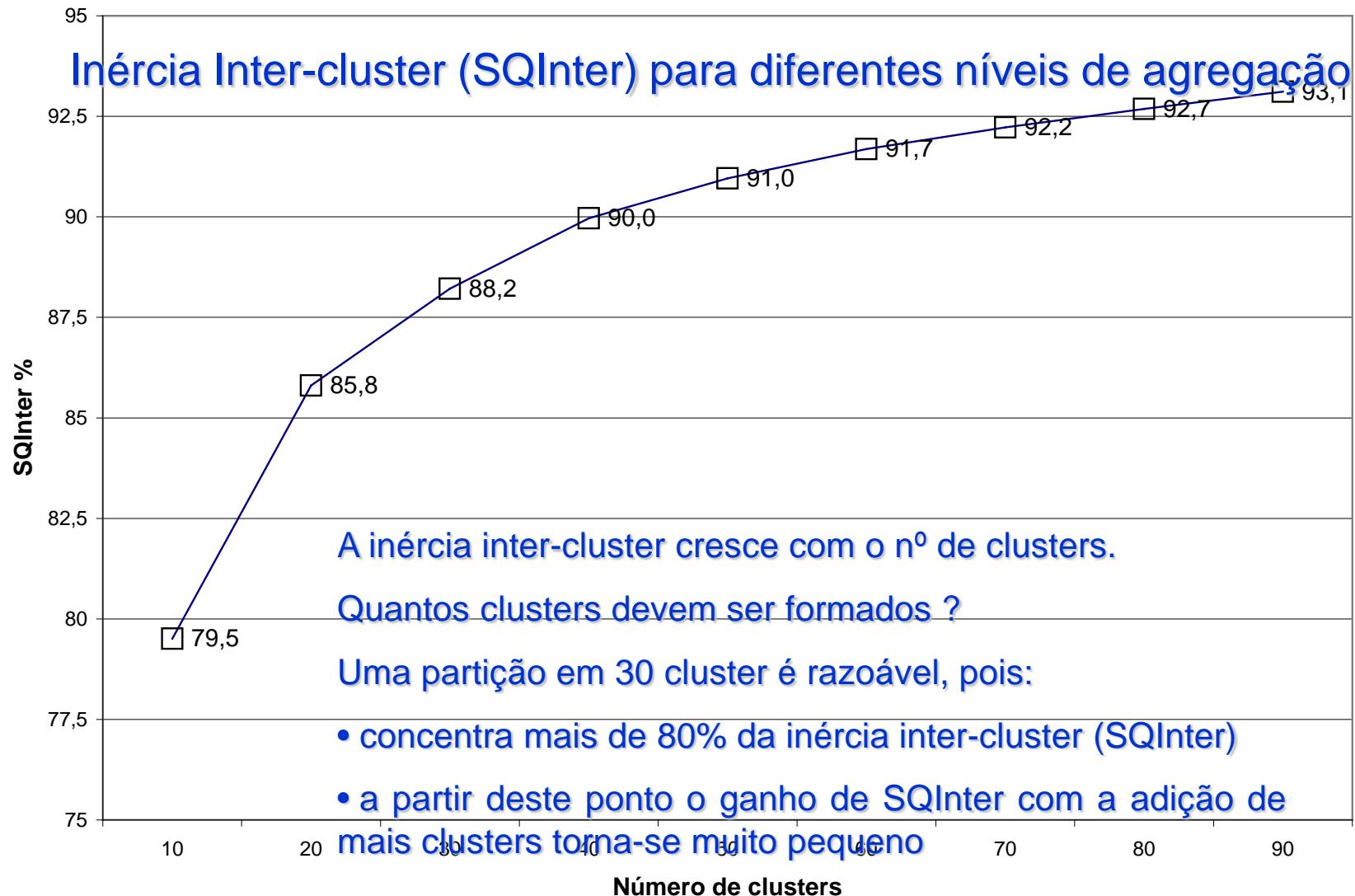
Yardstick (DEC e FEC médios ou medianos)

Benchmarking (menores DEC e FEC)

Percentil (percentil das distribuições do DEC e FEC)

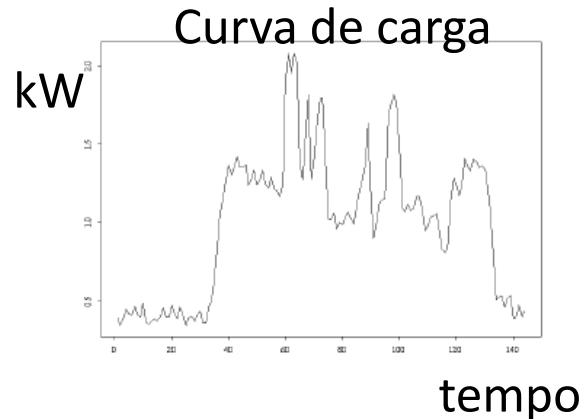
Exemplo 3 – Quantos cluster?

% Inéria interclasse em função do número de agrupamentos



Exemplo 4 - Caracterização da carga elétrica

- Os perfis horários da demanda de eletricidade nas diversas classes de consumo (residencial, comercial, industrial) e pontos da rede de distribuição de energia elétrica são informações fundamentais na operação e na expansão de um sistema elétrica.
- Perfis típicos (tipologias) são obtidos através da análise de uma grande quantidade de medições de curvas de carga
- Usos das tipologias :
 - ✓ tarifação dos sistemas de distribuição
 - ✓ gerenciamento pelo lado da demanda

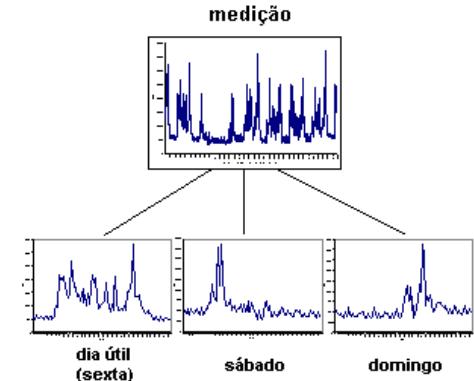


Exemplo 4 - Caracterização da carga elétrica

Etapas da construção de tipologias de curvas de carga

1 – Coleta de medição de curvas de carga de clientes e redes

2 - Inspeção visual de cada medição para identificar três curvas diárias características do ponto de medição: um dia útil , um sábado e um domingo típicos.



3 - Uso de algum algoritmo de análise de agrupamentos para fazer a identificação das tipologias em cada classe e faixa de consumo ou tipo de rede



4 - Ajuste das tipologias ao mercado de energia elétrica da classe que as curvas representam (MWh)



Rio de Janeiro, v.7, n.1, p. 29-54, janeiro a abril de 2015

CONSTRUINDO TIPOLOGIAS DE CURVAS DE CARGA COM O PROGRAMA R

Resumo

Os perfis típicos de curvas de carga de consumidores e redes constituem informações fundamentais para a determinação das tarifas de uso dos sistemas de distribuição de energia elétrica. Destaca-se que a sinalização horária das tarifas é determinada em grande parte pelos perfis típicos da demanda por eletricidade. Neste artigo é apresentada uma implementação computacional em ambiente R dos métodos K-Means, K-Medoides e Ward, três métodos estatísticos multivariados para análise de agrupamentos, úteis na identificação de perfis típicos da demanda horária por eletricidade, uma etapa crítica do processo de revisão tarifária das distribuidoras de energia elétrica. O presente artigo contribui no sentido de fornecer uma alternativa eficaz e econômica para a construção das tipologias de curvas de carga.

Palavras-Chave: Análise de agrupamentos, K-Means, K-Medoides, Método de Ward, curvas de carga, tipologias

Exemplo 4 - Tipologias de curvas de carga (Pessanha et al, 2015)

Importação de dados

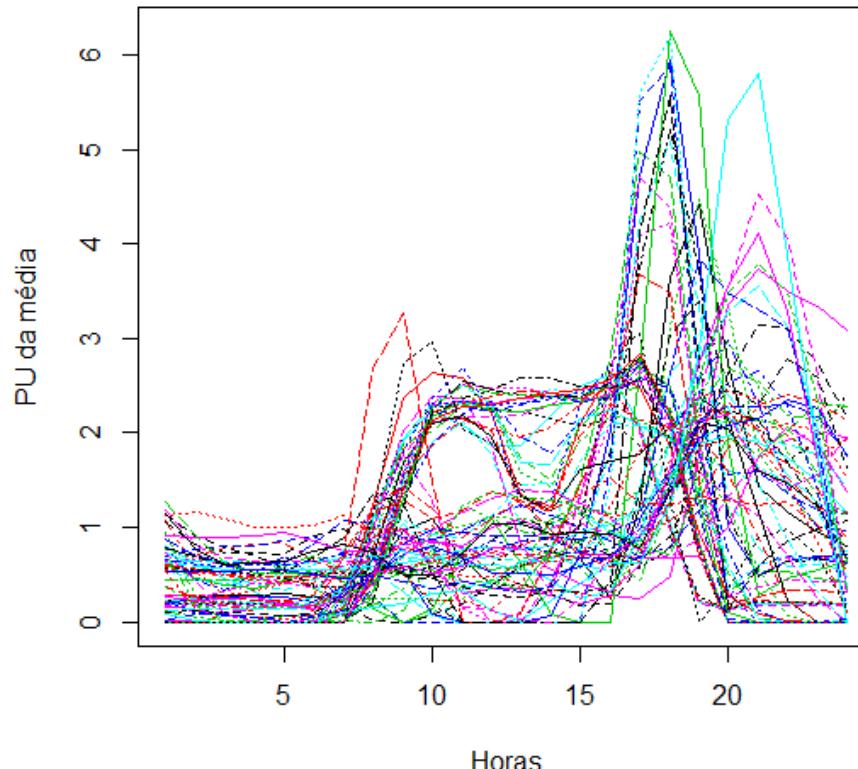
```
setwd("c:/curso R/aula4") # estabelece o diretório de trabalho  
dados = read.csv2("curvas_de_carga.csv",sep=";",dec=". ",header=T,row.names=1)
```

dimensões da matriz de dados > `dim(dados)` 74 curvas (linhas)

```
dim(dados)  
[1] 74 24 24 demandas horárias (colunas)
```

gráfico

```
matplot(matrix(seq(1,24,1),ncol=1),t(dados),type='l',ylab='PU da média',xlab='Horas')
```



Exemplo 4 - K Means no R

partição em 3 clusters

```
resultado.kmeans = kmeans(dados,centers=3,iter.max=1000,nstart=10)
```

resultados guardados no objeto resultado.kmeans

```
names(resultado.kmeans)
```

```
> names(resultado.kmeans)
[1] "cluster"      "centers"       "totss"        "withinss"      "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
```

número de curvas em cada cluster

```
resultado.kmeans$size
```

```
> resultado.kmeans$size
[1] 37 11 26
```

inércia total TSS

```
resultado.kmeans$totss
```

```
> resultado.kmeans$totss
[1] 1218.154
```

inércia entre os agrupamentos BSS

```
resultado.kmeans$betweenss
```

```
> resultado.kmeans$betweenss
[1] 866.7029
```

inércia dentro dos agrupamentos WSS

```
resultado.kmeans$tot.withinss
```

```
> resultado.kmeans$tot.withinss
[1] 351.4511
```

inércia em cada agrupamento

```
resultado.kmeans$withinss
```

```
> resultado.kmeans$withinss
```

```
[1] 216.72942 68.10920 66.61251
```

Exemplo 4 - K Means no R Composição dos clusters

alocação das curvas nos 3 clusters

resultado.kmeans\$cluster

```
> resultado.kmeans$cluster
```

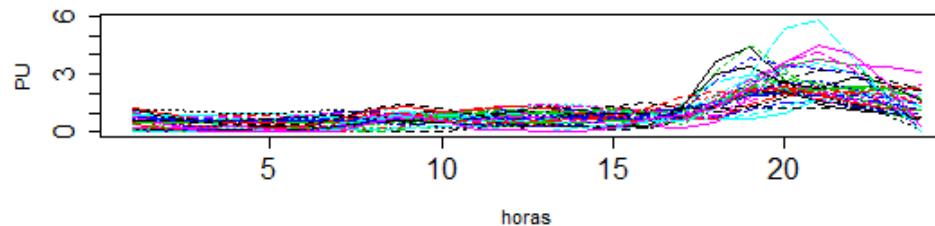
Cliente 1	Cliente 2	Cliente 3	Cliente 4	Cliente 5	Cliente 6	Cliente 7
1	1	1	1	1	1	1
Cliente 8	Cliente 9	Cliente 10	Cliente 11	Cliente 12	Cliente 13	Cliente 14
1	1	1	3	3	3	3
Cliente 15	Cliente 16	Cliente 17	Cliente 18	Cliente 19	Cliente 20	Cliente 21
2	2	2	2	2	2	2
Cliente 22	Cliente 23	Cliente 24	Cliente 25	Cliente 26	Cliente 27	Cliente 28
2	2	2	2	3	3	3
Cliente 29	Cliente 30	Cliente 31	Cliente 32	Cliente 33	Cliente 34	Cliente 35
3	3	3	3	3	3	3
Cliente 36	Cliente 37	Cliente 38	Cliente 39	Cliente 40	Cliente 41	Cliente 42
1	1	1	1	1	1	1
Cliente 43	Cliente 44	Cliente 45	Cliente 46	Cliente 47	Cliente 48	Cliente 49
1	1	1	1	1	3	3
Cliente 50	Cliente 51	Cliente 52	Cliente 53	Cliente 54	Cliente 55	Cliente 56
3	3	3	3	3	3	3
Cliente 57	Cliente 58	Cliente 59	Cliente 60	Cliente 61	Cliente 62	Cliente 63
3	3	1	1	1	1	1
Cliente 64	Cliente 65	Cliente 66	Cliente 67	Cliente 68	Cliente 69	Cliente 70
1	1	1	1	1	1	1
Cliente 71	Cliente 72	Cliente 73	Cliente 74			
1	1	3	1			

Exemplo 4 - K Means no R Visualização dos clusters

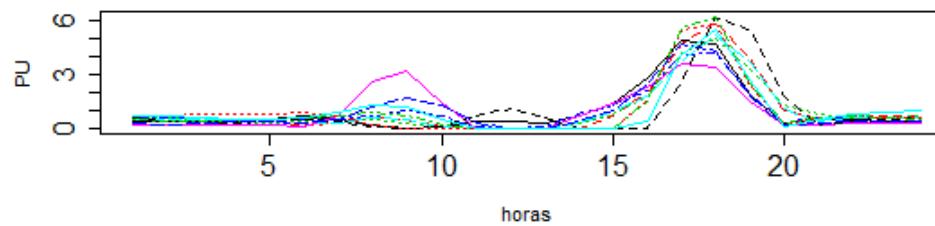
Visualização dos 3 clusters

```
par(mfrow=c(3,1))
for (i in 1:3) {
  cluster = which(resultado.kmeans$cluster ==i)
  matplot(matrix(seq(1,24,1),ncol=1),t(dados[cluster,]),type='l',ylab='PU',xlab='horas'
  ,main=paste('cluster',i),cex.main=2,cex.axis=1.5)
}
```

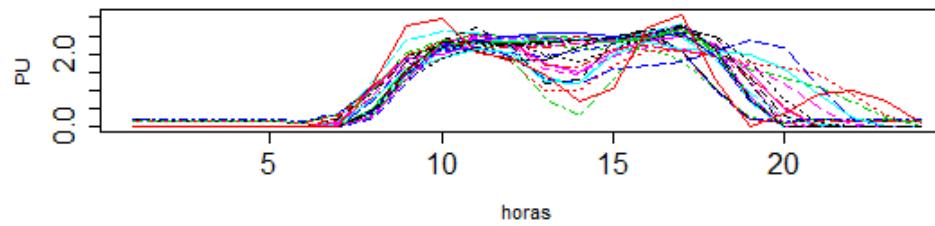
cluster 1



cluster 2



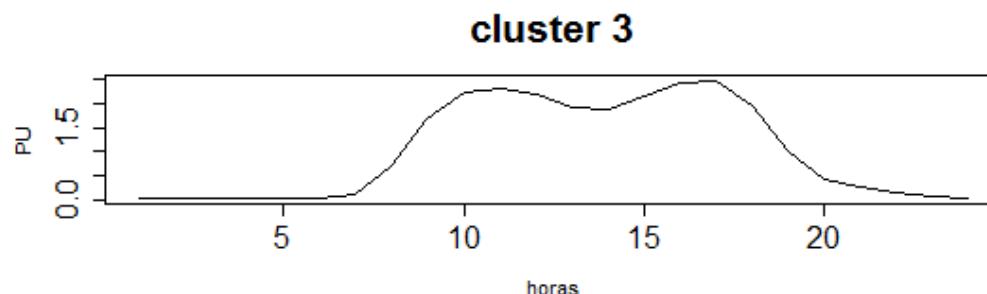
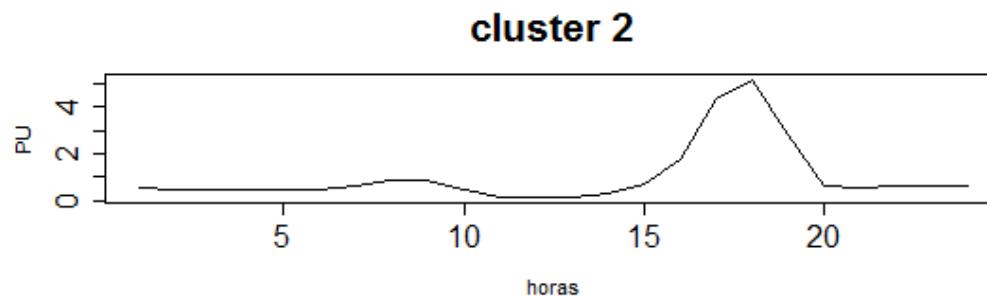
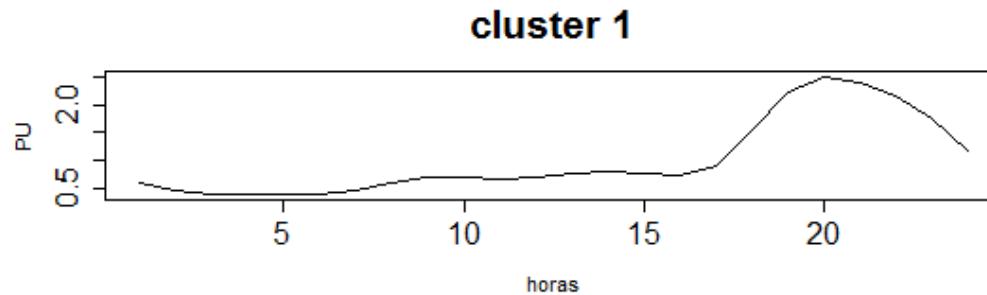
cluster 3



Exemplo 4 - K Means no R Visualização dos centroides

Centroides dos 3 clusters

```
par(mfrow=c(3,1))
for (i in 1:3) {
  centro=resultado.kmeans$center[i,]
  plot(centro,type='l',ylab='PU',xlab='horas',main=paste('cluster',i),cex.main=2,cex.a
    xis=1.5)
}
```



Exemplo 4 - Identificação do melhor número de clusters

O número adequado de clusters (k) deve maximizar o pseudo-F

$$pseudo-F = \frac{BSS/(k-1)}{WSS/(N-k)}$$

Quadrados médios entre clusters

Quadrados médios dentro dos clusters

```
pseudoF = numeric(0)
bss = numeric(0)
nobs=dim(dados)[1] # número de curvas de carga
for (i in 2:9) {
  resultado.kmeans = kmeans(dados,i,nstart=10) # partitiona em i clusters
  # pseudo F para a solução com i clusters
  auxpseudoF = (resultado.kmeans$betweenss/(i-1))/(resultado.kmeans$tot.withinss/(nobs-i))
  # parcela da BSS na inércia total na solução com i clusters
  auxbss = resultado.kmeans$betweenss/resultado.kmeans$totss
  # vetores com as estatísticas PseudoF e BSS em várias soluções
  pseudoF = c(pseudoF,auxpseudoF)
  bss = c(bss,auxbss)
}
par(mfrow=c(1,2))
plot(seq(2,9,1),pseudoF,xlab= 'número de clusters ',ylab= 'pseudoF ',pch=1)
plot(seq(2,9,1),bss*100,xlab= 'número de clusters ',ylab= '%BSS ',pch=1)
```

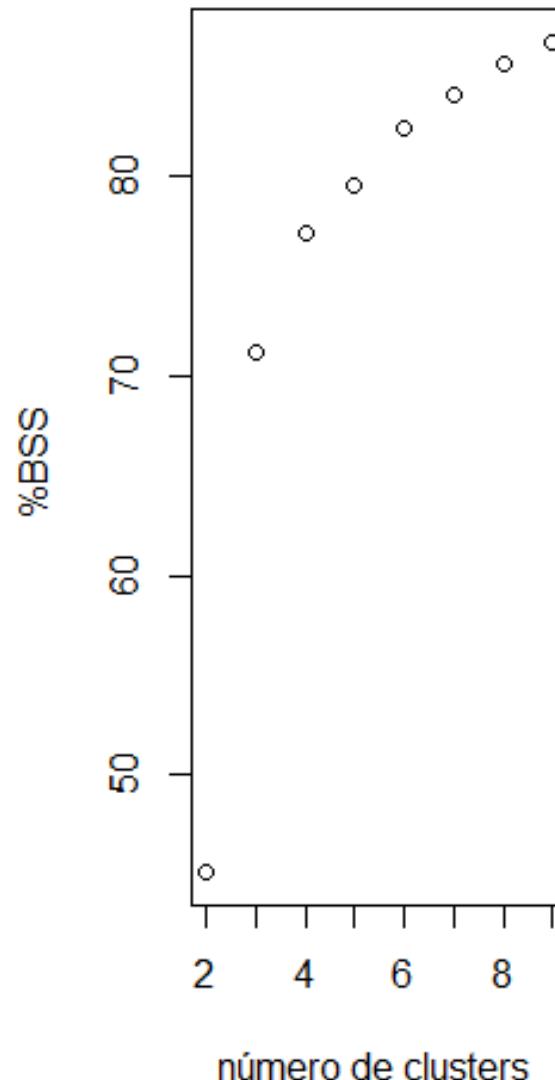
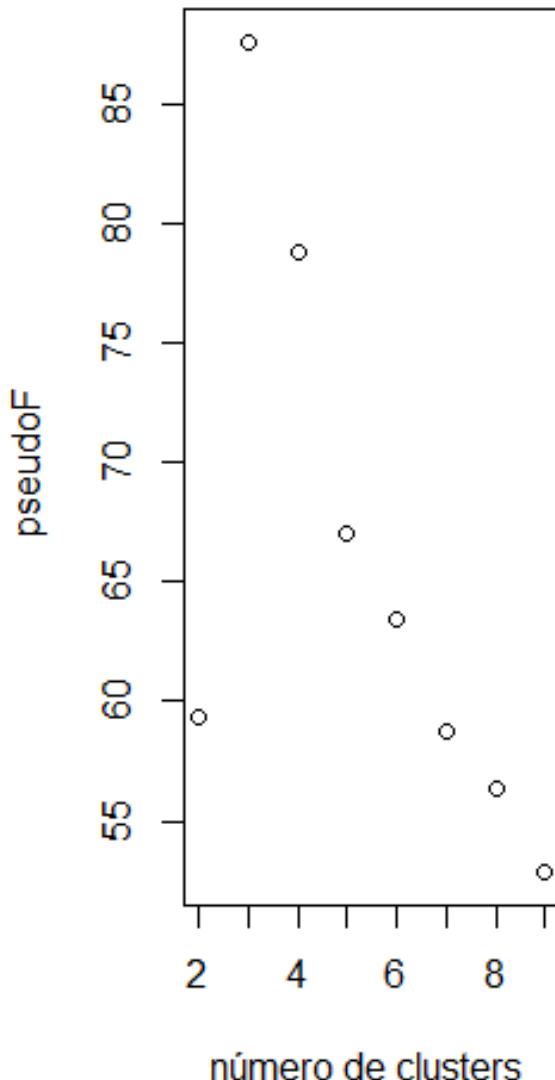


Lattin,J.; Carrol, J.D. & Green, P.E. (2011). Análise de Dados Multivariados. São Paulo: Cengage Learning.

Exemplo 4 - Identificação do melhor número de clusters

Melhor solução com 3 clusters

Solução com 3 cluster maximiza o Pseudo F e a BSS concentra mais de 70% da inércia total



Nuvens dinâmicas

Método não hierárquico semelhante ao K-Means, porém o conjunto de K centros de gravidade dos clusters é substituído por um sistema de K núcleos, cada um formado por um conjunto de q indivíduos.

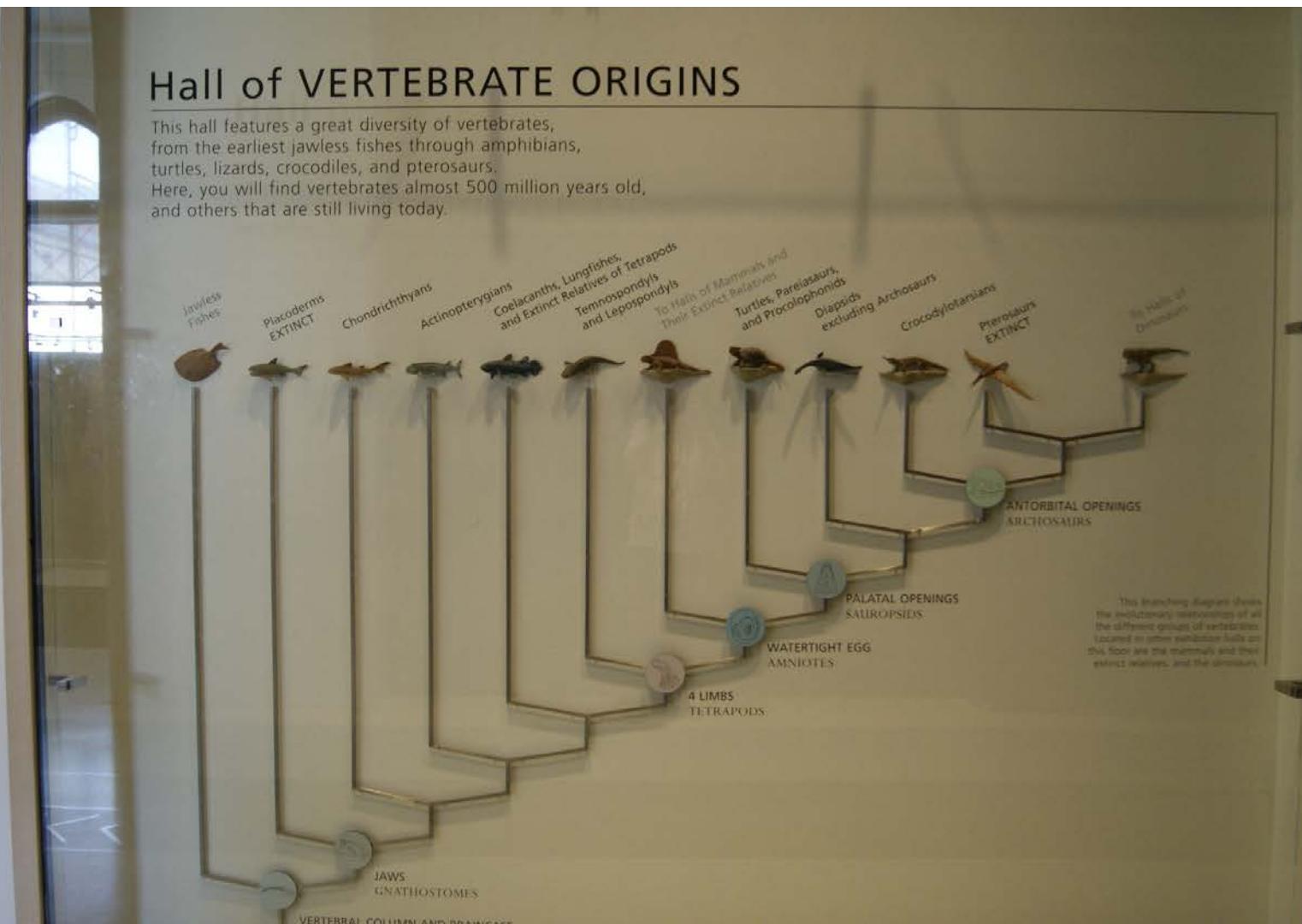
A idéia é que por ser formado por um conjunto de indivíduos, o núcleo tem um poder descritivo maior que o centro de gravidade do K-Means formado por apenas um objeto do núcleo

Os núcleos possuem um nº constante de q elementos. Tanto o número de elementos dos núcleos (q) quanto o número de núcleos (K) são parâmetros de entrada.

O sistema inicial de K núcleos pode ser fixado por uma tiragem aleatória dos elementos da população.

Definidos os núcleos iniciais o algoritmo das nuvens dinâmicas classifica os objetos nos núcleos mais próximos formando clusters de objetos. Em seguida, os núcleo são atualizados. Com os novos núcleos, novos clusters são definidos e novos núcleos são gerados e assim sucessivamente. As etapas de constituição dos clusters a atualização dos núcleos são executadas iterativamente até a convergência do algoritmo quando os clusters já não se modificam.

Métodos hierárquicos: Métodos de encadeamento, Método de Ward

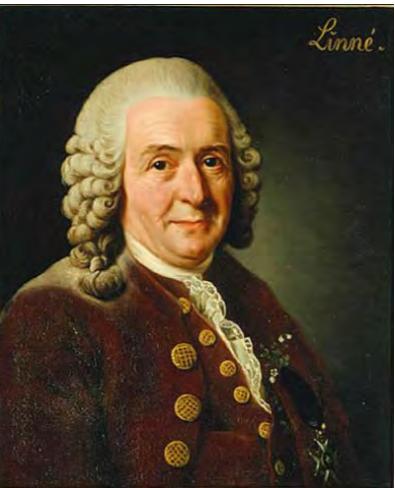


This branching diagram shows the evolutionary relationships of all the different groups of vertebrates. Located in other exhibition halls on this floor are the mammals and their extinct relatives, and the dinosaurs.

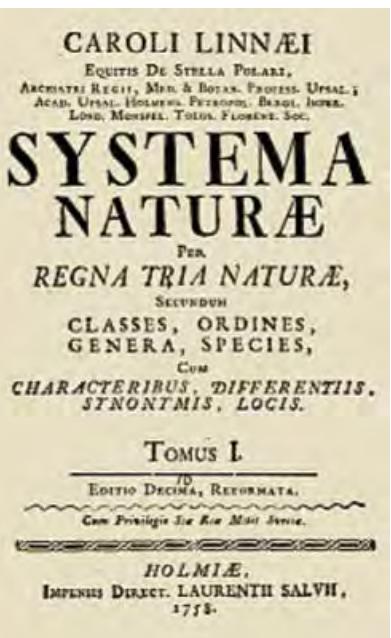
→ Miriam
Inform
Museu
← Hall of
Librar
David
Hall

Métodos hierárquicos

Linné.

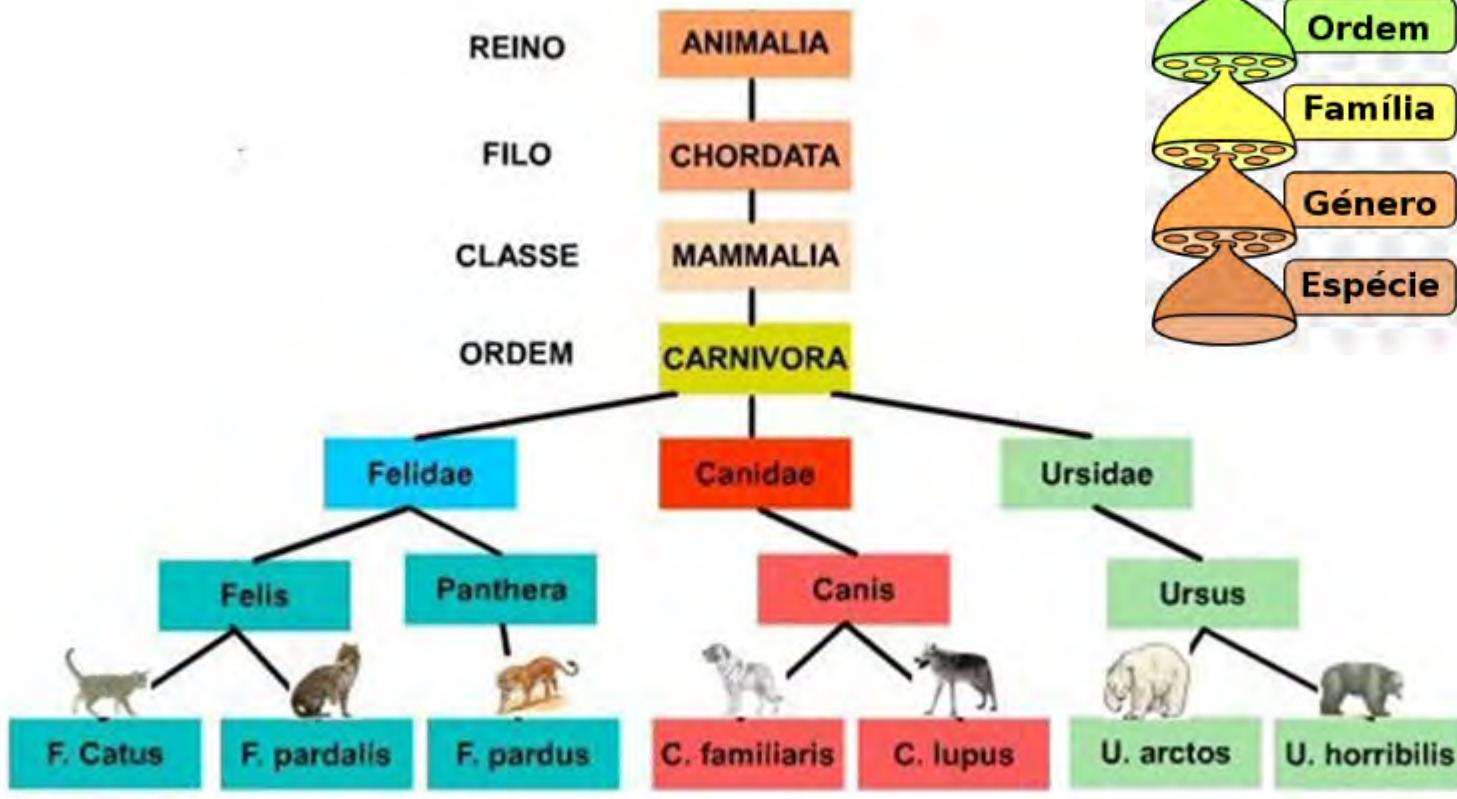


Carl von Linné
1707 - 1778



Taxonomia de Lineu

Classificação hierárquica dos seres vivos



Métodos hierárquicos

Aglomerativo (método mais comum)

- No início cada objeto forma um *cluster* que sucessivamente sofre uma série de fusões com outros *clusters* até que no final todos os objetos estejam em um único agrupamento.
- Um *cluster* formado em uma dada interação corresponde a união de *clusters* formados em passos anteriores.

Divisivo

- No início há apenas um *cluster* formado pelo conjunto de objetos que é dividido sucessivamente até que no final cada *cluster* contenha apenas um objeto.
- *Clusters* formados em uma dada interação correspondem a fragmentação de um *cluster* formado no passo anterior.

Métodos hierárquicos divisivos



Métodos hierárquicos aglomerativos





Métodos hierárquicos aglomerativos

- Encadeamento Simples ou vizinho mais próximo
(Single linkage ou nearest neighbor)
- Encadeamento Completo ou vizinho mais longe
(Complete linkage ou furthest neighbor)
- Encadeamento Médio (*Average linkage*)
- Método de Ward (*Ward's method*)
- Método do Centroide (*Centroid method*)

Diferem na maneira de como calculam as distâncias entre os *clusters*.



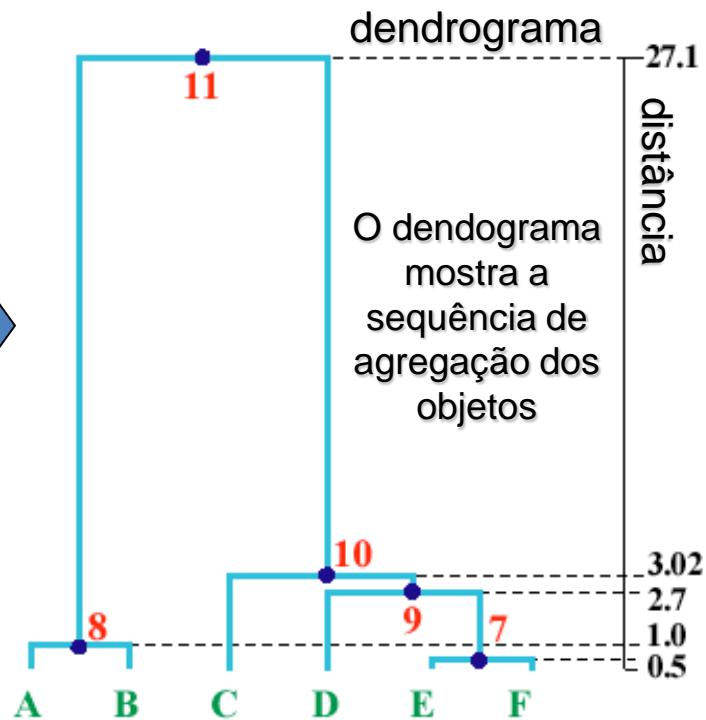
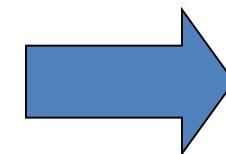
Métodos hierárquicos aglomerativos

Particionam um conjunto com N objetos seqüencialmente em 1,2,3,4 até N clusters, obtendo no final uma estrutura em árvore, semelhante as classificações zoológicas (reino, espécies, gêneros, famílias, ordem, etc.).

Produzem uma série de partições encaixadas.

