



**Data Mining**

**Partitioning Clustering (k-means)**

**Note on Profiling**

**17/11/2021**

**NOVA-IMS**

Fernando Lucas Bação  
[bacao@isegi.unl.pt](mailto:bacao@isegi.unl.pt)  
<http://www.isegi.unl.pt/bacao>

Instituto Superior de Estatística e Gestão de Informação  
Universidade Nova de Lisboa

1



**AGENDA**

- Cluster analysis
  - Clustering techniques
    - Partitioning Methods (kmeans and k-meadoids)
    - **A Note on Profiling**

Instituto Superior de Estatística e Gestão de Informação  
Universidade Nova de Lisboa

2

**NOVA**  
**IMS**  
Information Management School



# Profiling

Instituto Superior de Estatística e Gestão de Informação  
Universidade Nova de Lisboa



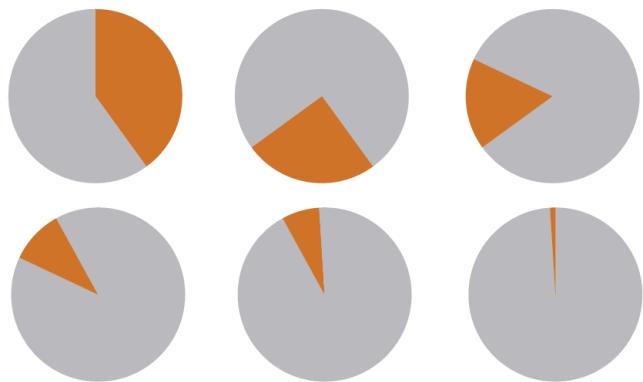
3

**NOVA**  
**IMS**  
Information Management School

## Clustering

- Profiling (size of the clusters)**

Part-to-Whole Mini-Pie Charts



Instituto Superior de Estatística e Gestão de Informação  
Universidade Nova de Lisboa

4

**NOVA  
IMS**  
Information Management School

## Clustering

- Profiling (comparing averages)

A scatter plot illustrating the comparison of averages for four variables (A, B, C, D) across two clusters. The x-axis represents the 'Normalized Mean' from 0 to 1, and the y-axis represents the 'Variables' A, B, C, and D. Two cluster averages are shown: a 'Database average' (represented by a grey vertical bar) and a 'Cluster average' (represented by a dark red square). Data points are represented by blue squares.

Variable	Database Average (Normalized Mean)	Cluster Average (Normalized Mean)
A	~0.85	~0.95
B	~0.15	~0.15
C	~0.35	~0.35
D	~0.25	~0.25

Instituto Superior de Estatística e Gestão de Informação  
Universidade Nova de Lisboa

5

**NOVA  
IMS**  
Information Management School

## Clustering

- Profiling (comparing profiles)

Four line graphs comparing trends in different health categories from 1975 to 2010. Each graph shows a central blue line with data points and surrounding grey lines representing confidence intervals. The graphs are labeled: Circulatory, Mental, Musculoskeletal, and Cancer.

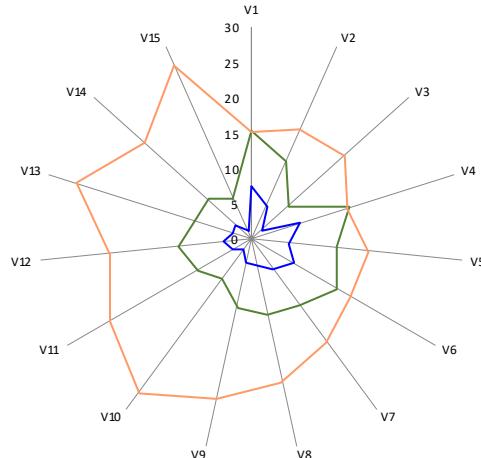
Category	Year	Value
Circulatory	1975	32
	2010	11
Mental	1975	11
	2010	23
Musculoskeletal	1975	17
	2010	26
Cancer	1975	10
	2010	14

JA Schwabish "An Economist's Guide to Visualizing Data", The Journal of Economic Perspectives, 2014

Instituto Superior de Estatística e Gestão de Informação  
Universidade Nova de Lisboa

