# Model based approach for utility max and medium

Questo metodo utilizza distribuzioni gaussiane per definire il numero di clusters. I risultati suggerisce la costruzione di 3 cluster con il metodo VEV

Mclust VEV (ellipsoidal, equal shape) model with 3 components:

log.likelihood n df BIC ICL

3621.836 68 428 5437.723 5437.723

Clustering table:

1 2 3

16 44 8



Figura 1: BIC - The “best model” is selected using the Bayesian Information Criterion or BIC. A large BIC score indicates strong evidence for the corresponding model.

Mixing probabilities:

1 2 3

0.2352941 0.6470588 0.1176471

Means:

[,1] [,2] [,3]

wm\_EL 0.0733125 0.091772727 0.017875

w\_EL 0.0733125 0.189568182 0.017875

wm\_ET 0.0633125 0.068454545 0.015750

w\_ET 0.0633125 0.048681818 0.015750

wm\_HT 0.0770625 0.084295455 0.020750

w\_HT 0.0749375 0.088568182 0.019000

wm\_HL 0.0415000 0.113250000 0.014125

w\_HL 0.0436875 0.062204545 0.015875

wm\_FS 0.2136875 0.112318182 0.041125

w\_FS 0.2143125 0.095818182 0.041125

wm\_LS 0.1420000 0.100772727 0.029500

w\_LS 0.1420000 0.117159091 0.029500

wm\_CL 0.0000000 0.006795455 0.003750

w\_CL 0.0000000 0.009681818 0.000000

wm\_OT 0.1393125 0.271295455 0.026500

w\_OT 0.3890625 0.385681818 0.861000

To visualise the clustering structure and the geometric characteristics induced by an estimated Gaussian finite mixture model we may project the data onto a suitable dimension reduction subspace. The function MclustDR() implements the methodology introduced in [Scrucca (2010)](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5096736/#R58). The estimated directions which span the reduced subspace are defined as a set of linear combinations of the original features, ordered by importance as quantified by the associated eigenvalues.



Figura 2: Dimension reduction contour plot of estimated mixture densities on the projection subspace estimated with MclustDR.

-----------------------------------------------------------------

Dimension reduction for model-based clustering and classification

-----------------------------------------------------------------

Mixture model type: Mclust (VEV, 3)

Clusters n

1 16

2 44

3 8

Estimated basis vectors:

Dir1 Dir2

wm\_EL -0.0074591 -0.0067197

w\_EL -0.1915406 -0.2385033

wm\_ET 0.0349016 -0.0548816

w\_ET 0.1657752 -0.7005038

wm\_HT -0.0200808 0.0136669

w\_HT -0.1681336 -0.2403694

wm\_HL -0.0201676 0.0115869

w\_HL -0.1888760 0.1313844

wm\_FS 0.0125940 0.0304077

w\_FS -0.3952475 0.1558014

wm\_LS -0.0297661 0.0838565

w\_LS -0.1545048 -0.3086920

wm\_CL 0.4832073 -0.3982100

w\_CL -0.6555765 0.1726927

wm\_OT -0.1406721 -0.1094731

w\_OT -0.0748904 -0.2146938

Dir1 Dir2

Eigenvalues 1.2518 0.93718

Cum. % 57.1855 100.00000

# Three popular methods for determining the optimal number of clusters

In this section we describe the three most popular methods including: i) Elbow method, ii) silhouette method and iii) gap statistic.

See <http://www.sthda.com/english/wiki/determining-the-optimal-number-of-clusters-3-must-known-methods-unsupervised-machine-learning#three-popular-methods-for-determining-the-optimal-number-of-clusters>

## 5.1 Elbow method

## for Kmeans

The **total within-cluster sum of square (wss)** measures the compactness of the clustering and we want it to be as small as possible.



Il numero ideale di cluster è 4 con kmeans

the optimal number of clusters K for k-means clustering is 4

## PAM

**Con pam il numero di cluster scende a 2!**

## HC

5 cluster



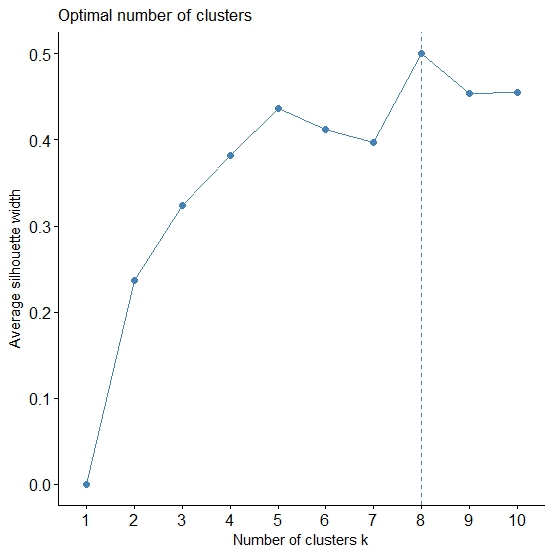
## Conclusioni:

Per il metodo elbow abbiamo avuto risultati diversi: 2-3-8 cluster rispettivamente per kmeans, pam e hc clustering

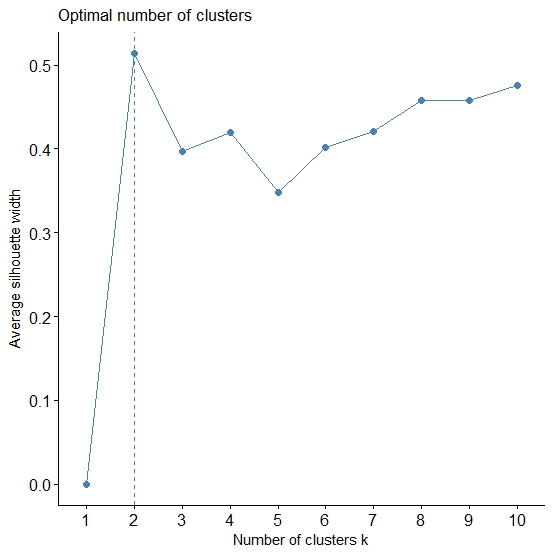
# The average silhouette method

It measures the quality of a clustering. That is, it determines how well each object lies within its cluster. A high average silhouette width indicates a good clustering.

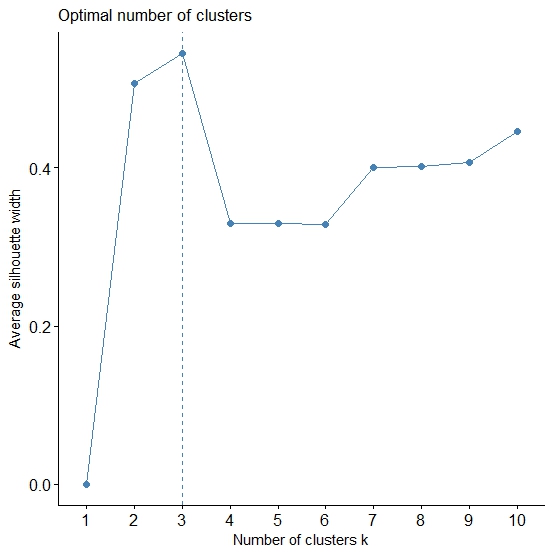
Average silhouette method computes the average silhouette of observations for different values of k. The optimal number of clusters k is the one that maximize the average silhouette over a range of possible values for k (Kaufman and Rousseeuw [1990]).



For kmeans the silhouette method, suggest 8 clusters



For Pam the silhouette method suggest 2 cluster



The silhouette for hc method suggest 3 clusters.

Proviamo ora con il GAP statistic method.

# The gap statistic method

The gap statistic compares the total within intracluster variation for different values of k with their expected values under null reference distribution of the data, i.e. a distribution with no obvious clustering.

Tale metodo restituisce sempre un alto numero di cluster: >10

# The NbClust count

The nbclust function racchiude un alto numero di indici e definisce il numero ottimale di cluster.

In the plot of Hubert index, we seek a significant knee that corresponds to a

significant increase of the value of the measure i.e the significant peak in Hubert

index second differences plot.

\*\*\* : The D index is a graphical method of determining the number of clusters.

In the plot of D index, we seek a significant knee (the significant peak in Dindex

second differences plot) that corresponds to a significant increase of the value of

the measure.



\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* Among all indices:

