Clamix - specificity of variable in cluster

R-Forge / Clamix [https://r-forge.r-project.org/scm/viewvc.php/pkg/R/clamix4.R?view=markup&revision=2&root=clamix&pathrev=2]

Problem

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

This morning (October 31, 2012) I had the idea to define for a selected variable V its **specificity** s(V,C) for a cluster C as

$s(V,C) = 1/2 \int |p_U(t) - p_C(t)| \, dt$

or in discrete case

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

Geometrically S(V,C) is the half area of the symmetric difference of the areas bellow the distribution of values of V on set of units U and the distribution of values of V on the cluster C. See Figure 1.

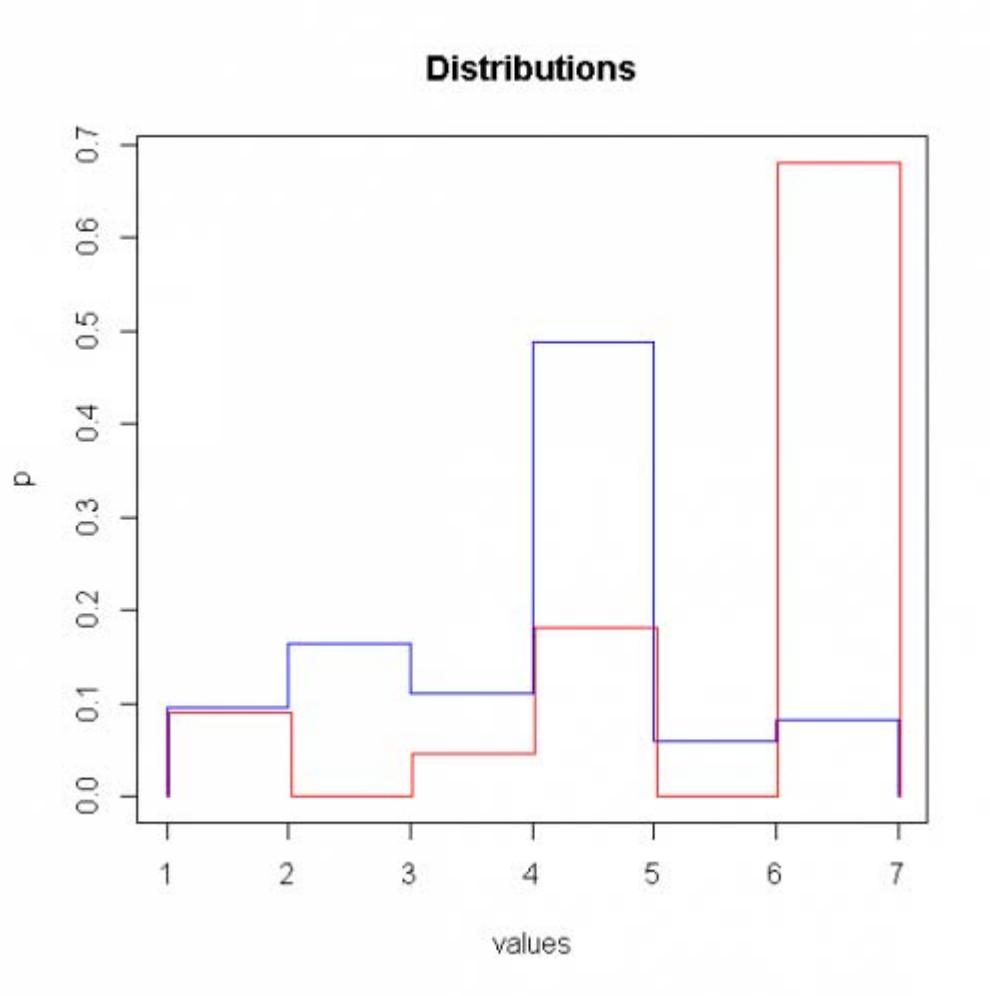
The specificity s(V,C) has the following properties:

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

Proof of 1.:

