BACHELOR DEGREE IN INFORMATICS (UPC).
COURSE 18-19 Q1 – LABORATORY TEST – QUIZ1-Re-SIT
Anàlisi de Dades i Explotació de la Informació (ADEI).
(Date: 5/11/2018 18:00-20:00 h   Place: Room A5S113)

Lecturer: Lídia Montero Mercadé
Office: Edifici C5 D207
Norms: Calculator, statistical tables and R Studio reference documents included in ATENEA are allowed. Internet access, emailing and chatting is strictly forbidden. Mobile phones should be switched off.
Quiz duration: 1h 30 min
Date for posting marks: Before 12/11/18, to be posted at Subject's ATENEA WEB page.
Open-office: 12/11/18 at 17:00 (C5-207).

Problem 1 Hobbies dataset: All questions account for 1 point

The data used refers to a questionnaire about hobbies where 8403 individuals were asked about their hobbies (18 questions). The characterization of the individuals was made considering the sex (male, female), age (15-25, 26-35, 36-45, 46-55, 56-65, 66-75, 76-85, 86- 100), civil status (single, married, widowed, divorced, remarried) and profession (worker, unqualified worker, technician, foreman, senior management, employees, others). And, finally, a quantitative variable indicates the number of hobbies practiced in the 18 possible options. The data matrix contains 8403 rows and 23 columns. The rows represent the individuals, the columns represent the different questions. The first 18 binary questions about different hobbies (practiced or not) and the following 4 factors are categorical variables of a socioeconomic nature. The last variable 23 is a quantitative variable with the total number of hobbies of each individual. A Multiple Correspondence Analysis taken into account characterization variables as supplementary is going to be analyzed.

1. How many axes do you need to retain for explaining 80% of the total inertial? Which is the total inertia?

You would need to account for axis 1 to 15. Total inertial is 1.166667

2. Which 3 factors have the greatest correlation with the two factorial axes?

Interpreting eta2 score for Categorical variable output: Cinema (0.39), Show (0.38), Exhibition (0.40), Travelling (0.35) and Gardening (0.45) have the highest correlation. Gardening with second axis and Exhibition, Cinema, Show and Travelling with the first axis.

3. Explain which activities are more frequent between the youngest population group.

On MCA factor map for supplementary categories the young group is clearly a rare class located on the right-down first plane area. After, MCA factor map for active categories has to be examined to realize that closest categories in first factorial plane location are Sports, Computer, Show and Cinema.

4. Are there any activities that commonly appear for those persons practicing fishing? Characterize the sociological profile of such a group.

Yes, MCA factor map for active categories has to be examined to realize that closest categories to actively fishing are knitting and gardening. According to MCA factor for supplementary categories corresponds to middle age people 55-65.

5. Which is the meaning of the first factorial axe?

It is a size axis indicating the number of simultaneous hobbies according to supplementary quantitative variable number of hobbies. According to MCA factor map age is also represented from none to many hobbies as age group decreases.

A Hierarchical Clustering is undertaken. A non-default criteria for selecting the number of clusters is chosen.

6. Interpret the default summary for the HCPC resulting object and determine the variables that characterize the partition. Indicate test performed for each variable.

Roughly as the cluster id increases the average number of simultaneous activities also increases: eta2 as a measure of correlation between Clustering classification and numeric variable number of activities is 0.75, thus very remarkable. A ChiSquared test for each factor (hobby) stated as H0: Factor is independent of Cluster Class and it is rejected for Travelleng, Reading, Listening.music, Cinema, Show, Exhibition, Computer, Sport, Walking, Collecting, Gardening and Cooking. For each cluster we should include the most representative simultaneous activities.

Approximated normal test for multinomial hypothesis involved in H0: Proportion of level factor i in cluster class j is equal to marginal proportion. For the different clusters:

Cluster 1- Contains old people without any hobby.

Cluster 2 – Contains a representation of married people, women in ages over 55 years practicing specific hobbies of knitting and gardening and none else.

Cluster 3 – Young group with Cinema, Listening.musing, Computers hobbies disliking Gardening and Collecting.

Cluster 4 – Contains people with Collecting, Playing.music, Listening.music and mechanic work.

Cluster 5- Contains people fond of Exhibition, Show, Cinema, Travelling, Sports and many other hobbies, single and active in labour market.

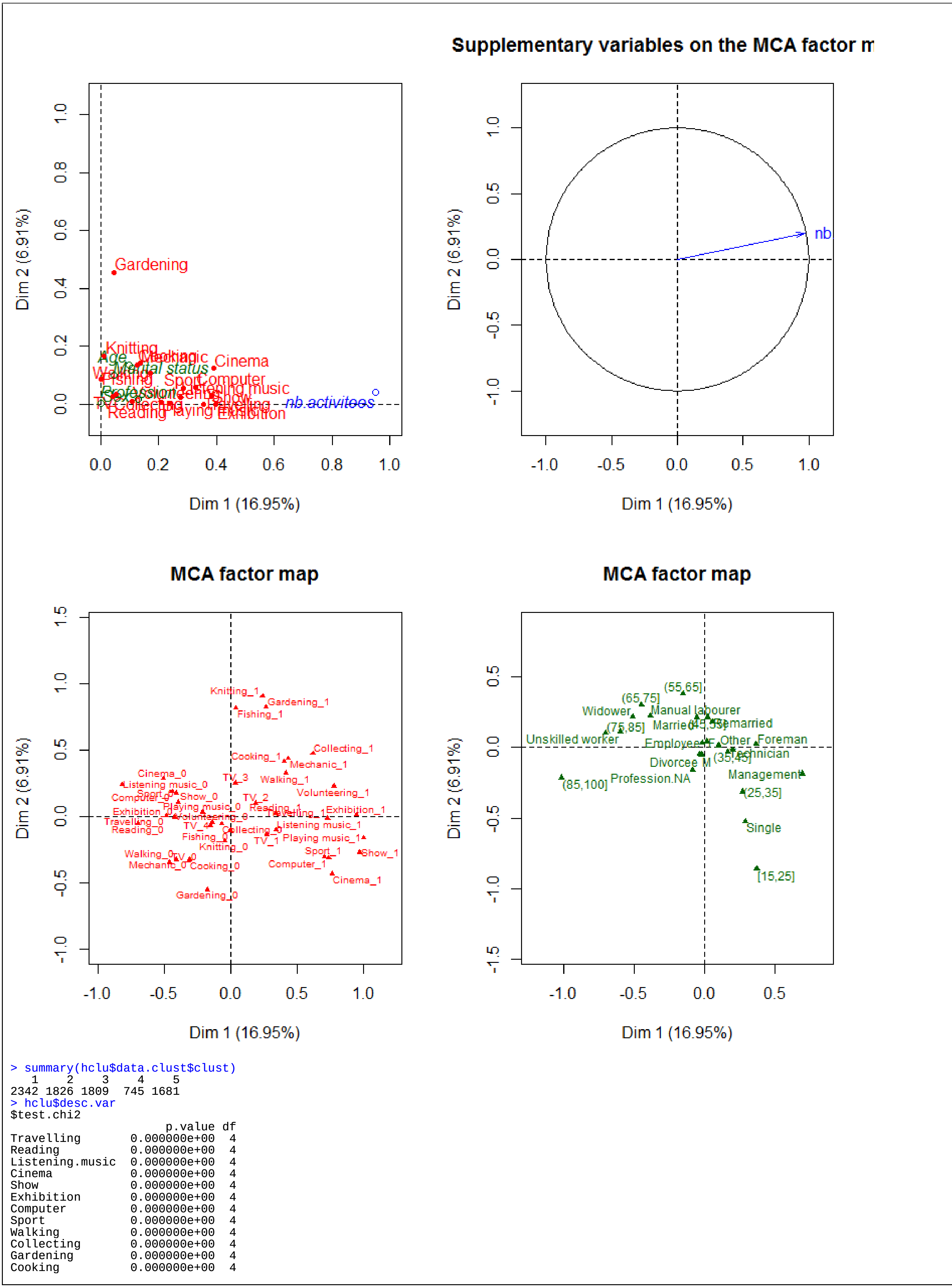
7. Which percentage of the inertia is explained by the selected clustering.