\* 4 proposed 2 as the best number of clusters

\* 7 proposed 3 as the best number of clusters

\* 1 proposed 4 as the best number of clusters

\* 2 proposed 5 as the best number of clusters

\* 1 proposed 6 as the best number of clusters

\* 1 proposed 9 as the best number of clusters

\* 2 proposed 10 as the best number of clusters

\* 1 proposed 12 as the best number of clusters

\* 5 proposed 15 as the best number of clusters

\*\*\*\*\* Conclusion \*\*\*\*\*

\* According to the majority rule, the best number of clusters is 3

$Best.nc

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | KL | CH | Hartigan | CCC | Scott | Marriot | TrCovW | TraceW | Friedman |  |
| Number\_clusters | 5 | 15 | 4 | 15 | 10 | 3 | 3 |  | 5 | 10. |
| Value\_Index | 2.746 | 72.4008 | 18.299 | 37.2717 | 679.9275 | 0 | 0.6302 |  | 1.346 | 142137.3 |
| Rubin | Cindex | DB | Silhouette | Duda | PseudoT2 | Beale | Ratkowsky | Ball |  |  |
| Number\_clusters | 12 | 15 | 2 | 15 | 3 | 3 | 9 | 6 | 3 |  |
| Value\_Index | -2.939 | 0.106 | 0.5452 | 0.5944 | 0.8191 | 11.4807 | -2.361 | 0.2894 | 2.4198 |  |
|  | PtBiserial | Frey | McClain | Dunn | Hubert | SDindex | Dindex | SDbw |  |  |
| Number\_clusters | 2 | 3 | 2 | 2 | 0 | 3 | 0 | 15 |  |  |
| Value\_Index | 0.5462 | 2.2067 | 0.1028 | 0.4696 | 0 | 5.8257 | 0 | 0.0812 |  |  |

$Best.partition

[1] 1 1 1 1 1 1 1 1 1 2 1 1 2 1 1 1 1 3 1 1 3 3 1 1 3 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 2 1 1

[45] 1 2 3 1 1 1 1 2 1 1 1 2 1 1 1 1 2 1 3 1 1 1 1 1

# Risultati

3 cluster è la suddivisione Migliore per questi dati.

# Analisi delle distanze

(http://www.sthda.com/english/wiki/visual-enhancement-of-clustering-analysis-unsupervised-machine-learning)



Figura 3: ordered dissimilarity matrix image (ODI)

The ordered dissimilarity matrix image (ODI) displays the [clustering tendency](http://www.sthda.com/english/wiki/assessing-clustering-tendency-a-vital-issue-unsupervised-machine-learning) of the dataset. Similar objects are close to one another. Red color corresponds to small distance and blue color indicates big distance between observation.

## Eclust:

qual è il n. migliore di cluster? Usa sempre gap statistics (grafico del gap statistics è uguale!).

Per Kmeans il numero migliore è 10



Per hclust 10 cluster







# Kmeans

**Primo test con 3 cluster (come suggeriscono I dati) e con 8 cluster**

> u.kmeans$cluster

[1] 3 3 1 3 3 3 3 1 1 1 1 1 1 3 3 1 1 2 1 3 2 2 1 3 2 3 3 3 1 1 1 1 3 3 1 3 3 3 3 3 3 1 3 1

[45] 3 1 2 1 3 1 1 1 1 3 3 1 1 1 3 3 1 1 2 3 3 3 1 1

> u.kmeans$centers

wm\_EL w\_EL wm\_ET w\_ET wm\_HT w\_HT wm\_HL w\_HL

1 0.08693548 0.06951613 0.03722581 0.03229032 0.06806452 0.06758065 0.0826129 0.05754839

2 0.01933333 0.79200000 0.01600000 0.01600000 0.05250000 0.02100000 0.0140000 0.01400000

3 0.08203226 0.08870968 0.09358065 0.07045161 0.08654839 0.09764516 0.1004839 0.05467742

wm\_FS w\_FS wm\_LS w\_LS wm\_CL w\_CL wm\_OT w\_OT

1 0.06970968 0.06980645 0.07067742 0.06690323 0.0009677419 0.00000000 0.2620323 0.63154839

2 0.04300000 0.04300000 0.03000000 0.03000000 0.0000000000 0.00000000 0.0260000 0.08433333

3 0.20229032 0.17909677 0.14745161 0.17448387 0.0096451613 0.01374194 0.1967419 0.32254839

>



> u.kmeans$cluster

[1] 7 1 5 7 3 7 1 3 5 4 5 5 4 6 2 5 5 8 7 7 8 8 5 7 8 7 7 1 2 7 2 4 1 7 2 2 3 7 7 1 7 4 1 5

[45] 7 4 8 7 1 5 2 4 7 7 1 4 4 2 7 1 4 7 8 2 1 7 2 2

> u.kmeans$centers

wm\_EL w\_EL wm\_ET w\_ET wm\_HT w\_HT wm\_HL w\_HL

1 0.09020000 0.11090000 0.08290000 0.08590000 0.1082000 0.14600000 0.06180000 0.07380000

2 0.12960000 0.07560000 0.03890000 0.03890000 0.0467000 0.04670000 0.02580000 0.02580000

3 0.09900000 0.09900000 0.36700000 0.06700000 0.1056667 0.10566667 0.87666667 0.08766667

4 0.02022222 0.02022222 0.01688889 0.01688889 0.0220000 0.02044444 0.01466667 0.01622222

5 0.12055556 0.12055556 0.04266667 0.04266667 0.1411111 0.14100000 0.13711111 0.13722222