$s(V,C) = 1/2 \int |p_U(t) - p_C(t)| \, dt \leq 1/2 \int (p_U(t) + p_C(t)) \, dt = 1/2 (\int p_U(t) \, dt + \int p_C(t) \, dt) = (1+1)/2 = 1$

```
> fU <- c(71,123,83,365,44,62)
> fC <- c(2,0,1,4,0,15)
> (pU <- fU/sum(fU))
[1] 0.09491979 0.16443850 0.11096257 0.48796791 0.05882353 0.08288770
> (pC <- fC/sum(fC))
[1] 0.09090909 0.00000000 0.04545455 0.18181818 0.00000000 0.68181818
> (r <- sum(abs(pU-pC))/2)
[1] 0.5989305
> plot(c(1,7),c(0,max(max(pU),max(pC))),type="n",main="Distributions",
+   xlab="values",ylab="p")
> lines(c(0,pU),type="S",col="blue")
> lines(c(7,7),c(pU[6],0),col="blue")
> lines((1:7)+0.02,c(0,pC),type="S",col="red")
> lines(c(7,7)+0.02,c(pC[6],0),col="red")
```



For identifying the most characteristic values I would try with the index

$\max(p_U(v), p_C(v)) / \min(p_U(v), p_C(v))$

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

Example / Cars

I put the data and the code to specific.zip - I hope that there is all that is needed

I included functions `plotDistri` and `specific` into Clamix2 thus producing Clamix3.

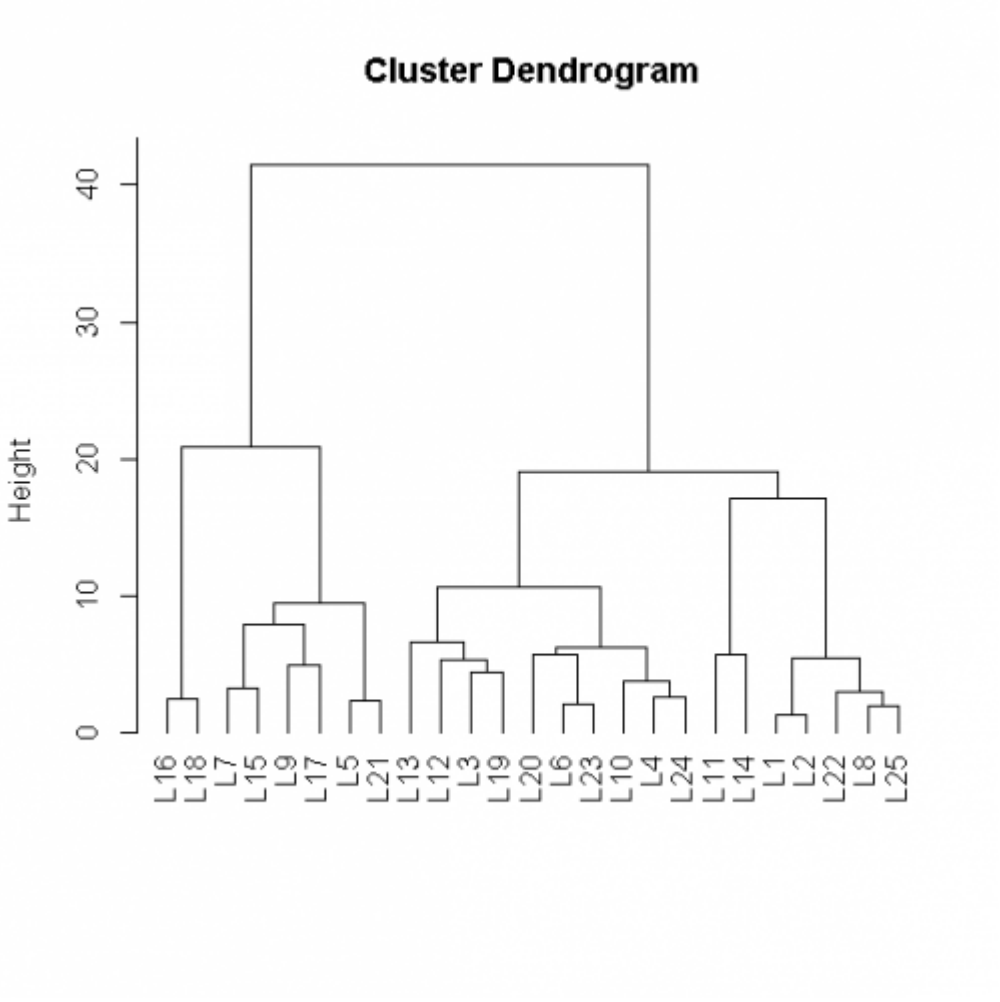
```
plotDistri <- function(pU,pC){
  ln <- length(pU)
  plot(c(1,ln),c(0,max(max(pU),max(pC))),type="n",main="Distributions",
    xlab="values",ylab="p")
  lines(c(0,pU[-ln]),type="S",col="blue")
  lines(c(ln,ln),c(pU[ln-1],0),col="blue")
  lines((1:ln)+0.02,c(0,pC[-ln]),type="S",col="red")
  lines(c(ln,ln)+0.02,c(pC[ln-1],0),col="red")
}

specific <- function(leader,var){
  Lq <- L[[leader]]
  names(Lq) <- names(total)
  q <- Lq[[var]]
  names(q) <- names(pU[[var]])
  ln <- length(q)
  q <- q/q[ln]
  c <- q[-ln]
  u <- pU[[var]][-ln]
  print(u)
  print(c)
  print(pmax(u,c)/pmin(u,c))
  plotDistri(pU[[var]],q)
```

```
}
```

