Lab 5

Gustav Sternelöv

October 6, 2016

Assignment 1

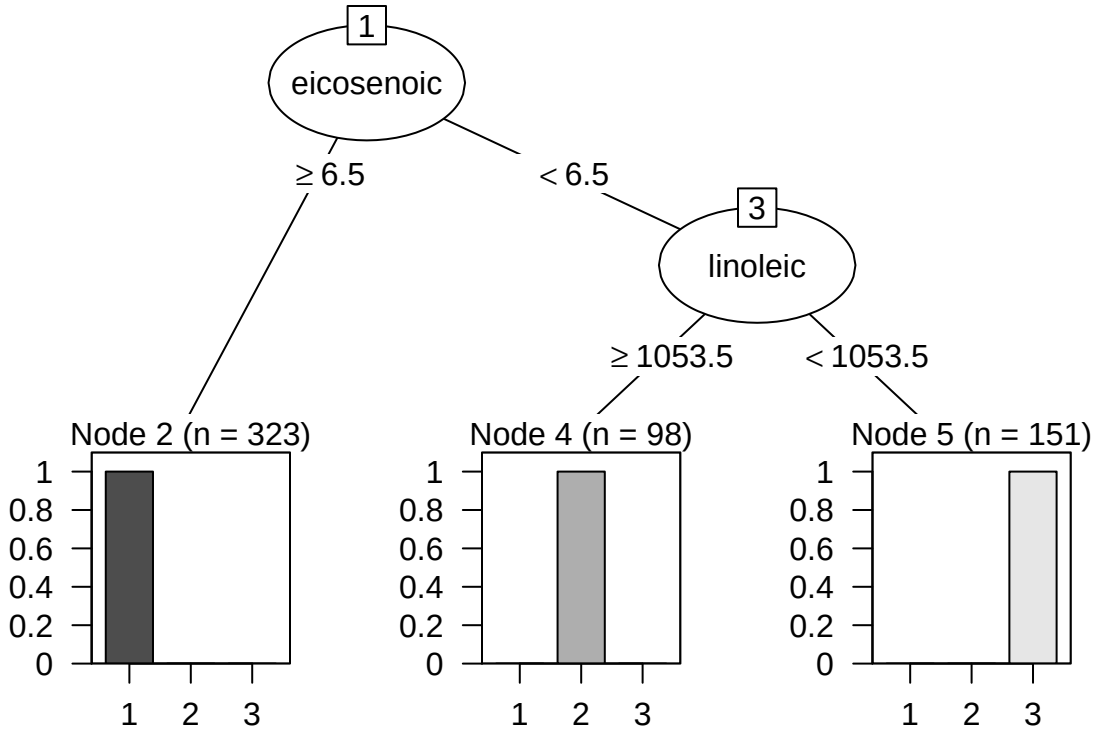
1.1

The data set *olive* is imported and the variable *Region* is converted to a factor with the follwing code.

```
olive <- read.csv("C:\\Users\\Gustav\\Documents\\Visualization\\Lab5\\olive.csv")
olive$Region <- as.factor(olive$Region)
```

1.2

A decision tree is fitted with region as target and all acids as inputs. The output of the model is presented graphically.