Less than 47%. Exactly, 42.65% since it should be calculated using

> sum(hclu$call$t$inert.gain[1:4])/hclu$call$t$within[1] or (hclu$call$t$within[1]-hclu$call$t$within[5])/hclu$call$t$within[1]

Output

```
> library(FactoMineR)
> library(car)
> data(hobbies)
> df<-hobbies[,c(9,1:8,10:23)]
> summary(hobbies)
 Reading  Listening music Cinema   Show     Exhibition Computer Sport    Walking  Travelling
 0:2757   0:2456          0:5044   0:5978   0:5808     0:5245   0:5308   0:4228   0:5040    
 1:5646   1:5947          1:3359   1:2425   1:2595     1:3158   1:3095   1:4175   1:3363    

 Playing music Collecting Volunteering Mechanic Gardening Knitting Cooking  Fishing  TV    
 0:6943        0:7541     0:7118       0:4864   0:5047    0:6990   0:4717   0:7458   0:1017
 1:1460        1: 862     1:1285       1:3539   1:3356    1:1413   1:3686   1: 945   1:1223

 Sex          Age          Marital status            Profession    nb.activitees   
 M:3787   (45,55]:1837   Single    :2140   Employee          :2552   Min.   : 0.000  
 F:4616   (35,45]:1646   Married   :4333   Manual labourer :1161   1st Qu.: 4.000  
          (25,35]:1302   Widower   : 734   Management        :1052   Median : 7.000  
          (55,65]:1257   Divorcee  : 792   Unskilled worker: 792   Mean   : 6.866  
          (65,75]: 937   Remarried: 404   Foreman           : 735   3rd Qu.: 9.000  
          [15,25]: 857                     (Other)           : 613   Max.   :16.000  
          (Other): 567                     NA's              :1498                   

> res.mca <- MCA(df,quali.sup=c(19:22),quanti.sup=23)
> summary(res.mca,nb.dec=2,ncp=2,nbind = 0,nbelements=Inf)

Call: MCA(X = df, quanti.sup = 23, quali.sup = c(19:22))

Eigenvalues
                      Dim.1  Dim.2  Dim.3  Dim.4  Dim.5  Dim.6  Dim.7  Dim.8  Dim.9 Dim.10 Dim.11
Variance               0.20   0.08   0.07   0.06   0.06   0.06   0.06   0.05   0.05   0.05   0.05
% of var.             16.95   6.91   6.17   5.39   5.01   4.78   4.76   4.57   4.55   4.21   3.99
Cumulative % of var.  16.95  23.86  30.03  35.42  40.43  45.22  49.98  54.55  59.09  63.30  67.29
                     Dim.12 Dim.13 Dim.14 Dim.15 Dim.16 Dim.17 Dim.18 Dim.19 Dim.20 Dim.21
Variance               0.05   0.04   0.04   0.04   0.04   0.04   0.04   0.03   0.03   0.03
% of var.              3.86   3.73   3.72   3.50   3.26   3.20   3.10   3.00   2.77   2.58
Cumulative % of var.  71.15  74.88  78.60  82.09  85.35  88.55  91.66  94.65  97.42 100.00

Categories
                     Dim.1    ctr   cos2 v.test    Dim.2    ctr   cos2 v.test  
Travelling_0      |  -0.49   3.99   0.35 -54.61 |   0.01   0.00   0.00   0.92 |
Travelling_1      |   0.73   5.98   0.35  54.61 |  -0.01   0.00   0.00  -0.92 |
Reading_0         |  -0.70   4.50   0.24 -44.77 |  -0.05   0.06   0.00  -3.25 |
Reading_1         |   0.34   2.20   0.24  44.77 |   0.02   0.03   0.00   3.25 |
Listening music_0 |  -0.82   5.48   0.28 -48.11 |   0.24   1.17   0.02  14.20 |
Listening music_1 |   0.34   2.26   0.28  48.11 |  -0.10   0.48   0.02 -14.20 |
Cinema_0          |  -0.51   4.37   0.39 -57.17 |   0.29   3.40   0.12  32.20 |
Cinema_1          |   0.76   6.56   0.39  57.17 |  -0.43   5.10   0.12 -32.20 |
Show_0            |  -0.39   3.11   0.38 -56.75 |   0.11   0.59   0.03  15.73 |
Show_1            |   0.97   7.66   0.38  56.75 |  -0.27   1.44   0.03 -15.73 |
Exhibition_0      |  -0.42   3.46   0.40 -57.88 |  -0.01   0.00   0.00  -0.75 |
Exhibition_1      |   0.94   7.75   0.40  57.88 |   0.01   0.00   0.00   0.75 |
Computer_0        |  -0.44   3.46   0.33 -52.45 |   0.19   1.51   0.06  22.11 |
Computer_1        |   0.74   5.74   0.33  52.45 |  -0.31   2.50   0.06 -22.11 |
Sport_0           |  -0.41   2.97   0.29 -49.09 |   0.18   1.36   0.05  21.19 |
Sport_1           |   0.70   5.09   0.29  49.09 |  -0.30   2.33   0.05 -21.19 |
Walking_0         |  -0.41   2.40   0.17 -38.03 |  -0.32   3.65   0.11 -29.95 |
Walking_1         |   0.42   2.43   0.17  38.03 |   0.33   3.70   0.11  29.95 |
Playing music_0   |  -0.21   1.02   0.21 -41.93 |   0.03   0.06   0.01   6.75 |
Playing music_1   |   1.00   4.86   0.21  41.93 |  -0.16   0.31   0.01  -6.75 |
Collecting_0      |  -0.07   0.13   0.04 -19.13 |  -0.05   0.18   0.03 -14.76 |
Collecting_1      |   0.62   1.10   0.04  19.13 |   0.48   1.60   0.03  14.76 |
Volunteering_0    |  -0.14   0.47   0.11 -30.23 |  -0.04   0.10   0.01  -8.90 |
Volunteering_1    |   0.78   2.59   0.11  30.23 |   0.23   0.55   0.01   8.90 |
Mechanic_0        |  -0.31   1.60   0.13 -33.67 |  -0.32   4.07   0.14 -34.35 |
Mechanic_1        |   0.43   2.19   0.13  33.67 |   0.44   5.60   0.14  34.35 |
Gardening_0       |  -0.18   0.53   0.05 -19.86 |  -0.55  12.46   0.45 -61.70 |
Gardening_1       |   0.27   0.79   0.05  19.86 |   0.83  18.74   0.45  61.70 |
Knitting_0        |  -0.05   0.05   0.01  -9.80 |  -0.18   1.92   0.17 -37.37 |
Knitting_1        |   0.24   0.27   0.01   9.80 |   0.91   9.52   0.17  37.37 |
Cooking_0         |  -0.31   1.55   0.13 -32.46 |  -0.33   4.09   0.14 -33.73 |
Cooking_1         |   0.40   1.98   0.13  32.46 |   0.42   5.24   0.14  33.73 |
Fishing_0         |   0.00   0.00   0.00  -1.22 |  -0.10   0.66   0.08 -26.68 |
Fishing_1         |   0.04   0.00   0.00   1.22 |   0.82   5.18   0.08  26.68 |
TV_0              |  -0.46   0.72   0.03 -15.63 |  -0.35   1.00   0.02 -11.78 |
TV_1              |   0.27   0.31   0.01  10.34 |  -0.13   0.18   0.00  -5.08 |
TV_2              |   0.19   0.26   0.01  10.22 |   0.10   0.18   0.00   5.41 |
TV_3              |   0.03   0.01   0.00   1.59 |   0.25   0.93   0.02  12.00 |
TV_4              |  -0.15   0.17   0.01  -8.29 |  -0.07   0.08   0.00  -3.68 |

Categorical variables (eta2)
                    Dim.1 Dim.2
Travelling        |  0.35  0.00 |
Reading           |  0.24  0.00 |
Listening music   |  0.28  0.02 |
Cinema            |  0.39  0.12 |
Show              |  0.38  0.03 |
Exhibition        |  0.40  0.00 |
Computer          |  0.33  0.06 |
Sport             |  0.29  0.05 |
