# Agrupamento de Séries Temporais e sua Aplicação na Análise de Processos Geofísicos e Ambientais


Manuel G. Scotto

Departamento de Matemática
IST, ULisboa


DM
DEPARTAMENTO
DE MATEMÁTICA
TÉCNICO LISBOA

## Conteúdo

- Agrupamento de objetos;
- Métodos de agrupamento de séries temporais;
- Aplicações (**dados geofísicos e ambientais**);
- Bibliografia.

# Agrupamento de Séries Temporais

## Agrupamento de objetos

- O processo de agrupamento tem por finalidade **criar grupos** de dados/objetos, de acordo com o seu grau de semelhança.

## Agrupamento de objetos

- O processo de agrupamento tem por finalidade **criar grupos** de dados/objetos, de acordo com o seu grau de semelhança.

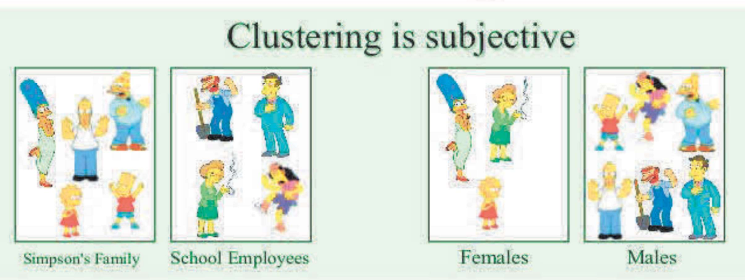
## Agrupamento de objetos

- O agrupamento é feito de tal forma que os objetos pertencentes à mesma classe sejam o mais semelhantes entre si do que objetos em classes diferentes, de acordo com algum critério definido *a priori*.

DM
DEPARTAMENTO
DE MATEMÁTICA
TÉCNICO LISBOA

## Agrupamento de objetos

- O agrupamento é feito de tal forma que os objetos pertencentes à mesma classe sejam o mais semelhantes entre si do que objetos em classes diferentes, de acordo com algum critério definido *a priori*.




DM DEPARTAMENTO DE MATEMÁTICA TÉCNICO LISBOA

# Agrupamento de Séries Temporais


DM
DEPARTAMENTO
DE MATEMÁTICA
TÉCNICO LISBOA

# Agrupamento de Séries Temporais

Caixa Geral de Depositos

Santander

BPI

Millennium bcp

Montepio

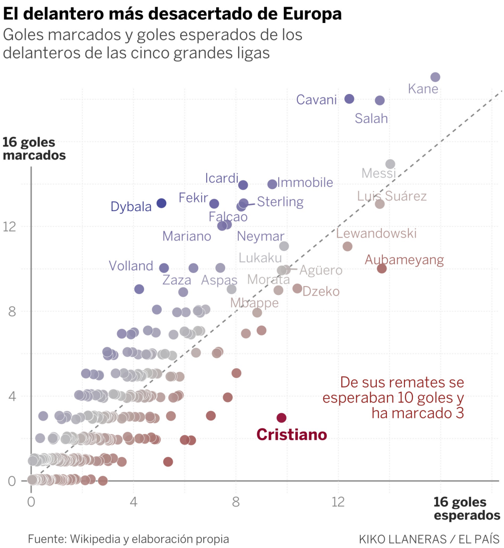
BancoBIC

NOVO BANCO

CA Crédito Agrícola

DM DEPARTAMENTO DE MATEMÁTICA TÉCNICO LISBOA

# Agrupamento de Séries Temporais




DM DEPARTAMENTO DE MATEMÁTICA TÉCNICO LISBOA

## Etapas no agrupamento de objetos

- Seleção de objetos e variáveis;
- Seleção da medida de proximidade;
- Escolha do método de formação de clusters;
- Apresentação e discussão dos resultados.

# Seleção de objetos e variáveis

| Code Client | Type Client | Sector | Cod-Sector | Age | Connection | Loans | Money |
|---|---|---|---|---|---|---|---|

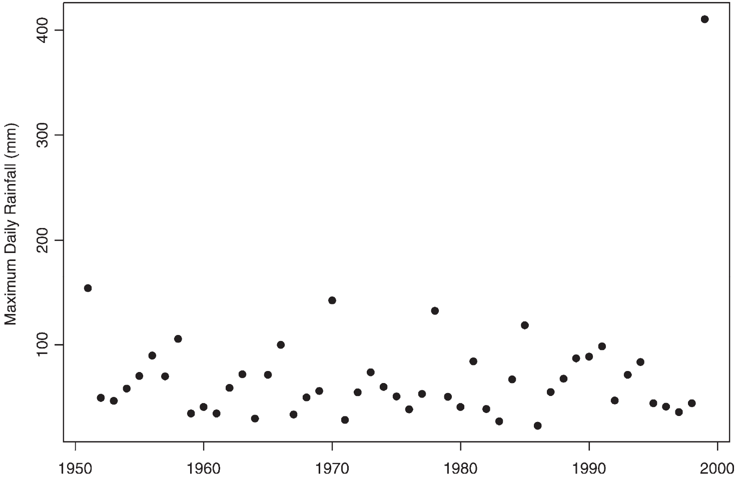
# Seleção de objetos e variáveis

| Code Client | Type Client | Sector | Cod-Sector | Age | Connection | Loans | Money |
|---|---|---|---|---|---|---|---|

# Seleção de objetos e variáveis

| Code Client | Type Client | Sector | Cod-Sector | Age | Connection | Loans | Money |
|---|---|---|---|---|---|---|---|