O resultado é apresentado na forma de uma árvore de classificação conhecida por dendrograma

objetos		
	X	Y
A	1	1
B	2	2
C	3,5	4,5
D	5,5	3
E	6	5
F	5	5

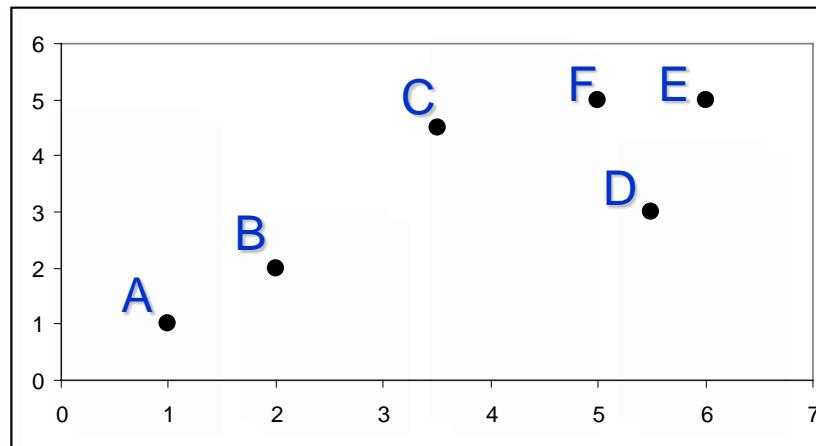
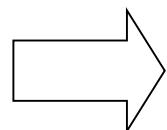




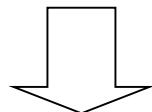
Métodos hierárquicos aglomerativos

Objetos

	X	Y
A	1	1
B	2	2
C	3,5	4,5
D	5,5	3
E	6	5
F	5	5



Solução inicial com 6 clusters, cada um com 1 objeto



Matriz de distâncias entre os clusters
(matriz simétrica)

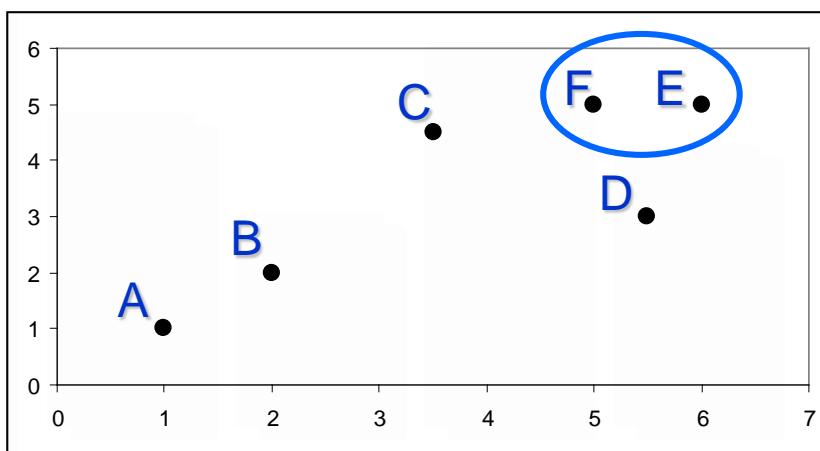
	A	B	C	D	E	F
A	0					
B	1	0				
C	9,250	4,250	0			
D	12,125	6,625	3,125	0		
E	20,500	12,500	3,250	2,125	0	
F	16,000	9,000	1,250	2,125	0,500	0

Qual o par de objetos mais próximos
(objetos mais semelhantes) ?

É o par E,F, logo estes objetos são
os primeiros a serem agrupados



Métodos hierárquicos aglomerativos



Solução intermediária com 5 clusters.

As distâncias passam a ser calculadas em relação ao cluster e não mais em relação aos seus objetos

Atualiza matriz de distâncias

Matriz de distâncias entre os clusters
(matriz simétrica)

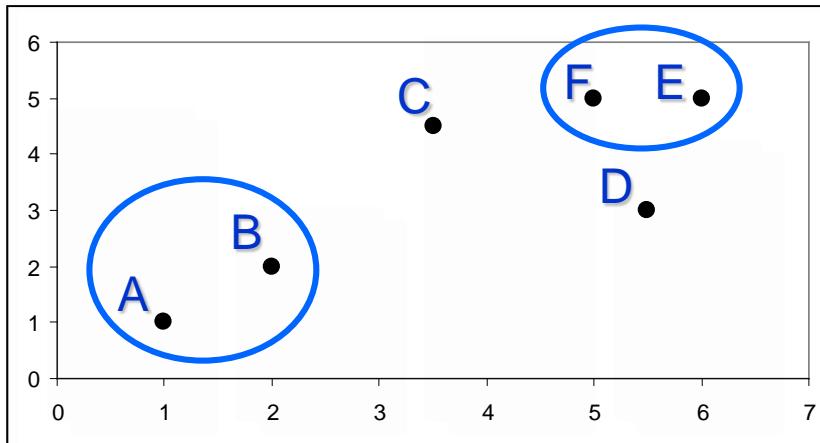
	A	B	C	D	E-F
A	0				
B	1,000	0			
C	9,250	4,250	0		
D	12,125	6,625	3,125	0	
E-F	24,167	14,167	2,833	2,667	0

Qual o par de objetos mais próximos?

É o par A,B, logo estes objetos são os próximos a serem agrupados



Métodos hierárquicos aglomerativos



Solução
intermediária com
4 clusters

↓ Atualiza matriz de distâncias

Matriz de distâncias entre os clusters
(matriz simétrica)

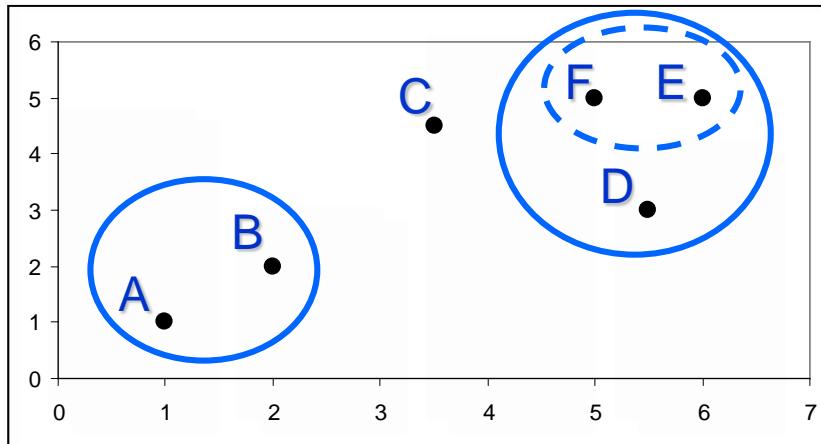
	A-B	C	D	E-F
A-B	0			
C	8,667	0		
D	12,167	3,125	0	
E-F	28,250	2,833	2,667	0

Qual o par de objetos mais
próximos?

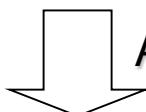
É o par (E,F),D logo estes objetos
são os próximos a serem agrupados



Métodos hierárquicos aglomerativos



Solução
intermediária com
3 clusters



Atualiza matriz de distâncias

Matriz de distâncias entre os clusters
(matriz simétrica)

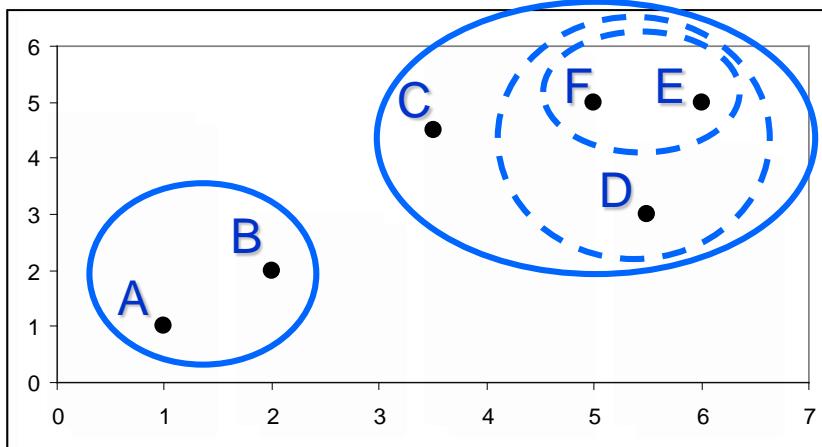
	A-B	C	E-F-D
A-B	0		
C	8,67	0	
E-F-D	28,83	3,02	0

Qual o par de objetos mais próximos?

É o par (E,F,D),C logo estes objetos
são os próximos a serem agrupados



Métodos hierárquicos aglomerativos



Solução
intermediária com
2 clusters

↓ Atualiza matriz de distâncias

Matriz de distâncias entre os clusters
(matriz simétrica)

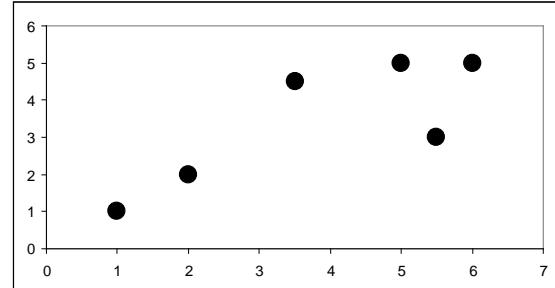
	A-B	E-F-D-C
A-B	0	
E-F-D-C		0

O próximo passo seria agrupar todos os objetos em um único cluster (solução trivial).

27,35 é a distância entre os dois clusters

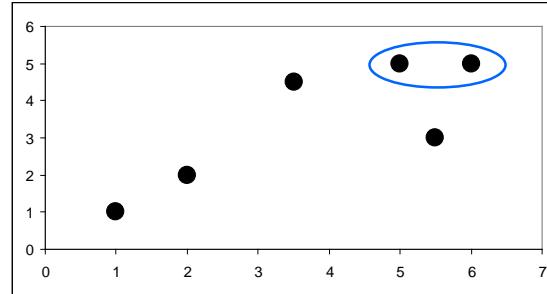
Métodos hierárquicos aglomerativos

6 objetos = 6 clusters



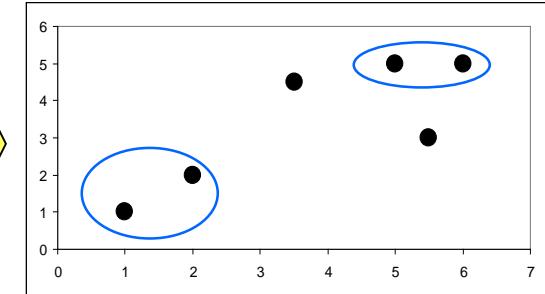
Matriz de distâncias 6x6
Menor distância = 0,5

5 clusters



Matriz de distâncias 5x5
Menor distância = 1

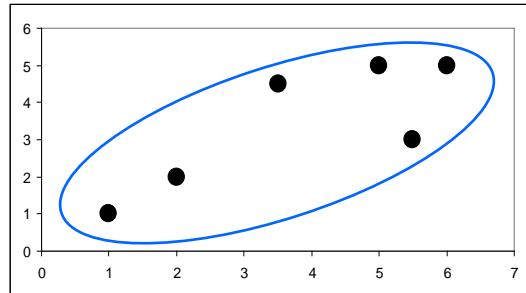
4 clusters



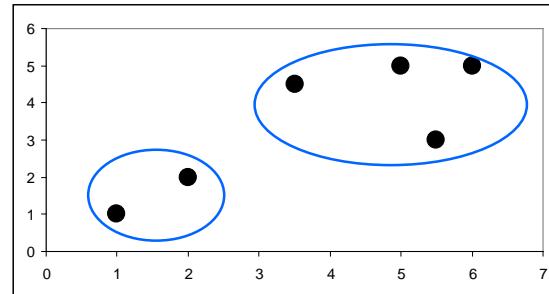
Matriz de distâncias 4x4
Menor distância = 2,667

A cada iteração o número de *clusters* diminui de uma unidade e os novos agrupamentos tornam-se mais heterogêneos internamente.

1 cluster

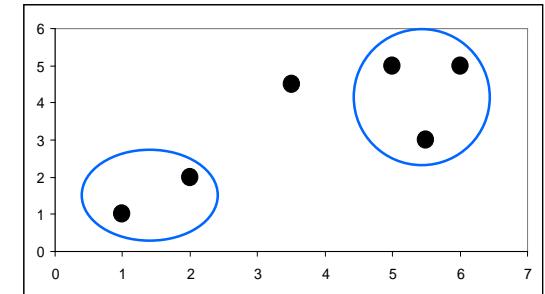


2 clusters

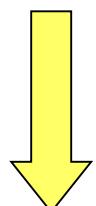


Matriz de distâncias 2x2
Menor distância = 27,35

3 clusters



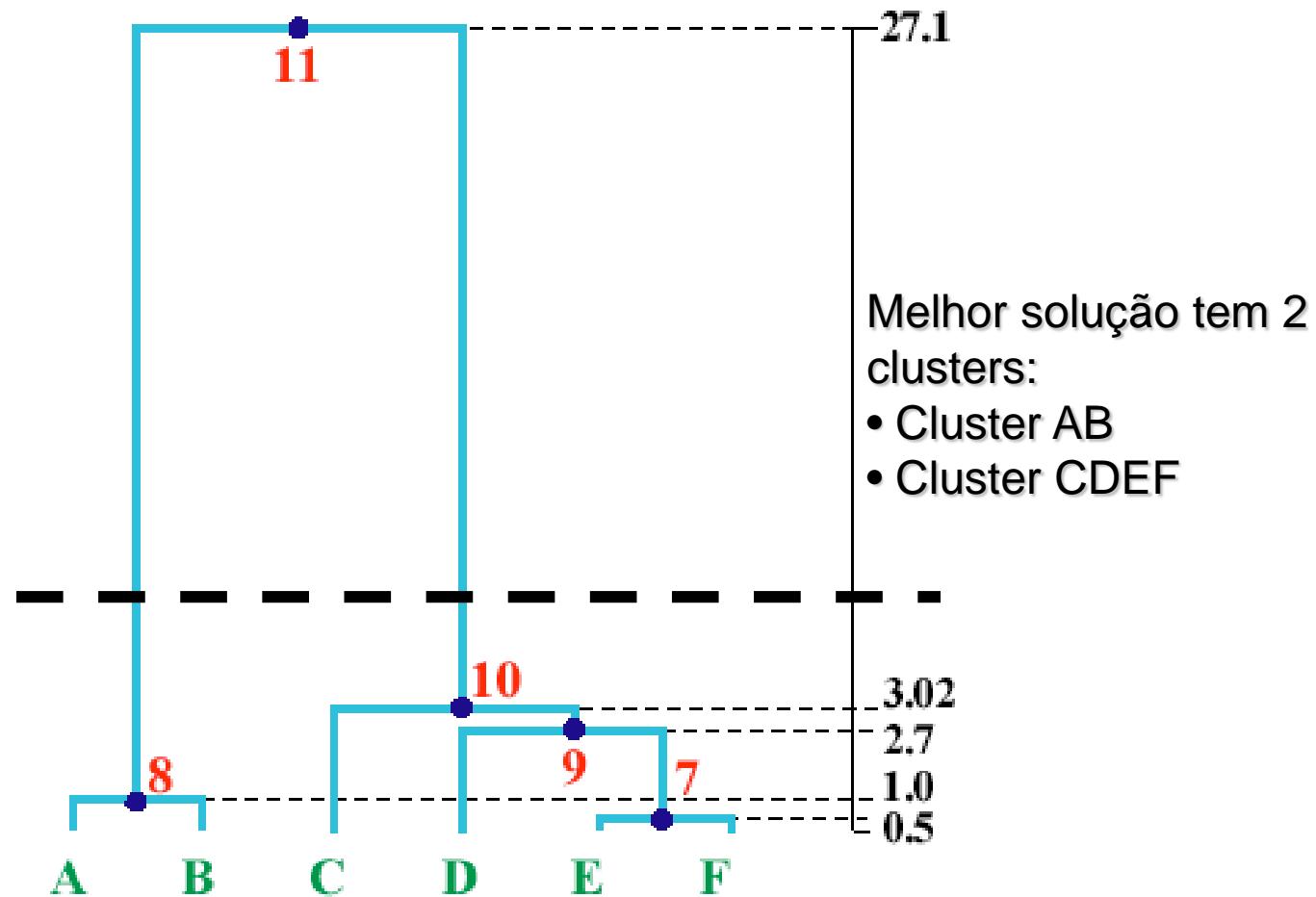
Matriz de distâncias 3x3
Menor distância = 3,02





Métodos hierárquicos aglomerativos

A sequência das agregações e as distâncias em que elas ocorrem são descritas no dendrograma, um gráfico útil na definição do número de agrupamentos.





Métodos hierárquicos aglomerativos

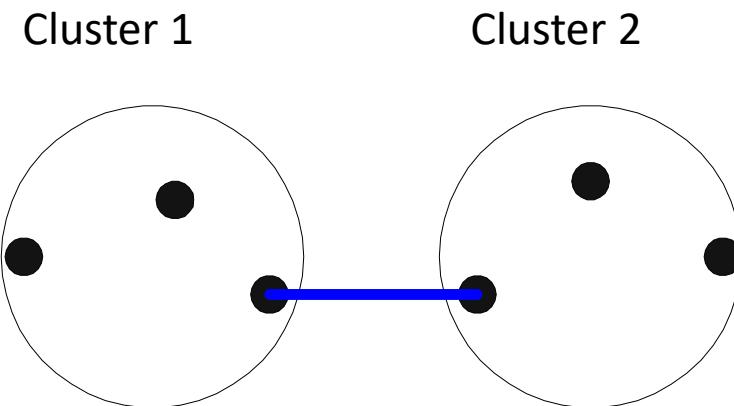
Algoritmo

- 1) Inicie com N *clusters*, cada um contendo apenas um objeto e construa a matriz de distâncias de ordem N .
- 2) Identifique o menor elemento da matriz de distâncias para encontrar o par de *clusters* mais similares.
- 3) Reúna os dois *clusters* identificados na etapa 2 em um único *cluster* e atualiza a matriz de distâncias, retirando as linhas e colunas relativas aos dois *clusters* identificados em 2 e incluindo a linha e coluna com as distâncias entre os demais *clusters* e o novo *cluster* formado. Note que a ordem da matriz de distâncias diminui de uma unidade a cada vez que a etapa 3 é executada
- 4) Repita os passos 2 e 3 até que reste apenas um *cluster*. A cada iteração guarde a identificação dos *clusters* que foram fundidos e também a distância entre eles, estas informações serão utilizadas na montagem do dendrograma.

4.2 Métodos hierárquicos

Encadeamento simples (*single linkage*)

Utiliza a distância mínima entre dois objetos para definir a distância entre dois *clusters*.



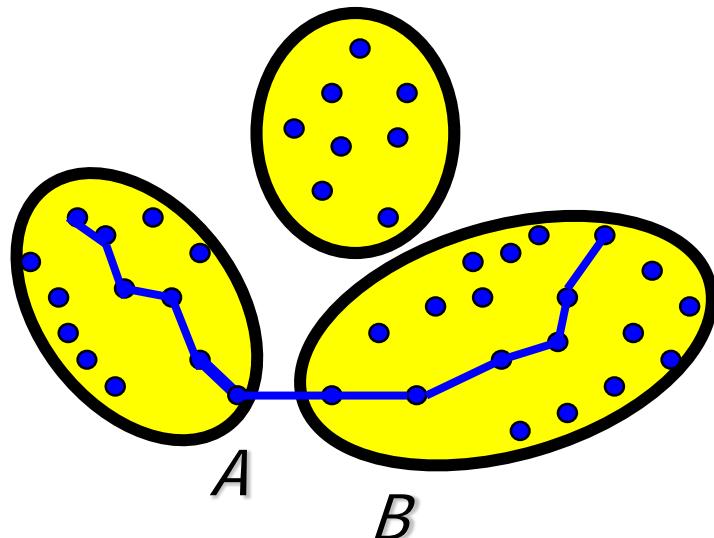
4.2 Métodos hierárquicos

Encadeamento simples (*single linkage*)

Quando os objetos estão pobemente estruturados, o *single linkage* pode reunir em um *cluster* elementos bastante diferentes, desde que haja entre eles uma cadeia de outros elementos que sejam semelhantes entre si (efeito de cadeia).

Exemplo de como a ligação individual pode agregar pontos distintos A e B

Tende a concentrar a maior parte dos objetos em um pequeno número de *clusters* e formar muitos *clusters* com poucos objetos.



Métodos hierárquicos - Encadeamento Simples (Reis, 2001)

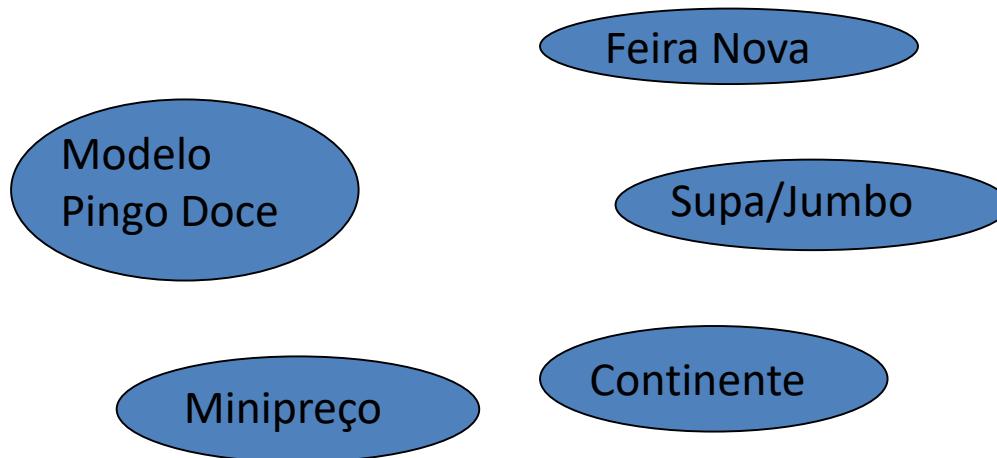
Agrupe as 6 empresas pelo método do encadeamento simples

1ª iteração

Modelo (1)	Pingo Doce (2)	Feira Nova (3)	Supa / Jumbo (4)	Minipreço (5)	Continente (6)	
0						Modelo (1)
0,3	0					Pingo Doce (2)
12,2	9,2	0				Feira Nova (3)
12,7	9,9	0,4	0			Supa / Jumbo (4)
2,3	1,9	15,2	17,1	0		Minipreço (5)
9,8	7,4	4,3	3,4	13,8	0	Continente (6)

Menor distância = $d_{12} = 0,3$ entre Modelo e Pingo Doce

5 clusters



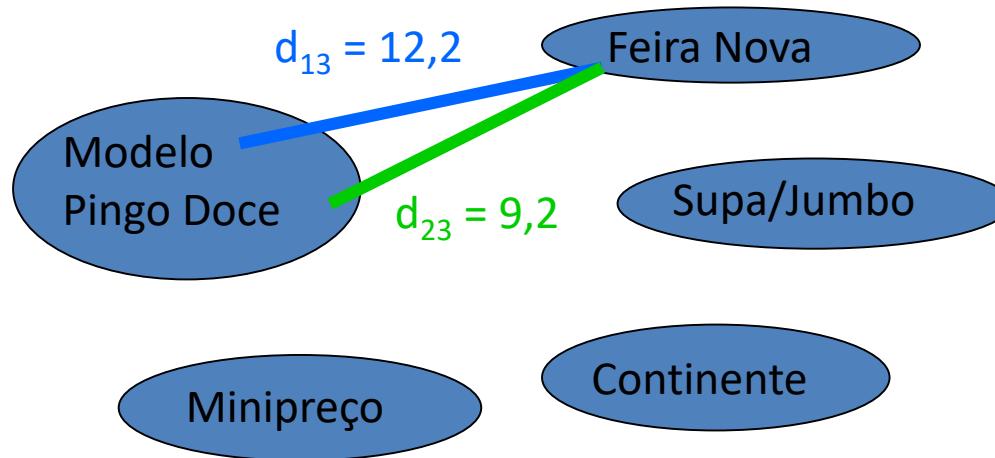
Métodos hierárquicos - Encadeamento Simples (Reis, 2001)

Atualização da matriz de distâncias

Modelo (1)	Pingo Doce (2)	Feira Nova (3)	Supa / Jumbo (4)	Minipreço (5)	Continente (6)
0					
0,3	0				
12,2	9,2	0			
12,7	9,9	0,4	0		
2,3	1,9	15,2	17,1	0	
9,8	7,4	4,3	3,4	13,8	0

D =

Modelo (1)
Pingo Doce (2)
Feira Nova (3)
Supa / Jumbo (4)
Minipreço (5)
Continente (6)



Qual a distância entre Feira Nova e o cluster Modelo/Pingo Doce ?

$$d_{12,3} = ?$$

Pelo encadeamento simples = $d_{12,3} = \min(12,2 ; 9,2) = 9,2$

Métodos hierárquicos - Encadeamento Simples (Reis, 2001)

Atualização da matriz de distâncias

D =

Modelo (1)	Pingo Doce (2)	Feira Nova (3)	Supa / Jumbo (4)	Minipreço (5)	Continente (6)	
Modelo (1)	0,3	12,2	12,7	2,3	9,8	Modelo (1)
Pingo Doce (2)	0	9,2	9,9	1,9	7,4	Pingo Doce (2)
Feira Nova (3)	12,2	0	0,4	15,2	4,3	Feira Nova (3)
Supa / Jumbo (4)	12,7	9,9	0	17,1	3,4	Supa / Jumbo (4)
Minipreço (5)	2,3	1,9	15,2	0	13,8	Minipreço (5)
Continente (6)	9,8	7,4	4,3	13,8	0	Continente (6)

Atualização

D =

Pingo Doce	Modelo (1,2)	Feira Nova (3)	Supa / Jumbo (4)	Minipreço (5)	Continente (6)	
Pingo Doce	0	9,2	9,9	1,9	7,4	Modelo, Pingo Doce (1,2)
Modelo (1,2)	9,2	0	0,4	15,2	4,3	Feira Nova (3)
Feira Nova (3)	9,9	0,4	0	17,1	3,4	Supa / Jumbo (4)
Supa / Jumbo (4)	1,9	15,2	17,1	0	13,8	Minipreço (5)
Minipreço (5)	7,4	4,3	3,4	13,8	0	Continente (6)

$$9,2 = \min(12,2 ; 9,2)$$

$$9,9 = \min(12,7 ; 9,9)$$

$$1,9 = \min(2,3 ; 1,9)$$

$$7,4 = \min(9,8 ; 7,4)$$

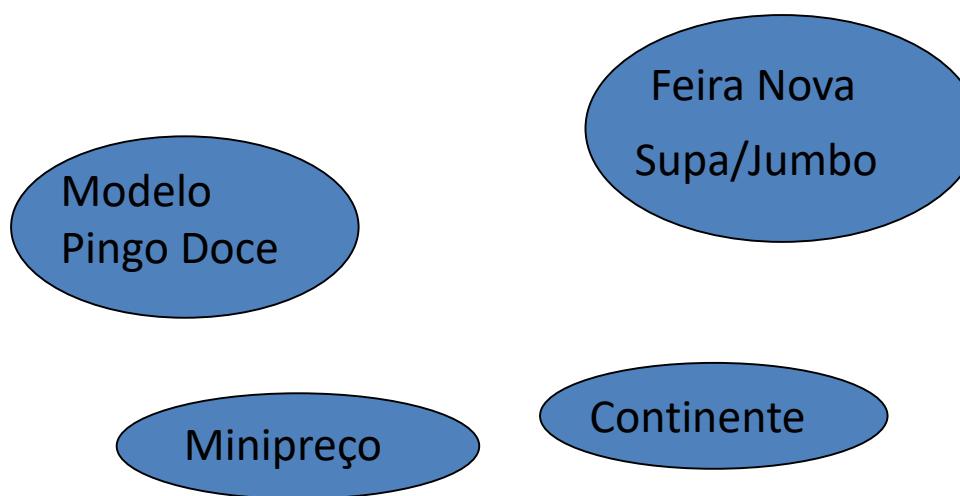
Métodos hierárquicos - Encadeamento Simples (Reis, 2001)

2ª iteração

Pingo Doce		Feira Nova (3)	Supa / Jumbo (4)	Minipreço (5)	Continente (6)	
Modelo (1,2)						Modelo, Pingo Doce (1,2)
	0					Feira Nova (3)
9,2		0				Supa / Jumbo (4)
9,9	0,4		0			Minipreço (5)
1,9	15,2	17,1	0			Continente (6)
7,4	4,3	3,4	13,8	0		

Menor distância = $d_{34} = 0,4$ entre Feira Nova e Supa/Jumbo

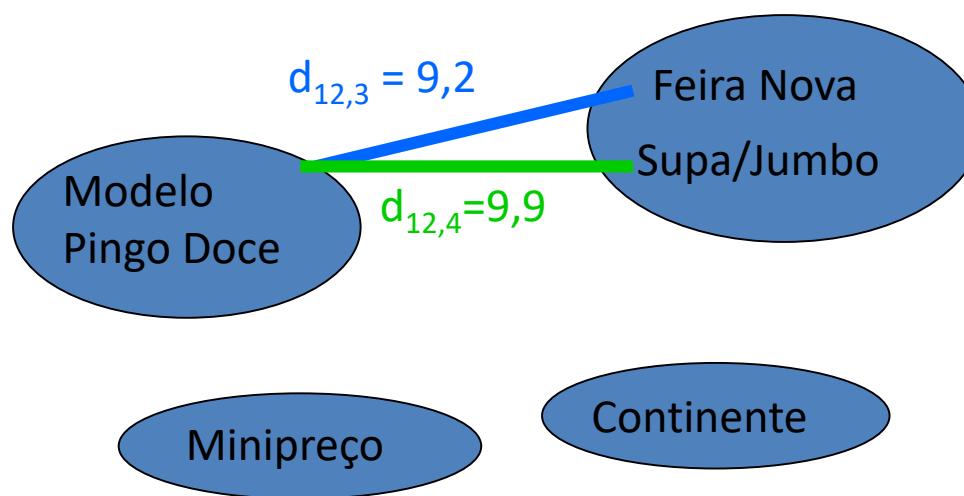
4 clusters



Métodos hierárquicos - Encadeamento Simples (Reis, 2001)

Atualização da matriz de distâncias

Pingo Doce Modelo (1,2)	Feira Nova (3)	Supa / Jumbo (4)	Minipreço (5)	Continente (6)	
0					Modelo, Pingo Doce (1,2)
9,2	0				Feira Nova (3)
9,9	0,4	0			Supa / Jumbo (4)
1,9	15,2	17,1	0		Minipreço (5)
7,4	4,3	3,4	13,8	0	Continente (6)



Qual a distância entre os *clusters* Feira Nova/Supa/Jumbo e Modelo/Pingo Doce ?
 $d_{12,34} = ?$

Pelo encadeamento simples = $d_{12,34} = \min(9,2 ; 9,9) = 9,2$

Métodos hierárquicos - Encadeamento Simples (Reis, 2001)

Atualização da matriz de distâncias



D =

Pingo Doce Modelo (1,2)	Feira Nova (3)	Supa / Jumbo (4)	Minipreço (5)	Continente (6)	
0	9,2	9,9	1,9	7,4	Modelo, Pingo Doce (1,2)
9,2	0	0,4	15,2	4,3	Feira Nova (3)
9,9	0,4	0	17,1	3,4	Supa / Jumbo (4)
1,9	15,2	17,1	0	13,8	Minipreço (5)
7,4	4,3	3,4	13,8	0	Continente (6)

Atualização



D =

Pingo Doce Modelo (1,2)	Supa / Jumbo Feira Nova (3,4)	Minipreço (5)	Continente (6)	
0	9,2	1,9	7,4	Modelo, Pingo Doce (1,2)
9,2	0	15,2	3,4	Feira Nova, Supa/Jumbo (3,4)
1,9	15,2	0	13,8	Minipreço (5)
7,4	3,4	13,8	0	Continente (6)

$$9,2 = \min(9,2 ; 9,9)$$

$$15,2 = \min(15,2 ; 17,1)$$

$$3,4 = \min(4,3 ; 3,4)$$

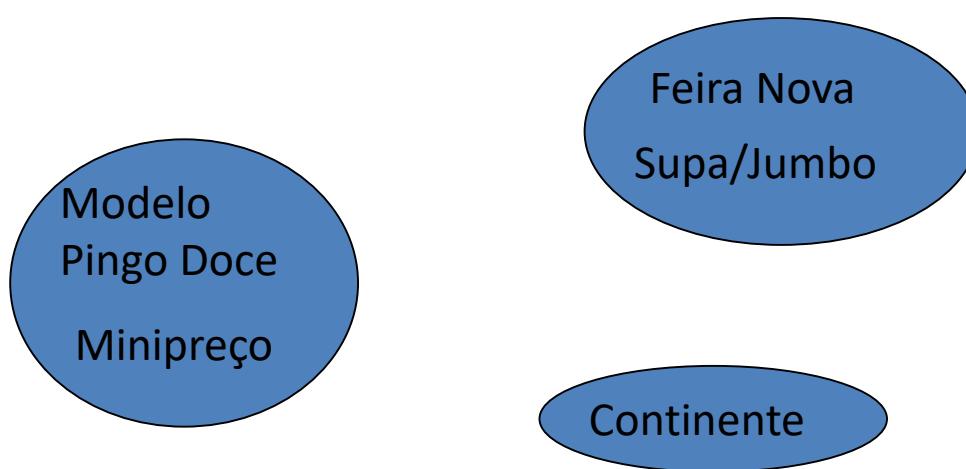
Métodos hierárquicos - Encadeamento Simples (Reis, 2001)

3ª iteração

Pingo Doce Modelo (1,2)	Supa / Jumbo Feira Nova (3,4)	Minipreço (5)	Continente (6)	
0				Modelo, Pingo Doce (1,2)
9,2	0			Feira Nova, Supa/Jumbo (3,4)
1,9	15,2	0		Minipreço (5)
7,4	3,4	13,8	0	Continente (6)

Menor distância = $d_{12,5} = 1,9$ entre Minipreço e o cluster Modelo/Pingo Doce

3 clusters



Métodos hierárquicos - Encadeamento Simples (Reis, 2001)

Atualização da matriz de distâncias



Pingo Doce Modelo (1,2)	Supa / Jumbo Feira Nova (3,4)	Minipreço (5)	Continente (6)	
0	9,2	1,9	7,4	Modelo, Pingo Doce (1,2)
9,2	0	15,2	3,4	Feira Nova, Supa/Jumbo (3,4)
1,9	15,2	0	13,8	Minipreço (5)
7,4	3,4	13,8	0	Continente (6)

Atualização

Minipreço Pingo Doce Modelo (1,2,5)	Supa / Jumbo Feira Nova (3,4)	Continente (6)	
0	9,2	7,4	Modelo, Pingo Doce, Minipreço (1,2,5)
9,2	0	3,4	Feira Nova, Supa/Jumbo (3,4)
7,4	3,4	0	Continente (6)

$$9,2 = \min(9,2 ; 15,2)$$

$$7,4 = \min(7,4 ; 13,8)$$

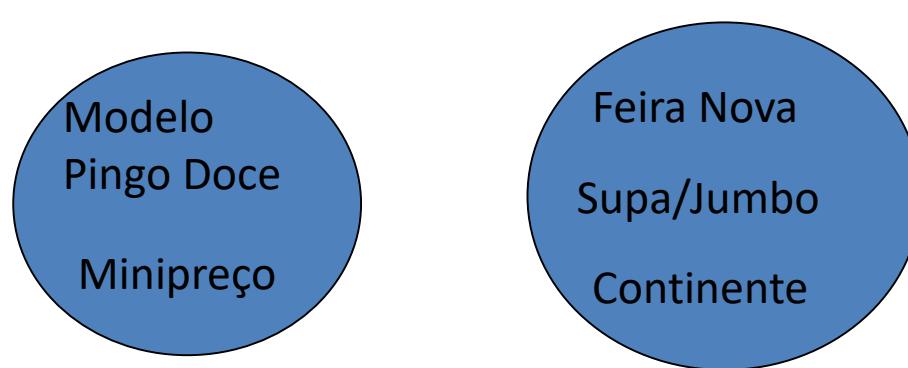
Métodos hierárquicos - Encadeamento Simples (Reis, 2001)

4^a iteração

	Minipreço		
Pingo Doce		Supa / Jumbo	
Modelo		Feira Nova	
(1,2,5)		(3,4)	(6)
D =	0		
	9,2	0	Modelo, Pingo Doce, Minipreço (1,2,5) Feira Nova, Supa/Jumbo (3,4)
	7,4	3,4	Continente (6)

Menor distância = $d_{34,6} = 3,4$ entre Continente e o cluster Feira Nova/Supa/Jumbo

2 clusters



Métodos hierárquicos - Encadeamento Simples (Reis, 2001)

Atualização da matriz de distâncias



	Minipreço Pingo Doce Modelo (1,2,5)	Supa / Jumbo Feira Nova (3,4)	Continente (6)	
Minipreço Pingo Doce Modelo (1,2,5)	0	9,2	7,4	Modelo, Pingo Doce, Minipreço (1,2,5)
Supa / Jumbo Feira Nova (3,4)	9,2	0	3,4	Feira Nova, Supa/Jumbo (3,4)
Continente (6)	7,4	3,4	0	Continente (6)

Atualização

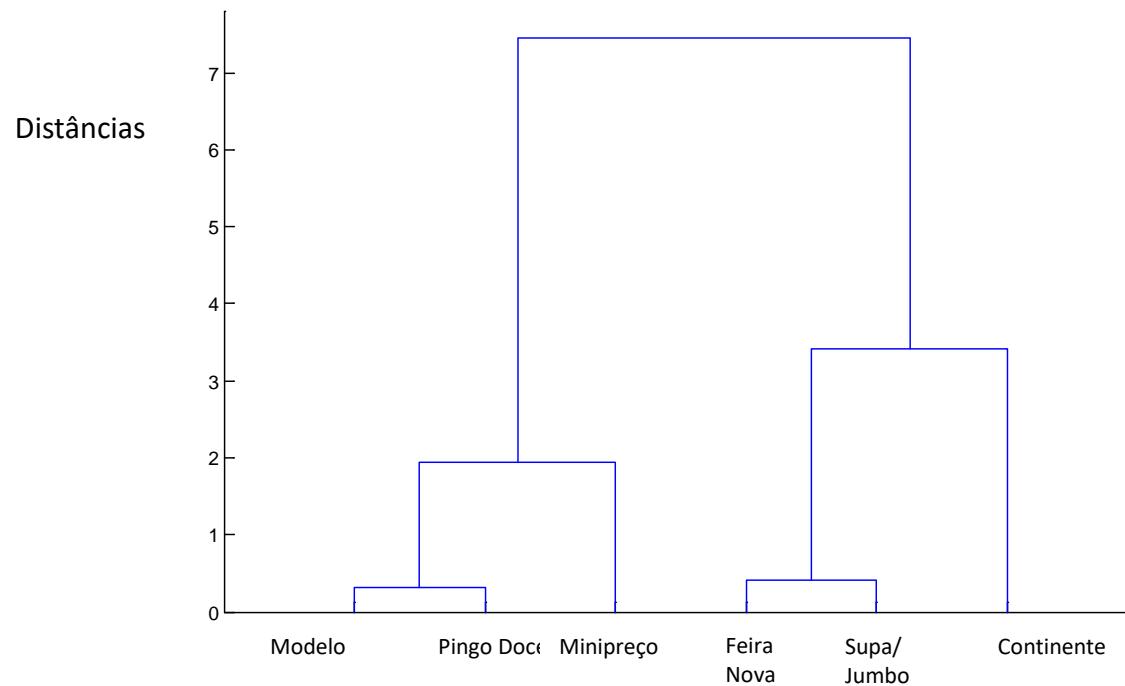
	Minipreço Pingo Doce Modelo (1,2,5)	Continente Supa / Jumbo Feira Nova (3,4,6)	
Minipreço Pingo Doce Modelo (1,2,5)	0	7,4	Modelo, Pingo Doce, Minipreço (1,2,5)
Supa / Jumbo Feira Nova (3,4,6)	7,4	0	Feira Nova, Supa/Jumbo, Continente (3,4,6)
Continente			

$$7,4 = \min(7,4 ; 9,2)$$

Métodos hierárquicos - Encadeamento Simples (Reis, 2001)

Seqüência de agrupamento e dendrograma

Passo	Distâncias	Nº de clusters	Clusters
1	$d_{1,2} = 0,3$	5	12 / 3 / 4 / 5 / 6
2	$d_{3,4} = 0,4$	4	12 / 34 / 5 / 6
3	$d_{12,,5} = 1,9$	3	125 / 34 / 6
4	$d_{34,,6} = 3,4$	2	125 / 346
5	$d_{125,,346} = 7,4$	1	123456



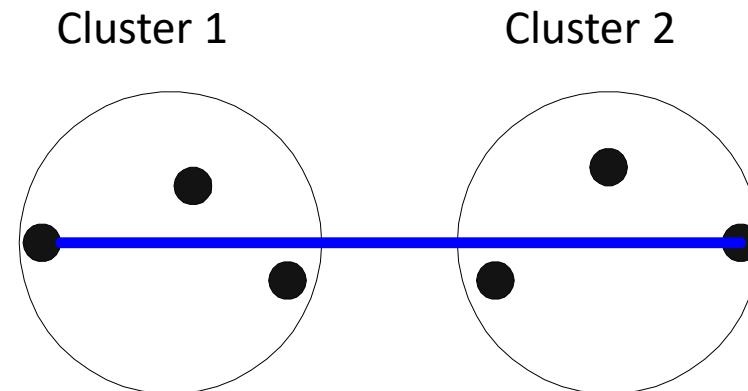
Métodos hierárquicos - Encadeamento Completo (Reis, 2001)

Encadeamento completo (complete linkage)

A distância entre dois *clusters* é definida pela maior distância entre dois objetos, um em cada *cluster*.