Walking           |  0.17  0.11 |
Playing music     |  0.21  0.01 |
Collecting        |  0.04  0.03 |
Volunteering      |  0.11  0.01 |
Mechanic          |  0.13  0.14 |
Gardening         |  0.05  0.45 |
Knitting          |  0.01  0.17 |
Cooking           |  0.13  0.14 |
Fishing           |  0.00  0.08 |
TV                |  0.05  0.03 |

Supplementary categories
                     Dim.1   cos2 v.test    Dim.2   cos2 v.test
F                 |   0.02   0.00   1.78 |   0.04   0.00   4.25 |
M                 |  -0.02   0.00  -1.78 |  -0.05   0.00  -4.25 |
(25,35]           |   0.27   0.01  10.49 |  -0.31   0.02 -12.36 |
(35,45]           |   0.20   0.01   9.09 |  -0.02   0.00  -0.92 |
(45,55]           |   0.02   0.00   1.06 |   0.21   0.01  10.31 |
(55,65]           |  -0.15   0.00  -5.88 |   0.38   0.03  14.60 |
(65,75]           |  -0.45   0.03 -14.53 |   0.30   0.01   9.79 |
(75,85]           |  -0.70   0.03 -15.86 |   0.10   0.00   2.28 |
(85,100]          |  -1.01   0.01  -9.40 |  -0.21   0.00  -1.99 |
[15,25]           |   0.37   0.02  11.42 |  -0.86   0.08 -26.58 |
Divorcee          |  -0.03   0.00  -1.00 |  -0.05   0.00  -1.44 |
Married           |  -0.06   0.00  -5.20 |   0.21   0.05  20.19 |
Remarried         |   0.05   0.00   1.10 |   0.18   0.00   3.71 |
Single            |   0.29   0.03  15.46 |  -0.52   0.09 -28.00 |
Widower           |  -0.51   0.02 -14.44 |   0.22   0.00   6.14 |
Employee          |  -0.02   0.00  -1.02 |   0.03   0.00   1.91 |
Foreman           |   0.37   0.01  10.38 |   0.02   0.00   0.61 |
Management        |   0.69   0.07  24.02 |  -0.19   0.00  -6.47 |
Manual labourer   |  -0.38   0.02 -14.04 |   0.22   0.01   8.19 |
Other             |   0.10   0.00   1.48 |   0.01   0.00   0.19 |
Profession.NA     |  -0.08   0.00  -3.61 |  -0.16   0.01  -6.82 |
Technician        |   0.17   0.00   3.39 |  -0.03   0.00  -0.63 |
Unskilled worker  |  -0.60   0.04 -17.61 |   0.11   0.00   3.35 |

Supplementary categorical variables (eta2)
                    Dim.1 Dim.2
Sex               |  0.00  0.00 |
Age               |  0.10  0.13 |
Marital status    |  0.05  0.10 |
Profession        |  0.13  0.02 |

Supplementary continuous variable
                    Dim.1   Dim.2
nb.activitees     |  0.98 |  0.20 |

> hclu<-HCPC(res.mca,5,order=T)
```



```
> summary(hclu$data.clust$clust)
   1    2    3    4    5
2342 1826 1809  745 1681
> hclu$desc.var
$test.chi2
                    p.value df
Travelling      0.000000e+00  4
Reading         0.000000e+00  4
Listening.music 0.000000e+00  4
Cinema          0.000000e+00  4
Show            0.000000e+00  4
Exhibition      0.000000e+00  4
Computer        0.000000e+00  4
Sport           0.000000e+00  4
Walking         0.000000e+00  4
Collecting      0.000000e+00  4
Gardening       0.000000e+00  4
Cooking         0.000000e+00  4
```

```
Age             6.421570e-309 28
Playing.music   6.766882e-299  4
Knitting        4.533775e-283  4
Mechanic        5.865693e-223  4
Profession      2.951388e-201 28
TV              2.716815e-200 16
Volunteering    5.469591e-191  4
Marital.status  5.916806e-163 16
Sex              4.649210e-41  4
Fishing          1.001452e-17  4

$category
$category$`1`
                                    Cla/Mod   Mod/Cla    Global       p.value     v.test
Cinema=Cinema_0                   42.724029 92.015371 60.026181  0.000000e+00        Inf
Listening.music=Listening music_0 60.749186 63.706234 29.227657  0.000000e+00        Inf
Reading=Reading_0                 55.785274 65.670367 32.809711  0.000000e+00        Inf
Computer=Computer_0               41.029552 91.887276 62.418184 1.316184e-307  37.490487
Exhibition=Exhibition_0           38.515840 95.516652 69.118172 2.690667e-289  36.349349
Cooking=Cooking_0                 42.548230 85.695986 56.134714 6.414214e-277  35.557273
Travelling=Travelling_0           41.071429 88.385995 59.978579 3.329326e-269  35.054504
Show=Show_0                       37.437270 95.559351 71.141259 2.546257e-258  34.332771
Walking=Walking_0                 42.147588 76.088813 50.315364 6.525446e-197  29.936852
Sport=Sport_0                     38.376036 86.976943 63.167916 6.265111e-194  29.706841
TV=TV_0                           59.980334 26.046114 12.102820 3.683673e-117  23.010212
Mechanic=Mechanic_0               36.883224 76.601196 57.884089 2.096690e-108  22.118526
Gardening=Gardening_0             36.417674 78.479932 60.061883 1.384870e-107  22.033184
Playing.music=Playing music_0     32.305920 95.772844 82.625253 4.206511e-107  21.982805
Knitting=Knitting_0               31.959943 95.388557 83.184577  1.203445e-93  20.528286
Collecting=Collecting_0           30.407108 97.907771 89.741759  7.818547e-68  17.403079
Volunteering=Volunteering_0       30.429896 92.485056 84.707842  6.453900e-39  13.048820
Age=(75,85]                       52.697095 10.845431  5.736047  2.526769e-32  11.836535
Profession=Manual labourer        39.793282 19.726729 13.816494  2.661782e-21   9.475228
…
TV=TV_2                           20.686456 19.043553 25.657503  1.329668e-18  -8.803202
TV=TV_3                           19.267606 14.602904 21.123408  8.331736e-21  -9.355358
Marital.status=Single             19.485981 17.805295 25.467095  9.876458e-25 -10.267466
Age=[15,25]                       12.718786  4.654142 10.198739  5.607008e-29 -11.171742
Profession=Management             12.357414  5.550811 12.519338  8.812159e-38 -12.848124
Volunteering=Volunteering_1       13.696498  7.514944 15.292158  6.453900e-39 -13.048820
Collecting=Collecting_1            5.684455  2.092229 10.258241  7.818547e-68 -17.403079
Knitting=Knitting_1                7.643312  4.611443 16.815423  1.203445e-93 -20.528286
Playing.music=Playing music_1      6.780822  4.227156 17.374747 4.206511e-107 -21.982805
Gardening=Gardening_1             15.017878 21.520068 39.938117 1.384870e-107 -22.033184
Mechanic=Mechanic_1               15.484600 23.398804 42.115911 2.096690e-108 -22.118526
Sport=Sport_1                      9.854604 13.023057 36.832084 6.265111e-194 -29.706841