And here is the code for clustering leaders from `cars25.rez`

```
setwd("C:/Users/Batagelj/work/clamix/clamix.R")
source("C:\\Users\\Batagelj\\work\\clamix\\clamix.R\\clamix3.R")
load("./cars2/cars25.rez")
load("./cars2/cars.so")
load("./cars2/cars.meta")
alpha <-rep(1/nVar,nVar)
hc <- hclustSO(rez$leaders)
plot(hc,hang=-1)
long[rez$clust==9]
L <- rez$leaders
total <- computeTotal(L)
objects()
```



and for producing ***specificity*** table S.

```
pU <- total
for(j in 1:nVar) pU[[j]] <- pU[[j]]/pU[[j]][[length(pU[[j]])]]
S <- matrix(0,nrow=length(L),ncol=length(total),
  dimnames=list(names(L),names(total)))
for(i in 1:length(L)){
  pC <- L[[i]]
  for(j in 1:nVar) {
    ln <- length(pC[[j]])
    pC[[j]] <- pC[[j]]/pC[[j]][[ln]]
    S[i,j] <- sum(abs(pU[[j]][-ln]-pC[[j]][-ln]))/2
  }
}
for(i in 1:length(L)) {
  cat(i,names(L)[i],"\n");
  print(sort(S[i,],decreasing=TRUE)[1:7])
}
```

Here is the list of the 7 most specific variables for each leader:

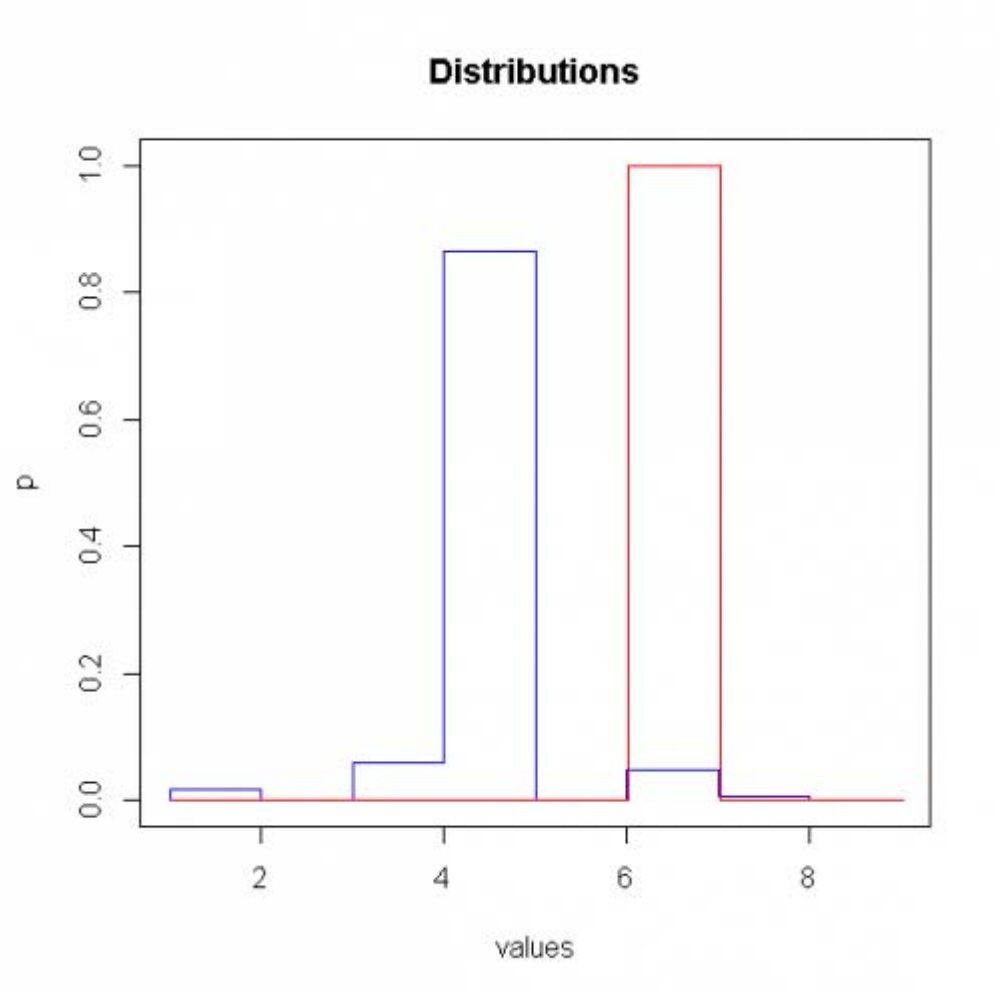
```
1 L1
 NumPassen       type rpm_maxTor     height   displace minFuelCon     weight
 0.9510749  0.8784285  0.8724981  0.8472943  0.8465530  0.8421053  0.8376575
2 L2
     type NumPassen    height wheelbase    weight   maxLoad     width
0.9329496 0.9225715 0.8276864 0.7862469 0.7004026 0.6820593 0.5931772
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
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
>
```

Now we can select from the ***specificity*** table S interesting variables for selected cluster and using the function `specific` try to provide its characteristics.

