outlier detection

```
setwd('C:/Users/Elena/Desktop/Elena/Polimi/MAGISTRALE/Nonparametric statistics/Progetto/github reposito
dataset_xsectional <- read.csv("oasis_cross-sectional.csv", header = T)
dataset_longitudinal <- read.csv("oasis_longitudinal.csv", header = T)
```

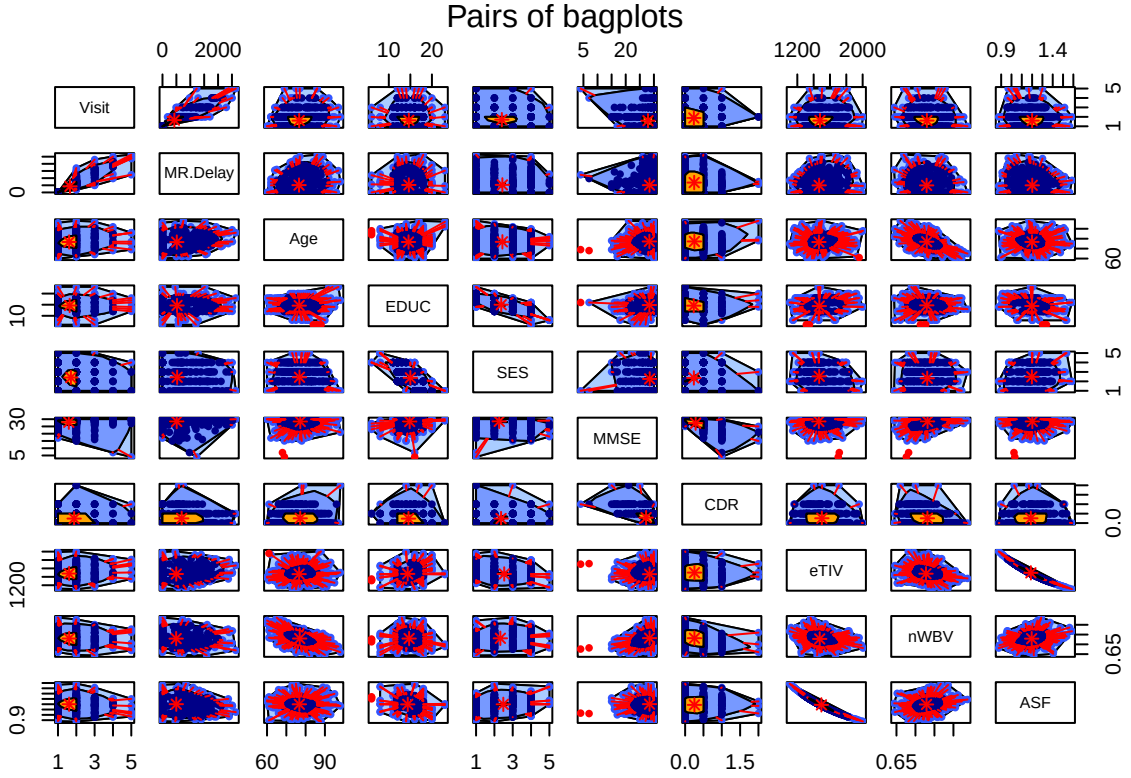
```
library(DepthProc)
```

```
## Warning: il pacchetto 'DepthProc' è stato creato con R versione 4.1.3
## Caricamento del pacchetto richiesto: ggplot2
## Caricamento del pacchetto richiesto: Rcpp
## Warning: il pacchetto 'Rcpp' è stato creato con R versione 4.1.3
## Caricamento del pacchetto richiesto: rrcov
## Warning: il pacchetto 'rrcov' è stato creato con R versione 4.1.3
## Caricamento del pacchetto richiesto: robustbase
## Warning: il pacchetto 'robustbase' è stato creato con R versione 4.1.3
## Scalable Robust Estimators with High Breakdown Point (version 1.6-2)
## Caricamento del pacchetto richiesto: MASS
## Caricamento del pacchetto richiesto: np
## Nonparametric Kernel Methods for Mixed Datatypes (version 0.60-11)
## [vignette("np_faq",package="np") provides answers to frequently asked questions]
## [vignette("np",package="np") an overview]
## [vignette("entropy_np",package="np") an overview of entropy-based methods]
## Warning in .recacheSubclasses(def@className, def, env): undefined subclass
## "packedMatrix" of class "replValueSp"; definition not updated
##
## Caricamento pacchetto: 'DepthProc'
## Il seguente oggetto è mascherato da 'package:base':
##
##     as.matrix
```

```
library(aplpack)
```

To visualize the outliers in this context we retort to a bagplot matrix:

```
bagplot_matrix <- aplpack::bagplot.pairs(dataset_longitudinal, main = 'Pairs of bagplots')  # bagplot d
```



Using this bagplot we want to try to understan which comparisons are reasonable to find outliers using the depth measures.

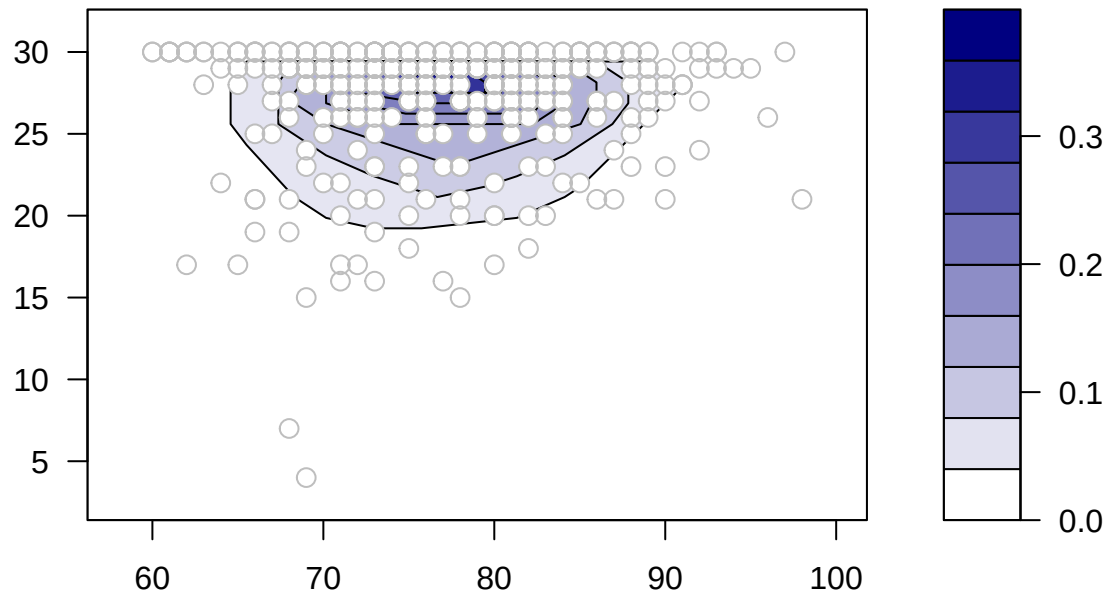
Indeed the outlier detection becomes more difficult when the dimension increases, as

- some outliers may be wrongly flagged as genuine points
- some good points may be wrongly flagged as outliers

These phenomena are respectively denoted as **masking** and **swamping**, we will cover it in details during the robust statistics analysis

```
# demented <- dataset_longitudinal[which(dataset_longitudinal$Group == 'Demented'),]
# nondemented <- dataset_longitudinal[which(dataset_longitudinal$Group == 'Nondemented'),]
#
# ddPlot(x = demented ,y=nondemented,depth_params = list(method='Tukey'))
```

```
depthContour(
  data.frame(dataset_longitudinal$Age, dataset_longitudinal$MMSE),
  depth_params = list(method = 'Tukey'),
  points = TRUE,
  colors = colorRampPalette(c('white', 'navy')),
  levels = 10,
  pdmedian = F,
  graph_params = list(cex=.01, pch=1),
  pmean = F
)
```

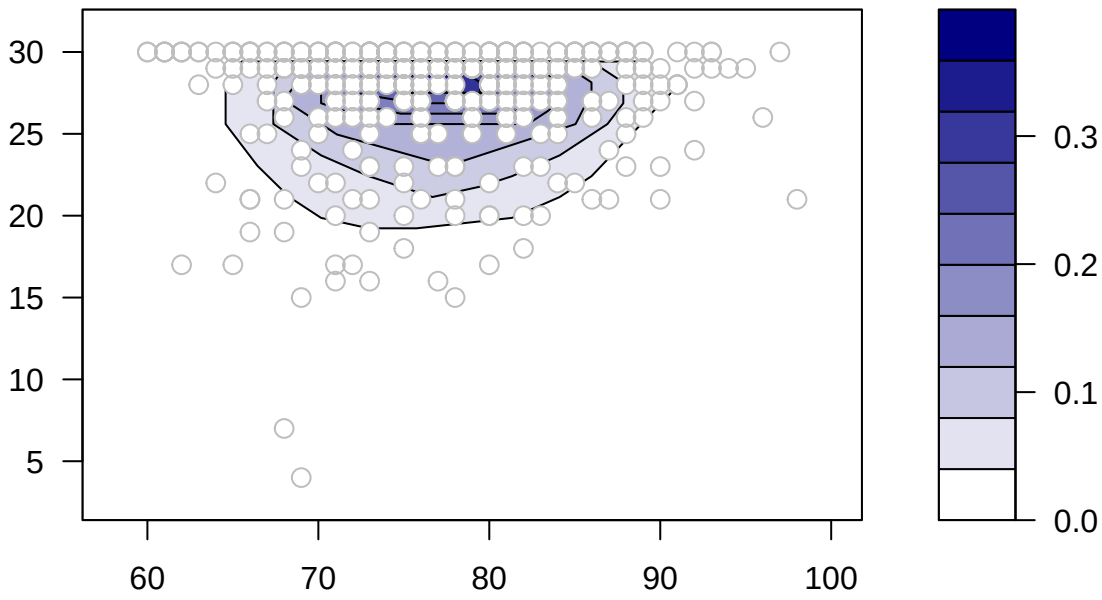



```
#{r} depthContour(   data.frame(dataset_longitudinal$Age, dataset_longitudinal$MMSE),
depth_params = list(method = 'spatial'),   points = TRUE,   colors = colorRampPalette(c('white',
'navy')),   levels = 10,   pdmedian = F,   graph_params = list(cex=.01, pch=1),   pmean =
F ) #
```

```
#{r} depthContour(   as.matrix(cbind(dataset_longitudinal$Age, dataset_longitudinal$MMSE)),
depth_params = list(method = 'LP'),   points = TRUE,   colors = colorRampPalette(c('white',
'navy')),   levels = 10,   pdmedian = F,   graph_params = list(cex=.01, pch=1),   pmean =
F ) #
```

