European
Commission

Principal components of distributional SD

A. Irpino, R. Verde
ESTP Cologne 14-16 May 2024

Principal Component Analysis of distributional data

We use the HistDAWass R package for showing the main procedures.

```r
# install.packages("HistDAWass",dependencies = T) #Installing the first time
library(HistDAWass)
```

The main procedures available for the PCA are:

- WH.1d.PCA for the analysis of a single distributional variable.
- WH.MultiplePCA for analysing more than one distributional variable.

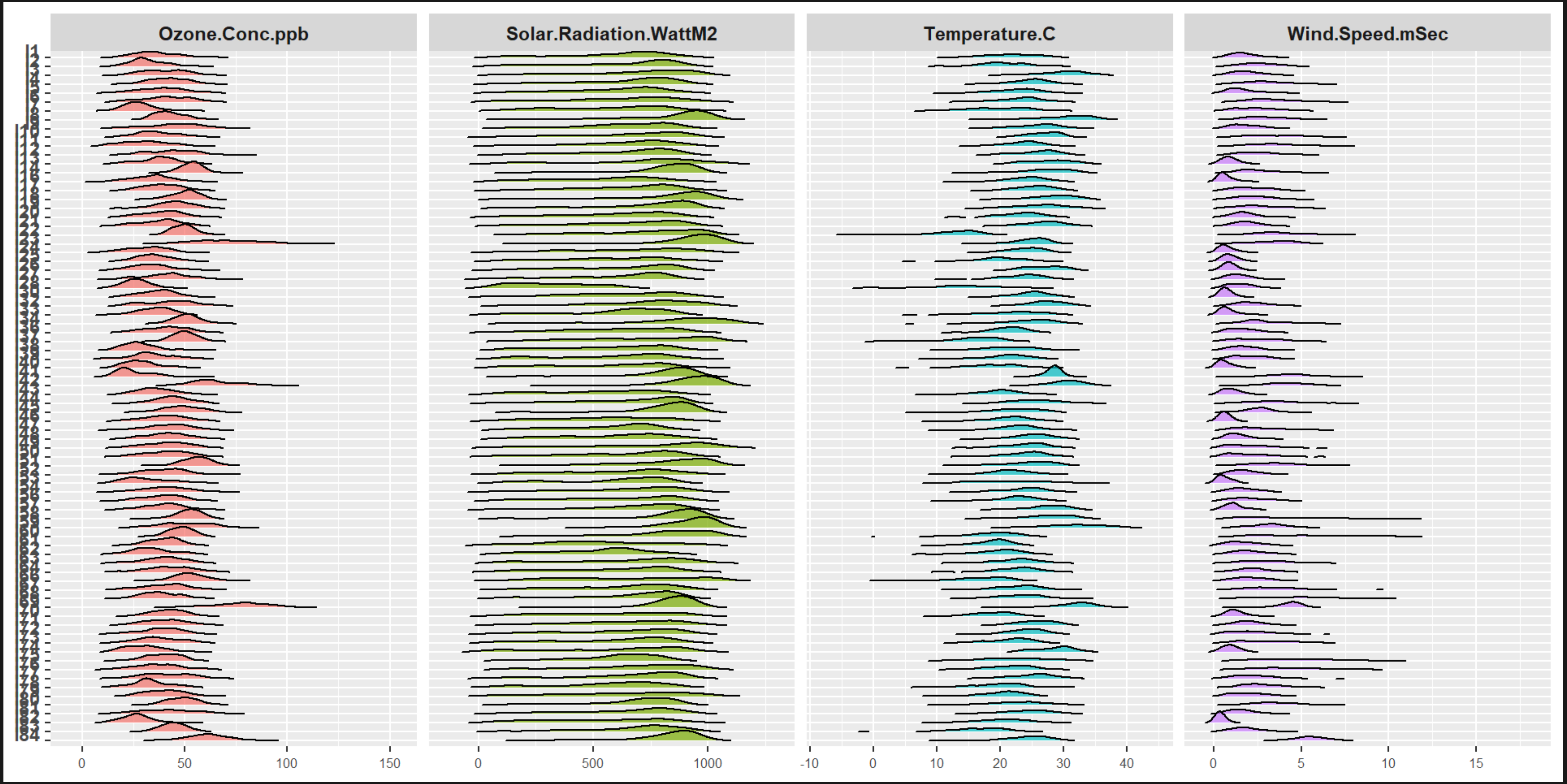
The Ozone dataset

The dataset contains MatH (matrix of histogram-valued data) object This data set list 78 stations located in the USA recording four variables, without missing data.

```r
get.MatH.main.info(OzoneFull)
```

```
$nrows
[1] 78

$ncols
[1] 4

$rownames
 [1] "I1"  "I2"  "I3"  "I4"  "I5"  "I6"  "I7"  "I8"  "I9"  "I10" "I11" "I12"
[13] "I13" "I14" "I16" "I17" "I18" "I19" "I20" "I21" "I22" "I23" "I24" "I25"
[25] "I26" "I27" "I28" "I30" "I32" "I33" "I34" "I36" "I37" "I38" "I39" "I40"
[37] "I41" "I42" "I43" "I44" "I45" "I46" "I47" "I48" "I49" "I50" "I51" "I52"
[49] "I53" "I54" "I56" "I57" "I58" "I59" "I60" "I61" "I62" "I63" "I64" "I66"
[61] "I67" "I68" "I69" "I70" "I71" "I72" "I73" "I74" "I75" "I76" "I77" "I78"
[73] "I79" "I80" "I81" "I82" "I83" "I84"

$varnames
[1] "Ozone.Conc.ppb"         "Temperature.C"          "Solar.Radiation.WattM2"
[4] "Wind.Speed.mSec"
```