1 / NumPassen

specificity = 0.9510749

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

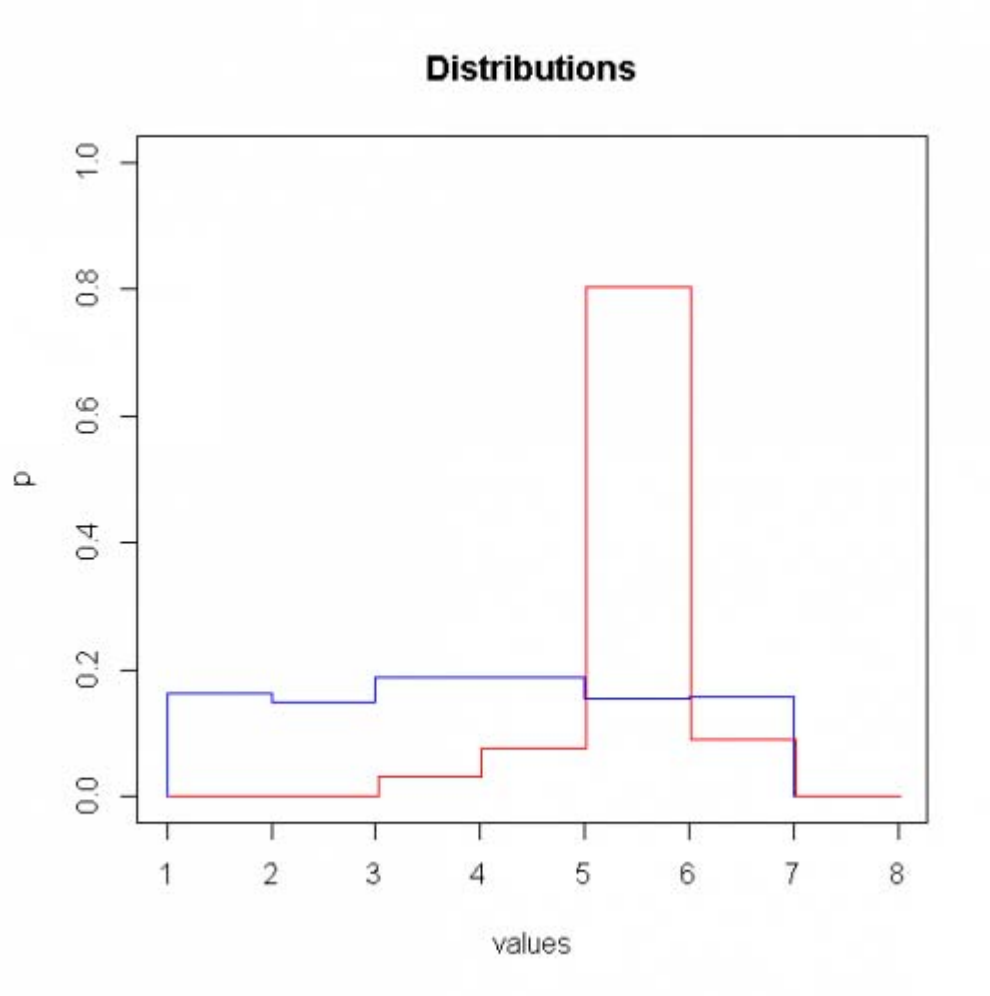


All cars in cluster 1 have the value NumPassen=7 .

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
>
```

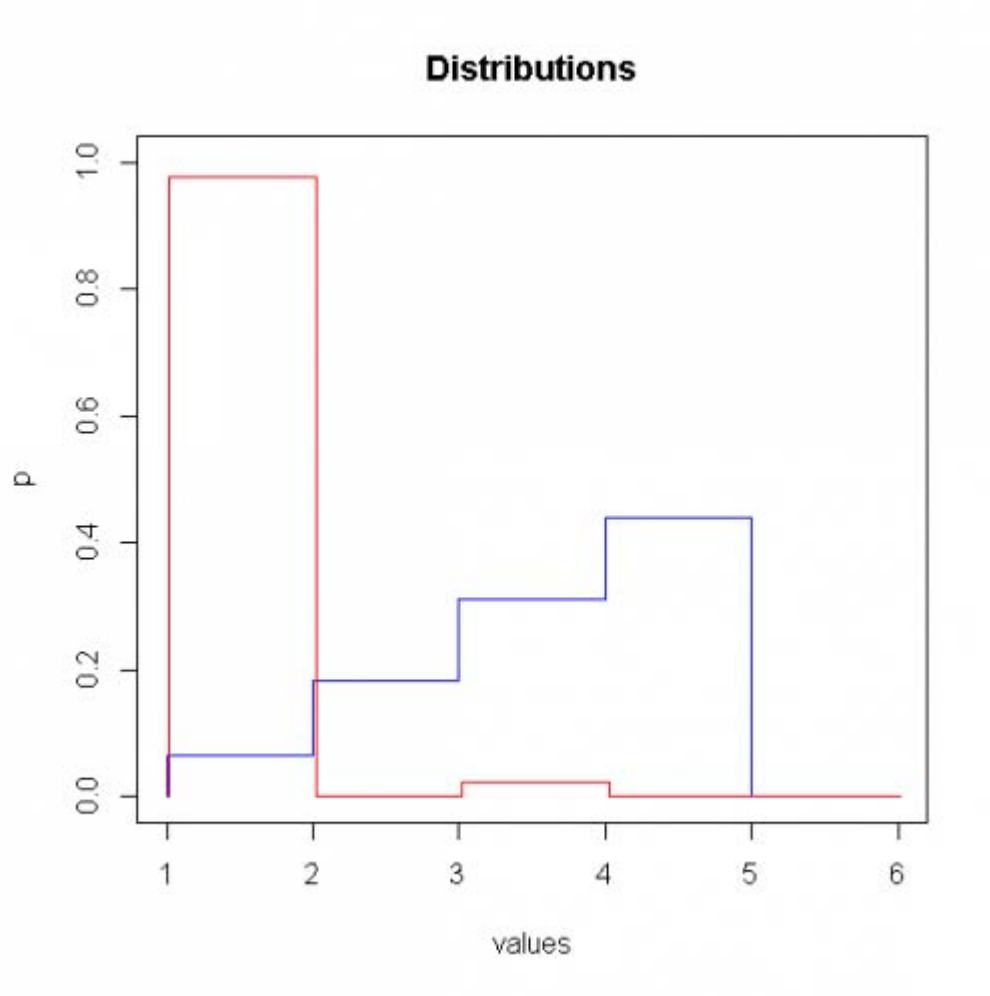


Most of the cars in cluster 10 have the maxSpeed in the interval (200,215].
No car in this cluster has maxSpeed in the interval [130,174].

14 / NumDoors

specificity = 0.9111175