6 0.10100000 0.10100000 0.05300000 0.05300000 0.0690000 0.06900000 0.05500000 0.05500000

7 0.06875000 0.06875000 0.05735000 0.05735000 0.0695000 0.06780000 0.03745000 0.03920000

8 0.01933333 0.79200000 0.01600000 0.01600000 0.0525000 0.02100000 0.01400000 0.01400000

wm\_FS w\_FS wm\_LS w\_LS wm\_CL w\_CL wm\_OT w\_OT

1 0.14610000 0.14610000 0.14340000 0.227200000 0.016300000 0.02900000 0.17950000 0.18400000

2 0.12660000 0.12660000 0.16490000 0.164900000 0.000000000 0.00000000 0.37550000 0.52150000

3 0.10633333 0.10633333 0.08733333 0.087333333 0.045333333 0.04533333 0.30700000 0.40200000

4 0.04522222 0.04522222 0.03533333 0.035333333 0.003333333 0.00000000 0.03855556 0.84577778

5 0.03655556 0.03688889 0.02055556 0.007555556 0.000000000 0.00000000 0.44544444 0.49766667

6 0.81000000 0.08100000 0.20300000 0.203000000 0.000000000 0.00000000 0.43900000 0.43900000

7 0.19200000 0.19250000 0.13555000 0.135550000 0.000000000 0.00000000 0.14780000 0.43930000

8 0.04300000 0.04300000 0.03000000 0.030000000 0.000000000 0.00000000 0.02600000 0.08433333



# Pam

Medoids:

ID wm\_EL w\_EL wm\_ET w\_ET wm\_HT w\_HT wm\_HL w\_HL wm\_FS w\_FS wm\_LS w\_LS wm\_CL w\_CL

[1,] 6 0.073 0.073 0.052 0.052 0.066 0.066 0.043 0.043 0.138 0.138 0.141 0.141 0 0

[2,] 61 0.018 0.018 0.016 0.016 0.021 0.021 0.014 0.014 0.043 0.043 0.030 0.030 0 0

[3,] 47 0.018 0.798 0.016 0.016 0.021 0.021 0.014 0.014 0.043 0.043 0.030 0.030 0 0

wm\_OT w\_OT

[1,] 0.215 0.487

[2,] 0.026 0.858

[3,] 0.026 0.079

Clustering vector:

[1] 1 1 1 1 1 1 1 1 1 2 1 1 2 1 1 1 1 3 1 1 3 3 1 1 3 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 2 1 1

[45] 1 2 3 1 1 1 1 2 1 1 1 2 2 1 1 1 2 1 3 1 1 1 1 1

Objective function:

build swap

0.2624854 0.2624854



# Hclust





> cut3

[1] 1 1 1 1 2 1 1 2 1 1 1 1 1 1 1 1 1 3 1 1 3 3 1 1 3 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1

[45] 1 1 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 1 1 1 1 1

> aggregate(utilities,list(cut3),mean)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Group.1 | wm\_EL | w\_EL | wm\_ET | w\_ET | wm\_HT | w\_HT | wm\_HL |
| 1 | 0.08374576 | 0.07810169 | 0.0500678 | 0.05057627 | 0.07586441 | 0.08144068 | 0.05162712 |
| 2 | 0.09900000 | 0.09900000 | 0.3670000 | 0.06700000 | 0.10566667 | 0.10566667 | 0.87666667 |
| 3 | 0.01933333 | 0.79200000 | 0.0160000 | 0.01600000 | 0.05250000 | 0.02100000 | 0.01400000 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| w\_HL | wm\_FS | w\_FS | wm\_LS | w\_LS | wm\_CL | w\_CL | wm\_OT | w\_OT |
| 0.05450847 | 0.1375085 | 0.1253729 | 0.11016949 | 0.12238983 | 0.003271186 | 0.004915254 | 0.2254407 | 0.48086441 |
| 0.08766667 | 0.1063333 | 0.1063333 | 0.08733333 | 0.08733333 | 0.045333333 | 0.045333333 | 0.3070000 | 0.40200000 |
| 0.01400000 | 0.0430000 | 0.0430000 | 0.03000000 | 0.03000000 | 0.000000000 | 0.000000000 | 0.0260000 | 0.08433333 |



Hclust con correlation method converted in dinstances. (http://www.sthda.com/english/wiki/hierarchical-clustering-essentials-unsupervised-machine-learning)



Figura 4: Hclust with dinstance matrix from the correlation (spearman) – ranking alternatives

[1] 1 2 3 1 1 1 1 3 3 1 3 3 1 1 1 3 3 1 1 1 1 1 3 1 1 1 1 1 2 1 1 1 1 1 1 1 3 1 1 3 1