The data



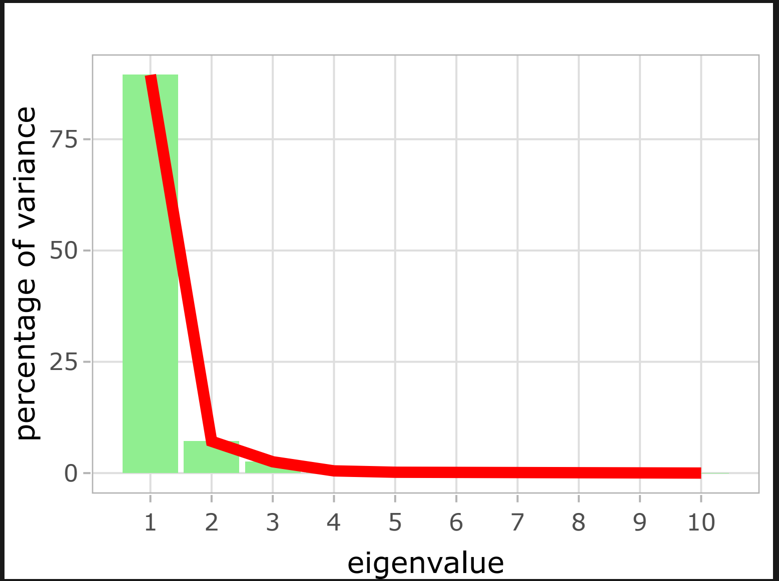
Ozone.Conc.ppb
Solar.Radiation.WattM2
Temperature.C
Wind.Speed.mSec
0
50
100
150
0
500
1000
-10
0
10
20
30
40
0
5
10
15
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84

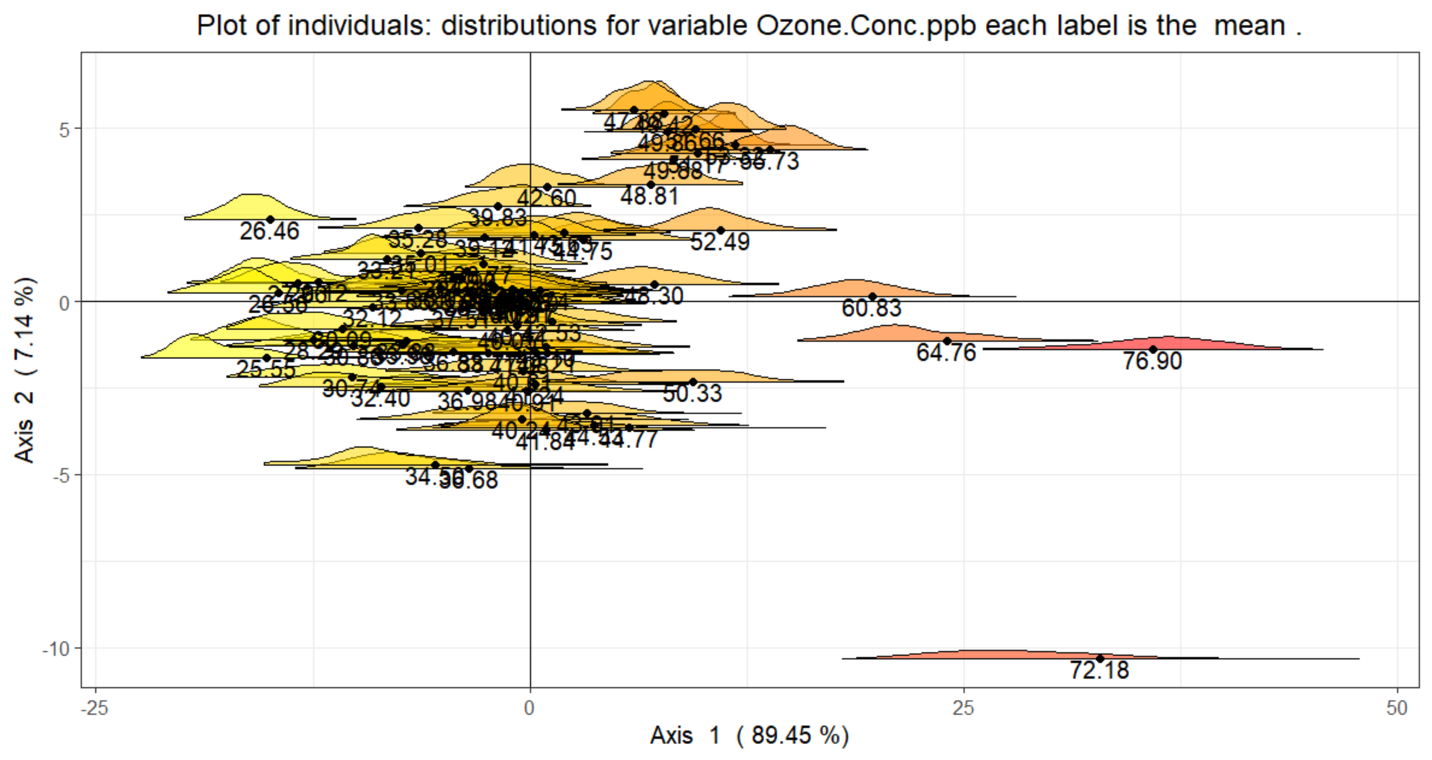
1d PCA: the analysis of `Ozone COncentration ppb`

```
OZ_1d_PCA<-WH.1d.PCA(OzoneFull,1,quantiles=20)
```

```
We do a PCA on variable --->  Ozone.Conc.ppb
```