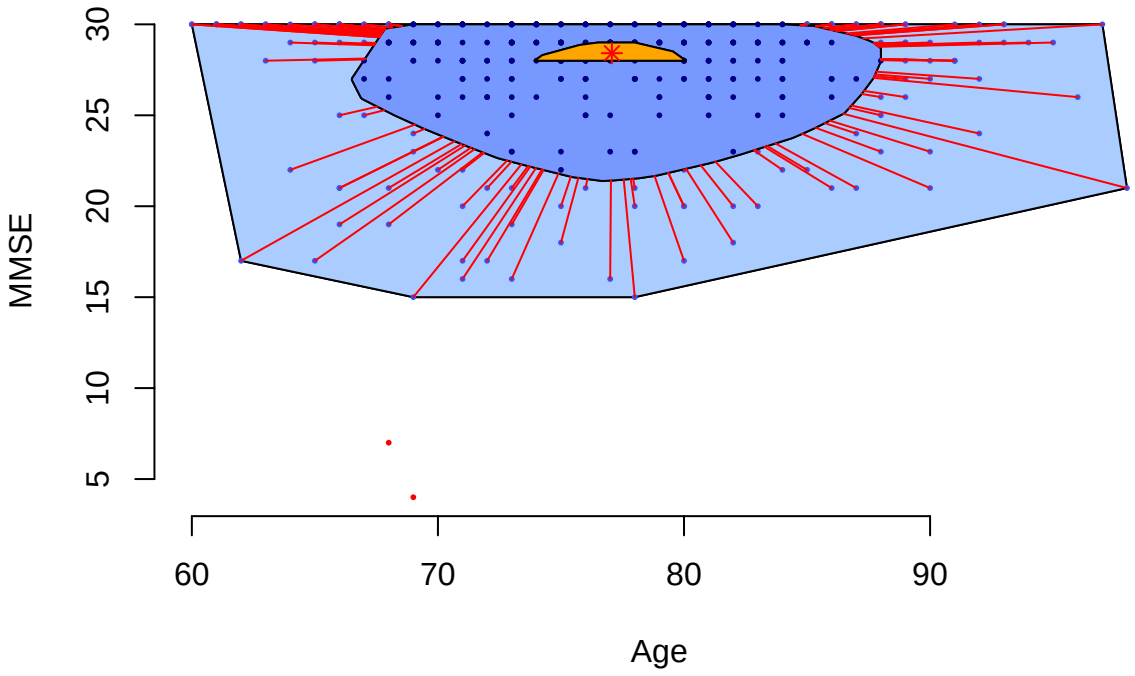
```
#{r} depthContour(   cbind(dataset_longitudinal$Age, dataset_longitudinal$MMSE),   depth_params
= list(method = 'Mahalanobis'),   points = TRUE,   colors = colorRampPalette(c('white',
'navy')),   levels = 10,   pdmedian = F,   graph_params = list(cex=.01, pch=1),   pmean =
F ) #
```

```r
depthContour(
  data.frame(dataset_longitudinal$Age, dataset_longitudinal$MMSE),
  depth_params = list(method = 'Tukey'),
  points = TRUE,
  colors = colorRampPalette(c('white', 'navy')),
  levels = 10,
  pdmedian = F,
  graph_params = list(cex=.01, pch=1),
  pmean = F
)
```




```
bags1 <- bagplot(data.frame(dataset_longitudinal$Age, dataset_longitudinal$MMSE), xlab = "Age", ylab =
```

```
## [1] "Warning: NA elements have been exchanged by median values!!"
```




```
outlying_obs1 <- bags1$pxy.outlier

outlying_obs1
```

```
##       x y
## [1,] 68 7
## [2,] 69 4
```

```
which(dataset_longitudinal$Age==outlying_obs1[,1] & dataset_longitudinal$MMSE==outlying_obs1[,2])
```

```
## Warning in dataset_longitudinal$Age == outlying_obs1[, 1]: la lunghezza più
## lunga dell'oggetto non è un multiplo della lunghezza più corta dell'oggetto
```

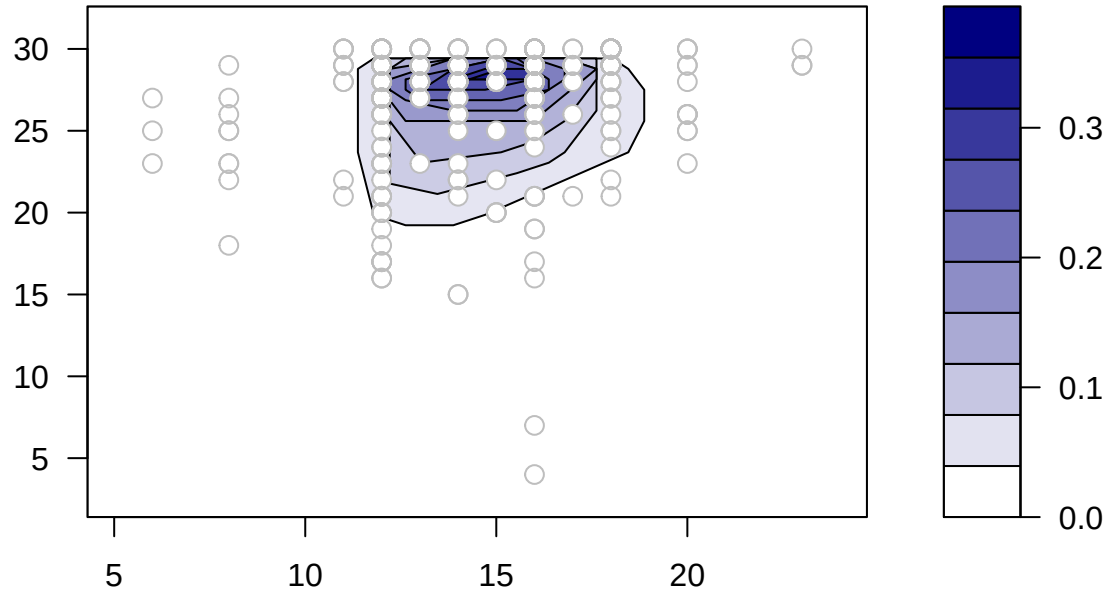
```
## Warning in dataset_longitudinal$MMSE == outlying_obs1[, 2]: la lunghezza più
## lunga dell'oggetto non è un multiplo della lunghezza più corta dell'oggetto
```

```
## [1] 101 102
```

```
#{r} depthContour(   data.frame(dataset_longitudinal$EDUC, dataset_longitudinal$MMSE),
depth_params = list(method = 'Mahalanobis'),   points = TRUE,   colors = colorRampPalette(c('white',
'navy')),   levels = 10,   pdmedian = F,   graph_params = list(cex=.01, pch=1),   pmean =
F ) #
```

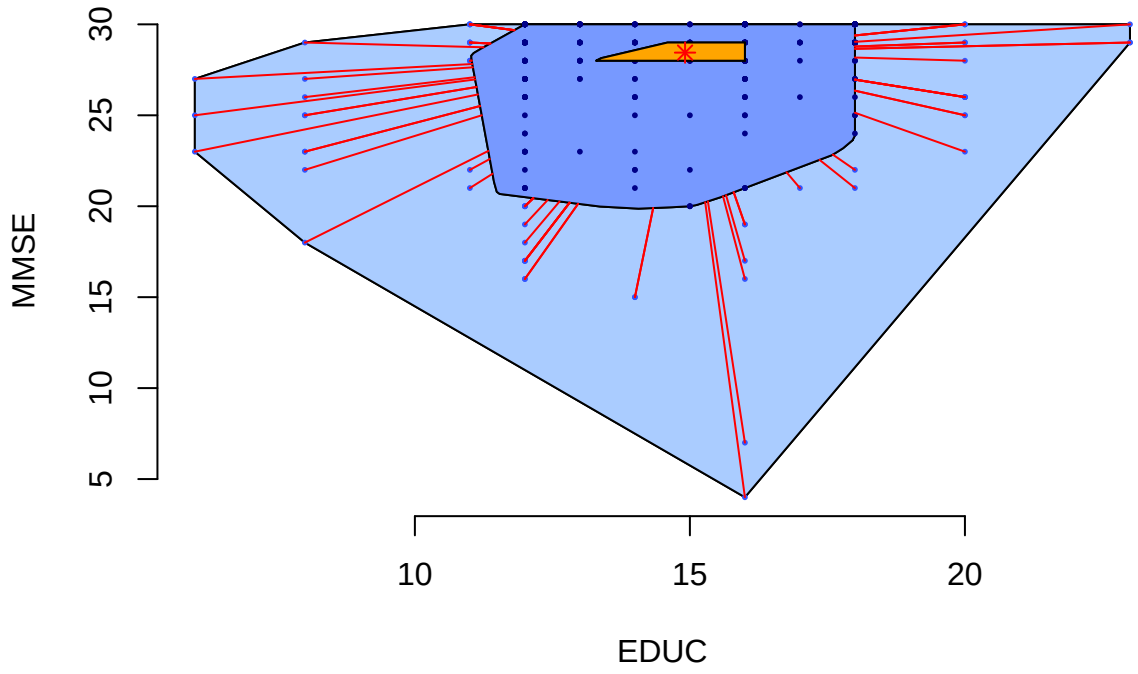
```
depthContour(
  data.frame(dataset_longitudinal$EDUC, dataset_longitudinal$MMSE),
  depth_params = list(method = 'Tukey'),
  points = TRUE,
  colors = colorRampPalette(c('white', 'navy')),
  levels = 10,
```

```
  pdmedian = F,
  graph_params = list(cex=.01, pch=1),
  pmean = F
)
```




```
bags2 <- bagplot(  data.frame(dataset_longitudinal$EDUC, dataset_longitudinal$MMSE), xlab = "EDUC", ylab
```

```
## [1] "Warning: NA elements have been exchanged by median values!!"
```