Walking=Walking_1                 13.413174 23.911187 49.684636 6.525446e-197 -29.936852
Show=Show_1                        4.288660  4.440649 28.858741 2.546257e-258 -34.332771
Travelling=Travelling_1            8.088017 11.614005 40.021421 3.329326e-269 -35.054504
Cooking=Cooking_1                  9.088443 14.304014 43.865286 6.414214e-277 -35.557273
Exhibition=Exhibition_1            4.046243  4.483348 30.881828 2.690667e-289 -36.349349
Computer=Computer_1                6.016466  8.112724 37.581816 1.316184e-307 -37.490487
Cinema=Cinema_1                    5.567133  7.984629 39.973819  0.000000e+00       -Inf
Listening.music=Listening music_1 14.292921 36.293766 70.772343  0.000000e+00       -Inf
Reading=Reading_1                 14.240170 34.329633 67.190289  0.000000e+00       -Inf

$category$`2`
                                    Cla/Mod   Mod/Cla    Global       p.value     v.test
Knitting=Knitting_1               53.786270 41.621030 16.815423 1.377952e-192  29.602734
Gardening=Gardening_1             38.051251 69.934283 39.938117 2.199103e-190  29.431076
Cinema=Cinema_0                   31.245044 86.308872 60.026181 1.707922e-165  27.417256
Computer=Computer_0               30.104862 86.473165 62.418184 1.286536e-142  25.426443
Cooking=Cooking_1                 34.102008 68.838992 43.865286 5.439270e-131  24.352614
Sport=Sport_0                     28.730219 83.515882 63.167916 4.069312e-101  21.348006
Show=Show_0                       26.513884 86.801752 71.141259  7.671659e-70  17.665945
Collecting=Collecting_0           23.896035 98.685652 89.741759  1.303331e-63  16.837165
Playing.music=Playing music_0     24.585914 93.483023 82.625253  2.153615e-51  15.081187
Walking=Walking_1                 27.832335 63.636364 49.684636  9.244088e-42  13.538675
Sex=F                             27.058059 68.400876 54.932762  6.314048e-40  13.224758
Exhibition=Exhibition_0           25.103306 79.846659 69.118172  7.667694e-31  11.546737
Marital.status=Married            26.217401 62.212486 51.564917  4.875761e-25  10.335360
Age=(55,65]                       32.537788 22.398686 14.958943  3.674170e-22   9.679805
Age=(65,75]                       33.404482 17.141292 11.150779  1.587237e-18   8.783317
…
Age=(25,35]                       12.749616  9.090909 15.494466  2.691558e-19  -8.980657
Profession=Management             11.026616  6.352683 12.519338  7.798216e-22  -9.602561
TV=TV_0                            9.341200  5.202629 12.102820  1.781694e-28 -11.068596
Exhibition=Exhibition_1           14.181118 20.153341 30.881828  7.667694e-31 -11.546737
Sex=M                             15.236335 31.599124 45.067238  6.314048e-40 -13.224758
Walking=Walking_0                 15.704825 36.363636 50.315364  9.244088e-42 -13.538675
Playing.music=Playing music_1      8.150685  6.516977 17.374747  2.153615e-51 -15.081187
Age=[15,25]                        3.383897  1.588171 10.198739  3.870482e-58 -16.074185
Collecting=Collecting_1            2.784223  1.314348 10.258241  1.303331e-63 -16.837165
Show=Show_1                        9.938144 13.198248 28.858741  7.671659e-70 -17.665945
Marital.status=Single              8.738318 10.240964 25.467095  2.682805e-73 -18.109478
Sport=Sport_1                      9.725363 16.484118 36.832084 4.069312e-101 -21.348006
Cooking=Cooking_0                 12.062752 31.161008 56.134714 5.439270e-131 -24.352614
Computer=Computer_1                7.821406 13.526835 37.581816 1.286536e-142 -25.426443
Cinema=Cinema_1                    7.442691 13.691128 39.973819 1.707922e-165 -27.417256
Gardening=Gardening_0             10.877749 30.065717 60.061883 2.199103e-190 -29.431076
Knitting=Knitting_0               15.250358 58.378970 83.184577 1.377952e-192 -29.602734

$category$`3`
                                    Cla/Mod    Mod/Cla    Global       p.value     v.test
Gardening=Gardening_0             31.008520 86.5118850 60.061883 5.379101e-166  27.459308
Cinema=Cinema_1                   36.528729 67.8275290 39.973819 4.920732e-162  27.125549
Listening.music=Listening music_1 27.694636 91.0447761 70.772343 1.154070e-119  23.258977
Computer=Computer_1               34.420519 60.0884467 37.581816 2.739813e-107  22.002260
Age=[15,25]                       51.575263 24.4333886 10.198739  1.555519e-94  20.627479
Knitting=Knitting_0               25.064378 96.8490879 83.184577  9.264004e-90  20.088709
Marital.status=Single             36.261682 42.8966280 25.467095  2.586557e-76  18.487852
Collecting=Collecting_0           23.763427 99.0602543 89.741759  2.060733e-70  17.739962
Sport=Sport_1                     31.276252 53.5102266 36.832084  3.454262e-60  16.364041
Walking=Walking_0                 27.199622 63.5710337 50.315364  1.842592e-37  12.790929
Volunteering=Volunteering_0       23.742624 93.4217800 84.707842  3.643955e-36  12.556886
Mechanic=Mechanic_0               25.534539 68.6567164 57.884089  3.028251e-26  10.598492
Age=(25,35]                       31.874040 22.9408513 15.494466  1.998704e-21   9.505092
…
```

```
Cooking=Cooking_1                 16.793272 34.2177999 43.865286  4.912044e-21  -9.411050
Age=(65,75]                        9.498399  4.9198452 11.150779  8.431923e-25 -10.282714
Marital.status=Married            17.009001 40.7407407 51.564917  2.206989e-25 -10.411082
Mechanic=Mechanic_1               16.021475 31.3432836 42.115911  3.028251e-26 -10.598492
Volunteering=Volunteering_1        9.260700  6.5782200 15.292158  3.643955e-36 -12.556886
Walking=Walking_1                 15.784431 36.4289663 49.684636  1.842592e-37 -12.790929
Sport=Sport_0                     15.844009 46.4897734 63.167916  3.454262e-60 -16.364041
Collecting=Collecting_1            1.972158  0.9397457 10.258241  2.060733e-70 -17.739962
Knitting=Knitting_1                4.033970  3.1509121 16.815423  9.264004e-90 -20.088709
Computer=Computer_0               13.765491 39.9115533 62.418184 2.739813e-107 -22.002260
Listening.music=Listening music_0  6.596091  8.9552239 29.227657 1.154070e-119 -23.258977