```
> specific(14,'NumDoors')
         2          3          4          5         NA
0.06449222 0.18383988 0.31208302 0.43958488 0.00000000
         2          3          4          5         NA
0.97560976 0.00000000 0.02439024 0.00000000 0.00000000
       2        3        4        5       NA
15.12756      Inf 12.79540      Inf      NaN
```

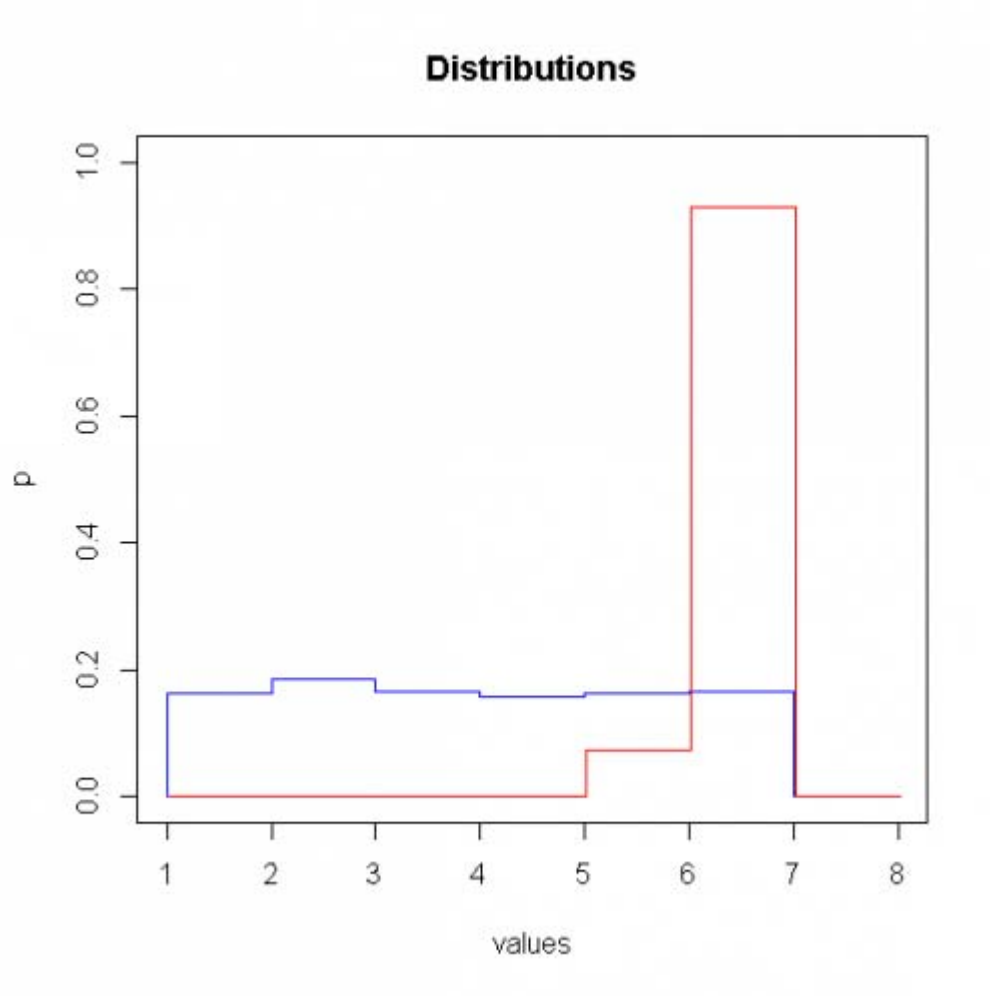


Most of the cars in cluster 14 have NumDoors=2.
A tiny part of them have also NumDoors=4.

19 / length

specificity = 0.7607460

