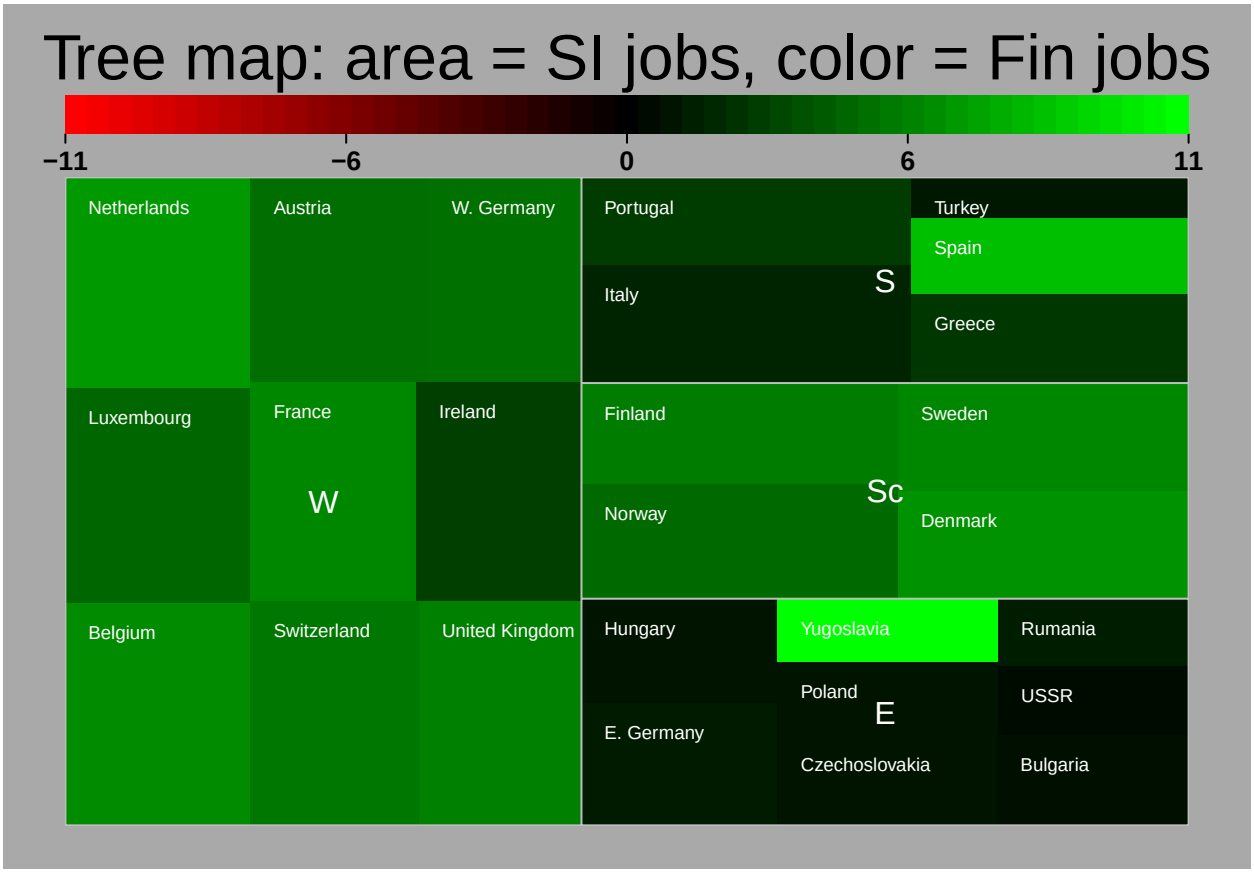
Lab report 2

Thomas Zhang

2016 M09 12

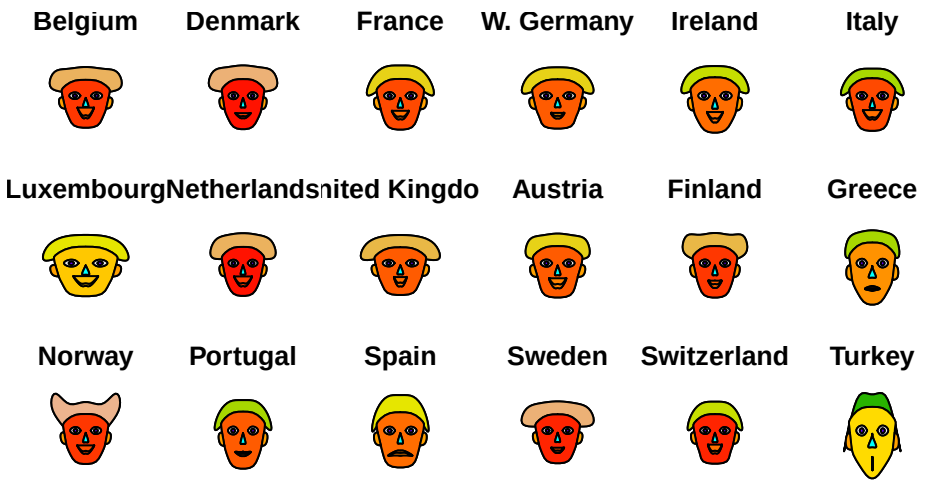
Assignment 1

OK, we create a tree map where the sizes are indicative of Percentage employed in service industries and the colors are indicative of Percentage employed in finance. We group the country observations by geographical group.



We see that the western european economies are fairly large in services, and they are also more developed than for instance the eastern european economies in finance. This seems reasonable to me.

Next, we do chernoff faces out of the variables in the `jobs.txt`. Constant dummy variables are used for the appearance of eyes, nose and ears.