[42] 1 1 3 1 1 1 1 1 3 1 1 1 1 1 1 1 2 1 3 1 1 1 1 1 1 1 1

# Cluster validation statistics

For hclust con eclust function per 3 cluster

$n

[1] 68

$cluster.number

[1] 3

$cluster.size

[1] 59 3 6

$min.cluster.size

[1] 3

$noisen

[1] 0

$diameter

[1] 1.0328480 0.8997261 0.1917107

$average.distance

[1] 0.44689852 0.74618342 0.07665788

$median.distance

[1] 0.45378189 0.79130525 0.03748333

$separation

[1] 0.6437002 0.7754096 0.6437002

$average.toother

[1] 0.9699687 1.0748541 0.9496638

$separation.matrix

[,1] [,2] [,3]

[1,] 0.0000000 0.7754096 0.6437002

[2,] 0.7754096 0.0000000 1.2255362

[3,] 0.6437002 1.2255362 0.0000000

$ave.between.matrix

[,1] [,2] [,3]

[1,] 0.0000000 1.049082 0.9304121

[2,] 1.0490819 0.000000 1.3282811

[3,] 0.9304121 1.328281 0.0000000

$average.between

[1] 0.9817167

$average.within

[1] 0.4442058

$n.between

[1] 549

$n.within

[1] 1729

$max.diameter

[1] 1.032848

$min.separation

[1] 0.6437002

$within.cluster.ss

[1] 7.704312

$clus.avg.silwidths

1 2 3

0.5195406 0.2891313 0.9173379

$avg.silwidth

[1] 0.5444752

$g2

NULL

$g3

NULL

$pearsongamma

[1] 0.7562655

$dunn

[1] 0.6232284

$dunn2

[1] 1.246895

$entropy

[1] 0.4750797

$wb.ratio

[1] 0.4524786

$ch

[1] 27.07752

$cwidegap

[1] 0.6326065 0.7913053 0.1898078

$widestgap

[1] 0.7913053

$sindex

[1] 0.6599352

$corrected.rand

NULL

$vi

NULL

# Fase 2 – Rough set analysis

Analysis of the relation between the cluster classification and the socioeconomic factors. This method aims to define which are the factors linked and which are the factors that are not relevant for this first classification.

The superreduct computation based on RST :

This function is a wrapper for computing different types of decision superreducts (i.e. attribute subsets which do not lose any information regarding the decisions but are not require to be irreducable).

A feature subset consisting of 6 attributes:

Hotel, genere, età, paese, professione, travelling.with

Il titolo di studio è ininfluente per tutti gli indici: gini, entrophy, nOfconflicts

# Fase 3 – cluster con aspetti personali

Metodo qui: <https://www.r-bloggers.com/clustering-mixed-data-types-in-r/>

Analisi silhouette con metodo pam. N. 9 cluster ipotetici (vedi figura con max silhouette width) 

**Medoidi (vedi medoidi3 file))**

Another benefit of the PAM algorithm with respect to interpretation is that the medoids serve as exemplars of each cluster.

Hotel genere età paese professione travelling.with wm\_EL w\_EL wm\_ET

41 Catalunya f 44 eu dip. pubblico couple 0.072 0.072 0.062

3 Angendras m 38 eu impiegato family 0.117 0.117 0.037

56 Margherita m 31 eu impiegato couple 0.018 0.018 0.016

w\_ET wm\_HT w\_HT wm\_HL w\_HL wm\_FS w\_FS wm\_LS w\_LS wm\_CL w\_CL wm\_OT w\_OT

41 0.062 0.076 0.076 0.042 0.042 0.206 0.206 0.137 0.137 0 0 0.133 0.404

3 0.037 0.134 0.134 0.129 0.129 0.031 0.031 0.000 0.000 0 0 0.444 0.554

56 0.016 0.021 0.021 0.014 0.014 0.043 0.043 0.030 0.030 0 0 0.026 0.858

****

Medoids:

ID

[1,] 41 41

[2,] 3 3

[3,] 56 56

Clustering vector:

[1] 1 2 2 1 2 1 1 2 2 3 2 2 3 2 1 2 2 3 1 1 3 3 2 1 3 1 1 1 1 1 1 3 1 1 3 1 1 1 1 2 1 3 3

[44] 2 1 3 3 1 1 2 1 3 1 1 1 3 3 3 1 2 3 1 3 3 1 1 3 3

Objective function:

build swap

0.1830427 0.1801762

Available components:

[1] "medoids" "id.med" "clustering" "objective" "isolation" "clusinfo"