Let’s see the scree plot and the first plane

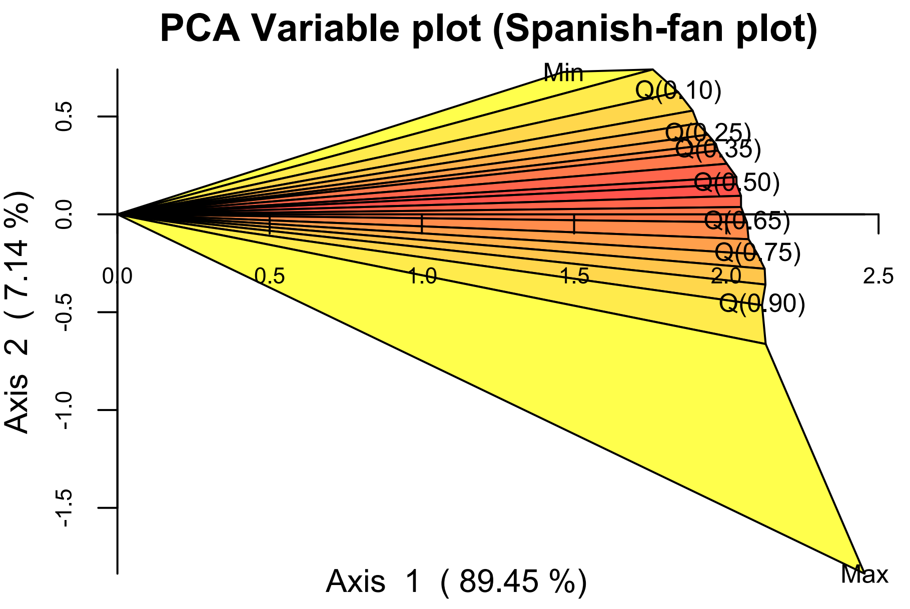


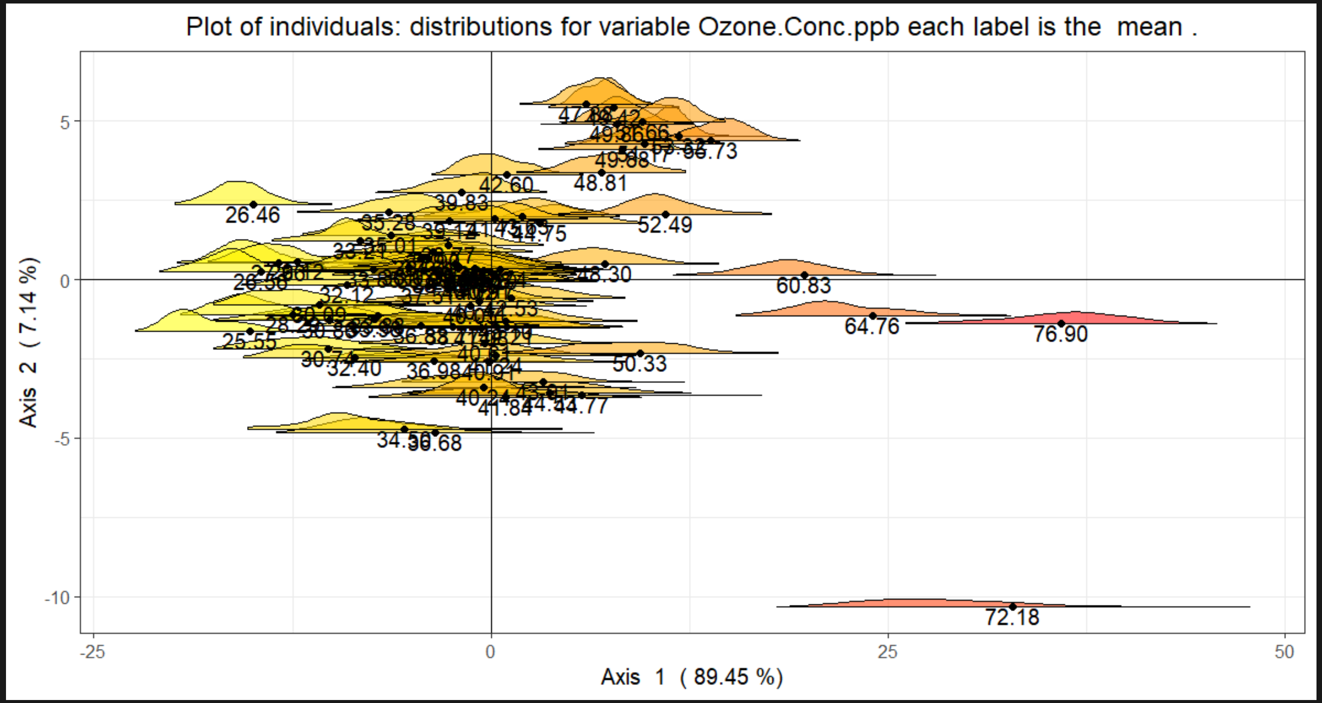


Stations on the first plane

Interpreting PCs

The Spanish fan plot for variable correlation





Stations on the first plane

PCA of all the variables

Now we take into consideration all the four variables and we start performing the multiple PCA using the function `WH.MultiplePCA`

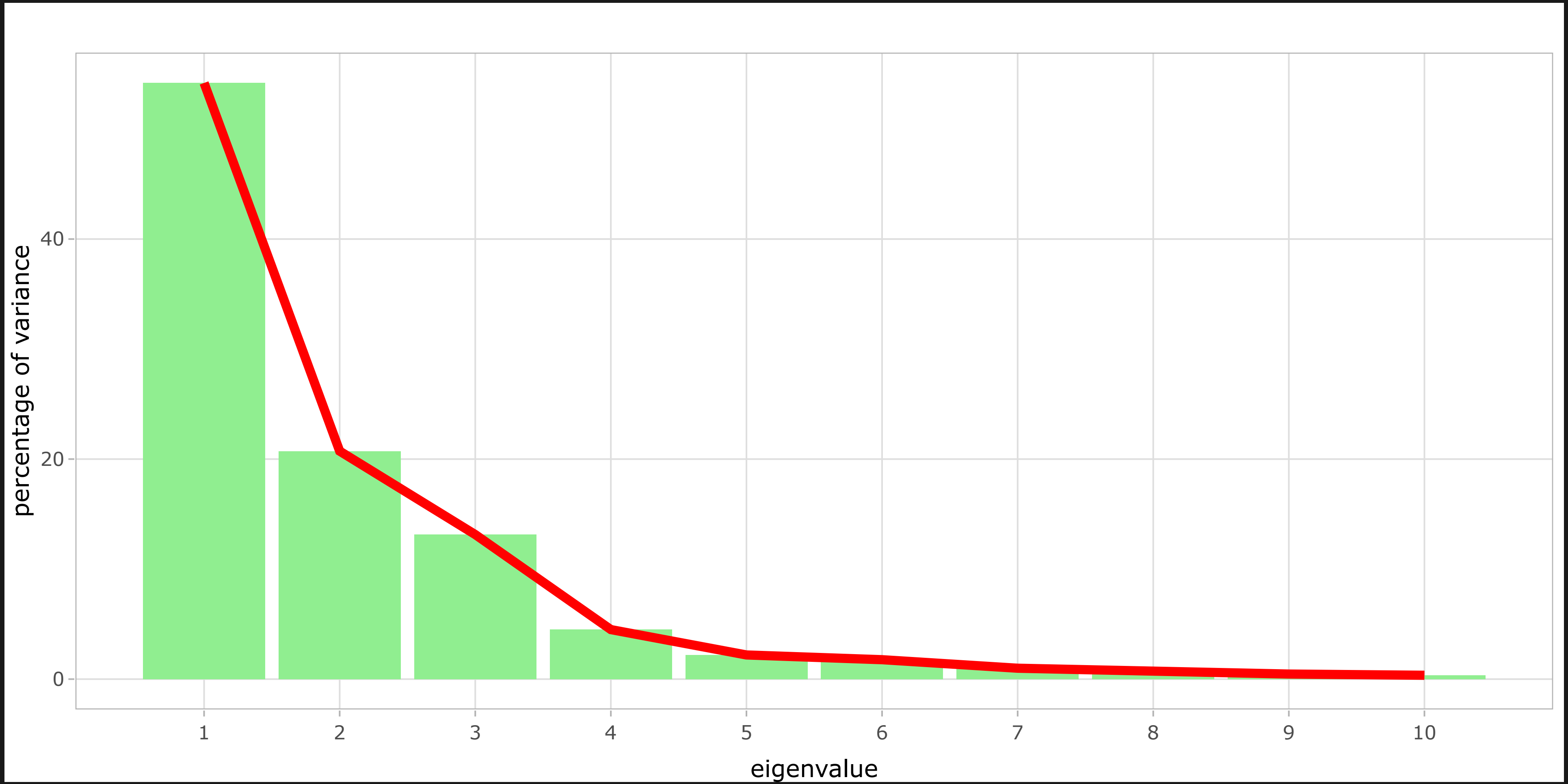
```r
OZ_PCA<-WH.MultiplePCA(data=OzoneFull,list.of.vars = c(1:4),quantiles = 20)
```

