Clustering of Modal Valued Symbolic Data

Vladimir Batagelj, Nataša Kejžar, Simona Korenjak-Černe

University of Ljubljana, Slovenia

The 3rd Workshop in Symbolic Data Analysis
ICAI School of Engineering of Comillas Pontifical University
Madrid, 7-9. November 2012

Clustering of modal SOs

An SO X is described by a list $X = [\mathbf{x}_i]$ of descriptions of variables V_i. Each variable is described with frequency distribution (*bar chart*) of its values

$$\mathbf{f}_{xi} = [f_{xi1}, f_{xi2}, \dots, f_{xik_i}].$$

With $\mathbf{x}_i = [p_{xi1}, p_{xi2}, \dots, p_{xik_i}]$ we denote the corresponding probability distribution $\sum_{j=1}^{k_i} p_{xij} = 1, \quad i = 1, \dots, m$.

The *criterion function* has a form

$$P(\mathbf{C}) = \sum_{C \in \mathbf{C}} p(C) \qquad \text{where} \qquad p(C) = \min_T \sum_{X \in C} d(X, T)$$

$T = [\mathbf{t}_i]$, $\mathbf{t}_i = [t_{i1}, t_{i2}, \dots, t_{ik_i}]$ is a cluster's *representative* and has the same form as SOs. T_C that minimizes $P(C)$ is called a *leader*.

Dissimilarity between SOs

The dissimilarity measure between SOs has a form

$$d(X, T) = \sum_i \alpha_i d(\mathbf{x}_i, \mathbf{t}_i), \quad \alpha_i \geq 0, \quad \sum_i \alpha_i = 1,$$

where

$$d(\mathbf{x}_i, \mathbf{t}_i) = \sum_{j=1}^{k_i} w_{xij} \delta(p_{xij}, t_{ij}), \quad w_{xij} \geq 0.$$

The weight w_{xij} can be for the same unit X different for each variable V_i (needed in descriptions of ego-centric networks, population pyramids, etc.).

Dissimilarities δ

	$\delta(x,t)$	t^*_{ij}
1	$(p_x - t)^2$	$\frac{P_{ij}}{A_{ij}}$
2	$(\frac{p_x - t}{t})^2$	$\frac{Q_{ij}}{P_{ij}}$
3	$\frac{(p_x - t)^2}{t}$	$\sqrt{\frac{Q_{ij}}{A_{ij}}}$
4	$(\frac{p_x - t}{p_x})^2$	$\frac{H_{ij}}{F_{ij}}$
5	$\frac{(p_x - t)^2}{p_x}$	$\frac{A_{ij}}{H_{ij}}$
6	$\frac{(p_x - t)^2}{p_x t}$	$\sqrt{\frac{P_{ij}}{H_{ij}}}$

$$A_{ij} = \sum_{X \in C} w_{xij} \qquad P_{ij} = \sum_{X \in C} w_{xij} p_{xij} \qquad Q_{ij} = \sum_{X \in C} w_{xij} p_{xij}^2$$

$$H_{ij} = \sum_{X \in C} \frac{w_{xij}}{p_{xij}} \qquad F_{ij} = \sum_{X \in C} \frac{w_{xij}}{p_{xij}^2}$$

Algorithms for Clustering of modal SOs

For solving the clustering problem: Determine the clustering $\mathbf{C}^*$

$$P(\mathbf{C}^*) = \min_{\mathbf{C} \in \Phi_k} P(\mathbf{C})$$

we adapted:

- leaders (dynamic clouds) algorithm
- hierarchical agglomerative clustering algorithm

Both algorithms are solving the same clustering problem.

The leaders algorithm is used to cluster large sets of units to obtain a smaller set of leaders.

The leaders are further clustered using the agglomerative algorithm to decide about the right number of clusters and to reveal the relations among clusters.

