

KID Function

Fabian Blasch

08.02.2022

Packages

```
# Packages
get.package <- function(package){

  lapply(package, \(x){
    # check if packages are installed and if not install them
    if(!require(x, character.only = T)){
      install.packages(x)
    }
    # call package
    library(x, character.only = T)
  })

}

# exec
get.package(c("png", "jpeg", "tabulizer", "pdftools", "raster", "rgdal", "sp",
             "cluster", "fastcluster"))

# since I will use Map() and lapply() for plotting I will wrap them in invisible()
invis.Map <- function(f, ...) invisible(Map(f, ...))
invis.lapply <- function(x, f, ...) invisible(lapply(x, f, ...))
```

Actual SRRI

We can obtain the actual SRRI from the file name. Later this data will be utilized to evaluate the classification accuracy of the applied methods.

```
# set
setwd("./../../KIDs")

# files
file_names <- list.files(pattern = ".pdf", recursive = T)

# create df
dat.valid.SRRI <- as.data.frame(cbind("KID" = file_names,
                                         "SRRI" = sapply(strsplit(sapply(strsplit(file_names, "_", fixed = T),
                                         function(x) x[length(x)]), ".", fixed = T), "[", 1)))

# split first col
```

```

dat.valid.SRRI[, "KAG"] <- sapply(strsplit(dat.valid.SRRI[, 1], "/"), "[", 1)
dat.valid.SRRI[, "KID"] <- sapply(strsplit(dat.valid.SRRI[, 1], "/"), "[", 2)

# order
dat.valid.SRRI <- dat.valid.SRRI[, c(3, 1, 2)]

# glimpse
head(dat.valid.SRRI, 7)

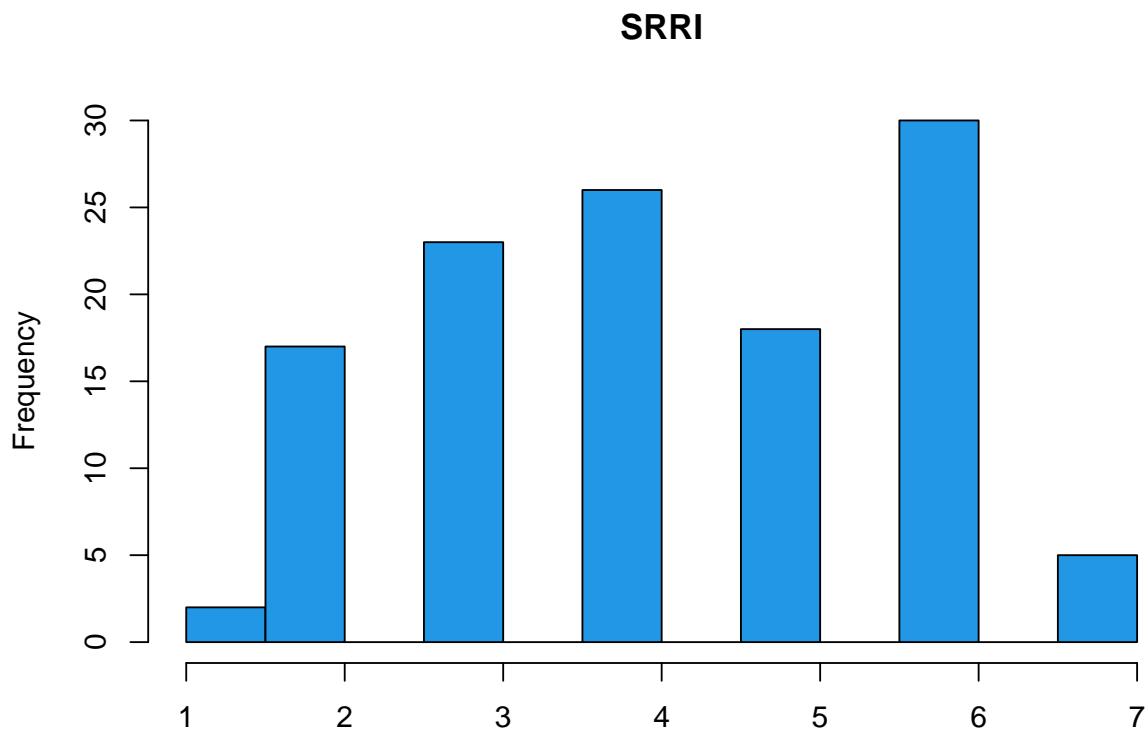
##      KAG          KID SRRI
## 1 Allianz ki-allakt_6.pdf    6
## 2 Allianz  ki-allap_6.pdf    6
## 3 Allianz ki-alleur_2.pdf    2
## 4 Allianz  ki-allna_6.pdf    6
## 5 Allianz ki-allnar_2.pdf    2
## 6 Allianz ki-allore_3.pdf    3
## 7 Allianz ki-allost_6.pdf    6

# dim
dim(dat.valid.SRRI)

## [1] 121   3

# Hist
hist(as.numeric(dat.valid.SRRI[, "SRRI"]), breaks = 10, main = "SRRI", col = 4, xlab = "")

```



Shade Color

To extract the SRRI the following colors are required and need to be converted to HEX.

```
# set
{setwd("./.../KIDs/Auxiliary")

# import
dat.col.KAG <- read.table(list.files(pattern = "RGB"),
                           col.names = c("KAG", "R", "G", "B"))

# add hex
sapply(as.data.frame(t(dat.col.KAG[, -1])), 
      function(x) do.call(rgb, as.list(c(x, maxColorValue = 255)))) -> HEX

# bind
dat.col.KAG <- cbind(dat.col.KAG, "HEX" = HEX)

# display
dat.col.KAG

# list of RG vectors for all KAGs
KAG_RGB <- setNames(lapply(as.data.frame(t(dat.col.KAG)[2:4, ]), function(x){
  as.numeric(x)
}), nm = dat.col.KAG[, 1])

# correct IQAM
dat.col.KAG[5, 5] <- "#959595"

# IQAM actually features changing color shemes across their KIDs
# therefore the format haws to change from df to list as we have to match
# against multiple colors

list.col.KAG <- as.list(dat.col.KAG[order(dat.col.KAG$KAG), ncol(dat.col.KAG)]) |>
  setNames(nm = dat.col.KAG$KAG[order(dat.col.KAG$KAG)])

# add second col to IQAM
list.col.KAG[["IQAM"]] <- c(list.col.KAG[["IQAM"]], "#949494")

# correct Schoellerbank col and add
list.col.KAG[["Schoellerbank"]] <- "#E3E3E3"
list.col.KAG[["Schoellerbank"]] <- c(list.col.KAG[["Schoellerbank"]], "#CBCBCB")

# correct Union
list.col.KAG[["Union"]] <- "#4F4E4E"

# write
# write.csv(dat.col.KAG, "KAG_COL_HEX.csv")
# saveRDS(list.col.KAG, file = "list_col_KAG.rds")
}
```

SRRI Extraction Function

Given a KID document this function aims to extract the SRRI from the standard graph (usually) located on the first of two pages.

```
# load package
setwd("./../Package/KIDs")
devtools::load_all()}
```

Tests

Starting with one KAG.

