

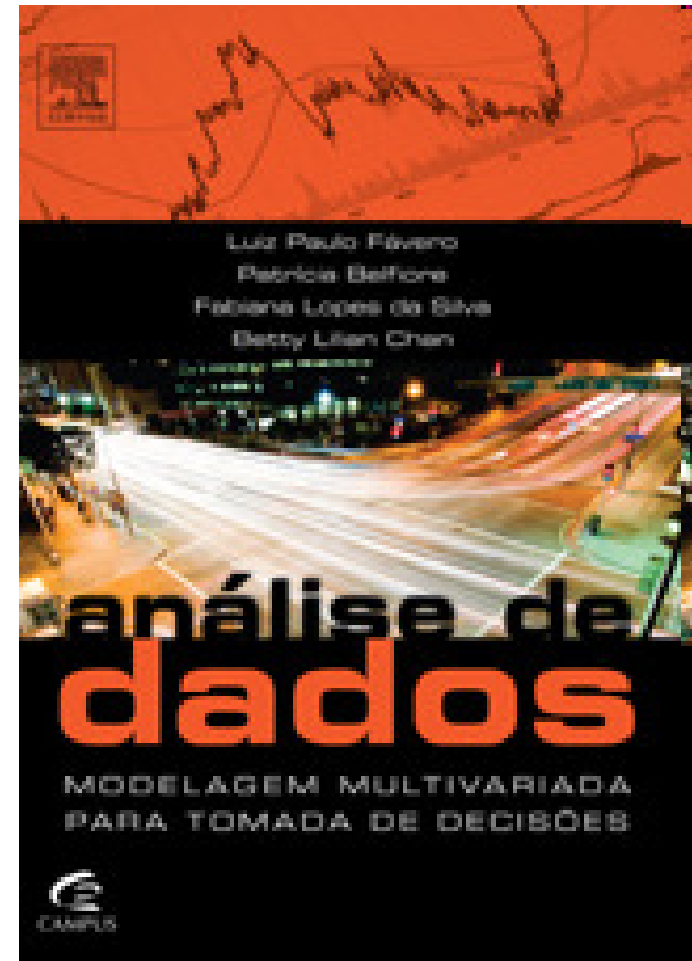
# **Análise de dados multivariados I**

## **Escalonamento**

## **Multidimensional**

## CAPÍTULO 9

### Escalonamento Multidimensional



*Análise de Dados: Modelagem Multivariada para Tomada de Decisões.* Luiz Paulo FÁVERO, Patrícia BELFIORE, Fabiana Lopes DE SILVA e Betty Lilian CHAN, Rio de Janeiro: Elsevier, 2009.

O ideal é acharmos uma representação gráfica dos objetos de modo que o estresse seja o menor possível.

JOHNSON e WICHERN (2007)

## **Neste tópico:**

- Utilização do escalonamento multidimensional.
- Forma do banco de dados.
- Dados de percepção, preferência e similaridade.
- Medidas de similaridade e dissimilaridade.
- Tipos de escalonamento multidimensional.
- Interpretar as representações gráficas.

## 1 Apresentação do Capítulo:

O EMD é uma técnica de interdependência que permite mapear distâncias entre objetos.

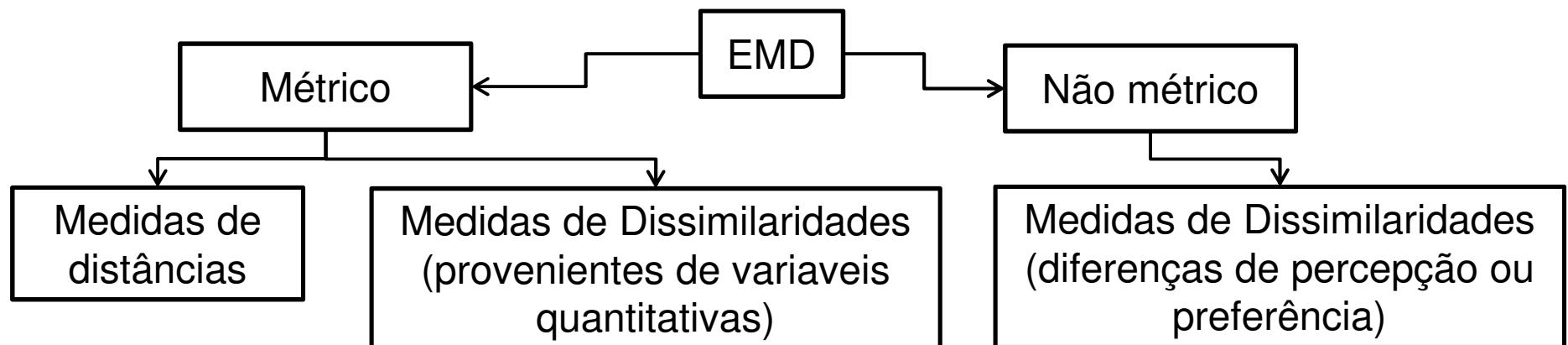
Será apresentado os EMD's não métrico e métrico.

Objetivos do capítulo:

- Introduzir conceitos do EMD.
- Aplicação da técnica.
- Discutir os resultados obtidos.

## 2 Introdução ao EMD

- O EMD é de fácil aplicação.
- A técnica é apropriada para representar graficamente  $n$  elementos em um espaço de dimensão menor do que o original, levando-se em conta a distância ou a similaridade que os elementos têm entre si.



**Observação:** Na análise de cluster hierárquicos pelo SPSS, no rodapé da matriz de proximidade gerada existe a informação de que aquelas distâncias euclidianas formam uma matriz de dissimilaridades (*this is a dissimilarity matrix*).

**Analogia entre o EMD e o uso de um mapa.**

Arquivo: cidadesBrasileiras.sav

Suponha que não temos a configuração geográfica, mas temos a informação das distâncias entre cidades

# Escalonamento Multidimensional (EMD)

CidadesBrasileiras.sav [DataSet1] - SPSS Statistics Data Editor

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

1 : Aracaju 0,0 Visible: 15 of 15 Variables

	Aracaju	Belém	BHorizonte	Brasília	Curitiba	Fortaleza	Maceió	Manaus	Natal	PAlegre	Recife	RJaneiro	SLuís	SPaulo	Salvador
1	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
2	1590,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
3	1350,00	2123,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
4	1271,00	1627,00	589,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
5	2010,00	2574,00	823,00	1087,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
6	810,00	1138,00	1860,00	1682,00	2598,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
7	202,00	1632,00	1416,00	1566,00	2205,00	717,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
8	2574,00	1288,00	2569,00	1967,00	2634,00	2295,00	2670,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
9	606,00	1550,00	1800,00	1775,00	2580,00	444,00	435,00	2658,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
10	2520,00	3084,00	1370,00	1617,00	547,00	3126,00	2712,00	3987,00	3069,00	0,00	0,00	0,00	0,00	0,00	0,00
11	386,00	1680,00	1632,00	1632,00	2400,00	640,00	191,00	2823,00	252,00	3083,00	0,00	0,00	0,00	0,00	0,00
12	1485,00	2460,00	340,00	900,00	669,00	2190,00	1680,00	2854,00	2122,00	1133,00	1865,00	0,00	0,00	0,00	0,00
13	1237,00	493,00	1848,00	1530,00	2514,00	640,00	1200,00	1752,00	1035,00	3042,00	1197,00	2271,00	0,00	0,00	0,00
14	1740,00	2490,00	500,00	865,00	330,00	2238,00	1940,00	3100,00	2486,00	844,00	2135,00	364,00	2360,00	0,00	0,00
15	267,00	1695,00	980,00	1053,00	1734,00	1018,00	464,00	2617,00	870,00	2241,00	654,00	1220,00	1290,00	1486,00	0,00
16															
17															
18															
19															
20															
21															