Elimina a possibilidade do efeito de encadeamento.

É fortemente afetado por *outliers*.



Métodos hierárquicos - Encadeamento Completo (Reis, 2001)

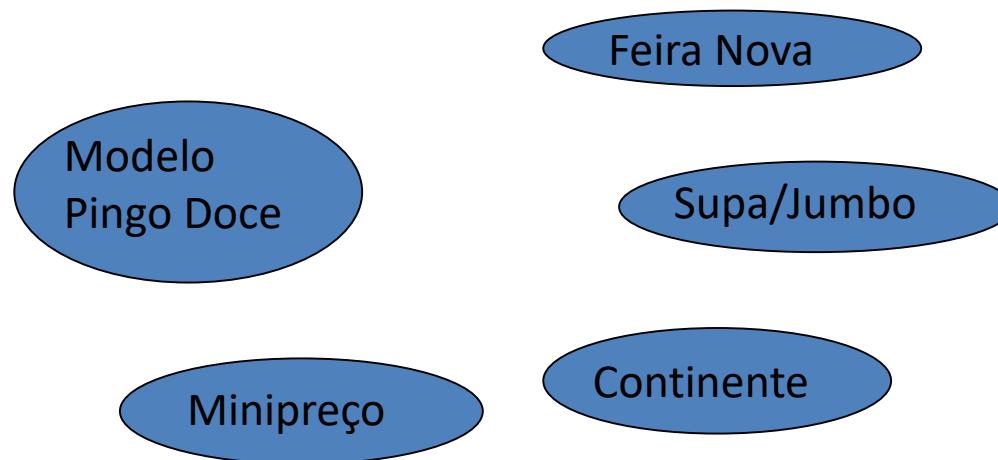
Exemplo: Agrupar as 6 empresas pelo método do encadeamento simples

1^a iteração

Modelo (1)	Pingo Doce (2)	Feira Nova (3)	Supa / Jumbo (4)	Minipreço (5)	Continente (6)	
0						Modelo (1)
0,3	0					Pingo Doce (2)
12,2	9,2	0				Feira Nova (3)
12,7	9,9	0,4	0			Supa / Jumbo (4)
2,3	1,9	15,2	17,1	0		Minipreço (5)
9,8	7,4	4,3	3,4	13,8	0	Continente (6)

Menor distância = $d_{12} = 0,3$ entre Modelo e Pingo Doce

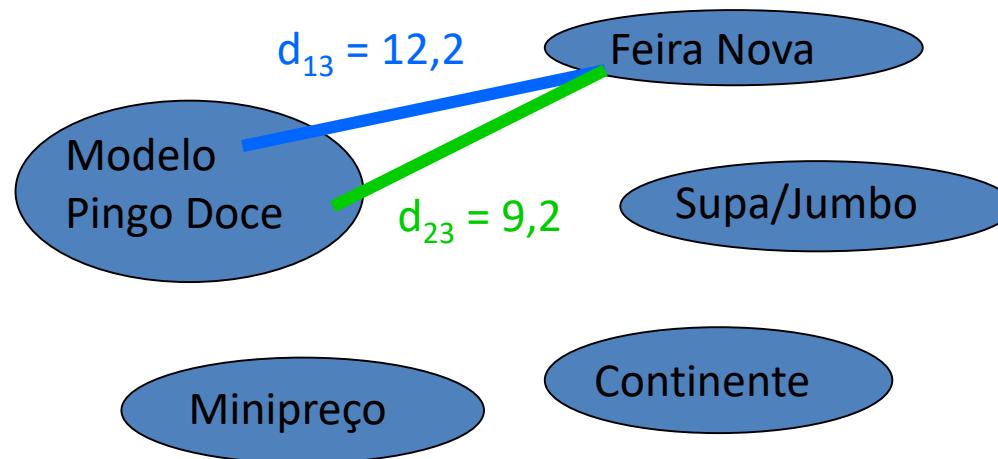
5 clusters



Métodos hierárquicos - Encadeamento Completo (Reis, 2001)

Atualização da matriz de distâncias

Modelo (1)	Pingo Doce (2)	Feira Nova (3)	Supa / Jumbo (4)	Minipreço (5)	Continente (6)	
0	0,3	12,2	9,2	0		Modelo (1)
0,3	0	12,7	9,9	0,4	0	Pingo Doce (2)
12,2	9,2	0	0,4	0	Feira Nova (3)	
9,2	0	0,4	0	0	Supa / Jumbo (4)	
0	12,7	0,4	0	0	Minipreço (5)	
	9,9	0	17,1	0	Continente (6)	
	0,4	15,2	17,1	0		Continente (6)
	0	15,2	17,1	0		Continente (6)
12,2	9,2	0	0	0		Continente (6)
9,2	0	0	0	0		Continente (6)
0	0	0	0	0		Continente (6)
0	0	0	0	0		Continente (6)



Qual a distância entre Feira Nova e o cluster Modelo/Pingo Doce ?

$$d_{12,3} = ?$$

Pelo encadeamento completo = $d_{12,3} = \max(12,2 ; 9,2) = 12,2$

Métodos hierárquicos - Encadeamento Completo (Reis, 2001)

Atualização da matriz de distâncias



	Modelo (1)	Pingo Doce (2)	Feira Nova (3)	Supa / Jumbo (4)	Minipreço (5)	Continente (6)	
Modelo (1)	0	0,3	12,2	12,7	2,3	9,8	Modelo (1)
Pingo Doce (2)	0,3	0	9,2	9,9	1,9	7,4	Pingo Doce (2)
Feira Nova (3)	12,2	9,2	0	0,4	15,2	4,3	Feira Nova (3)
Supa / Jumbo (4)	12,7	9,9	0,4	0	17,1	3,4	Supa / Jumbo (4)
Minipreço (5)	2,3	1,9	15,2	17,1	0	13,8	Minipreço (5)
Continente (6)	9,8	7,4	4,3	3,4	13,8	0	Continente (6)

Atualização

	Pingo Doce					
	Modelo (1,2)	Feira Nova (3)	Supa / Jumbo (4)	Minipreço (5)	Continente (6)	
Modelo (1,2)	0	12,2	12,7	2,3	9,8	Modelo, Pingo Doce (1,2)
Feira Nova (3)	12,2	0	0,4	15,2	4,3	Feira Nova (3)
Supa / Jumbo (4)	12,7	0,4	0	17,1	3,4	Supa / Jumbo (4)
Minipreço (5)	2,3	15,2	17,1	0	13,8	Minipreço (5)
Continente (6)	9,8	4,3	3,4	13,8	0	Continente (6)

$$12,2 = \max(12,2 ; 9,2)$$

$$12,7 = \max(12,7 ; 9,9)$$

$$2,3 = \max(2,3 ; 1,9)$$

$$9,8 = \max(9,8 ; 7,4)$$

Métodos hierárquicos - Encadeamento Completo (Reis, 2001)

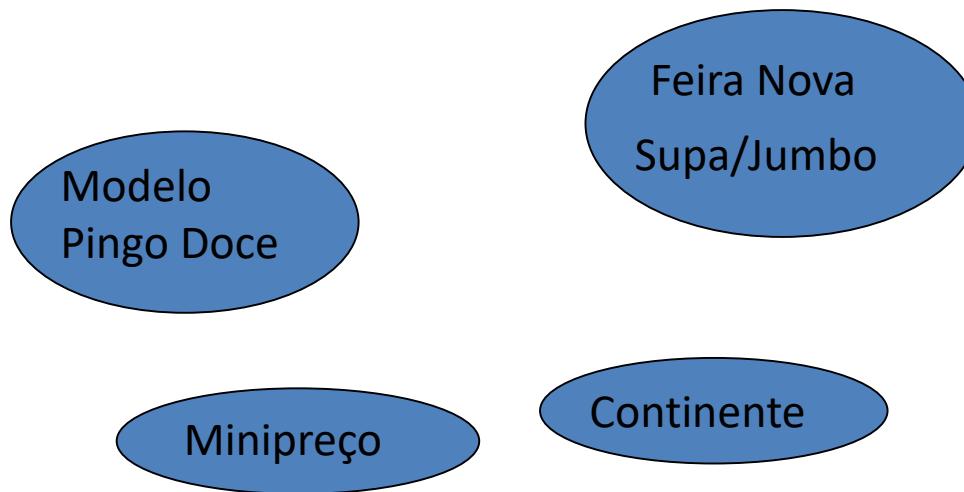
2ª iteração

Pingo Doce				
Modelo (1,2)		Feira Nova (3)	Supa / Jumbo (4)	Minipreço (5) Continente (6)
D =	0			
	12,2	0		
	12,7	0,4	0	
	2,3	15,2	17,1	0
	9,8	4,3	3,4	13,8 0

Modelo, Pingo Doce (1,2)
Feira Nova (3)
Supa / Jumbo (4)
Minipreço (5)
Continente (6)

Menor distância = $d_{34} = 0,4$ entre Feira Nova e Supa/Jumbo

4 clusters



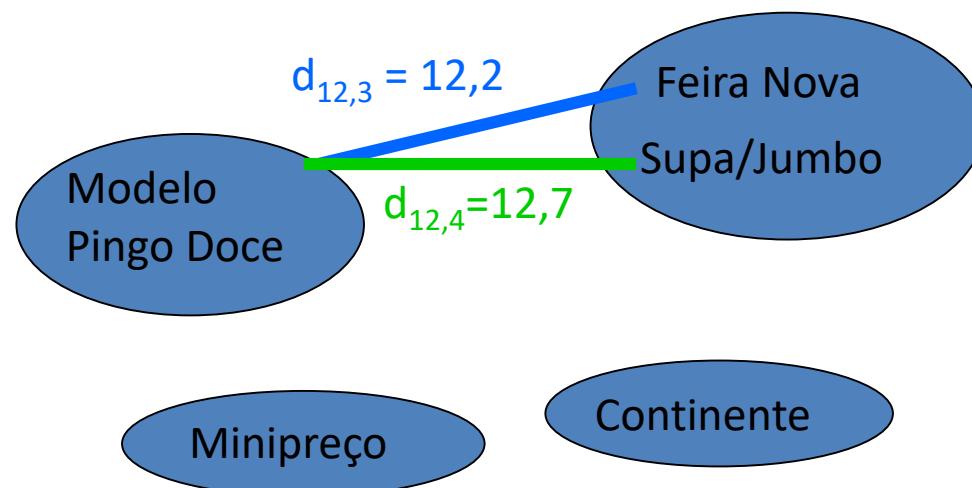
Métodos hierárquicos - Encadeamento Completo (Reis, 2001)

Atualização da matriz de distâncias

Pingo Doce Modelo (1,2)	Feira Nova (3)	Supa / Jumbo (4)	Minipreço (5)	Continente (6)
0				
12,2	0			
12,7	0,4	0		
2,3	15,2	17,1	0	
9,8	4,3	3,4	13,8	0

D =

Modelo, Pingo Doce (1,2)
Feira Nova (3)
Supa / Jumbo (4)
Minipreço (5)
Continente (6)



Qual a distância entre os *clusters* Feira Nova/Supa/Jumbo e Modelo/Pingo Doce ?
 $d_{12,34} = ?$

Pelo encadeamento completo = $d_{12,34} = \max(12,2 ; 12,7) = 12,7$

Métodos hierárquicos - Encadeamento Completo (Reis, 2001)

Atualização da matriz de distâncias

D =

Pingo Doce Modelo (1,2)	Feira Nova (3)	Supa / Jumbo (4)	Minipreço (5)	Continente (6)	
0	12,2	12,7	2,3	9,8	Modelo, Pingo Doce (1,2)
12,2	0	0,4	15,2	4,3	Feira Nova (3)
12,7	0,4	0	17,1	3,4	Supa / Jumbo (4)
2,3	15,2	17,1	0	13,8	Minipreço (5)
9,8	4,3	3,4	13,8	0	Continente (6)

Atualização

D =

Pingo Doce Modelo (1,2)	Supa / Jumbo Feira Nova (3,4)	Minipreço (5)	Continente (6)	
0	12,7	2,3	9,8	Modelo, Pingo Doce (1,2)
12,7	0	17,1	4,3	Feira Nova, Supa/Jumbo (3,4)
2,3	17,1	0	13,8	Minipreço (5)
9,8	4,3	13,8	0	Continente (6)

$$12,7 = \max(12,2 ; 12,7)$$

$$17,1 = \max(15,2 ; 17,1)$$

$$4,3 = \max(4,3 ; 3,4)$$

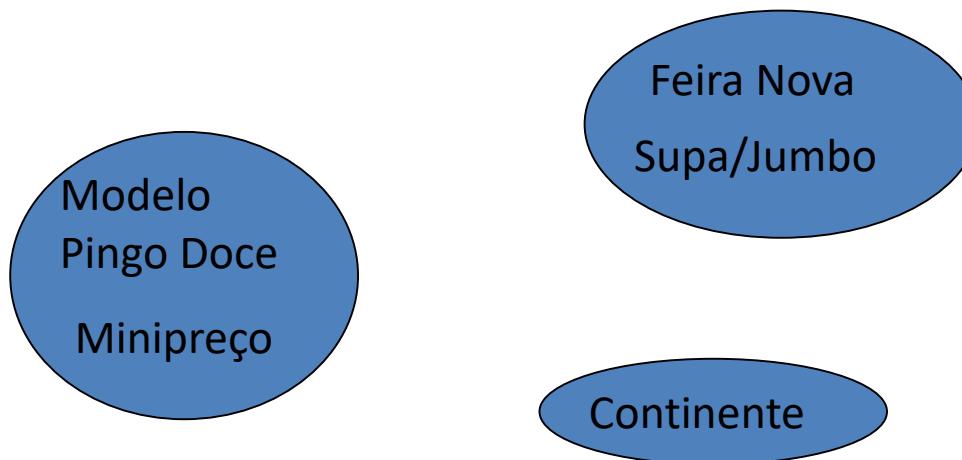
Métodos hierárquicos - Encadeamento Completo (Reis, 2001)

3ª iteração

Pingo Doce Modelo (1,2)	Supa / Jumbo Feira Nova (3,4)	Minipreço (5)	Continente (6)	
0				Modelo, Pingo Doce (1,2)
12,7	0			Feira Nova, Supa/Jumbo (3,4)
2,3	17,1	0		Minipreço (5)
9,8	4,3	13,8	0	Continente (6)

Menor distância = $d_{12,5} = 2,3$ entre Minipreço e o cluster Modelo/Pingo Doce

3 clusters



Métodos hierárquicos - Encadeamento Completo (Reis, 2001)

Atualização da matriz de distâncias



D =

Pingo Doce Modelo (1,2)	Supa / Jumbo Feira Nova (3,4)	Minipreço (5)	Continente (6)	
0	12,7	2,3	9,8	Modelo, Pingo Doce (1,2)
12,7	0	17,1	4,3	Feira Nova, Supa/Jumbo (3,4)
2,3	17,1	0	13,8	Minipreço (5)
9,8	4,3	13,8	0	Continente (6)

Atualização

D =

Minipreço Pingo Doce Modelo (1,2,5)	Supa / Jumbo Feira Nova (3,4)	Continente (6)	
0	17,1	13,8	Modelo, Pingo Doce, Minipreço (1,2,5)
17,1	0	4,3	Feira Nova, Supa/Jumbo (3,4)
13,8	4,3	0	Continente (6)

$$17,1 = \max(12,7 ; 17,1)$$
$$13,8 = \max(9,8 ; 13,8)$$

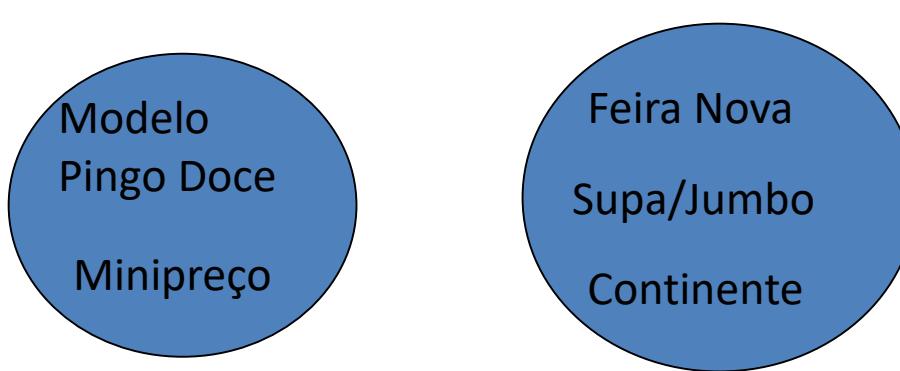
Métodos hierárquicos - Encadeamento Completo (Reis, 2001)

4^a iteração

	Minipreço			
Pingo Doce		Supa / Jumbo		
Modelo		Feira Nova		
(1,2,5)		(3,4)	(6)	
D =	0	17,1	13,8	Modelo, Pingo Doce, Minipreço (1,2,5)
	17,1	0	4,3	Feira Nova, Supa/Jumbo (3,4)
	13,8	4,3	0	Continente (6)

Menor distância = $d_{34,6} = 3,4$ entre Continente e o cluster Feira Nova/Supa/Jumbo

2 clusters



Métodos hierárquicos - Encadeamento Completo (Reis, 2001)

Atualização da matriz de distâncias



		Minipreço	Supa / Jumbo	Continente	
		Pingo Doce	Feira Nova	(6)	
Ping		Modelo (1,2,5)	(3,4)		
D =		0	17,1	13,8	Modelo, Pingo Doce, Minipreço (1,2,5)
		17,1	0	4,3	Feira Nova, Supa/Jumbo (3,4)
		13,8	4,3	0	Continente (6)

Atualização

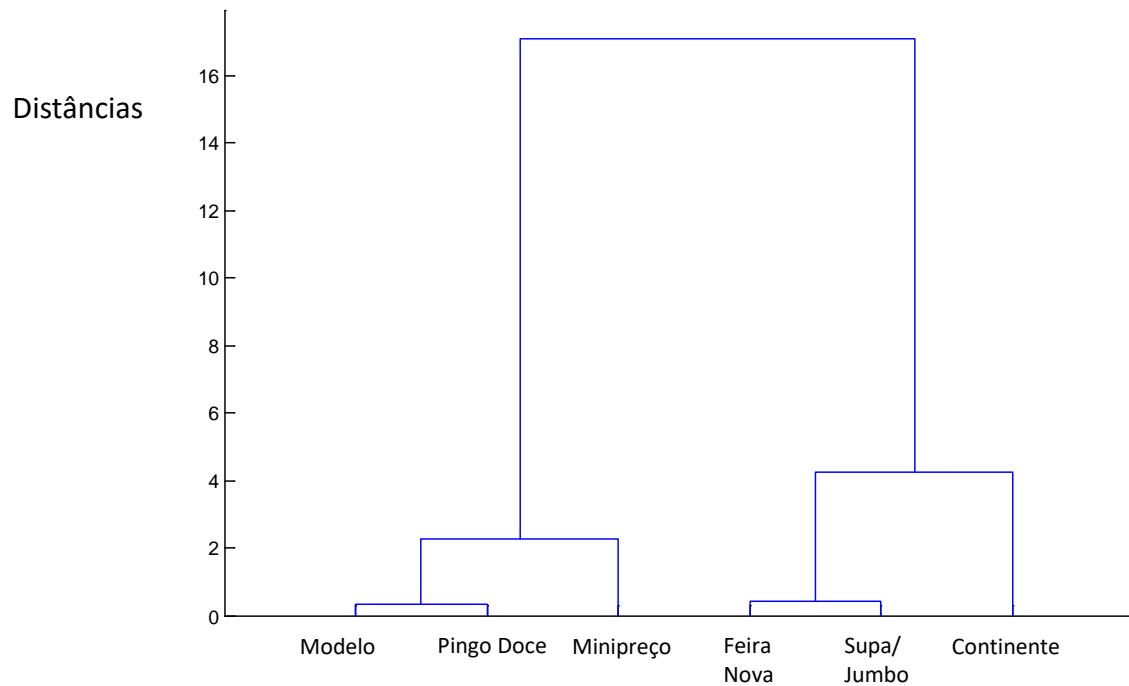
		Minipreço	Continente	
		Pingo Doce	Supa / Jumbo	
Ping		Modelo (1,2,5)	Feira Nova (3,4,6)	
D =		0	17,1	Modelo, Pingo Doce, Minipreço (1,2,5)
		17,1	0	Feira Nova, Supa/Jumbo, Continente (3,4,6)

$$17,1 = \max(17,1 ; 13,8)$$

Métodos hierárquicos - Encadeamento Completo (Reis, 2001)

Seqüência de agrupamento e dendrograma

Passo	Distâncias	Nº de clusters	Clusters
1	$d_{1,2} = 0,3$	5	12 / 3 / 4 / 5 / 6
2	$d_{3,4} = 0,4$	4	12 / 34 / 5 / 6
3	$d_{12,,5} = 2,3$	3	125 / 34 / 6
4	$d_{34,,6} = 4,3$	2	125 / 346
5	$d_{125,,346} = 17,1$	1	123456



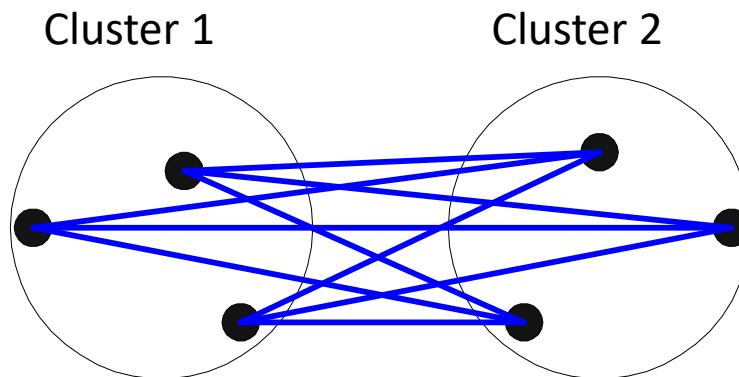
Métodos hierárquicos - Encadeamento Médio (Reis, 2001)

Encadeamento médio (average linkage)

A distância entre dois *clusters* é definida pela média das distâncias entre os elementos dos dois grupos.

Tende a enviesar em direção à produção de clusters com aproximadamente a mesma variância.

É menos afetado por *outliers*.



Métodos hierárquicos - Encadeamento Médio (Reis, 2001)

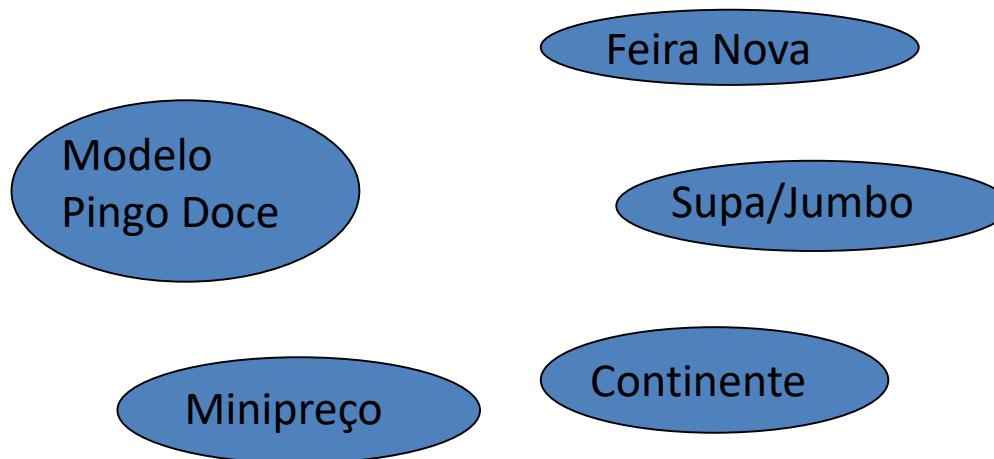
Exemplo: Agrupar as 6 empresas pelo método do encadeamento simples

1^a iteração

Modelo (1)	Pingo Doce (2)	Feira Nova (3)	Supa / Jumbo (4)	Minipreço (5)	Continente (6)	
0						Modelo (1)
0,3	0					Pingo Doce (2)
12,2	9,2	0				Feira Nova (3)
12,7	9,9	0,4	0			Supa / Jumbo (4)
2,3	1,9	15,2	17,1	0		Minipreço (5)
9,8	7,4	4,3	3,4	13,8	0	Continente (6)

Menor distância = $d_{12} = 0,3$ entre Modelo e Pingo Doce

5 clusters



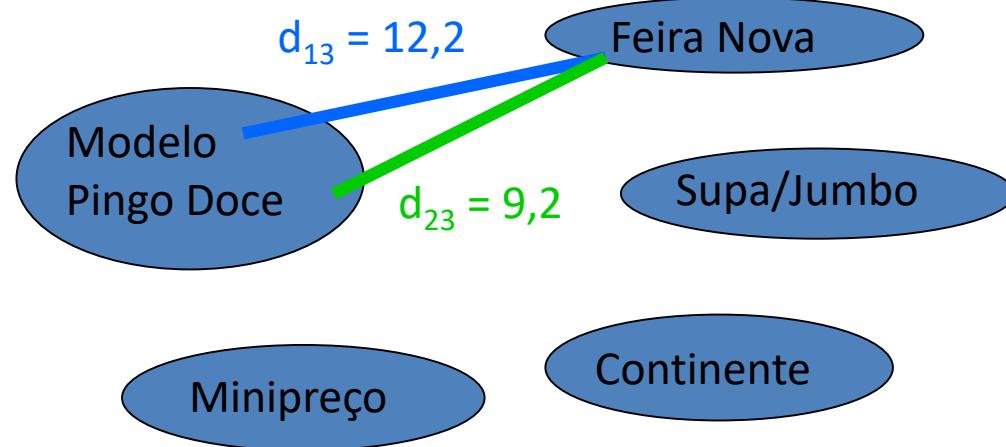
Métodos hierárquicos - Encadeamento Médio (Reis, 2001)

Atualização da matriz de distâncias

Modelo (1)	Pingo Doce (2)	Feira Nova (3)	Supa / Jumbo (4)	Minipreço (5)	Continente (6)
0					
0,3	0				
12,2	9,2	0			
12,7	9,9	0,4	0		
2,3	1,9	15,2	17,1	0	
9,8	7,4	4,3	3,4	13,8	0

D =

Modelo (1)
Pinga Doce (2)
Feira Nova (3)
Supa / Jumbo (4)
Minipreço (5)
Continente (6)



Qual a distância entre Feira Nova e o cluster Modelo/Pingo Doce ?

$$d_{12,3} = ?$$

Pelo encadeamento médio = $d_{12,3} = \text{média}(12,2 ; 9,2) = 10,7$

Métodos hierárquicos - Encadeamento Médio (Reis, 2001)

Atualização da matriz de distâncias



	Modelo (1)	Pingo Doce (2)	Feira Nova (3)	Supa / Jumbo (4)	Minipreço (5)	Continente (6)	
Modelo (1)	0	0,3	12,2	12,7	2,3	9,8	Modelo (1)
Pingo Doce (2)	0,3	0	9,2	9,9	1,9	7,4	Pingo Doce (2)
Feira Nova (3)	12,2	9,2	0	0,4	15,2	4,3	Feira Nova (3)
Supa / Jumbo (4)	12,7	9,9	0,4	0	17,1	3,4	Supa / Jumbo (4)
Minipreço (5)	2,3	1,9	15,2	17,1	0	13,8	Minipreço (5)
Continente (6)	9,8	7,4	4,3	3,4	13,8	0	Continente (6)

Atualização

	Pingo Doce					
	Modelo (1,2)	Feira Nova (3)	Supa / Jumbo (4)	Minipreço (5)	Continente (6)	
Modelo (1,2)	0	10,7	11,3	2,1	8,6	Modelo, Pingo Doce (1,2)
Feira Nova (3)	10,7	0	0,4	15,2	4,3	Feira Nova (3)
Supa / Jumbo (4)	11,3	0,4	0	17,1	3,4	Supa / Jumbo (4)
Minipreço (5)	2,1	15,2	17,1	0	13,8	Minipreço (5)
Continente (6)	8,6	4,3	3,4	13,8	0	Continente (6)

$$10,7 = \text{média}(12,2 ; 9,2)$$

$$11,3 = \text{média}(12,7 ; 9,9)$$

$$2,1 = \text{média}(2,3 ; 1,9)$$

$$8,6 = \text{média}(9,8 ; 7,4)$$

Métodos hierárquicos - Encadeamento Médio (Reis, 2001)

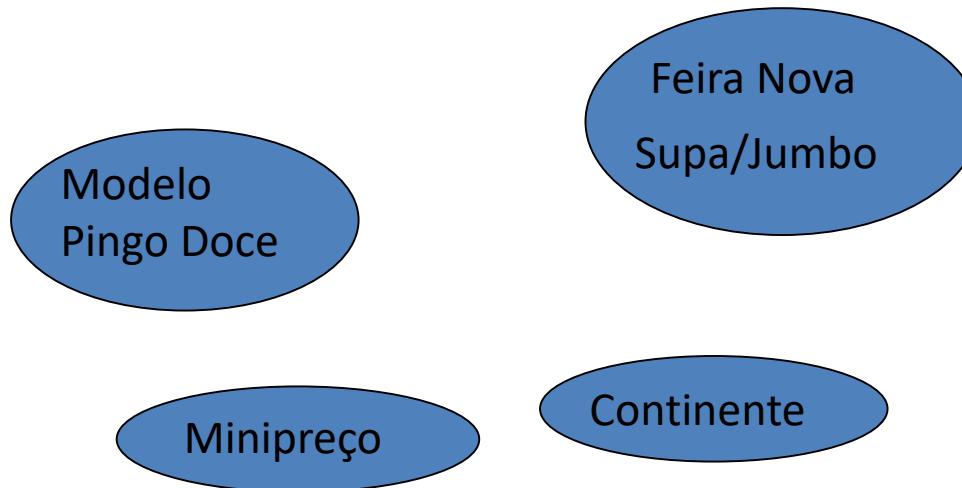
2ª iteração

Pingo Doce				
		Modelo (1,2) Feira Nova (3) Supa / Jumbo (4)	Minipreço (5)	Continente (6)
D =	0			
	10,7	0		
	11,3	0,4	0	
	2,1	15,2	17,1	0
	8,6	4,3	3,4	13,8
				0

Modelo, Pingo Doce (1,2)
Feira Nova (3)
Supa / Jumbo (4)
Minipreço (5)
Continente (6)

Menor distância = $d_{34} = 0,4$ entre Feira Nova e Supa/Jumbo

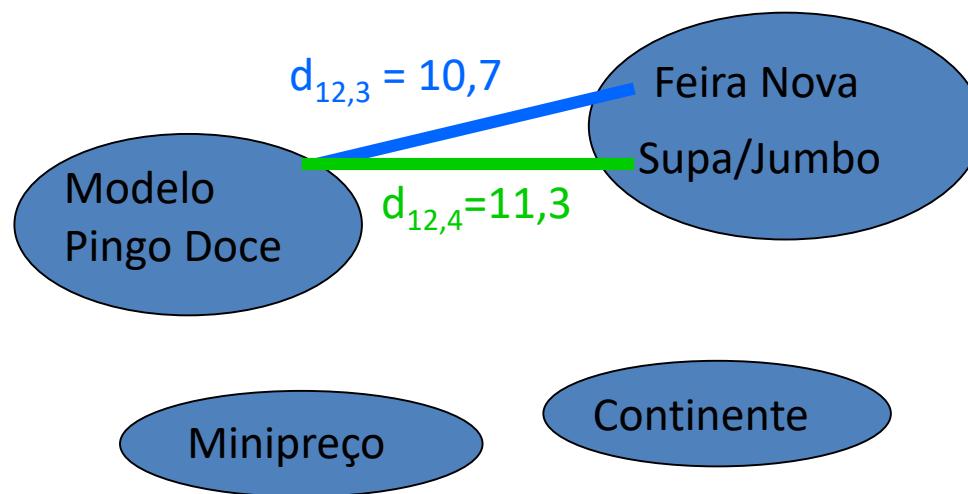
4 clusters



Métodos hierárquicos - Encadeamento Médio (Reis, 2001)

Atualização da matriz de distâncias

Pingo Doce Modelo (1,2)	Feira Nova (3)	Supa / Jumbo (4)	Minipreço (5)	Continente (6)	
0					Modelo, Pingo Doce (1,2)
10,7	0				Feira Nova (3)
11,3	0,4	0			Supa / Jumbo (4)
2,1	15,2	17,1	0		Minipreço (5)
8,6	4,3	3,4	13,8	0	Continente (6)



Qual a distância entre os *clusters* Feira Nova/Supa/Jumbo e Modelo/Pingo Doce ?
 $d_{12,34} = ?$

Pelo encadeamento médio = $d_{12,34} = \text{média}(10,7 ; 11,3) = 11$

Métodos hierárquicos - Encadeamento Médio (Reis, 2001)

Atualização da matriz de distâncias



Pingo Doce							
Modelo (1,2)		Feira Nova (3)	Supa / Jumbo (4)	Minipreço (5)	Continente (6)		
0	10,7	11,3		2,1	8,6	Modelo, Pingo Doce (1,2)	
10,7	0	0,4		15,2	4,3	Feira Nova (3)	
11,3	0,4	0		17,1	3,4	Supa / Jumbo (4)	
2,1	15,2	17,1		0	13,8	Minipreço (5)	
8,6	4,3	3,4		13,8	0	Continente (6)	

Atualização

Pingo Doce		Supa / Jumbo				
Modelo (1,2)		Feira Nova (3,4)	Minipreço (5)	Continente (6)		
0	11	2,1	8,6		Modelo, Pingo Doce (1,2)	
11	0	16,15	3,85		Feira Nova, Supa/Jumbo (3,4)	
2,1	16,15	0	13,8		Minipreço (5)	
8,6	3,85	13,8	0		Continente (6)	

$$11 = \text{média}(10,7 ; 11,3)$$

$$16,15 = \text{média}(15,2 ; 17,1)$$

$$3,85 = \text{média}(4,3 ; 3,4)$$

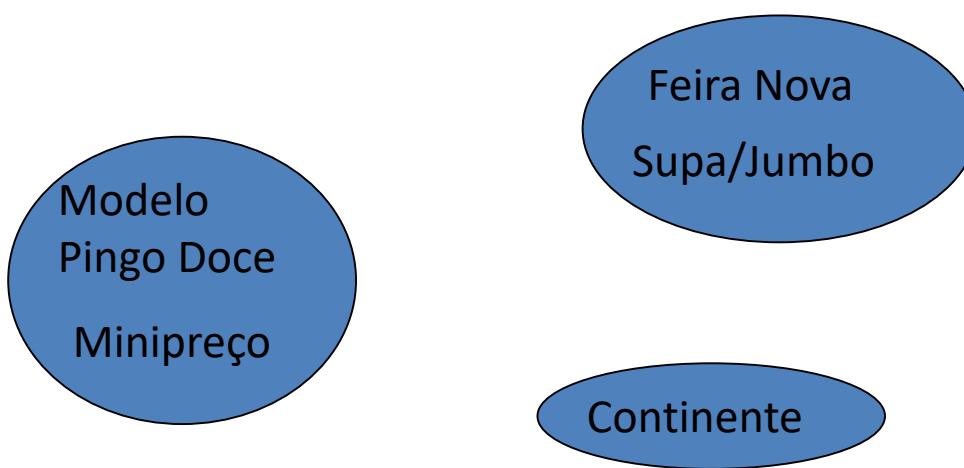
Métodos hierárquicos - Encadeamento Médio (Reis, 2001)

3ª iteração

Pingo Doce Modelo (1,2)	Supa / Jumbo Feira Nova (3,4)	Minipreço (5)	Continente (6)	
0				Modelo, Pingo Doce (1,2)
11	0			Feira Nova, Supa/Jumbo (3,4)
2,1	16,15	0		Minipreço (5)
8,6	3,85	13,8	0	Continente (6)

Menor distância = $d_{12,5} = 2,1$ entre Minipreço e o cluster Modelo/Pingo Doce

3 clusters



Métodos hierárquicos - Encadeamento Médio (Reis, 2001)

Atualização da matriz de distâncias



D =

Pingo Doce Modelo (1,2)	Supa / Jumbo Feira Nova (3,4)	Minipreço (5)	Continente (6)	
0	11	2,1	8,6	Modelo, Pingo Doce (1,2)
11	0	16,15	3,85	Feira Nova, Supa/Jumbo (3,4)
2,1	16,15	0	13,8	Minipreço (5)
8,6	3,85	13,8	0	Continente (6)

Atualização

D =

Minipreço Pingo Doce Modelo (1,2,5)	Supa / Jumbo Feira Nova (3,4)	Continente (6)	
0	12,7	10,3	Modelo, Pingo Doce, Minipreço (1,2,5)
12,7	0	3,85	Feira Nova, Supa/Jumbo (3,4)
10,3	3,85	0	Continente (6)

$$12,7 = \text{média}(11 ; 16,15)$$

$$10,3 = \text{média}(8,6 ; 13,8)$$

Métodos hierárquicos - Encadeamento Médio (Reis, 2001)

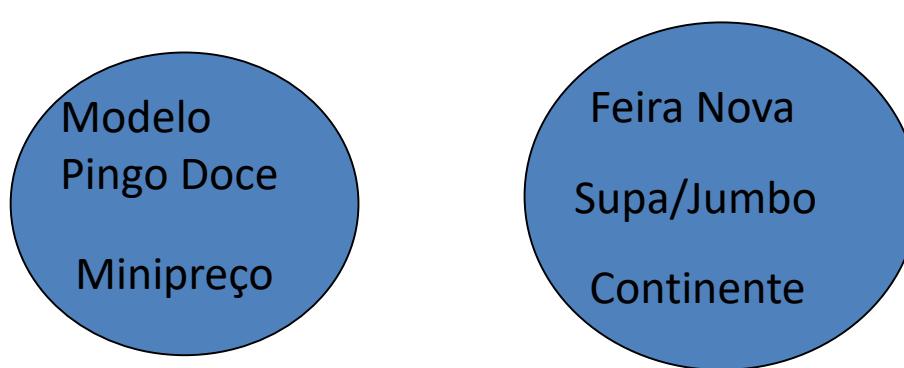
4ª iteração

Minipreço			
Pingo Doce	Supa / Jumbo	Feira Nova	Continente
Modelo			
(1,2,5)		(3,4)	(6)
0			
12,7	0		Modelo, Pingo Doce, Minipreço (1,2,5) Feira Nova, Supa/Jumbo (3,4)
10,3	3,85	0	Continente (6)

D =

Menor distância = $d_{34,6} = 3,85$ entre Continente e o cluster Feira Nova/Supa/Jumbo

2 clusters



Métodos hierárquicos - Encadeamento Médio (Reis, 2001)

Atualização da matriz de distâncias

	Minipreço	Supa / Jumbo	Continente	
Pingo Doce				
Modelo	(1,2,5)	(3,4)	(6)	
0	12,7	10,3		Modelo, Pingo Doce, Minipreço (1,2,5)
12,7	0	3,85		Feira Nova, Supa/Jumbo (3,4)
10,3	3,85	0		Continente (6)

Atualização

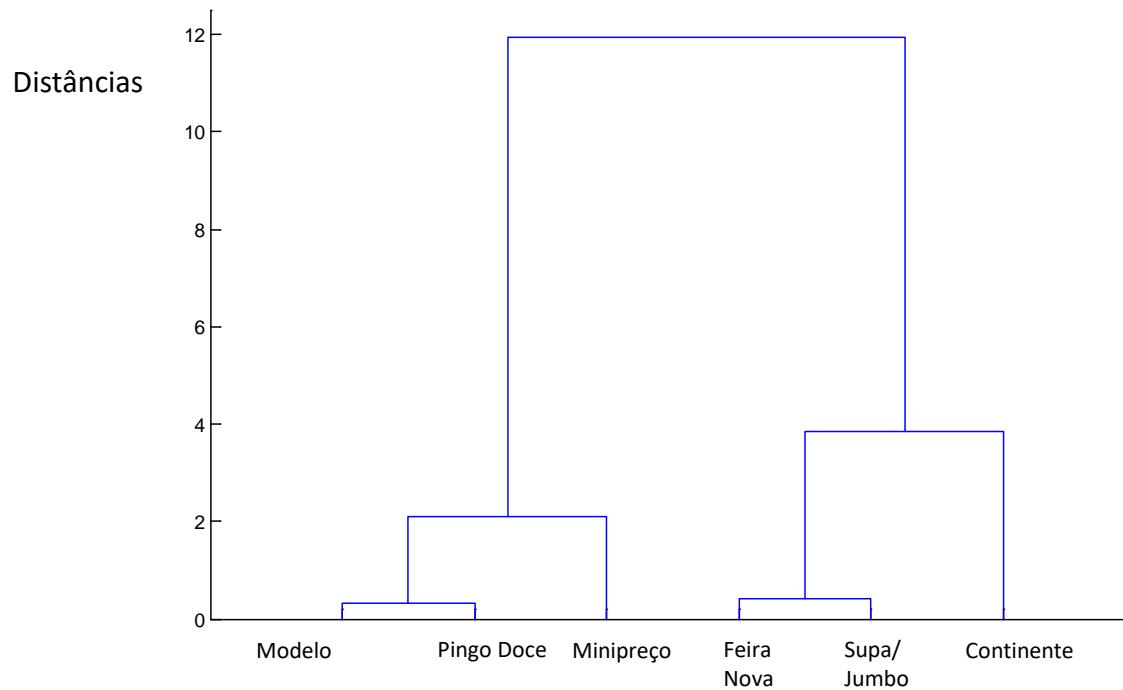
	Minipreço	Continente	
Pingo Doce			
Modelo	(1,2,5)	Supa / Jumbo	
0	11,5		Modelo, Pingo Doce, Minipreço (1,2,5)
11,5	0		Feira Nova, Supa/Jumbo, Continente (3,4,6)

$$11,5 = \text{média}(12,7 ; 10,3)$$

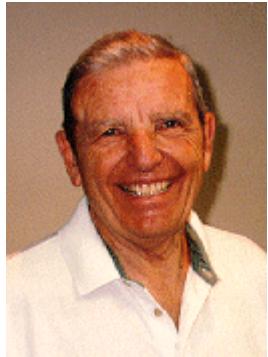
Métodos hierárquicos - Encadeamento Médio (Reis, 2001)

Seqüência de agrupamento e dendrograma

Passo	Distâncias	Nº de clusters	Clusters
1	$d_{1,2} = 0,3$	5	12 / 3 / 4 / 5 / 6
2	$d_{3,4} = 0,4$	4	12 / 34 / 5 / 6
3	$d_{12,,5} = 2,1$	3	125 / 34 / 6
4	$d_{34,,6} = 3,85$	2	125 / 346
5	$d_{125,,346} = 11.5$	1	123456



Método de Ward



Joe H. Ward Jr.
1926 - 2011

Ward, J. H., Jr. (1963), "Hierarchical Grouping to Optimize an Objective Function", Journal of the American Statistical Association, 58, 236–244.