The code executes a Multiple PCA and produces a set of textual and graphical outputs. The code make use of the `FactoMiner` package which is specialized for dimension reduction techniques and the output interpretation.

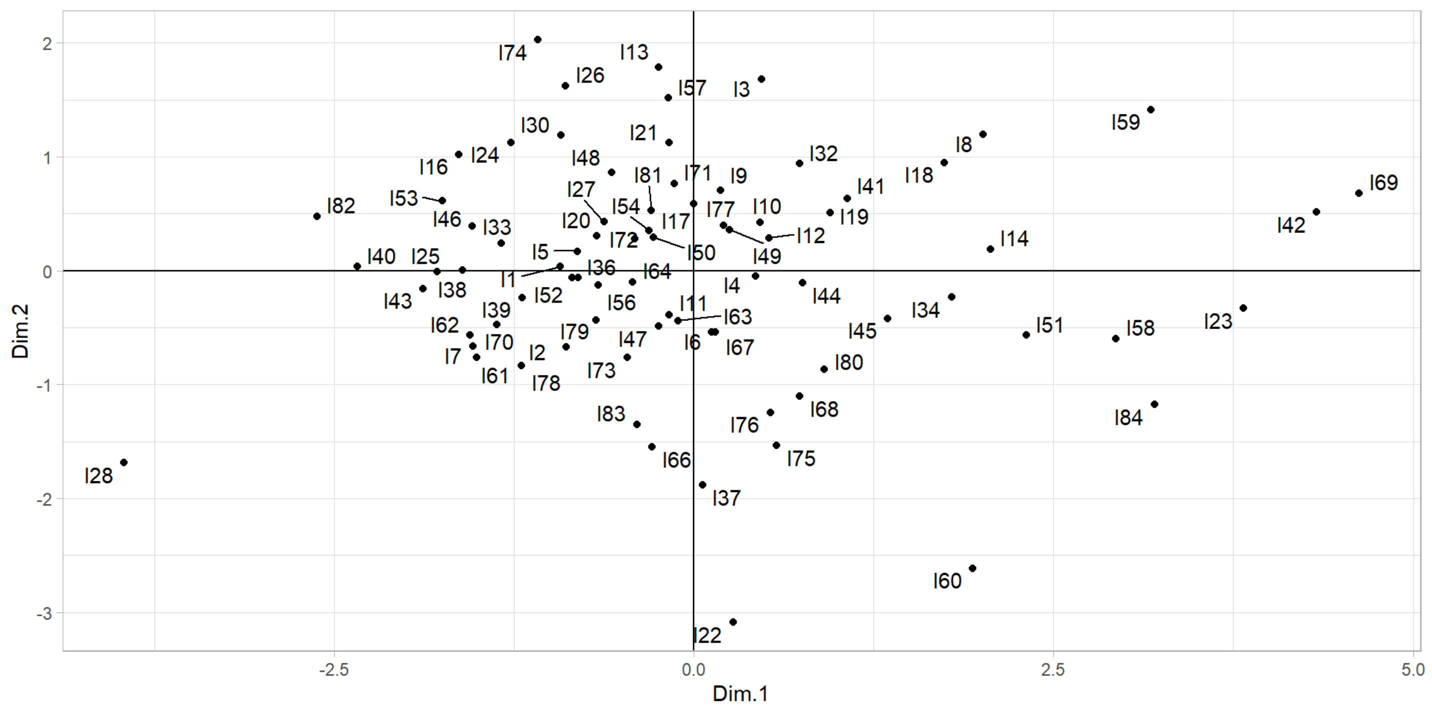
Ouput of the procedure

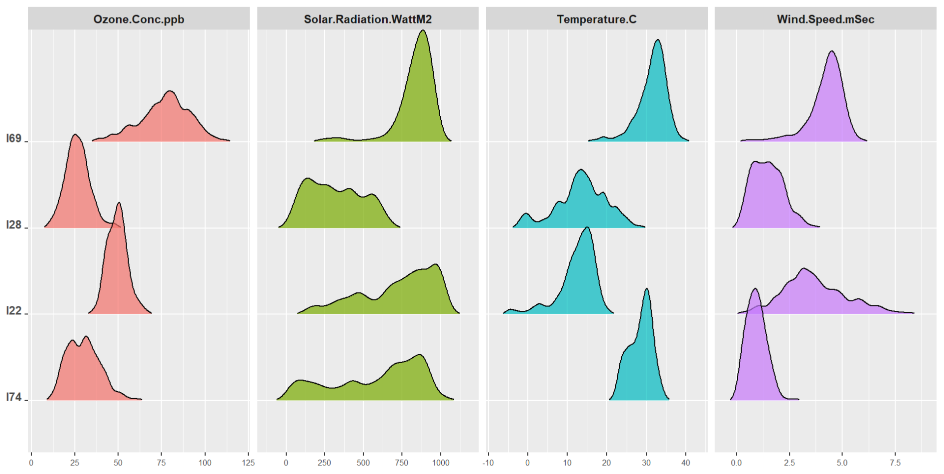
```
**Results of the Multiple Factor Analysis (MFA)**
The analysis was performed on 78 individuals, described by 84 variables
*Results are available in the following objects :

  name                   description
1 "$eig"                 "eigenvalues"
2 "$separate.analyses"   "separate analyses for each group of variables"
3 "$group"               "results for all the groups"
4 "$partial.axes"        "results for the partial axes"
5 "$inertia.ratio"       "inertia ratio"
6 "$ind"                 "results for the individuals"
7 "$quanti.var"          "results for the quantitative variables"
8 "$summary.quanti"      "summary for the quantitative variables"
9 "$global.pca"          "results for the global PCA"
```