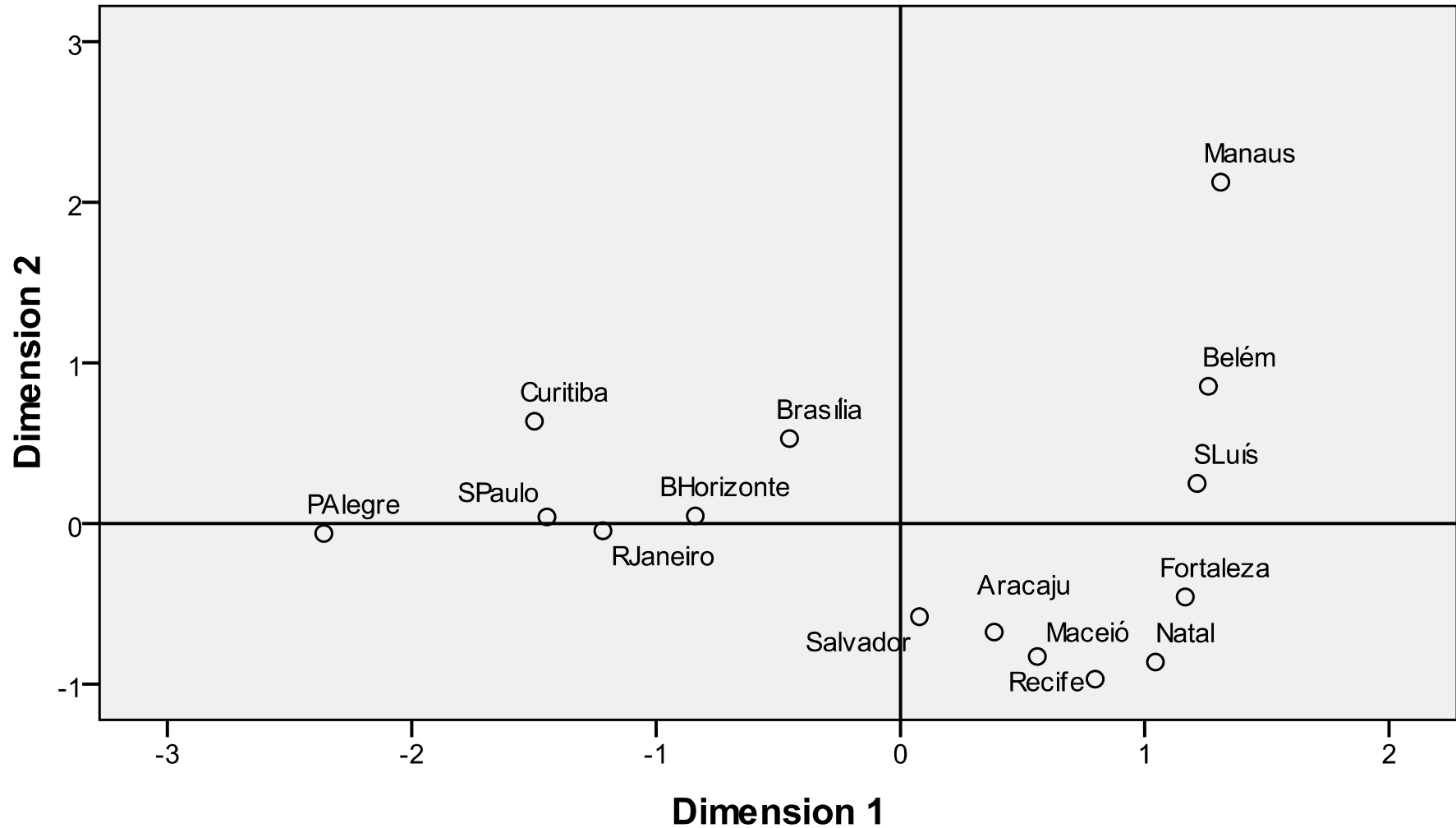
Data View Variable View

SPSS Statistics Processor is ready

Pela técnica é criado um novo sistema de coordenadas que facilita a interpretação das proximidades dos objetos

# Escalonamento Multidimensional (EMD)

## Euclidean distance model



Disposição das distâncias entre 15 cidades brasileiras



## 3 Modelagem do Escalonamento Multidimensional

- Para  $N$  objetos de uma matriz de similaridade temos  
 $M = N(N-1) / 2$  distâncias (ou dissimilaridades) entre pares de objetos.
- A similaridade entre pares de objetos é tal que:

$$S_{i_1 j_1} < S_{i_2 j_2} < \dots < S_{i_M j_M}$$

- A distância ou dissimilaridade é tal que:

$$d_{i_1 j_1} > d_{i_2 j_2} > \dots > d_{i_M j_M}$$

**Ex.** Uma matriz de Similaridades de pares de quatro estímulos (objetos)

Similaridades	Estímulo 1	Estímulo 2	Estímulo 3	Estímulo 4
Estímulo 1	-			
Estímulo 2	4	-		
Estímulo 3	1	6	-	
Estímulo 4	3	5	2	-

Dissimilaridades  $\delta_{ij} = (M + 1) - s_{ij}$

**Ex.** Matriz de dissimilaridades de pares de quatro estímulos (objetos)

Dissimilaridades	Estímulo 1	Estímulo 2	Estímulo 3	Estímulo 4
Estímulo 1	-			
Estímulo 2	3	-		
Estímulo 3	6	1	-	
Estímulo 4	4	2	5	-

## Escalonamento Multidimensional (EMD)

A partir de uma **Matriz de Dissimilaridades** –  $\Delta$  (por exemplo,

para  $n = 4$ )

$$\Delta = \begin{pmatrix} \delta_{11} & \delta_{12} & \delta_{13} & \delta_{14} \\ \delta_{21} & \delta_{22} & \delta_{23} & \delta_{24} \\ \delta_{31} & \delta_{32} & \delta_{33} & \delta_{34} \\ \delta_{41} & \delta_{42} & \delta_{43} & \delta_{44} \end{pmatrix}$$

a técnica de EMD fornece como resultado uma matriz retangular  $n \times m$ , sendo  $m$  o número de dimensões.