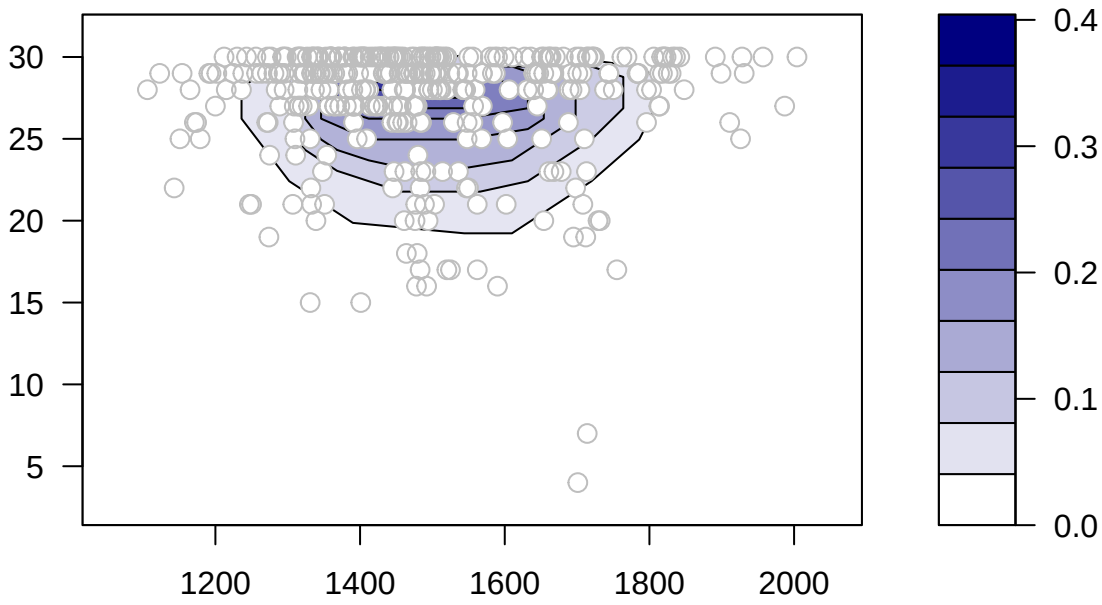

```
outlying_obs2 <- bags2$pxy.outlier

outlying_obs2
```

```
## NULL
```

```
which(dataset_longitudinal$EDUC==outlying_obs2[,1] & dataset_longitudinal$MMSE==outlying_obs2[,2])
```

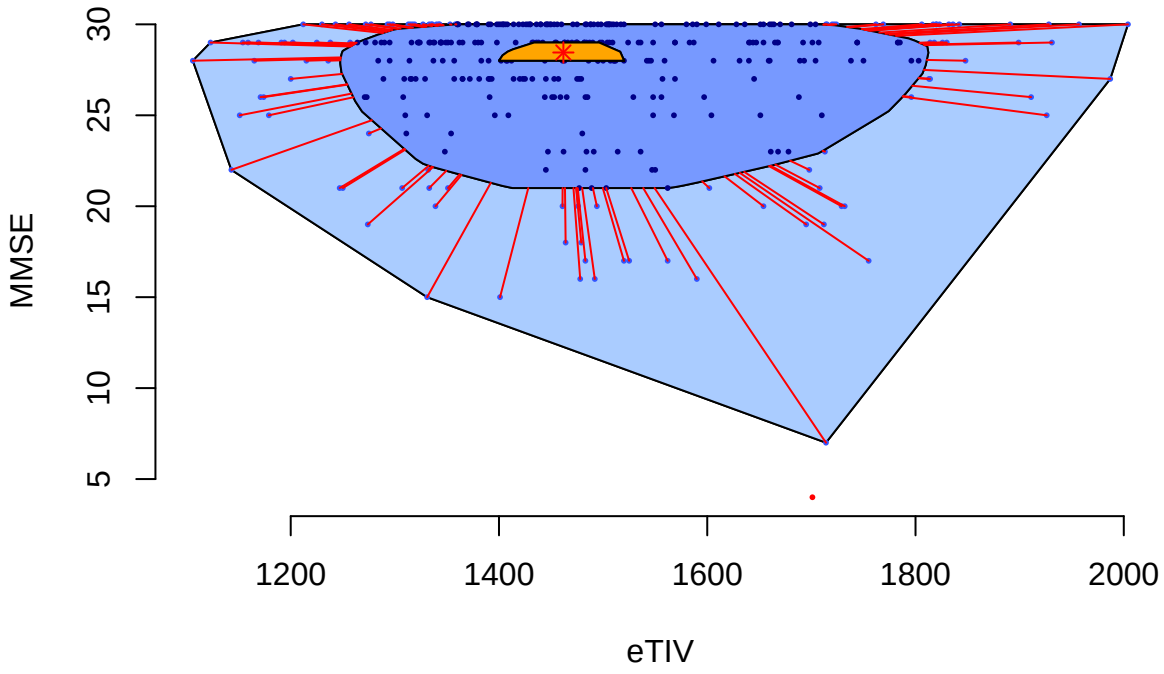
```
## integer(0)
```

```
depthContour(
  data.frame(dataset_longitudinal$eTIV, dataset_longitudinal$MMSE),
  depth_params = list(method = 'Tukey'),
  points = TRUE,
  colors = colorRampPalette(c('white', 'navy')),
  levels = 10,
  pdmedian = F,
  graph_params = list(cex=.01, pch=1),
  pmean = F
)
```




```
bags3 <- bagplot(data.frame(dataset_longitudinal$eTIV, dataset_longitudinal$MMSE), xlab = "eTIV", ylab =
```

```
## [1] "Warning: NA elements have been exchanged by median values!!"
```



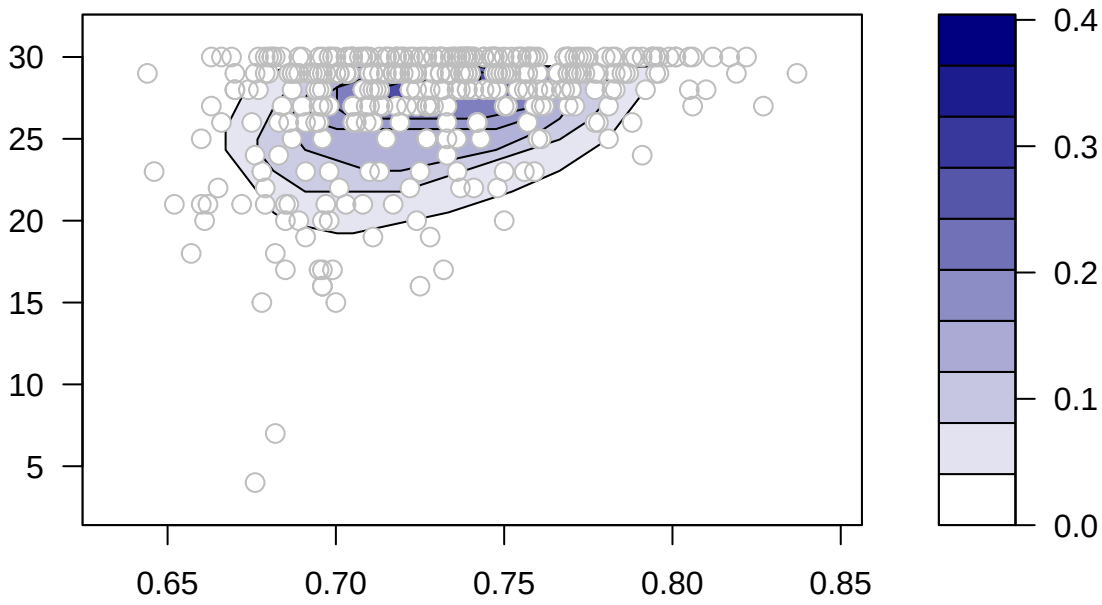

```
outlying_obs3 <- bags3$pxy.outlier

outlying_obs3
```

```
##         x y
## [1,] 1701 4
```

```
which(dataset_longitudinal$eTIV==outlying_obs3[,1] & dataset_longitudinal$MMSE==outlying_obs3[,2])
```

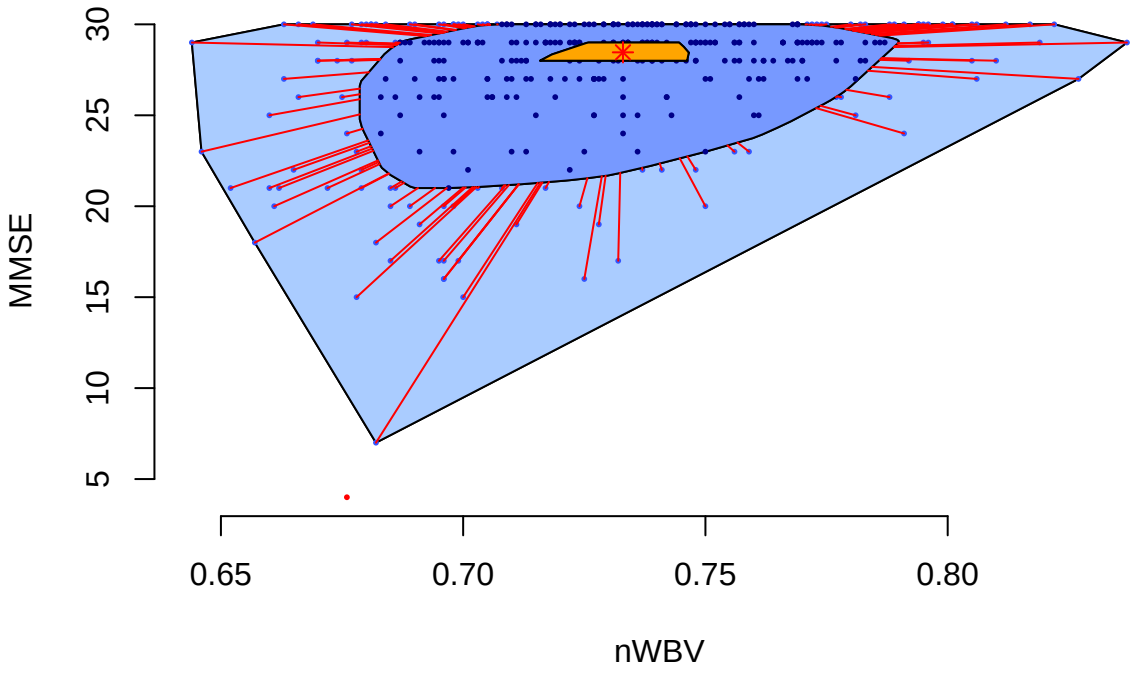
```
## [1] 102
```

```
depthContour(
  data.frame(dataset_longitudinal$nWBV, dataset_longitudinal$MMSE),
  depth_params = list(method = 'Tukey'),
  points = TRUE,
  colors = colorRampPalette(c('white', 'navy')),
  levels = 10,
  pdmedian = F,
  graph_params = list(cex=.01, pch=1),
  pmean = F
)
```




```
bags4 <- bagplot(data.frame(dataset_longitudinal$nWBV, dataset_longitudinal$MMSE), xlab = "nWBV", ylab =
```

```
## [1] "Warning: NA elements have been exchanged by median values!!"
```