Sheme of analysis

raw data

⇓

ENCODE

⇓

unified data

⇓

MAKE SOs

⇓

SOs - lists of distributions

⇓

⇓

`leaderSO`

⇓

clustering and cluster leaders

⇓

`hclustSO`

⇓

hierarchy and cluster leaders

⇓

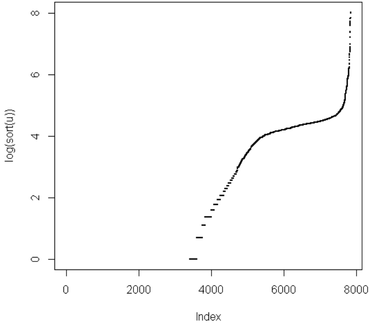
ANALYSIS

⇓

dendrogram, reports

Encoding variables: Cholesterol

```
> v <- food$Cholestrl
> u <- v[!is.na(v)]
> plot(log(sort(u)),pch=20,cex=0.5)
> (brks <- quantile(u[u>0], seq(0,1,1/9)))
       0% 11.11111% 22.22222% 33.33333% 44.44444% 55.55556%
        1         4        10        27        53        66
66.66667% 77.77778% 88.88889%      100%
       76        86       102      3100
> r <- findInterval(v,brks)
> r[r==10] <- 9
> (T <- c(as.vector(table(r)),length(r[is.na(r)])))
 [1] 3413  415  554  503  486  491  483  494  488  507  360
> a <- c("0",as.character(brks))
> names(T) <- c(paste("[",a[1:10],",",a[2:11],")",sep=""),"NA")
> T
     [0,1)      [1,4)     [4,10)    [10,27)    [27,53)    [53,66)
      3413        415        554        503        486        491
   [66,76)    [76,86)   [86,102) [102,3100)         NA
       483        494        488        507        360
```



Specificity of variable in cluster

In program Clamix we still don't have a good (final) answer to the question: which variables (and their values) are characteristic (specific) for a given cluster C ?

An approach is to define for a selected variable V its *specificity* $s(V, C)$ for a cluster C as

$$s(V, C) = 1/2 \int_{-\infty}^{\infty} |p_U(t) - p_C(t)| dt$$

or in discrete case

$$s(V, C) = 1/2 \sum_{v} |p_U(v) - p_C(v)|$$

where p_U is the distribution of values of V on set of units U; and p_C is the distribution of values of V on the cluster C.

Specificity of variable in cluster

The specificity $s(V, C)$ has the following properties:

- $0 \leq s(V, C) \leq 1$
- if $p_U = p_C$ then $s(V, c) = 0$; values of V on C are random sample from the values of V on U.
- if p_U and p_C are disjoint then $s(V, c) = 1$.

For identifying the most characteristic values v of variable V on C we compute the index

$$\frac{\max(p_U(v), p_C(v))}{\min(p_U(v), p_C(v))}$$

and select some values with (very) large value of this index.

Cars 1997

The raw data were obtained from Cars Catalog 1997 based on Katalog Avtomobilov '97 / Posebna priloga Dela in Slovenskih novic April '97 (by Janko Blagojevič). Transformation into symbolic objects (SOs) by Vladimir Batagelj, 29. July 2010.

```
> load("./cars2/cars.so")
> load("./cars2/cars.meta")
> length(SOs)
[1] 1349
> length(SOs[[1]])
[1] 26
> names(namedSO)
 [1] "price"      "type"       "NumDoors"   "NumPassen"  "motorsite"
 [6] "drive"      "length"     "width"      "height"     "wheelbase"
[11] "luggage"    "enlarLugg"  "fuelCapac"  "weight"     "maxLoad"
[16] "displace"   "maxPowKW"   "maxPowKM"   "rpm_maxPow" "maxTorque"
[21] "rpm_maxTor" "transmiss"  "breaks"     "minFuelCon" "accelTime"
[26] "maxSpeed"
```