A matriz  $X$  corresponde à solução com duas dimensões:

$$X = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \\ x_{41} & x_{42} \end{pmatrix}$$

## Escalonamento Multidimensional (EMD)

Uma fórmula geral para distância é a distância de Minkowski:

$$d_{ij} = \left( \sum_{p=1}^m (x_{ip} - x_{jp})^q \right)^{1/q}$$

A estimação das distâncias correspondentes à todos os objetos proporciona uma nova matriz, matriz  $D$ .

$$D = \begin{pmatrix} d_{11} & d_{12} & d_{13} & d_{14} \\ d_{21} & d_{22} & d_{23} & d_{24} \\ d_{31} & d_{32} & d_{33} & d_{34} \\ d_{41} & d_{42} & d_{43} & d_{44} \end{pmatrix}$$

A solução do EMD deve ser tal que exista uma correspondência máxima entre as distâncias de objetos provenientes da matriz  $\Delta$  e as distâncias obtidas pela matriz  $D$ .

Kruskal (1964) propôs uma medida de adequação de ajuste para avaliar o quanto as distâncias derivadas dos dados de dissimilaridades  $f(\delta_{ij})$  se aproximam daquelas originais fornecidas pelos respondentes (transformadas).

Medida: STRESS (STandardized Residual Sum of Squares)

$$Stress = \sqrt{\frac{\sum_i \sum_j (f(\delta_{ij}) - d_{ij})^2}{\sum_i \sum_j d_{ij}^2}}$$

Quanto maior o valor do STRESS, pior o ajuste

Valores de referência para o STRESS

STRESS	Adequação do Ajuste
20%	Pobre
10%	Razoável
5%	Bom
2,5%	Excelente
0%	Perfeito

Outra medida: SSTRESS - coeficiente de Young, encontrado no algoritmo ALSCAL (SPSS).

$$SStress = \sqrt{\frac{\sum_i \sum_j (f^2(\delta_{ij}) - d_{ij}^2)^2}{\sum_i \sum_j d_{ij}^4}}$$

Ainda para medir a qualidade do ajuste: índice RSQ. Correlação quadrática ( $R^2$ ) entre as distâncias originais e as derivadas dos dados de dissimilaridade.

$$RSQ = \frac{\left( \sum_i \sum_j [f(\delta_{ij}) - f(\delta_{..})] \cdot [d_{ij} - d_{..}] \right)^2}{\left( \sum_i \sum_j [f(\delta_{ij}) - f(\delta_{..})]^2 \right) \cdot \left( \sum_i \sum_j [d_{ij} - d_{..}]^2 \right)}$$

Os subscritos (..) representam a média do elemento correspondente ao sub-índice.

## 4 EMD não métrico: Um exemplo prático

Pesquisa: Avaliar a percepção entre “proximidades” de marcas de automóveis, para uma quantidade de seis marcas.

(Note que não estamos falando em termos de distâncias)

Marca	Nome
1	Peugeot
2	Renault
3	Citroën
4	Toyota
5	Honda
6	Fiat

Dados em escala ordinal. Dados de preferência ou percepção quanto à proximidade dos pares que estão sendo avaliados.



## 4 EMD não métrico: Um exemplo prático

### 4.1 Preparação da Modelagem

A partir da pesquisa obtivemos uma hierarquia de similaridades, apresentada abaixo.

	Peugeot	Renault	Citroën	Toyota	Honda	Fiat
Peugeot						
Renault	14					
Citroën	3	4				
Toyota	12	10	7			
Honda	13	11	6	15		
Fiat	8	9	5	2	1	

Matriz de Similaridades entre pares de Marcas

## 4 EMD não métrico: Um exemplo prático

As medidas de dissimilaridades são então obtidas da tabela anterior:

	Peugeot	Renault	Citroën	Toyota	Honda	Fiat
Peugeot						
Renault	2					
Citroën	13	12				
Toyota	4	6	9			
Honda	3	5	10	1		
Fiat	8	7	11	14	15	

Matriz de Dissimilaridades entre pares de Marcas

Esta análise é mais direta. Por ex. as marcas mais próximas são Toyota e Honda.

No SPSS: arquivo **MarcasAutomoveis.sav**

## 4 EMD não métrico: Um exemplo prático

MarcasAutomoveis.sav [DataSet1] - SPSS Statistics Data Editor

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

1 : Peugeot

Visible: 6 of 6 Variables

	Peugeot	Renault	Citroën	Toyota	Honda	Fiat	var
1	.	.	.	.	.	.	.
2	2,00	.	.	.	.	.	.
3	13,00	12,00	.	.	.	.	.
4	4,00	6,00	9,00	.	.	.	.
5	3,00	.	.	.	.	.	.
6	8,00	.	.	.	.	.	.
7	.	.	.	.	.	.	.
8	.	.	.	.	.	.	.
9	.	.	.	.	.	.	.

Data View Variable View

MarcasAutomoveis.sav [DataSet1] - SPSS Statistics Data Editor

File Edit View Data Transform **Analyze** Graphs Utilities Add-ons Window Help

1 : Peugeot

Visible: 6 of 6 Variables

	Peugeot	Renault	Citroën	Toyota	Honda	Fiat	var
1	.	.	.	.	.	.	.
2	2,00	.	.	.	.	.	.
3	13,00	.	.	.	.	.	.
4	4,00	.	.	.	.	.	.
5	3,00	.	.	.	.	.	.
6	8,00	.	.	.	.	.	.
7	.	.	.	.	.	.	.
8	.	.	.	.	.	.	.
9	.	.	.	.	.	.	.