```
> specific(19,'length')
[2600,4010] (4010,4245] (4245,4470] (4470,4555] (4555,4761] (4761,6000]          NA
  0.1616012   0.1845812   0.1645663   0.1586360   0.1638251   0.1667902   0.0000000
[2600,4010] (4010,4245] (4245,4470] (4470,4555] (4555,4761] (4761,6000]          NA
 0.00000000  0.00000000  0.00000000  0.00000000  0.07246377  0.92753623  0.00000000
[2600,4010] (4010,4245] (4245,4470] (4470,4555] (4555,4761] (4761,6000]          NA
        Inf         Inf         Inf         Inf    2.260786    5.561095         NaN
```

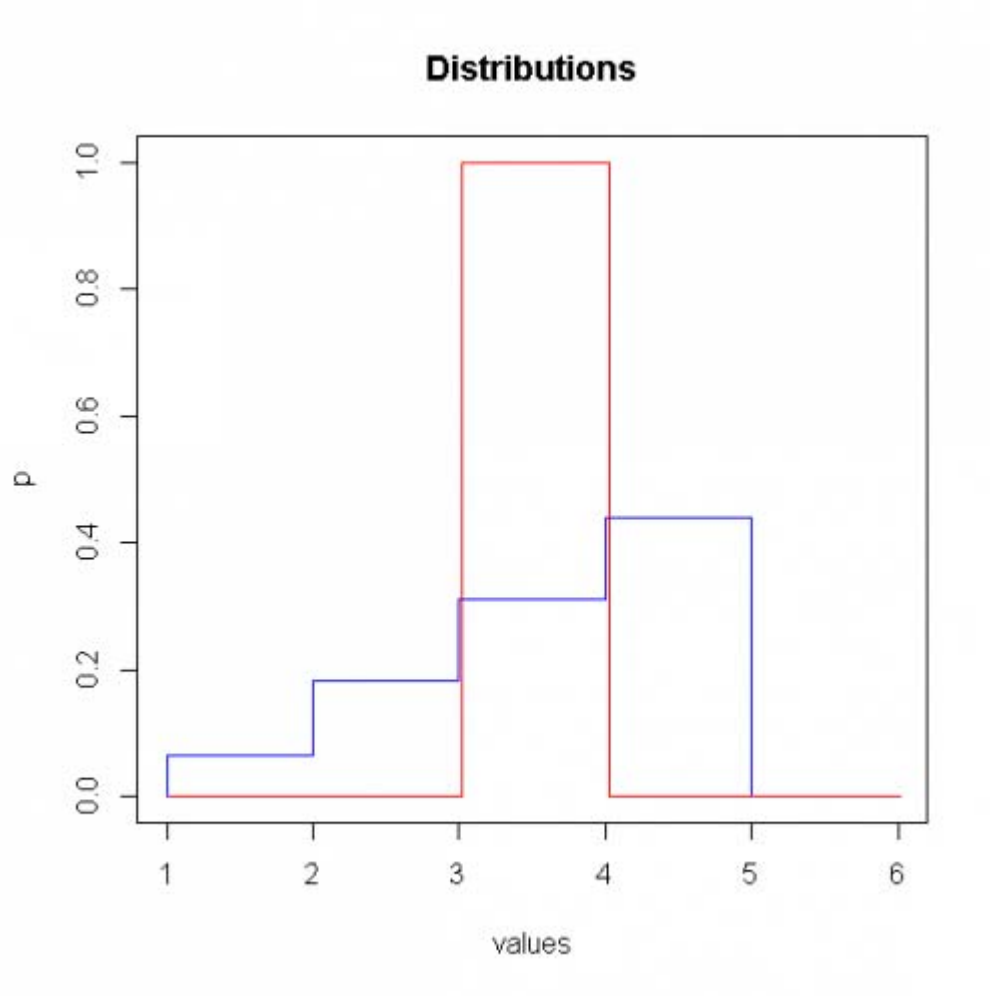


All cars from cluster 19 have length in the interval (4555,6000].
Most of them in the interval (4761,6000].

19 / NumDoors

specificity = 0.6879170