Specificities in clustering of cars / part 1

```
1 L1
 NumPassen       type rpm_maxTor    height   displace minFuelCon     weight
 0.9510749  0.8784285  0.8724981 0.8472943  0.8465530  0.8421053  0.8376575
2 L2
     type NumPassen    height wheelbase    weight   maxLoad     width
0.9329496 0.9225715 0.8276864 0.7862469 0.7004026 0.6820593 0.5931772
3 L3
fuelCapac wheelbase     drive     width    length    weight   luggage
0.8223112 0.8030377 0.7758308 0.7418734 0.6767976 0.6753150 0.6427614
4 L4
maxTorque  maxPowKW  maxPowKM  displace    weight  maxSpeed     price
0.7388487 0.6975537 0.6939879 0.6008026 0.5518519 0.5518519 0.5136627
5 L5
 maxPowKW  maxPowKM maxTorque accelTime fuelCapac     price  maxSpeed
0.7548909 0.7541496 0.6530764 0.6436974 0.6194766 0.5819286 0.5597061
6 L6
rpm_maxTor rpm_maxPow     weight   displace minFuelCon  maxTorque      price
 0.8302446  0.7863762  0.6962830  0.6839458  0.6641957  0.6538176  0.6515938
7 L7
     type maxTorque    height  displace  maxSpeed  NumDoors  maxPowKW
0.7636739 0.6730912 0.6249364 0.5647466 0.5631186 0.5604151 0.5352113
8 L8
     type     drive    height  maxSpeed   maxLoad fuelCapac    weight
0.9058710 0.8732543 0.8472943 0.8176575 0.7369311 0.6093996 0.6093847
9 L9
  displace  maxTorque   maxPowKM   maxPowKW      price  accelTime minFuelCon
 0.6499158  0.6258036  0.6206146  0.6206146  0.5664196  0.4966575  0.4959162
10 L10
  maxSpeed   maxPowKW   maxPowKM  enlarLugg       type  maxTorque rpm_maxTor
 0.6473594  0.6369477  0.6109127  0.5828785  0.5797785  0.5769931  0.5430173
11 L11
 maxPowKW     price  displace  maxSpeed    weight    length wheelbase
0.8419041 0.8317272 0.8036959 0.7721593 0.7662290 0.7617812 0.7450704
12 L12
     type    length fuelCapac     drive   maxLoad wheelbase   luggage
0.8421053 0.8128016 0.7382339 0.7367513 0.6600505 0.6489766 0.5693106
```

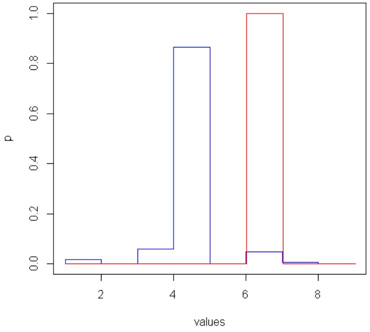
Specificities in clustering of cars / part 2

```
13 L13
 maxPowKM maxTorque  maxPowKW accelTime wheelbase  maxSpeed     width
0.6809837 0.6790215 0.6664196 0.6316749 0.6103257 0.6055640 0.5570226
14 L14
 NumDoors      type  maxPowKW  maxSpeed    height  maxPowKM     price
0.9111175 0.9014808 0.8561898 0.8435878 0.8354337 0.8157081 0.8073370
15 L15
 maxPowKM  maxPowKW maxTorque    weight     price  displace  maxSpeed
0.8078356 0.7796666 0.6671709 0.5999439 0.5945846 0.5803799 0.5201651
16 L16
  maxLoad  maxSpeed wheelbase     width fuelCapac maxTorque  maxPowKW
0.8398814 0.8376575 0.8369162 0.8361749 0.8257969 0.8228317 0.7983692
17 L17
enlarLugg  NumDoors      type    length     price  displace maxTorque
0.6586360 0.6022027 0.5984962 0.5925765 0.5471460 0.4960606 0.4959229
18 L18
 maxPowKW fuelCapac  maxPowKM    length     price     width    weight
0.7983692 0.7978249 0.7976279 0.7894478 0.7177137 0.6683428 0.6638743
19 L19
   length  NumDoors      type    weight enlarLugg wheelbase   luggage
0.7607460 0.6879170 0.6842105 0.6767547 0.6586360 0.6114137 0.6011431
20 L20
rpm_maxTor rpm_maxPow  fuelCapac   maxPowKW   maxPowKM minFuelCon     weight
 0.7847901  0.7766359  0.7529146  0.6956163  0.6901745  0.6822731  0.6553002
21 L21
   weight  maxSpeed  maxPowKW     price accelTime  maxPowKM    length
0.8097283 0.6901408 0.6530764 0.6370078 0.6283522 0.6241660 0.6095365
22 L22
      type  fuelCapac     height     weight      drive  maxTorque minFuelCon
 0.9258710  0.8695330  0.8472943  0.8376575  0.8242887  0.7812939  0.7731397
23 L23
fuelCapac    length   luggage wheelbase   maxLoad  NumDoors     width
0.8065508 0.7991379 0.7546605 0.7502402 0.7206161 0.6879170 0.6864344
24 L24
   length wheelbase fuelCapac     width      type   luggage    weight
0.8413640 0.7451674 0.7214461 0.7108970 0.6882591 0.6590067 0.6083709
25 L25
     type    height wheelbase     drive  NumDoors    length   luggage
0.8703155 0.8472943 0.8309859 0.7265876 0.7050490 0.6538176 0.6389918
```