# Seleção de objetos e variáveis




DM DEPARTAMENTO DE MATEMÁTICA TÉCNICO LISBOA

| Measure | Formula |
|---|---|
| D1: Euclidean distance | $d_{ij} = \left[\sum_{k=1}^{p} w_k^2 (x_{ik} - x_{jk})^2\right]^{1/2}$ |
| D2: City block distance | $d_{ij} = \sum_{k=1}^{p} w_k \lvert x_{ik} - x_{jk}\rvert$ |
| D3: Minkowski distance | $d_{ij} = \left(\sum_{k=1}^{p} w_k^r \lvert x_{ik} - x_{jk}\rvert^r\right)^{1/r} \quad (r \geq 1)$ |
| D4: Canberra distance (Lance and Williams, 1966) | $d_{ij} = \begin{cases} 0 & \text{for } x_{ik} = x_{jk} = 0 \\ \sum_{k=1}^{p} w_k \lvert x_{ik} - x_{jk}\rvert / (\lvert x_{ik}\rvert + \lvert x_{jk}\rvert) & \text{for } x_{ik} \neq 0 \text{ or } x_{jk} \neq 0 \end{cases}$ |
| D5: Pearson correlation | $\delta_{ij} = (1 - \phi_{ij})/2$ with $\phi_{ij} = \sum_{k=1}^{p} w_k (x_{ik} - \bar{x}_{i\cdot})(x_{jk} - \bar{x}_{j\cdot}) \Big/ \left[\sum_{k=1}^{p} w_k (x_{ik} - \bar{x}_{i\cdot})^2 \sum_{k=1}^{p} w_k (x_{jk} - \bar{x}_{j\cdot})^2\right]^{1/2}$ where $\bar{x}_{i\cdot} = \sum_{k=1}^{p} w_k x_{ik} \Big/ \sum_{k=1}^{p} w_k$ |
| D6: Angular separation | $\delta_{ij} = (1 - \phi_{ij})/2$ with $\phi_{ij} = \sum_{k=1}^{p} w_k x_{ik} x_{jk} \Big/ \left(\sum_{k=1}^{p} w_k x_{ik}^2 \sum_{k=1}^{p} w_k x_{jk}^2\right)^{1/2}$ |

DM DEPARTAMENTO DE MATEMÁTICA TÉCNICO LISBOA

**Tipo de métodos**

- **Métodos não hierárquicos:** os métodos não hierárquicos criam os $k$ melhores grupos, em que $k$ é imposto à partida;
- **Métodos hierárquicos**

# Escolha do Método de Formação de Grupos

## Métodos não hierárquicos

Os métodos por partição constroem $k$ grupos e classificam os elementos nos $k$ grupos de acordo com os seguintes requisitos:

DM DEPARTAMENTO DE MATEMÁTICA TÉCNICO LISBOA

## Métodos não hierárquicos

Os métodos por partição constroem $k$ grupos e classificam os elementos nos $k$ grupos de acordo com os seguintes requisitos:

- cada grupo contem pelo menos um elemento;
- cada elemento pertence a um só grupo.

## Métodos não hierárquicos

Os métodos por partição constroem $k$ grupos e classificam os elementos nos $k$ grupos de acordo com os seguintes requisitos:

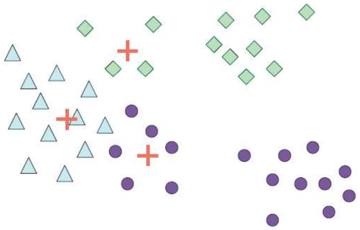
- cada grupo contem pelo menos um elemento;
- cada elemento pertence a um só grupo.

O algoritmo procura uma partição de tal forma que os objetos pertencentes a um mesmo grupo estejam o **mais próximo** possível e que casos em grupos diferentes estejam o **mais afastados** possível (princípio de coesão interna e isolamento externo).

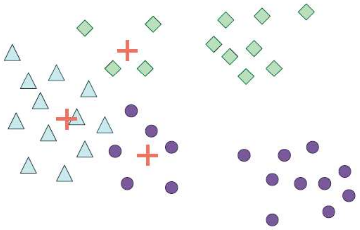
**Algoritmo**

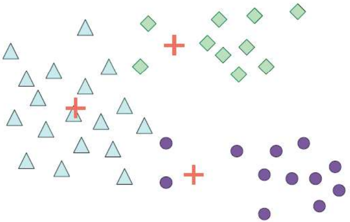
1. Selecionar a partição inicial;
2. Colocar cada objeto no grupo que tem o centróide mais próximo dele.
3. Recalcular os centróides dos novos grupos;
4. Repetir os passos 2 e 3 até não haver mais recolocações.