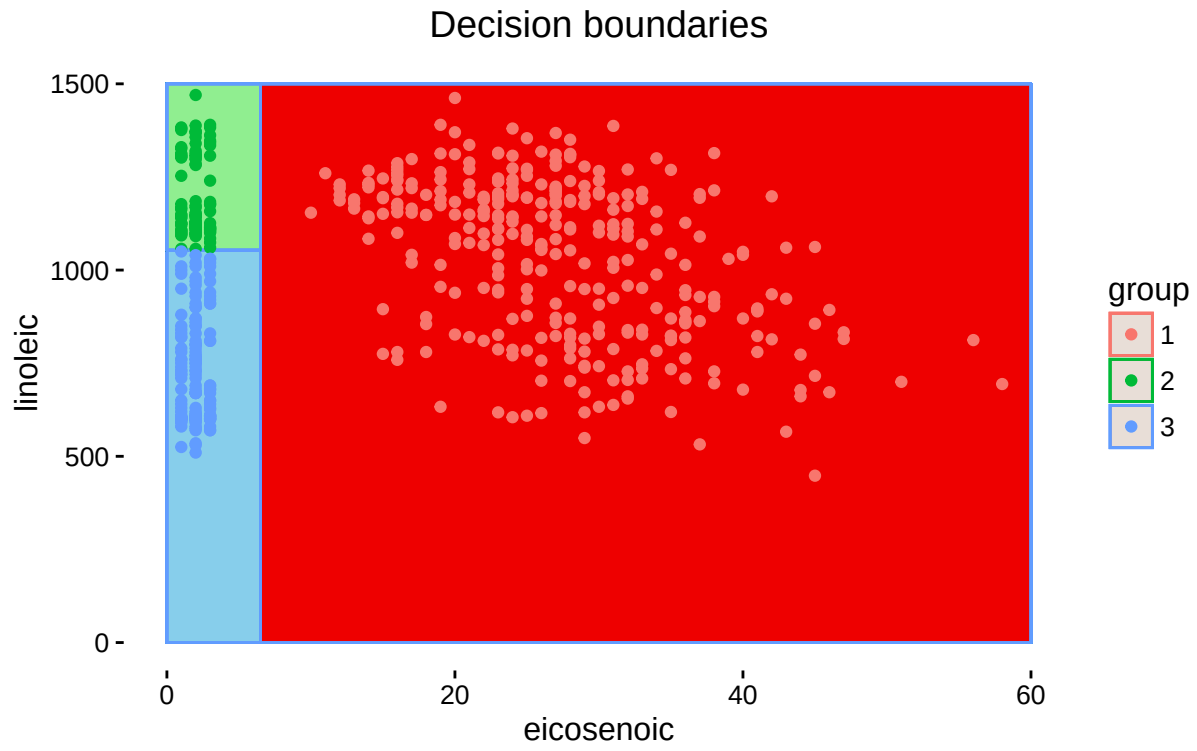
The two acids selected for decision making is eicosenoic and linoleic. If the former acid is higher than or equal to 6.5 the observation is classified to the region 1. A eicosenoic value lower 6.5 and a linoleic value higher than or equal to 1053.5 is connected to region 2. If the linoleic value instead is lower than 1053.5 the tree predicts the observation to come from region 3. The depth of the tree is equal to 2.

The fitted model is applied on the data and the results is presented in a confusion matrix. 0 observations are misclassified by the model.

```
##
##        1   2   3
##    1 323   0   0
##    2   0  98   0
##    3   0   0 151
```

1.3

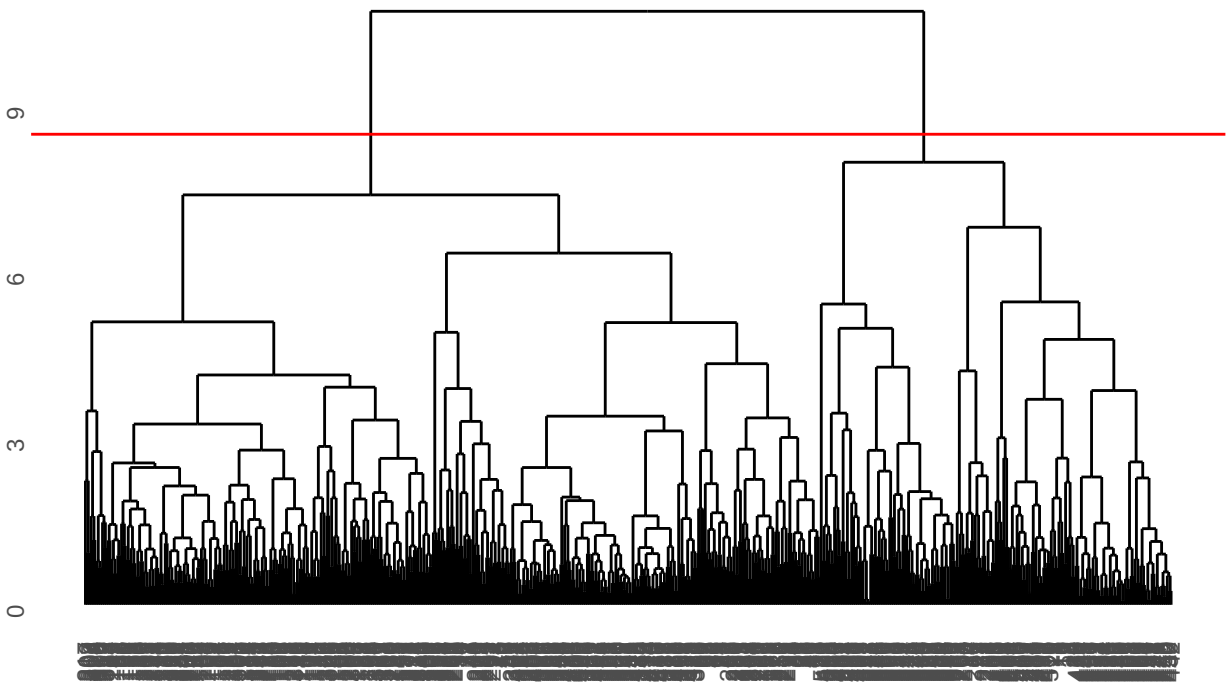
The most reasonable choice of acids for demonstrating the classification results is eicosenoic and linoleic. These two variables are the ones with impact on the decisions made by the fitted model.