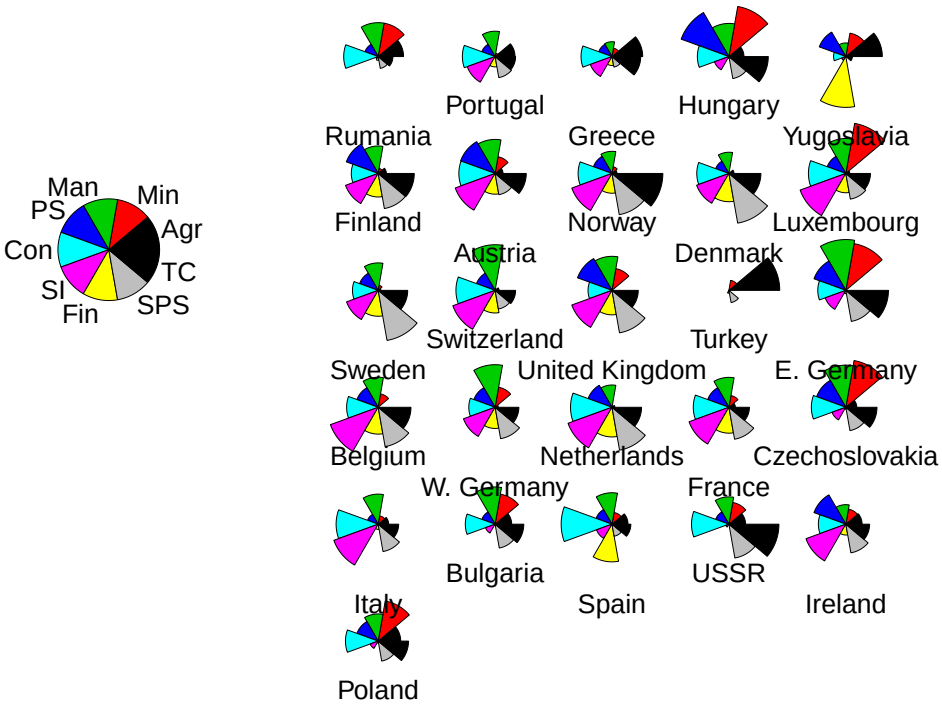



```
## effect of variables:
##  modified item       Var
```

```
##  "height of face   " "Agr"
##  "width of face    " "Min"
##  "structure of face" "Man"
##  "height of mouth  " "PS"
##  "width of mouth   " "Con"
##  "smiling          " "SI"
##  "height of eyes   " "ones"
##  "width of eyes    " "ones"
##  "height of hair   " "Fin"
##  "width of hair    "  "SPS"
##  "style of hair    "  "TC"
##  "height of nose   "  "ones"
##  "width of nose    "  "ones"
##  "width of ear     "  "ones"
##  "height of ear    "  "ones"
```

In my view, We see that there are two main clusters among the faces. It is roughly split between west and east europe.