Cinema=Cinema_0                   11.538462 32.1724710 60.026181 4.920732e-162 -27.125549
Gardening=Gardening_1              7.270560 13.4881150 39.938117 5.379101e-166 -27.459308

$category$`4`
                                    Cla/Mod   Mod/Cla    Global      p.value     v.test
Collecting=Collecting_1           73.897912 85.503356 10.258241 0.000000e+00        Inf
Playing.music=Playing music_1     18.835616 36.912752 17.374747 3.650584e-41  13.437396
Mechanic=Mechanic_1               13.478384 64.026846 42.115911 2.097106e-36  12.600537
Volunteering=Volunteering_1       16.108949 27.785235 15.292158 3.955222e-20   9.189268
Listening.music=Listening music_1 10.576761 84.429530 70.772343 1.321124e-19   9.058619
Exhibition=Exhibition_1           13.179191 45.906040 30.881828 1.957675e-19   9.015616
Computer=Computer_1               12.349588 52.348993 37.581816 9.901284e-18   8.575086
Fishing=Fishing_1                 15.555556 19.731544 11.245984 1.213513e-12   7.103824
Reading=Reading_1                 10.201913 77.315436 67.190289 2.310993e-10   6.339108
…

Fishing=Fishing_0                  8.018235 80.268456 88.754016 1.213513e-12  -7.103824
Computer=Computer_0                6.768351 47.651007 62.418184 9.901284e-18  -8.575086
Exhibition=Exhibition_0            6.938705 54.093960 69.118172 1.957675e-19  -9.015616
Listening.music=Listening music_0  4.723127 15.570470 29.227657 1.321124e-19  -9.058619
Volunteering=Volunteering_0        7.558303 72.214765 84.707842 3.955222e-20  -9.189268
TV=TV_1                            2.044154  3.355705 14.554326 8.760305e-26 -10.498673
Mechanic=Mechanic_0                5.509868 35.973154 57.884089 2.097106e-36 -12.600537
Playing.music=Playing music_0      6.769408 63.087248 82.625253 3.650584e-41 -13.437396
Collecting=Collecting_0            1.432171 14.496644 89.741759 0.000000e+00       -Inf

$category$`5`
                                    Cla/Mod   Mod/Cla    Global       p.value     v.test
Exhibition=Exhibition_1           50.712909 78.286734 30.881828  0.000000e+00        Inf
Show=Show_1                       52.618557 75.907198 28.858741  0.000000e+00        Inf
Cinema=Cinema_1                   42.125633 84.176086 39.973819  0.000000e+00        Inf
Travelling=Travelling_1           42.491823 85.008923 40.021421  0.000000e+00        Inf
Sport=Sport_1                     41.615509 76.621059 36.832084 2.214018e-307  37.476622
Computer=Computer_1               39.392020 74.003569 37.581816 9.754161e-256  34.159227
Walking=Walking_1                 32.335329 80.309340 49.684636 3.333808e-183  28.864489
Reading=Reading_1                 27.878144 93.634741 67.190289 9.317077e-181  28.668911
Playing.music=Playing music_1     48.972603 42.534206 17.374747 7.539792e-173  28.027372
Listening.music=Listening music_1 26.769800 94.705532 70.772343 9.942971e-162  27.099640
Mechanic=Mechanic_1               32.325516 68.054729 42.115911 1.534386e-127  24.024720
Volunteering=Volunteering_1       46.770428 35.752528 15.292158 6.687073e-127  23.963475
Cooking=Cooking_1                 30.358112 66.567519 43.865286  1.275326e-97  20.968380
Profession=Management             45.342205 28.375967 12.519338  3.089491e-90  20.143165
Gardening=Gardening_1             29.410012 58.715051 39.938117  7.722217e-68  17.403789
TV=TV_1                           32.461161 23.616895 14.554326  4.067648e-29  11.200211
Marital.status=Single             26.635514 33.908388 25.467095  3.976567e-18   8.679459
….
Marital.status=Widower             7.356948  3.212374  8.734976  4.265027e-23  -9.897572
Age=(75,85]                        3.526971  1.011303  5.736047  1.089065e-27 -10.905156
Profession=Manual labourer         7.407407  5.116002 13.816494  6.865617e-37 -12.688304
Profession=Unskilled worker        3.914141  1.844140  9.425205  3.572894e-43 -13.775623
Gardening=Gardening_0             13.750743 41.284949 60.061883  7.722217e-68 -17.403789
Cooking=Cooking_0                 11.914352 33.432481 56.134714  1.275326e-97 -20.968380
Volunteering=Volunteering_0       15.172801 64.247472 84.707842 6.687073e-127 -23.963475
Mechanic=Mechanic_0               11.040296 31.945271 57.884089 1.534386e-127 -24.024720
Listening.music=Listening music_0  3.623779  5.294468 29.227657 9.942971e-162 -27.099640
Playing.music=Playing music_0     13.913294 57.465794 82.625253 7.539792e-173 -28.027372
Reading=Reading_0                  3.881030  6.365259 32.809711 9.317077e-181 -28.668911
Walking=Walking_0                  7.828761 19.690660 50.315364 3.333808e-183 -28.864489
Computer=Computer_0                8.331745 25.996431 62.418184 9.754161e-256 -34.159227
Sport=Sport_0                      7.403919 23.378941 63.167916 2.214018e-307 -37.476622
Exhibition=Exhibition_0            6.284435 21.713266 69.118172  0.000000e+00       -Inf
Show=Show_0                        6.774841 24.092802 71.141259  0.000000e+00       -Inf
Cinema=Cinema_0                    5.273592 15.823914 60.026181  0.000000e+00       -Inf
Travelling=Travelling_0            5.000000 14.991077 59.978579  0.000000e+00       -Inf


$quanti.var
                   Eta2 P-value
nb.activitees 0.7569577       0

$quanti
$quanti$`1`
                 v.test Mean in category Overall mean sd in category Overall sd p.value
nb.activitees -64.08247         3.061913        6.866       1.305579   3.382391       0

$quanti$`2`
                 v.test Mean in category Overall mean sd in category Overall sd     p.value
nb.activitees -3.153909         6.645126        6.866       1.745287   3.382391 0.001610995

$quanti$`3`
NULL

$quanti$`4`
                v.test Mean in category Overall mean sd in category Overall sd      p.value
nb.activitees 18.19659         9.018792        6.866       2.473412   3.382391 5.492415e-74

$quanti$`5`
                v.test Mean in category Overall mean sd in category Overall sd p.value
nb.activitees 62.13467         11.45092        6.866       1.743553   3.382391       0


> sum(hclu$call$t$inert.gain[1:5])/hclu$call$t$within[1]
[1] 0.4707469
```