[7] "silinfo" "diss" "call"

**Con 8 cluster invece i medoidi sono**

Hotel genere età paese professione travelling.with wm\_EL w\_EL wm\_ET

54 Margherita f 50 eu dip. pubblico couple 0.080 0.080 0.070

9 Angendras m 55 eu dip. pubblico couple 0.117 0.117 0.037

6 Angendras m 60 italia impiegato couple 0.073 0.073 0.052

56 Margherita m 31 eu impiegato couple 0.018 0.018 0.016

18 Angendras m 68 eu pensionato family 0.018 0.804 0.016

26 B&B f 38 extraeu impiegato alone 0.069 0.069 0.059

48 Catalunya m 42 eu libero professionista couple 0.061 0.061 0.046

37 Catalunya f 55 italia dip. pubblico group 0.101 0.101 0.101

w\_ET wm\_HT w\_HT wm\_HL w\_HL wm\_FS w\_FS wm\_LS w\_LS wm\_CL w\_CL wm\_OT w\_OT

54 0.070 0.087 0.087 0.051 0.051 0.219 0.219 0.147 0.147 0.000 0.000 0.140 0.347

9 0.037 0.134 0.134 0.129 0.129 0.031 0.031 0.000 0.000 0.000 0.000 0.444 0.553

6 0.052 0.066 0.066 0.043 0.043 0.138 0.138 0.141 0.141 0.000 0.000 0.215 0.487

56 0.016 0.021 0.021 0.014 0.014 0.043 0.043 0.030 0.030 0.000 0.000 0.026 0.858

18 0.016 0.021 0.021 0.014 0.014 0.043 0.043 0.030 0.030 0.000 0.000 0.026 0.072

26 0.059 0.072 0.072 0.039 0.039 0.200 0.200 0.133 0.133 0.000 0.000 0.131 0.428

48 0.046 0.061 0.061 0.042 0.042 0.110 0.110 0.110 0.110 0.000 0.000 0.154 0.569

37 0.101 0.114 0.114 0.950 0.095 0.113 0.113 0.078 0.078 0.136 0.136 0.262 0.262

****

Medoids:

ID

[1,] 54 54

[2,] 9 9

[3,] 6 6

[4,] 56 56

[5,] 18 18

[6,] 26 26

[7,] 48 48

[8,] 37 37

Clustering vector:

[1] 1 2 2 3 3 3 1 2 2 4 2 2 4 3 3 2 2 5 1 3 5 5 6 6 5 6 6 1 6 6 1 4 3 7 7 7 8 3 1 3 1 4 7

[44] 2 1 4 5 7 7 2 1 4 1 1 1 4 4 7 1 7 4 3 4 3 1 6 7 7

Objective function:

build swap

0.1313776 0.1313776

## ORA FACCIO L’ANALISI CON HCLUST



Osservando il dendrogramma il metodo ward restituisce cluster meglio distinti.







Divido in cluster in 4 e faccio l’analisi interna separatamente per ognuno per capire quale suddivisione sia migliore con clusterstat method (fatto anche per il metodo completo)

Grp=

[1] 1 2 2 1 1 1 3 2 2 4 2 2 4 1 1 2 2 4 1 1 4 4 3 3 4 3 3 3 3 3 3 4 1 1 1 1 3 1 3 2 3 4 1

[44] 2 3 4 4 1 1 2 3 4 3 3 3 4 3 1 3 3 4 3 4 1 3 3 1 1

**Media dei dati(senza dati categorici)**

Group.1 Hotel genere età paese professione travelling.with wm\_EL w\_EL

1 1 NA NA 46.65000 NA NA NA 0.09940000 0.07240000

2 2 NA NA 44.36364 NA NA NA 0.12454545 0.12663636

3 3 NA NA 49.13043 NA NA NA 0.07552174 0.08352174

4 4 NA NA 49.64286 NA NA NA 0.01850000 0.34964286

wm\_ET w\_ET wm\_HT w\_HT wm\_HL w\_HL wm\_FS w\_FS

1 0.08905000 0.05170000 0.06150000 0.08405000 0.07140000 0.03540000 0.19335000 0.15690000

2 0.06690909 0.05572727 0.14727273 0.14054545 0.19500000 0.13400000 0.03545455 0.03572727

3 0.06139130 0.06139130 0.07726087 0.07578261 0.08652174 0.05086957 0.16721739 0.16765217