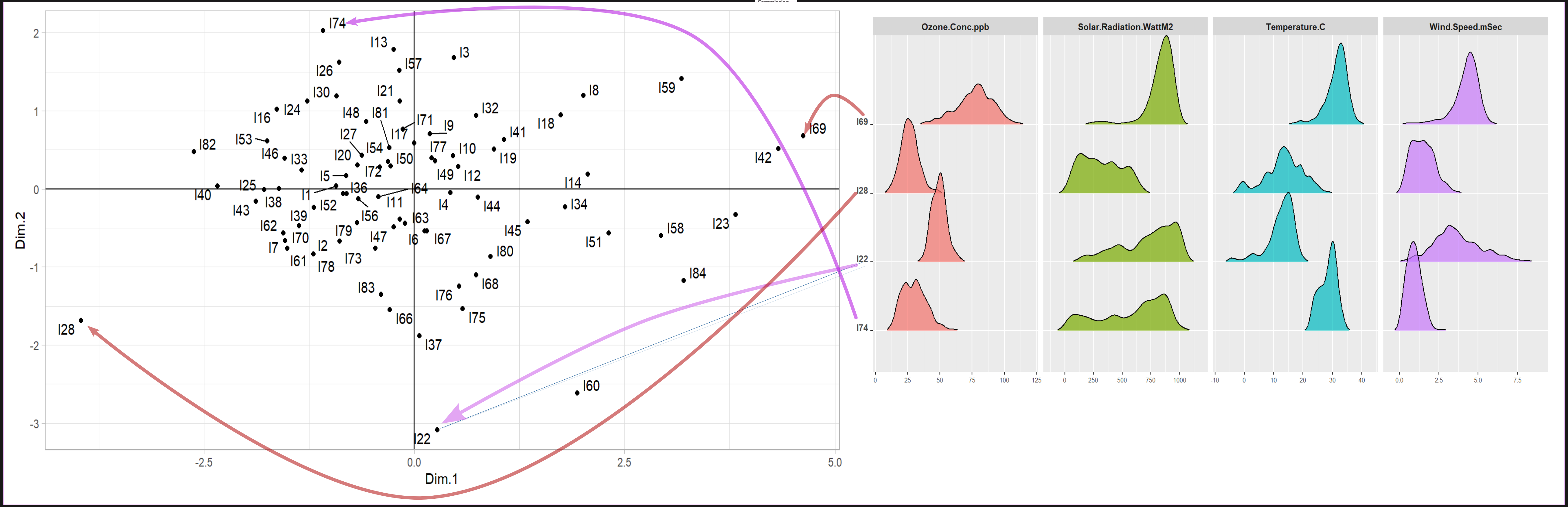
The analysis of the eigenvalues: the scree-plot



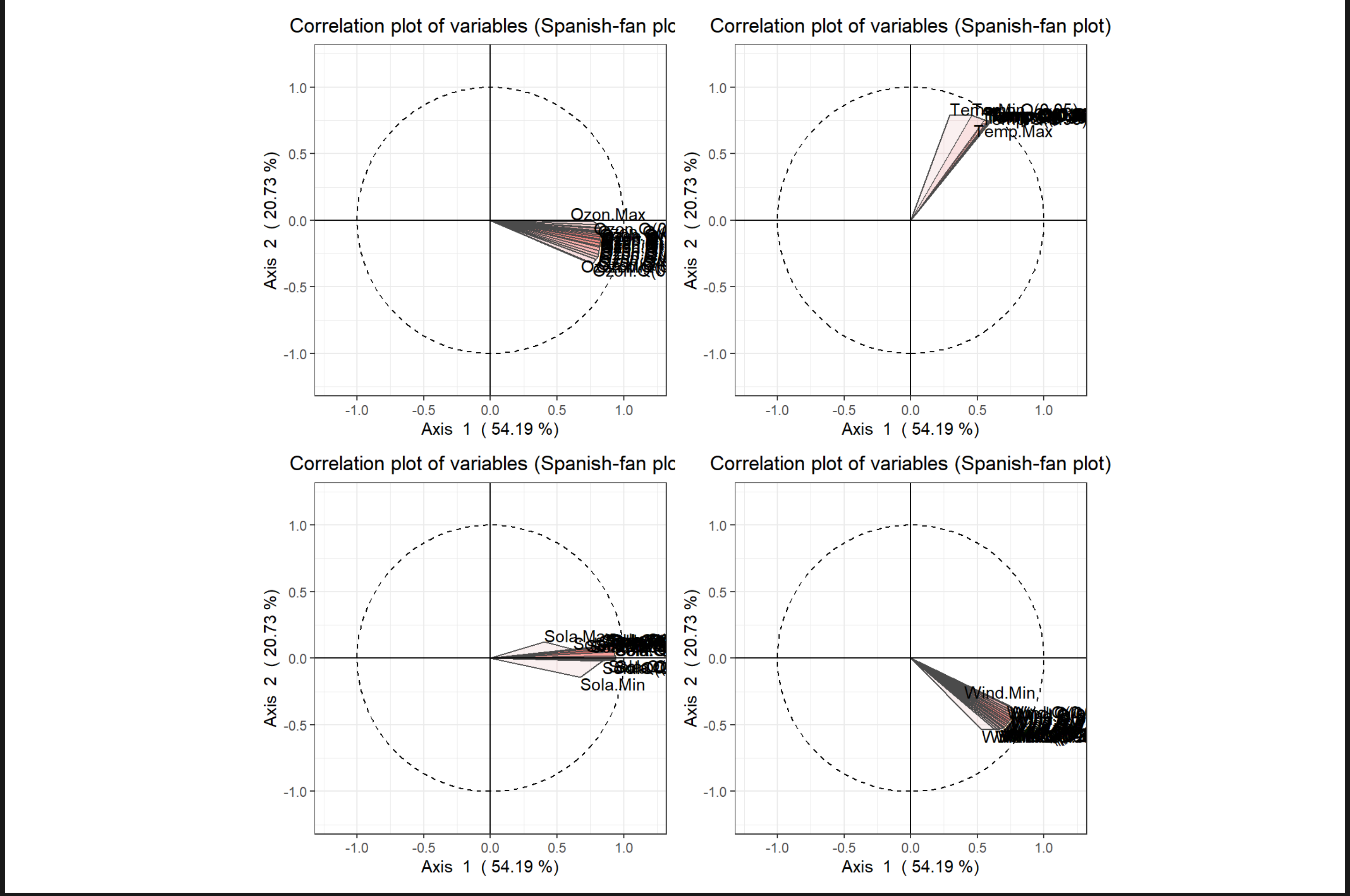
The plot of individuals (the 78 stations) on the first plane