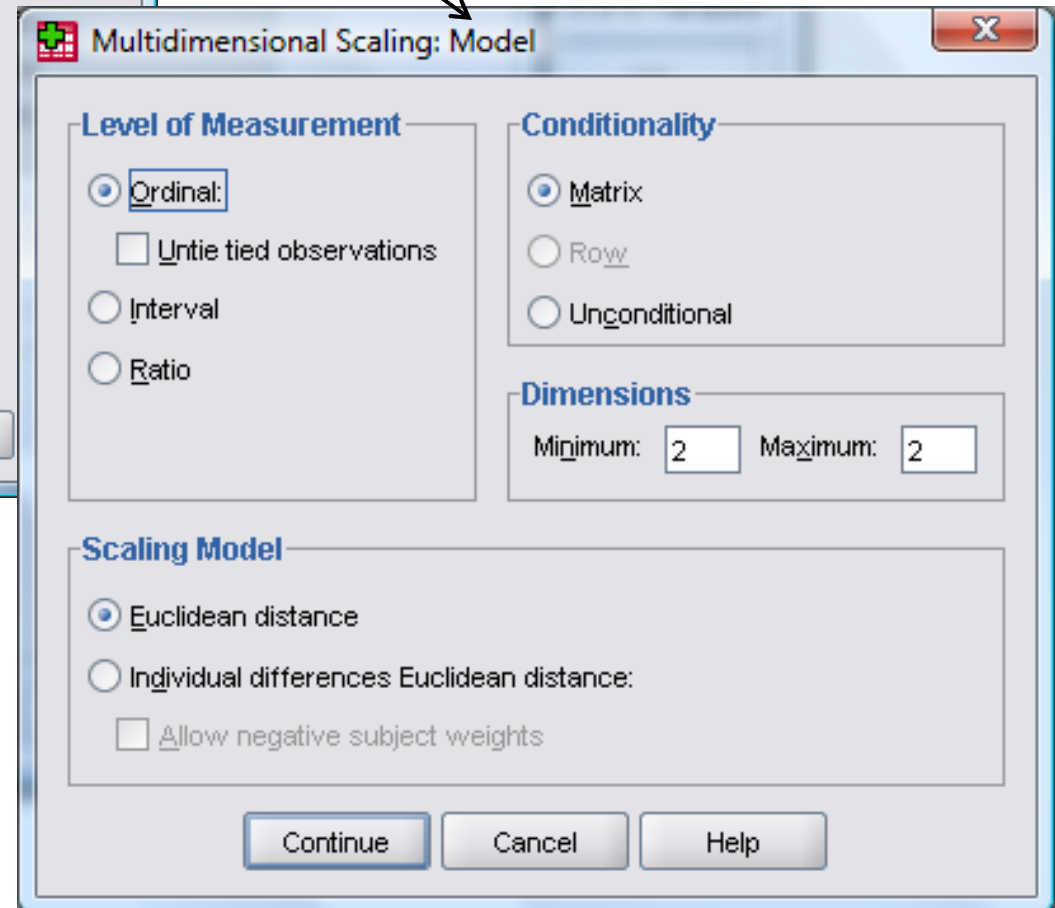
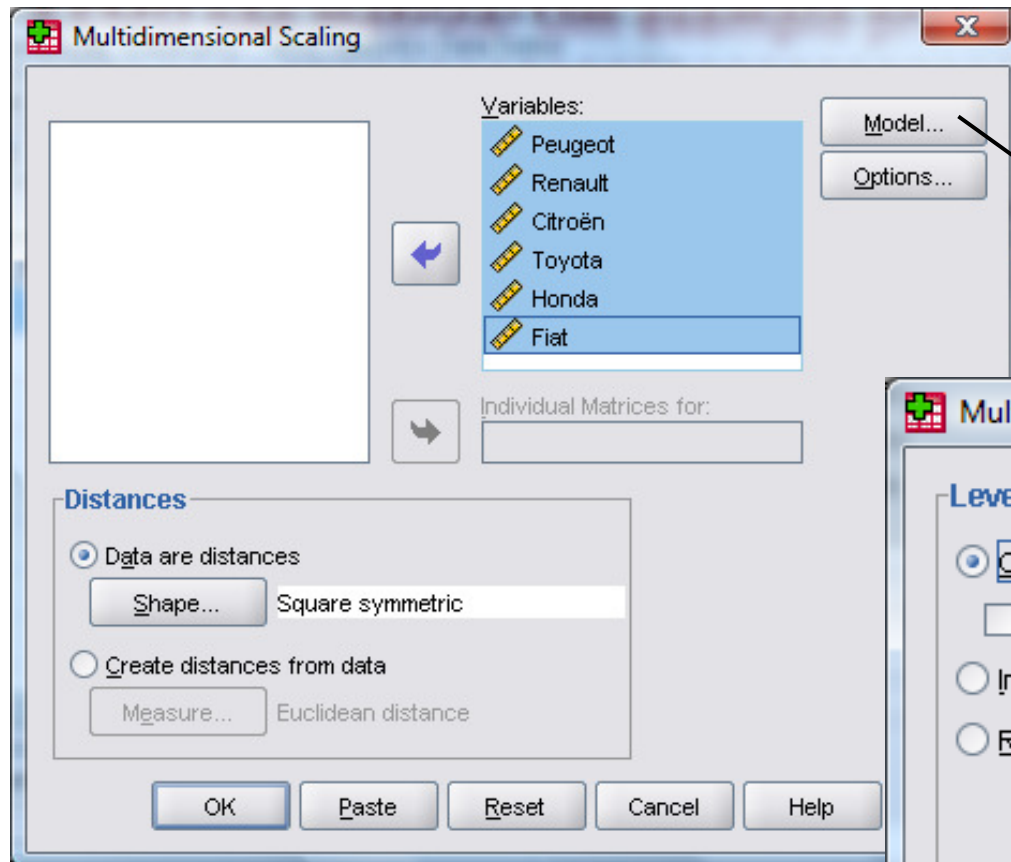
Data View Variable View

Multidimensional Scaling (ALSCAL)...

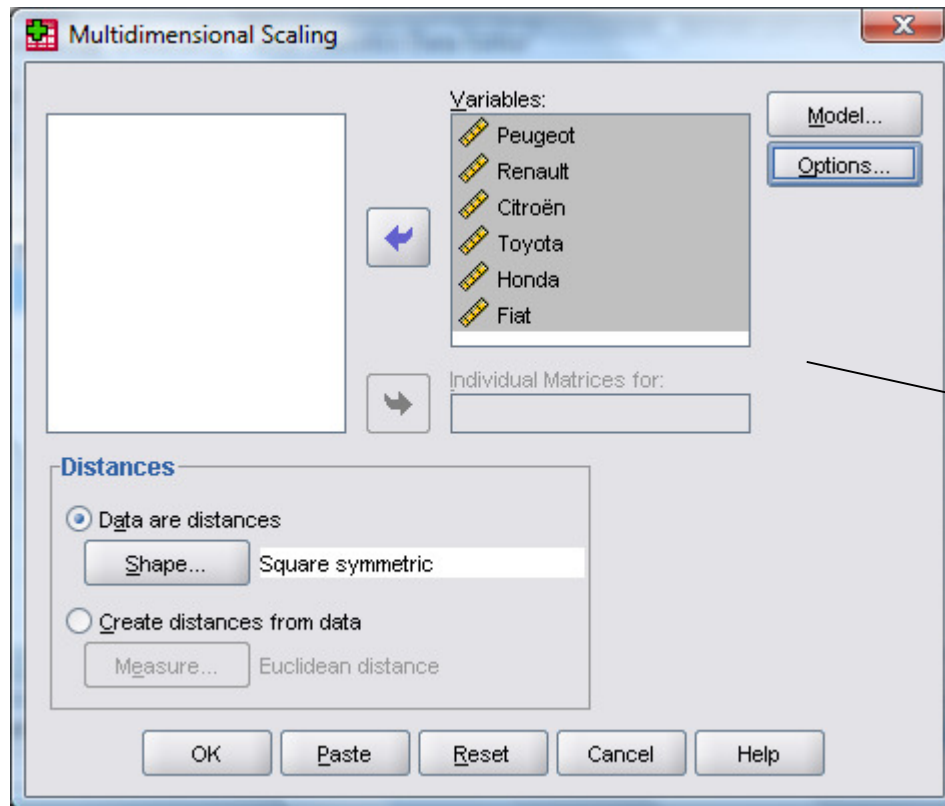
- Reports
- Descriptive Statistics
- Compare Means
- General Linear Model
- Correlate
- Regression
- Classify
- Dimension Reduction
- Scale**
  - Reliability Analysis...
  - Multidimensional Scaling (ALSCAL)...**
- Nonparametric Tests
- Forecasting
- Multiple Response
- Quality Control
- ROC Curve...
- Amos 7...