Erste

```
# setwd to file that contains KIDs
setwd("./../../KIDs")

# safe dirs
dirs <- list.dirs()[-c(1, 4)] # remove hardcode later

# colors
col <- dat.col.KAG[order(dat.col.KAG[, "KAG"]), c("KAG", "HEX")]
col[5, 1] <- "Kepler Fonds"

# test Erste
Map(function(x){

  # change dir to respective KAG
  {setwd(x)

  # pdfs
  file_nom <- list.files(pattern = ".pdf")}

  # FUN over all .pdfs
  lapply(file_nom, function(z){
    SRRI_ext_loc(doc = z, col = dat.col.KAG[4, 5])
  })

}, dirs[3])) -> erste.test

# extracted SRRI
cbind(dat.valid.SRRI[dat.valid.SRRI[, "KAG"] == "Erste", ],
      "Extracted" = sapply(erste.test[[1]], "[[", 2)) -> res

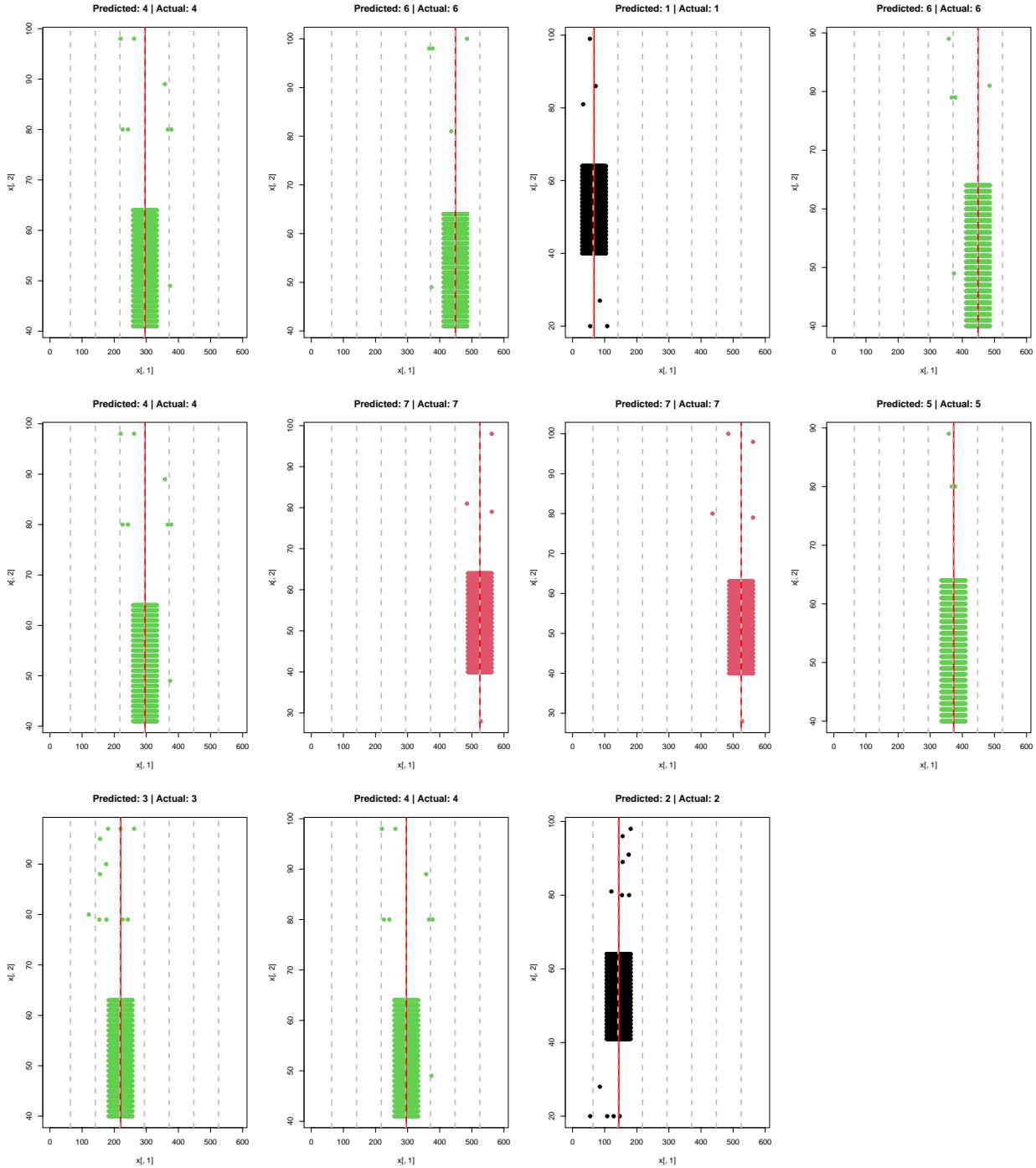
# align
par(mfrow = c(3, 4))

# plot
invis.Map(function(x, y, z, l, k, m){

  {plot(x[, 1], x[, 2], col = x[, ncol(x)], pch = 19, main = paste("Predicted:", z, "| Actual:", l),
        xlim = c(1, 590))
  abline(v = y, col = "red", lwd = 2)
  lapply(k, function(x) abline(v = x, col = "grey", lwd = 2, lty = 2))}

}, lapply(erste.test[[1]], "[[", 3), sapply(erste.test[[1]], "[[", 4), res[, 4], res[, 3],
```

```
lapply(erste.test[[1]], "[", 5), lapply(erste.test[[1]], "[", 5))
```



In the case of Erste the SRRI extraction works perfectly. Now the remaining KAGs will be examined.

```
# setwd to file that contains the first KIDs
setwd("./.../KIDs/Allianz")

# store Errors
utils::capture.output(
```

```

# Map over dirs
Map(function(x, y){

  # set
  {setwd(paste0("./../", x))

  # pdfs
  file_nom <- list.files(pattern = ".pdf")}

  # lapply over all .pdfs
  lapply(file_nom, function(z){

    # extract and error handle
    try(SRRI_ext_loc(doc = z, col = y, tol = 50), silent = F)

  })

}, dirs, list.col.KAG) -> test

, type = "message")

# error index
lapply(test, function(x){

  # error ind
  which(sapply(x, class) == "try-error")

}) -> err.tmp

# retrieve error throwing funds with ind
do.call(rbind, Map(function(x, y, z){

  if(length(y) > 0){

    {setwd(paste0("./../", z))

    # .pdfs
    file_nom <- list.files(pattern = ".pdf")}

    # subset
    cbind(rep(z, length(y)),
          file_nom[y],
          sapply(x[y], "[", 1))

  } else {
    cbind(NA, NA, "No errors.")
  }

}, test, err.tmp, dirs)) -> dat.err

