Lab 2 - Visualization

Gustav Sternelöv

8 september 2016

Assignment 1

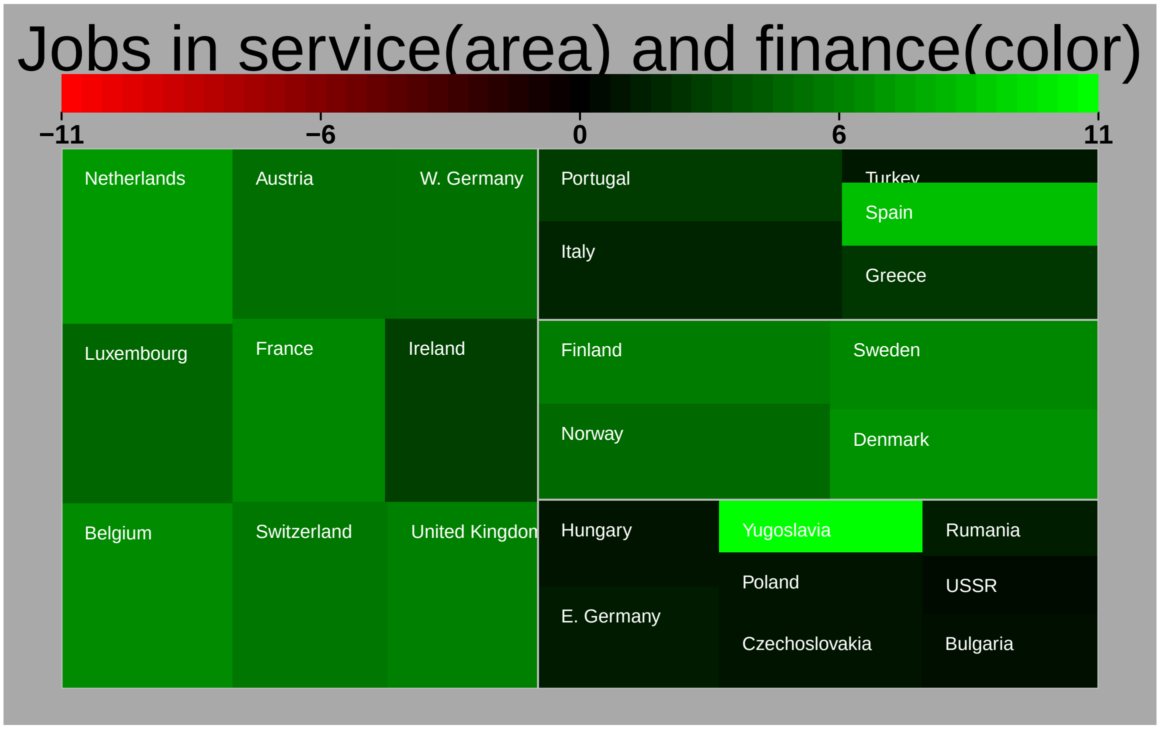
1)

The column *group* is added to the data set. The first six observations in the data set are shown below.

```
##       Country  Agr Min  Man  PS  Con   SI Fin  SPS  TC group
## 1     Belgium  3.3 0.9 27.6 0.9  8.2 19.1 6.2 26.6 7.2     W
## 2     Denmark  9.2 0.1 21.8 0.6  8.3 14.6 6.5 32.2 7.1    Sc
## 3      France 10.8 0.8 27.5 0.9  8.9 16.8 6.0 22.6 5.7     W
## 4 W. Germany  6.7 1.3 35.8 0.9  7.3 14.4 5.0 22.3 6.1     W
## 5     Ireland 23.2 1.0 20.7 1.3  7.5 16.8 2.8 20.8 6.1     W
## 6       Italy 15.9 0.6 27.6 0.5 10.0 18.1 1.6 20.1 5.7     S
```

2)

A tree map of the data is plotted where the rectangles are sorted after group, size is given by percentage employed in service areas and color shows percentage employed in finance areas.