<http://iv.slis.indiana.edu/sw/data/ward.pdf>

A inércia dentro dos *clusters* WSS aumenta com a fusão de dois *clusters* em cada etapa dos métodos hierárquicos aglomerativos.

Para minimizar este aumento inexorável, o método de Ward agrupa o par de *clusters* que minimiza o incremento na WSS.

Agraga os clusters *i* e *j* se a distância d_{ij} entre eles é mínima

$$d_{ij} = \frac{n_i n_j}{n_i + n_j} \left\| \bar{X}_i - \bar{X}_j \right\|^2$$

n_i = número de elementos em *i*
 n_j = número de elementos em *j*

Quadrado da distância euclidiana entre os centros de *i* e *j*

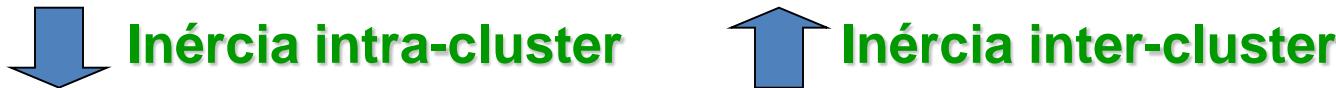


Métodos hierárquicos aglomerativos: Método de Ward

Inicialmente, cada cluster é composto por apenas um objeto (N clusters).
a inércia intra-cluster é nula e a inércia inter-cluster é igual a inércia total.

No final do processo de aglomeração, todos os objetos são agrupados em um único cluster.

a inércia inter-cluster é nula e a inércia intra-cluster é igual a inércia total.



A cada iteração, a fusão de dois clusters transfere uma parcela Δ da inércia inter-cluster para a inércia intra-cluster, aumentando a heterogeneidade interna dos agrupamentos, deteriorando a qualidade da partição:

$$\Delta = n_i d_{(G_i, G_{i \cup j})}^2 + n_j d_{(G_j, G_{i \cup j})}^2$$

Parcela resultante da agregação dos cluster i e j
 $G_{i \cup j}$ é o centro de gravidade da fusão dos clusters i e j

O método de Ward agrupa os dois *cluster* que minimizam esta parcela Δ .



Métodos hierárquicos aglomerativos

Para minimizar o incremento da inércia intra cluster a cada interação, o método de Ward trabalha com a seguinte métrica na matriz de distâncias.

$$d_{ij} = \frac{n_i n_j}{n_i + n_j} d^2(G_i, G_j)$$

n_i = nº de elementos do cluster i

n_j = nº de elementos do cluster j

G_i = centro de gravidade do cluster i

G_j = centro de gravidade do cluster j

$d^2(G_i, G_j)$ = quadrado da distância euclidiana entre G_i e G_j



Métodos hierárquicos aglomerativos: Método de Ward

Dados

	X	Y
A	1	1
B	2	2
C	3,5	4,5
D	5,5	3
E	6	5
F	5	5



Matriz de distâncias

	A	B	C	D	E	F
A	0					
B	1	0				
C	9,250	4,250	0			
D	12,125	6,625	3,125	0		
E	20,500	12,500	3,250	2,125	0	
F	16,000	9,000	1,250	2,125	0,500	0

$$d_{AC} = \frac{n_A n_C}{n_A + n_C} d^2(G_A, G_C) = \frac{1 \cdot 1}{1 + 1} \left[(1 - 3,5)^2 + (1 - 4,5)^2 \right] = 9,25$$



Métodos hierárquicos aglomerativos: Método de Ward

Exemplo: 1^a iteração

Matriz de distâncias

	A	B	C	D	E	F
A	0					
B	1	0				
C	9,250	4,250	0			
D	12,125	6,625	3,125	0		
E	20,500	12,500	3,250	2,125	0	
F	16,000	9,000	1,250	2,125	0,500	0

Menor distância entre E e F
Forme cluster EF
Centro de gravidade do cluster EF

$$G_{EF} = \left(\frac{x_E + x_F}{2}; \frac{y_E + y_F}{2} \right) = (5,5;5)$$

Atualiza matriz

	A	B	C	D	E-F
A	0				
B	1,000	0			
C	9,250	4,250	0		
D	12,125	6,625	3,125	0	
E-F	24,167	14,167	2,833	2,667	0

$$d_{EF,A} = \frac{n_{EF}n_A}{n_{EF} + n_A} d^2(G_{EF}, G_A) = \frac{2 \cdot 1}{2 + 1} [(5,5 - 1)^2 + (5 - 1)^2] = 24,167$$



Métodos hierárquicos aglomerativos: Método de Ward

Exemplo: 2^a iteração

Matriz de distâncias

	A	B	C	D	E-F
A	0				
B	1,000	0			
C	9,250	4,250	0		
D	12,125	6,625	3,125	0	
E-F	24,167	14,167	2,833	2,667	0

Menor distância entre A e B
Forme cluster AB
Centro de gravidade do cluster AB

$$G_{AB} = \left(\frac{x_A + x_B}{2}, \frac{y_A + y_B}{2} \right) = (1,5; 1,5)$$

	A-B	C	D	E-F
A-B	0			
C	8,667	0		
D	12,167	3,125	0	
E-F	28,250	2,833	2,667	0

Atualiza matriz



$$d_{EF,AB} = \frac{n_{EF} n_{AB}}{n_{EF} + n_{AB}} d^2(G_{EF}, G_{AB}) = \frac{2 \cdot 2}{2 + 2} \left[(5,5 - 1,5)^2 + (5 - 1,5)^2 \right] = 28,250$$



Métodos hierárquicos aglomerativos: Método de Ward

Exemplo: 3^a iteração

Matriz de distâncias

	A-B	C	D	E-F
A-B	0			
C	8,667	0		
D	12,167	3,125	0	
E-F	28,250	2,833	2,667	0

Menor distância entre E-F e D
Forme cluster EFD
Centro de gravidade do cluster EFD

$$G_{EFD} = \left(\frac{x_E + x_F + x_D}{3}; \frac{y_E + y_F + y_D}{3} \right) = (5,5; 4,33)$$

Atualiza matriz

	A-B	C	E-F-D
A-B	0		
C	8,67	0	
E-F-D	28,83	3,02	0

$$d_{EFD,C} = \frac{n_{EFD}n_C}{n_{EFD} + n_C} d^2(G_{EFD}, G_C) = \frac{3 \cdot 1}{3 + 1} [(5,5 - 3,5)^2 + (4,33 - 4,5)^2] = 3,02$$



Métodos hierárquicos aglomerativos: Método de Ward

Exemplo: 4^a iteração

Matriz de distâncias

	A-B	C	E-F-D
A-B	0		
C	8,67	0	
E-F-D	28,83	3,02	0

**Menor distância entre EFD e C
Forme cluster EFDC
Centro de gravidade do cluster EFDC**

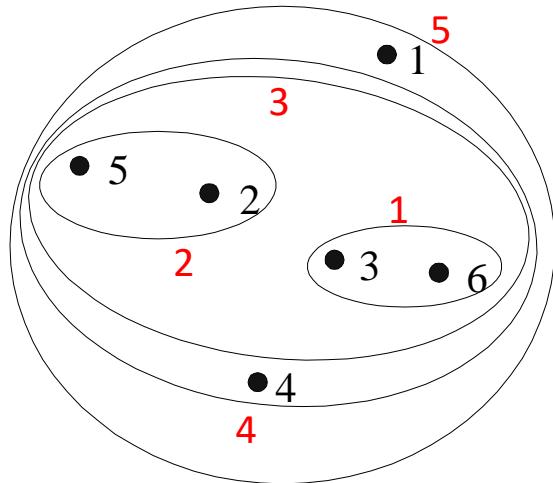
$$G_{EFDC} = \left(\frac{x_E + x_F + x_D + x_C}{4}; \frac{y_E + y_F + y_D + y_C}{4} \right) = (5; 4,375)$$

Atualiza matriz

	A-B	E-F-D-C
A-B	0	
E-F-D-C	27,35	0

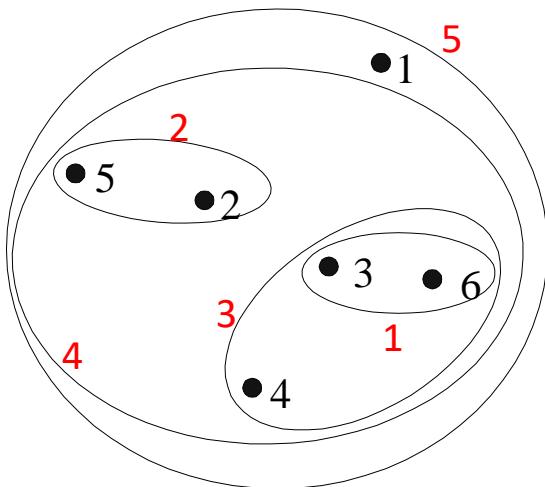
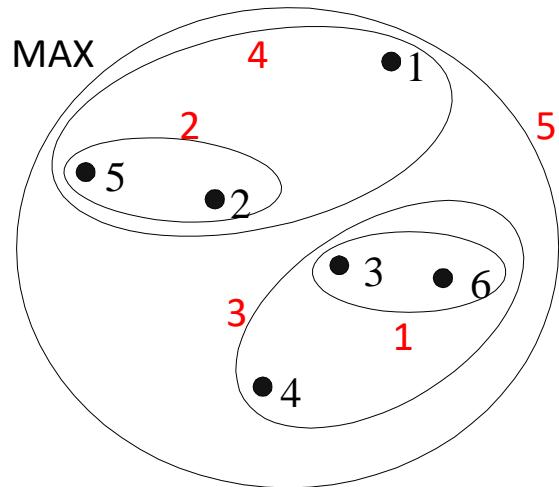
$$d_{EFDC,AB} = \frac{n_{EFDC} n_{AB}}{n_{EFDC} + n_{AB}} d^2(G_{EFDC}, G_{AB}) = \frac{4 \cdot 2}{4 + 2} \left[(5 - 1,5)^2 + (4,375 - 1,5)^2 \right] = 27,35$$

Comparação de métodos hierárquicos



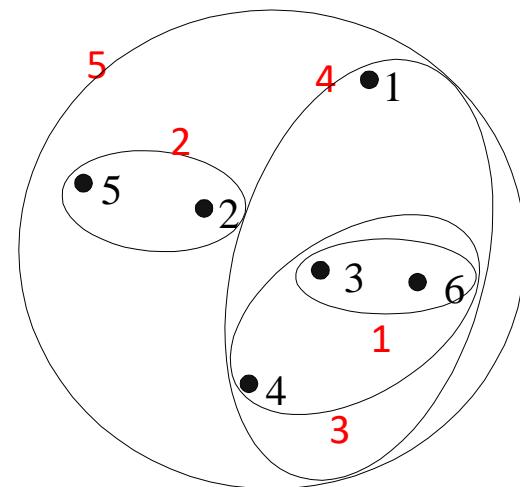
Simples MIN

Completo MAX



Médio

Ward's Method



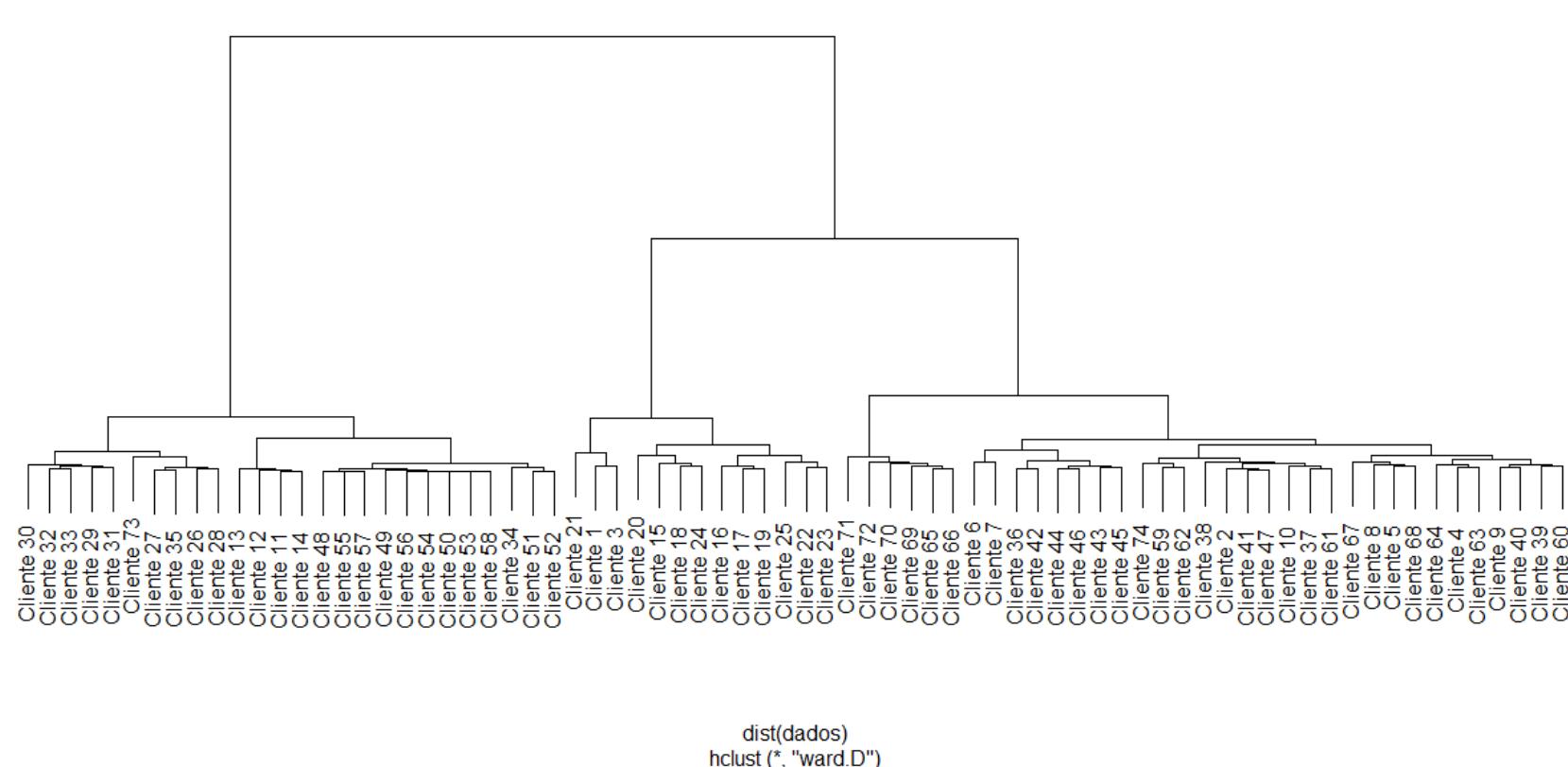
Exemplo Tipologias de curvas de carga (Pessanha et al, 2015)

classificação pelo método de Ward

```
resultado.hc = hclust(dist(dados),method="ward.D",members=NULL)
```

dendrograma

```
plot(resultado.hc,ylab="Distâncias",main="")
```



Exemplo Tipologias de curvas de carga (Pessanha et al, 2015)

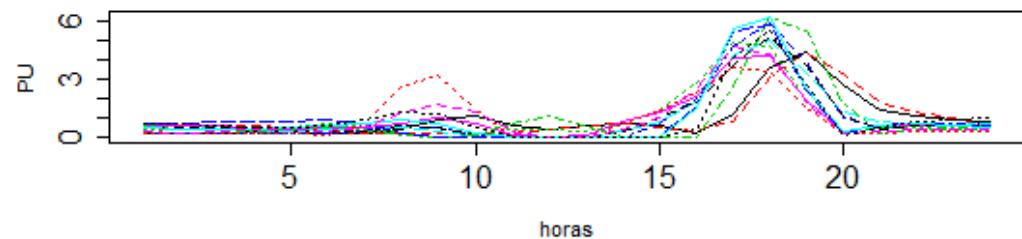
seleciona solução com 3 clusters

clusters = cutree(resultado.hc,k=3)

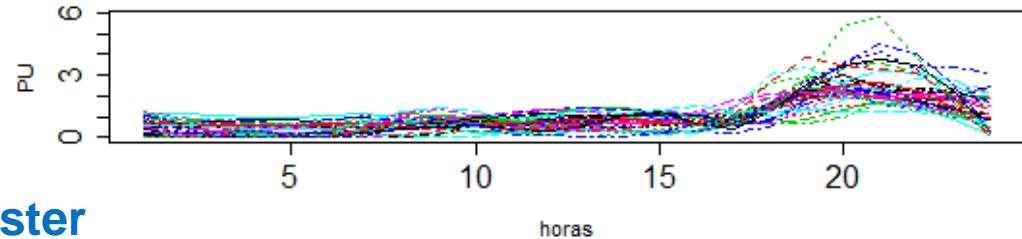
```
> clusters
 Cliente 1 Cliente 2 Cliente 3 Cliente 4 Cliente 5 Cliente 6 Cliente 7
      1          2          1          2          2          2          2
 Cliente 8 Cliente 9 Cliente 10 Cliente 11 Cliente 12 Cliente 13 Cliente 14
      2          2          2          3          3          3          3
 Cliente 15 Cliente 16 Cliente 17 Cliente 18 Cliente 19 Cliente 20 Cliente 21
      1          1          1          1          1          1          1
 Cliente 22 Cliente 23 Cliente 24 Cliente 25 Cliente 26 Cliente 27 Cliente 28
      1          1          1          1          3          3          3
 Cliente 29 Cliente 30 Cliente 31 Cliente 32 Cliente 33 Cliente 34 Cliente 35
      3          3          3          3          3          3          3
 Cliente 36 Cliente 37 Cliente 38 Cliente 39 Cliente 40 Cliente 41 Cliente 42
      2          2          2          2          2          2          2
 Cliente 43 Cliente 44 Cliente 45 Cliente 46 Cliente 47 Cliente 48 Cliente 49
      2          2          2          2          2          3          3
 Cliente 50 Cliente 51 Cliente 52 Cliente 53 Cliente 54 Cliente 55 Cliente 56
      3          3          3          3          3          3          3
 Cliente 57 Cliente 58 Cliente 59 Cliente 60 Cliente 61 Cliente 62 Cliente 63
      3          3          2          2          2          2          2
 Cliente 64 Cliente 65 Cliente 66 Cliente 67 Cliente 68 Cliente 69 Cliente 70
      2          2          2          2          2          2          2
 Cliente 71 Cliente 72 Cliente 73 Cliente 74
      2          2          3          2
```

Exemplo Tipologias de curvas de carga (Pessanha et al, 2015)

cluster 1



cluster 2



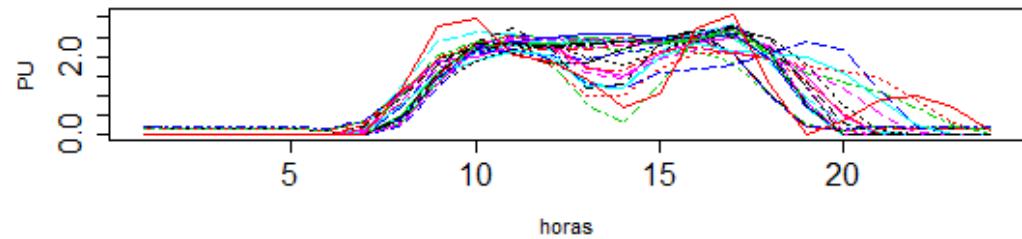
número de elementos em cada cluster

```
table(clusters) clusters
  1   2   3
 13  35  26
```

curvas em cada cluster

```
par(mfrow=c(3,1))
for (i in 1:3) {
  cluster = which(clusters==i)
  matplot(matrix(seq(1,24,1),ncol=1),t(dados[cluster,]),type='l',ylab='PU',xlab='horas'
  ,main=paste('cluster',i),cex.main=2,cex.axis=1.5)
}
```

cluster 3



Métodos hierárquicos aglomerativos

Vantagens:

- Rápidos e exigem menos tempo de processamento.
- Apresentam resultados para diferentes níveis de agregação.

Desvantagens:

- A alocação de um objeto em um cluster é irrevogável, ou seja, uma vez o objeto incluído em um cluster, ele nunca será removido e ligado a outro agrupamento. Combinações anteriores indesejáveis persistem no decorrer da análise e levam a resultados artificiais. Não existe a possibilidade de realocação de objetos que possam ter sido incorretamente agrupados nos estágios anteriores.
- Impacto substancial dos outliers. Ward é o menos suscetível aos outliers.
- Não são apropriados para analisar uma amostra muita extensa, pois a medida que o tamanho da amostra aumenta, a necessidade de armazenamento da matriz de distâncias cresce drasticamente.



Métodos hierárquicos x Métodos não hierárquicos

Métodos Hierárquicos são preferidos quando:

- Serão analisadas varias alternativas de agrupamento.
- O tamanho da amostra é moderado (de 300 a 1000 objetos)

Métodos não-hierárquicos são preferidos quando:

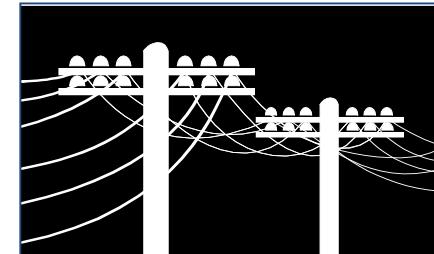
- O número de grupos é conhecido.
- Presença dos outliers, desde que os métodos não-hierárquicos são menos influenciados por outliers.
- Há um grande nº de objetos a serem agrupados.

Método de Ward no R

Para fins de regulação econômica da distribuição de energia elétrica é interessante segmentar as empresas distribuidoras em clusters de forma a permitir análises comparativas entre empresas semelhantes.

Considere uma matriz de dados formada por 59 concessionárias de distribuição que atuam no setor elétrico brasileiro, cada uma descrita por 9 variáveis (Fonte: Aneel):

- Mercado atendido (MWh)
- Número de consumidores
- Tamanho da rede de distribuição
- Densidade de consumidores (consumidores/km de rede)
- Consumo por unidades consumidora – CPC (MWh/consumidor)
- Índice de complexidade no combate às perdas não técnicas
- Composição do mercado por classe de tendão BT, MT e AT.



Método de Ward: Código R

```
# Leitura do arquivo de dados
dados = read.csv("c:/cursoR/distribuidoras.csv",sep=",",dec=".",
header=TRUE)
empresas=dados[,2:10]
rownames(empresas)=dados[,1]

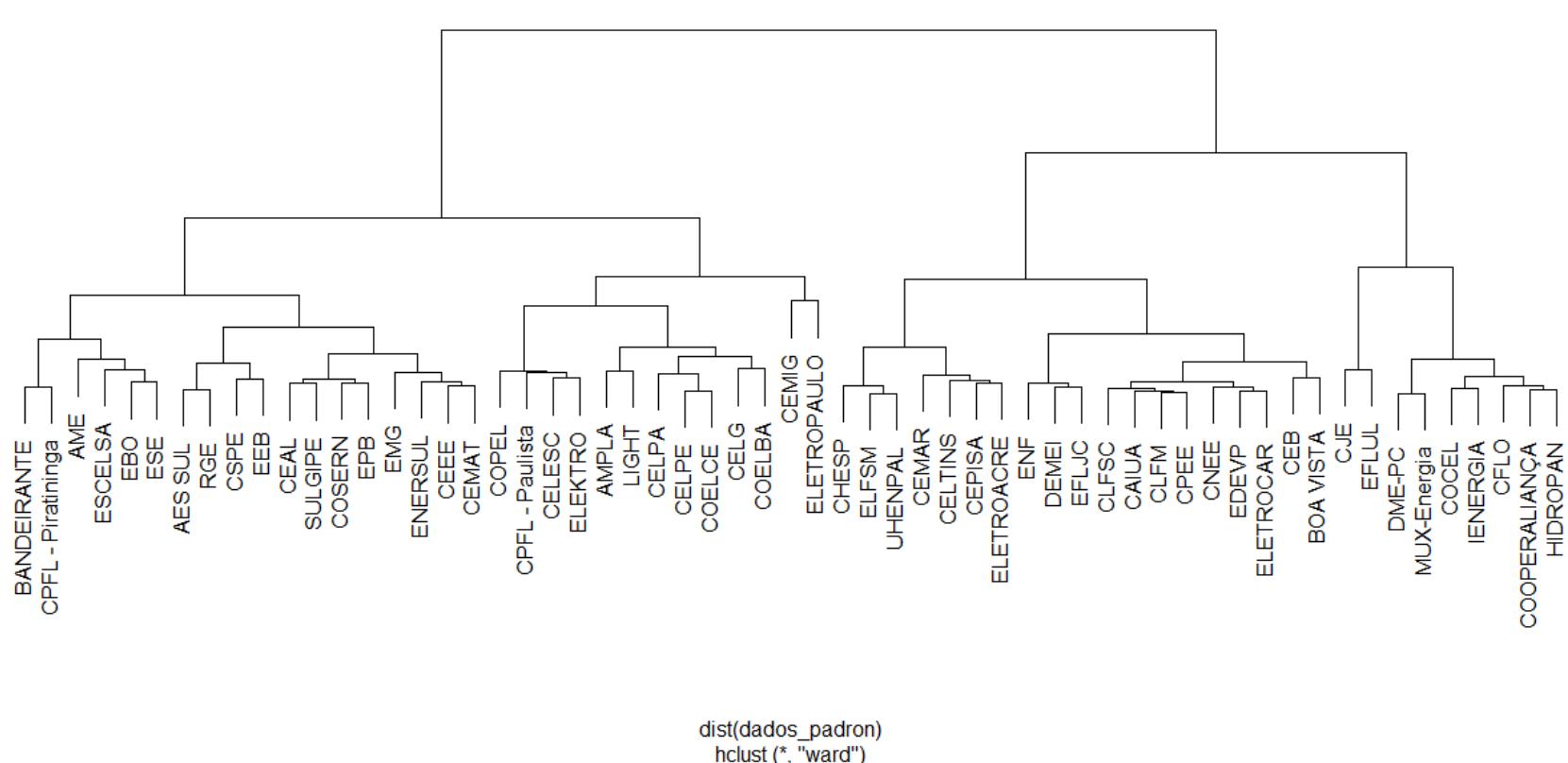
# Padronização dos dados
dados_padron=scale(empresas,center=TRUE,scale=TRUE)

# Execução do método de Ward
saída_Ward=hclust(dist(dados_padron),method="ward")

# Faz o dendrograma
plot(saída_Ward,main="Dendrograma",ylab="distâncias")
```

Método de Ward: Código R

Dendrograma



Método de Ward: Código R

Mostra as partições para diferentes níveis de agregação, por exemplo, de 2 a 5 clusters. **classe=cutree(saida_Ward,k=2:5)**

	2	3	4	5
AES SUL	1	1	1	1
AMPLA	1	1	2	2
BANDEIRANTE	1	1	1	1
CEB	2	2	3	3
CEEE	1	1	1	1
CELESC	1	1	2	2
CEMIG	1	1	2	2
COPEL	1	1	2	2
CPFL - Paulista	1	1	2	2
CPFL - Piratininga	1	1	1	1
ELEKTRO	1	1	2	2
ELETROPAULO	1	1	2	2
ESCELSA	1	1	1	1
LIGHT	1	1	2	2
RGE	1	1	1	1
AME	1	1	1	1
CAIUA	2	2	3	3
CEAL	1	1	1	1
CELG	1	1	2	2
CELPA	1	1	2	2
CELPE	1	1	2	2
CELTINS	2	2	3	3
CEMAR	2	2	3	3
CEMAT	1	1	1	1
CEPISA	2	2	3	3
CLFSC	2	2	3	3
CNEE	2	2	3	3
COELBA	1	1	2	2
COELCE	1	1	2	2

Número de agrupamentos

COSERN	1	1	1	1
CSPE	1	1	1	1
EBO	1	1	1	1
EDEVP	2	2	3	3
EEB	1	1	1	1
EMG	1	1	1	1
ENERSUL	1	1	1	1
EPB	1	1	1	1
ESE	1	1	1	1
SULGIPE	1	1	1	1
BOA VISTA	2	2	3	3
CFLO	2	3	4	4
CHESP	2	2	3	3
CJE	2	3	4	5
CLFM	2	2	3	3
COCEL	2	3	4	4
COOPERALIANÇA	2	3	4	4
CPEE	2	2	3	3
DEMEI	2	2	3	3
DME-PC	2	3	4	4
EFLJC	2	2	3	3
EFLUL	2	3	4	5
ELETROACRE	2	2	3	3
ELETROCAR	2	2	3	3
ELFSM	2	2	3	3
ENF	2	2	3	3
HIDROPAN	2	3	4	4
IENERGIA	2	3	4	4
MUX-Energia	2	3	4	4
UHENPAL	2	2	3	3

Na partição em 5 clusters
a Cosern pertence ao cluster 1

Na partição em 2 clusters
a Eletroacre pertence ao cluster 2

Método de Ward: Código R

```
# Total de empresas em cada cluster na solução com 3 agrupamentos  
table(classe[,2])
```

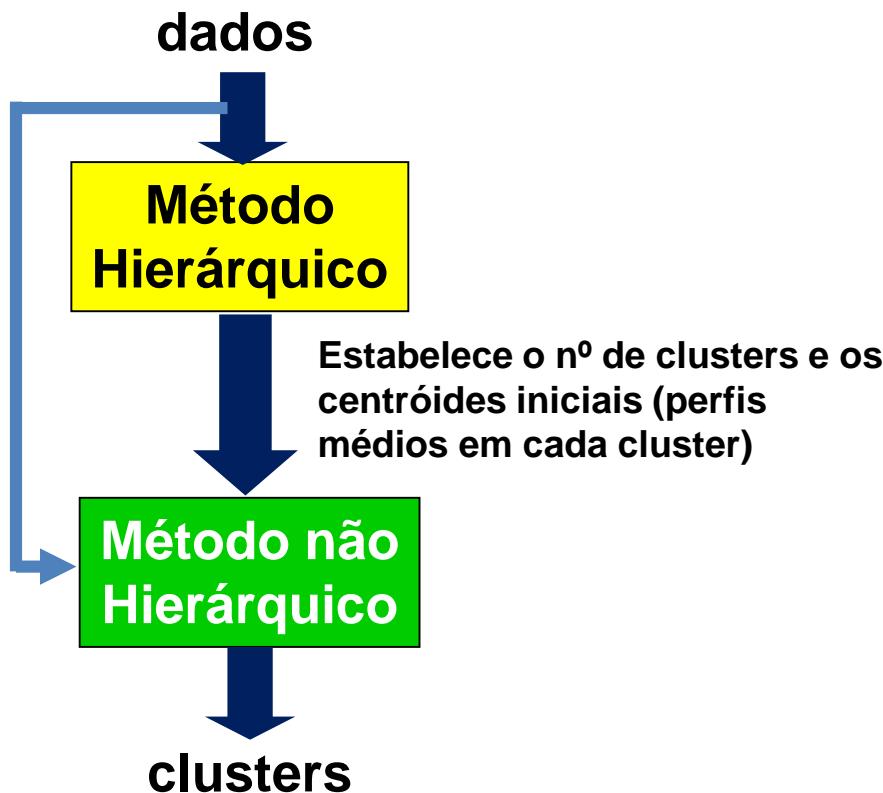
	1	2	3
	31	19	9

```
# Empresas no cluster 1
```

```
empresas1=which(classe[,2]==1)
```

	AES SUL	AMPLA	BANDEIRANTE	CEEE
	1	2	3	5
	CELESC	CEMIG	COPEL	CPFL - Paulista
	6	7	8	9
CPFL - Piratininga	10	ELEKTRO	ELETROPAULO	ESCELSA
	11		12	13
LIGHT		RGE	AME	CEAL
	14	15	16	18
CELG		CELPA	CELPE	CEMAT
	19	20	21	24
COELBA	28	COELCE	COSERN	CSPE
	29		30	31
EBO		EEB	EMG	ENERSUL
	32	34	35	36
EPB		ESE	SULGIPE	
	37	38	39	

Aproveita as vantagens dos métodos hierárquicos e não hierárquicos.



Estratégia adotada no SNACC



Mapa Auto organizáveis

Fuzzy C Means Method

DBSCAN

Teoria clássica dos conjuntos (Lógica Clássica)

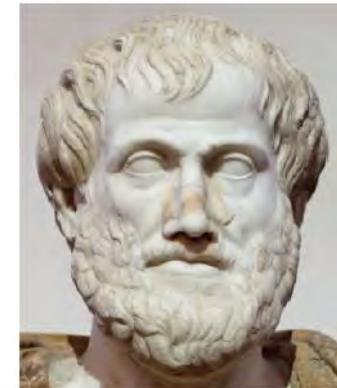
Os elementos são classificados em conjuntos bem definidos (crisp set)

Um elemento pertence ou não a um determinado conjunto.

Exemplo:

Consumidor adimplente: paga a conta até a data do vencimento

Consumidor inadimplente: não paga a conta até a data do vencimento



Aristóteles

384 a.C. - 322 a.C.

Lógica Booleana: VERDADEIRO ou FALSO.

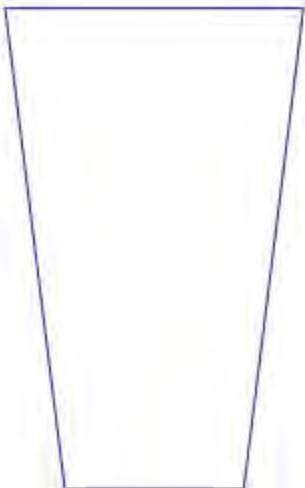
O grau de pertinência de um objeto ou indivíduo a um determinado conjunto é uma variável binária:

“1” se elemento pertence a um conjunto

“0” se elemento não pertence a um conjunto

Um consumidor que PAGA a conta em dia tem grau de pertinência 1 no conjunto adimplente e grau de pertinência 0 no conjunto inadimplente.