I74
I13
I3
I26
I57
I30
I21
I8
I59
I16
I24
I48
I81
I32
I71
I9
I18
I17
I41
I69
I82
I53
I27
I46
I33
I54
I77
I10
I20
I50
I19
I42
I5
I72
I49
I12
I14
I25
I36
I40
I64
I1
I38
I11
I43
I52
I4
I44
I34
I39
I56
I63
I62
I79
I45
I23
I70
I47
I6
I67
I58
I7
I2
I51
I61
I73
I80
I78
I84
I83
I68
I76
I28
I75
I66
I37
I60
I22
2
1
0
-1
-2
-3
Dim.2
-2.5
0.0
2.5
5.0
Dim.1
Ozone.Conc.ppb
Solar.Radiation.WattM2
Temperature.C
Wind.Speed.mSec
I69
I28
I22
I74
0
25
50
75
100
125
0
250
500
750
1000
-10
0
10
20
30
40
0.0
2.5
5.0
7.5

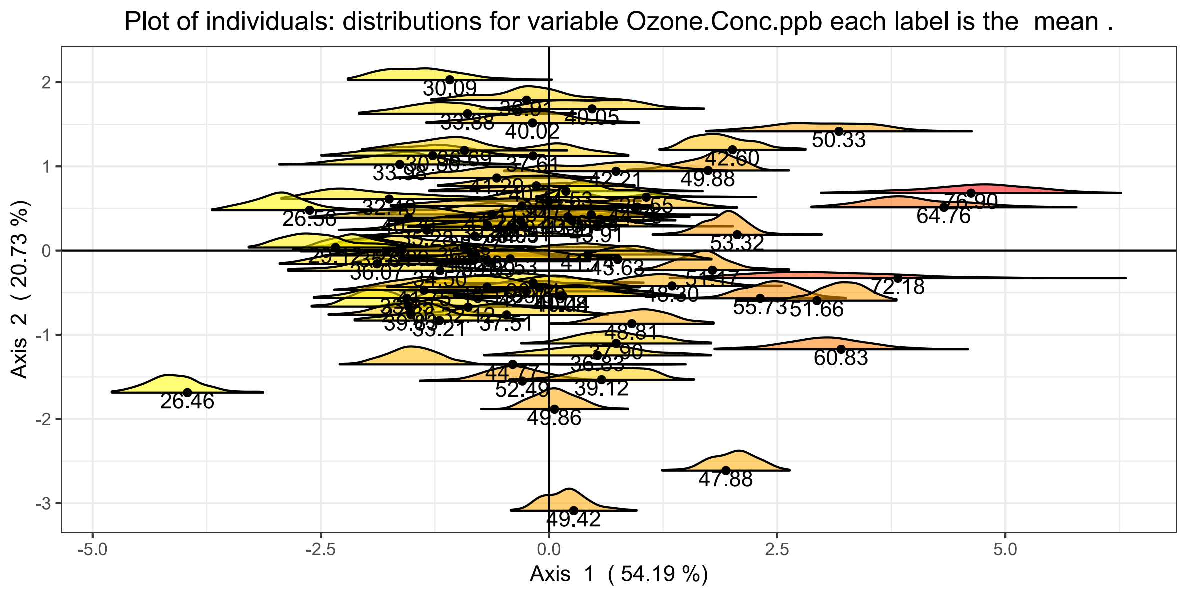
The plot(s) of variables on the first plane: the spanish fun plots

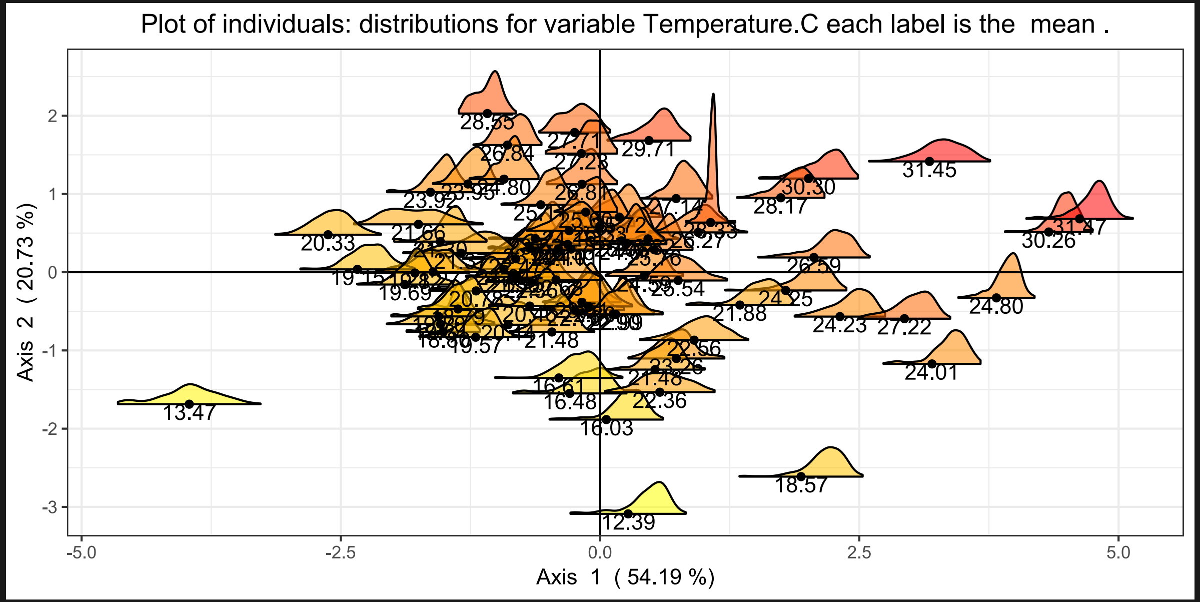


Some comments:

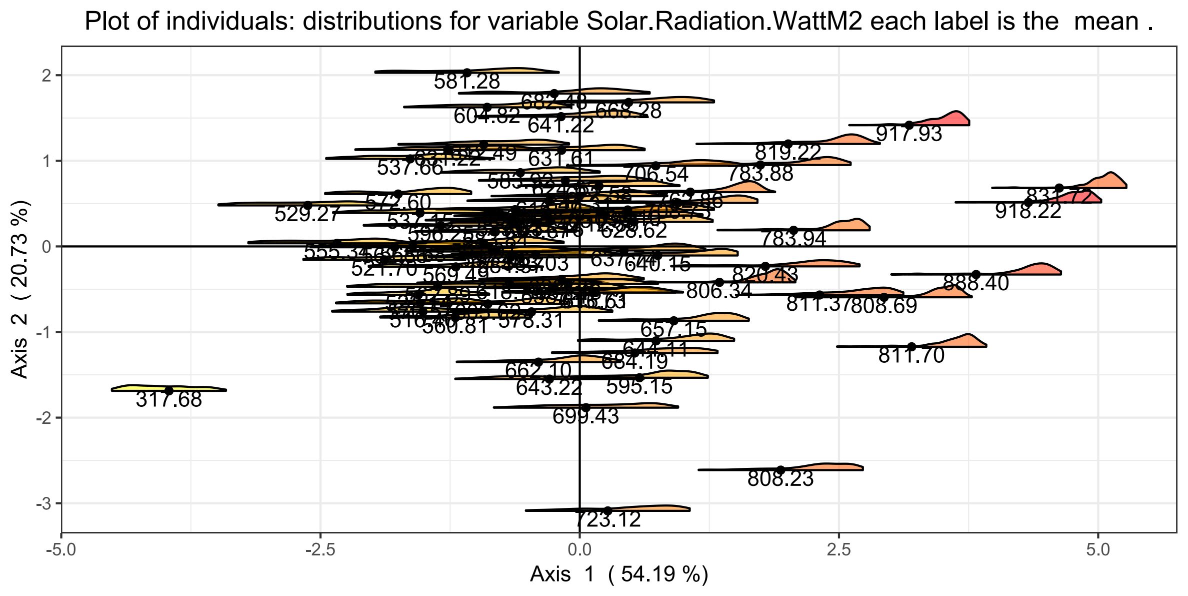
- The horizontal direction is highly correlated with the Ozone and solar radiation. And both, Ozone and Solar Radiation, are highly and positively correlated.
- The vertical dimension is more related to the Temperature.
- Temperature is rather uncorrelated to the Wind Speed (The two fans are almost at 90 degrees).
- Wind Speed is moderately and positively correlated with the Ozone, but the correlation decrease from the minimum of the wind speed to the maximum.

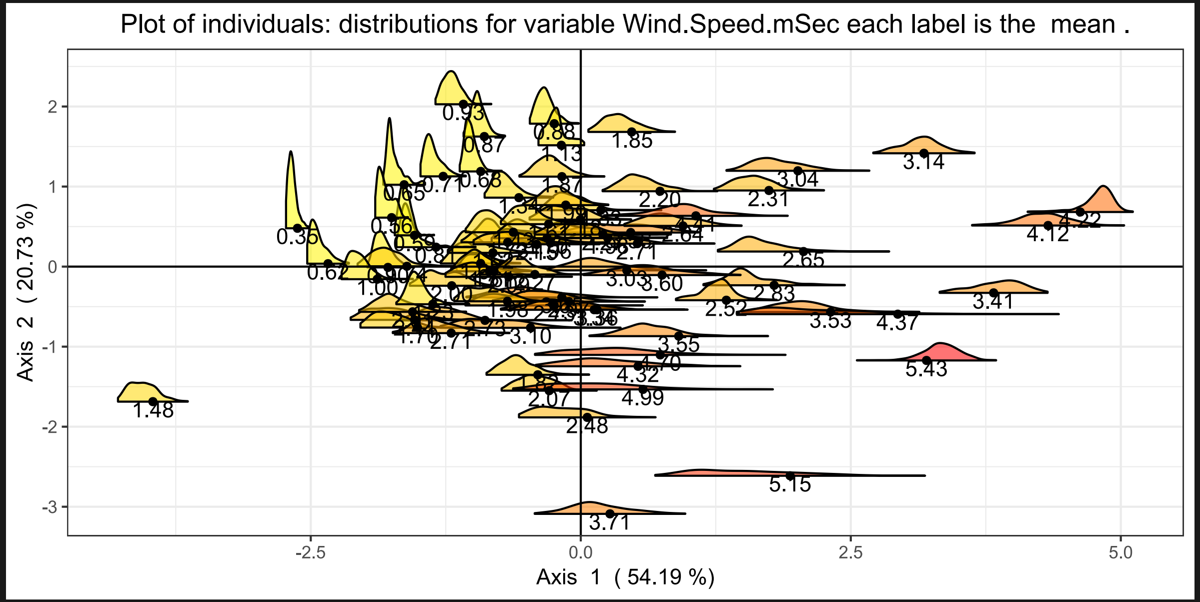
The plot of distributions for each variable 1


Plot of individuals: distributions for variable Ozone.Conc.ppb each label is the mean .