Cluster 1: NumPassen

specificity = 0.9510749

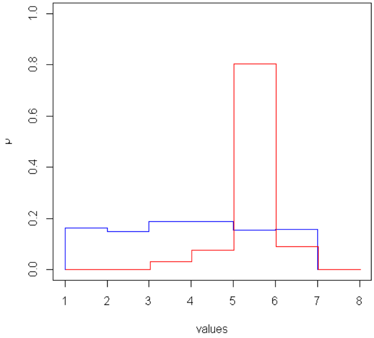
```
> specific(1,'NumPassen')
          2           3           4           5           6           7
0.018532246 0.000000000 0.059303188 0.864343958 0.001482580 0.048925130
          8          NA
0.007412898 0.000000000
 2  3  4  5  6  7  8 NA
 0  0  0  0  0  1  0  0
       2        3        4        5        6        7        8       NA
     Inf      NaN      Inf      Inf      Inf 20.43939      Inf      NaN
```



All cars in cluster 1 have the value NumPassen=7 .

Cluster 10: maxSpeed

```
specificity = 0.6473594
> specific(10,'maxSpeed')
[130,163] (163,174] (174,187] (187,200] (200,215] (215,400]        NA
0.1623425 0.1475167 0.1890289 0.1890289 0.1556709 0.1564122 0.0000000
 [130,163]  (163,174]  (174,187]  (187,200]  (200,215]  (215,400]         NA
0.00000000 0.00000000 0.03030303 0.07575758 0.80303030 0.09090909 0.00000000
[130,163] (163,174] (174,187] (187,200] (200,215] (215,400]        NA
      Inf       Inf  6.237954  2.495182  5.158514  1.720534       NaN
```



Most of the cars in cluster 10 have the maxSpeed in the interval (200,215].
No car in this cluster has maxSpeed in the interval [130,174].

Clustering of footballer careers

Dataset properties:

- all transfers/loans (all moves) in a career of a football player that **was recruited into the EPL** (in between the seasons 1992/93 and 2006/07)
- 3,749 players (with nationality, position) that moved between 2,301 clubs (with ranks)
- player success is regarded as **rank of a club** for which he plays (1 . . . best rank, 100 . . . worst rank)

Clustering of footballer careers

Dataset properties:

- all transfers/loans (all moves) in a career of a football player that **was recruited into the EPL** (in between the seasons 1992/93 and 2006/07)
- 3,749 players (with nationality, position) that moved between 2,301 clubs (with ranks)
- player success is regarded as **rank of a club** for which he plays (1 . . . best rank, 100 . . . worst rank)

- player has to be **observed long enough** to contribute his part to the most common career movements (not interested in injuries)
- careers for players from **19 to 30 years of age** (must turn 30 at least at the end of 2006/07)
- 1,287 players

Symbolic object for footballers

$$X_{player} = [x_1, x_2, \ldots, x_n]$$

x_i mean rank of a player in the i-th yearly interval

n number of yearly intervals (11)