# $k$-means

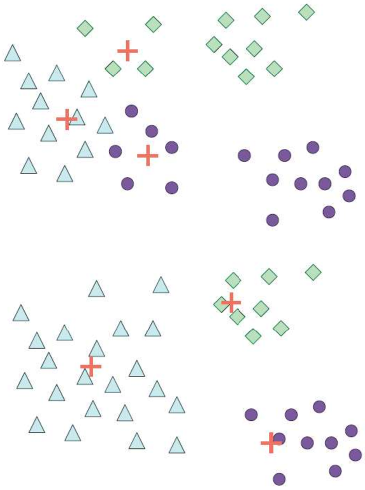


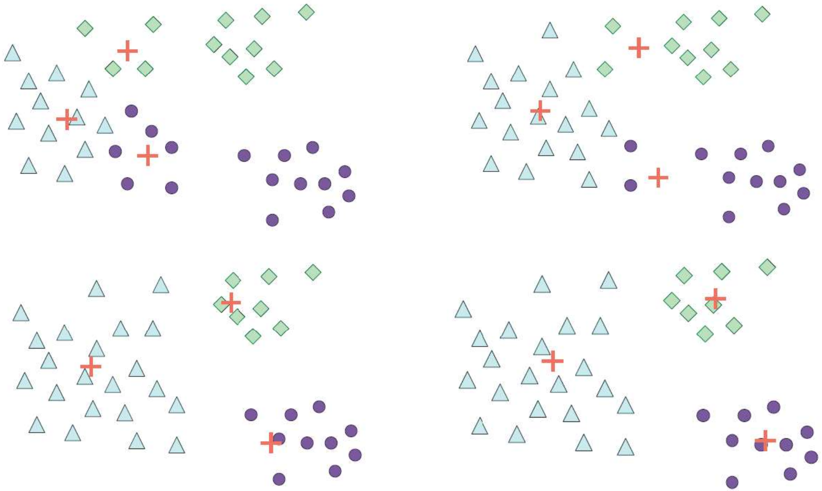
# *k*-means





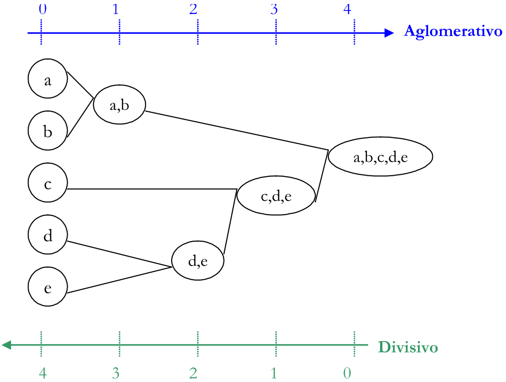
# $k$-means





## Métodos hierárquicos (MH)

Nos MH existem **2 procedimentos:** aglomerativo e divisivo



## Métodos hierárquicos

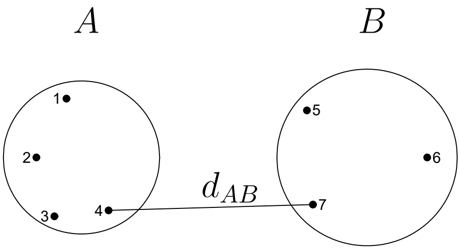
- Método do vizinho **mais próximo** (*single linkage*);
- Método do vizinho **mais afastado** (*complete linkage*);
- Método de ligação **por média** (*average linkage*);
- Método de Ward.

DM DEPARTAMENTO DE MATEMÁTICA TÉCNICO LISBOA

# Escolha do Método de Formação de Grupos

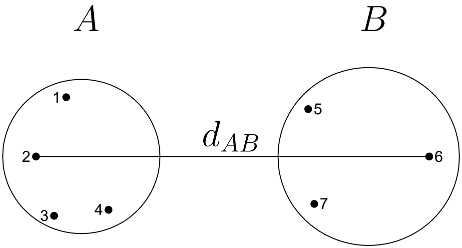
Método do vizinho **mais próximo** (*single linkage*)

$$d_{AB} = \min_{a \in A, b \in B} \{d(a, b)\}.$$



Método do vizinho **mais afastado** (*complete linkage*)

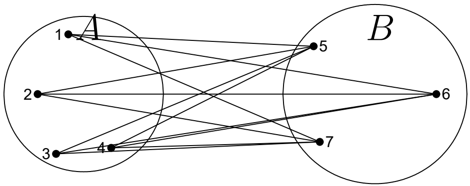
$$d_{AB} = \max_{a \in A, b \in B} \{d(a,b)\}.$$



# Escolha do Método de Formação de Grupos

Método de ligação **por média** (*average linkage*);

$$d_{AB} = (|A||B|)^{-1} \sum_{a \in A} \sum_{b \in B} d(a, b).$$



### Agrupamento de objetos

- Embora uma parte significativa dos métodos de agrupamento propostos na literatura sejam para classificar dados sem estrutura de dependência temporal, nos últimos anos têm vindo a ser propostos vários métodos para **classificar séries temporais**.

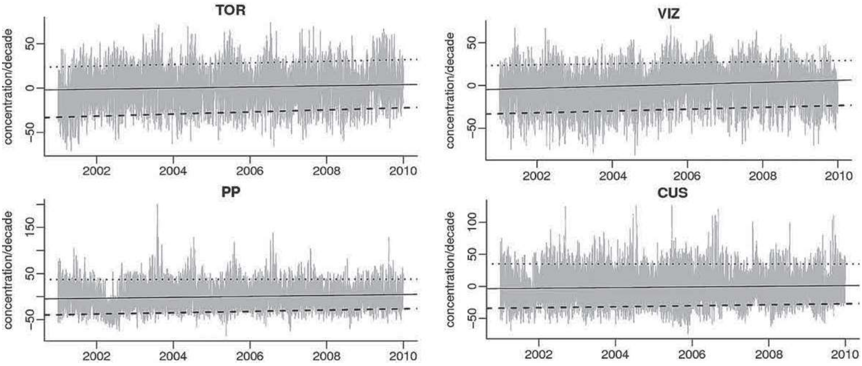
## Série temporal?

**Wikipedia:** Uma série temporal é uma sequência de realizações (observações) de uma variável ao longo do tempo. Dito de outra forma, é uma sequência de pontos (dados numéricos) em ordem sucessiva, geralmente ocorrendo em intervalos uniformes. Portanto, uma série temporal é uma sequência de números coletados em intervalos regulares durante um período de tempo.

DM
DEPARTAMENTO
DE MATEMÁTICA
TÉCNICO LISBOA

# Exemplos de Séries Temporais

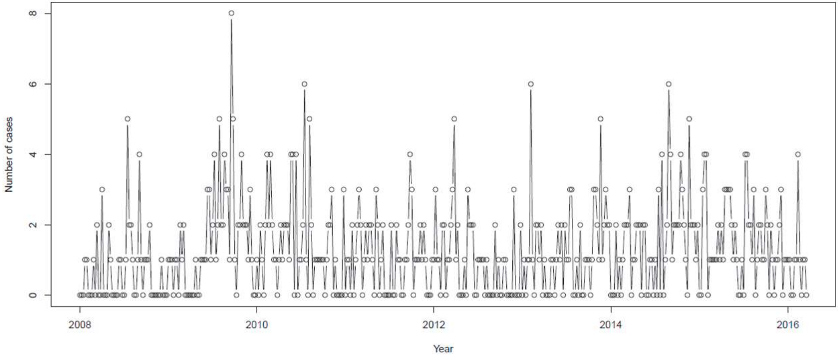
- Concentrações horárias, em $\mu gm^{-3}$, de $O_3$ em Els Torms (TOR), Víznar (VIZ), Paio Pires (PP) e Custóias (CUS)