4 0.01585714 0.01585714 0.03435714 0.01985714 0.01407143 0.01507143 0.04192857 0.04192857

wm\_LS w\_LS wm\_CL w\_CL wm\_OT w\_OT

1 0.15170000 0.18850000 0.000000000 0.00000000 0.25635000 0.4110000

2 0.04445455 0.01545455 0.000000000 0.01154545 0.42127273 0.4670909

3 0.13056522 0.14378261 0.013000000 0.01300000 0.18473913 0.4056957

4 0.02971429 0.02971429 0.002142857 0.00000000 0.02628571 0.5281429

Computes a number of distance based statistics, which can be used for cluster validation, comparison between clusterings and decision about the number of clusters: cluster sizes, cluster diameters, average distances within and between clusters, cluster separation, biggest within cluster gap, average silhouette widths, the Calinski and Harabasz index, a Pearson version of Hubert’s gamma coefficient, the Dunn index and two indexes to assess the similarity of two clusterings, namely the corrected Rand index and Meila’s VI

$n

[1] 68

$cluster.number

[1] 4

$cluster.size

[1] 20 11 23 14

$min.cluster.size

[1] 11

$noisen

[1] 0

$diameter

[1] 0.4187507 0.3279487 0.4602439 0.3430013

$average.distance

[1] 0.2282049 0.2023359 0.2487106 0.2051567

$median.distance

[1] 0.2241055 0.1926326 0.2433466 0.2036885

$separation

[1] 0.09907151 0.15909596 0.09907151 0.14917228

$average.toother

[1] 0.2950421 0.3287210 0.3151403 0.3288537

$separation.matrix

[,1] [,2] [,3] [,4]

[1,] 0.00000000 0.1713946 0.09907151 0.1491723

[2,] 0.17139459 0.0000000 0.15909596 0.2019070

[3,] 0.09907151 0.1590960 0.00000000 0.1756220

[4,] 0.14917228 0.2019070 0.17562201 0.0000000

$ave.between.matrix

[,1] [,2] [,3] [,4]

[1,] 0.0000000 0.3003863 0.2796063 0.3162019

[2,] 0.3003863 0.0000000 0.3507857 0.3329500

[3,] 0.2796063 0.3507857 0.0000000 0.3378961

[4,] 0.3162019 0.3329500 0.3378961 0.0000000

$average.between

[1] 0.3150184

$average.within

[1] 0.2310364

$n.between

[1] 1689

$n.within

[1] 589

$max.diameter

[1] 0.4602439

$min.separation

[1] 0.09907151

$within.cluster.ss

[1] 1.813741

$clus.avg.silwidths

1 2 3 4

0.15067557 0.32347114 0.09027805 0.33983029

$avg.silwidth

[1] 0.1971428

$g2

NULL

$g3

NULL

$pearsongamma

[1] 0.4643643

$dunn

[1] 0.2152587

$dunn2

[1] 1.124223

$entropy

[1] 1.346645

$wb.ratio

[1] 0.7334062

$ch

[1] 15.0346

$cwidegap

[1] 0.2482360 0.2538270 0.2994880 0.1674932

$widestgap

[1] 0.299488

$sindex

[1] 0.1135788

$corrected.rand

NULL

$vi

NULL

## Sintesi caratteristiche socioeconomiche rispetto ai cluster

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **grp** | **Afcamera** | **Angendras** | **B&B** | **Catalunya** | **Margherita** |
| 1 | 2 | 8 | 0 | 8 | 2 |
| 2 | 0 | 8 | 0 | 2 | 1 |
| 3 | 0 | 1 | 8 | 4 | 10 |
| 4 | 0 | 5 | 1 | 4 | 4 |

|  |  |  |  |
| --- | --- | --- | --- |
| **grp** | **eu** | **extraeu** | **italia** |
| 1 | 13 | 1 | 6 |
| 2 | 8 | 0 | 3 |
| 3 | 14 | 2 | 7 |
| 4 | 13 | 0 | 1 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **grp** | **altro** | **casalinga** | **dip.pubblico** | **impiegato** | **insegnante** | **liberoProfessionista** |
| 1 | 0 | 0 | 0 | 9 | 0 | 6 |
| 2 | 0 | 0 | 1 | 5 | 1 | 3 |
| 3 | 1 | 0 | 12 | 6 | 0 | 3 |
| 4 | 0 | 3 | 2 | 4 | 0 | 3 |

|  |  |  |  |
| --- | --- | --- | --- |
| **grp** | **operaio** | **pensionato** | **studente** |
| 1 | 1 | 3 | 1 |
| 2 | 1 | 0 | 0 |
| 3 | 1 | 0 | 0 |
| 4 | 0 | 2 | 0 |

|  |  |  |
| --- | --- | --- |
| **grp** | **f** | **m** |
| 1 | 3 | 17 |
| 2 | 0 | 11 |
| 3 | 18 | 5 |
| 4 | 5 | 9 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **grp** | **alone** | **colleague** | **couple** | **family** | **friends** | **group** |
| 1 | 1 | 1 | 15 | 2 | 1 | 0 |
| 2 | 1 | 0 | 7 | 2 | 1 | 0 |
| 3 | 8 | 0 | 5 | 5 | 3 | 2 |
| 4 | 2 | 1 | 7 | 3 | 1 | 0 |