SPSS Statistics Processor is ready

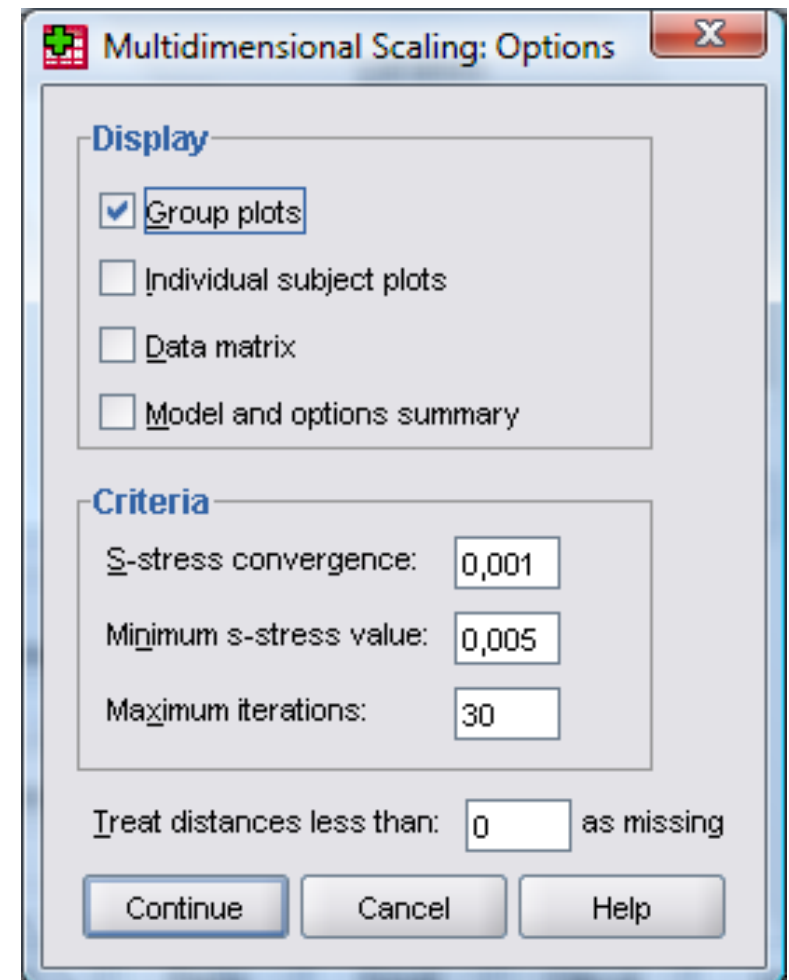
## 4 EMD não métrico: Um exemplo prático



## 4 EMD não métrico: Um exemplo prático



É importante escolher o mapa perceptual (group plots)



## 4 EMD não métrico: Um exemplo prático

### 4.2 análise dos resultados

#### Medidas de qualidade do ajuste

Stress	SStress	RSQ
0,00366	0,0049	0,99992

#### Coordenadas de cada estímulo, para duas dimensões

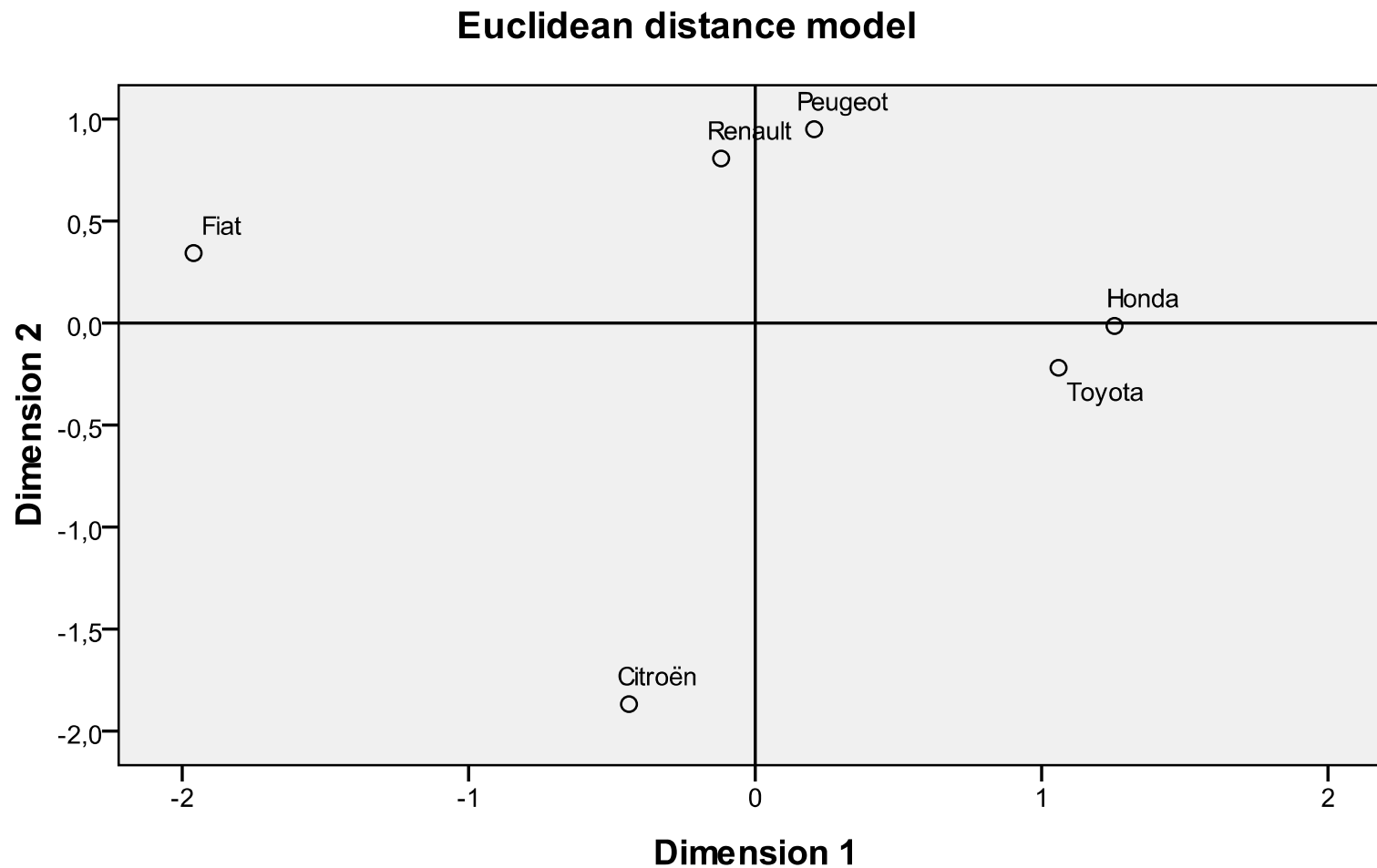
Configuration derived in 2 dimension

		Stimulus Coordinates Dimension	
Stimulus Number	Stimulus Name	1	2
1	Peugeot	0,2063	0,9503
2	Renault	-0,1189	0,8071
3	Citroën	-0,4406	-1,8678
4	Toyota	1,0589	-0,2192
5	Honda	1,2549	-0,0139
6	Fiat	-1,9605	0,3435