It can be problematic to use this plot for detecting the three regions in terms of clustering as the blue and green points probably not would be separated. Instead, out of this plot, the clustering probably would yield two clusters.

1.4

The data is scaled and a hierarchical clustering with complete link is performed.



A dendrogram is produced and it can be seen that it is reasonable to say that there are 2 natural clusters in data. However, as given in the exercise are 3 clusters created. *Rggobi* is used for creating a 2D-tour in order to look at acid contents in the oils for different clusters.

Assignment 2

2.1-2

A motion chart is created by using the package *googleVis* and the following code.

```
Oilcoal$Year <- as.numeric(Oilcoal$Year)
mCh <- gvisMotionChart(Oilcoal, idvar="Country", timevar="Year")
plot(mCh)
```

Some interesting patterns are observed in the motion chart and snapshots of these interesting parts are presented below.

Figure 1 shows how Chinas consumption of oil and coal has evolved from 2002 to 2009. The most remarkable feature of the picture is how fast the consumption of coal increases during these years. This is an effect of the rapid development of the economy in china that was present during this period. As the economy developes the demand for energy increases and coal is a cheap energy source.

Figure 2 shows how the oil consumption decreased in the US from the late 70's until the middle of the 80's. This was probably connected to the oil crisis in the world that took place in that period. The crisis was an effect of the increased price on petroluem, which in turn was an effect of the disturbances in the Middle East.

The last chart in this section, figuer 3, again focuses on the US and this time on the decrease in oil consumption from 2007 to 2009. This could perhaps have been an effect of the financial crisis in the US which started in 2007 and lasted for a couple of years.

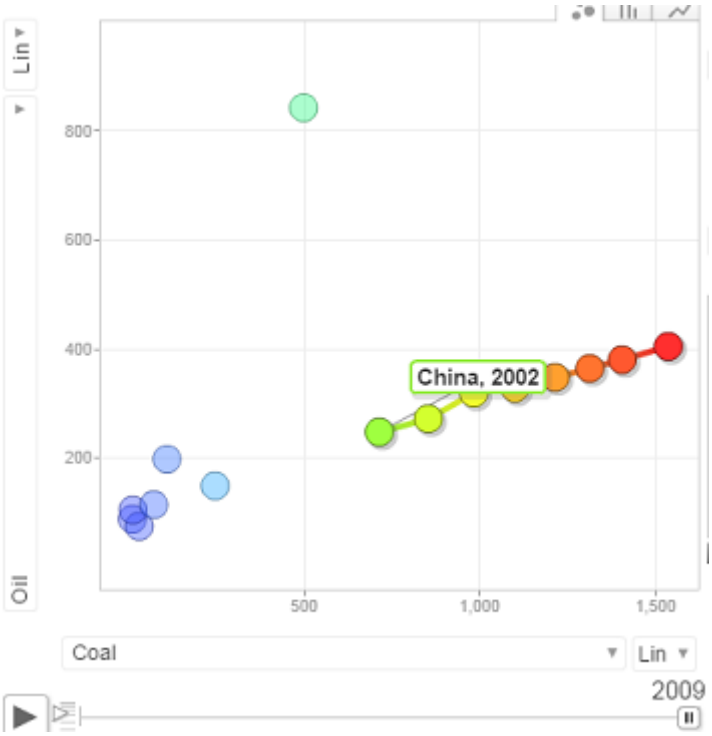


Figure 1:

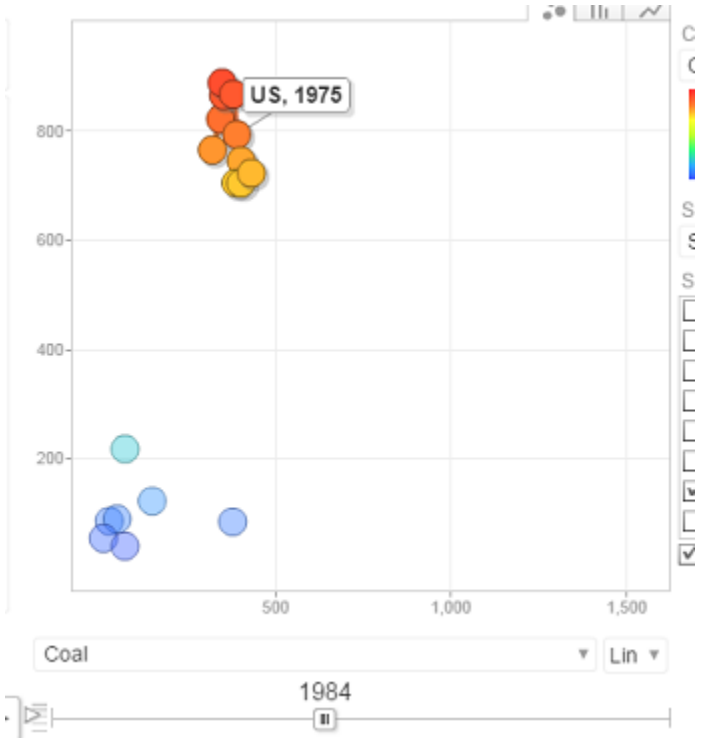


Figure 2:

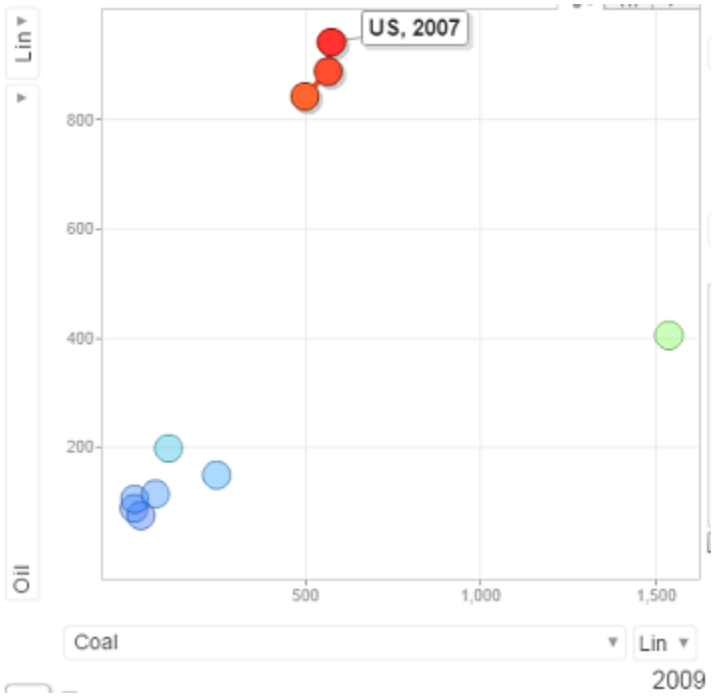


Figure 3:

2.3

Two countries that had similar patterns were France and the United Kingdom. A trace plot with their consumption over time is therefore created.

It is a bit hard to see the development during time on this snapshot, but if the whole motion chart is viewed it is easier to see that patterns are similar.

2.4

When watching the whole motion chart I think it in some sense is more informative than a combination of time series plots per country. This since with the motion chart it is possible to see the can see the development for both oil and coal simultaneously. Hence, the consumption of each energy source and development over time becomes more apparent in a motion chart.

2.5

The spline model with Oil_p as response and year and country as predictors is fitted.