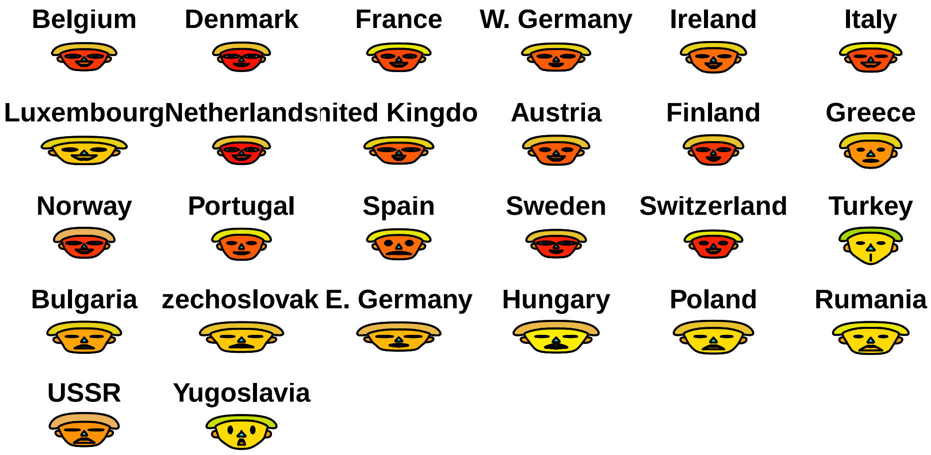


The results given by the tree map seem to be very reasonable. Finance and service are in general more common areas to be working within in western Europe and in Scandinavia compared to in eastern Europe.

This is especially true for the finance sector which, except for in Yugoslavia, is employing a very low percentage of the workers in the eastern Europe. Looking at the countries in the southern Europe it can be seen that the service industry is big in Italy while finance jobs is more common in Spain. The results for the southern countries are also thought to be reasonable.

3)

Chernoff faces are produced and all the quantitative variables are used. Since the use of Chernoff faces demands 15 variables and the data set just contains 9 variables, 6 extra variables are added to the data set. For all observations is the value for these extra variables set to 1. Hence, the new variables will have no effect on the results since the parts of the face they control will be the same for all observations.



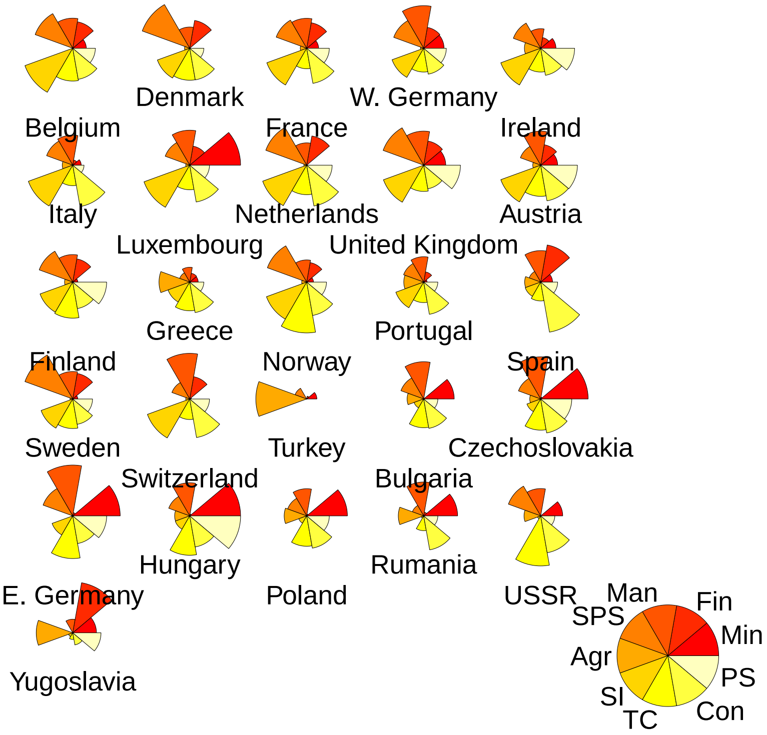
The Chernoff faces can be used for analysing which countries that are similar and if there are any outliers. It can easily be seen that the eastern countreis are fairly similar. The exception is once again Yugoslavia which looks like an outlier, and so does also Turkey. For the other groups of countries are the differencies inside the groups relatively small. The exceptions are Luxembourg who looks like an outlier and Netherlands who appear to be more similar the Scandinavian countries than the other Western countries.

4)

The variables are reordered by using single link hierarchical clustering and the columns in the data set are now ordered as follows.