DM
DEPARTAMENTO DE MATEMÁTICA
TÉCNICO LISBOA

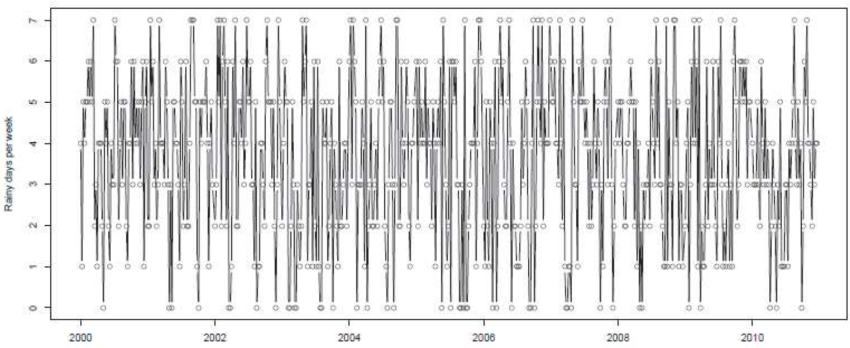
# Exemplos de Séries Temporais

- Número semanal de novos casos de Human Papilloma Virus (HPV) num hospital de Girona (Catalunya/España)



# Exemplos de Séries Temporais

- Número semanal de dias com chuva em Bremen (Alemanha) durante o período 2000-2010



# Agrupamento de Séries Temporais

## Agrupamento de séries temporais

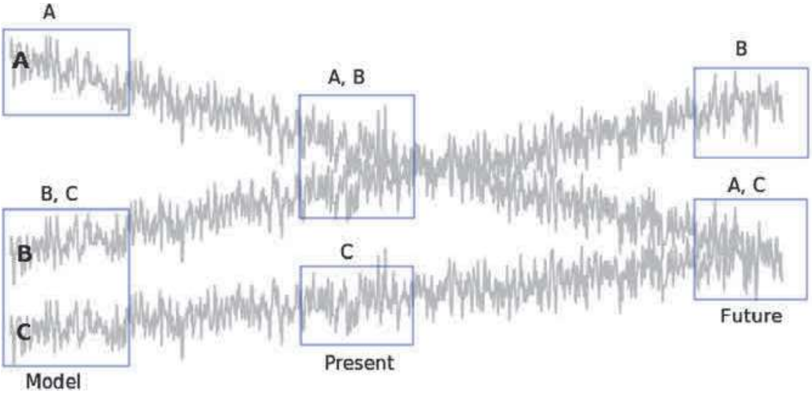
- A **análise de séries temporais** geofísicas e ambientais tem suscitado um interesse crescente pela sua relevância científica e socioeconómica e pelo seu papel na identificação, compreensão e mitigação das alterações climáticas.
- A disponibilidade cada vez mais generalizada, e de forma aberta e gratuita, de dados meteorológicos de estações, satélites torna cada vez mais evidente a necessidade de ferramentas para resumir a informação contida nesse enorme volume de dados.
- Neste contexto, as técnicas para agrupamento de séries temporais assumem um papel de relevo.

# Agrupamento de Séries Temporais

## Agrupamento de séries temporais

- Uma **questão fundamental** que surge sempre em qualquer processo de agrupamento de séries temporais é definir a noção de **semelhança** entre as séries.
- No contexto da classificação de séries temporais a definição de tal medida torna-se particularmente complexa, devido ao caráter dinâmico das séries.

# Agrupamento de Séries Temporais



## Agrupamento de séries temporais

Nos últimos anos têm sido propostas um conjunto alargado de medidas de similaridade para o agrupamento de séries temporais que, de grosso modo, podem ser agrupadas em 4 categorias:

- baseadas em **modelos** (*model-based*);
- baseadas em **características** das séries (*feature-based*);
- baseadas em **previsões** (*future-information-based*);
- baseadas em medidas de **complexidade** (*complexity-based*).

## Medidas baseadas em modelos

- O procedimento neste caso é ajustar primeiro um modelo paramétrico a cada uma das séries e representar cada uma delas através do correspondente vetor de estimativas pontuais dos parâmetros do modelo.
- A seguir constrói-se uma medida de distância entre o conjunto de vetores dois a dois.

## Distância de Piccolo

Para processos AR(I)MA invertíveis, Piccolo (1990) introduziu a distância de Minkowski de segunda ordem em que os vetores de parâmetros contêm as estimativas pontuais dos parâmetros na sua representação AR:

DM DEPARTAMENTO DE MATEMÁTICA TÉCNICO LISBOA

## Distância de Piccolo

Para processos AR(I)MA invertíveis, Piccolo (1990) introduziu a distância de Minkowski de segunda ordem em que os vetores de parâmetros contêm as estimativas pontuais dos parâmetros na sua representação AR:

- Um processo estacionário $(X_t)$ admite uma representação ARMA$(p, q)$, se $X_t$ verificar

$$X_t = \psi_1 X_{t-1} + \cdots + \psi_p X_{t-p} + Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q},$$

onde $Z_t \sim WN(0, \sigma_Z^2)$ e $\psi_p \neq 0$, $\theta_q \neq 0$.