```

Now that we have identified all KIDs for which the extraction failed, we can proceed to see if the classification was correct for the remaining kids.

```

# Plot
Map(function(x, y){

  sapply(y, function(z){
    # cond
    if(class(z) == "try-error"){
      return(NA)
    } else {
      z[[2]]
    }
  }) -> tmp

  # match
  cbind(dat.valid.SRRI[dat.valid.SRRI[, "KAG"] == x, ],
        "Extracted" = tmp)

}, col[, 1], test) -> tef

# plot

# over KAGs
invis.Map(function(m, n){

  # arrange
  par(mfrow = c(ceiling(length(m) / 4), 4))

  # over KIDs
  invis.Map(function(x, y, z, k){

    if(class(x) == "try-error"){

      # plot empty for KIDs that remain unclassified for now
      plot(NULL, xlim = c(0, 1), ylim = c(0, 1), main = paste(k, "\n", "Error"))

    } else {
      # build tmp vars for plotting
      plot.coo <- x[[3]]
      med <- x[[4]]
      scal <- x[[5]]
      pred <- y
      act <- z
      fund <- k

      # plot
      plot(plot.coo[, 1], plot.coo[, 2], col = plot.coo[, ncol(plot.coo)], pch = 19,
            xlim = c(1, 590),
            main = paste(fund, "\n", "Predicted:", pred, "| Actual:", act))

      # median
      abline(v = med, col = "red", lty = 1, lwd = 2)

      # Scale
      lapply(scal, function(s) abline(v = s, col = "grey", lwd = 2, lty = 2))

    }
  })
})
}