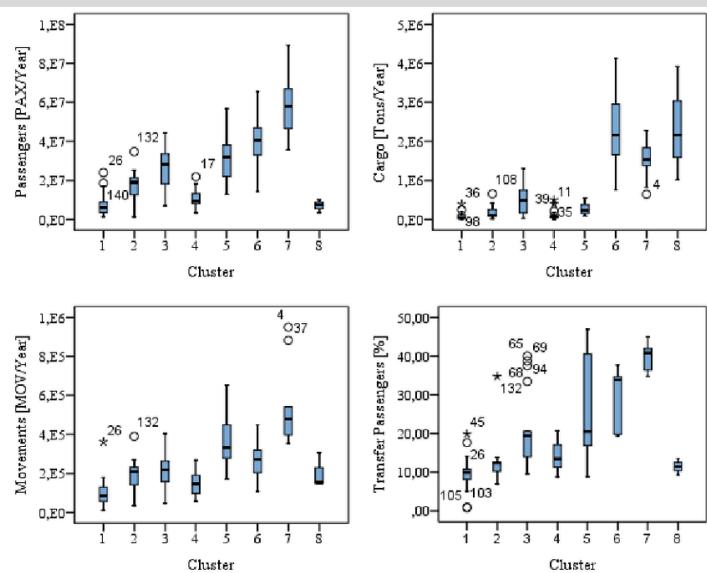
6

- Profiling (comparing profiles)



Instituto Superior de Estatística e Gestão de Informação  
Universidade Nova de Lisboa

7



Instituto Superior de Estatística e Gestão de Informação  
Universidade Nova de Lisboa

8

The image shows a scatter plot with four data points labeled 1, 2, 3, and 4. Point 1 is at the far left, point 2 is in the middle-left, point 3 is in the middle-right, and point 4 is at the far right. A callout box contains the text: "Helps building a sense of clusters and the distances between them".

- Profiling (multidimensional scaling)

9

**NOVA**  
**IMS**  
Information Management School

## Clustering

- Profiling (multidimensional scaling)

The diagram illustrates the results of a multidimensional scaling (MDS) analysis for four customer segments. The segments are represented as boxes, and arrows indicate their approximate positions in the MDS space.

- Good Customers**  
54.3 days (DPC/NC)  
144.7 eur/day (TC/DPC)  
7.761 eur/purchase (TC/NC)  
190.4 days since last purchase
- Recent**  
60 days (DPC/NC)  
146.8 eur/day (TC/DPC)  
8.367 eur/purchase (TC/NC)  
161.6 days since last purchase
- Gold**  
33 days (DPC/NC)  
400 eur/day (TC/DPC)  
14.155 eur/purchase (TC/NC)  
127.7 days since last purchase
- Recovering**  
181.6 days (DPC/NC)  
38.2 eur/day (TC/DPC)  
7.041 eur/purchase (TC/NC)  
1128.8 days since last purchase

Arrows indicate the approximate positions of the segments in the MDS space:

- An arrow points from **Good Customers** to **Recent**, labeled with an asterisk and the number 4.
- An arrow points from **Recent** to **Recovering**, labeled with an asterisk and the number 2.
- An arrow points from **Gold** to **Recovering**, labeled with an asterisk and the number 3.

10

**NOVA  
IMS**  
Information Management School

### Clustering

- Profiling (comparing variables not used)

**Used and unused variable distribution**

Instituto Superior de Estatística e Gestão de Informação  
Universidade Nova de Lisboa

11

**NOVA  
IMS**  
Information Management School

### Clustering

- Profiling (leverage)

**Leverage**  
Ratio  
Sales(%) / Individuals(%)

Cluster	Leverage Ratio
C1	8.36
C2	1.96
C3	0.52
C4	0.48

Instituto Superior de Estatística e Gestão de Informação  
Universidade Nova de Lisboa