Let us reorder the countries (left to right, up to down, like reading a book) using hierachical clustering and plot the variables as segment charts and compare.

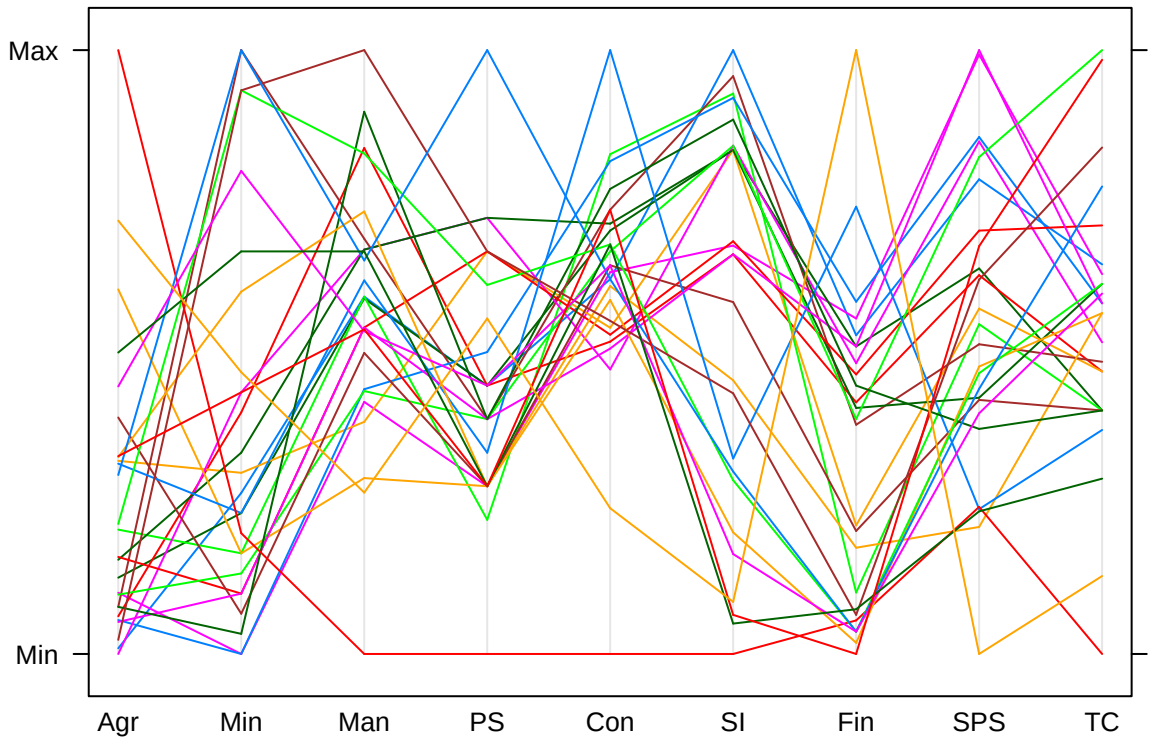


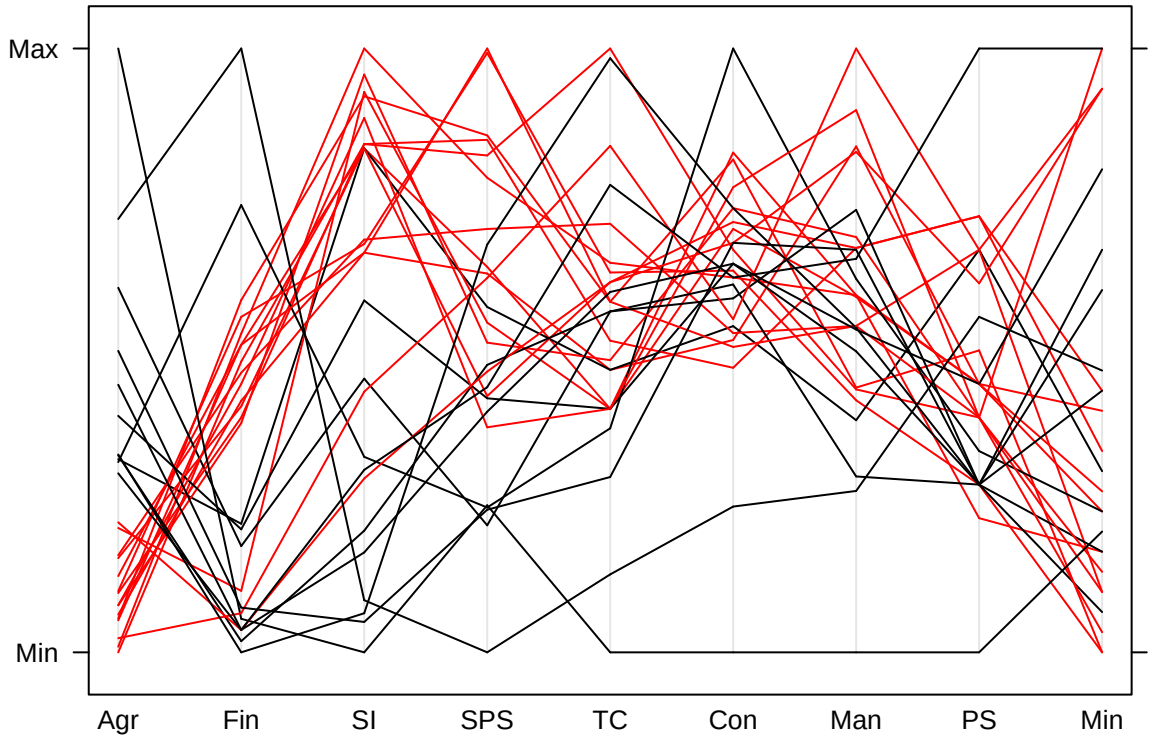
Honestly, these segment charts reinforces a north-western european cluster large in SI and Con employed percentages, while the other countries range from balanced sectors to overweight in Agr employed percentage.

We now make three parallel coordinate plots of the countries over their quantitative employed-in-various-sectors percentage variables. The first parallel plot is unsorted, the second is sorted by the travelling salesman