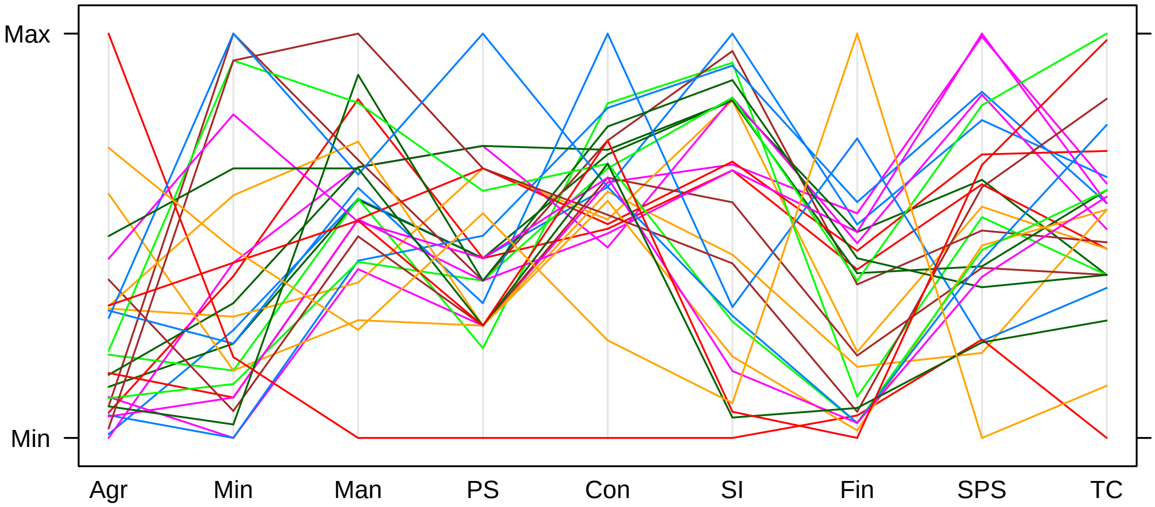
```
##   Min Fin  Man  SPS Agr   SI  TC Con  PS
## 1 0.9 6.2 27.6 26.6 3.3 19.1 7.2 8.2 0.9
```

Then, segment charts are produced by using the reordered data and all the quantitative variables in it.



Yugoslavia, Turkey and perhaps also Luxembourg still looks like outliers. The interpretation of the segment charts is pretty much the same as for the Chernoff faces. Mainly, two different groups can be seen where the first consists of the Scandinavian and western countries and the second group consists of the major part of the eastern countries.

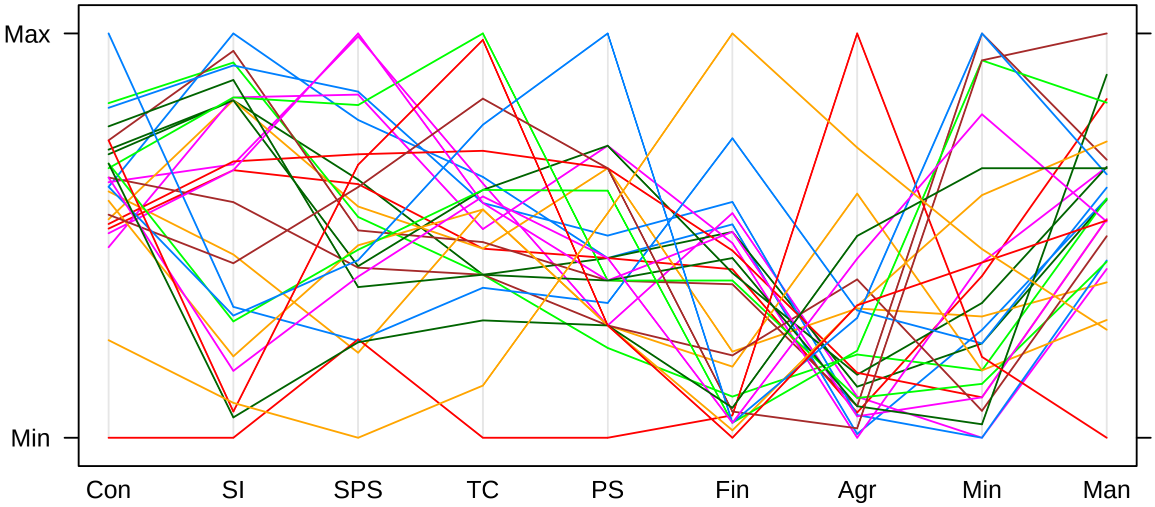
5)