2.6

Three extra data points is inserted for each year. So, for example, the year 1965 now has the data points 1965, 1965.25, 1965.50 and 1965.75. The *predict.Krig* function is used for making predictions at each time point and a loop that produces a stacked bar chart at each time point is written.

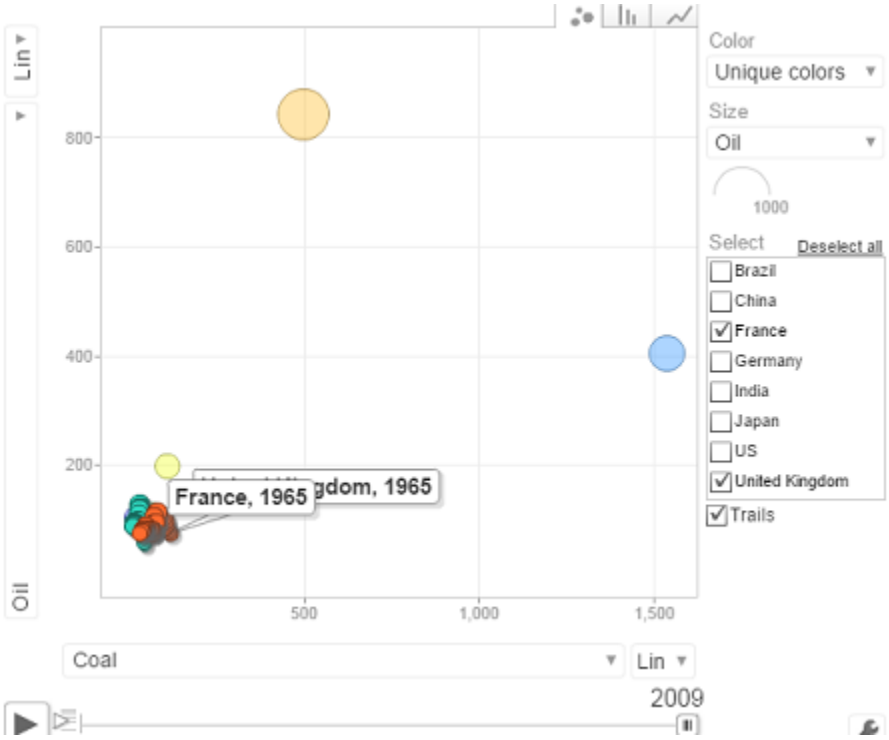


Figure 4:

```
for(i in 1:180){
    print(ggplot(extFrame[c((1:8)+8*(i-1), 1441:1448+8*(i-1)),], aes(y=OilP, x=Country,
  fill=ConsumptionP)) +  geom_bar(stat="identity") + theme_classic() +
  scale_y_continuous(labels = percent)+scale_fill_manual(values=c("royalblue","darkorange"))+
    scale_x_discrete(labels=c("1"= "Brazil" ,"2"= "China" ,"3"= "France" ,"4"= "Germany" ,
                              "5"= "India" ,"6"= "Japan" ,"7"= "United Kingdom" ,"8"= "US"))+
    labs(y="Percentage of\nconsumption", title=paste("Year:",AllFrame[(1)+8*(i-1),2])) +
    theme(axis.title.y = element_text(angle=0),axis.text.x = element_text(angle=15))
    )
  }
```

2.7

The advantage with this visualization is that the proportional use of each energy source becomes clearer. For example, it can be seen that the percentage of coal energy in China has been fairly constant. This can be compared to the motion chart where the increase of coal consumption during the 2000's were apparent and the increase in oil consumption not were noticed at all, at least not by me.

A disadvantage with this visualization is that it not gives any information about the magnitude of the consumption.