```

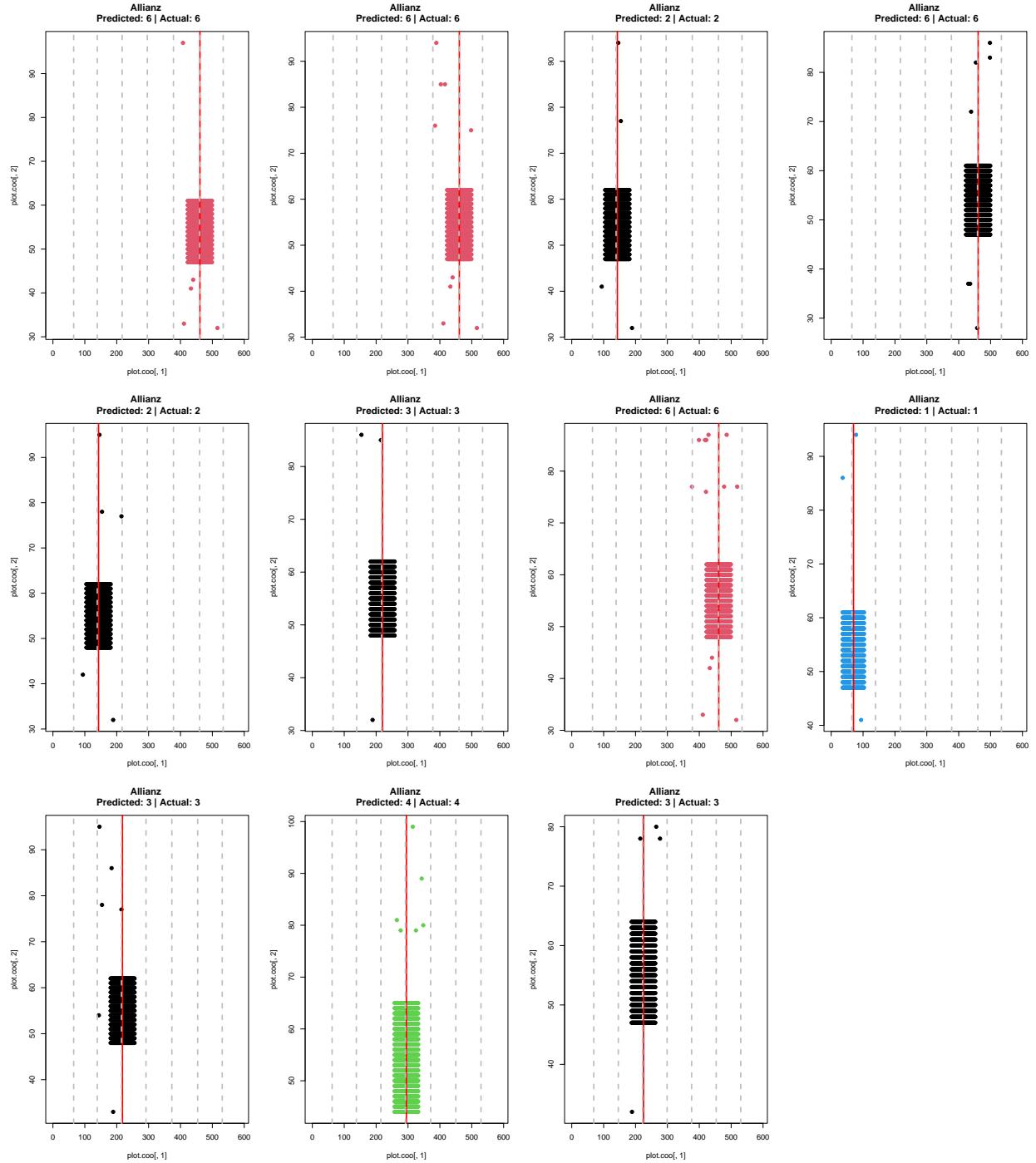
```

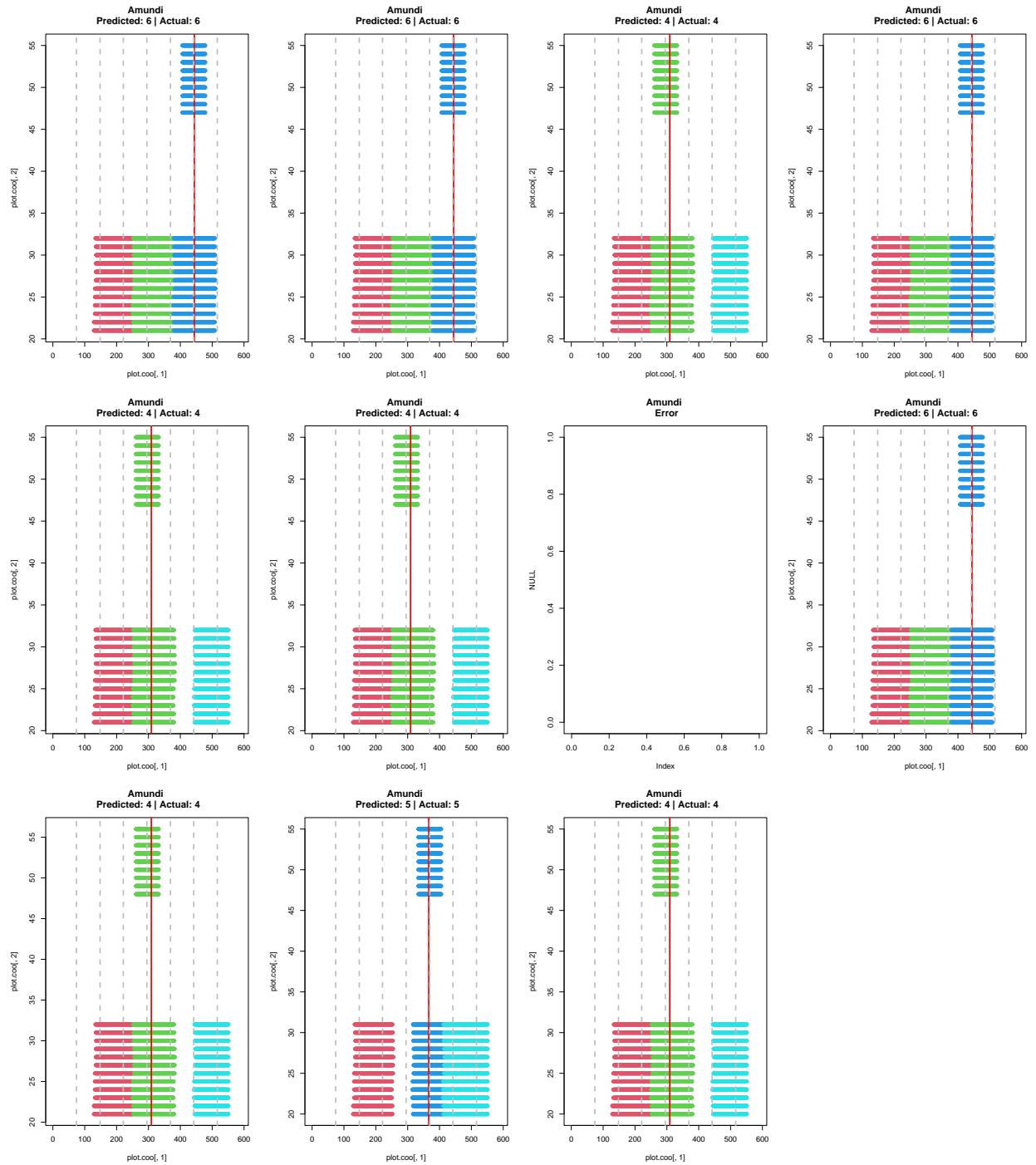
        }

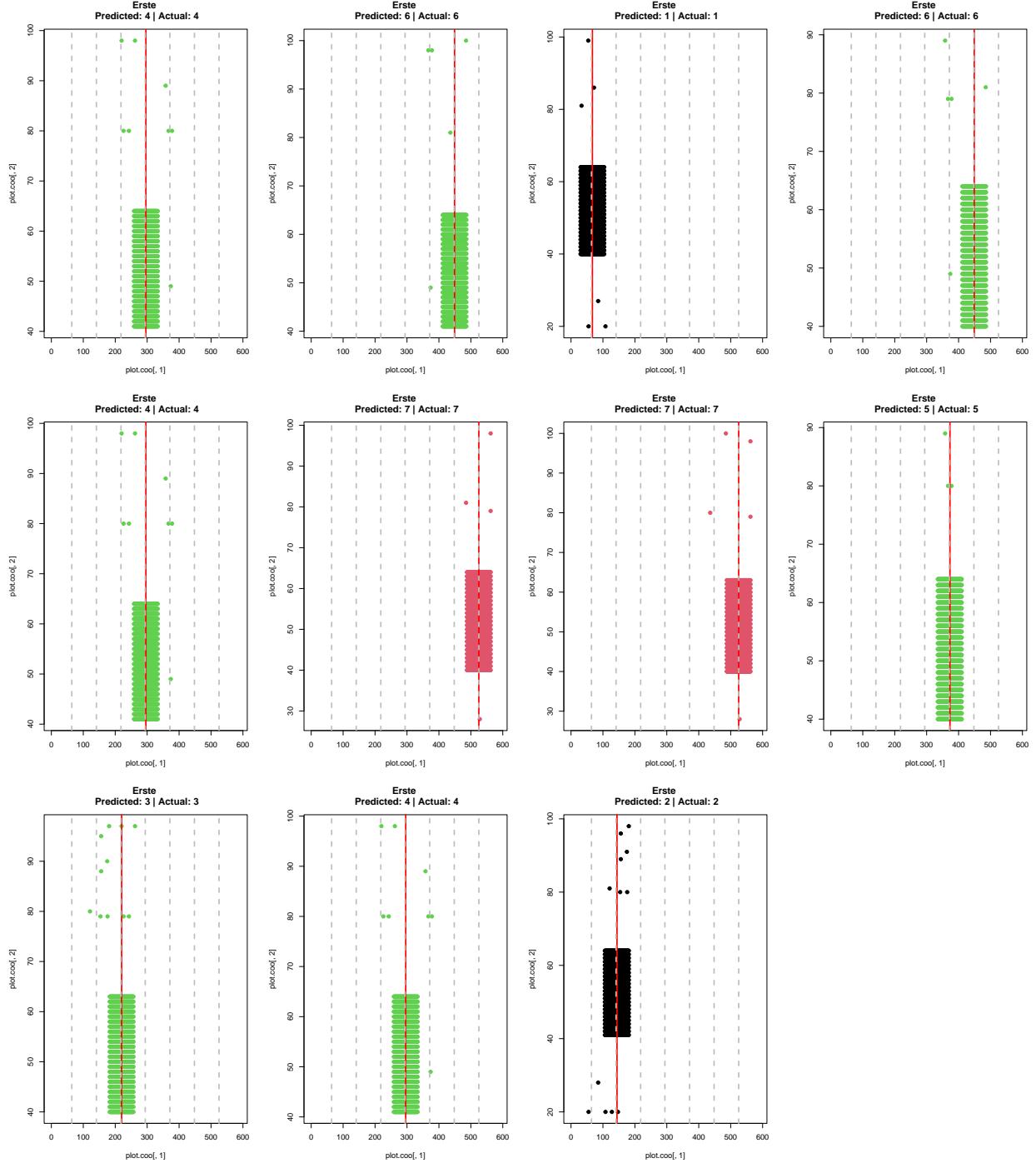