It is hard to detect any clear clusters by looking at the parallel plot. It is possible that one can be seen as a bunch of the lines takes similar values for the variables *PS* and *Con*, which also are the variables that perhaps could be correlated. Regarding outliers one is clear, the red line which has low values for the variables *Man*, *PS*, *Con* and *SI*.

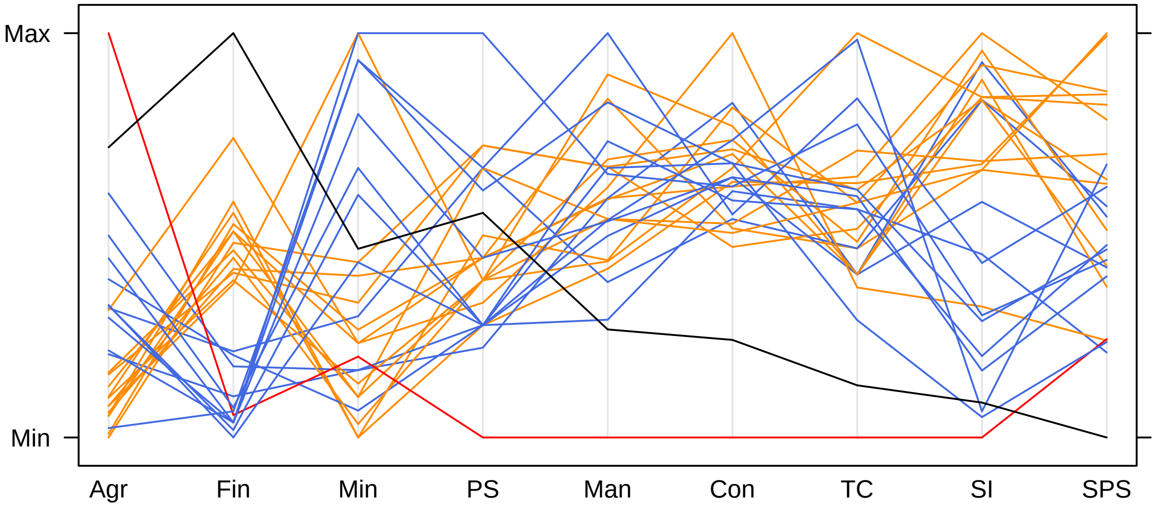
6)

The instructions given in the lab are performed and a parallel plot with the permuted columns is created.



Compared to the parallel plot in *5)* I find it easier to distinguish clusters in the plot above. By looking at for example *SI* and *Fin* it appears to be two different clusters in the plot.

7)



The two clusters found are mainly given by the variable *Fin*, but also by the variables *SI* and *Arg*. Two outliers are found where the one coloured in red has very low values for the variables *PS*, *Man*, *Con*, *TC* and *SI*. The second outlier has a very high value for *Fin* and very low for *SPS*, *TC* and *Con*.
The first cluster, coloured in blue, consists of the countries which are included in the table below. In this group are all the eastern countries but Yugoslavia and all southern countries but Spain and Turkey. There is also one of the counries in the western region, Ireland.

```
##           Country group
## 5         Ireland     W
## 6           Italy     S
## 12         Greece     S
## 14       Portugal     S
## 19       Bulgaria     E
## 20 Czechoslovakia     E
## 21     E. Germany     E
## 22        Hungary     E
## 23         Poland     E
## 24        Rumania     E
## 25           USSR     E
```

The second cluster, coloured in orange, are presented in the next table. In this cluster are all of the Scandinavian countries and all the western except Ireland. Spain is the only of the southern countries in this cluster.

```
##           Country group
## 1         Belgium     W
## 2         Denmark    Sc
## 3          France     W
## 4      W. Germany     W
## 7      Luxembourg     W
## 8     Netherlands     W
## 9  United Kingdom     W
## 10        Austria     W
## 11        Finland    Sc
## 13         Norway    Sc
## 15          Spain     S
## 16         Sweden    Sc
## 17    Switzerland     W
```