Results

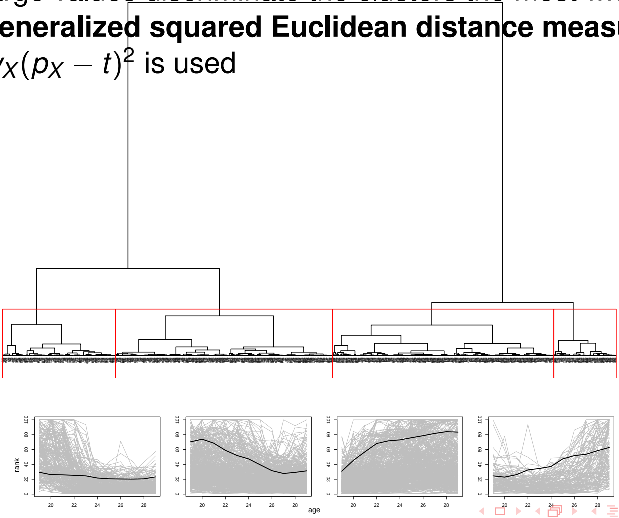
- adapted **hierarchical** clustering procedure

Results

- adapted **hierarchical** clustering procedure
- large values discriminate the clusters the most when **generalized squared Euclidean distance measure** $w_X(p_X - t)^2$ is used

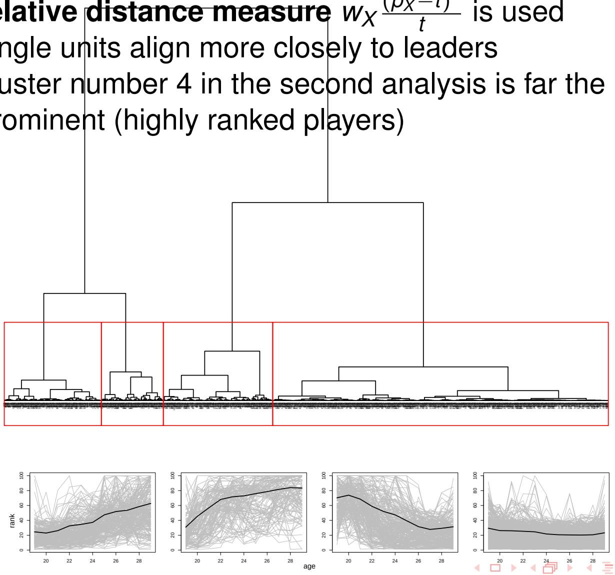
Results

- adapted **hierarchical** clustering procedure
- large values discriminate the clusters the most when **generalized squared Euclidean distance measure** $w_X(p_X - t)^2$ is used

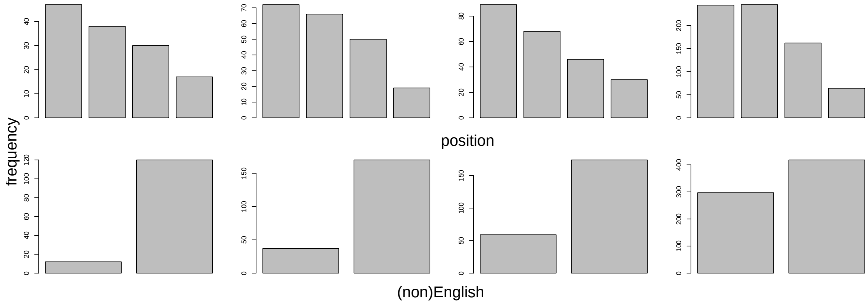


Results

- **relative distance measure** $w_X \frac{(p_X - t)^2}{t}$ is used
- single units align more closely to leaders
- cluster number 4 in the second analysis is far the most prominent (highly ranked players)



Supplementary variables



- positions: 1 . . . defender, 2 . . . midfielder, 3 . . . forward, 4 . . . goalkeeper
- nonEnglish vs. ENG, WAL, SCO, NIR or IRL

The European Social Survey data

- ESS (European Social Survey) is academically-driven social survey (est. 2001)
- biennial cross-sectional survey
- covers more than thirty nations, fifth round with over 50,000 respondents
- each round with about 300 questions (662 variables in round 5)
- studies attitudes, beliefs and behaviour patterns of EU populations