Um consumidor que NÃO PAGA a conta em dia tem grau de pertinência 0 no conjunto adimplente e grau de pertinência 1 no conjunto inadimplente.



Copo vazio



Copo cheio

O que dizer do copo do meio ?

O copo está meio cheio e meio vazio.

O copo tem grau de pertinência de 50% no conjunto cheio e grau de pertinência de 50% no conjunto vazio

Teoria dos conjuntos difusos (Lógica Fuzzy)

Proposta por Zadeh em 1965

INFORMATION AND CONTROL 8, 338–353 (1965)

Fuzzy Sets*

L. A. ZADEH

*Department of Electrical Engineering and Electronics Research Laboratory,
University of California, Berkeley, California*

A fuzzy set is a class of objects with a continuum of grades of membership. Such a set is characterized by a membership (characteristic) function which assigns to each object a grade of membership ranging between zero and one. The notions of inclusion, union, intersection, complement, relation, convexity, etc., are extended to such sets, and various properties of these notions in the context of fuzzy sets are established. In particular, a separation theorem for convex fuzzy sets is proved without requiring that the fuzzy sets be disjoint.



Lotfi Asker Zadeh
Baku, Azerbaijão 1921

Zadeh, L.A. Fuzzy Sets, Information and Control, vol. 8, n.3, p. 338-353, June, 1965

<https://people.eecs.berkeley.edu/~zadeh/papers/Fuzzy%20Sets-Information%20and%20Control-1965.pdf>

Teoria dos conjuntos difusos (Lógica Fuzzy)

Trata de questões associadas à imprecisão na descrição das propriedades de um fenômeno (vagueza da informação), por exemplo, afirmações vagas:

- O indivíduo é idoso
- O indivíduo é alto
- O transporte é muito rápido
- A temperatura está quente
- A chuva está forte
- O consumo domiciliar é baixo
- A participação da conta de luz no orçamento familiar é alta

Os exemplos acima são informações linguísticas, ou seja, sentenças expressas em linguagem natural (baixo, moderado baixo, médio, moderado alto, alto)

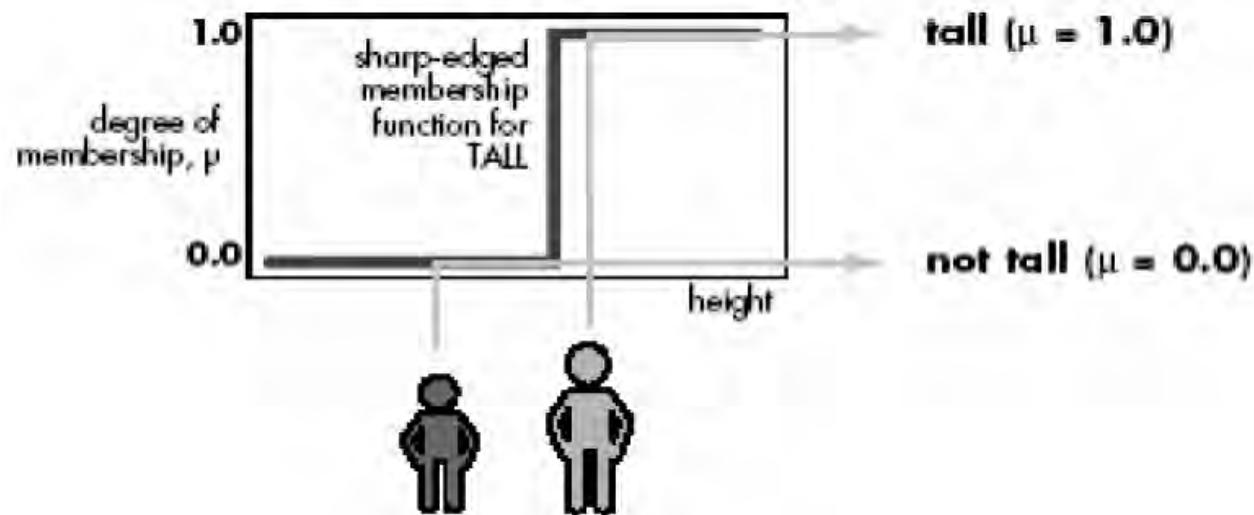
A Lógica Fuzzy fornece os fundamentos para efetuar o raciocínio aproximado com proposições imprecisas, permitindo tratar dados vagos ou imprecisos.

Lógica Fuzzy x Teoria clássicas dos conjuntos

Função de pertinência
do conjunto alto na
Teoria clássica dos
conjuntos (crisp set)

Transição abrupta

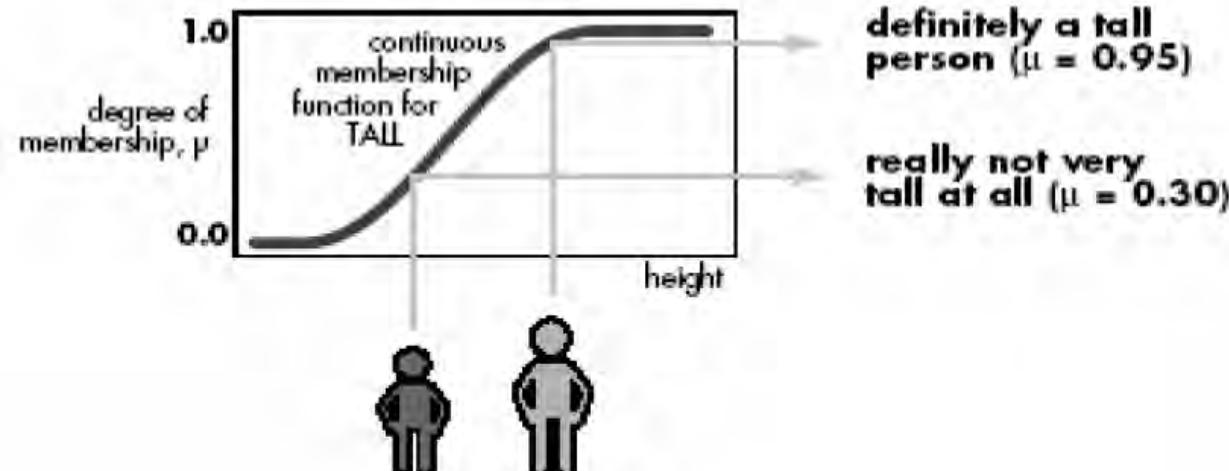
Conjunto clássico pessoas altas



Função de pertinência do
conjunto fuzzy alto (fuzzy
set)

Transição gradual

Conjunto fuzzy pessoas altas

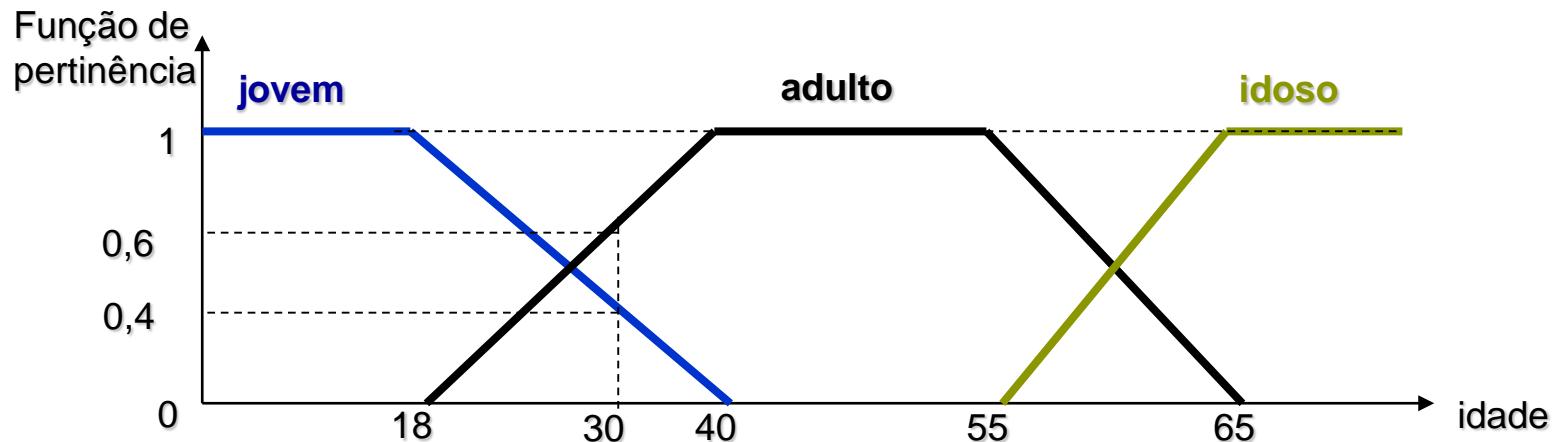


Fuzzy C Means Method

Introduz a lógica fuzzy na particão de um conjunto com n objetos em k cluster.

O que é lógica fuzzy ?

Exemplo: A seguir temos três conjuntos fuzzy (jovem, adulto e idoso) e suas respectivas funções de pertinência



Um indivíduo com idade menor ou igual a 18 anos certamente é jovem.

Um indivíduo com idade superior a 65 anos certamente é idoso.

Mas, o que dizer de um indivíduo com 30 anos de idade: ele é jovem ou adulto?

Neste caso podemos dizer que este indivíduo pertence ao conjunto jovem com pertinência de 0,4 a ao conjunto adulto com pertinência 0,6 (um indivíduo pode pertencer a vários conjuntos fuzzy, mas com diferentes graus de pertinência)

Fuzzy C Means Method

A aplicação da lógica fuzzy na análise de agrupamentos implica em assumir que um mesmo objeto pode pertencer a todos os clusters porém com diferentes graus de pertinência.

No K-Means cada objeto pode pertencer a apenas um cluster, logo a função de pertinência é binária: 0 não pertence , 1 pertence.

No método fuzzy (FCM) a função de pertinência não é binária, mas um número entre 0 e 1 que indica a “força” com que o objeto pertence ao cluster.

A soma dos graus de pertinência de um objeto aos diferentes clusters deve ser unitária.

Fuzzy C Means Method

No FCM a partição ótima resulta da solução do seguinte problema de otimização resolvido iterativamente com o objetivo de encontrar os centróides dos clusters de tal forma que a inércia intra-cluster seja mínima:

$$\text{Min} \quad J = \sum_{h=1}^K \sum_{i=1}^n u_{ij}^m \|x_i - C_h\|^2 \quad \text{Função objetivo é uma versão da inércia intra cluster}$$

s.a. $\sum_{h=1}^k u_{1h} = 1 \quad \text{Soma do grau de pertinência do objeto #1 nos k clusters é 1}$

...

$$\sum_{h=1}^k u_{nh} = 1 \quad \text{Soma do grau de pertinência do objeto #n nos k clusters é 1}$$

K é o número de clusters (K é um parâmetro de entrada)

m é a constante de fuzzyficação, em geral um número entre 1,25 e 2 (dado de entrada)

n é o número de objetos

u_{ih} é o grau de pertinência do objeto i ao cluster h

x_i é o objeto i

C_h é o centro de gravidade do cluster h .

Fuzzy C Means Method

Algoritmo do FCM

Passo 1 - Inicialize os graus de pertinência u_{ih} com valores aleatórios entre 0 e 1, observando a restrição de soma unitária dos valores referentes a cada objeto i ao longo dos k clusters. Esta etapa é conhecida como *fuzzyficação*.

Passo 2 - Calcule as coordenadas dos k centróides: $C_h = \sum_{i=1}^n u_{ih}^m x_i \Bigg/ \sum_{i=1}^n u_{ih}^m$

Passo 3 – Calcule o valor da função objetivo, pare se o valor estiver abaixo de uma tolerância ou se a melhoria em relação à iteração anterior for desprezível.

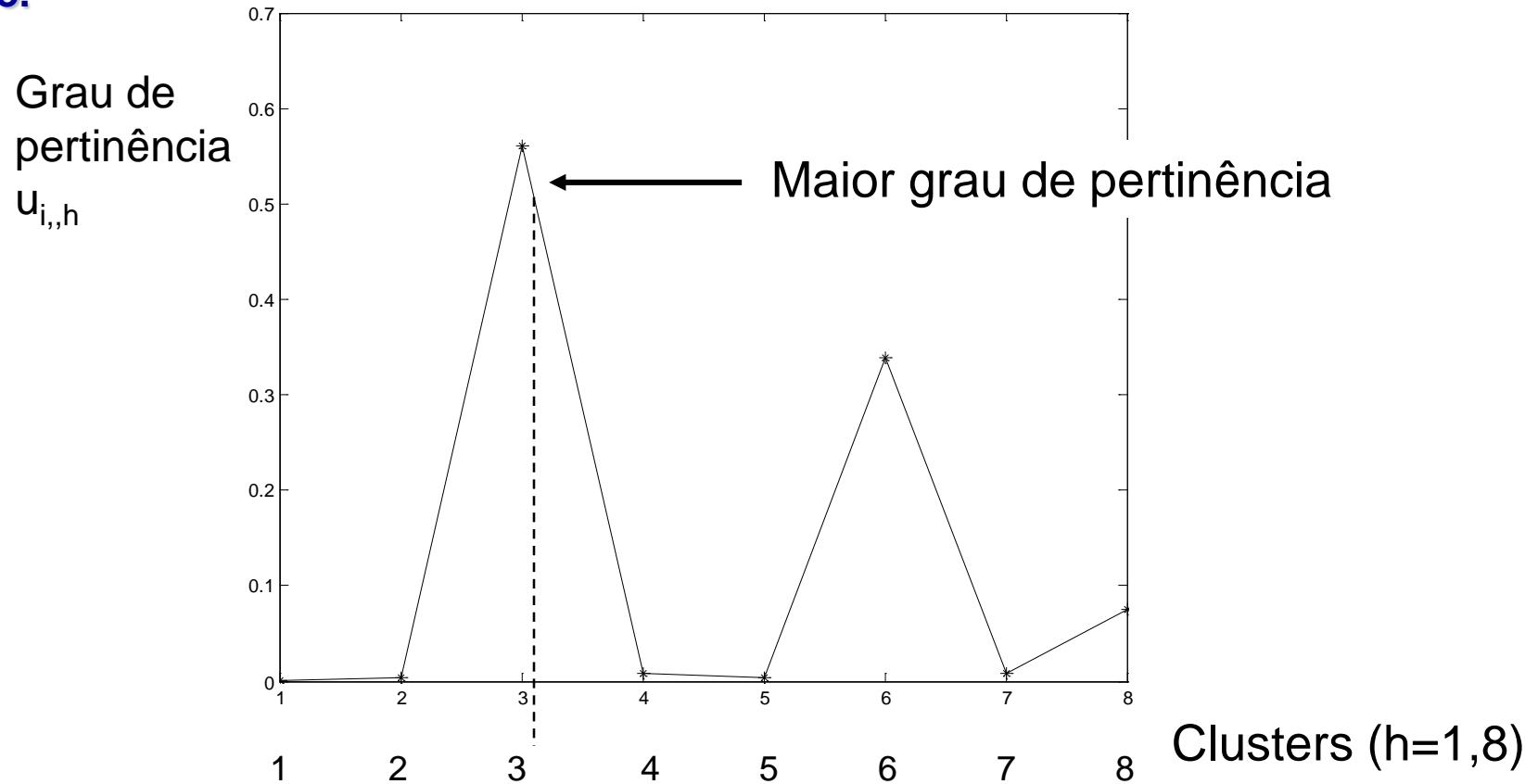
Passo 4 – Atualize os graus de pertinência $u_{ij} = 1 \Bigg/ \sum_{t=1}^k \left(\frac{\|x_i - c_j\|}{\|x_i - c_t\|} \right)^{\frac{2}{m-1}}$

Passo 5 – Retorne ao passo 2.

Após a convergência do algoritmo, são definidos os centróides dos clusters e o grau de pertinência de cada objeto i em cada cluster h (u_{ih}).

Fuzzy C Means Method

Exemplo de função de pertinência para um objeto i obtida após a convergência do algoritmo.



Cada objeto deve ser alocado no *cluster* onde apresenta maior grau de pertinência (etapa conhecida como *defuzzyficação*).

No exemplo acima, o objeto deve ser alocado no cluster 3.

Fuzzy C Means Method

O número ideal de clusters (k) é aquele que minimiza a medida de compacidade e separação.

$$CS = \frac{\sum_{h=1}^k \sum_{i=1}^n u_{ih}^m \|x_i - C_h\|^2}{n \cdot \min \|distância_entre_dois_centroides\|^2} \rightarrow$$

compacidade $\left\{ \begin{array}{l} \text{Mede a heterogeneidade} \\ \text{dentro dos clusters} \end{array} \right.$

separação $\left\{ \begin{array}{l} \text{Mede separação} \\ \text{dos clusters} \end{array} \right.$

Solução ideal deve satisfazer dois requisitos:

- Mínima heterogeneidade interna (numerador pequeno)
- Máxima separação dos clusters (denominador grande)

Mínimo CS

Fuzzy C Means Method

- Exemplo: Aplicação do FCM em um conjunto com 125 curvas de dias úteis de consumidores BT com intervalo de integração de 15 minutos (cada curva é um vetor com 96 pontos).
- Adotou-se uma constante m igual 1,25
- Foram testados diferentes número de agrupamentos

Número de classes	CS
2	1,0977
4	0,7315
6	0,8180
8	0,7034
10	0,8630
12	0,8809
14	0,9510
16	0,9393

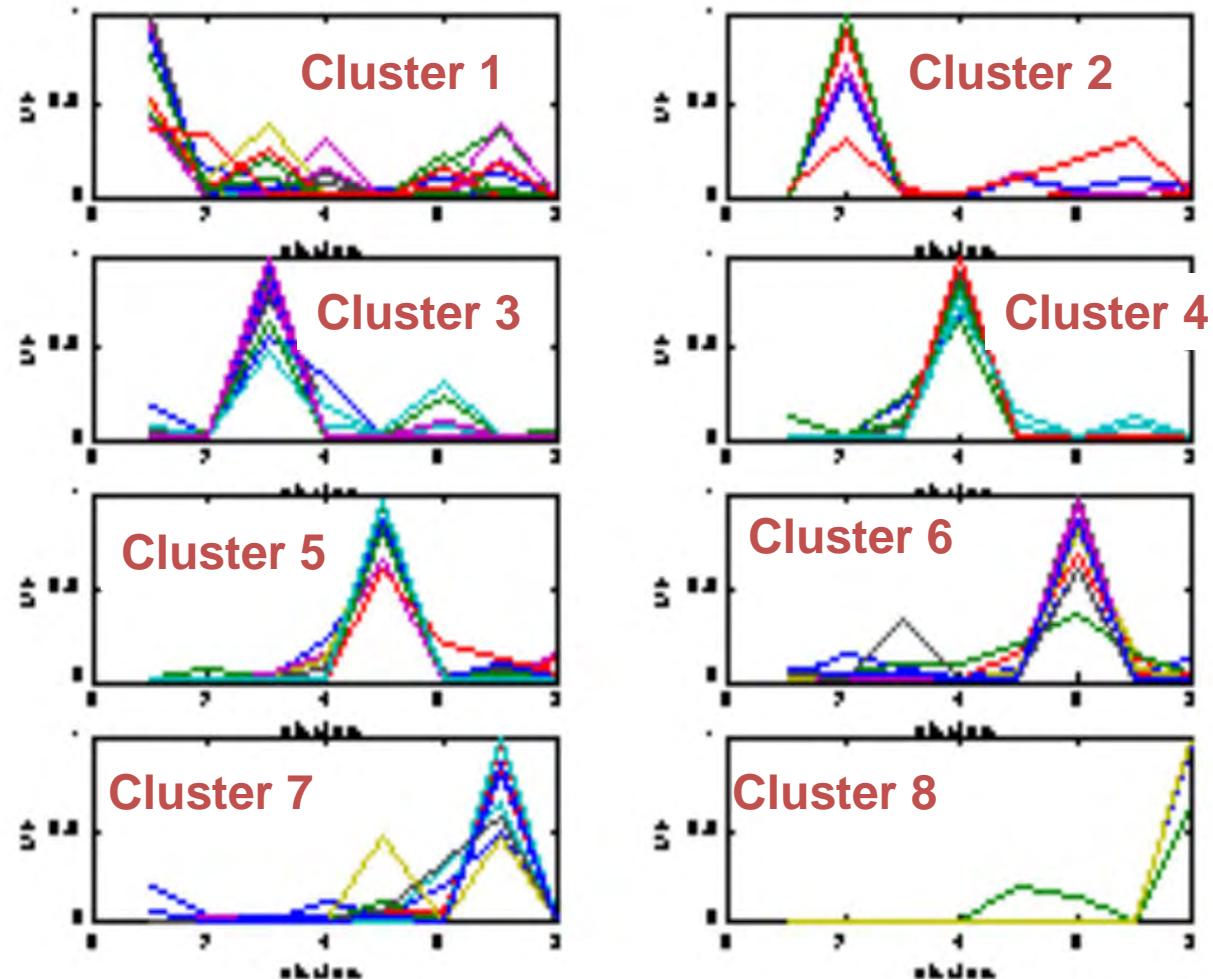
Mínimo CS
Melhor partição tem 8 clusters



Fuzzy C Means Method

- Defuzzyficação pelo máximo
- Cada curva de carga está associada com uma função de pertinência

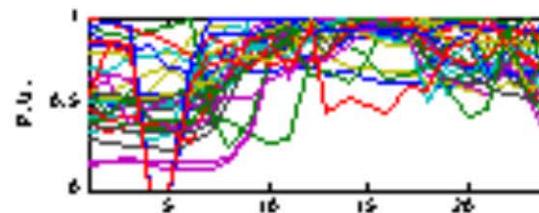
Funções de pertinência classificadas pela máxima pertinência



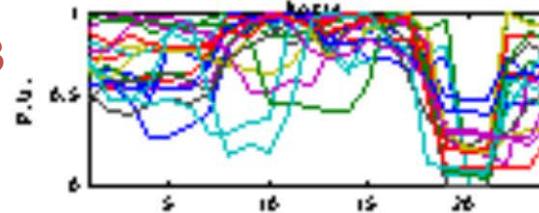
Fuzzy C Means Method

Classificação das curvas de carga

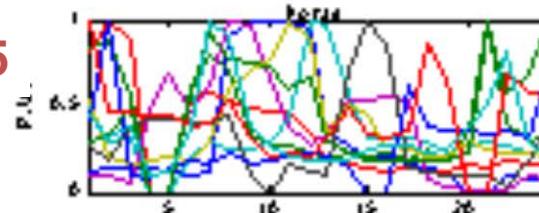
Cluster 1



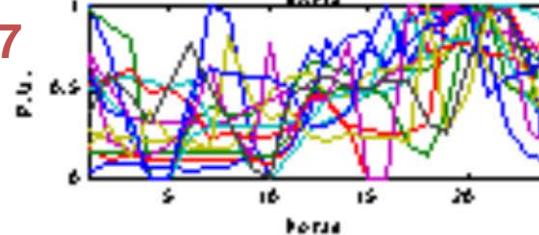
Cluster 3



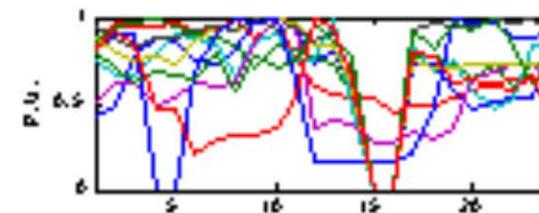
Cluster 5



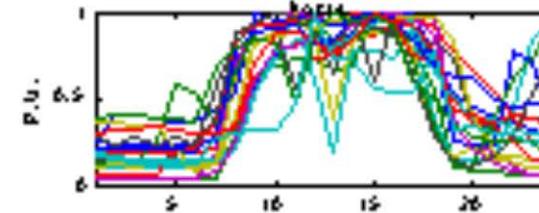
Cluster 7



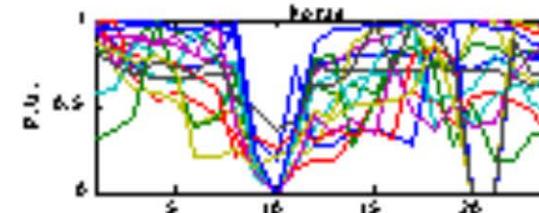
Cluster 2



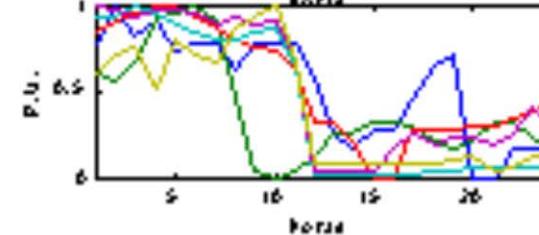
Cluster 4



Cluster 6



Cluster 8



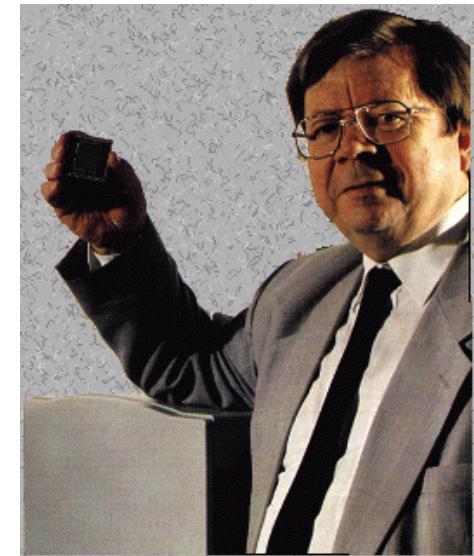
Rede auto-organizável

Proposto por Teuvo Kohonen em 1982

Rede neural com treinamento não supervisionado

A rede aprende as similaridades entre os padrões de entrada
(reconhecimento de padrões)

Útil na análise de agrupamentos (*cluster analysis*)



Teuvo Kohonen

Biol. Cybern. 43, 59–69 (1982)

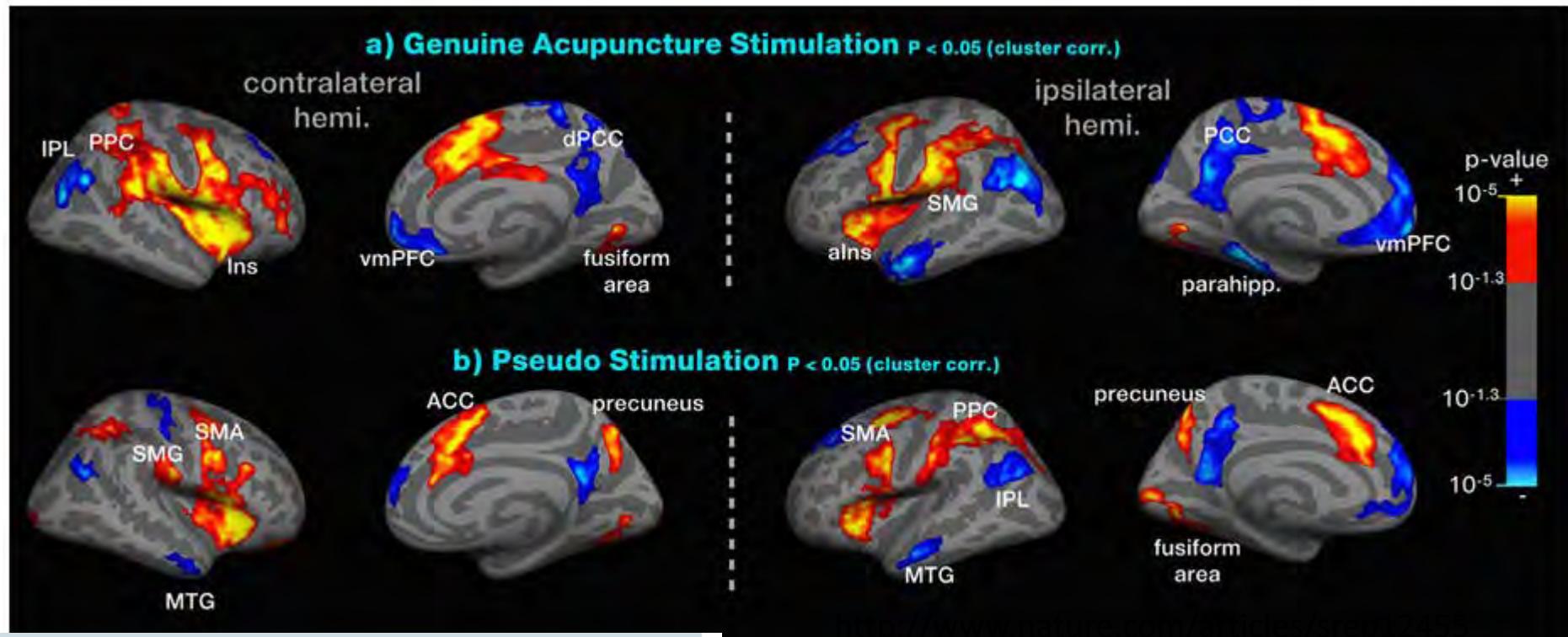
**Biological
Cybernetics**
© Springer-Verlag 1982

Self-Organized Formation of Topologically Correct Feature Maps

Teuvo Kohonen

Department of Technical Physics, Helsinki University of Technology, Espoo, Finland

Inspiração neurofisiológica



SCIENTIFIC
REPORTS

Altmetric: 5 Views: 1,092

More detail »

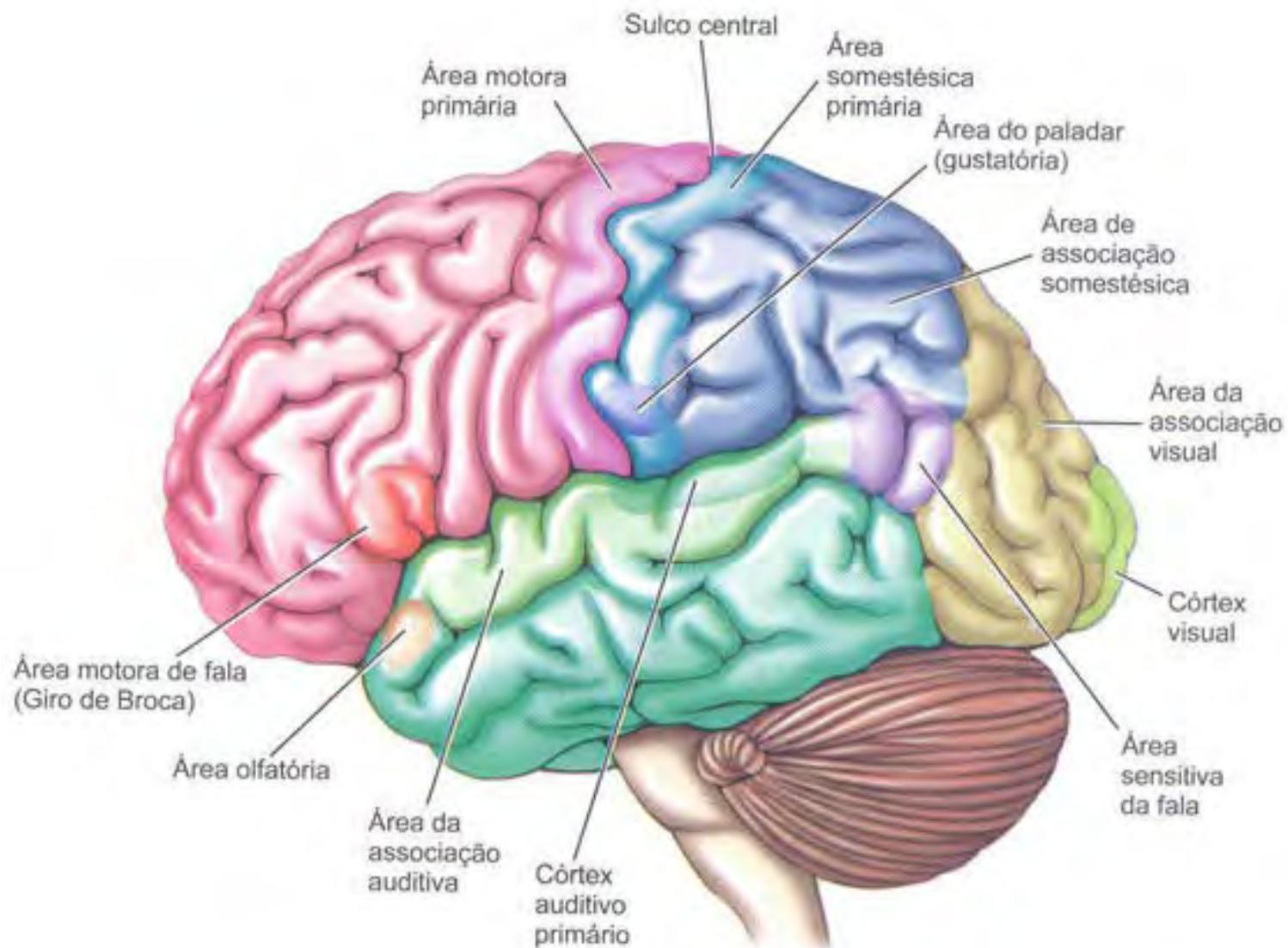
Article | OPEN

Cortical Activation Patterns of Bodily Attention triggered by Acupuncture Stimulation

Won-Mo Jung, In-Seon Lee ✉, Christian Wallraven, Yeon-Hee Ryu, Hi-Joon Park & Younbyoung Chae ✉

Ativação cortical relacionada à acupuntura

Inspiração neurofisiológica



Mapas auto-organizáveis

Idealizadas a partir da analogia com a região do córtex cerebral humano.

O córtex cerebral possui regiões responsáveis por funções específicas.

Existem regiões, por exemplo, dedicadas à fala, à visão, ao controle motor, etc.

Os neurônios estão espacialmente ordenados, e assim neurônios topologicamente próximos tendem a responder a padrões ou estímulos semelhantes.

Para uma determinada ativação cerebral, o grau de ativação dos neurônios diminui à medida que se aumenta a distância da região de ativação inicial.



Por exemplo, na figura ao lado, o centro de excitação em amarelo é rodeado por uma vizinhança que reage a esta excitação, ou seja, as regiões especializadas.

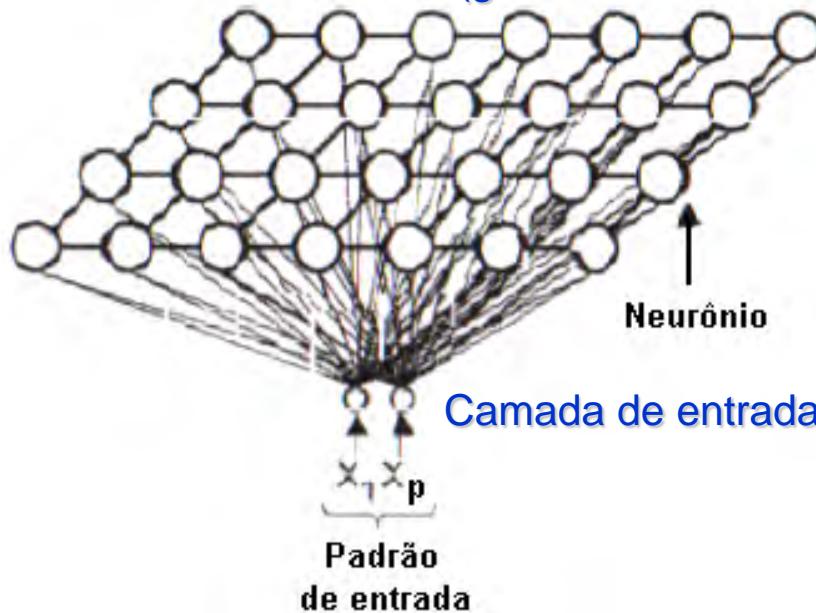
Este comportamento apresenta potencialidade para visualização de dados multivariados e mineração de dados, por exemplo, na análise de agrupamentos.

Mapas auto-organizáveis

Rede neural com duas camadas:

- Camada de entrada: recebe os padrões de entrada ou estímulos (no nosso caso, cada padrão de entrada é uma curva típica de dia útil de um cliente/rede)
- Camada de saída: grade de neurônio ou mapa de Kohonen (mapa topologicamente ordenado).

Camada de saída (grade de neurônios)



curva diária com demandas registradas a cada 15 minutos ($p = 96$ pontos)

Cada neurônio possui um 'endereço' na grade e todos recebem os mesmos sinais provenientes da entrada $x = \{ x_1, \dots, x_p \}$.

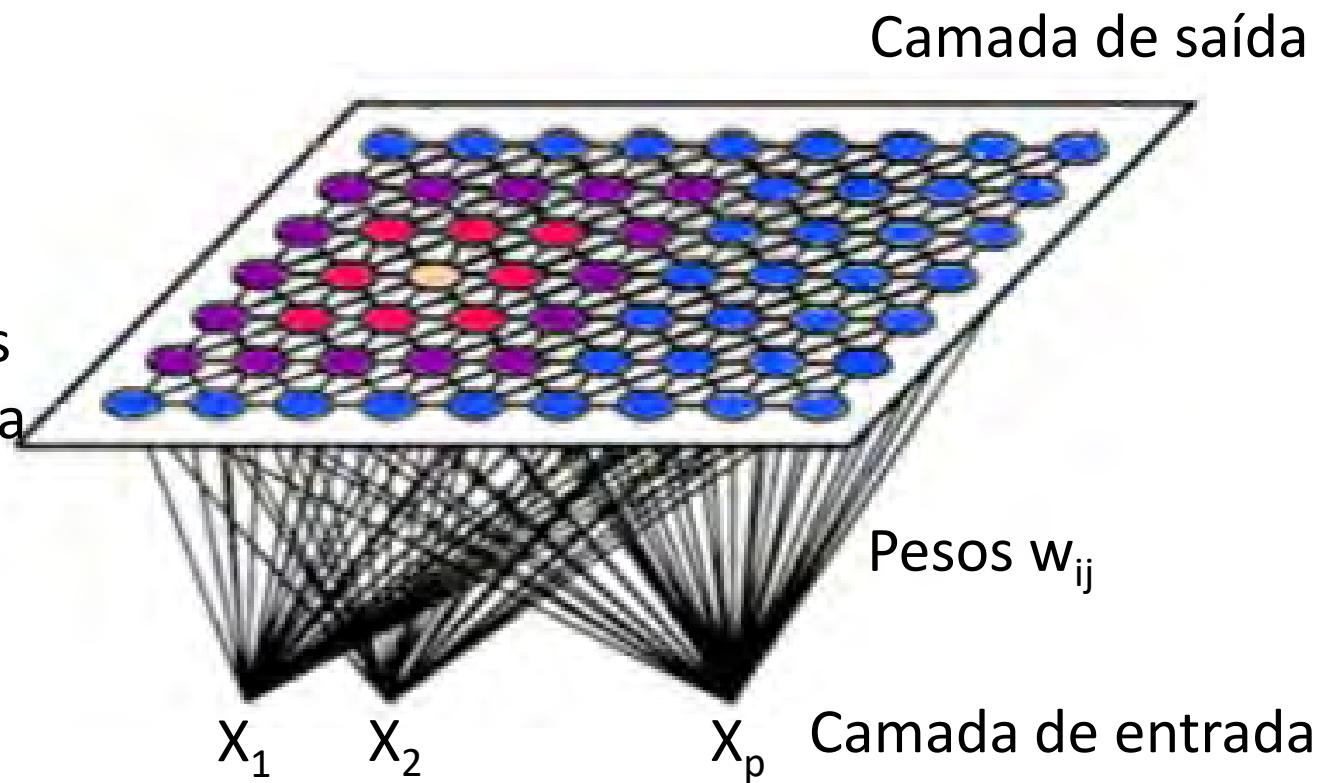
As entradas são ponderadas por pesos nas sinapses, cujos valores são determinados após o treinamento da rede.