The table with the two outliers is shown below. This table contains the last southern country, Turkey, and the only eastern country that not was in the first cluster, Yugoslavia.

```
##       Country group
## 18     Turkey     S
## 26 Yugoslavia     E
```

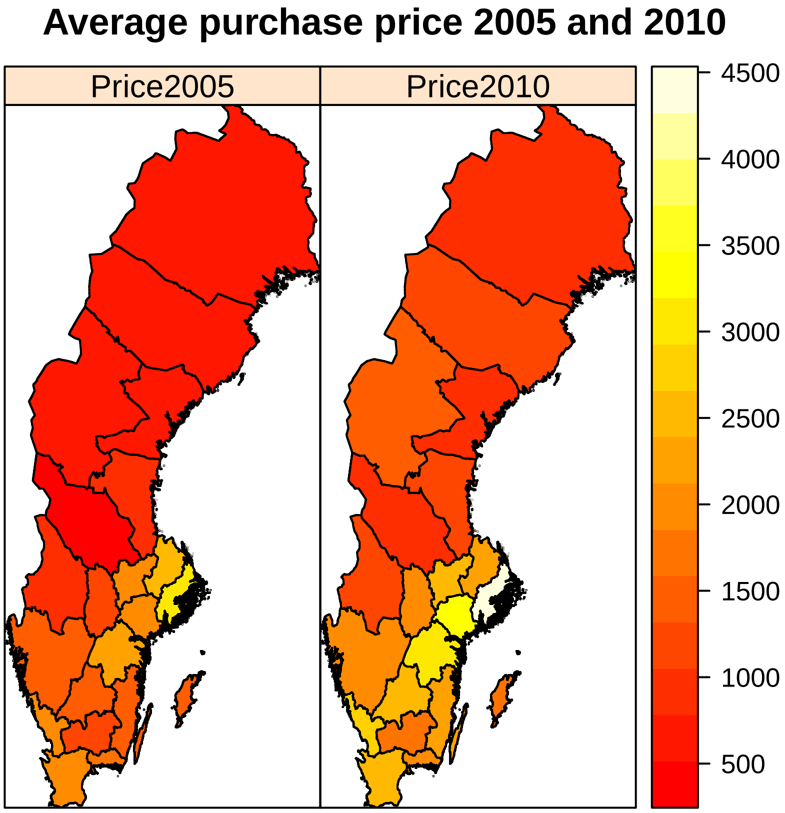
In general, the clusters found are quite well related to the group variable. As mentioned above, most of the eastern countires are in one cluster and most, or all, of the western and Scandinavian countries are in another cluster. It is only for the southern countries a weak relation between the clusters and the group variable can be seen.
An interpretation of these results are that countries in different regions of Europe are quite similar to each other regarding which sectors the inhabitants are employeed in. In eastern Europe more inhabitants are employeed in the agriculture industry and less in the finance industry. The finance industry and the service industry are in general employing a higher percentage of the inhabitants in the western and Scandinavian countries.

8)

The last parallel plot, the one in *7)*, was by far the easiest one to analyze. First and foremost thanks to the use of colours to indicate clusters which helps a lot. In the first two parallel plots there were a lot of different colours and lines which made it hard to see patterns and to analyze the plot. Regarding the clusters found by the parallel coordinates so are they very much the same as the clusters identified by looking at the chernoff faces and the segment charts.

Assignment 2

1-3)



From 2005 to 2010 the prices has increased in all regions. The same pattern can be seen in the plots for both years with the lowest average prices found for the northern regions. The, on average, most expensive region is Stockholm and thereafter the regions south of Stockholm. In general are the regions with higher prices those who are most populated, for example Stockholm and Skåne.