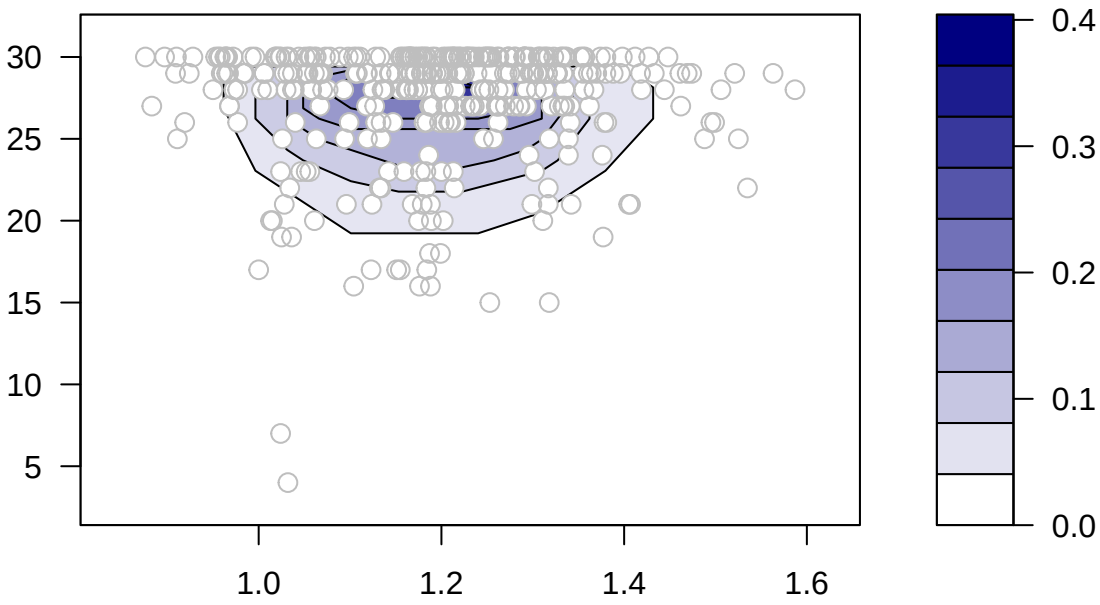

```
outlying_obs4 <- bags4$pxy.outlier

outlying_obs4
```

```
##          x y
## [1,] 0.676 4
```

```
which(dataset_longitudinal$nWBV==outlying_obs4[,1] & dataset_longitudinal$MMSE==outlying_obs4[,2])
```

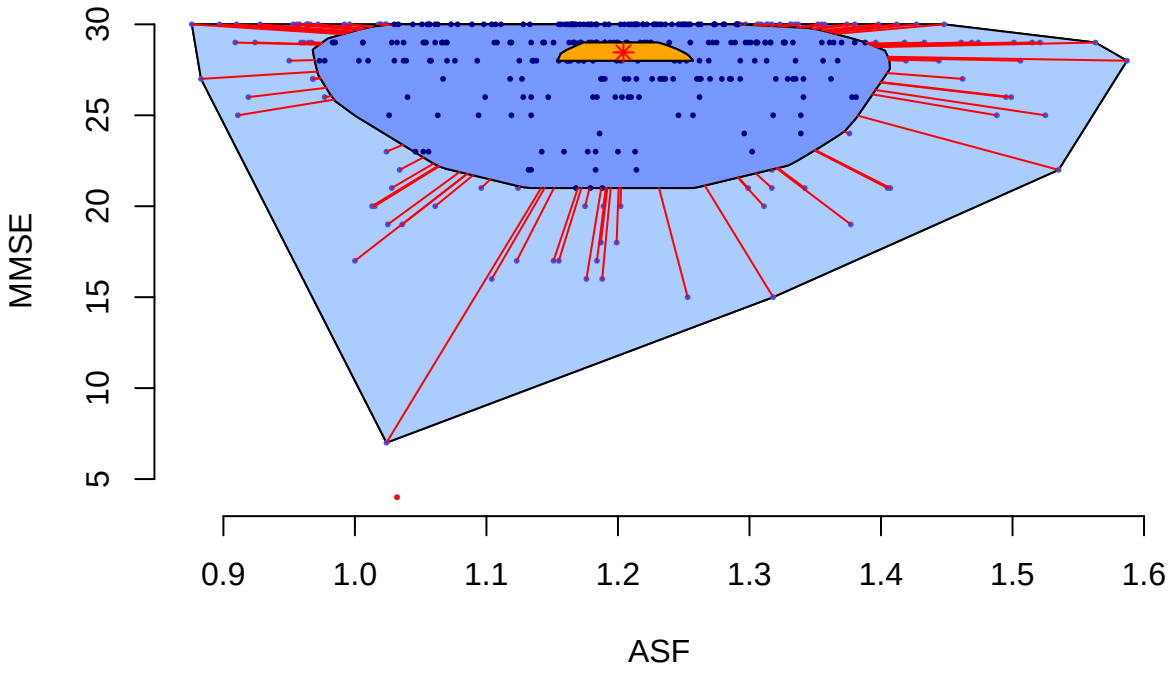
```
## [1] 102
```

```
depthContour(
  data.frame(dataset_longitudinal$ASF, dataset_longitudinal$MMSE),
  depth_params = list(method = 'Tukey'),
  points = TRUE,
  colors = colorRampPalette(c('white', 'navy')),
  levels = 10,
  pdmedian = F,
  graph_params = list(cex=.01, pch=1),
  pmean = F
)
```




```
bags5 <- bagplot(data.frame(dataset_longitudinal$ASF, dataset_longitudinal$MMSE), xlab = "ASF", ylab =
```

```
## [1] "Warning: NA elements have been exchanged by median values!!"
```



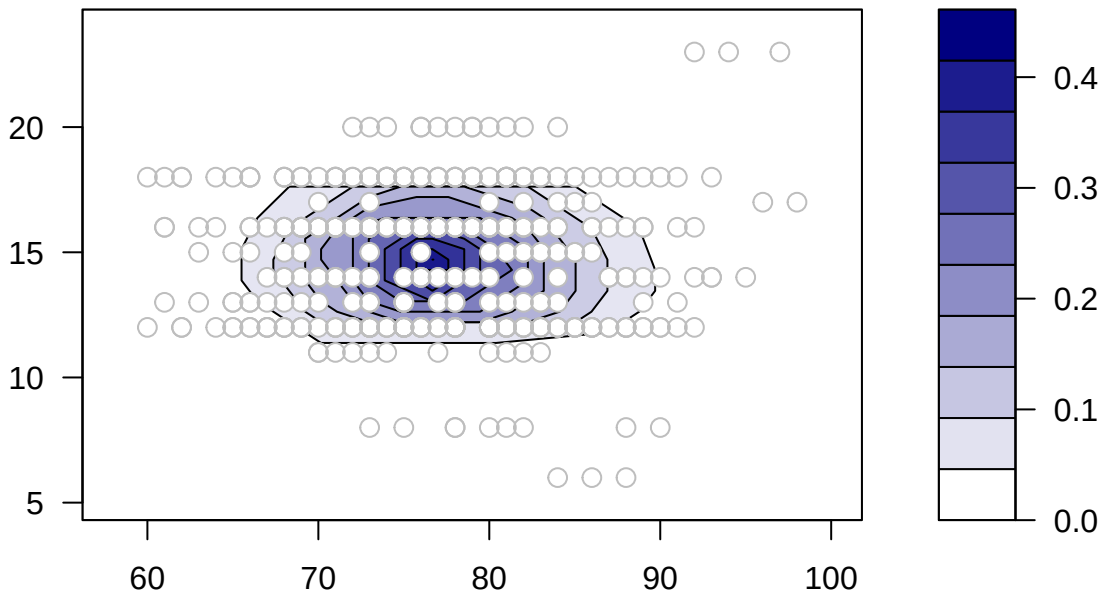

```
outlying_obs5 <- bags5$pxy.outlier

outlying_obs5
```

```
##          x y
## [1,] 1.032 4
```