Problem 2 Geomorphology dataset: All qüestions account for 1 point

The data used refer to geomorphology analysis. It is a data frame with 75 rows (the number of samples) and 11 columns: **drift** column corresponds to the target variable to be accounted for. Nevertheless, **p20** variable has to be investigated to understand the meaning. Source: The dataset is analysed in: http://www.sciencedirect.com/science/article/pii/S0169555X11006362.

1. Indicate the R command to perform a Principal Component Analysis on geomorphology data considering supplementary variable/s.

```
res.pca = PCA(geomorphology, quali.sup = 4, quanti.sup=3, graph=F) #
```

To use as supplementary variables categorical target drift (fourth column) and quantitavive p20 (third column).

2. Describe feature selection for geomorphology drift. Global factor and variable relations have to be detailed and hypothesis test explained.

Output for catdes() has to be interpreted. Numeric variables globally related to drift target are obtained by checking H0: Eta 2 =0 (based on a Fisher test) that it is rejected for Wind.Effect, Latitude, Terrain.Ruggedness.Index and Altitude. Also p20 var is globally affecting drift.

For each category in drift target a test on means in category compared to global mean for all numeric variables is stated (H0. Mean(X in category) = Global mean (X) based on a t-Student test) and those numeric variables rejecting H0 a drift category are indicated:

- *For Drift-Beach: Latitude in group is over global mean latitude and Altitude in group is under global mean altitude.*
- *For Drift-Diamic: Wind.effect, Altitude and p20 in group are over global means.*
- *For Drift-Kame: Valley.depth is significantly over the global mean and p20, Latitude and Wind.effect in group are under global means.*
- *For Drift-Landslide: Terrain.Ruggedness.Index is over global mean.*
- *For Drift-Terraces: Latitude and Wind.effect in group are under global means.*
- *For Drift-Organic soil no remarkable differences with respect to global means are found for any numeric variable in dataset.*

3. Describe numeric feature selection for the p20 characteristic. Hypothesis tests should be explained.

Output for condes() has to be interpreted.

P20 variable is globally and directly related to Altitude and Block.size.median. More intense is relation to Altitude than to Block.size.median. A correlation test based on Pearson correlation coefficient is checked where H0: Correlation(X,p20) = 0 and it is rejected for both Altitude and Block.size.median.

P20 is globally associated with factor Drift according to an R2 Test (H0: R2=0 – based on Fisher test for general linear models).

4. Describe the profile for the p20 characteristic. Detail significant levels and effects on means of numeric variable.

Drift.Diamict and Drift.Kame categories for Drift show a P20 mean 3.91 units over global P20 mean and -7.14 units under global P20, respectively. A t.Student test type for means checked: H0: Mean(P20) in Drift.category = Global P20 mean and it is rejected for Diamict and Kame Drift levels.

So, P20 is globally related to Altitude and Block.size.median and P20 is over the mean for Diamict Drift terrains while P20 is under the mean for Kame Drift terrains.

P20 meaning does not become clear.

5. Determine the number of significant axes according to Kaiser's rule and the resulting explained inertia.

Only dimensions 1 to 3 show an eigenvalue over 1 and since normalized PCA is used according to Kaiser's rule (strictly applied) these axes should be retained. They explain 58% of the total inertia of data.