Arquitetura do Mapa auto-organizável

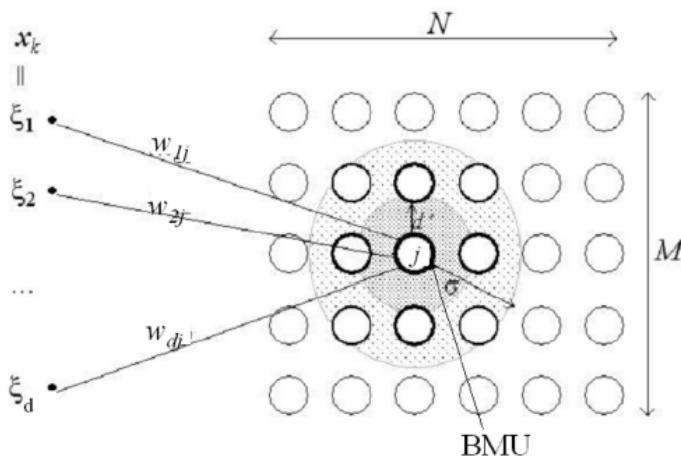
Em geral tem apenas duas camadas interconectadas por pesos sinápticos adaptáveis:

- camada de entrada com p neurônios
- camada de saída com q neurônios em uma grade bidimensional (mapa)

Cada entrada é
conectada com
todos os neurônios
da camada de saída



Mapas auto-organizáveis



Durante o treinamento da rede neural os pesos das sinapses (w) são ajustados de tal forma que os neurônios se especializam na detecção de um conjunto de padrões de entrada e se organizam topologicamente, fazendo com que os padrões detectados por um dado neurônio estejam relacionados com a posição do neurônio no reticulado.

Padrões de entrada semelhantes são detectados por neurônios vizinhos.

Resumindo, a rede auto-organizável projeta um conjunto de dados com N objetos caracterizados por p atributos em um plano (um mapa), de tal forma que na projeção a proximidade dos objetos no R^p seja preservada no R^2 .

Algortimo de treinamento

- 1) Inicialize a taxa de aprendizado, o tamanho da região de vizinhança e faça a Inicialização aleatória dos pesos.
- 2) Para cada padrão de treinamento (vetor $x_i \forall i=1,N$) faça
 - 2.1 Defina o neurônio vencedor
 - 2.2 Atualize os pesos do neurônio vencedor e dos vizinhos. Os pesos destes neurônios ficam mais próximos do padrão de entrada.
 - 2.3 Se o número de ciclos for múltiplo de N reduza a taxa de aprendizagem e a vizinhança do neurônio vencedor.
- 3) Repita o passo 2 até a convergência do mapa.

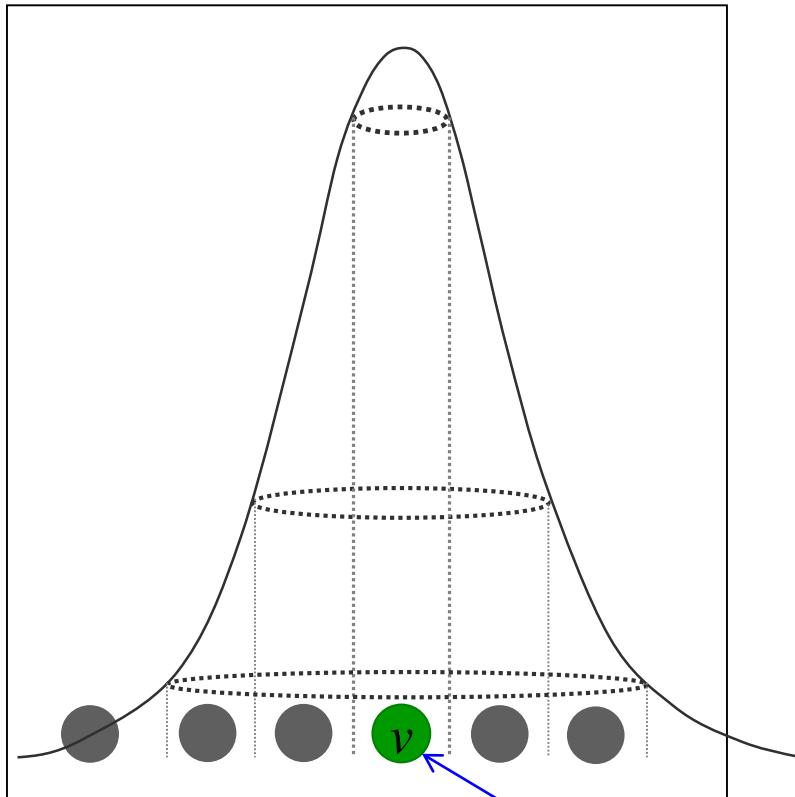
Duas fases: Ordenação e Convergência

Ordenação: Agrupa os neurônios em clusters (taxa de aprendizagem e região de vizinhança seguem trajetórias decrescentes)

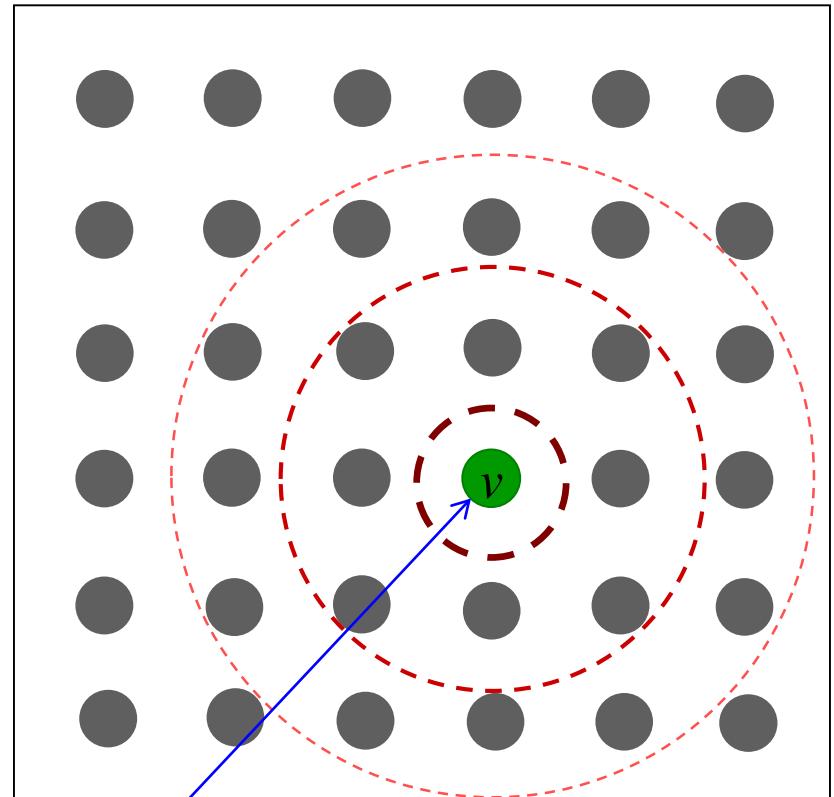
Convergência: Refinamento do mapa topológica construído na ordenação (valores muito reduzidos da taxa de aprendizagem e vizinhança)

Redução da região de vizinhança aos longo das iterações

Região de vizinhança



Região de vizinhança



Neurônio vencedor

Redução da região de vizinhança aos longo das iterações

Região de vizinhança gaussiana do neurônio vencedor

$$h_{vj}(t) = \exp\left(-\frac{\|\mathbf{r}_v - \mathbf{r}_j\|}{2\sigma^2(t)}\right)$$

distância entre o neurônio v vencedor e o neurônio j que está sendo atualizado

Define a largura da vizinhança e deve ser decrescente no tempo.

$$\sigma(t) = \sigma(0) \exp\left(-\frac{t}{\tau_1}\right) \rightarrow h_{vj}(t) \rightarrow 0 \text{ quando } t \rightarrow \infty$$

$$\tau_1 = \frac{\text{Número_de_Iterações}}{\log \sigma(0)}$$

Redução da taxa de aprendizagem

A taxa de aprendizagem decresce com o tempo, para que as adaptações sejam cada vez mais “finas”.

$$\alpha(t) = \alpha_0 \exp\left(-\frac{t}{\tau_2}\right),$$

número total de iterações

Ajuste dos pesos

$$\mathbf{w}_j(t+1) = \mathbf{w}_j(t) + \alpha(t)h_{vj}(t)[\mathbf{x} - \mathbf{w}_j(t)]$$

Diagram illustrating the weight update rule:

- $\mathbf{w}_j(t+1)$: Vetor peso atualizado
- $\mathbf{w}_j(t)$: Vetor peso anterior
- $\alpha(t)$: Taxa de aprendizagem
- $h_{vj}(t)$: Vizinhança
- $\mathbf{x} - \mathbf{w}_j(t)$: Adaptação
- Δw : Total weight change

The diagram shows the weight update rule for a neuron. It consists of several terms: the updated weight vector $\mathbf{w}_j(t+1)$, the previous weight vector $\mathbf{w}_j(t)$, the learning rate $\alpha(t)$, the neighborhood function $h_{vj}(t)$, the input vector \mathbf{x} , and the difference vector $\mathbf{x} - \mathbf{w}_j(t)$. A bracket above the equation indicates the total weight change Δw . Arrows point from each term to its corresponding label below the equation.

Esta regra move o vetor de peso do neurônio vencedor em direção ao padrão de entrada

Algortimo de treinamento

Ao final do treinamento os neurônios organizam-se em um mapa topologicamente ordenado, em que padrões semelhantes ativam neurônios vizinhos.

Após o treinamento os neurônios especializaram-se em detectar características dos padrões de entrada.

SOM no R

Package ‘kohonen’

September 4, 2015

Version 2.0.19

Title Supervised and Unsupervised Self-Organising Maps

Author Ron Wehrens

Maintainer Ron Wehrens <ron.wehrens@gmail.com>

Description Functions to train supervised and self-organising maps (SOMs). Also interrogation of the maps and prediction using trained maps are supported. The name of the package refers to Teuvo Kohonen, the inventor of the SOM.



Exemplo Tipologias de curvas de carga (Pessanha et al, 2015)

```
# Importação de dados
```

```
setwd("c:/curso R/aula5") # estabelece o diretório de trabalho
```

```
dados = read.csv2("curvas_de_carga.csv",sep=";",dec=".")header=T, row.names=1)
```

```
# dimensões da matriz de dados
```

```
dim(dados)
```

```
# gráfico dos perfis
```

```
matplot(matrix(seq(1,24,1),ncol=1),t(dados),type='l',ylab='PU da média',xlab='Horas')
```

```
# Mapa de Kohonen de tamanho 3 x 3
```

```
library(kohonen)
```

```
tamanho=3
```

```
dados=as.matrix(dados)
```

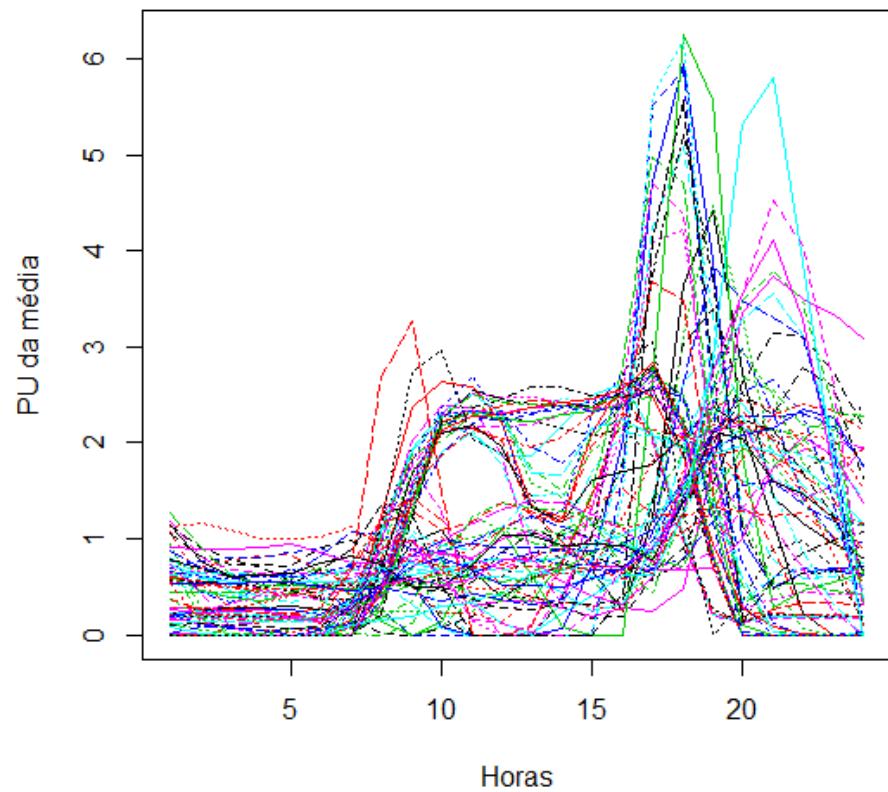
```
mapa=som(dados,grid=somgrid(tamanho,tamanho),rlen=10000,alpha=c(0.05,0.01))
```

```
windows()
```

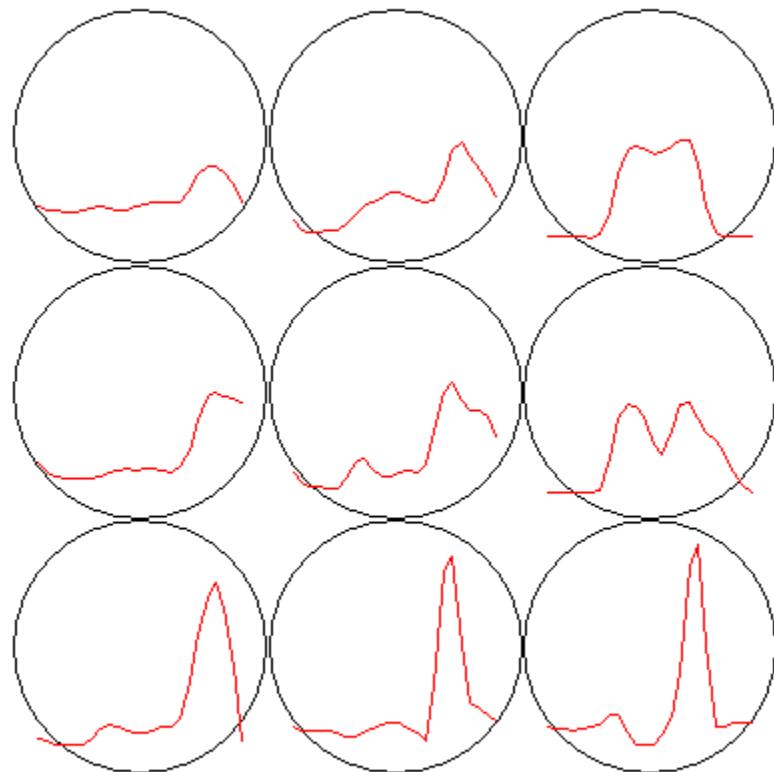
```
plot(mapa)
```

Exemplo Tipologias de curvas de carga (Pessanha et al, 2015)

Perfis de carga



Mapa 3 x 3



Exemplo Tipologias de curvas de carga (Pessanha et al, 2015)

```
ncurvas = length(mapa$unit.classif)
classe=mapa$unit.classif
numel=numeric(0)
for (i in 1:(tamanho*tamanho)) {
  indice=which(classe==i)
  numel=c(numel,length(indice))
  if (i==1) {
    if (length(indice)>1) {
      tipo=apply(dados[indice,],2,"mean")
    } else {
      tipo=dados[indice,]
    }
  } else {
    if (length(indice)>1) {
      vetor=apply(dados[indice,],2,"mean")
    } else {
      vetor=dados[indice,]
    }
    tipo=rbind(tipo,vetor)
  }
}
```

Cálculo dos centroides em cada neurônio (cluster)

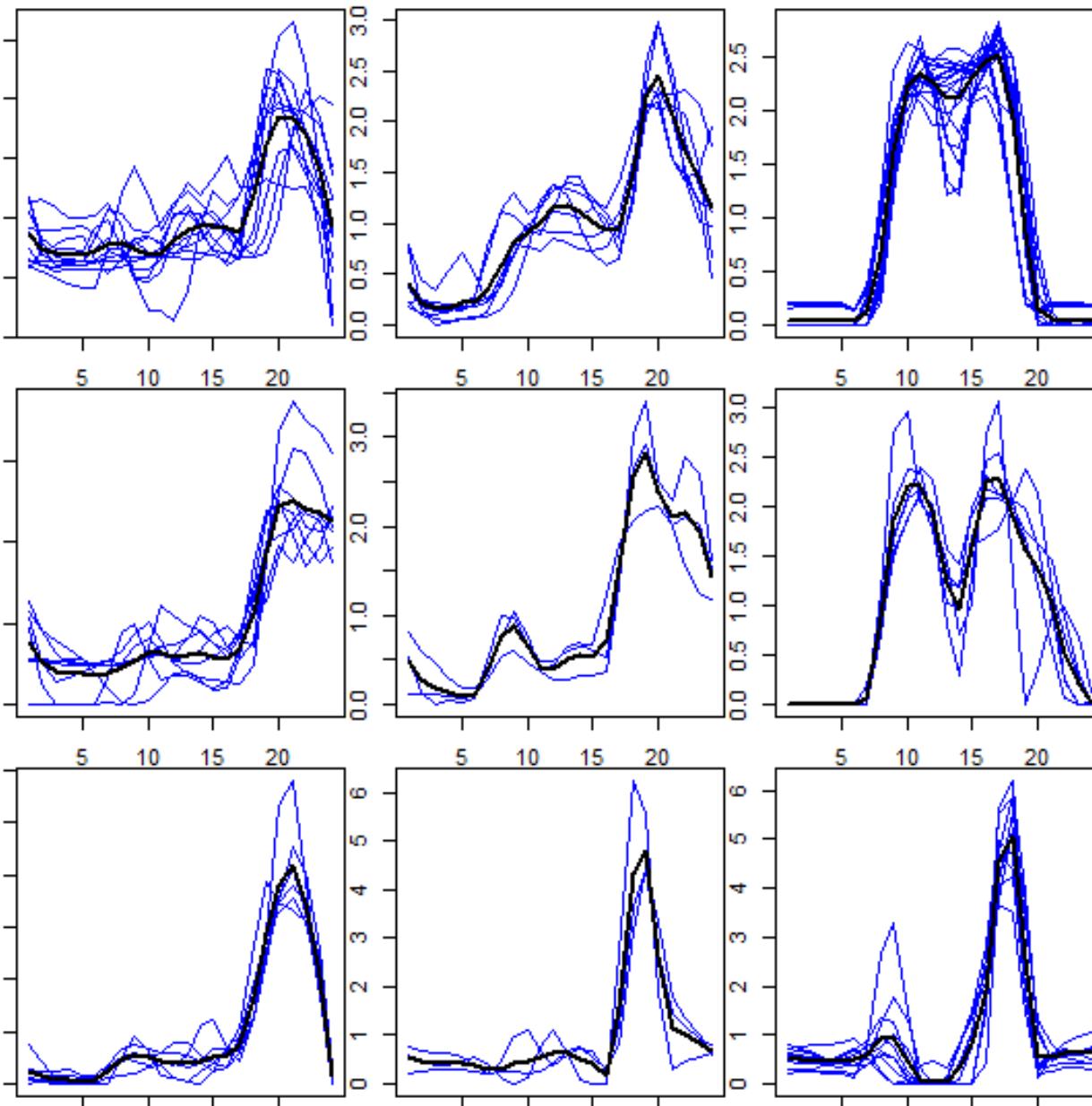
Os centroides (perfis horários) são as linhas da matriz tipo

Exemplo Tipologias de curvas de carga (Pessanha et al, 2015)

```
windows()
par(mfrow=c(tamanho,tamanho))
par(mar=c(1,1,1,1))
indices<-apply(matrix(seq(tamanho^2:1),tamanho,tamanho),1,rev)
vecindices<-matrix(t(indices),1,tamanho^2)
for (i in 1:tamanho^2) {
  aux<-which(classe==vecindices[i])
  teto<-max(dados[aux,])
  piso<-min(dados[aux,])
  ncurvas<-length(aux)
  if (ncurvas>0) {
    plot(tipo[vecindices[i],],type = "l",ylim=c(piso,teto),ylab= "", xlab = "",lty = "solid",col="black",lwd=2)
    for (j in 1:ncurvas) {
      lines(dados[aux[j],],col = "blue", lty = 1,ylab = "", xlab = "")
    }
    lines(tipo[vecindices[i],],type = "l",main = paste("cluster ",vecindices[i]),ylab = "", xlab = "",lty = "solid",col="black",lwd=2)
  }
}
```

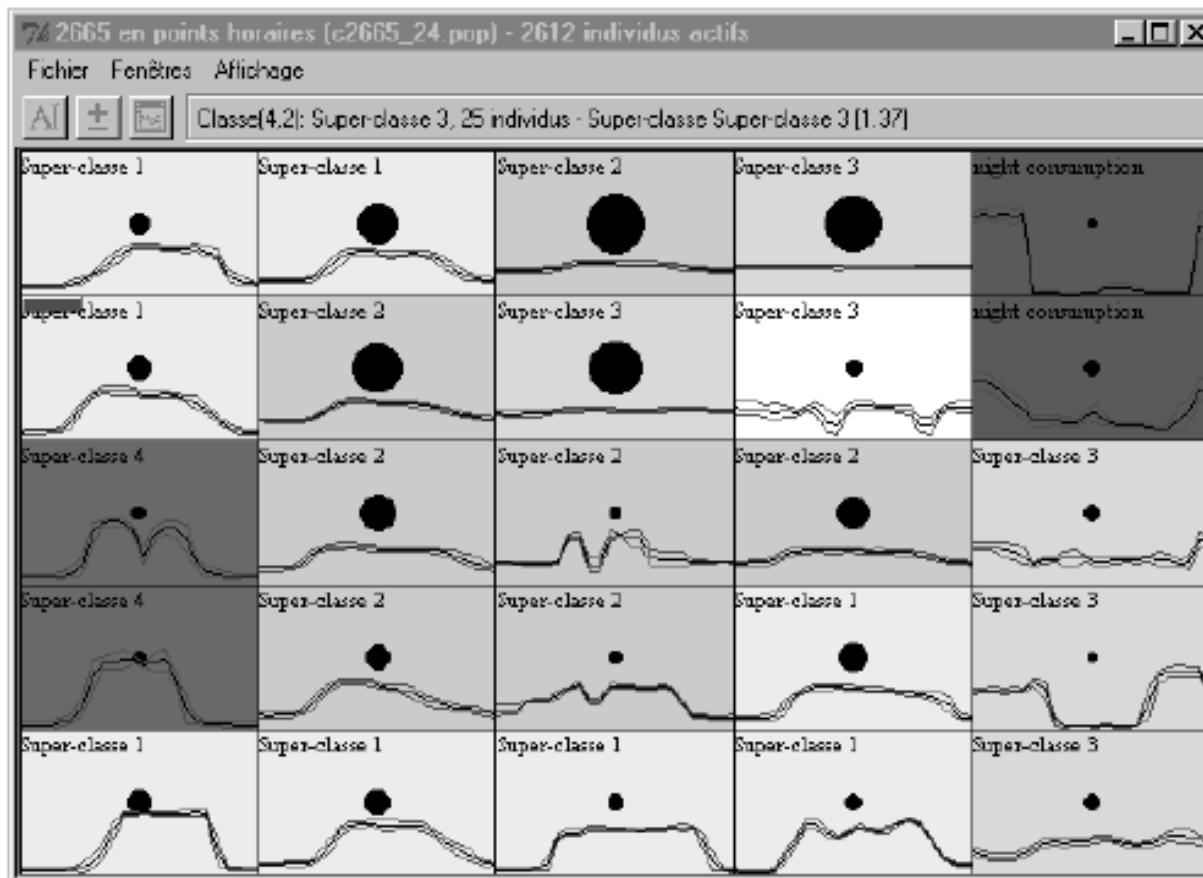
Gráfico das curvas em cada cluster
e dos centroides

Exemplo Tipologias de curvas de carga (Pessanha et al, 2015)



Mapas auto-organizáveis

- O “Courboscope” desenvolvido pela Electricité de France (EDF) é um bom exemplo de sistema computacional para construção de tipologias de curva de carga baseado em Mapa de Kohonen.



Mapas auto-organizáveis

Aplicação na segmentação das 45 maiores distribuidoras do SEB.

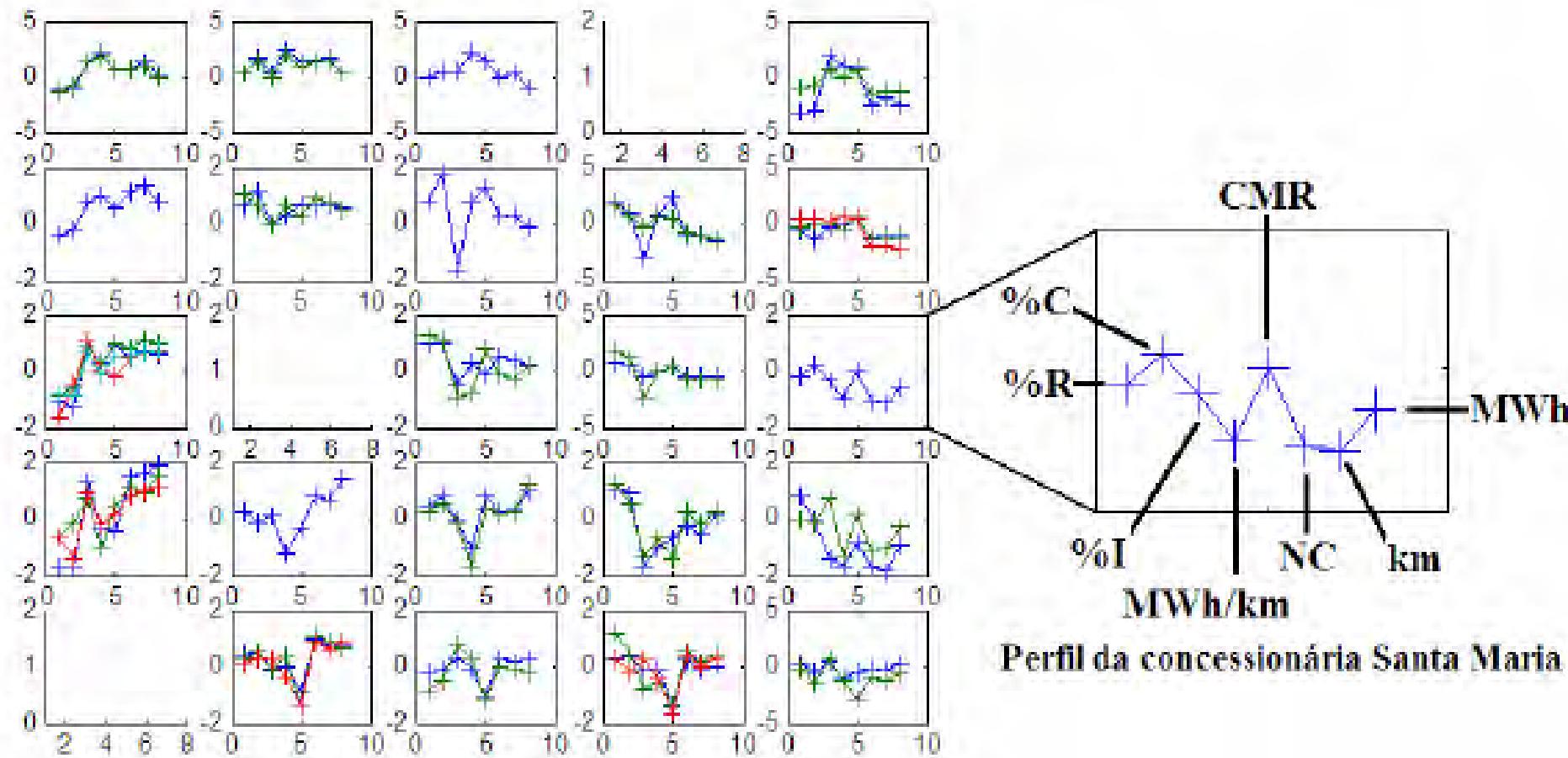
Cada concessionária foi caracterizada por 8 atributos, todos referentes ao ano 2000, que caracterizam três dimensões:

Dimensões	Variáveis
Estrutura	<ul style="list-style-type: none">participação da classe residencial no consumo (%R)participação da classe comercial no consumo (%C)participação da classe industrial no consumo (%I)
Concentração	<ul style="list-style-type: none">tamanho da rede de distribuição (km)carregamento da rede (MWh/km)consumo médio residencial (CMR)
Tamanho	<ul style="list-style-type: none">nº de unidades consumidoras (NC)energia elétrica distribuída (MWh)

Nesta segmentação foi considerado um mapa de Kohonen 5x5, cujo treinamento efetuado pelo MATLAB convergiu após 20.000 iterações.

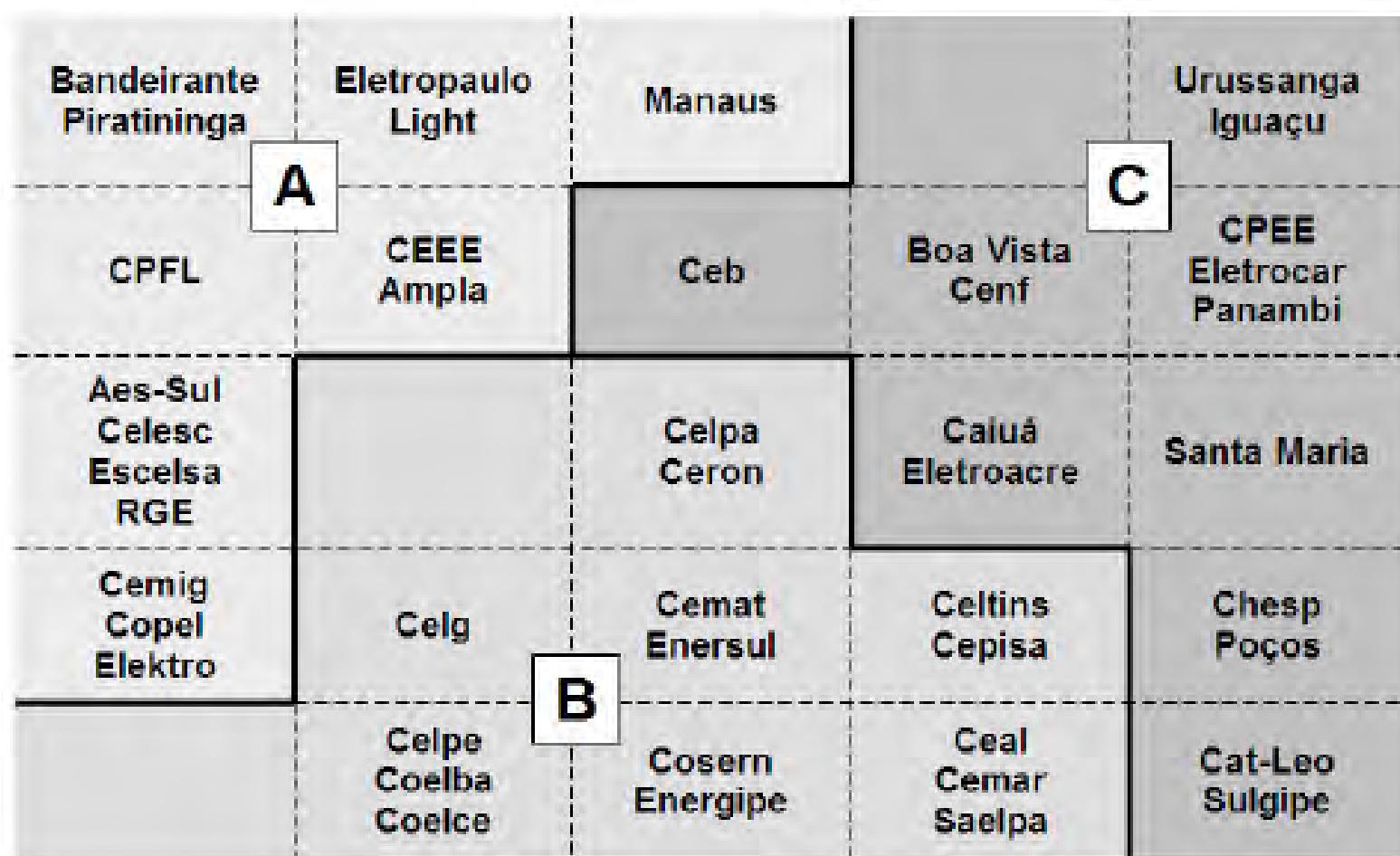
Mapas auto-organizáveis

Mapa 5x5, cada quadrado representa um neurônio e cada curva representa o vetor de atributos de uma distribuidora.



Mapas auto-organizáveis

Mapa 5x5 com os nomes das empresas e proposta de agregação dos neurônios nas superclasses A,B e C.



Comparação das técnicas de análise de agrupamentos na construção de tipologias de curvas de carga

Objetivo: Classificar um conjunto com 125 medições de curvas de carga de consumidores de baixa tensão

Cada curva representa o comportamento do consumidor em um dia útil e tem 96 pontos (valores de demanda amostrados a cada 15 minutos)

Cada curva foi dividida pelo seu valor médio para que a classificação seja efetuada com base no perfil da curva e não nos valores absolutos da demanda.

Nº de clusters igual a 8, o máximo recomendado pelo Programa de Revisão Tarifária (1994)

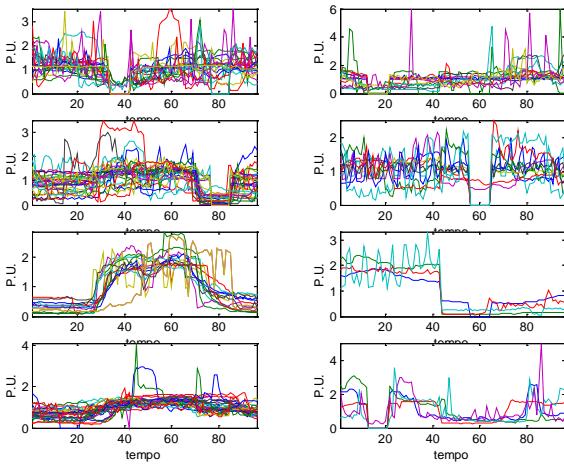
Foram comparados cinco algoritmos:

- Ward (TARDIST)
- k-Means MATLAB
- NUDYC-DESCR2 (SNACC)
- FCM (Fuzzy Clustering Method) MATLAB
- Mapa Auto-organizável (rede neural) MATLAB

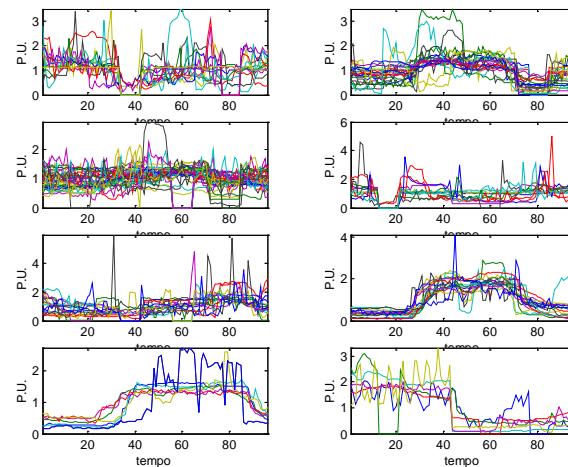


Comparação das técnicas de análise de agrupamentos na construção de tipologias de curvas de carga

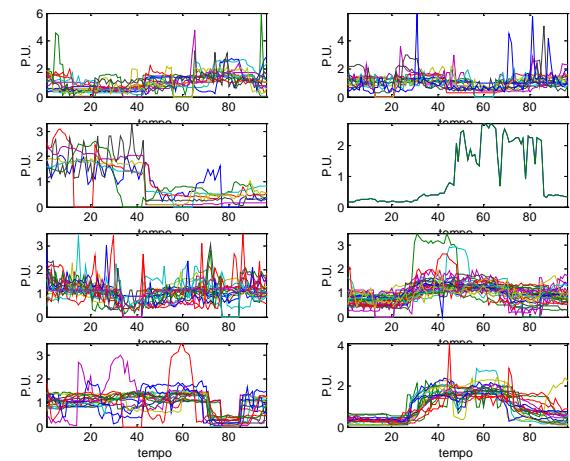
WARD



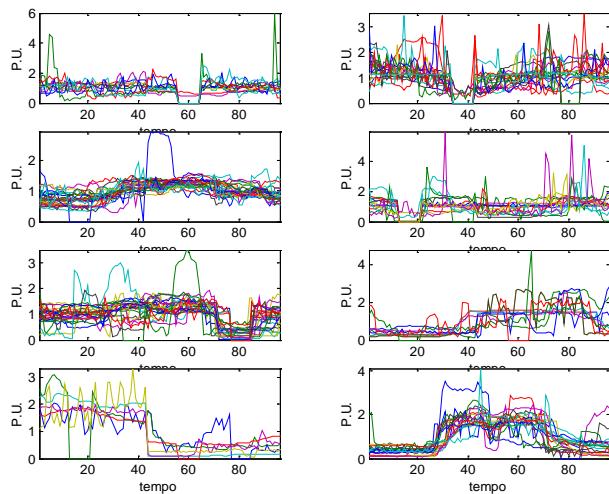
NUDYC-DESCR2



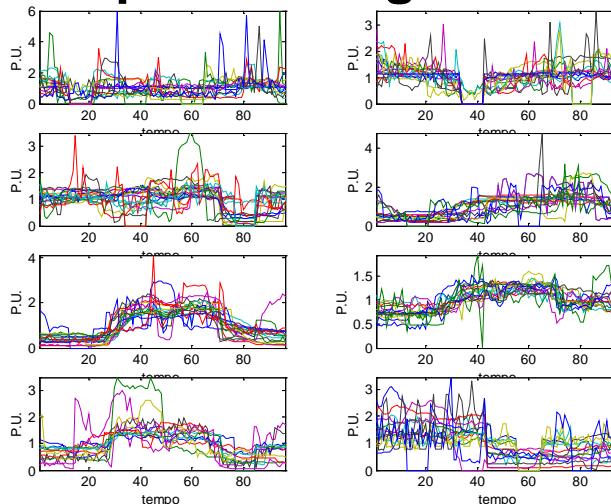
K - MEANS



FCM



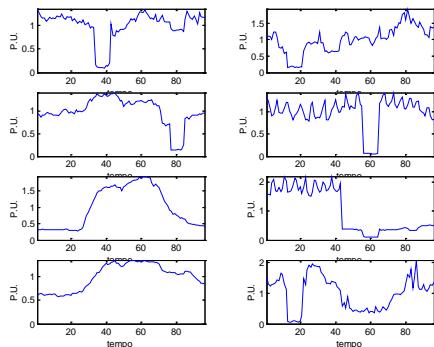
Mapa Auto-organizável



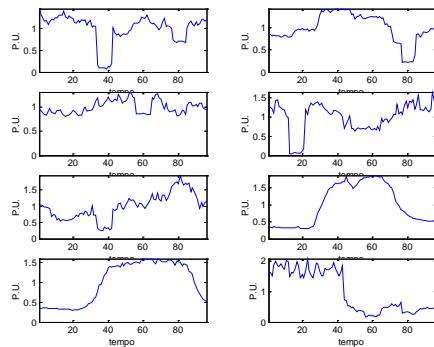


Comparação das técnicas de análise de agrupamentos na construção de tipologias de curvas de carga

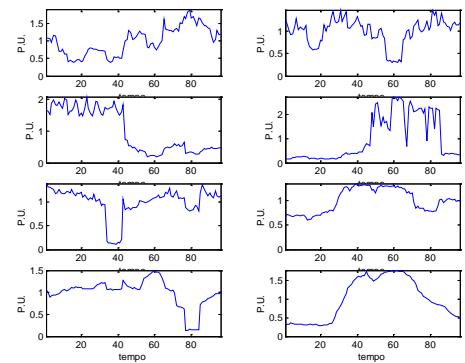
WARD



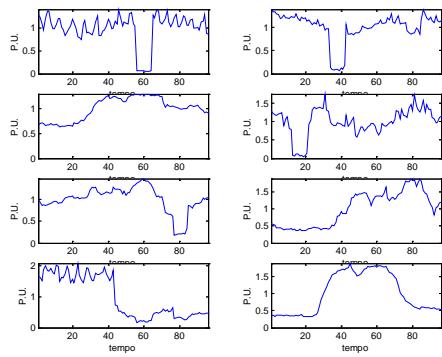
NUDYC-DESCR2



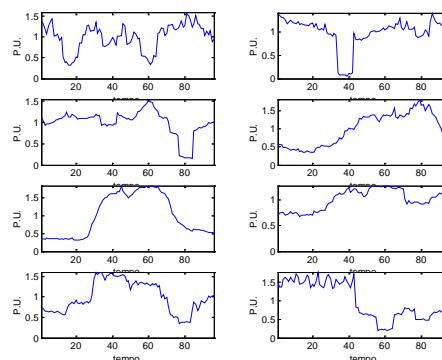
K - MEANS



FCM



Mapa Auto-organizável



Pacote clvalid (Brock et al, 2008)



Journal of Statistical Software

March 2008, Volume 25, Issue 4.