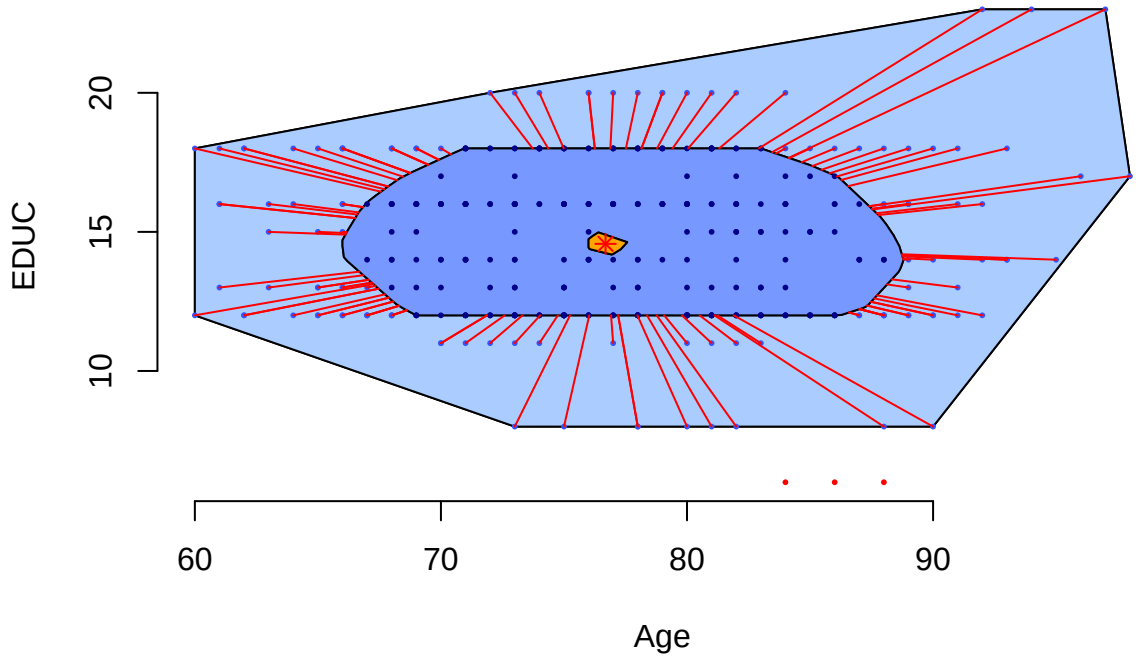
```
which(dataset_longitudinal$ASF==outlying_obs5[,1] & dataset_longitudinal$MMSE==outlying_obs5[,2])
```

```
## [1] 102
```

```
depthContour(
  data.frame(dataset_longitudinal$Age,dataset_longitudinal$EDUC ),
  depth_params = list(method = 'Tukey'),
  points = TRUE,
  colors = colorRampPalette(c('white', 'navy')),
  levels = 10,
  pdmedian = F,
  graph_params = list(cex=.01, pch=1),
  pmean = F
)
```




```
bags6 <- bagplot(data.frame(dataset_longitudinal$Age,dataset_longitudinal$EDUC ), xlab = "Age", ylab =
```




```
outlying_obs6 <- bags6$pxy.outlier

outlying_obs6
```

```
##       x y
## [1,] 84 6
## [2,] 86 6
## [3,] 88 6
```

```
which(dataset_longitudinal$Age==outlying_obs6[,1] & dataset_longitudinal$EDUC==outlying_obs6[,2])
```

```
## Warning in dataset_longitudinal$Age == outlying_obs6[, 1]: la lunghezza più
## lunga dell'oggetto non è un multiplo della lunghezza più corta dell'oggetto
```

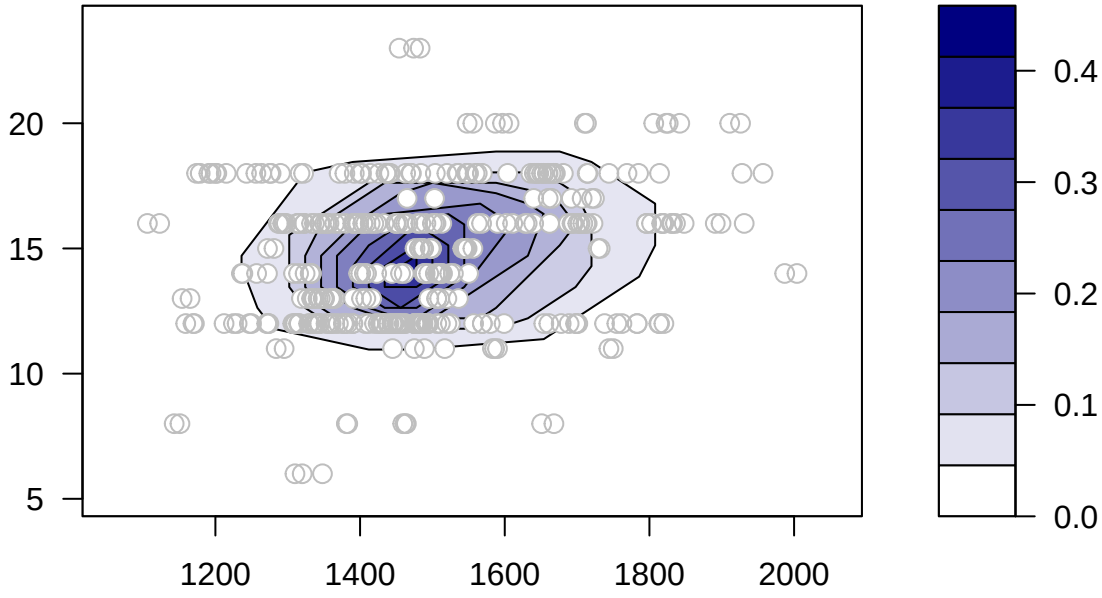
```
## Warning in dataset_longitudinal$EDUC == outlying_obs6[, 2]: la lunghezza più
## lunga dell'oggetto non è un multiplo della lunghezza più corta dell'oggetto
```

```
## [1] 79 80 81
```

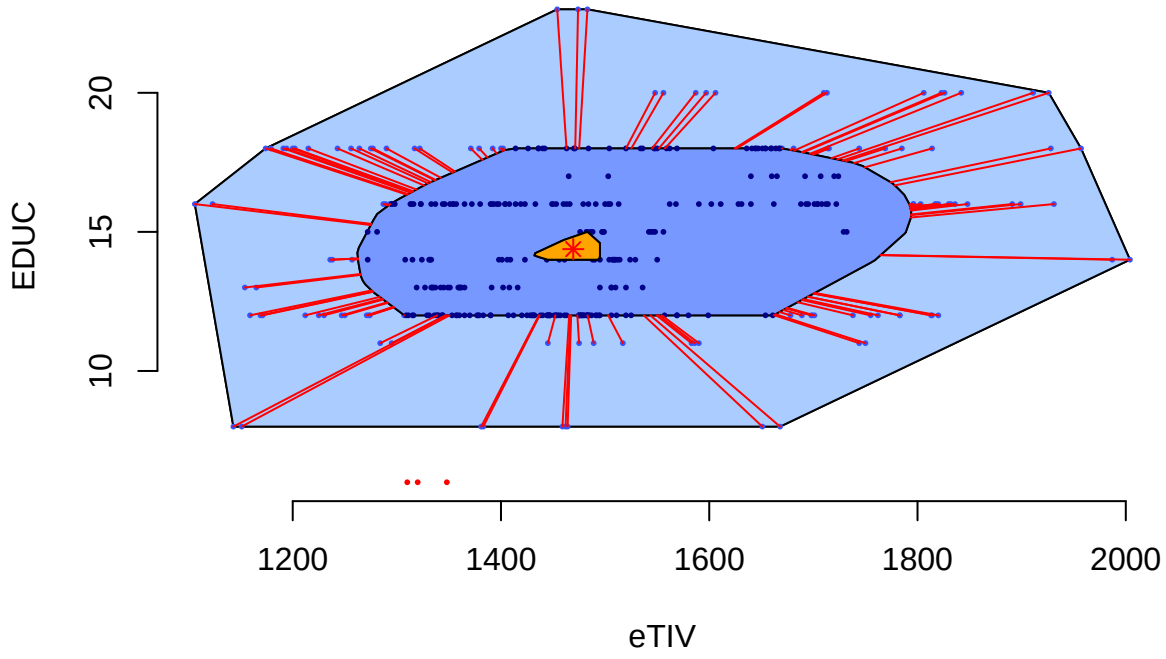
We can notice that these three persons have an eduction really lower wth respect to the others.

```
depthContour(
  data.frame(dataset_longitudinal$eTIV, dataset_longitudinal$EDUC),
  depth_params = list(method = 'Tukey'),
  points = TRUE,
  colors = colorRampPalette(c('white', 'navy')),
  levels = 10,
  pdmedian = F,
  graph_params = list(cex=.01, pch=1),
```

```
  pmean = F
)
```




```
bags7 <- bagplot(data.frame(dataset_longitudinal$eTIV, dataset_longitudinal$EDUC), xlab = "eTIV", ylab =
```




```
outlying_obs7 <- bags7$pxy.outlier

outlying_obs7
```

```
##          x y
## [1,] 1310 6
## [2,] 1320 6
## [3,] 1348 6
```

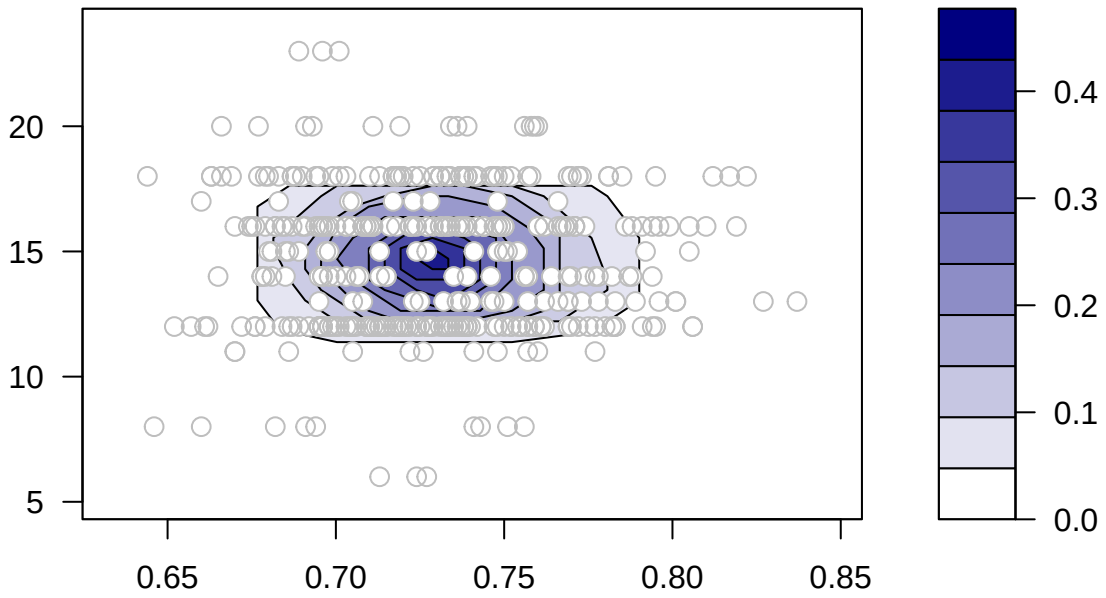
```
which(dataset_longitudinal$eTIV==outlying_obs7[,1] & dataset_longitudinal$EDUC==outlying_obs7[,2])
```

```
## Warning in dataset_longitudinal$eTIV == outlying_obs7[, 1]: la lunghezza più
## lunga dell'oggetto non è un multiplo della lunghezza più corta dell'oggetto
```