6. Which are the variables with the best representation for the two first principal components?

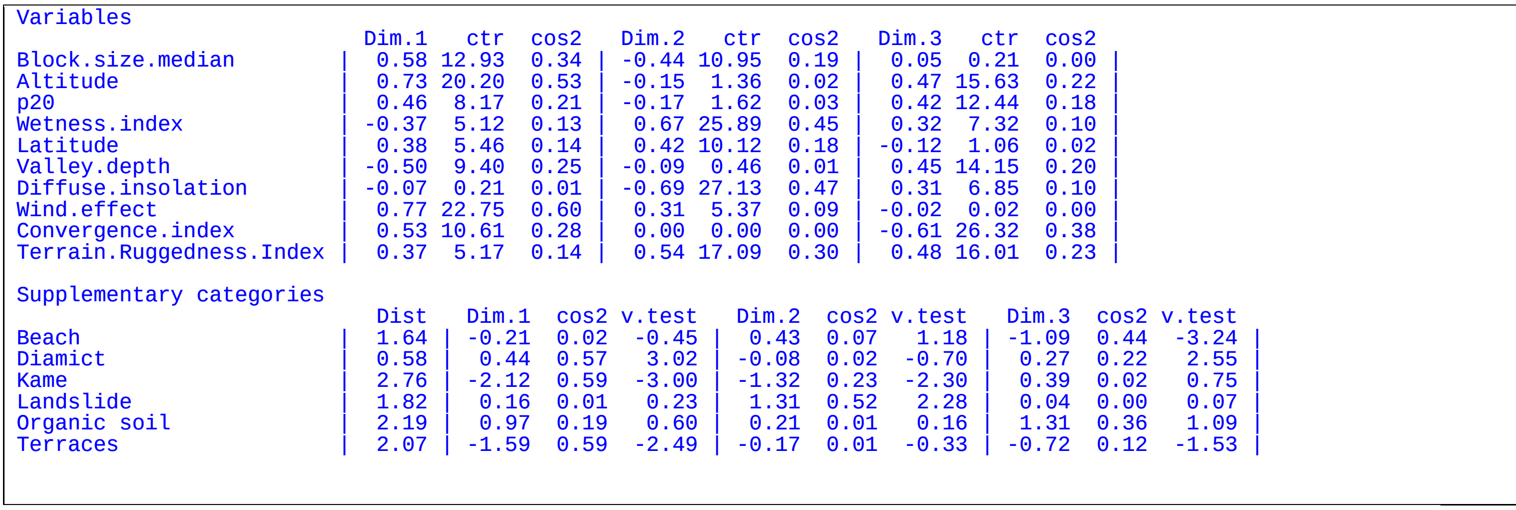
Best representation is analyzed through cos2 (squared correlation to factorial axes). Wind.effect (0.6), Altitude (0.53) and less Block.size.Median characterize first axis, all of them positively correlated. For second axis Diffuse.insolation (0.49), Wetness.index (0.45) and Terrain.Ruggedness.Index (0.30) characterize the meaning of the axis; Diffuse.insolation is inversely related to Wetness.index and Terrain.Ruggedness.Index.

7. Determine the most contributive observations for each axes in the first factorial plane.

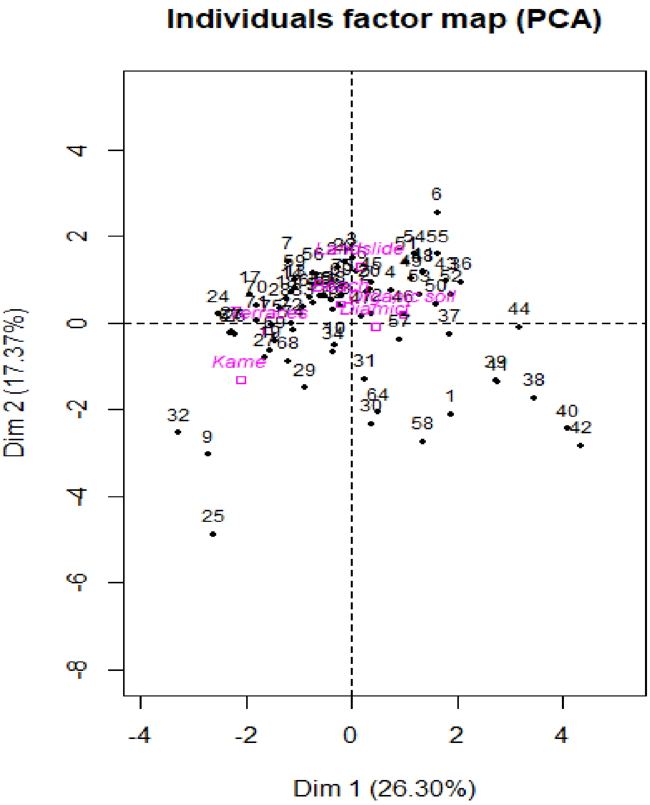
We have to take a look to individuals plot. Most contributive observations in PCA are those with absolute higher coordinates. 32, 9, 25 (on the left) and 38, 40, 42 (on the right side) of first axis and 6 and 25 for the second axis.