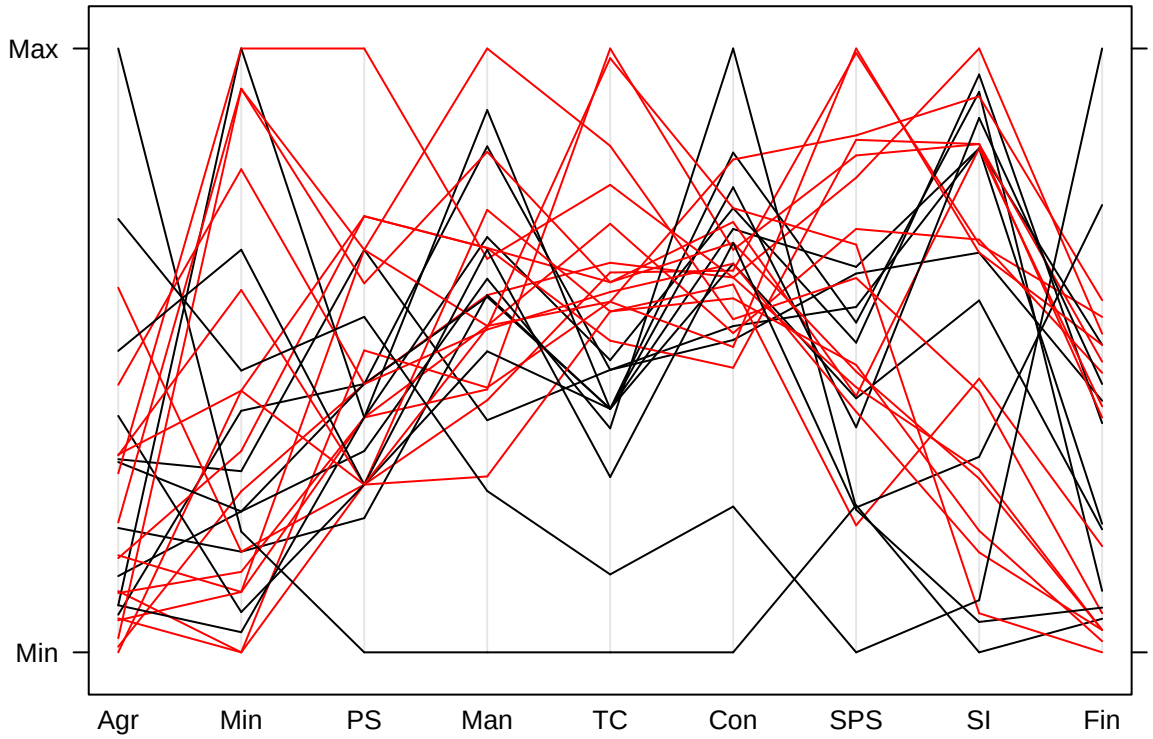
algorithm between most correlated variables and the third is sorted by hierachical clustering seriation on the same distance measure as the second.

For the sorted parallel coordinate plots we try to split the countries into two clusters based on observation on a suitable variable. That variable is Agr for second plot and TC for third plot. We list the countries in the red colored clusters.






```
##  [1] "Belgium"        "Denmark"        "France"         "W. Germany"
##  [5] "Italy"          "Luxembourg"     "Netherlands"    "United Kingdom"
##  [9] "Austria"        "Finland"        "Norway"         "Sweden"
## [13] "Switzerland"    "Czechoslovakia" "E. Germany"
```




```
##  [1] "Belgium"        "Denmark"        "Netherlands"    "United Kingdom"
##  [5] "Austria"        "Finland"        "Greece"         "Norway"
##  [9] "Sweden"         "Bulgaria"       "Czechoslovakia" "E. Germany"
## [13] "Hungary"        "Poland"         "USSR"
```

From the first plot, we can clearly see that Turkey, which scores low on many variables, is a clear outlier. It also seems as if sectors Man and Con are correlated with eachother.

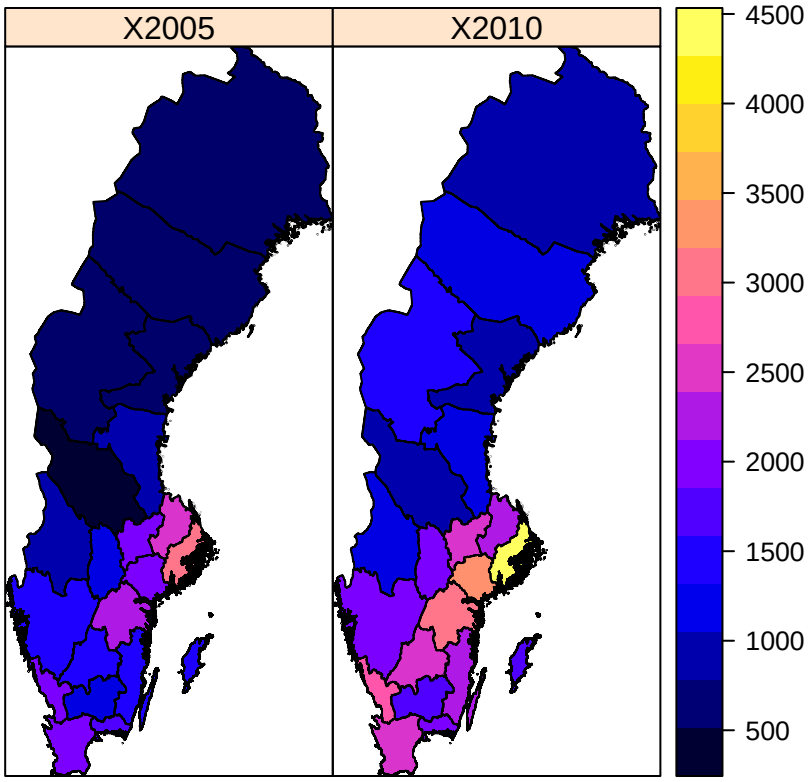
From the second plot, we see that low percentages of employment in the Agr sector is characteristic of the north-western european countries.