## 4 EMD não métrico: Um exemplo prático

### Representação Gráfica das coordenadas (bi-dimensional)

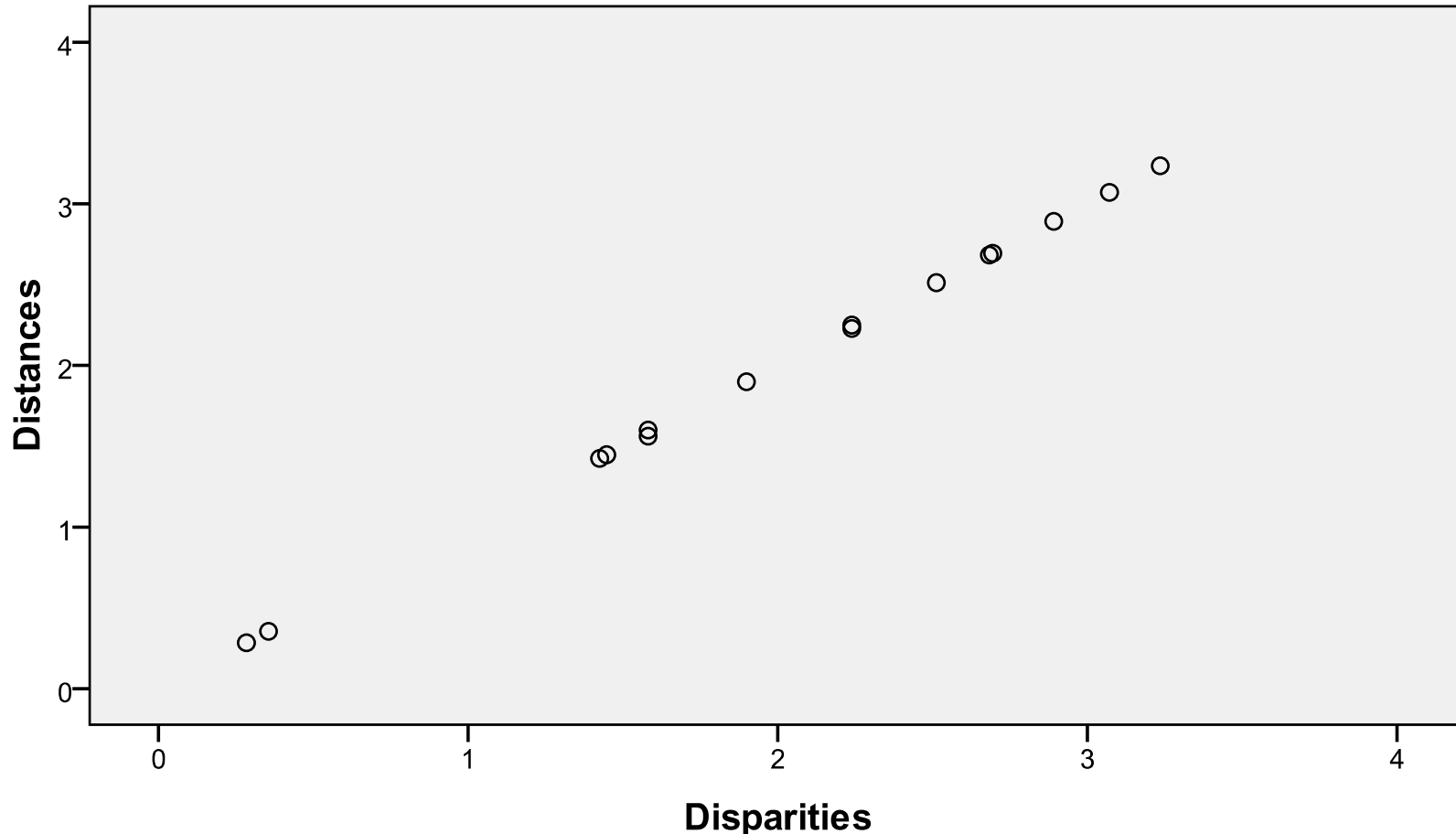
Derived Stimulus Configuration



## 4 EMD não métrico: Um exemplo prático

**Gráfico de Ajuste Linear** entre as distâncias derivadas dos dados de dissimilaridades e as distâncias originais transformadas. RSQ alto.

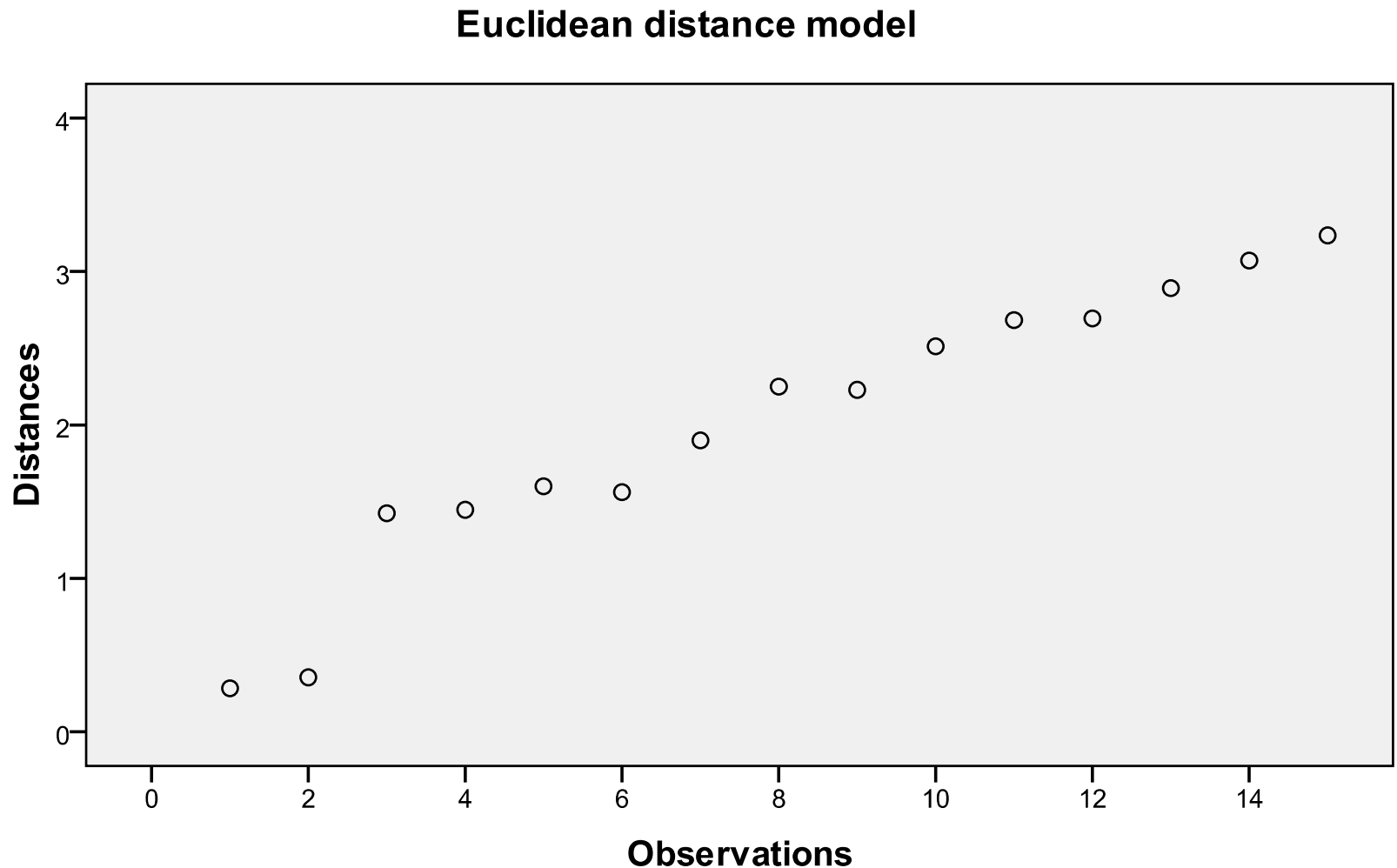
Scatterplot of Linear Fit  
Euclidean distance model