## Distância de Piccolo

O processo $(X_t)$ é invertível se verifica

$$Z_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}, \ \forall t.$$

## Distância de Piccolo

O processo $(X_t)$ é invertível se verifica

$$Z_t = \sum_{j=0}^{\infty} \pi_j X_{t-j}, \ \forall t.$$

- **Distância de Piccolo:**

$$d^{Pic}(\boldsymbol{X}_N, \boldsymbol{Y}_N) := \sqrt{\sum_{i=1}^{p} (\pi_{i,X} - \pi_{i,Y})^2},$$

$p = \max(p_1, p_2)$, sendo $p_1, p_2 \in \mathbb{N}$ as ordens dos modelos ajustados às séries $\boldsymbol{X}_N$ e $\boldsymbol{Y}_N$, na sua representação $\mathtt{AR}$, e

## Distância de Piccolo

- **Distância de Piccolo (cont):**

$$\pi_X = [\pi_{1,X} \;\; \pi_{2,X} \cdots \pi_{p_1,X}]'$$

e

$$\pi_Y = [\pi_{1,Y} \;\; \pi_{2,Y} \cdots \pi_{p_2,Y}]'$$

os correspondentes parâmetros associados a cada um dos modelos.

Note-se que $\pi_{i,X} = 0$ se $i > p_1$ e $\pi_{i,Y} = 0$ se $i > p_2$.

## Distância de Maharaj

Maharaj (1996) considerou uma extensão da medida de Piccolo em que as distâncias entre os parâmetros dos modelos são ponderadas pelas matrizes de autocovariância, $W_X(p)$ e $W_Y(p)$, dos modelos autorregressivos ajustados às séries $\boldsymbol{X}_N$ e $\boldsymbol{Y}_N$, e as correspondentes variâncias do ruído branco, $\sigma^2_{Z_X}$ e $\sigma^2_{Z_Y}$, associado a cada um dos modelos.

DM DEPARTAMENTO DE MATEMÁTICA TÉCNICO LISBOA

## Distância de Maharaj

- **Distância de Maharaj:**

$$d^{Mah}(\boldsymbol{X}_N, \boldsymbol{Y}_N) := \sqrt{N}(\pi_X - \pi_Y)' W^{-1}(\pi_X - \pi_Y),$$

sendo

$$W = \sigma^2_{Z_X} W_X^{-1}(p) + \sigma^2_{Z_Y} W_Y^{-1}(p).$$

DM DEPARTAMENTO DE MATEMÁTICA TÉCNICO LISBOA

## Medidas baseadas em características das séries

Galeano e Peña (2000) propuseram a seguinte versão ponderada da distância euclidiana para comparar funções de autocorrelação amostrais, considerando um desfasamento de $h$ unidades, obtidas a partir de séries $\boldsymbol{X}_N$ e $\boldsymbol{Y}_N$.

## Distância de Galeano e Peña

- **Distância de Galeano e Peña:**

$$d^{ACF}(\boldsymbol{X}_N, \boldsymbol{Y}_N) := \sqrt{(\hat{\rho}_X - \hat{\rho}_Y)' \Omega (\hat{\rho}_X - \hat{\rho}_Y)},$$

sendo $\Omega$ uma qualquer matriz de pesos.

## Distância de Galeano e Peña

- **Distância de Galeano e Peña (cont):**

  Caiado et al. (2006) consideraram **três** definições para $\Omega$:

  - ser a matriz identidade. Neste caso a medida $d^{ACF}$ torna-se na distância euclidiana.
  - A segunda é considerar pesos com um decaimento geométrico sendo que, nesse caso, a medida $d^{ACF}$ é dada pela expressão

$$\sqrt{\sum_{i=1}^{h} \alpha(1-\alpha)^i(\hat{\rho}_X(i) - \hat{\rho}_Y(i))^2}, \; 0 < \alpha < 1.$$

## Distância de Galeano e Peña

- **Distância de Galeano e Peña (cont):**

  Caiado et al. (2006):

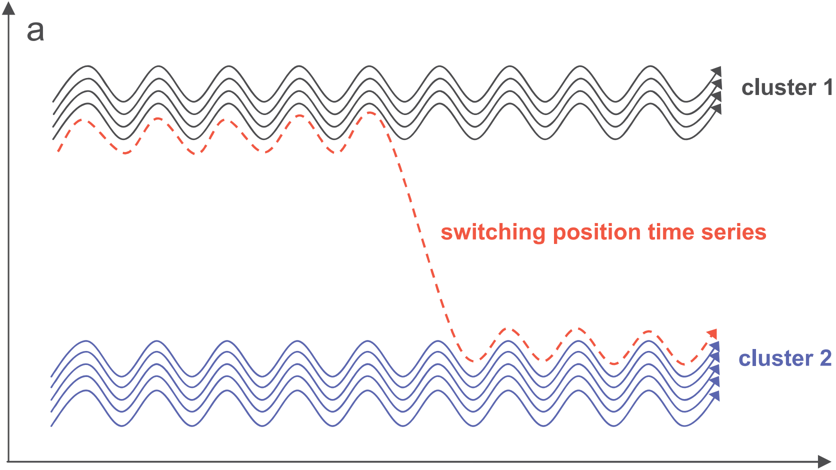
  - a terceira proposta é usar a distância de Mahalanobis entre as autocorrelações, sendo o $\Omega$ a matriz de covariâncias amostrais dos coeficientes de autocorrelação com elementos dados pela fórmula de Bartlett truncada.