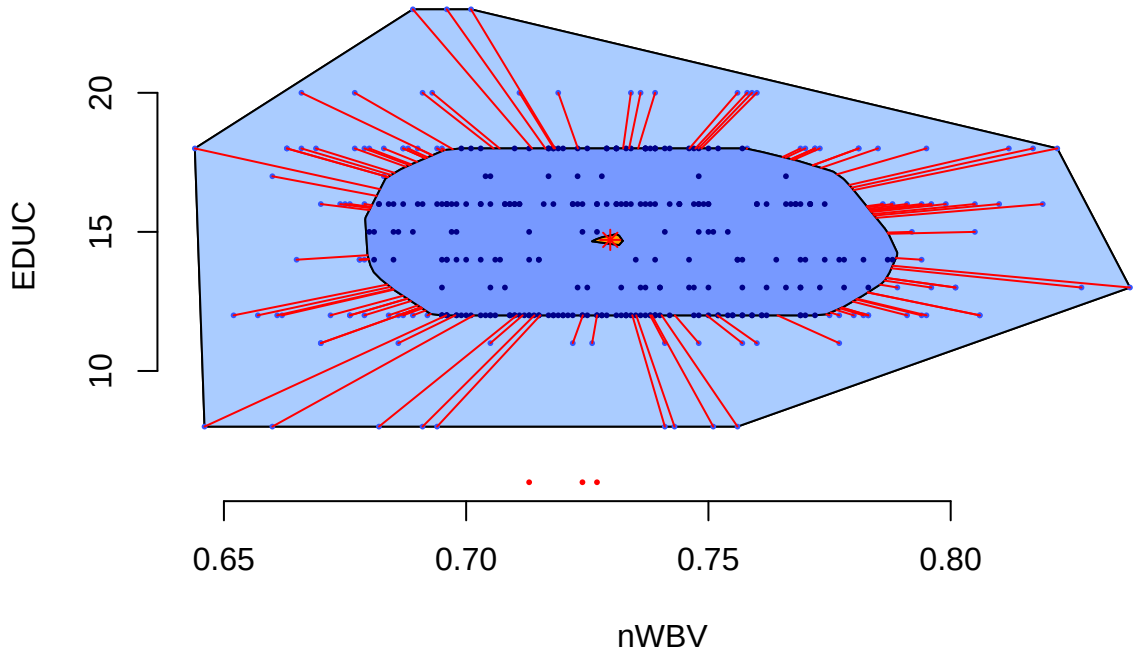
```
## Warning in dataset_longitudinal$EDUC == outlying_obs7[, 2]: la lunghezza più
## lunga dell'oggetto non è un multiplo della lunghezza più corta dell'oggetto
```

```
## [1] 79 80 81
```

```
depthContour(
  data.frame(dataset_longitudinal$nWBV, dataset_longitudinal$EDUC),
  depth_params = list(method = 'Tukey'),
  points = TRUE,
  colors = colorRampPalette(c('white', 'navy')),
  levels = 10,
  pdmedian = F,
  graph_params = list(cex=.01, pch=1),
  pmean = F
)
```




```
bags8 <- bagplot(data.frame(dataset_longitudinal$nWBV, dataset_longitudinal$EDUC), xlab = "nWBV", ylab =
```




```
outlying_obs8 <- bags8$pxy.outlier

outlying_obs8
```

```
##          x y
## [1,] 0.727 6
## [2,] 0.724 6
## [3,] 0.713 6
```

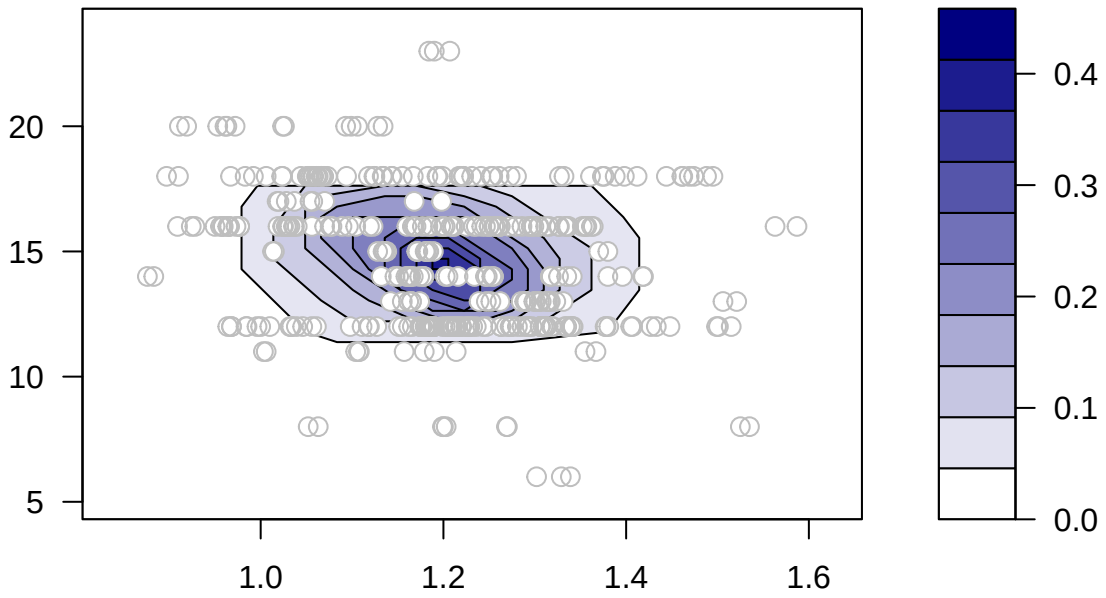
```
which(dataset_longitudinal$nWBV==outlying_obs8[,1] & dataset_longitudinal$EDUC==outlying_obs8[,2])
```

```
## Warning in dataset_longitudinal$nWBV == outlying_obs8[, 1]: la lunghezza più
## lunga dell'oggetto non è un multiplo della lunghezza più corta dell'oggetto
```

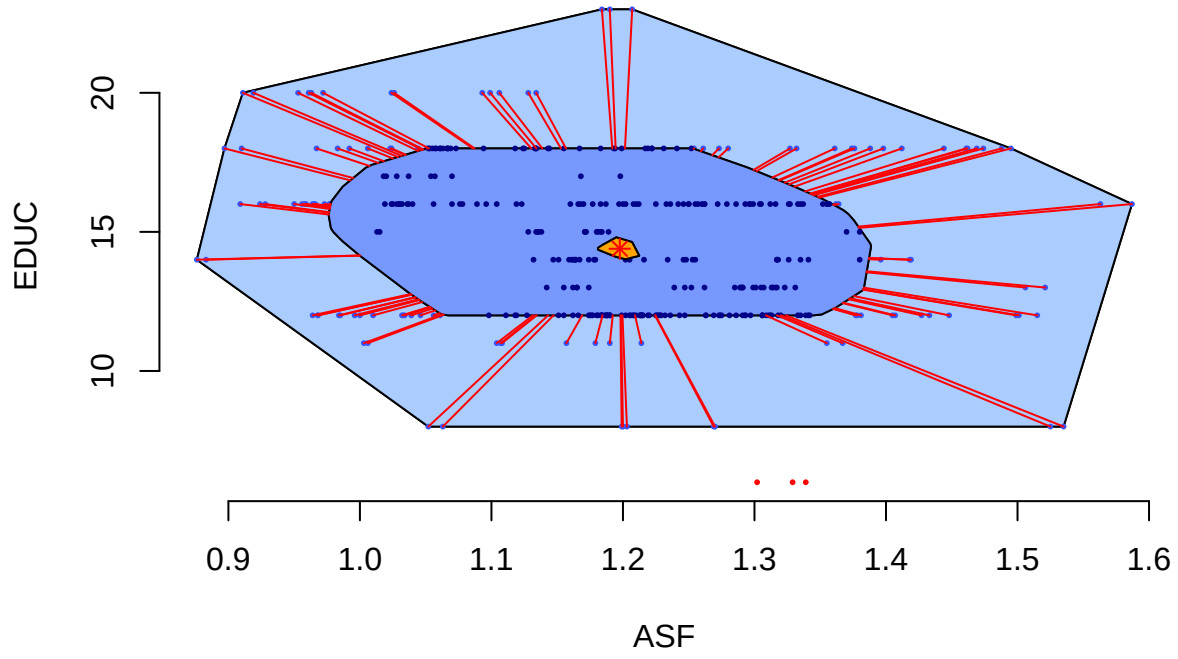
```
## Warning in dataset_longitudinal$EDUC == outlying_obs8[, 2]: la lunghezza più
## lunga dell'oggetto non è un multiplo della lunghezza più corta dell'oggetto
```

```
## [1] 79 80 81
```

```
depthContour(
  data.frame(dataset_longitudinal$ASF, dataset_longitudinal$EDUC),
  depth_params = list(method = 'Tukey'),
  points = TRUE,
  colors = colorRampPalette(c('white', 'navy')),
  levels = 10,
  pdmedian = F,
  graph_params = list(cex=.01, pch=1),
  pmean = F
)
```




```
bags9 <- bagplot(data.frame(dataset_longitudinal$ASF, dataset_longitudinal$EDUC), xlab = "ASF", ylab =
```




```
outlying_obs9 <- bags9$pxy.outlier

outlying_obs9
```

```
##          x y
## [1,] 1.339 6
## [2,] 1.329 6
## [3,] 1.302 6
```

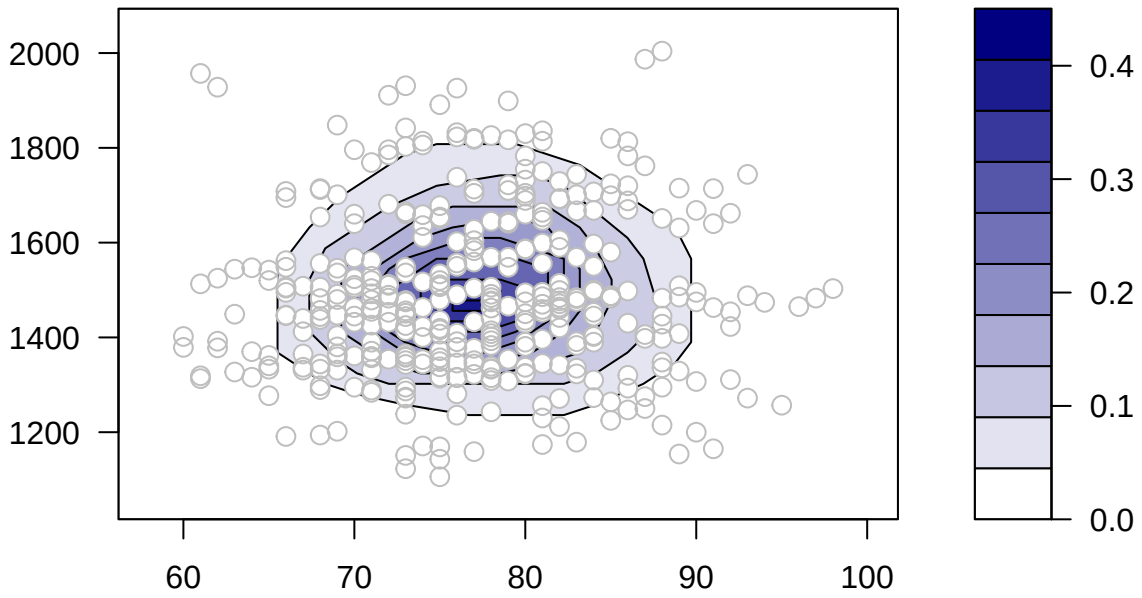
```
which(dataset_longitudinal$ASF==outlying_obs9[,1] & dataset_longitudinal$EDUC==outlying_obs9[,2])
```

```
## Warning in dataset_longitudinal$ASF == outlying_obs9[, 1]: la lunghezza più
## lunga dell'oggetto non è un multiplo della lunghezza più corta dell'oggetto
```