```
> specific(19,'NumDoors')
         2          3          4          5         NA
0.06449222 0.18383988 0.31208302 0.43958488 0.00000000
 2  3  4  5 NA
 0  0  1  0  0
       2         3         4         5        NA
     Inf       Inf 3.204276       Inf       NaN
```

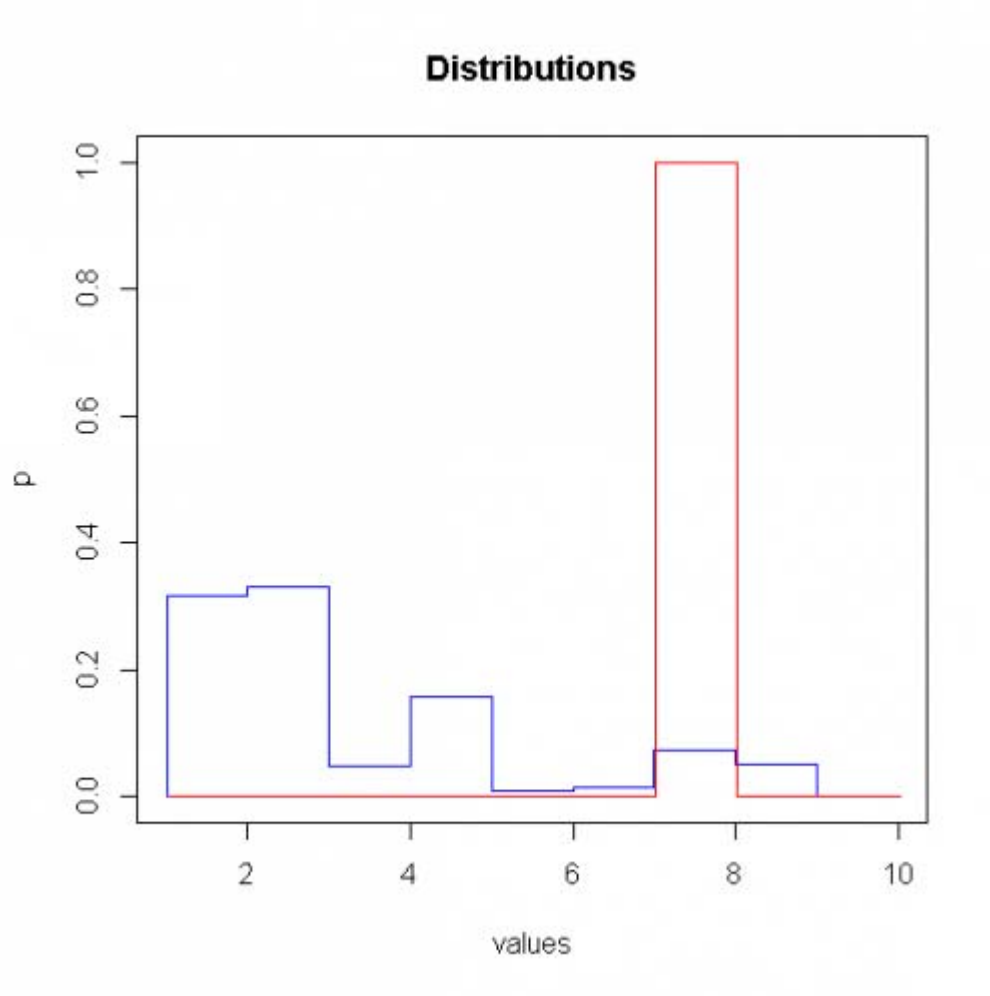


All cars from cluster 19 have NumDoors=4.

22 / type

specificity = 0.9258710

```
> specific(22,'type')
         LI          KL          EN          KA          KB          RO          TE
0.315789474 0.330615271 0.047442550 0.157894737 0.008154188 0.014084507 0.074128984
         KU        <NA>
0.051890289 0.000000000
  LI   KL   EN   KA   KB   RO   TE   KU <NA>
   0    0    0    0    0    0    1    0    0
   LI    KL    EN    KA    KB    RO    TE    KU  <NA>
  Inf   Inf   Inf   Inf   Inf   Inf 13.49   Inf   NaN
```



All cars in the cluster 22 have type=TE.

notes/clamix.txt · Last modified: 2012/11/02 02:25 by batagelj