<http://www.jstatsoft.org/>

clValid: An R Package for Cluster Validation

Guy Brock
Universtiy of Louisville

Vasyl Pihur
Universtiy of Louisville

Susmita Datta
Universtiy of Louisville

Somnath Datta
Universtiy of Louisville

<http://www.jstatsoft.org/v25/i04/paper>

<http://cran.r-project.org/web/packages/clValid/clValid.pdf>



Pacote clValid

Disponibiliza uma variedade de métodos de análise de agrupamentos e métricas de validação e estabilidade dos clusters

Medidas de validação e estabilidade

Avaliam se os clusters são compactos, bem separados e estáveis.

Medidas de validação:

- 1) **conectividade**: relativa ao grau de vizinhança entre objetos em um mesmo cluster, varia entre 0 e infinito e quanto menor melhor.
- 2) **silhueta**: homogeneidade interna, assume valores entre -1 e 1 e quanto mais próximo de 1 melhor.
- 3) **índice de Dunn**: quantifica a separação entre os agrupamentos, assume valores entre 0 e 1 e quanto maior melhor.

Medidas de estabilidade:

- 1) **APN - *average proportion of non-overlap***: proporção média de observações não classificadas no mesmo cluster nos casos com dados completos e incompletos. Assume valor no intervalo [0,1], próximos de 0 indicam agrupamentos consistentes.
- 2) **AD - *average distance***: distância média entre observações classificadas no mesmo cluster nos casos com dados completos e incompletos. Assume valores não negativos, sendo preferíveis valores próximos de zero.
- 3) **ADM - *average distance between means***: distância média entre os centroides quando as observações estão em um mesmo cluster. Assume valores não negativos, sendo preferíveis valores próximos de zero.
- 4) **FOM - *figure of merit***: medida do erro cometido ao usar os centroides como estimativas das observações na coluna removida. Assume valores não negativos, sendo preferíveis valores próximos de zero.

Índice de Dunn

É a razão da menor distância entre duas observações em *clusters* distintos pela maior distância entre dois *clusters*.

$$Dunn = \frac{\text{menor distância entre duas observações em cluster distintos}}{\text{maior distância entre dois clusters}}$$

O índice de Dunn assume valores no intervalo [0,1] e visando obter agrupamentos bem separados, o índice de Dunn deve ser maximizado.

Medida de Validação: Conectividade

Considere um conjunto com N elementos descritos por p variáveis e classificados em K clusters C_1, \dots, C_k .

Seja L o número de vizinhos de uma observação x_i , por exemplo, se $L=1$, apenas o elemento mais próximo de x_i é considerado seu vizinho. Denotando os L vizinhos de x_i por $nn_{i,j}, j=1, L$, define-se a seguinte variável $z_{i,j}$:

$z_{i,j} = 0$, se x_i e $nn_{i,j}$ pertencem ao mesmo cluster

$z_{i,j} = 1/j$, se x_i e $nn_{i,j}$ não pertencem ao mesmo cluster

Assim, a medida de conectividade para uma solução com K clusters é dada por:

$$\text{Conectividade} = \sum_{i=1}^N \sum_{j=1}^L z_{i,j}$$

A conectividade varia entre 0 e infinito e na partição ideal em k agrupamentos deve ser minimizada.

Medida de Validação: Silhueta

O comprimento da silhueta é a média das siluetas das observações.

A silhueta de uma observação x_i , permite avaliar se a mesma foi bem classificada entre os K agrupamentos possíveis.

A silhueta de uma observação x_i classificada em um cluster k é dada por:

$$S_i = \frac{b_i - a_i}{\max(b_i, a_i)}$$

a_i = média das distâncias entre x_i e as observações no mesmo cluster k

b_i = média das distâncias entre x_i e as observações no agrupamento vizinho mais próximo do cluster k .

A silhueta de uma observação assume valores no intervalo $[-1, +1]$, sendo desejável um valor próximo de $+1$, indicando que a observação x_i está mais próxima das observações do *cluster* em que ela foi alocada e não do *cluster* vizinho.

Uma boa partição dos objetos em K clusters deve maximizar a silhueta.

Medidas de estabilidade: APN, AD, ADM, FOM

Tais medidas são calculadas a partir dos resultados gerados pela análise de agrupamentos do conjunto completo de dados (p variáveis) e dos resultados da análise de agrupamentos aplicada em conjuntos de dados sem uma das variáveis ($p-1$ variáveis), na qual uma coluna é removida por vez

Exemplo 7 Classificação das distribuidoras (continuação)

```
valida=clValid(empresas,2:6,clMethods=c("hierarchical","kmeans"),validation="internal")
summary(valida)
```

```
Clustering Methods:
  hierarchical kmeans
```

```
Cluster sizes:
  2 3 4 5 6
```

```
Validation Measures:
```

		2	3	4	5	6
hierarchical	Connectivity	6.4004	8.8210	12.8464	18.1500	20.5706
	Dunn	0.2741	0.2741	0.1437	0.2618	0.2618
	Silhouette	0.7751	0.7256	0.7218	0.6833	0.6753
kmeans	Connectivity	7.0810	11.4940	16.0909	20.4333	28.3048
	Dunn	0.2034	0.0948	0.0430	0.1212	0.0957
	Silhouette	0.7715	0.7007	0.7208	0.7106	0.6667

```
Optimal Scores:
```

	Score	Method	Clusters
Connectivity	6.4004	hierarchical	2
Dunn	0.2741	hierarchical	2
Silhouette	0.7751	hierarchical	2

Medidas de validação

Exemplo 7 Classificação das distribuidoras (continuação)

```
valida=clValid(empresas,2:6,clMethods=c("hierarchical","kmeans"),validation="stability")
summary(valida)
```

Clustering Methods:
hierarchical kmeans

Cluster sizes:
2 3 4 5 6

Validation Measures:

		2	3	4	5	6
hierarchical	APN	0.0045	0.0229	0.0139	0.0136	0.0426
	AD	5158060.1721	4490404.7244	2264247.3897	1858919.5068	1695933.4000
	ADM	287645.1235	318333.8310	218983.1538	210165.3194	240389.3293
	FOM	825602.4534	553429.1330	451316.3516	433025.5158	316203.6791
kmeans	APN	0.0171	0.0195	0.0289	0.0327	0.0357
	AD	4935621.9749	3467652.2243	2114200.9002	1570133.1221	1401266.3819
	ADM	276714.6096	227193.2554	231269.7065	193599.4717	216750.5940
	FOM	784134.7403	651545.6608	451099.5523	341061.9743	370141.4422

Optimal Scores:

	Score	Method	Clusters
APN	0.0045	hierarchical	2
AD	1401266.3819	kmeans	6
ADM	193599.4717	kmeans	5
FOM	316203.6791	hierarchical	6

Medidas de estabilidade

Exemplo 8 Tipologias de curvas de carga (Continuação)

```
library(clValid) # carrega o pacote clValid
```

```
# medidas de validação das soluções fornecidas pelos métodos hierárquico e
```

```
# Kmeans de 2 a 9 clusters
```

```
intern=clValid(dados,2:9,clMethods=c("hierarchical","kmeans"),validation="internal")
```

```
summary(intern)
```

```
> summary(intern)
```

```
Clustering Methods:  
hierarchical kmeans
```

```
Cluster sizes:  
2 3 4 5 6 7 8 9
```

```
Validation Measures:
```

	2	3	4	5	6	7	8	9
--	---	---	---	---	---	---	---	---

hierarchical	Connectivity	1.9151	2.1262	5.1940	8.1913	17.1262	20.9214	22.1516	25.1917
	Dunn	0.3685	0.3581	0.3581	0.3494	0.3971	0.4422	0.4566	0.4566
	Silhouette	0.3535	0.5530	0.5232	0.4923	0.4285	0.4265	0.4151	0.3370
kmeans	Connectivity	1.9151	2.1262	9.1560	16.7853	26.1714	28.8060	30.0361	38.2671
	Dunn	0.3685	0.3581	0.2931	0.2570	0.3624	0.3464	0.4040	0.2411
	Silhouette	0.3535	0.5530	0.4930	0.4562	0.4429	0.4166	0.4125	0.3543

Medidas de validação

```
Optimal Scores:
```

	Score	Method	Clusters
Connectivity	1.9151	hierarchical	2
Dunn	0.4566	hierarchical	8
Silhouette	0.5530	hierarchical	3

Exemplo 8 Tipologias de curvas de carga (Continuação)

estabilidade das soluções fornecidas pelos métodos hierárquico e Kmeans

de 2 a 9 clusters

```
estavel=clValid(dados,2:9,clMethods=c("hierarchical","kmeans"),validation="stability")
summary(estavel)
```

```
> summary(estavel)
```

```
Clustering Methods:
  hierarchical kmeans
```

```
Cluster sizes:
  2 3 4 5 6 7 8 9
```

```
Validation Measures:
```

	2	3	4	5	6	7	8	9
hierarchical	APN 0.0344 0.0032 0.0063 0.0254 0.0163 0.0217 0.0149 0.0197							
	AD 4.4765 2.7432 2.6538 2.5725 2.3175 2.1856 2.0783 2.0289							
	ADM 0.2597 0.0249 0.0377 0.1061 0.1340 0.1284 0.0693 0.0975							
	FOM 0.7192 0.4351 0.4129 0.4082 0.3851 0.3637 0.3490 0.3483							
kmeans	APN 0.0379 0.0000 0.0114 0.0430 0.0112 0.0337 0.0238 0.0361							
	AD 4.4798 2.7342 2.6141 2.5267 2.2085 2.1479 2.0446 1.9030							
	ADM 0.2638 0.0000 0.0612 0.1785 0.0545 0.1321 0.0856 0.1810							
	FOM 0.7192 0.4173 0.4128 0.3927 0.3577 0.3578 0.3499 0.3398							

```
Optimal Scores:
```

	Score	Method	Clusters
APN	0.0000	kmeans	3
AD	1.9030	kmeans	9
ADM	0.0000	kmeans	3
FOM	0.3398	kmeans	9

Medidas de estabilidade

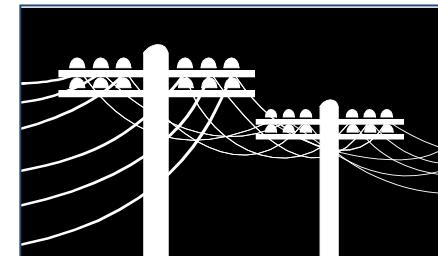


Medidas de Validação no R

Para fins de regulação econômica da distribuição de energia elétrica é interessante segmentar as empresas distribuidoras em clusters de forma a permitir análises comparativas entre empresas semelhantes.

Considere uma matriz de dados formada por 59 concessionárias de distribuição que atuam no setor elétrico brasileiro, cada uma descrita por 9 variáveis (Fonte: Aneel):

- Mercado atendido (MWh)
- Número de consumidores
- Tamanho da rede de distribuição
- Densidade de consumidores (consumidores/km de rede)
- Consumo por unidades consumidora – CPC (MWh/consumidor)
- Índice de complexidade no combate às perdas não técnicas
- Composição do mercado por classe de tendão BT, MT e AT.



Medidas de Validação Interna no R

```
valida=clValid(empresas,2:6,clMethods=c("hierarchical","kmeans"),validation="internal")
summary(valida)
```

```
Clustering Methods:
hierarchical kmeans

Cluster sizes:
2 3 4 5 6

Validation Measures:
              2      3      4      5      6
hierarchical
  Connectivity 6.4004 8.8210 12.8464 18.1500 20.5706
  Dunn         0.2741 0.2741 0.1437 0.2618 0.2618
  Silhouette   0.7751 0.7256 0.7218 0.6833 0.6753
kmeans
  Connectivity 7.0810 11.4940 16.0909 20.4333 28.3048
  Dunn         0.2034 0.0948 0.0430 0.1212 0.0957
  Silhouette   0.7715 0.7007 0.7208 0.7106 0.6667

Optimal Scores:
          Score Method Clusters
Connectivity 6.4004 hierarchical 2
Dunn         0.2741 hierarchical 2
Silhouette   0.7751 hierarchical 2
```



Gráfico da silhueta

Função disponível no pacote cluster

```
silhueta = silhouette(cutree(saida_Ward,k=2),dist(dados_padron))
plot(silhueta,main="")
summary(silhueta)
```

Individual silhouette widths:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-0.2467	0.1185	0.2482	0.2217	0.3580	0.4648

↑
Comprimento médio da
silhueta

(maior silhueta média
ocorre na solução com
dois agrupamentos)

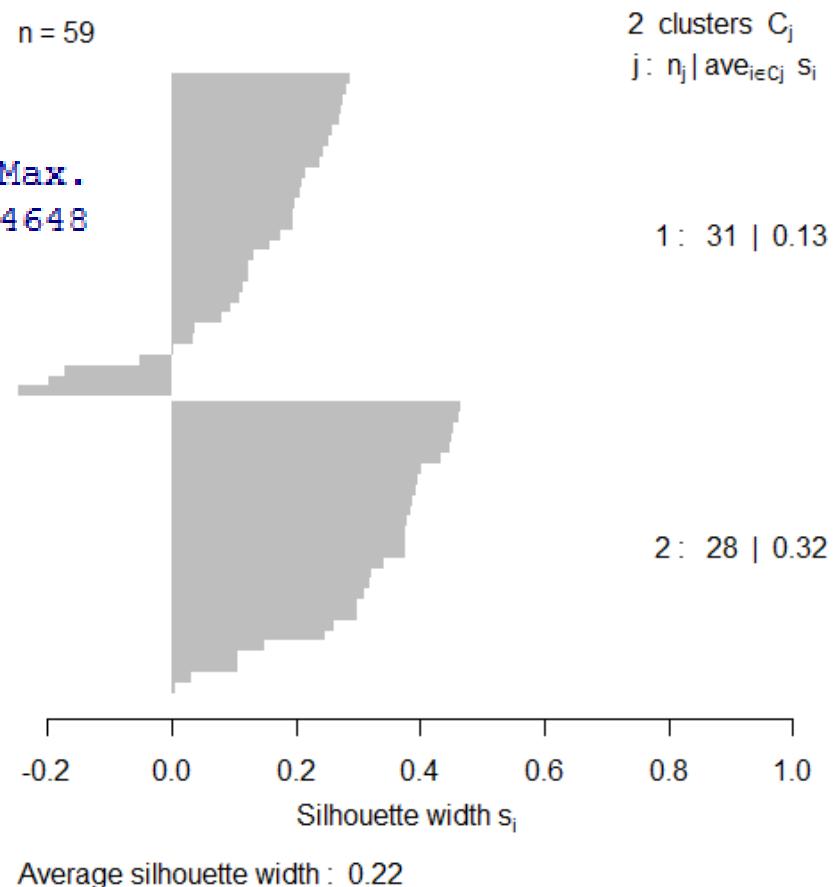




Gráfico da silhueta

Função disponível no pacote cluster

```
silhueta = silhouette(cutree(saida_Ward,k=3),dist(dados_padron))
```

```
plot(silhueta,main="")
```

```
summary(silhueta)
```

Individual silhouette widths:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-0.37730	0.08392	0.23810	0.21520	0.37860	0.59190

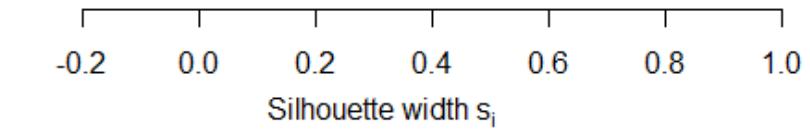
n = 59

3 clusters C_j
 $j : n_j | \text{ave}_{i \in C_j} s_i$

1 : 31 | 0.06

2 : 19 | 0.43

3 : 9 | 0.30



Average silhouette width : 0.22



Gráfico da silhueta

Função disponível no pacote cluster

```
silhueta = silhouette(cutree(saida_Ward,k=4),dist(dados_padron))
```

```
plot(silhueta,main="")
```

```
summary(silhueta)
```

Individual silhouette widths:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-0.23590	0.08545	0.24110	0.21600	0.36850	0.51250

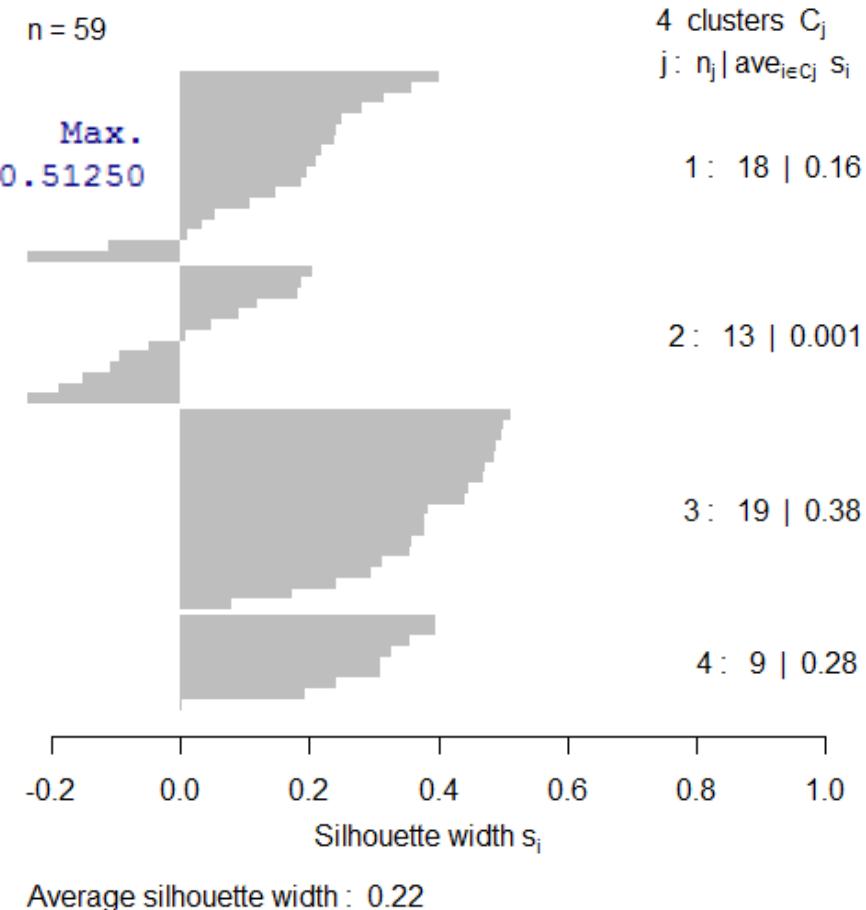




Gráfico da silhueta

Função disponível no pacote cluster

```
silhueta = silhouette(cutree(saida_Ward,k=5),dist(dados_padron))
```

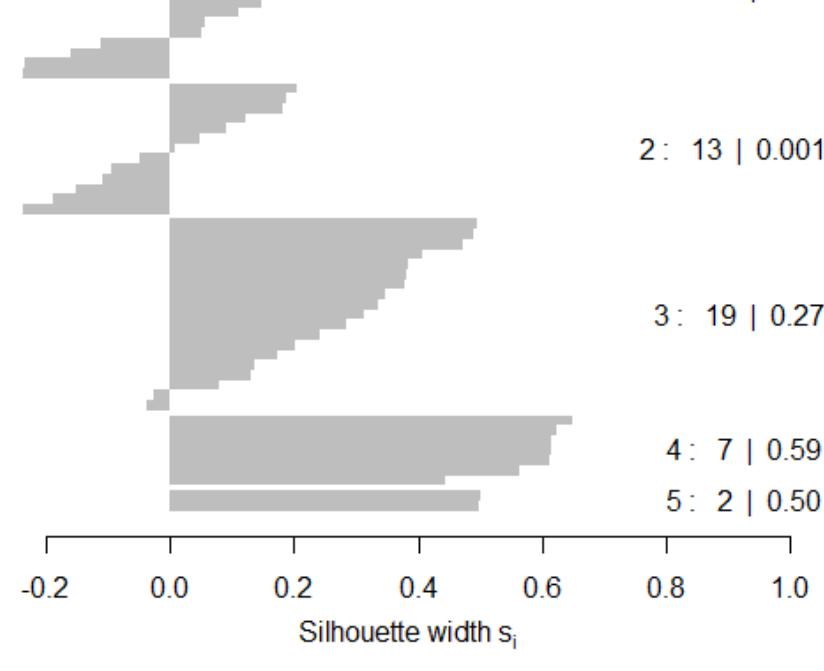
```
plot(silhueta,main="")
```

```
summary(silhueta)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-0.23660	0.05332	0.20200	0.21150	0.38220	0.64860

n = 59

5 clusters C_j
 $j : n_j | \text{ave}_{i \in C_j} s_i$
1 : 18 | 0.12



Validação da estabilidade R

```
valida=clValid(empresas,2:6,clMethods=c("hierarchical","kmeans"),validation="stability")
summary(valida)
```