12

- Exploring two solutions

		Sofist Digital	Variedade Serviço	Preço	Status_PC	ProAtividade	Tempo Internet e RS	Sustentabilidade	Life Status	Propensao_Pfi delizacao	Switching	
k=5	FREQ	F1_SD	F2_VS	F3_P	F4_SPC	F5_PA	F6_IRS	F7_CC	F8_LS	F9_FI	F10_AB	
1	222	-0,40	<b>-0,54</b>	<b>-0,31</b>	-0,29	0,13	<b>0,65</b>	<b>-1,67</b>	-0,20	-0,09	-0,03	Pouco Envolvidos e pouco sensíveis ao preço
2	463	<b>0,70</b>	<b>0,44</b>	<b>0,38</b>	0,13	<b>-0,32</b>	<b>-0,37</b>	-0,19	<b>0,31</b>	<b>-0,58</b>	0,03	Social Customers
3	469	<b>0,58</b>	-0,16	0,15	<b>0,48</b>	<b>0,23</b>	0,00	0,09	<b>-0,22</b>	<b>0,98</b>	<b>-0,04</b>	Sofisticado com Propensao Fidelizacao
4	428	<b>-1,47</b>	0,27	0,25	-0,13	0,01	<b>-0,25</b>	0,29	0,15	0,15	<b>0,04</b>	Info Excluidos
5	420	0,29	-0,30	<b>-0,68</b>	<b>-0,39</b>	0,02	0,31	<b>0,69</b>	-0,16	<b>-0,56</b>	-0,02	Sustentaveis
k=6	FREQ	F1_SD	F2_VS	F3_P	F4_SPC	F5_PA	F6_IRS	F7_CC	F8_LS	F9_FI	F10_AB	
1	241	0,12	<b>-0,30</b>	-0,09	-0,20	-0,04	<b>1,34</b>	<b>-1,29</b>	-0,03	<b>-0,13</b>	-0,24	Pouco Envolvidos
2	134	-0,16	0,11	0,02	-0,15	-0,10	0,13	0,10	0,00	0,07	<b>2,84</b>	Switchers
3	285	0,43	<b>0,48</b>	<b>-1,28</b>	-0,16	<b>0,52</b>	<b>-0,28</b>	0,05	0,04	<b>0,23</b>	-0,07	Exigentes
4	426	<b>-1,42</b>	0,22	<b>0,30</b>	-0,20	-0,03	<b>-0,27</b>	0,10	0,01	<b>0,21</b>	-0,27	Info Excluidos
5	414	0,39	<b>-0,77</b>	0,20	<b>0,92</b>	0,19	<b>-0,23</b>	0,15	<b>0,21</b>	0,15	-0,16	Status/Conservadores
6	502	<b>0,63</b>	<b>0,30</b>	<b>0,34</b>	<b>-0,37</b>	<b>-0,38</b>	-0,10	<b>0,35</b>	<b>-0,19</b>	<b>-0,39</b>	-0,24	Social Customers

Instituto Superior de Estatística e Gestão de Informação  
Universidade Nova de Lisboa

13

- Exploring two solutions with different k

K=7 VS K=8	Pouco envolvidos	Switchers	Informados / Social Customers	Fidelizáveis Sofisticados	Info Excluidos	Status	Proativos	Desprendido	Grand Total
Pouco envolvidos	71	106	23	13	45	1	1	49	203
Switchers	2	4	4	2	1	3	2	122	
Informados / Social Customers	4	219	68	2	20	1	64	378	
Fidelizáveis Sofisticados	72	105	191	4	1	2	69	444	
Status/Conservadores	7	5	39	13	157	42	263		
Proativos	5	15	20	3	153	19	215		
Info Excluidos	12	4	309	1	1	50	377		
Grand Total	161	106	383	339	378	181	158	296	2002

K=6 VS K=8	Pouco envolvidos	Switchers	Informados / Social Customers	Fidelizáveis Sofisticados	Info Excluidos	Status	Proativos	Desprendido	Grand Total
Pouco Envolvidos	146	16	24	8	17	7	23	241	
Switchers	2	105	6	11	1	2	4	3	134
Info Excluidos	1	1	26	56	13	27	60	100	285
Status/Conservadores	3	13	22	50	17	11	25	426	
Social Customers	9	322	49	5	98	58	60	83	502
Grand Total	161	106	383	339	378	181	158	296	2002

Instituto Superior de Estatística e Gestão de Informação  
Universidade Nova de Lisboa

14

**NOVA  
IMS**  
Information Management School

## Clustering

- **Consolidation of two solutions**

The diagram illustrates the process of consolidating two segments into a larger number of consolidations. It features three main components: 'Value Segment (4)' at the top left, 'Buying Segment (4)' below it, and 'Consolidations (16)' at the bottom right. Each segment is represented by a blue rounded rectangle containing four small colored squares (blue, orange, red, green). Arrows point from both the 'Value Segment' and the 'Buying Segment' towards the 'Consolidations' box, indicating their merging into a total of 16 consolidations.

Instituto Superior de Estatística e Gestão de Informação  
Universidade Nova de Lisboa

15