## Hclust Ward con 5 cluster

Questa suddivisione si rivela migliore internamente rispetto alla precedente in 4 sia all’average silhouette sia per l’analisi di alter component.

grp.5

[1] 1 2 2 1 1 1 3 2 2 4 2 2 4 1 1 2 2 4 1 1 4 4 5 5 4 5 5 3 5 5 5 4 1 1 1 1 3 1

[39] 3 2 3 4 1 2 3 4 4 1 1 2 5 4 5 3 3 4 5 1 3 3 4 5 4 1 3 5 1 1



Figura : Visualizzazione grafica per component principali.

# Multiple regression: personal characteristics and Cluster

Call:

lm(formula = dati$Cluster ~ dati$Hotel + dati$genere + dati$età +

dati$paese + dati$professione + dati$travelling.with)

Residuals:

Min 1Q Median 3Q Max

-1.8910 -0.6569 0.0000 0.5097 2.0741

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 2.967480 1.781964 1.665 0.1027

dati$HotelAngendras 0.736282 0.905460 0.813 0.4203

dati$HotelB&B 2.140571 1.148048 1.865 0.0686 .

dati$HotelCatalunya 0.578004 0.948061 0.610 0.5451

dati$HotelMargherita 1.613865 0.979830 1.647 0.1064

dati$generem -0.505911 0.380741 -1.329 0.1905

dati$età -0.002270 0.016715 -0.136 0.8926

dati$paeseextraeu -0.299471 0.809427 -0.370 0.7131

dati$paeseitalia -0.003621 0.371556 -0.010 0.9923

dati$professionecasalinga 1.598184 1.507380 1.060 0.2946

dati$professionedip. pubblico 0.604666 1.320560 0.458 0.6492

dati$professioneimpiegato -0.037314 1.331938 -0.028 0.9778

dati$professioneinsegnante -0.141920 1.897135 -0.075 0.9407

dati$professionelibero professionista -0.059738 1.337606 -0.045 0.9646

dati$professioneoperaio -0.568807 1.418172 -0.401 0.6902

dati$professionepensionato 0.208677 1.470257 0.142 0.8878

dati$professionestudente -1.241505 1.797416 -0.691 0.4932

dati$travelling.withcolleague -0.408505 1.013032 -0.403 0.6886

dati$travelling.withcouple -1.414582 0.561612 -2.519 0.0153 \*

dati$travelling.withfamily -0.661496 0.615007 -1.076 0.2877

dati$travelling.withfriends -0.942419 0.700718 -1.345 0.1852

dati$travelling.withgroup -1.017802 1.126828 -0.903 0.3711

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.146 on 46 degrees of freedom

Multiple R-squared: 0.5986, Adjusted R-squared: 0.4153

F-statistic: 3.266 on 21 and 46 DF, p-value: 0.0004016

Call:

lm(formula = dati1$Cluster ~ dati1$cost + dati1$environment +

dati1$weather + dati1$health + dati1$traditions + dati1$events +

dati1$sport + dati1$fun + dati1$study + dati1$work + dati1$relax +

dati1$fAf)

Residuals:

Min 1Q Median 3Q Max

-2.9352 -1.0607 0.0648 0.9167 3.3629

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.13137 1.07914 2.902 0.00533 \*\*

dati1$cost 0.06541 0.14694 0.445 0.65797

dati1$environment -0.01797 0.14299 -0.126 0.90042

dati1$weather -0.07446 0.18145 -0.410 0.68315

dati1$health 0.25751 0.21539 1.196 0.23701

dati1$traditions -0.10331 0.17884 -0.578 0.56586

dati1$events -0.43123 0.22580 -1.910 0.06138 .

dati1$sport 0.21789 0.18646 1.169 0.24761

dati1$fun -0.28640 0.25737 -1.113 0.27063

dati1$study -0.05753 0.28701 -0.200 0.84188

dati1$work 0.25006 0.21684 1.153 0.25381

dati1$relax 0.09272 0.15179 0.611 0.54383

dati1$fAf -0.10913 0.16759 -0.651 0.51765

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.492 on 55 degrees of freedom

Multiple R-squared: 0.1861, Adjusted R-squared: 0.008472

F-statistic: 1.048 on 12 and 55 DF, p-value: 0.4208