```
Clustering Methods:
  hierarchical kmeans

Cluster sizes:
  2 3 4 5 6

Validation Measures:
          2           3           4           5           6
hierarchical APN      0.0045      0.0229      0.0139      0.0136      0.0426
                  AD  5158060.1721  4490404.7244  2264247.3897  1858919.5068  1695933.4000
                  ADM 287645.1235  318333.8310  218983.1538  210165.3194  240389.3293
                  FOM 825602.4534  553429.1330  451316.3516  433025.5158  316203.6791
kmeans       APN      0.0171      0.0195      0.0289      0.0327      0.0357
                  AD  4935621.9749  3467652.2243  2114200.9002  1570133.1221  1401266.3819
                  ADM 276714.6096  227193.2554  231269.7065  193599.4717  216750.5940
                  FOM 784134.7403  651545.6608  451099.5523  341061.9743  370141.4422
```

Optimal Scores:

	Score	Method	Clusters
APN	0.0045	hierarchical	2
AD	1401266.3819	kmeans	6
ADM	193599.4717	kmeans	5
FOM	316203.6791	hierarchical	6

Previsão de carga

Previsões do perfil de carga em bases horárias para um dia à frente constituem informações essenciais para a programação diária da operação do Sistema Interligado Nacional.

Cronograma para envio da previsão de carga para a elaboração da programação diária da operação eletroenergética.

Dia de elaboração da programação (dia em que a previsão deverá ser disponibilizada)	Carga prevista para o dia	Carga prevista para o dia (antecipação da programação)
2ª feira	3ª feira	4ª feira
3ª feira	4ª feira	5ª feira
4ª feira	5ª feira	6ª feira
5ª feira	6ª feira e sábado	domingo
6ª feira	domingo e 2ª feira	3ª feira

Fonte: Submódulo 5.4 Consolidação da previsão de carga para a programação diária da operação eletroenergética e para a programação de intervenções em instalações da rede de operação

Previsão de carga

A metodologia porposta explora as similaridades entre os perfis de carga.

As similaridades permitem construir previsores simplificados, porém com razoável capacidade preditiva.

Na metodologia proposta são empregadas técnicas de aprendizado supervisionado e não supervisionado para quantificar a similaridade entre os perfis de carga e estabelecer uma previsão.

Previsão de carga

Seja $D(h,d)$ a demanda em uma hora h em um dia d .

A demanda pode ser expressa da seguinte maneira:

$$D(h,d) = D(d) \cdot \frac{D(h,d)}{D(d)} \quad \forall h = 1,24$$

Demanda média do dia d

Previsão determinada por uma rede neural feedforward

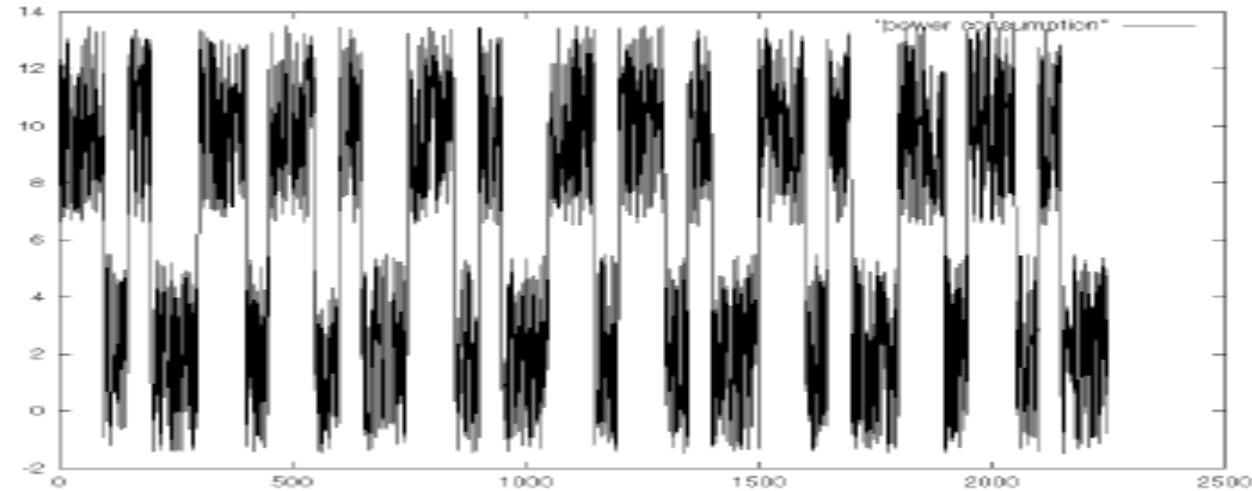
Perfil horário da carga normalizado pela demanda média diária

Previsão determinada pelo algoritmo de Wang Mendel aplicado na interpretação do mapa gerado por uma rede SOM

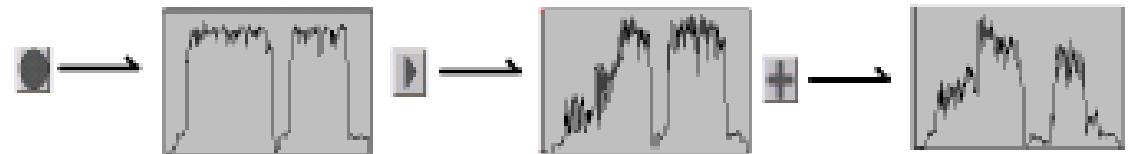
A rede neural *feedforward* e o algoritmo de Wang Mendel são técnicas de aprendizagem superviosanada, enquanto a rede SOM emprega o aprendizado não supervisionado.

Previsão de carga

Série temporal da carga formada pela sequência dos perfis padronizados



Cluster identificados pela rede SOM e marcados por símbolos (números)



Representação simbólica



Hébrail, G. 14th Computational Statistics Symposium, Utrecht, Netherlands, 2000

Experimento computacional

Dados: série da carga horária no Submercado Nordeste no período de 1/3/2010 até 31/12/203 (1402 dias).

Período de treinamento: 1/3/2010 até 30/9/2013 (1310 dias).

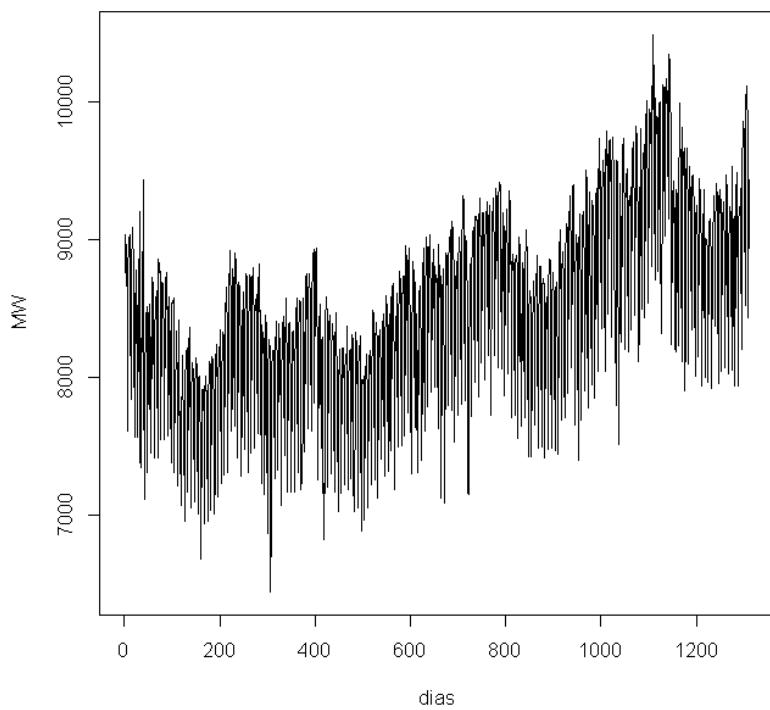
Período de validação: 1/10/2013 e 31/12/2013 (92 dias).

Todos as etapas da metodologia foram implementadas em ambiente R (R Core Team, 2014), um software livre (<https://www.r-project.org/>).

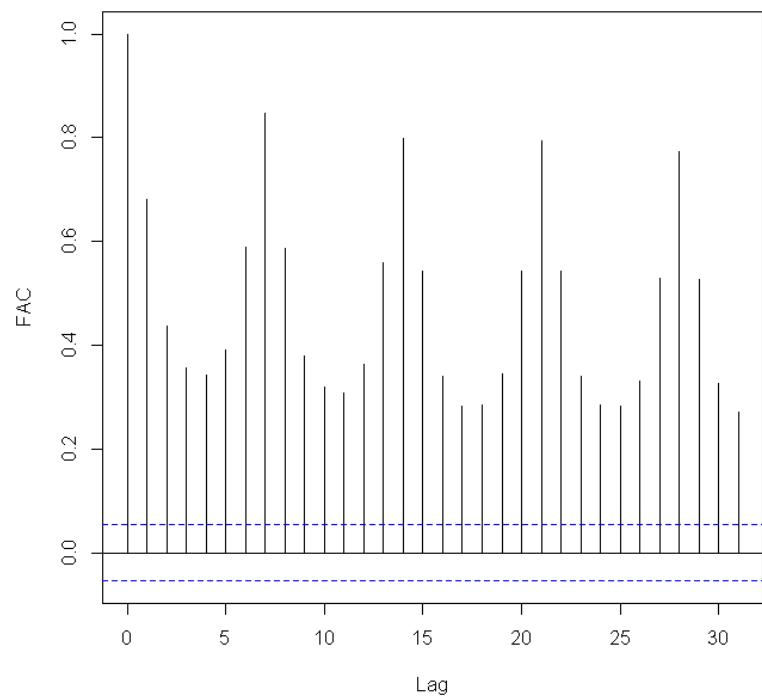


Experimento computacional

Série temporal da demanda média diária
MW



Função de autocorrelação (FAC) das diferenças da demanda média diária



Rede neural artificial para previsão da demanda média diária

Demandâna média dos dias anteriores

d-1, d-7, d-14 e d-21

Mês do dia d, 11 variáveis dummies

Dia da semana, 6 variáveis dummies

Tipo do dia, 4 variáveis dummies

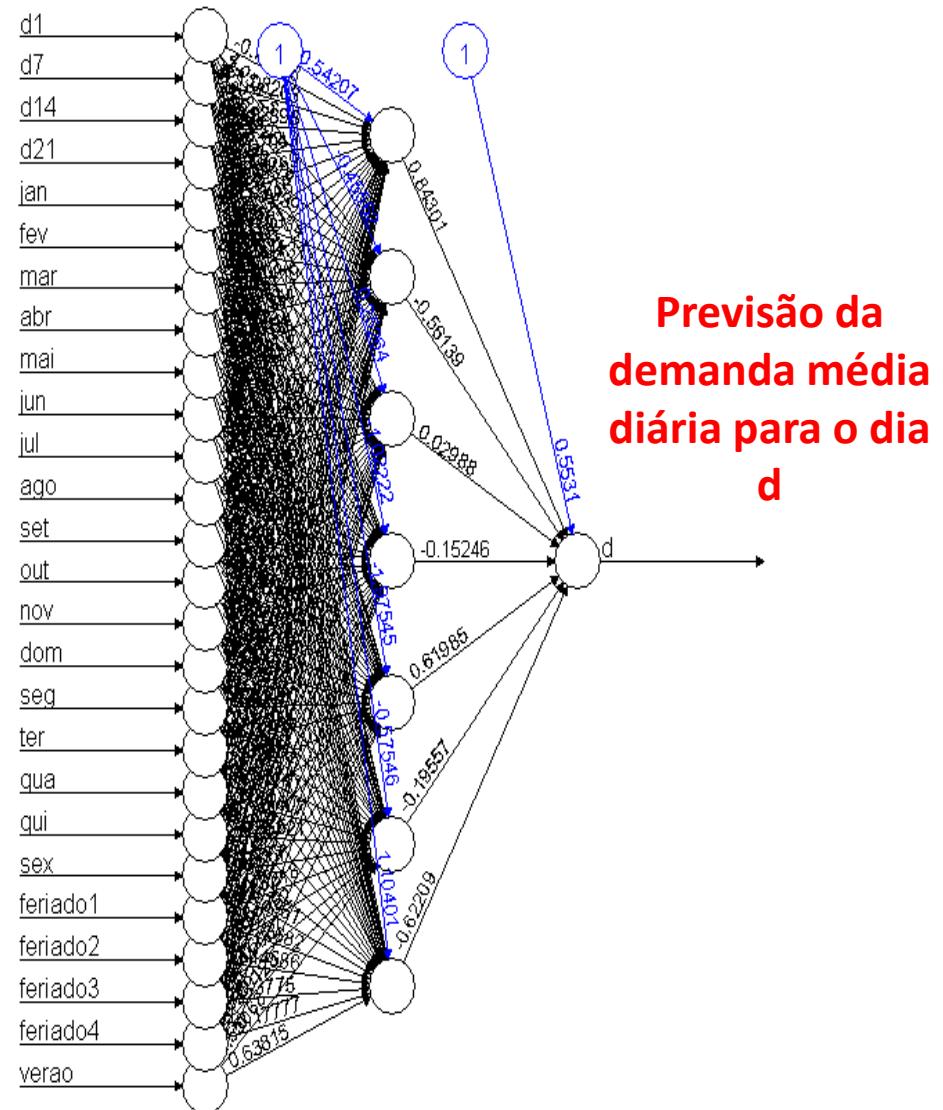
feriado1: feriado

feriado2: carnaval

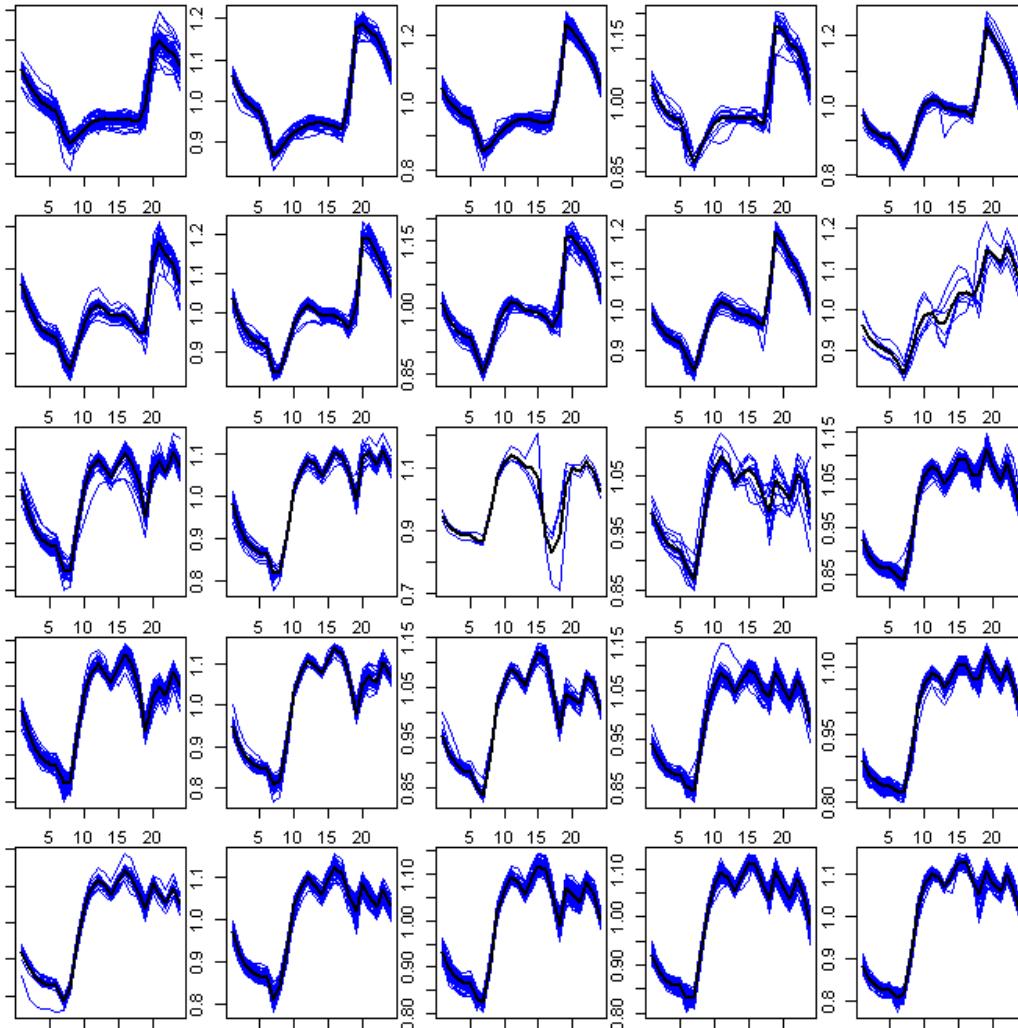
feriado3: quarta-feria de cinzas

feriado4: véspera de feriado

Horário de verão, 1 variável dummy



Mapa de Kohonen dos perfis de carga normalizados pelas respectivas demandas médias

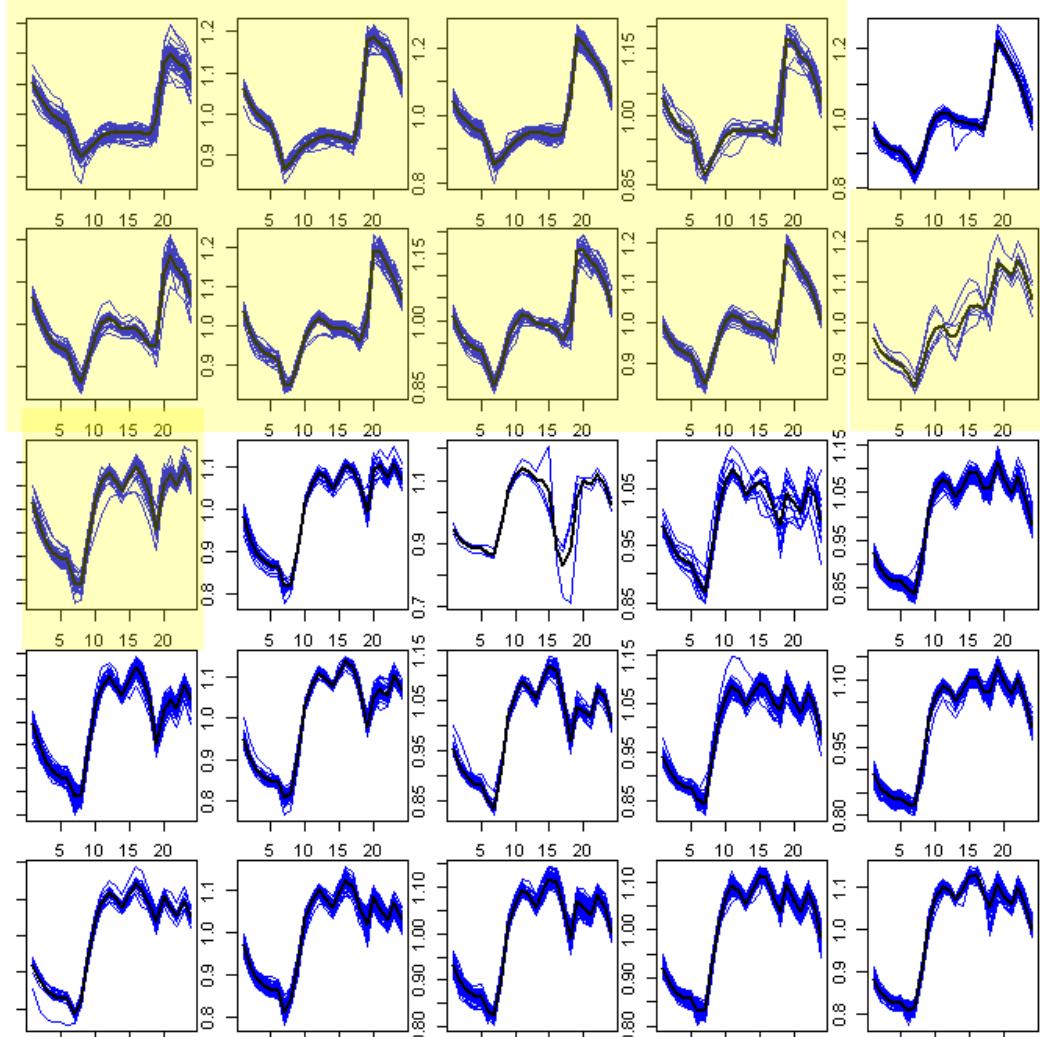


Os 1310 perfis diários foram agrupados em 25 *clusters* (perfis típicos).

Frequência relativa (%)

4.9	3.1	7.6	1.2	4.1
2.1	2.2	2.1	4.0	0.5
2.4	1.5	0.2	1.0	11.0
5.7	1.9	2.6	5.3	5.1
1.8	6.0	6.3	12.4	5.0

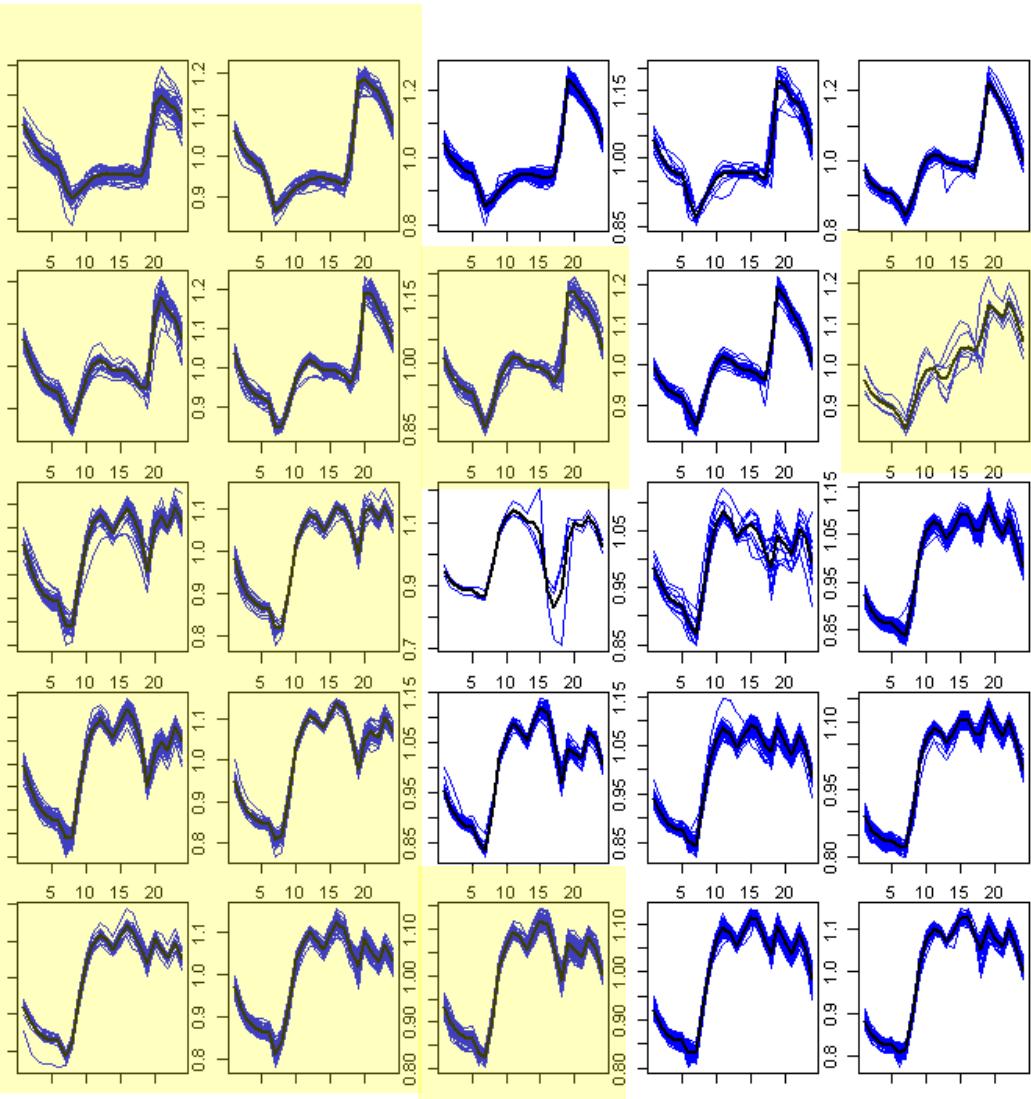
Mapa de Kohonen dos perfis de carga normalizados pelas respectivas demandas médias



Incidência de feriados (%)

17.2	12.5	4.0	81.2	0.0
7.1	10.3	7.1	1.9	33.3
3.2	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0

Mapa de Kohonen dos perfis de carga normalizados pelas respectivas demandas médias



Incidência do horário de verão (%)

98.4	10.0	0.0	0	0.0
100.0	96.6	7.1	0	16.7
100.0	100.0	0.0	0	0.0
100.0	100.0	0.0	0	0.0
100.0	100.0	2.4	0	0.0

Previsão do perfil de carga por meio do algoritmo de Wang Mendel

Variáveis explicativas do perfil de carga do dia d :

- perfis dos dias $d-1$, $d-7$, $d-14$ e $d-21$
- tipo do dia d (normal, véspera de feriado, feriado ou carnaval)
- período do ano (horário normal ou horário de verão)

Por meio do algoritmo de Wang Mendel foi gerada uma base contendo 683 regras fuzzy.

Por exemplo, para previsão do perfil em um dia útil logo após um domingo:

SE $perfil(d-1) \in \text{cluster 22}$ **E** $perfil(d-7) \in \text{cluster 3}$ **E** $perfil(d-14) \in \text{cluster 5}$ **E** $perfil(d-21) \in \text{cluster 3}$ **E** dia d não é dia normal **E** dia d não está no horário de verão **ENTÃO** $perfil(d) \in \text{cluster 10}$

Resultados

A metodologia proposta apresentou um desempenho ligeiramente superior ao método ingênuo.

A metodologia proposta conseguiu fazer previsões melhores nos feriados.

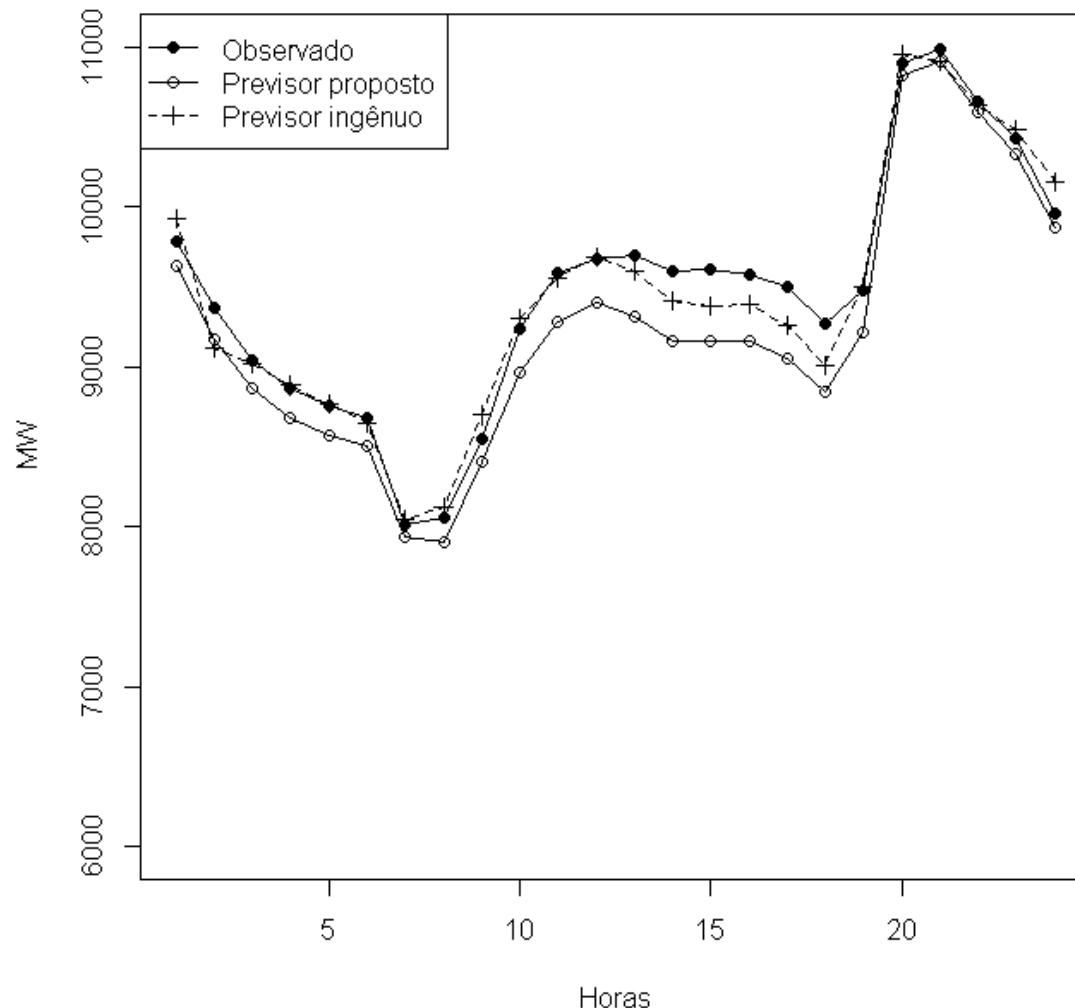
Previsor ingênuo: curva (dia d) = carga (dia d-7)

Desempenho para previsão 1 dia a frente

Metodologia	Período	MAPE	MAD
Proposta	Treinamento	1,77 %	150,42 MW
	Validação	2,08 %	194,25 MW
Previsor ingênuo	Treinamento	3,16 %	266,64 MW
	Validação	3,92 %	359,55 MW

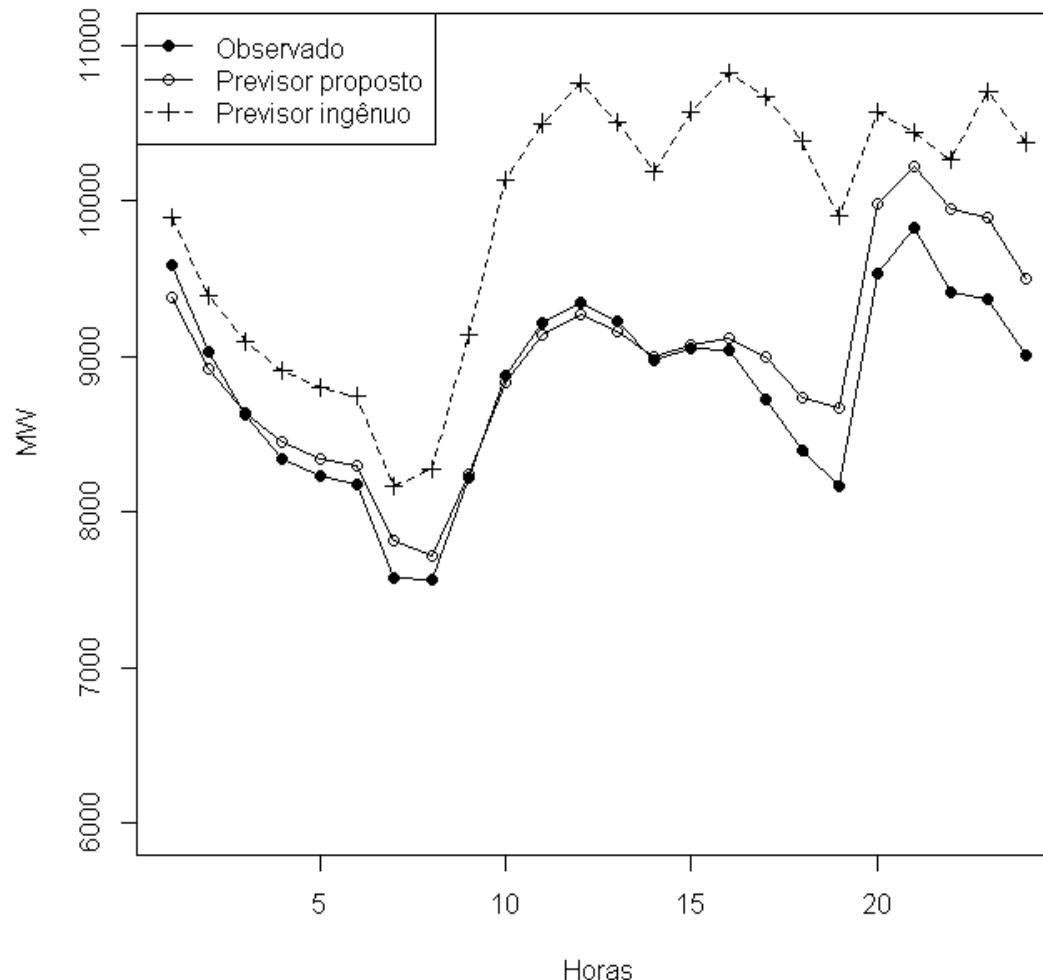
Resultados

Previsão 1 dia à frente para 30/11/2013

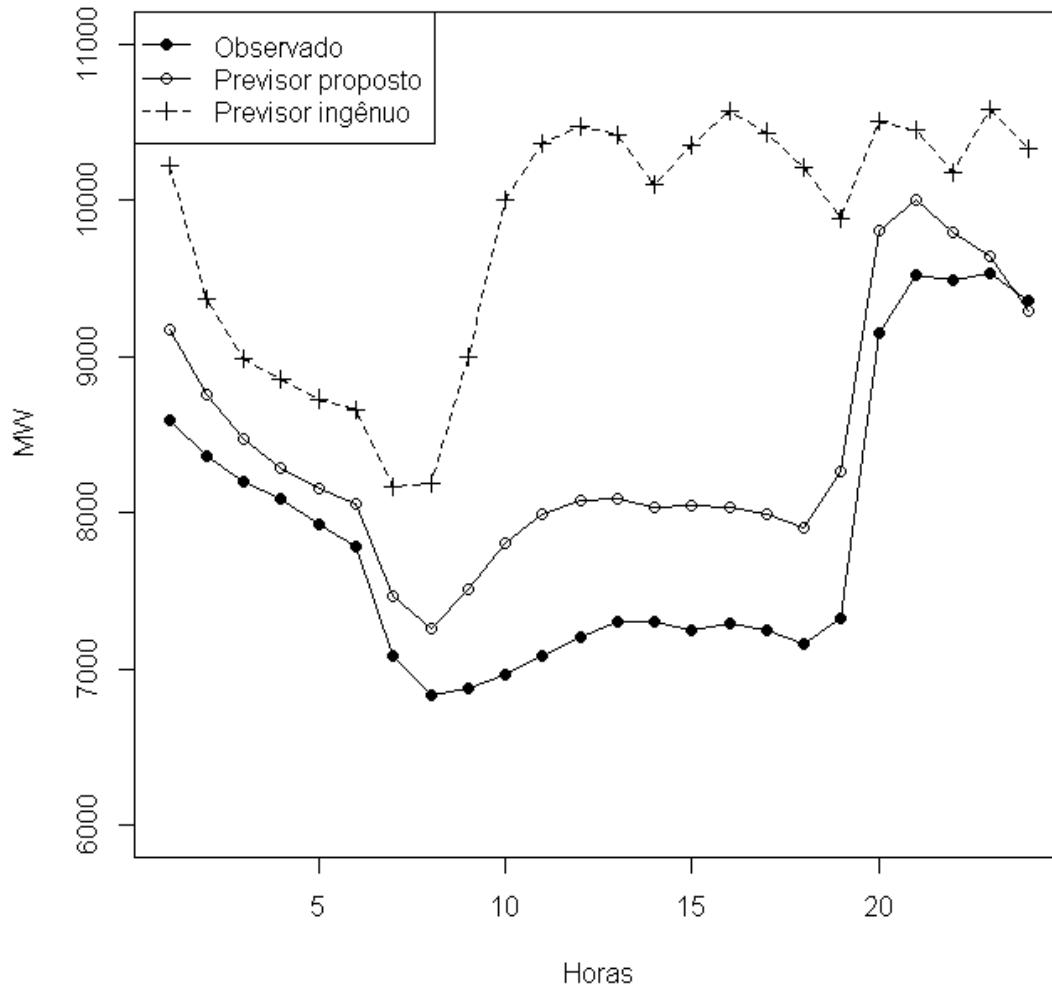


Resultados

Previsão 1 dia à frente para a véspera de Natal



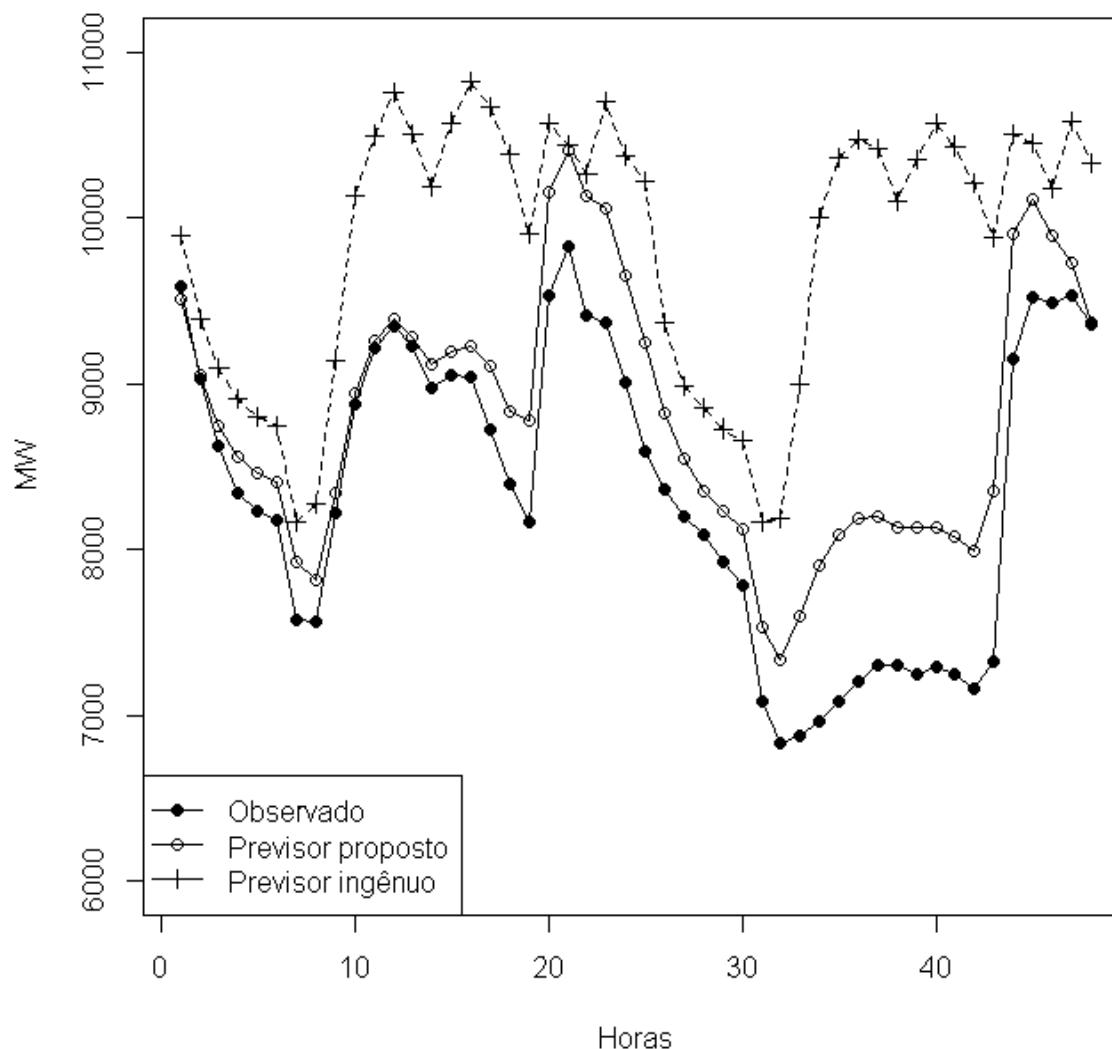
Previsão 1 dia à frente para o Natal



Desempenho para previsão 2 dias a frente

Metodologia	Período	MAPE	MAD
Proposta	Treinamento	2,09 %	177,58 MW
	Validação	2,66 %	248,27 MW
Previsor ingênuo	Treinamento	3,16%	266,57 MW
	Validação	3,93%	360,56 MW

Previsão 2 dias à frente para 23 e 24 de dezembro



Conclusões

A análise de agrupamentos é um método exploratório.

Útil na organização de um conjunto de dados em grupos de objetos que compartilham atributos semelhantes.

É uma das principais técnicas da mineração de dados

Conta com uma variedade de algoritmos.

Obrigado

José Francisco
UERJ/Cepel

professorjfmp@hotmail.com



Medidas de associação

Utilizadas para comparar objetos com características não métricas.

Por exemplo, para definir a semelhança entre indivíduos caracterizados por variáveis qualitativas do tipo booleano:

- 0 indica a ausência de uma característica para determinado indivíduo.
- 1 indica a presença da característica.

	TV	geladeira	freezer	computador	rádio	p-ésimo atributo		
2 indivíduos caracterizados por p variáveis binárias	indivíduo i	1	1	1	0	1	...	1
	indivíduo j	1	1	0	1	0	...	0

Medidas de associação baseiam-se no número de vezes que ocorrem concordâncias e/ou discordâncias entre os indivíduos.

Medidas de associação

	TV	geladeira	freezer	computador	rádio	p-ésimo atributo	
indivíduo i	1	1	1	0	1	...	1
indivíduo j	1	1	0	1	0	...	0

		indivíduo i		totais	
		1	0		
indivíduo j	1	a	b	a+b	
	0	c	d	c+d	
totais		a+c	b+d	$p=a+b+c+d$	



a = nº de atributos que assumem o valor 1 (estão presentes) em ambos os indivíduos

b = nº de atributos que assumem o valor 1 no indivíduo j e 0 no indivíduo i

c = nº de atributos que assumem o valor 0 no indivíduo j e 1 no indivíduo i

d = nº de atributos que assumem o valor 0 (estão ausentes) em ambos os indivíduos

Medidas de associação – Coeficiente de emparelhamento simples

		indivíduo i		totais	
		1	0		
indivíduo j	1	a	b	a+b	$Nº\ de\ concordâncias\ entre\ os\ dois\ indivíduos$
	0	c	d	c+d	
totais		a+c	b+d	$p=a+b+c+d$	

$s_{ij} = \frac{a + d}{a + b + c + d}$

$\leftarrow N^{\circ}\ de\ atributos$

s_{ij} mede a semelhança entre dois indivíduos

		indivíduo i		totais	
		1	0		
indivíduo j	1	a	b	a+b	$Nº\ de\ discordâncias\ entre\ os\ dois\ indivíduos$
	0	c	d	c+d	
totais		a+c	b+d	$p=a+b+c+d$	

$d_{ij} = \frac{b + c}{a + b + c + d}$

$\leftarrow N^{\circ}\ de\ atributos$

d_{ij} mede a diferença entre dois indivíduos

Medidas de associação – Coeficientes de Jaccard

$$s_{ij} = \frac{a}{a+b+c}$$

s_{ij} mede a semelhança entre dois indivíduos

$$d_{ij} = \frac{b+c}{a+b+c}$$

d_{ij} mede a diferença entre dois indivíduos

Evitam a contribuição da ausência conjunta de uma característica para o cálculo da distância entre dois indivíduos

Outras medidas de associação (Johnson & Wichern, 2002)

$$s_{ij} = \frac{2(a+d)}{2(a+d)+b+c}$$

Peso duplo às presenças e ausências simultâneas

$$s_{ij} = \frac{a+d}{a+d+2(b+c)}$$

Peso duplo às situações discordantes

$$s_{ij} = \frac{2a}{2a+b+c}$$

Peso duplo às presenças simultâneas e exclusão das ausências simultâneas

Peso duplo às situações discordantes e exclusão das ausências simultâneas

$$s_{ij} = \frac{a}{b+c}$$

Quociente entre presenças simultâneas e situações discordantes

Medidas de associação – estatística qui-quadrado



		indivíduo i		totais
		1	0	
indivíduo j	1	a	b	a+b
	0	c	d	c+d
totais		a+c	b+d	p=a+b+c+d

$$r_{ij} = \frac{\chi^2}{p} = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

Medidas de associação – Coeficiente de Gower

Permite a utilização de variáveis definidas em diferentes escalas de medida (binárias, nominais, ordinais e continuas), mas torna-se idêntico ao coeficiente de Jaccard quando as variáveis são todas binárias.

$$S_{ij} = \frac{\sum_{v=1}^p s_{ijv}}{\sum_{v=1}^p w_{ijv} s_{ijv}}$$

s_{ijv} é o valor da semelhança entre os indivíduos i e j para a variável v

w_{ijv} é a ponderação da variável v

$w_{ijv} = 1$ se a comparação para a variável v for considerada válida

$w_{ijv} = 0$ se a comparação para a variável v não for considerada válida, por exemplo, quando pelo menos um dos indivíduos apresenta uma não resposta (*missing value) para a variável em causa,

Medidas de associação – Exemplo (Johnson & Wichern, 2002)

Inglês	Norueguês	Dinamarquês	Alemão	Holandês	Francês	Espanhol	Italiano	Polonês	Húngaro	Finlandês
one	en	en	eins	een	un	uno	uno	jeden	egy	yksi
two	to	to	zwei	twee	deux	dos	due	dwa	ketto	kaksi
tree	tre	tre	drei	drie	troi	tres	tre	trzy	harom	kolme
four	fire	fire	vier	vier	quatre	cuatro	quattro	cztery	negy	neua
five	fem	fem	funf	vijf	cinq	cinco	cinque	piec	öt	viisi
six	seks	seks	sechs	zes	six	seis	sei	szese	hat	kuusi
seven	sju	syv	sieben	zeven	sept	siete	sette	siedem	het	seitseman
eight	atte	otte	acht	acht	huit	ochos	otto	osiem	nyolc	kahdeksan
nine	ni	ni	neun	negen	neuf	nueve	nove	dziewiec	kilenc	yhdeksan
ten	ti	ti	zehn	tien	dix	diez	dieci	dziesiec	tiz	kymmenen

Concordância na primeira letra (matriz de similaridade)

3.1 Medidas de associação – Exemplo (Johnson & Wichern, 2002)

Inglês	Norueguês	Dinamarquês	Alemão	Holandês	Francês	Espanhol	Italiano	Polonês	Húngaro	Finlandês
one	en	en	eins	een	un	uno	uno	jeden	egy	yksi
two	to	to	zwei	twee	deux	dos	due	dwa	ketto	kaksi
tree	tre	tre	drei	drie	troi	tres	tre	trzy	harom	kolme
four	fire	fire	vier	vier	quatre	cuatro	quattro	cztery	negy	neua
five	fem	fem	funf	vijf	cinq	cinco	cinque	piec	öt	viisi
six	seks	seks	sechs	zes	six	seis	sei	szese	hat	kuusi
seven	sju	syv	sieben	zeven	sept	siete	sette	siedem	het	seitseman
eight	atte	otte	acht	acht	huit	ochos	otto	osiem	nyolc	kahdeksan
nine	ni	ni	neun	negen	neuf	nueve	nove	dziewiec	kilenc	yhdeksan
ten	ti	ti	zehn	tien	dix	diez	dieci	dziesiec	tiz	kymmenen

Discordância na primeira letra (matriz de dissimilaridade)

Medidas de correlação

Medida de similaridade expressa pelo “coeficiente de correlação” entre os objetos.

Dados dois objetos i e j caracterizados por p variáveis

$$X_i^T = \begin{pmatrix} x_{i1} & \dots & x_{ip} \end{pmatrix}$$
$$X_j^T = \begin{pmatrix} x_{j1} & \dots & x_{jp} \end{pmatrix}$$

Vetor de médias $\bar{X} = \begin{pmatrix} \bar{x}_1 & \dots & \bar{x}_p \end{pmatrix}$

Coeficiente de correlação de Pearson entre os objetos i e j

$$r_{ij} = \frac{\sum_{v=1}^p (x_{iv} - \bar{x}_v)(x_{jv} - \bar{x}_v)}{\sqrt{\sum_{v=1}^p (x_{iv} - \bar{x}_v)^2} \cdot \sqrt{\sum_{v=1}^p (x_{jv} - \bar{x}_v)^2}} \quad -1 \leq r_{ij} \leq +1$$

Matriz de distâncias – Exemplo (Reis, 2001)

Com o objetivo de encontrar grupos estratégicos, pretende-se aplicar a análise de agrupamentos a uma amostra de seis grandes empresas comerciais, para as quais foram medidas as seguintes dimensões estratégicas:

X_1 = nº de lojas

X_2 = dimensão média das lojas (m^2)

X_3 = % da área alimentar no total das vendas

X_4 = montante da área alimentar no total de caixa (\$ 1000)

Matriz de dados

Empresas	nº de lojas	dimensão média das lojas m^2	% da área alimentar no total das vendas	montante da área alimentar no total de caixa
Modelo	42	830	95	12.000
Pingo Doce	33	1.215	92	11.500
Feira Nova	1	6.000	70	18.500
Supa/Jumbo	6	5.675	63	21.400
Minipreço	31	298	98	870
Continente	4	885	65	23.100
\bar{x}_j → Média	19,50	2.483,83	80,50	14.561,67
s_j → Desvio-padrão	17,81	2.616,34	16,16	8.228,90

Matriz de distâncias – Exemplo (Reis, 2001)

Variáveis em escalas diferentes \Rightarrow padronização

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \quad = \text{Valor padronizado da variável } j \text{ na } i\text{-ésima empresa}$$

Dados
padronizados

Empresas	nº de lojas	dimensão média das lojas m ²	% da área alimentar no total das vendas	montante da área alimentar no total de caixa (\$ 1000)
Modelo	1,26	-0,63	0,90	-0,31
Pingo Doce	0,76	-0,48	0,71	-0,37
Feira Nova	-1,04	1,34	-0,65	0,48
Supa / Jumbo	-0,76	1,22	-1,08	0,83
Minipreço	0,65	-0,84	1,08	-1,66
Continente	-0,87	-0,61	-0,96	1,04

Matriz de distâncias – Exemplo (Reis, 2001)

Construindo a matriz de distâncias

Empresas	nº de lojas (X_1)	dimensão média das lojas m ² (X_2)	% da área alimentar no total das vendas (X_3)	montante da área alimentar no total de caixa (1000 contos) (X_4)
Modelo (1)	1,26	-0,63	0,90	-0,31
Pingo Doce (2)	0,76	-0,48	0,71	-0,37
Feira Nova (3)	-1,04	1,34	-0,65	0,48
Supa / Jumbo (4)	-0,76	1,22	-1,08	0,83
Minipreço (5)	0,65	-0,84	1,08	-1,66
Continente (6)	-0,87	-0,61	-0,96	1,04

Métrica escolhida é o quadrado da distância Euclidiana

$$d_{12} = (1,26 - 0,76)^2 + (-0,63 - (-0,48))^2 + (0,9 - 0,71)^2 + (-0,31 - (-0,37))^2 = 0,3$$

	Modelo (1)	Pingo Doce (2)	Feira Nova (3)	Supa / Jumbo (4)	Minipreço (5)	Continente (6)
D =	0 0,3	0	0	0	0	0

Matriz de distâncias – Exemplo (Reis, 2001)

Construindo a matriz de distâncias

Empresas	nº de lojas (X_1)	dimensão média das lojas m ² (X_2)	% da área alimentar no total das vendas (X_3)	montante da área alimentar no total de caixa (1000 contos) (X_4)
Modelo (1)	1,26	-0,63	0,90	-0,31
Pingo Doce (2)	0,76	-0,48	0,71	-0,37
Feira Nova (3)	-1,04	1,34	-0,65	0,48
Supa / Jumbo (4)	-0,76	1,22	-1,08	0,83
Minipreço (5)	0,65	-0,84	1,08	-1,66
Continente (6)	-0,87	-0,61	-0,96	1,04

$$d_{13} = (1,26 - (-1,04))^2 + (-0,63 - 1,34)^2 + (0,9 - (-0,65))^2 + (-0,31 - 0,48)^2 = 12,2$$

	Modelo (1)	Pingo Doce (2)	Feira Nova (3)	Supa / Jumbo (4)	Minipreço (5)	Continente (6)	
D =	0						Modelo (1)
	0,3	0					Pingo Doce (2)
	12,2		0		0		Feira Nova (3)
						0	Supa / Jumbo (4)
						0	Minipreço (5)
						0	Continente (6)

Matriz de distâncias – Exemplo (Reis, 2001)

Construindo a matriz de distâncias

Empresas	nº de lojas (X_1)	dimensão média das lojas m ² (X_2)	% da área alimentar no total das vendas (X_3)	montante da área alimentar no total de caixa (1000 contos) (X_4)
Modelo (1)	1,26	-0,63	0,90	-0,31
Pingo Doce (2)	0,76	-0,48	0,71	-0,37
Feira Nova (3)	-1,04	1,34	-0,65	0,48
Supa/ Jumbo (4)	-0,76	1,22	-1,08	0,83
Minipreço (5)	0,65	-0,84	1,08	-1,66
Continente (6)	-0,87	-0,61	-0,96	1,04

$$d_{14} = (1,26 - (-0,76))^2 + (-0,63 - 1,22)^2 + (0,9 - (-1,08))^2 + (-0,31 - 0,83)^2 = 12,7$$

Modelo (1)	Pingo Doce (2)	Feira Nova (3)	Supa / Jumbo (4)	Minipreço (5)	Continente (6)	
0						Modelo (1)
0,3	0					Pingo Doce (2)
12,2		0				Feira Nova (3)
12,7			0			Supa / Jumbo (4)
				0		Minipreço (5)
					0	Continente (6)

Matriz de distâncias – Exemplo (Reis, 2001)

Construindo a matriz de distâncias

Empresas	nº de lojas (X_1)	dimensão média das lojas m ² (X_2)	% da área alimentar no total das vendas (X_3)	montante da área alimentar no total de caixa (1000 contos) (X_4)
Modelo (1)	1,26	-0,63	0,90	-0,31
Pingo Doce (2)	0,76	-0,48	0,71	-0,37
Feira Nova (3)	-1,04	1,34	-0,65	0,48
Supa/ Jumbo (4)	-0,76	1,22	-1,08	0,83
Minipreço (5)	0,65	-0,84	1,08	-1,66
Continente (6)	-0,87	-0,61	-0,96	1,04

Matriz de distâncias

	Modelo (1)	Pingo Doce (2)	Feira Nova (3)	Supa / Jumbo (4)	Minipreço (5)	Continente (6)	
D =	0	0,3	12,2	12,7	2,3	9,8	
	0	0	9,2	9,9	1,9	7,4	
		0	0	0,4	15,2	4,3	
				0	17,1	3,4	
					0	13,8	0
							Modelo (1) Pingo Doce (2) Feira Nova (3) Supa / Jumbo (4) Minipreço (5) Continente (6)