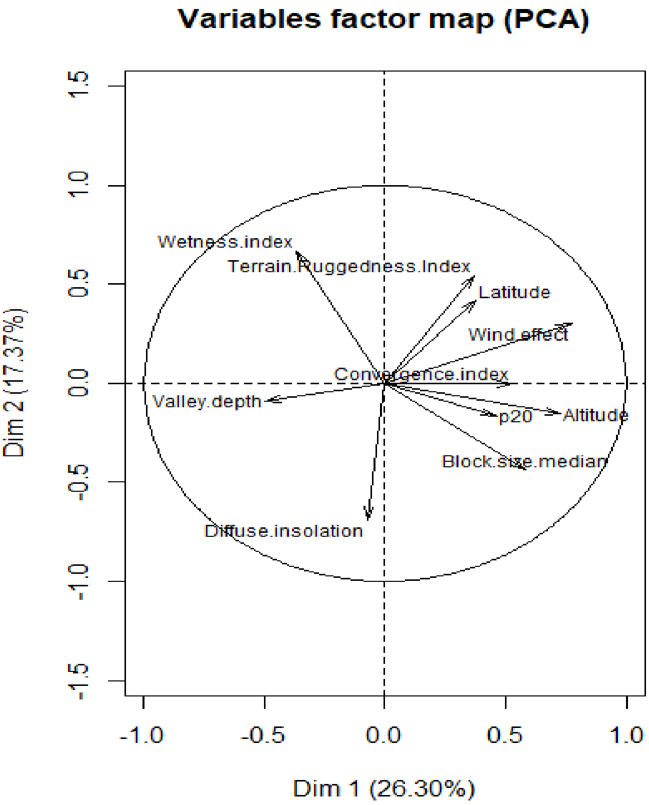
Observation 25 seems to be a multivariant outlier since it is far away from the center of gravity and show extreme coordinates in both first and second axis.

```
> library(FactoMineR)
> library(car)
> data(geomorphology)
> summary(geomorphology)
 Block.size.median    Altitude          p20                   Drift    Wetness.index       Latitude      Valley.depth
 Min.   : 1.30     Min.   : 15.0   Min.   : 5.25   Beach        :11   Min.   :-16.160   Min.   :65.40   Min.   :  0.94
 1st Qu.: 5.95     1st Qu.: 70.0   1st Qu.:12.96   Diamict      :47   1st Qu.:-10.522   1st Qu.:65.58   1st Qu.: 14.12
 Median : 7.40     Median :174.0   Median :20.71   Kame         : 5   Median : -8.569   Median :65.79   Median : 33.14
 Mean   :12.13     Mean   :278.1   Mean   :21.22   Landslide    : 5   Mean   : -8.792   Mean   :65.77   Mean   : 95.87
 3rd Qu.:10.60     3rd Qu.:422.5   3rd Qu.:28.50   Organic soil: 1   3rd Qu.: -6.576   3rd Qu.:65.98   3rd Qu.:134.30
 Max.   :68.90     Max.   :955.0   Max.   :44.39   Terraces     : 6   Max.   : -3.971   Max.   :66.11   Max.   :642.65
 Diffuse.insolation  Wind.effect     Convergence.index Terrain.Ruggedness.Index
 Min.   :0.4083     Min.   :0.6524   Min.   :-54.106   Min.   : 0.000
 1st Qu.:0.4125     1st Qu.:0.7643   1st Qu.: -1.600   1st Qu.: 1.113
 Median :0.4205     Median :0.8570   Median :  2.008   Median : 3.031
 Mean   :0.4460     Mean   :0.8275   Mean   :  3.549   Mean   : 4.207
 3rd Qu.:0.4442     3rd Qu.:0.8945   3rd Qu.: 10.353   3rd Qu.: 5.479
 Max.   :0.8402     Max.   :0.9555   Max.   : 66.059   Max.   :18.428
> names(geomorphology)
 [1] "Block.size.median"        "Altitude"                 "p20"                      "Drift"
 [5] "Wetness.index"            "Latitude"                 "Valley.depth"             "Diffuse.insolation"
 [9] "Wind.effect"              "Convergence.index"        "Terrain.Ruggedness.Index"
> condes(geomorphology,3)
$`quanti`
                  correlation      p.value
Altitude            0.4255372 0.0001412487
Block.size.median   0.2899094 0.0116373485

$quali
             R2     p.value
Drift 0.1927639 0.009991143

$category
         Estimate     p.value
Diamict  3.911609 0.001353439
Kame    -7.135752 0.035552596
 > catdes(geomorphology,num.var=4)

Link between the cluster variable and the quantitative variables
================================================================
                              Eta2      P-value
Wind.effect              0.4219796 2.858260e-07
Latitude                 0.3720476 4.180150e-06
Terrain.Ruggedness.Index 0.2221201 3.350538e-03
Altitude                 0.2192816 3.735253e-03
p20                      0.1927639 9.991143e-03

Description of each cluster by quantitative variables
=====================================================
$`Beach`
            v.test Mean in category Overall mean sd in category  Overall sd      p.value
Latitude  4.084264         66.00577     65.76955      0.1097896   0.2062583 4.421682e-05
Altitude -2.822595         55.63636    278.09333     30.8876931 281.0733794 4.763676e-03

$Diamict
              v.test Mean in category Overall mean sd in category   Overall sd      p.value
Wind.effect 3.720687        0.8565728    0.8275324     0.07494895   0.08698909 0.0001986817
Altitude    3.579329      368.3617021  278.0933333   310.28281029 281.07337943 0.0003444772
p20         3.125911       23.8193617   21.2161333     8.89468175   9.28156982 0.0017725520

$Kame
                v.test Mean in category Overall mean sd in category   Overall sd      p.value
Valley.depth  2.839554      259.2002844   95.8739521   113.69570977 132.23847835 0.0045176578
p20          -2.091633       12.7720000   21.2161333     4.42426220   9.28156982 0.0364713569
Latitude     -2.833715       65.5153308   65.7695543     0.03953775   0.20625832 0.0046010389
Wind.effect  -3.698889        0.6875788    0.8275324     0.03826181   0.08698909 0.0002165455

$Landslide
                           v.test Mean in category Overall mean sd in category Overall sd     p.value
Terrain.Ruggedness.Index 3.079081         9.884088     4.207276       5.965207   4.238725 0.002076402

$`Organic soil`
NULL

$Terraces
               v.test Mean in category Overall mean sd in category Overall sd      p.value
Latitude    -2.090087       65.5996092   65.7695543     0.18048700 0.20625832 0.0366100182
Wind.effect -3.496109        0.7076425    0.8275324     0.03401363 0.08698909 0.0004720967

> res.pca = PCA(geomorphology, quali.sup = 4,graph=F) #
> summary(res.pca,nb.dec=2,nbind = 0,nbelements=Inf)

Call:
PCA(X = geomorphology, quali.sup = 4, graph = F)


Eigenvalues
                      Dim.1  Dim.2  Dim.3  Dim.4  Dim.5  Dim.6  Dim.7  Dim.8  Dim.9 Dim.10
Variance               2.63   1.74   1.43   0.99   0.97   0.64   0.62   0.38   0.35   0.25
% of var.             26.30  17.37  14.29   9.95   9.75   6.38   6.18   3.81   3.48   2.50
Cumulative % of var.  26.30  43.67  57.96  67.90  77.65  84.02  90.21  94.02  97.50 100.00
```



```
Variables
                             Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr   cos2  
Block.size.median         |  0.58 12.93  0.34 | -0.44 10.95  0.19 |  0.05  0.21  0.00 |
Altitude                  |  0.73 20.20  0.53 | -0.15  1.36  0.02 |  0.47 15.63  0.22 |
p20                       |  0.46  8.17  0.21 | -0.17  1.62  0.03 |  0.42 12.44  0.18 |
Wetness.index             | -0.37  5.12  0.13 |  0.67 25.89  0.45 |  0.32  7.32  0.10 |
Latitude                  |  0.38  5.46  0.14 |  0.42 10.12  0.18 | -0.12  1.06  0.02 |
Valley.depth              | -0.50  9.40  0.25 | -0.09  0.46  0.01 |  0.45 14.15  0.20 |
Diffuse.insolation        | -0.07  0.21  0.01 | -0.69 27.13  0.47 |  0.31  6.85  0.10 |
Wind.effect               |  0.77 22.75  0.60 |  0.31  5.37  0.09 | -0.02  0.02  0.00 |
Convergence.index         |  0.53 10.61  0.28 |  0.00  0.00  0.00 | -0.61 26.32  0.38 |
Terrain.Ruggedness.Index  |  0.37  5.17  0.14 |  0.54 17.09  0.30 |  0.48 16.01  0.23 |

Supplementary categories
                              Dist    Dim.1   cos2 v.test    Dim.2   cos2 v.test    Dim.3   cos2 v.test  
Beach                     |  1.64 | -0.21  0.02  -0.45 |  0.43  0.07   1.18 | -1.09  0.44  -3.24 |
Diamict                   |  0.58 |  0.44  0.57   3.02 | -0.08  0.02  -0.70 |  0.27  0.22   2.55 |
Kame                      |  2.76 | -2.12  0.59  -3.00 | -1.32  0.23  -2.30 |  0.39  0.02   0.75 |
Landslide                 |  1.82 |  0.16  0.01   0.23 |  1.31  0.52   2.28 |  0.04  0.00   0.07 |
Organic soil              |  2.19 |  0.97  0.19   0.60 |  0.21  0.01   0.16 |  1.31  0.36   1.09 |
Terraces                  |  2.07 | -1.59  0.59  -2.49 | -0.17  0.01  -0.33 | -0.72  0.12  -1.53 |
```


Individuals factor map (PCA)
Dim 2 (17.37%)
Dim 1 (26.30%)
Landslide
Beach
Organic soil
Diamict
Terraces
Kame


Variables factor map (PCA)
Dim 2 (17.37%)
Dim 1 (26.30%)
Wetness.index
Terrain.Ruggedness.Index
Latitude
Wind.effect
Convergence.index
Valley.depth
p20
Altitude
Block.size.median
Diffuse.insolation