## 4 EMD não métrico: Um exemplo prático

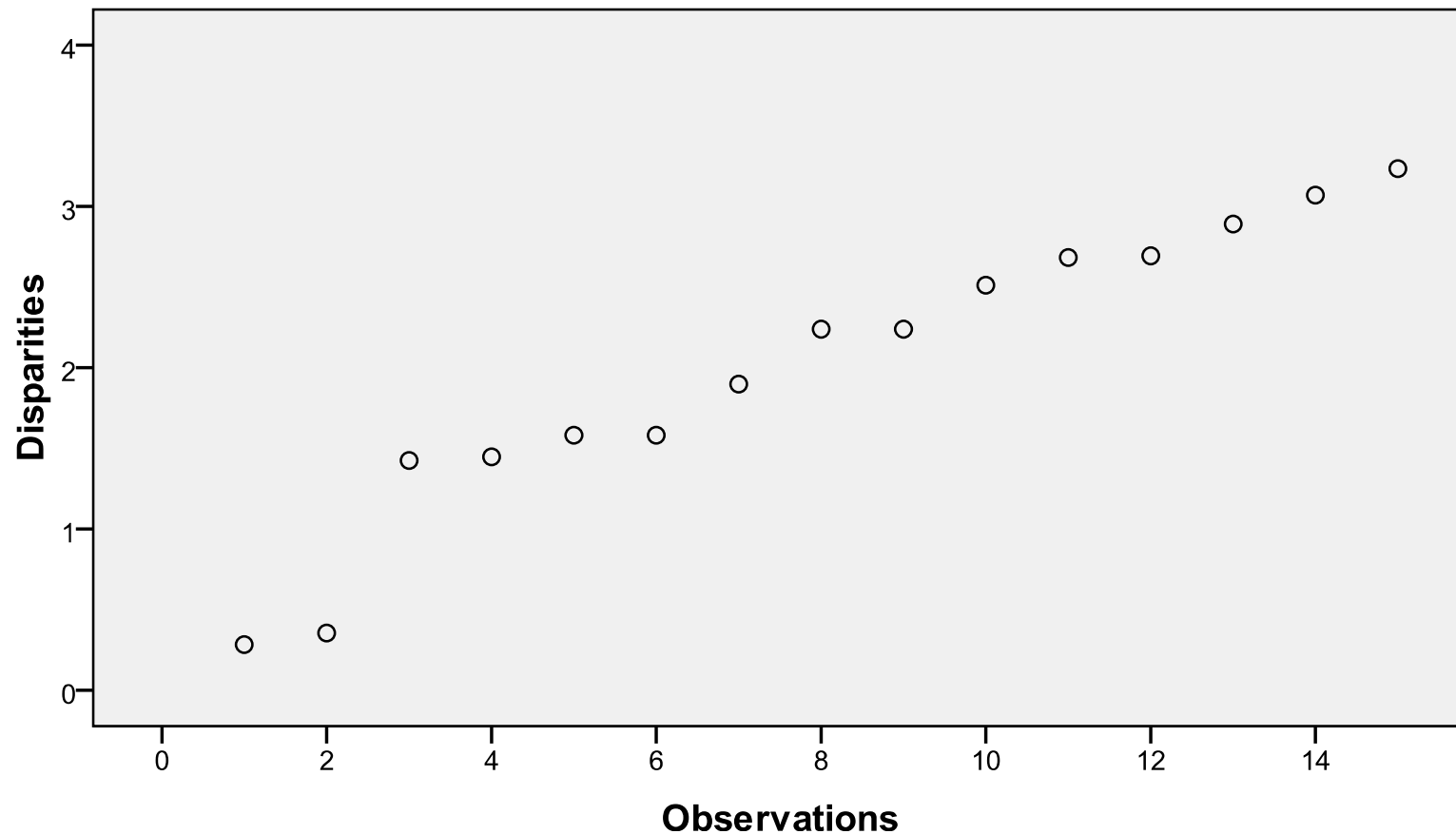
Relação entre as **distâncias derivadas dos dados de dissimilaridades** e a **posição inicial do Ranking**. Ajuste melhor para maiores números no ranking



## 4 EMD não métrico: Um exemplo prático

Relação entre as **distâncias originais transformadas (disparidades)** e a **posição inicial do Ranking**. O gráfico está na forma de escada

Transformation Scatterplot  
Euclidean distance model



## 5 EMD métrico: Um exemplo prático

---

- No EMD métrico os dois últimos gráficos não são oferecidos pelo software, pois não se trabalha com medidas na forma de ranking. Trabalha-se com medidas de distâncias ou dissimilaridades propriamente ditas, e não com medidas de preferência ou percepção.

- No EMD métrico os dados devem estar em uma escala quantitativa.