ESS — household data

Each respondent answers the following questions:

ESS — household data

Each respondent answers the following questions:

- categories of household residents
  {partner:1, offspring:2, parents:1, siblings:0, relatives:1, others:0}

ESS — household data

Each respondent answers the following questions:

- categories of household residents
  {partner:1, offspring:2, parents:1, siblings:0, relatives:1, others:0}
- gender
  {male:2, female:4}

ESS — household data

Each respondent answers the following questions:

- categories of household residents
  {partner:1, offspring:2, parents:1, siblings:0, relatives:1, others:0}
- gender
  {male:2, female:4}
- year of birth for every household resident
  {0–19 years:1, 20–34 years:1, 35–64 years:2, 65+ years:1}

ESS — household data

Each respondent answers the following questions:

- categories of household residents
  {partner:1, offspring:2, parents:1, siblings:0, relatives:1, others:0}
- gender
  {male:2, female:4}
- year of birth for every household resident
  {0–19 years:1, 20–34 years:1, 35–64 years:2, 65+ years:1}

We are interested in main household structures.

ESS — data as symbolic objects (SO)

- categories of household residents
  {respondent:1, partner:1, siblings:2, parents:1, siblings:0, relatives:1, others:0}
- gender
  {male:2, female:4}
- year of birth for every household resident
  {0–19 years:1, 20–34 years:1, 35–64 years:2, 65+ years:1, NA: 1}

ESS — data as symbolic objects (SO)

- categories of household residents
  {respondent:1, partner:1, siblings:2, parents:1, siblings:0, relatives:1, others:0}
- gender
  {male:2, female:4}
- year of birth for every household resident
  {0–19 years:1, 20–34 years:1, 35–64 years:2, 65+ years:1, NA: 1}
  641 respondents with missings

ESS — data as symbolic objects (SO)

- categories of household residents
  {respondent:1, partner:1, siblings:2, parents:1, siblings:0, relatives:1, others:0}
- gender
  {male:2, female:4}
- year of birth for every household resident
  {0–19 years:1, 20–34 years:1, 35–64 years:2, 65+ years:1, NA: 1}
  641 respondents with missings

Order of categories is not considered in the clustering process.

Clustering process

With large number of SOs (50,372 respondents to cluster)

1. cluster units with **non-hierarchical** method (to get smaller number of clusters and their leaders)
2. cluster clusters (i.e. leaders) with **hierarchical** method

Clustering process

With large number of SOs (50,372 respondents to cluster)

1. cluster units with **non-hierarchical** method (to get smaller number of clusters and their leaders)
2. cluster clusters (i.e. leaders) with **hierarchical** method

It is desired for methods to be harmonized (to base on the same criterion function), to solve the same optimization problem.

Clustering process

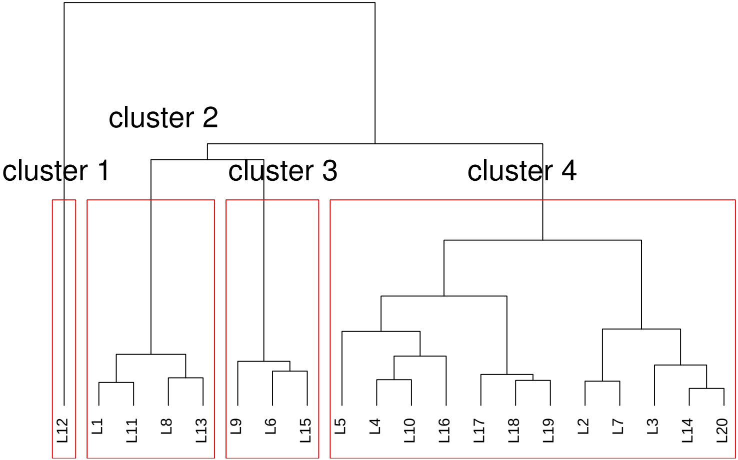
With large number of SOs (50,372 respondents to cluster)