    },m , n[, 4], n[, 3], n[, 1])

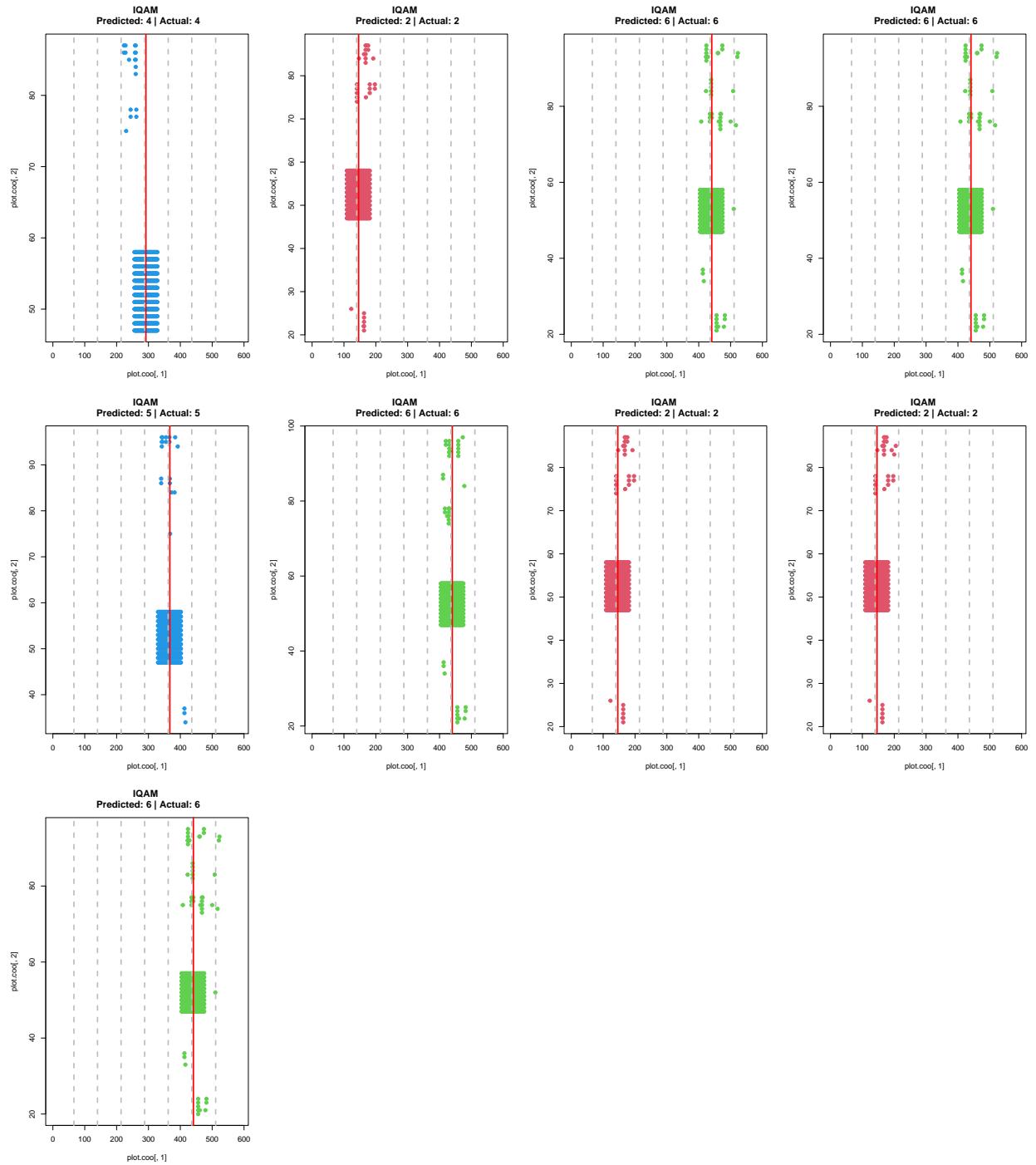
}, test, tef)

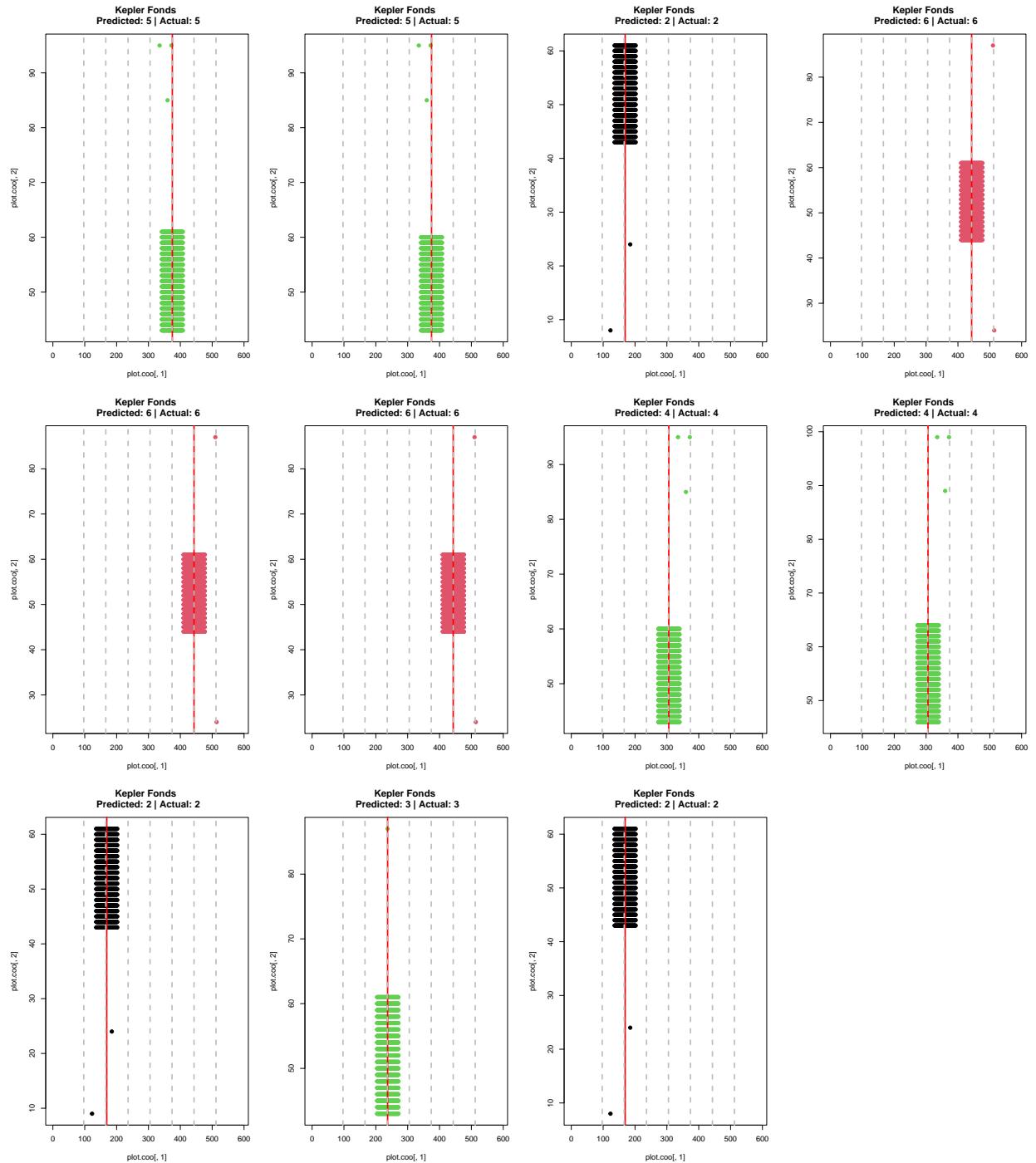
```