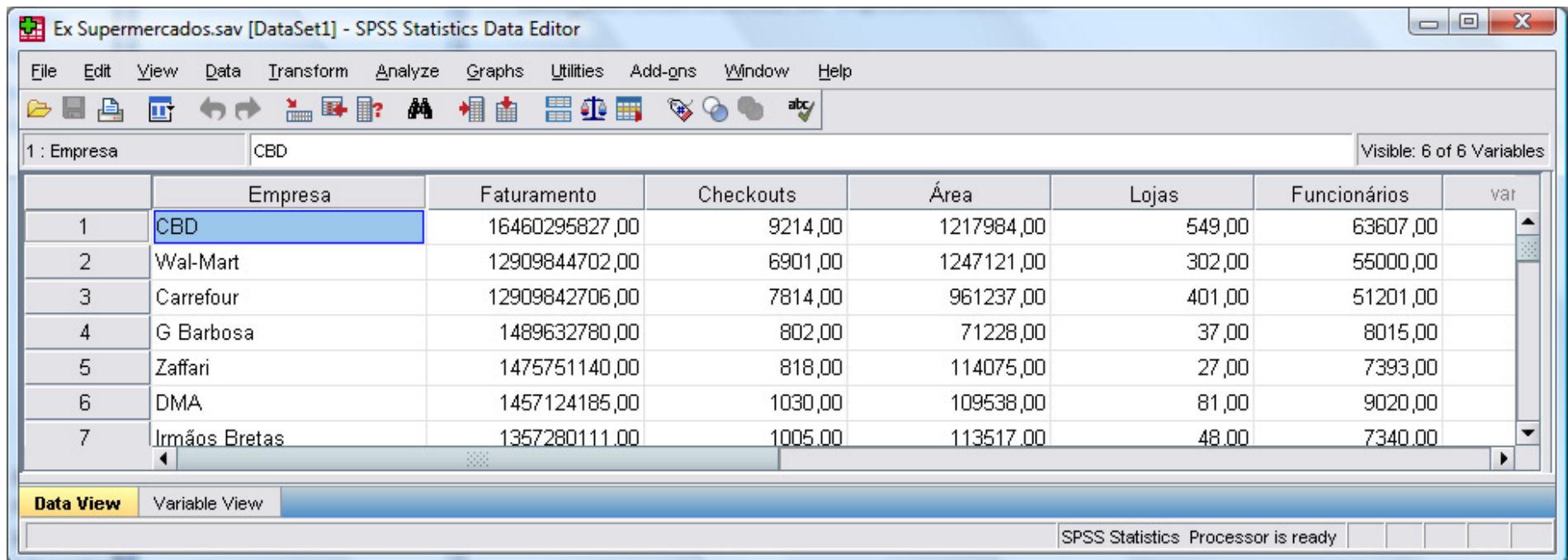
As distâncias ou correlações devem ser calculadas com variáveis padronizadas.

- No EMD não métrico os diferentes mapas são definidos por diferentes percepções, no EMD as diferenças estão baseadas nas variáveis.

Ex. duas pessoas podem apresentar grande similaridade entre altura e peso e enorme dissimilaridade entre renda e nível de escolaridade.

## 5 EMD métrico: Um exemplo prático

Exemplo: Dez maiores grupos supermercadistas brasileiros no ano de 2006. Arquivo **Ex Supermercados.sav**



Ex Supermercados.sav [DataSet1] - SPSS Statistics Data Editor

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

1 : Empresa CBD Visible: 6 of 6 Variables

	Empresa	Faturamento	Checkouts	Área	Lojas	Funcionários	var
1	CBD	16460295827,00	9214,00	1217984,00	549,00	63607,00	
2	Wal-Mart	12909844702,00	6901,00	1247121,00	302,00	55000,00	
3	Carrefour	12909842706,00	7814,00	961237,00	401,00	51201,00	
4	G Barbosa	1489632780,00	802,00	71228,00	37,00	8015,00	
5	Zaffari	1475751140,00	818,00	114075,00	27,00	7393,00	
6	DMA	1457124185,00	1030,00	109538,00	81,00	9020,00	
7	Irmãos Bretas	1357280111,00	1005,00	113517,00	48,00	7340,00	

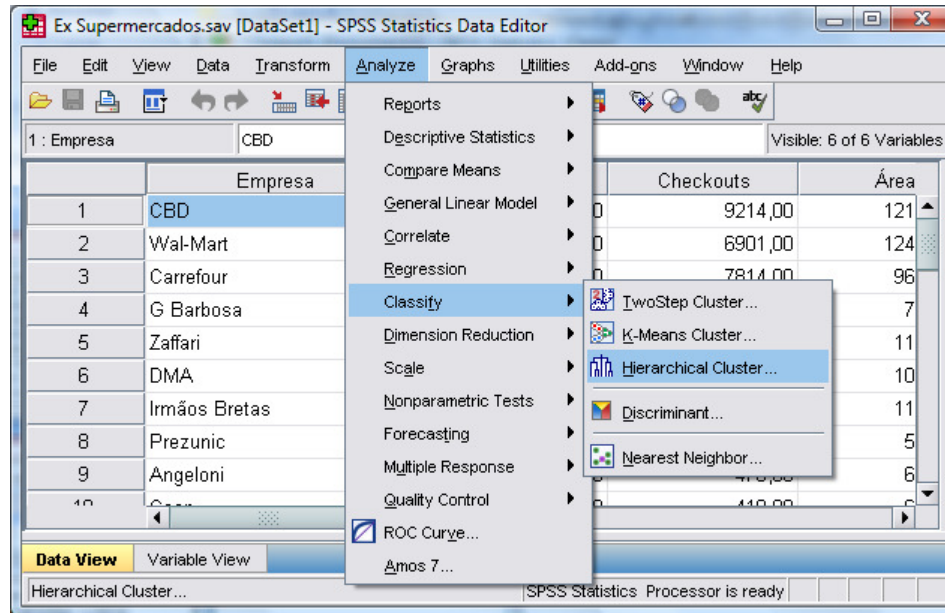
Data View Variable View

SPSS Statistics Processor is ready

Este banco de dados não é a base para a entrada do software para a elaboração do EMD. Precisa ser transformado em uma matriz de dissimilaridades ou distâncias

## 5 EMD métrico: Um exemplo prático

### Análise de cluster: Fornece medida de dissimilaridade



Proximity Matrix

Case	Euclidean Distance									
	1:CBD	2:Wal-Mart	3:Carrefour	4:G Barbosa	5:Zaffari	6:DMA	7:Irmãos Bretas	8:Prezunic	9:Angeloni	10:Coop
1:CBD	,000	1,584	1,254	5,344	5,346	5,159	5,278	5,451	5,489	5,548
2:Wal-Mart	1,584	,000	,812	4,137	4,120	3,981	4,072	4,241	4,269	4,335
3:Carrefour	1,254	,812	,000	4,115	4,115	3,937	4,050	4,222	4,261	4,321
4:G Barbosa	5,344	4,137	4,115	,000	,101	,251	,120	,117	,157	,237
5:Zaffari	5,346	4,120	4,115	,101	,000	,293	,122	,135	,157	,229
6:DMA	5,159	3,981	3,937	,251	,293	,000	,185	,338	,382	,431
7:Irmãos Bretas	5,278	4,072	4,050	,120	,122	,185	,000	,187	,235	,285
8:Prezunic	5,451	4,241	4,222	,117	,135	,338	,187	,000	,076	,128
9:Angeloni	5,489	4,269	4,261	,157	,157	,382	,235	,076	,000	,121
10:Coop	5,548	4,335	4,321	,237	,229	,431	,285	,128	,121	,000

This is a dissimilarity matrix