1. cluster units with **non-hierarchical** method (to get smaller number of clusters and their leaders)
   20 clusters
2. cluster clusters (i.e. leaders) with **hierarchical** method
   4 final clusters

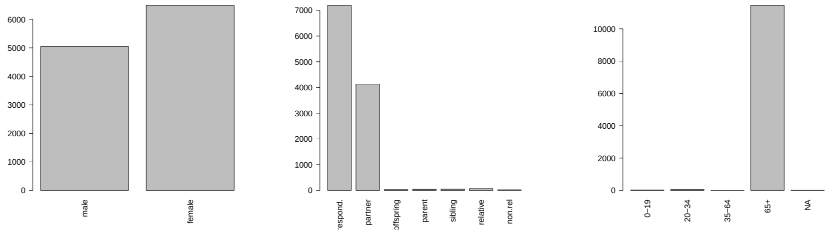
It is desired for methods to be harmonized (to base on the same criterion function), to solve the same optimization problem.

Custering with 5 age groups

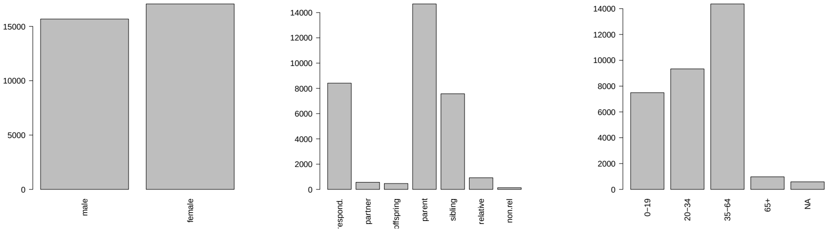
{0–19, 20–34, 35–64, 65+, NA}



Clusters (first two — small)

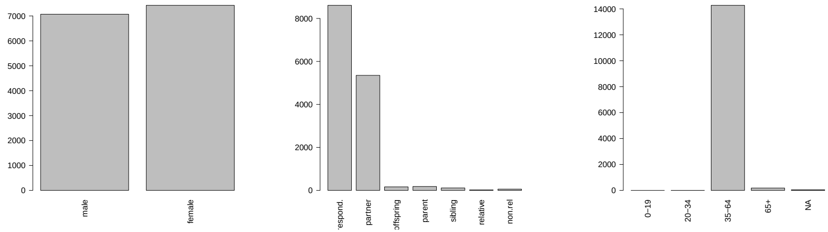


cluster 1

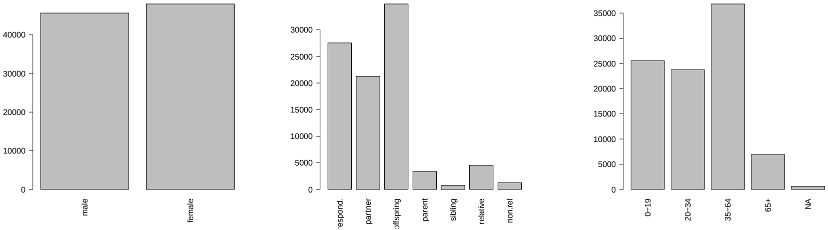


cluster 2

Clusters (second two — small and large)