## Importante

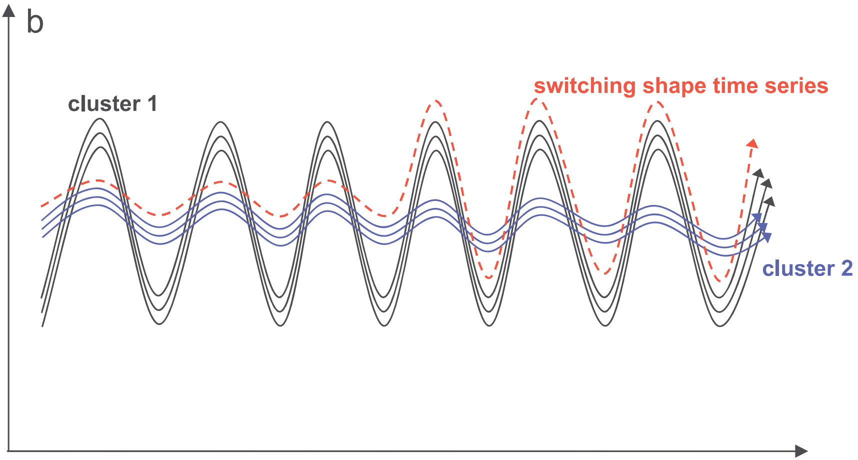
- Uma das **limitações** destas medidas é o facto de ter que se assumir que a estrutura de dependência do processo permanece inalterada ao longo do tempo.
- No entanto, em muito casos a dinâmica das séries temporais muda ao longo do tempo, seja por causa de mudanças de regime, por mudanças de forma ou por uma combinação das duas.
- Nestas situações torna-se possível que em determinados intervalos ao longo do tempo, uma série seja mais "provável" pertencer à uma determinada classe enquanto, noutras janelas temporais, seja mais "provável" pertencer a uma outra classe.

DM DEPARTAMENTO DE MATEMÁTICA TÉCNICO LISBOA

# D'Urso e Maharaj (2009)



# D'Urso e Maharaj (2009)



## Distância de D'Urso e Maharaj

D'Urso e Maharaj (2009) propuseram o método `A-FCM` (*Autocorrelation-based Fuzzy C-Mean*). Este método assenta no cálculo dos pesos associados à atribuição das séries em cada uma das $C$ classes definidas *a priori*. Este cálculo é feito a partir da minimização da função objetivo

$$\sum_{k=1}^{K}\sum_{c=1}^{C} u_{k,c}^{m} \sum_{r=1}^{h} (\hat{\rho}_{X_{(k)}}(r) - \hat{\rho}_c(r))^2,$$

## Distância de D'Urso e Maharaj (cont)

$$\sum_{k=1}^{K}\sum_{c=1}^{C} u_{k,c}^{m} \sum_{r=1}^{h}(\hat{\rho}_{X_{(k)}}(r) - \hat{\rho}_{c}(r))^2, \ (u_{k,c} > 0),$$

sujeita à restrição $\sum_{c=1}^{C} u_{k,c} = 1$, em que $u_{k,c}$ representa o grau de pertença da $k$-éssima série temporal na $c$-éssima classe, $m > 1$ controla o grau de imprecisão (*fuzziness*) da partição em $C$ classes, $\hat{\rho}_{X_{(k)}}(r)$ é o valor estimado da função de autocorrelação da $k$-éssima série temporal e $\hat{\rho}_{c}(r)$ representa o valor estimado da `ACF` da série temporal representativa (*centroid time series*) da $c$-éssima classe.

## Distância de Lafuente-Rego e Vilar (2016)

Uma maneira alternativa de agrupar séries temporais é comparar as suas ACF's quantílicas. A função de autocorrelação quantílica proporciona informação adicional sobre a estrutura de dependência entre as observações associadas a pares de quantis, para cada uma das séries e também entre séries.

Para um conjunto de quantis $q_{\tau_1}, q_{\tau_2}, \ldots, q_{\tau_r}$, associados às ordens $0 < \tau_1 < \cdots < \tau_r < 1$, definidas *a priori*, e um conjunto de desfasamentos $h_1 < \cdots < h_L$, seja

$$d^{QCF}(\boldsymbol{X}_N, \boldsymbol{Y}_N) := \sum_{i=1}^{L} \sum_{i=1}^{r} \sum_{j=1}^{r} \left( \hat{\gamma}_{h_i}^{\boldsymbol{X}_N}(\tau_i, \tau_j) - \hat{\gamma}_{h_i}^{\boldsymbol{Y}_N}(\tau_i, \tau_j) \right)^2,$$

## Distância de Lafuente-Rego e Vilar (2016) (cont)

com

$$\hat{\gamma}_l^Z(\tau,\tau') = \frac{1}{N-l}\sum_{t=1}^{N-l} I(Z_t \le \hat{q}_\tau) I(Z_{t+l} \le \hat{q}_{\tau'}) - \tau\tau',$$

sendo $I(\cdot)$ a função indicatriz e $\hat{q}_\tau$ e $\hat{q}_{\tau'}$ quantis empíricos.

Notar que $\hat{\gamma}_l^Z(\tau,\tau')$ é um estimador de

$\gamma_l^Z(\tau,\tau') = Cov\{I(Z_t \le q_\tau), I(Z_{t+l} \le q_{\tau'})\} = P(Z_t \le q_\tau, Z_{t+l} \le q_{\tau'}) - \tau\tau'.$

## Medidas baseadas em características das séries

Quando o propósito da classificação é agrupar séries temporais consoante o seu **grau de dependência extremal conjunta**, o adequado é considerar medidas de similaridade que incorporem informação sobre a relação de interdependência nas caudas.

## Distância de Durante et al., Luca e Zuccolotto

- **Durante et al. (2015) e De Luca e Zuccolotto (2011)** introduziram a medida

$$d^{TD}(\boldsymbol{X}_N, \boldsymbol{Y}_N) := -\log(\lambda_L),$$

sendo $\lambda_L$ o coeficiente de dependência de cauda esquerda,

$$\lambda_L := \lim_{x \to 0^+} P(U_1 \leq x | U_2 \leq x) = \lim_{x \to 0^+} \frac{C(x,x)}{x},$$

em que $U_1 = F(X)$ e $U_2 = F(Y)$, e $C(\cdot)$ representa uma função cópula. A medida $d^{TD} \geq 0$, sendo que valores reduzidos implicam séries fortemente dependentes na cauda.