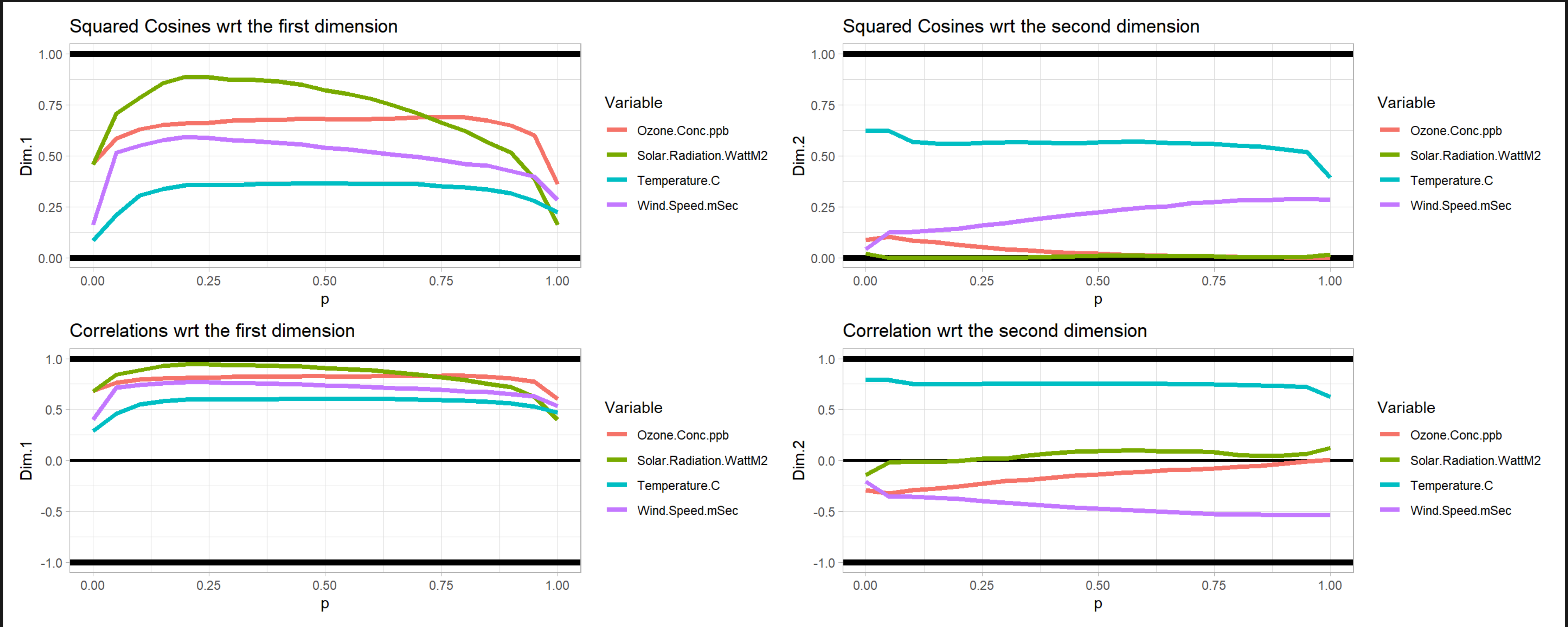

Plot of individuals: distributions for variable Temperature.C each label is the mean .

The plot of distributions for each variable 2





Tools for the interpretation of axes: plots of COS2 and correlations



Squared Cosines wrt the first dimension

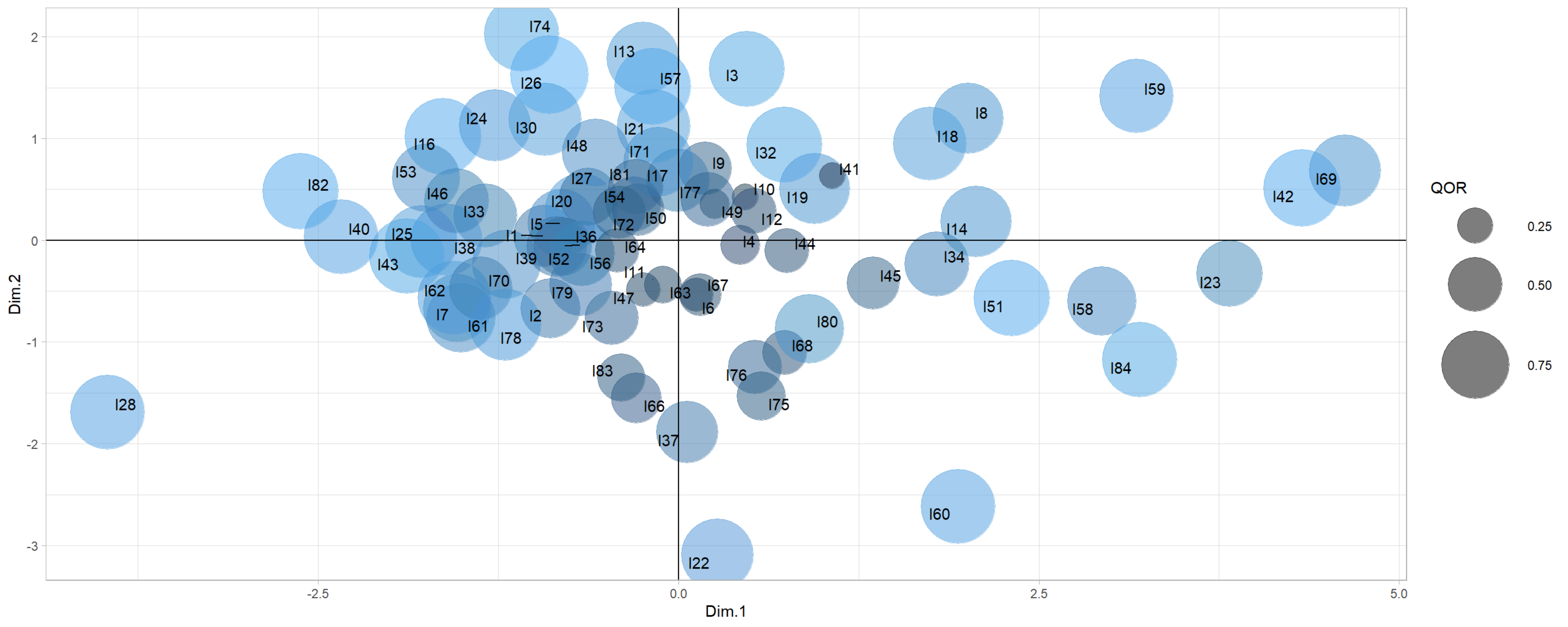
Squared Cosines wrt the second dimension

Correlations wrt the first dimension

Correlation wrt the second dimension

The quality of representation of individuals

The size of the balls is proportional to the quality of representation of points on the plane



Conclusions

- PCA for distributional data allows to discover more patterns in the data with recpect to the PCA on points;
- If data are intervals, you can consider them as uniform distributions, namely, a histogram with just one bin;
- If distributions have a discrete domain, it is easy to generalize the method.
- If distributions has a nominal support, you can’t use PCA (other methods are available).