R code

```r
## ---- echo=FALSE-------------------------------------------------------------
## 1)
jobs <- read.table(file = "C:\\Users\\Gustav\\Documents\\Visualization\\jobs.txt",
    header = TRUE, sep = "\t")
jobs$group <- 0
jobs[c(1, 3, 4, 5, 7, 8, 9, 10, 17), 11] <- "W"
jobs[c(2, 11, 13, 16), 11] <- "Sc"
jobs[c(6, 12, 14, 15, 18), 11] <- "S"
jobs[c(19, 20, 21, 22, 23, 24, 25, 26), 11] <- "E"
head(jobs)

## ---- echo=FALSE, fig.height=3.75, fig.width=6, warning=FALSE,
## message=FALSE, fig.align='center'---- 2)
library(portfolio)
map.market(id = jobs$Country, area = jobs$SI, group = jobs$group, color = jobs$Fin,
    main = "Jobs in service(area) and finance(color)", lab = c(F, T))


## ---- echo=FALSE, message=FALSE----------------------------------------------
## 3)
library(aplpack)
jobs[, 12:17] <- 1
par(mar = c(1.1, 4.1, 4.1, 2.1))
faces(as.matrix(jobs[, c(2:10, 12:17)]), labels = jobs$Country, print.info = FALSE)

## ---- echo=FALSE, warning=FALSE, message=FALSE--------------------------
library(gclus)
library(graphics)
jobs <- jobs[, 1:11]
d <- dist(t(scale(jobs[, 2:10])))
jobs_order <- order.single(d)
jobs2 <- jobs[, jobs_order + 1]
head(jobs2, 1)

## ---- echo=FALSE-------------------------------------------------------------
palette(heat.colors(9))
stars(jobs2, labels = as.character(jobs$Country), draw.segments = TRUE, key.labels = names(jobs2),
    key.loc = c(14, 1.85))


## ---- echo=FALSE, fig.height=3.5, fig.width=7, fig.align='center'--------
## 5)
parallelplot(jobs[, 2:10], horizontal.axis = FALSE)

## ---- echo=FALSE, fig.height=3.5, fig.width=7, fig.align='center',
## warning=FALSE---- 6)
set.seed(123445)
## a)
job_cor <- 1 - cor(jobs[, 2:10], )
## b)
library(TSP)
```

```
res <- solve_TSP(TSP(job_cor))
jobs2 <- jobs[, as.integer(res) + 1]
## c)
parallelplot(jobs2, horizontal.axis = FALSE)

## ---- echo=FALSE, message=FALSE, warning=FALSE, fig.height=3.5,
## fig.width=7, fig.align='center'----
library(seriation)
res2 <- seriate(as.dist(job_cor), method = "HC")
jobs_order2 <- get_order(res2)
jobs3 <- jobs[, jobs_order2 + 1]

color = 1 + (jobs3$Fin - min(jobs3$Fin) > 0.3 * (max(jobs3$Fin) - min(jobs3$Fin)))
color[18] <- 3
color[26] <- 4
color[color == 1] <- "royalblue"
color[color == 2] <- "darkorange"
color[color == 3] <- "red"
color[color == 4] <- "black"
parallelplot(jobs3, horizontal.axis = FALSE, col = color)

## ---- echo=FALSE---------------------------------------------------------
group = 1 + (jobs3$Fin - min(jobs3$Fin) > 0.3 * (max(jobs3$Fin) - min(jobs3$Fin)))
group[18] <- 3
group[26] <- 4
jobs$group2 <- group
subset(jobs, jobs$group2 == 1)[, c(1, 11)]

## ---- echo=FALSE---------------------------------------------------------
subset(jobs, jobs$group2 == 2)[, c(1, 11)]

## ---- echo=FALSE---------------------------------------------------------
subset(jobs, jobs$group2 == 3 | jobs$group2 == 4)[, c(1, 11)]

## ---- echo=FALSE, message=FALSE, warning=FALSE---------------------------
houseP <- read.csv("C:\\Users\\Gustav\\Documents\\Visualization\\prices_0510.csv",
    sep = ";")

map1 <- readRDS("C:\\Users\\Gustav\\Downloads\\SWE_adm1.rds")

temp <- map1@data
new <- merge(temp, houseP, by.x = "NAME_1", by.y = "County", all.y = F, all.x = T,
    sort = FALSE)

map1@data$Price2005 <- new$X2005
map1@data$Price2010 <- new$X2010

spplot(map1, zcol = c("Price2005", "Price2010"), main = "Average purchase price 2005 and 2010",
    col.regions = heat.colors(16))


## ----code=readLines(knitr::purl('C:\\Users\\Gustav\\Documents\\Visualization\\Lab2_viz.Rmd',documenta
## = 1)), eval = FALSE, tidy=TRUE----
```