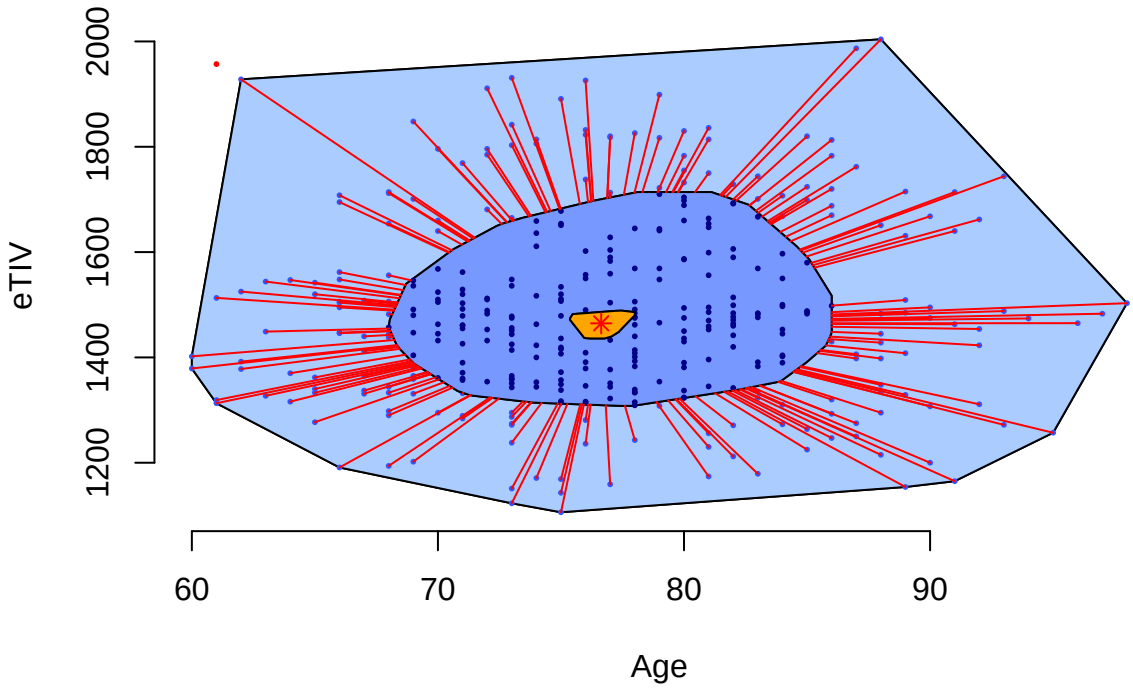
```
## Warning in dataset_longitudinal$EDUC == outlying_obs9[, 2]: la lunghezza più
## lunga dell'oggetto non è un multiplo della lunghezza più corta dell'oggetto
```

```
## [1] 79 80 81
```

```
depthContour(
  data.frame(dataset_longitudinal$Age, dataset_longitudinal$eTIV),
  depth_params = list(method = 'Tukey'),
  points = TRUE,
  colors = colorRampPalette(c('white', 'navy')),
  levels = 10,
  pdmedian = F,
  graph_params = list(cex=.01, pch=1),
  pmean = F
)
```




```
bags10 <- bagplot(data.frame(dataset_longitudinal$Age, dataset_longitudinal$eTIV), xlab = "Age", ylab =
```




```
outlying_obs10 <- bags10$pxy.outlier

outlying_obs10
```

```
##       x    y
## [1,] 61 1957
```

```
which(dataset_longitudinal$Age==outlying_obs10[,1] & dataset_longitudinal$eTIV==outlying_obs10[,2])
```

```
## [1] 140
```

The patient OAS2_0048 at line 101 results as anomalous in the fourth visit for MMSE vs Age.

The patient OAS2_0048 at line 102 results as anomalous in the fifth visit for both MMSE vs Age, and MMSE vs the last 3 columns (that are dependents).

The patient OAS2_0040 at lines 79, 80, 81 anomalous in all the 3 visits for EDUC vs Age and vs the last 3 columns.

The patient OAS2_0066 at line 140 results as anomalous in the first visit only for Age vs eTIV (!attention: non per le altre due colonne con cui etiv è dipendente! -> può essere il valore di etiv sbagliato nella riga 140? e soprattutto non LL visita successiva)

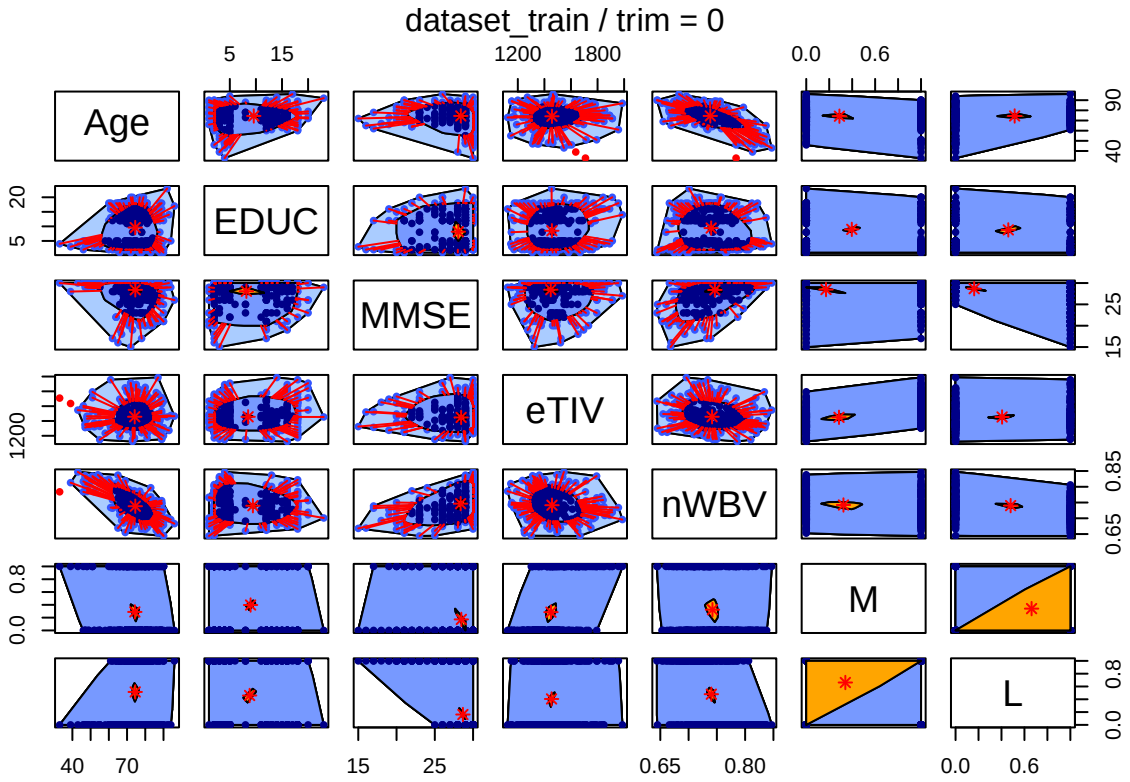
Now we'll do the same with the complete dataset considering only the first visit:

```
dataset_train <- read.csv("train.csv", header = T)
dataset_train <- dataset_train[,-1]
dataset_train$L <- ifelse(dataset_train$label == 'Dem', 1, 0)  # dummy variable for demented/nondemente
head(dataset_train)
```

```
##   M.F Age EDUC MMSE eTIV  nWBV  label M L
## 1   M  87   14   27 1987 0.696 Nondem 1 0
## 2   M  75   12   23 1678 0.736    Dem 1 1
## 3   F  88   18   28 1215 0.710 Nondem 0 0
## 4   M  80   12   28 1689 0.712 Nondem 1 0
## 5   M  71   16   28 1357 0.748    Dem 1 1
## 6   F  93   14   30 1272 0.698 Nondem 0 0
```

To visualize the outliers in this context we retort to a bagplot matrix:

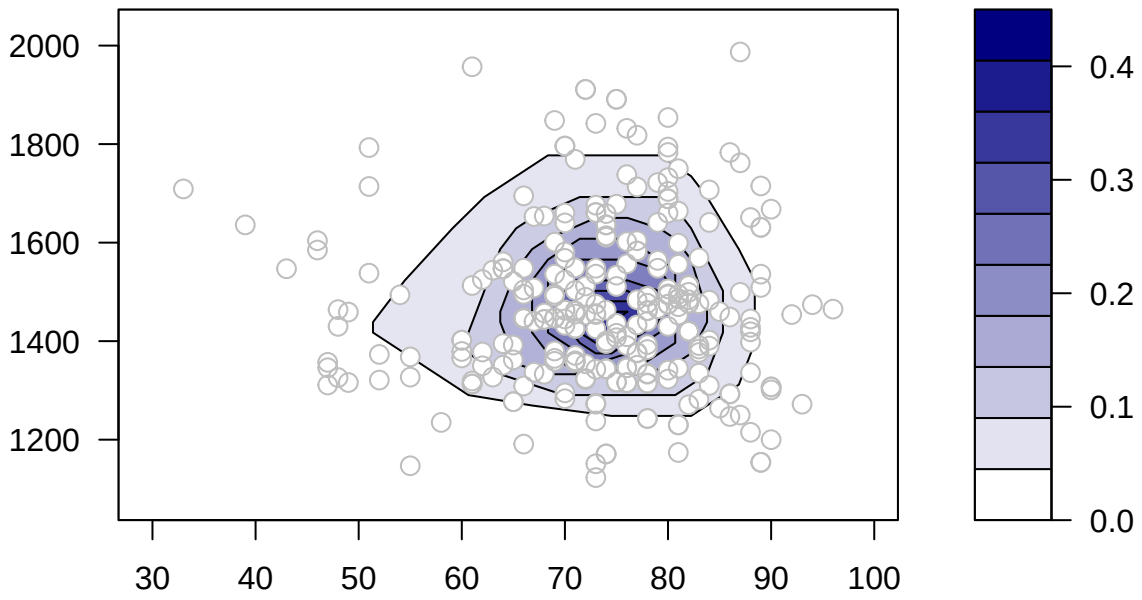
```
bagplot_matrix_train <- aplpack::bagplot.pairs(dataset_train)
```