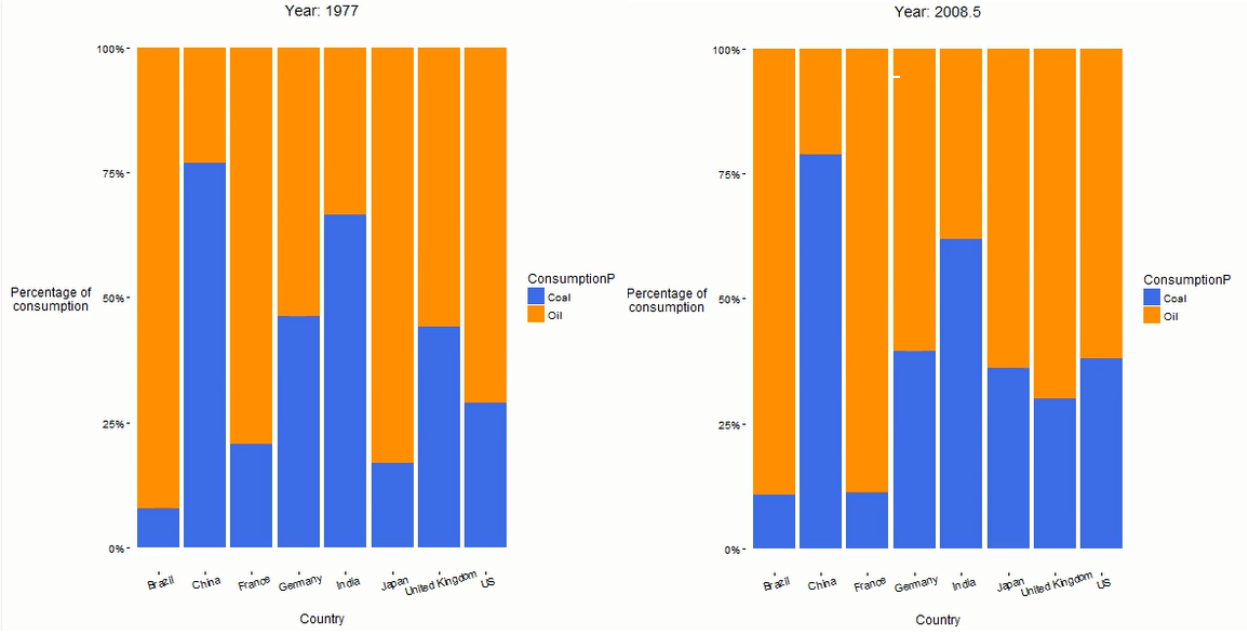


Figure 5:

R-code

```
## ---- echo=FALSE, warning=FALSE, message=FALSE--------------------------
## Assignment 1
library(rpart)
library(partykit)
library(ggplot2)
library(plyr)
library(scales)
library(fields)
library(animation)
library(ggdendro)

## ----------------------------------------------------------------------
olive <- read.csv("C:\\Users\\Gustav\\Documents\\Visualization\\Lab5\\olive.csv")
olive$Region <- as.factor(olive$Region)

## ---- echo=FALSE, fig.height=4------------------------------------------
dT <- rpart(Region ~ palmitic + palmitoleic + stearic + oleic + linoleic + linolenic +
    arachidic + eicosenoic, data = olive)

plot(as.party(dT))

## ---- echo=FALSE--------------------------------------------------------
classific <- predict(dT, olive[, 4:11])
classific <- data.frame(pred = c(classific[1:323, 1], classific[324:421, 2],
    classific[422:572, 3]))
classific[324:421, ] <- 2
classific[422:572, ] <- 3
```

```
table(classific[, 1], olive$Region)

## ---- echo=FALSE, fig.height=4-------------------------------------------
olive$group <- classific[, 1]
olive$group <- as.factor(olive$group)
ggplot(olive, aes(x = eicosenoic, y = linoleic, col = group)) + geom_rect(aes(xmin = 0,
    xmax = 6.5, ymin = 0, ymax = 1053.5), fill = "skyblue", alpha = 0.1) + geom_rect(aes(xmin = 0,
    xmax = 6.5, ymin = 1053.5, ymax = 1500), fill = "lightgreen", alpha = 0.1) +
    geom_rect(aes(xmin = 6.5, xmax = 60, ymin = 0, ymax = 1500), fill = "red2",
        alpha = 0.1) + geom_point() + theme_classic() + ggtitle("Decision boundaries")

## ---- echo=FALSE, fig.height=4-------------------------------------------
scaleOlive <- scale(olive[, 4:11])
compHc <- hclust(dist(scaleOlive), method = "complete")
# plot(compHc)
ggdendrogram(compHc, rotate = FALSE, size = 2) + geom_hline(yintercept = 8.5,
    col = "red")

## ---- echo=FALSE, warning=FALSE, message=FALSE---------------------------
## Assignment 2
library(googleVis)
library(XLConnect)
wb = loadWorkbook("C:\\Users\\Gustav\\Documents\\Visualization\\Lab5\\Oilcoal.xls")
Oilcoal = readWorksheet(wb, sheet = "Sheet2", header = TRUE)
# 2.1

## ---- message=FALSE------------------------------------------------------
Oilcoal$Year <- as.numeric(Oilcoal$Year)
mCh <- gvisMotionChart(Oilcoal, idvar = "Country", timevar = "Year")
plot(mCh)

## ---- echo=FALSE--------------------------------------------------------
## 2.5
Oilcoal$OilP <- Oilcoal$Oil/(Oilcoal$Oil + Oilcoal$Coal)
Oilcoal$Country <- as.factor(Oilcoal$Country)
Oilcoal$Country <- seq_along(levels(Oilcoal$Country))[Oilcoal$Country]

splineM <- Tps(x = as.matrix(Oilcoal[, 1:2]), Y = Oilcoal$OilP)

## ---- eval=FALSE--------------------------------------------------------
## for(i in 1:180){ print(ggplot(extFrame[c((1:8)+8*(i-1),
## 1441:1448+8*(i-1)),], aes(y=OilP, x=Country, fill=ConsumptionP)) +
## geom_bar(stat='identity') + theme_classic() + scale_y_continuous(labels =
## percent)+scale_fill_manual(values=c('royalblue','darkorange'))+
## scale_x_discrete(labels=c('1'= 'Brazil' ,'2'= 'China' ,'3'= 'France' ,'4'=
## 'Germany' , '5'= 'India' ,'6'= 'Japan' ,'7'= 'United Kingdom' ,'8'=
## 'US'))+ labs(y='Percentage of\nconsumption',
## title=paste('Year:',AllFrame[(1)+8*(i-1),2])) + theme(axis.title.y =
## element_text(angle=0),axis.text.x = element_text(angle=15)) ) }

## ---- echo=FALSE, eval=FALSE-------------------------------------------- #
## 2.6 - 2.7 k <- 0 yearSeq <- seq(0,0.75,0.25) extFrame <-
## data.frame(Country=rep(1:8, each=180), Year=0) for(i in 1:360){ for(j in
```

```
## 1:4){ k <- 1+k extFrame[k,2] <- Oilcoal$Year[i] + yearSeq[j] } }
## extFrame$OilP <- predict.Krig(object = splineM, x = as.matrix(extFrame))
## extFrame <- rbind(extFrame, data.frame(Country=extFrame[,1], Year=
## extFrame[,2], OilP= as.vector(1 - extFrame$OilP))) extFrame$ConsumptionP
## <- rep(c('Oil', 'Coal'), each=1440) extFrame$Country <-
## as.factor(extFrame$Country) extFrame <- AllFrame[ order(AllFrame[,4],
## AllFrame[,2]), ] ani.options(ffmpeg='C:\\Program
## Files\\ImageMagick-7.0.3-Q16\\ffmpeg.exe') saveVideo({ for(i in
## 1:180){ print(ggplot(extFrame[c((1:8)+8*(i-1), 1441:1448+8*(i-1)),],
## aes(y=OilP, x=Country, fill=ConsumptionP)) + geom_bar(stat='identity') +
## theme_classic() + scale_y_continuous(labels =
## percent)+scale_fill_manual(values=c('royalblue','darkorange'))+
## scale_x_discrete(labels=c('1'= 'Brazil' ,'2'= 'China' ,'3'= 'France' ,'4'=
## 'Germany' , '5'= 'India' ,'6'= 'Japan' ,'7'= 'United Kingdom' ,'8'=
## 'US'))+ labs(y='Percentage of\nconsumption',
## title=paste('Year:',AllFrame[(1)+8*(i-1),2])) + theme(axis.title.y =
## element_text(angle=0),axis.text.x = element_text(angle=15)) ) }
## },video.name='C:\\Users\\Gustav\\Documents\\Visualization\\Lab5\\stackedBars.mp4',
## interval=0.1, ani.width=600,ani.height=600 )

## ----code=readLines(knitr::purl('C:\\Users\\Gustav\\Documents\\Visualization\\Lab5\\Lab_5.Rmd',documer
## = 1)), eval = FALSE, tidy=TRUE---- NA
```