## Distância de Durante et al., Luca e Zuccolotto

De Luca e Zuccolotto (2011) propuseram

$$C(x_1, x_2) = 1 - \left\{1 - \left[(1-(1-x_1)^{\kappa})^{-\theta} + (1-(1-x_2)^{\kappa})^{-\theta} - 1\right]^{-1/\theta}\right\}^{\frac{1}{\kappa}}.$$

Neste caso $\lambda_L = 2^{-1/\theta}$. De Luca e Zuccolotto (2015) consideraram uma extensão do modelo anterior em que $\theta$ varia no tempo em função de um conjunto de covariáveis. De referir que de forma perfeitamente análoga, é possível definir a medida $d^{TD}$ considerando o coeficiente de dependência de cauda direita

$$\lambda_U := \lim_{x \to 1^-} P(U_1 > x | U_2 > x).$$

## Distância de D'Urso e Maharaj

- **Distância de D'Urso e Maharaj (2012):**

  D'Urso e Maharaj introduziram um método de agrupamento em que o afastamento entre pares de séries temporais é obtido através de uma medida que pondera, por um lado, a distância entre as variâncias de onduletas associadas a pares bivariados de séries temporais, para um conjunto de escalas definidas *a priori*; e uma outra equivalente para as correspondentes covariâncias de onduletas.

## Distância de D'Urso e Maharaj

- **Distância de D'Urso e Maharaj (2012) (cont):**

A dita medida entre o $i$-éssimo e o $j$-éssimo par $\boldsymbol{Z}_i := (\boldsymbol{X}_{i,N}, \boldsymbol{Y}_{i,N})$ e $\boldsymbol{Z}_j := (\boldsymbol{X}_{j,N}, \boldsymbol{Y}_{j,N})$, é da forma

$$d^W(\boldsymbol{Z}_i, \boldsymbol{Z}_j) := \left\{ (a_{WV} \cdot d_{wv}(\boldsymbol{Z}_i, \boldsymbol{Z}_j))^2 + (a_{WC} \cdot d_{wc}(\boldsymbol{Z}_i, \boldsymbol{Z}_j))^2 \right\}^{\frac{1}{2}},$$

em que os pesos $a_{WV}, a_{WC} \geq 0$ $(a_{WV} + a_{WC} = 1)$, e $d_{wv}$ e $d_{wc}$ correspondem às distâncias entre as variâncias e as covariâncias de onduletas associadas, entre o $i$-éssimo e o $j$-éssimo par, sendo dadas pelas expressões

## Distância de D'Urso e Maharaj

- **Distância de D'Urso e Maharaj (2012) (cont):**

$$d_{wv}(\boldsymbol{Z}_i, \boldsymbol{Z}_j) := \sum_{r=1}^{R} \|diag(\mathcal{C}_{\boldsymbol{Z}_i}(\nu_r)) - diag(\mathcal{C}_{\boldsymbol{Z}_j}(\nu_r))\|,$$

$$d_{wc}(\boldsymbol{Z}_i, \boldsymbol{Z}_j) := \sum_{r=1}^{R} \|\gamma_{\boldsymbol{Z}_i}(\nu_r) - \gamma_{\boldsymbol{Z}_j}(\nu_r)\|,$$

onde $R$ representa o número de escalas, $\mathcal{C}$ é a matriz de variâncias/covariâncias de onduletas, sendo a componente da covariância representada pela função $\gamma$, no termo $d_{wc}$.

### Medidas baseadas em previsões

Nos casos em que o objetivo do agrupamento relaciona-se com o desempenho das previsões futuras das séries temporais, torna-se necessário definir medidas de distância para **comparar densidades de previsão**.

DM
DEPARTAMENTO DE MATEMÁTICA
TÉCNICO LISBOA

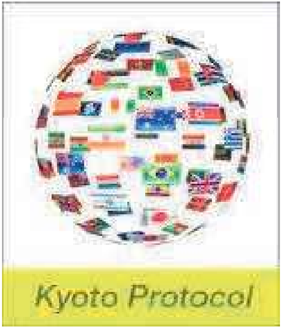
## Medidas baseadas em previsões

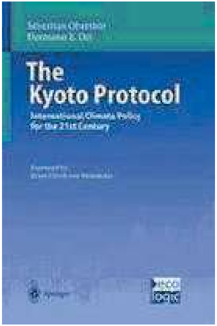
- **Distância de Alonso et al.:** com este objetivo, Alonso et al. (2006) introduziram a medida

$$d^F(\boldsymbol{X}_N, \boldsymbol{Y}_N) := \int \left(f_{X_{N+h}}(x) - f_{Y_{N+h}}(x)\right)^2 dx,$$

sendo $f(\cdot)$ a função de densidade de previsão $h$ passos à frente associada a um modelo autorregressivo de ordem $p$.