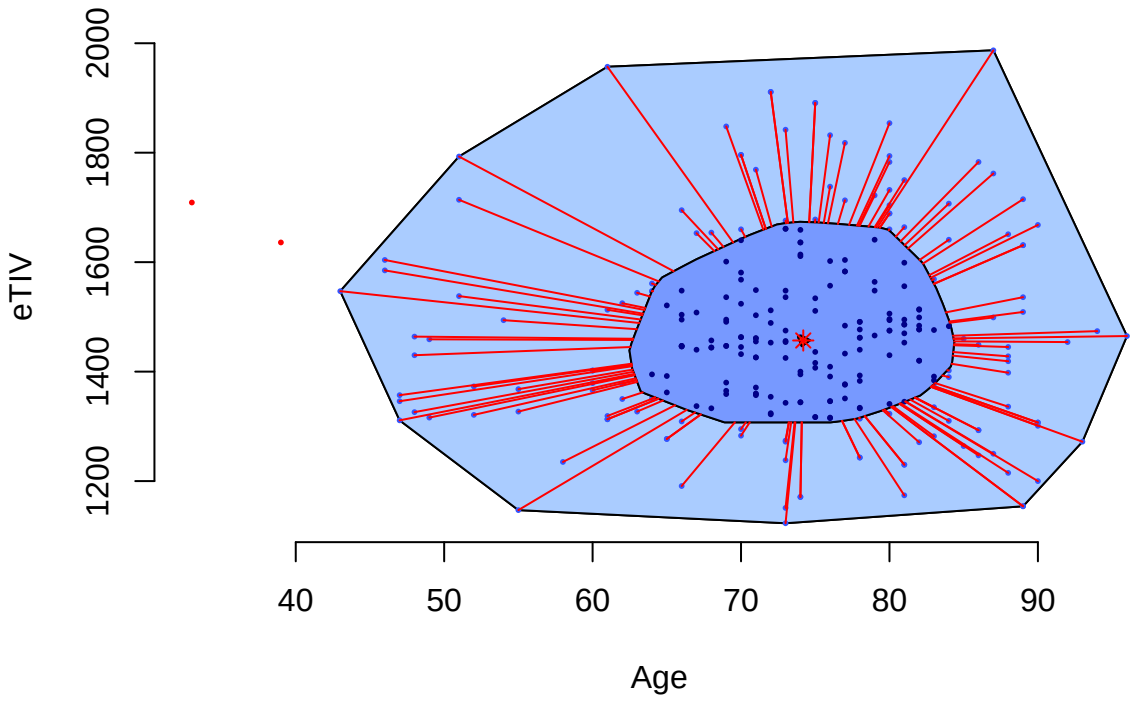
it seems to have problems only Age with eTIV and nWBV

Age vs eTIV:

```
depthContour(
  data.frame(dataset_train$Age, dataset_train$eTIV),
  depth_params = list(method = 'Tukey'),
  points = TRUE,
  colors = colorRampPalette(c('white', 'navy')),
  levels = 10,
  pdmedian = F,
  graph_params = list(cex=.01, pch=1),
  pmean = F
)
```




```
bags1_train <- bagplot(data.frame(dataset_train$Age, dataset_train$eTIV), xlab = "Age", ylab = "eTIV")
```




```
outlying_obs1_train <- bags1_train$pxy.outlier

outlying_obs1_train
```

```
##       x    y
## [1,] 39 1636
## [2,] 33 1709
```

```
which(dataset_train$Age==outlying_obs1_train[1,1] & dataset_train$eTIV==outlying_obs1_train[1,2])
```

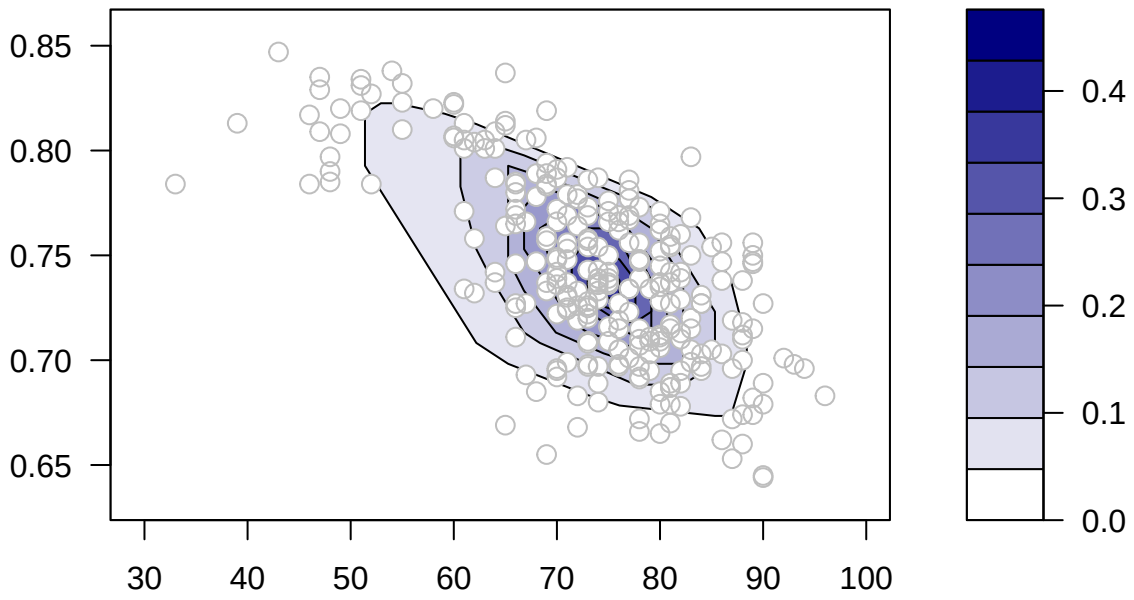
```
## [1] 145
```

```
which(dataset_train$Age==outlying_obs1_train[2,1] & dataset_train$eTIV==outlying_obs1_train[2,2])
```

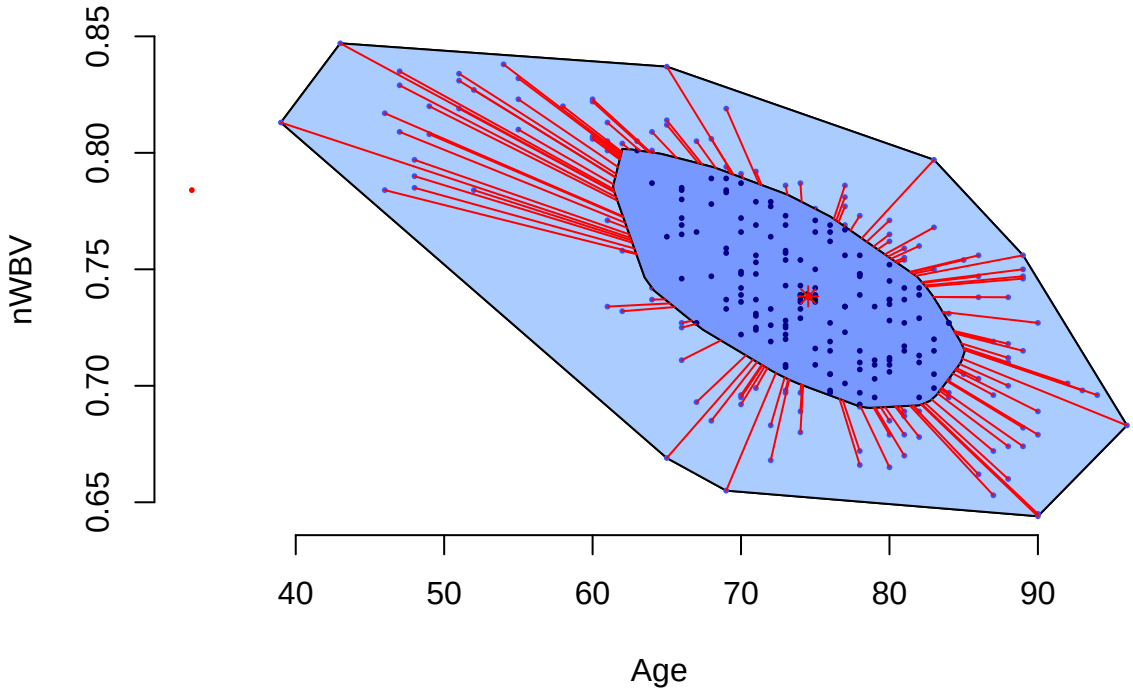
```
## [1] 175
```

Age vs nWBV:

```
depthContour(
  data.frame(dataset_train$Age, dataset_train$nWBV),
  depth_params = list(method = 'Tukey'),
  points = TRUE,
  colors = colorRampPalette(c('white', 'navy')),
  levels = 10,
  pdmedian = F,
  graph_params = list(cex=.01, pch=1),
  pmean = F
)
```




```
bags2_train <- bagplot(data.frame(dataset_train$Age, dataset_train$nWBV), xlab = "Age", ylab = "nWBV")
```




```
outlying_obs2_train <- bags2_train$pxy.outlier

outlying_obs2_train
```

```
##       x     y
## [1,] 33 0.784
```

```
which(dataset_train$Age==outlying_obs2_train[,1] & dataset_train$nWBV==outlying_obs2_train[,2])
```

```
## [1] 175
```

In the dataset 'train.csv': -at line 175 there is a patient anomalous both for Age vs eTIV and Age vs nWBV; -at line 145 there is a patient anomalous only for Age vs eTIV.