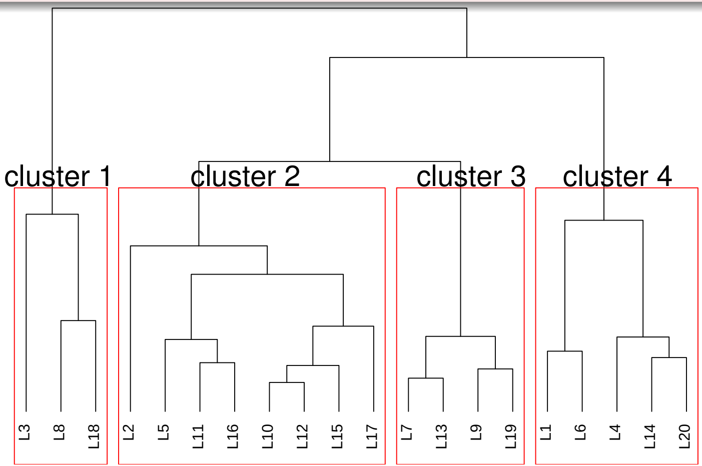
cluster 3



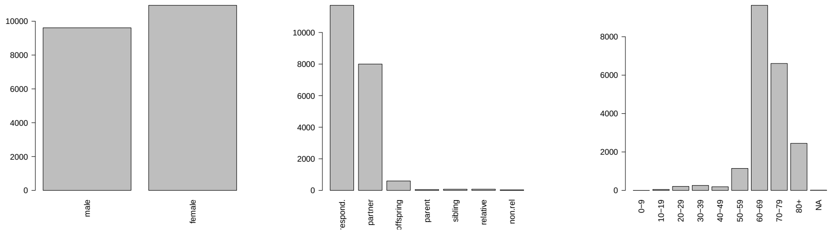
cluster 4

Custering with 10 age groups

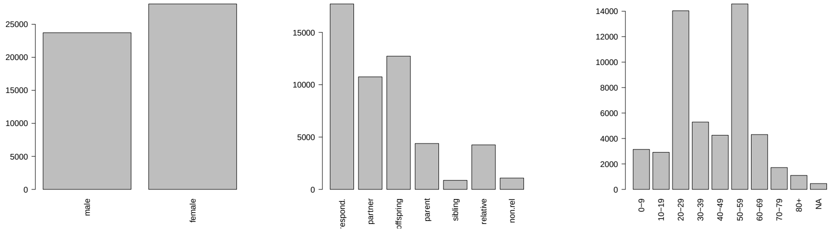
{0–9, 10–19, 20–29, 30–39, 40–49, 50–59, 60–69, 70–79, 80+, NA}



Clusters (first two)

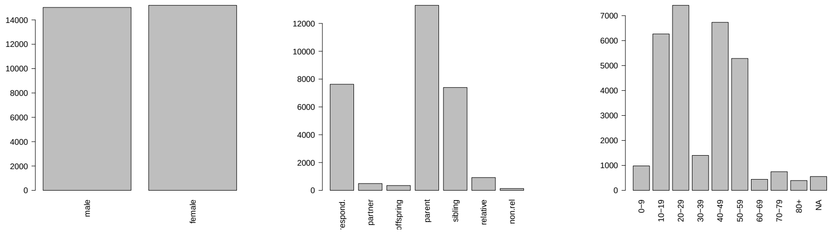


cluster 1

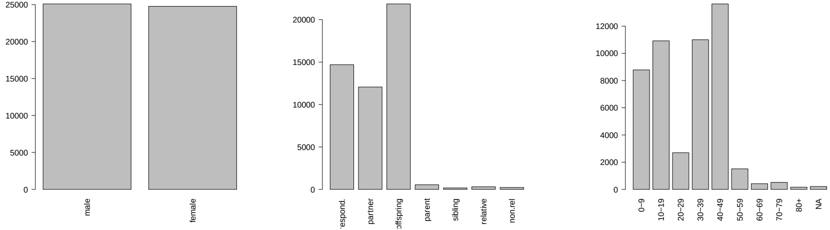


cluster 2

Clusters (second two)



cluster 3



cluster 4

Characteristics and considerations

- design and population weights included (results slightly differ if not)

Characteristics and considerations

- design and population weights included (results slightly differ if not)
- 10 runs of leaders algorithm used, results of best is presented, all 10 very similar

Characteristics and considerations

- design and population weights included (results slightly differ if not)
- 10 runs of leaders algorithm used, results of best is presented, all 10 very similar

- clustering with 5 age categories exhibits "chaining" in hierarchical algorithm (could be due to large span of category 3 (35-64))

Characteristics and considerations

- design and population weights included (results slightly differ if not)
- 10 runs of leaders algorithm used, results of best is presented, all 10 very similar

- clustering with 5 age categories exhibits "chaining" in hierarchical algorithm (could be due to large span of category 3 (35-64))
- more reasonable to use 10 age categories