# KYOTO PROTOCOL



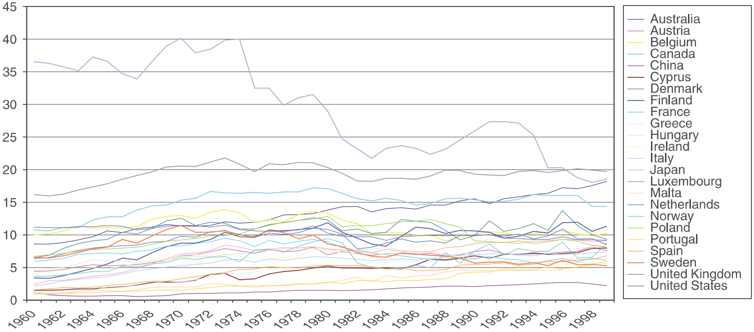

DM
DEPARTAMENTO
DE MATEMÁTICA
TÉCNICO LISBOA

## Kyoto protocol

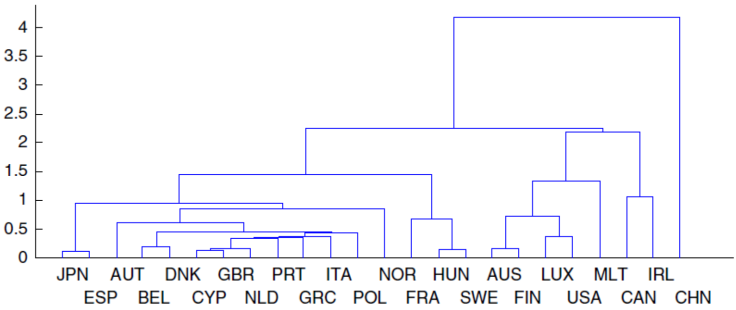
- The **objective** of the **Kyoto conference** was to establish a legally binding international agreement, whereby all the participating nations commit themselves to tackling the issue of global warming and greenhouse gas emissions;
- The target agreed upon was an average reduction of 5.2% from 1990 levels by the year 2012;
- Under the protocol, only the Annex I countries have committed themselves to national or joint reduction targets, that range from a joint reduction of 8% for the European Union and others, to 7% for the US (non-binding as the US is not a signatory), 6% for Japan and 0% for Russia.

DM DEPARTAMENTO DE MATEMÁTICA TÉCNICO LISBOA

# Agrupamento de Séries Temporais



# Agrupamento de Séries Temporais



## Medidas baseadas em previsões

Quando o que está em causa é o agrupamento das séries em termos de **previsões de longo prazo** a partir, por exemplo, da comparação de distribuições de valores de retorno associadas a períodos de retorno definidos *a priori*, uma abordagem que entre em linha de conta com as propriedades extremais das séries torna-se mais adequada.

# Agrupamento de Séries Temporais

## Medidas baseadas em previsões

- **Scotto et al. (2010)**: distância $L_2$ de Wasserstein ponderada

$$d^{EVT}(\boldsymbol{X}_N, \boldsymbol{Y}_N) := \left( \int_0^1 (F_{x_p}^{-1}(y|\boldsymbol{x}_N) - F_{y_p}^{-1}(y|\boldsymbol{y}_N))^2 y(1-y) dy \right)^{1/2},$$

em que $F_{z_p}$ representa a distribuição preditiva *a posteriori* do valor de retorno $z_p$ associado a um período de retorno de $1/p$ unidades de tempo.

## Distâncias baseadas em subamostragem

- **Alonso e Maharaj (2006)**: $\boldsymbol{X}_N \sim P_{\boldsymbol{X}}$ e $\boldsymbol{Y}_N \sim P_{\boldsymbol{Y}}$

## Distâncias baseadas em subamostragem

- **Alonso e Maharaj (2006)**: $\boldsymbol{X}_N \sim P_{\boldsymbol{X}}$ e $\boldsymbol{Y}_N \sim P_{\boldsymbol{Y}}$

  Testar $\mathcal{H}_0 : P_{\boldsymbol{X}} = P_{\boldsymbol{Y}}$ *vs* $\mathcal{H}_1 : P_{\boldsymbol{X}} \neq P_{\boldsymbol{Y}}$

## Distâncias baseadas em subamostragem

- **Alonso e Maharaj (2006)**: $\boldsymbol{X}_N \sim P_{\boldsymbol{X}}$ e $\boldsymbol{Y}_N \sim P_{\boldsymbol{Y}}$

  Testar $\mathcal{H}_0 : P_{\boldsymbol{X}} = P_{\boldsymbol{Y}}$ vs $\mathcal{H}_1 : P_{\boldsymbol{X}} \neq P_{\boldsymbol{Y}}$

  Seja $T_{N,m} := N\sum_{k=1}^{m}(\hat{\rho}_{\boldsymbol{X}}(k) - \hat{\rho}_{\boldsymbol{Y}}(k))^2$

## Distâncias baseadas em subamostragem:

1. Seja $\boldsymbol{X}_j = (X_j, X_{j+1}, \ldots, X_{j+l-1})$ e $\boldsymbol{Y}_j = (Y_j, Y_{j+1} \ldots, Y_{j+l-1})$ para $j = 1, 2, \ldots, N - l + 1$.
2. Para cada subamostra calcular
$$T^{(j)}_{l,m} := l \sum_{k=1}^{m} (\hat{\rho}_{\boldsymbol{X}_j}(k) - \hat{\rho}_{\boldsymbol{Y}_j}(k))^2.$$
3. Estimar a distribuição de $T_{N,m}$
$$\hat{G}_{N,l}(x) = \frac{1}{N-l+1} \sum_{j=1}^{N-l+1} I(T^{(j)}_{l,m} \leq x).$$
4. O valor crítico do teste, $q_{1-\alpha}$, é
$$q_{1-\alpha} := \inf\{x : \hat{G}_{N,l}(x) \geq 1 - \alpha\}.$$

Information Systems 53 (2015) 16–38

Contents lists available at ScienceDirect

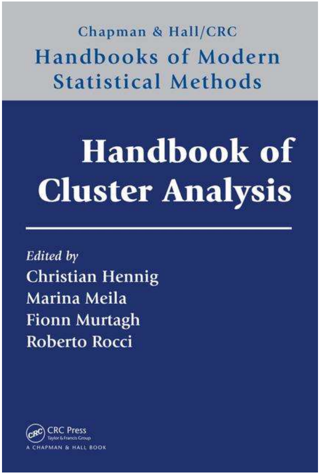
Information Systems

journal homepage: www.elsevier.com/locate/infosys

# Time-series clustering – A decade review

Saeed Aghabozorgi, Ali Seyed Shirkhorshidi *, Teh Ying Wah

*Department of Information System, Faculty of Computer Science and Information Technology, University of Malaya (UM), 50603 Kuala Lumpur, Malaysia*


Chapman & Hall/CRC
Handbooks of Modern Statistical Methods
Handbook of Cluster Analysis
Edited by
Christian Hennig
Marina Meila
Fionn Murtagh
Roberto Rocci
CRC Press
A CHAPMAN & HALL BOOK


DM
DEPARTAMENTO
DE MATEMÁTICA
TÉCNICO LISBOA