From the third plot, I think I can say that TC may not be such a good clustering variable as Agr.

Overall I think the second parallel coordinate plot was easiest to analyze. The clustering in the second plot corresponds well to the group variable and the results found in the segment charts and chernoff faces. Turkey is easy outlier to see in all visualizations as well, mainly thanks to its heavy realiance on the Agr(icultural) sector.

Assignment 2

We use data from SCB to plot the agricultural real estate prices in Sweden for the years 2005 and 2010 by county.



Generally, we see that the closer to the heart of darkness that is the government and finance center of Stockholm the real estate is located, the higher the average price is. This is true for both years 2005 and 2010. Why is it that people are willing to pay more for an *agricultural* real estate closer to a trash pile of human iniquity? Is it a mass delusion of an extraordinarily vain and stupid people, or maybe a convenience fee for the fat politico-capitalist scum who wish to pretend being simple farmers in their spare time? I will laugh the day the property ponzi scheme bursts, as all such bubble phenomena always do.

Appendix

R code

```
library(XLConnect)
library(gclus)
library(aplpack)
library(portfolio)
library(lattice)
library(TSP)
library(sp)
library(seriation)
jobs <- read.table("jobs.txt",sep="\t",header = TRUE)
map.market(id = jobs$Country, area = jobs$SI, group = jobs$Group,
           color = jobs$Fin, lab = c("group" = TRUE, "id" = TRUE),
           main = "Tree map: area = SI jobs, color = Fin jobs")
ones <- rep(1,26)
```

```
facesframe <- cbind(jobs[,2:7],ones,ones,jobs[,8:10],
                    ones,ones,ones,ones)
faces(facesframe, labels = jobs$Country,nrow.plot = 3)
countrydist <- dist(facesframe)
neworder <- order.single(countrydist)
segmentframe <- jobs[neworder,2:10]
par(mfrow = c(1,1))
stars(segmentframe,draw.segments = TRUE,labels = as.character(jobs$Country[neworder]),
      key.loc = c(-3,10))
parallelplot(jobs[,2:10],horizontal.axis = FALSE)

dd <- as.dist((1 - cor(jobs[,2:10]))/2)

res<-solve_TSP(TSP(dd))
colore = 1+(jobs$Agr - min(jobs$Agr) < 1/4 * (max(jobs$Agr) - min(jobs$Agr)))
parallelplot(jobs[,(1 + as.integer(res))],
             horizontal.axis = FALSE, col = colore)
as.character(jobs$Country[which(colore == 2)])

res2 <- get_order(seriate(dd))
color = 1+(jobs$TC - min(jobs$TC) > 1/2 * (max(jobs$TC) - min(jobs$TC)))
parallelplot(jobs[,(1 + as.integer(res2))],
             horizontal.axis = FALSE, col = color)
as.character(jobs$Country[which(color == 2)])
swecounties <- readRDS("SWE_adm1.rds")
temp <- swecounties@data
temp$NAME_1[12] <- "Ã-rebro"
names(temp)[6] <- "County"
wb <- loadWorkbook("BO0501C6.xlsx")
swehousedata = readWorksheet(wb,sheet = "BO0501C6")
newframe <- merge(temp, swehousedata,sort = FALSE)
swecounties@data$X2005=newframe$X2005
swecounties@data$X2010=newframe$X2010

spplot(swecounties, zcol=c("X2005","X2010"))
##
```