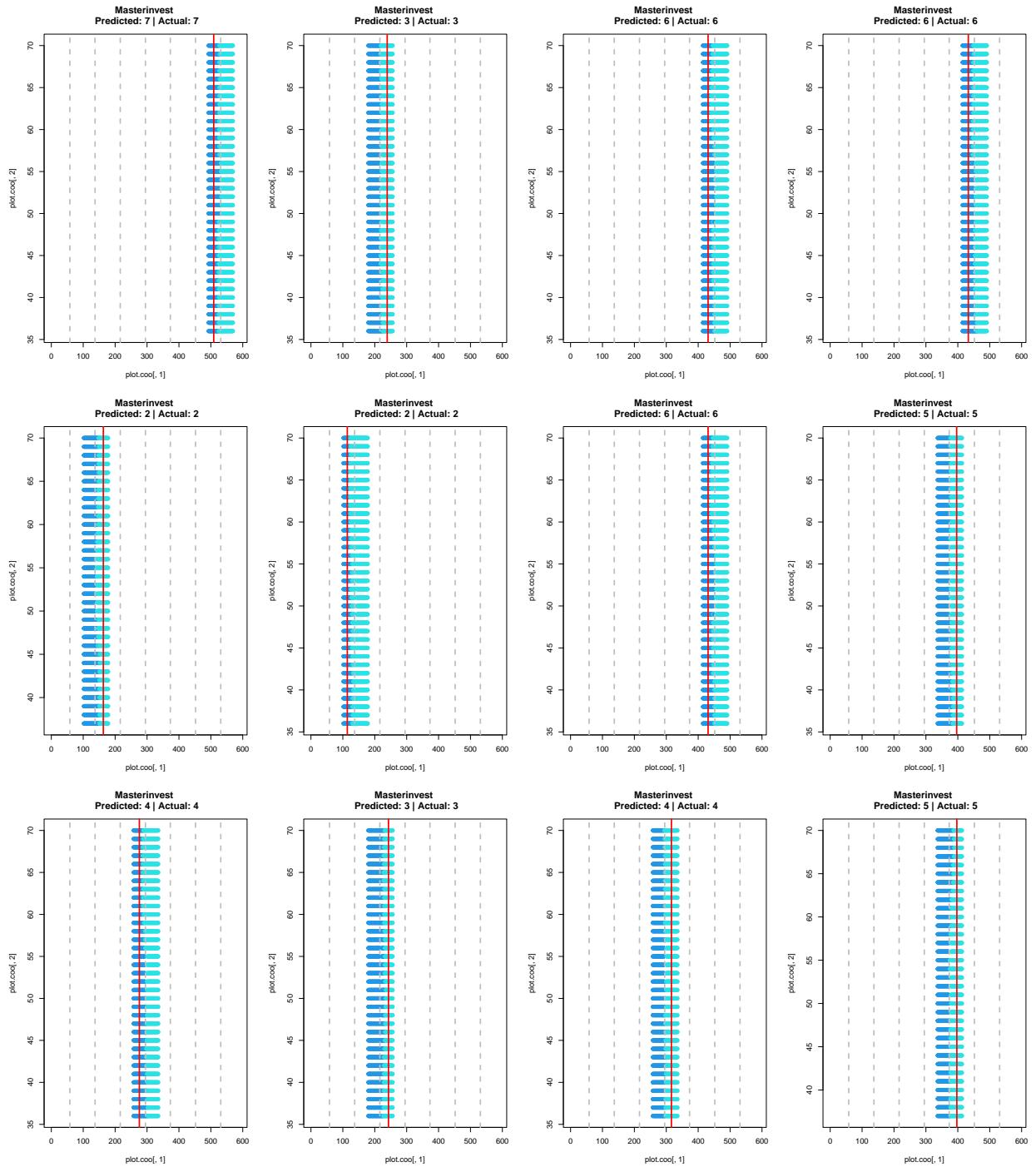


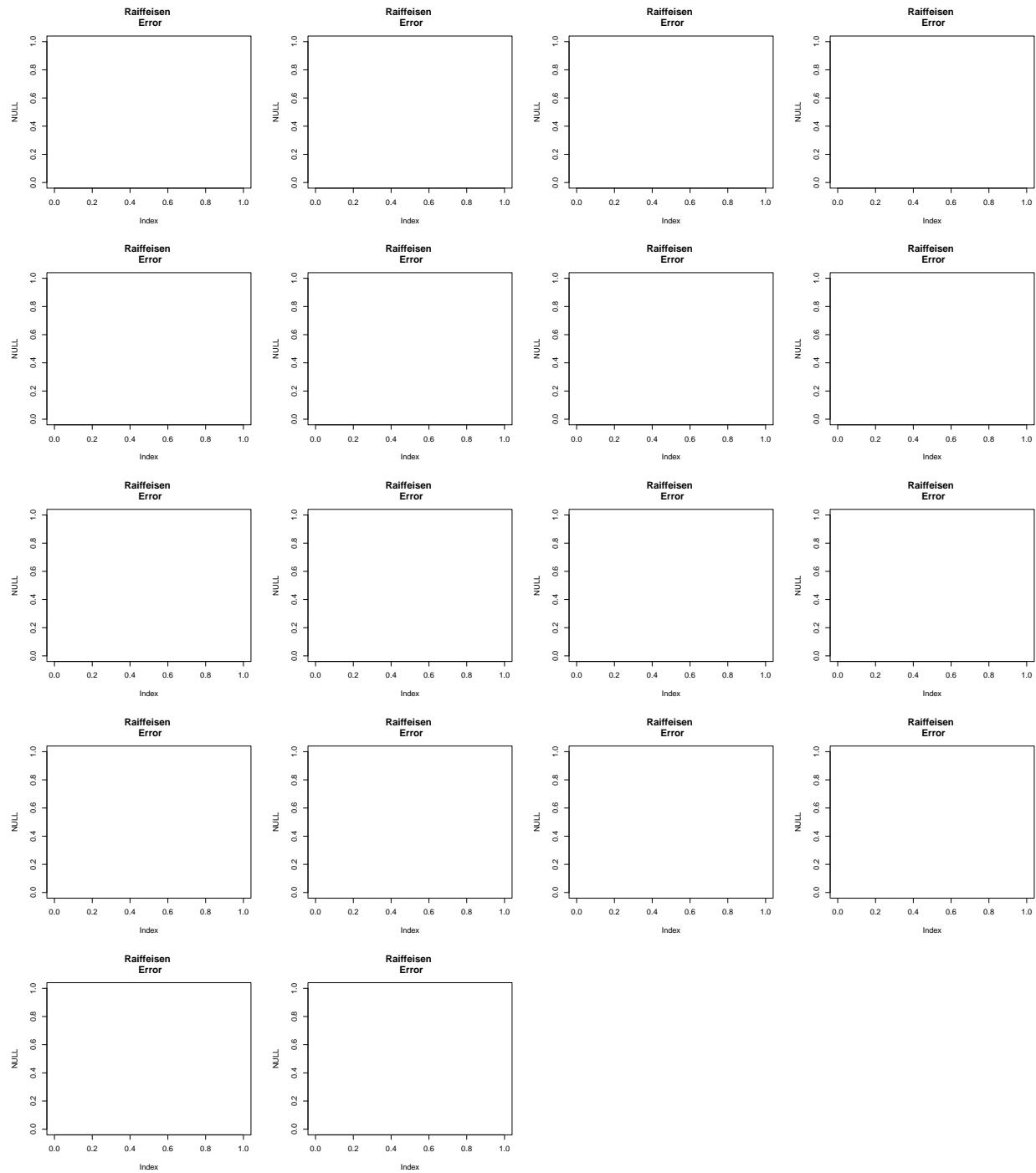


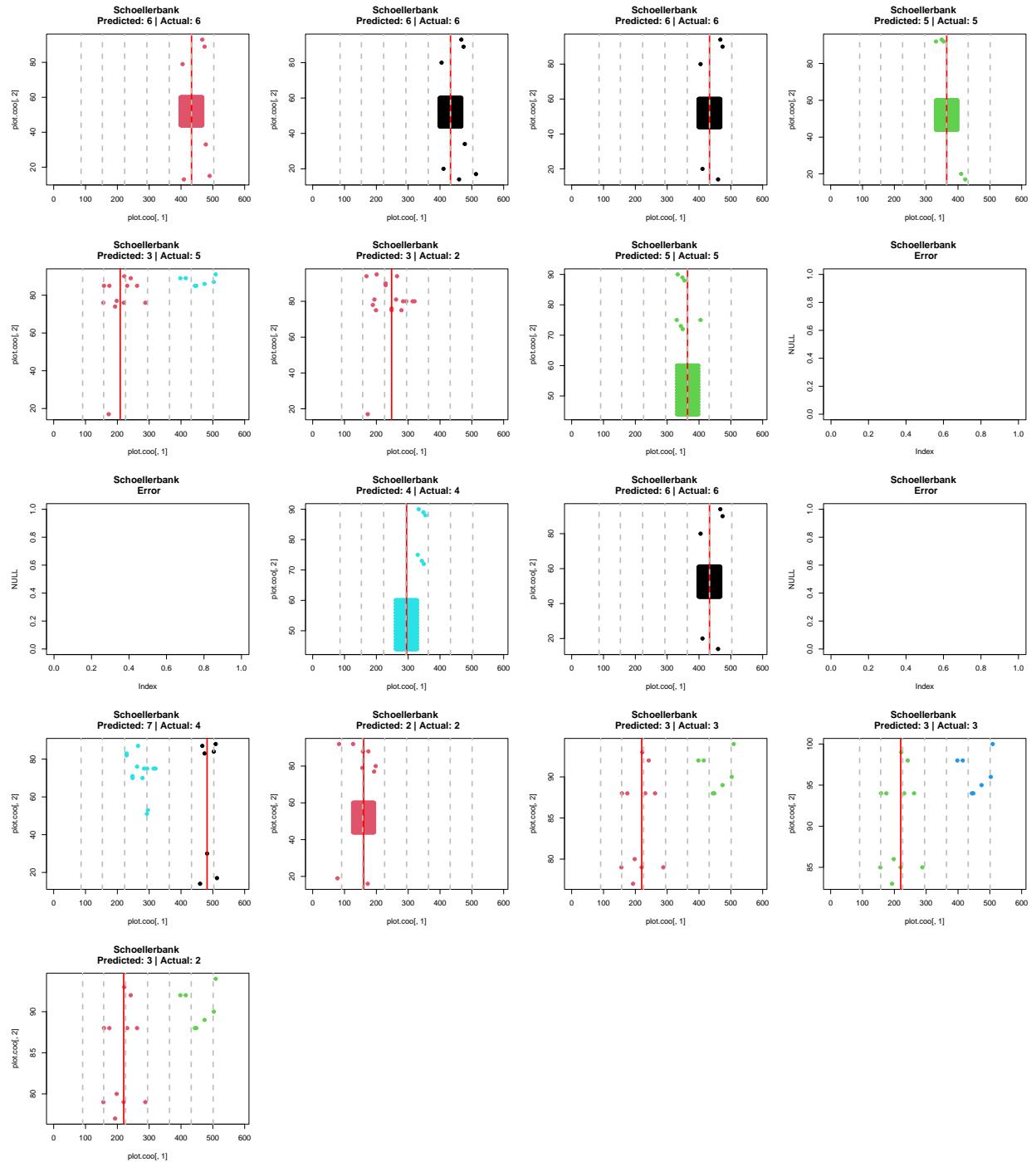


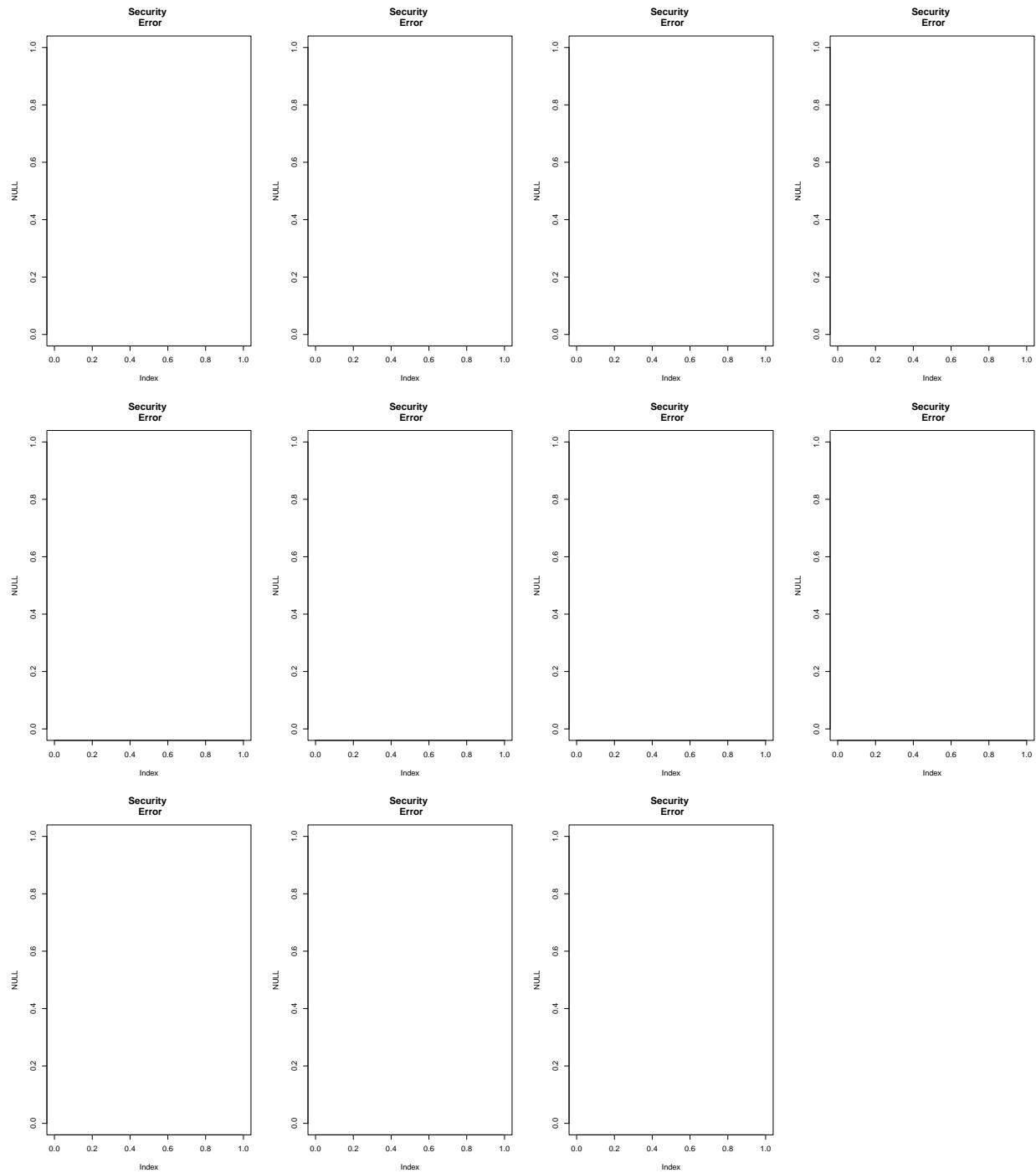


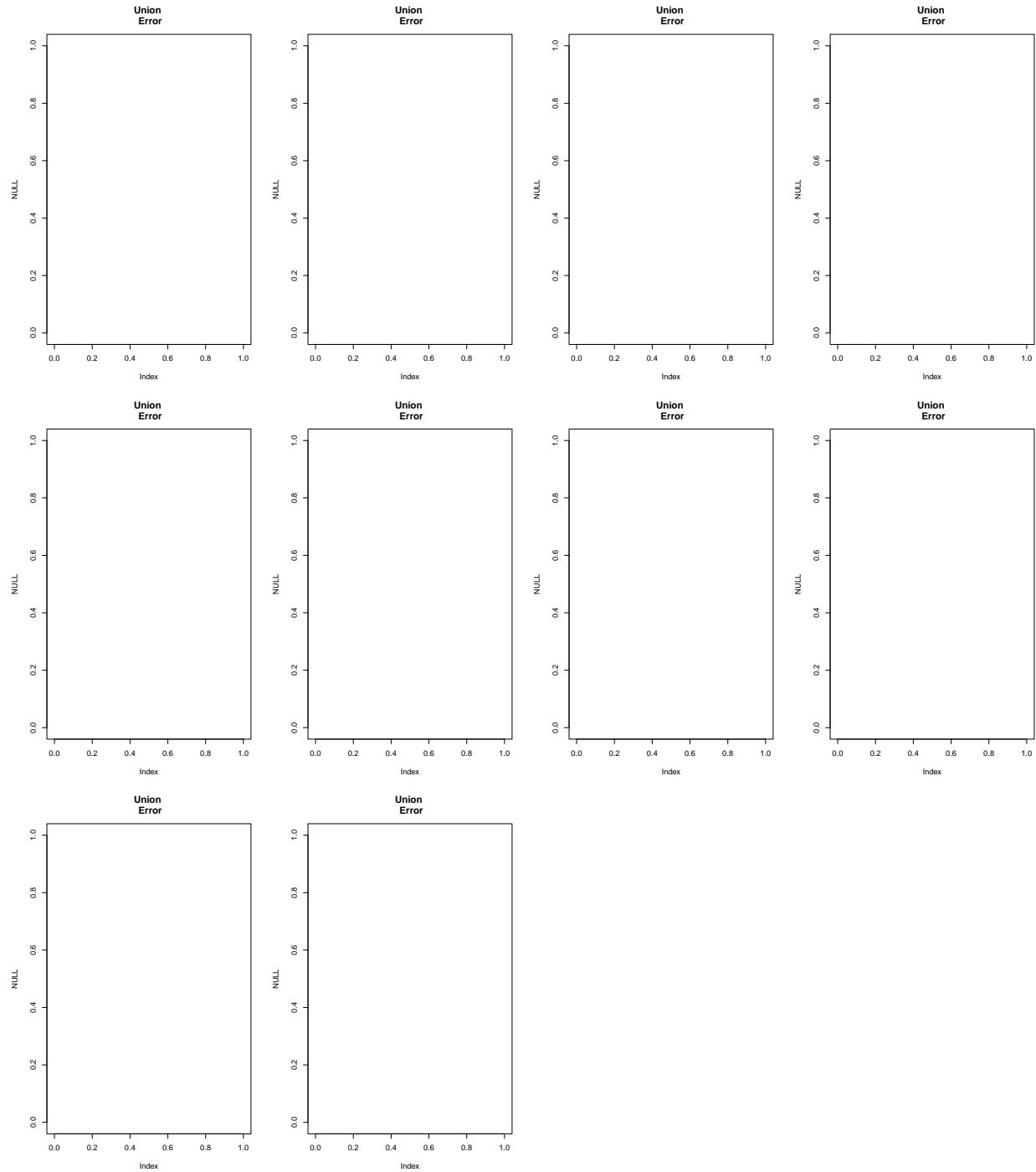












Problem description by KAG

- Allianz
 - Works for the given sample.
- Amundi
 - Works for every KID excluding one that is a scanned pdf. Unfortunately, all Amundi files have 2 bars in the same color as the SRRI shade, accordingly the weight of SRRI shade is reduced and the median no longer lies in the middle of the SRRI shading. For now all scanned PDFs will return a warning. Later one could implement an algorithm that detects whether bars cover the entire page and accordingly remove those bitmap entries.
- Erste
 - Works for the given sample.
- IQAM
 - Works for the given sample.
- Kepler Fonds
 - Works for the given sample.
- Masterinvest
 - Works for the given sample.
- Raiffeisen
 - Readout does not contain the full scale.
 - Check color.
 - Check last three PDFs for text detection issues.
- Schoellerbank
 - No box detected, check color.
 - Three PDFs threw an error, all because of text detection.
- Security
 - Readout does not contain the full scale.
 - Scale is completely off.
 - Rectangle is missing in some files.
 - Check cutoff.
 - Check color.
- Union
 - Check SRRI text detection.