World's Life Expectancy and Fertility Prediction

MACHINE LEARNING ALGORITHMS IN R

Abstract

Predict Life Expectancy and Fertility Rates using machine learning algorithms through R programming.

Sara Eugenia Rodríguez Reyes

sara.eug.rod.rey@gmail.com

Introduction

We have data of the World Bank of life expectancy and fertility rates of several countries around the world from 1960 to 2013.

We would like to predict how does this data will behave in the near future. To be able to achieve that, we need to organize the data in a way that it is easy to handle in order to perform an exploratory analysis, find patterns in the data and finally make a prediction of the behavior of the information.

The Data

Includes:

1. Average fertility data set (258 observations from different regions through 53 years)
2. Life expectancy data set (258 observations from different regions through 53 years)

Each Split by:

1. Zones (Extracted them from *data.worldbank.org*)
2. Countries

Note: consider the possibility that the information may or not be organized as it would normally do. The data sets are incomplete and a bit damaged.

Let's assume that the world has certain average $WA_{(X,t)}$, from the quantity X on time t, and that the world population can be divided into the next regions:

1. Sub-Saharan_Africa_(IFC_classification),
2. Europe_and_Central_Asia_(IFC_classification),
3. Latin_America_and_the_Caribbean_(IFC_classification),
   Middle_East_and_North_Africa_(IFC_classification),
4. East_Asia_and_the_Pacific_(IFC_classification),
5. North_America,

With averages, $SSA_{(X)}$, $ECA_{(X)}$, $LAC_{(X)}$, $MEA_{(X)}$, $EAP_{(X)}$, $NAM_{(X)}$ respectively.

So, the $WA_{(X,t)}$ average can be written in the following way:

$$WA_{(X,t)} = ssa * SA_{(X,t)} + eca * ECA_{(X,t)} + lac * LAC_{(X,t)} + mea * MEA_{(X,t)} + eap * EAP_{(X,t)} + nam * NAM_{(X,t)}$$

Where the lowercase variables represent the proportional world population that lives in certain regions and assuming those values are constants in time.

The Goal

Use the data to fit the values of the constants mentioned above using machine learning algorithms in order to predict a world life expectancy and fertility rate.

Also, test this ML models into many conclusions until (by comparison) a better or worse methodology is found explaining the reasons why it is or not the best way to describe the data.

Methodology

1. Data Cleaning: detecting and correcting (or removing) corrupt or inaccurate records from the data sets
2. Exploratory Analysis: analyzing the data sets to summarize their main characteristics, with visual methods.
3. Forecasting: making predictions of the future based on past and present data through machine learning algorithms
4. Conclusions

1. Data Cleaning

Many times, the data isn't ready to processing, that's why it's necessary to clean it. Selecting the data that has an important effect in the variable we want to predict and that they're not only noise.

1.1 Missing Values Analysis

These data sets are missing quite a lot of values, with variables ranging from 0-100% 'missingness'. Therefore making a prediction without taking action, may not be the best option and the results may get distorted.

The first approach would be to observe how many missing values are in each data set in order to decide whether it is convenient to eliminate those variables or use some method to replace them.

1.1.1 Missing Values by year

Since both data sets contain 258 observations, the year 2013 is deleted… it has no data.

Year	Fertility Missing Values	Life Expectancy Missing Values
2013	258	258

1.1.2 Missing Values by country/region

Those regions with 100% of missing values are being excluded, since we intend to use the model for prediction and those variables are inherently hard to get, then there's no sense including them in the model.

Variables being excluded:

- Mexico_and_Central_America
- Andorra
- Andean_Region
- American_Samoa
- Sub-Saharan_Africa_(IFC_classification)
- East_Asia_and_the_Pacific_(IFC_classification)
- Europe_and_Central_Asia_(IFC_classification)
- Latin_America_and_the_Caribbean_(IFC_classification)

- Middle_East_and_North_Africa_(IFC_classification)
- South_Asia_(IFC_classification)
- Cayman_Islands
- Not_classified
- Latin_America_and_the_Caribbean
- Monaco
- Northern_Mariana_Islands
- Southern_Cone_Extended
- Turks_and_Caicos_Islands
- Tuvalu

1.2 Missing Values Imputation

There are still variables with some missing values. However, we use "*missForest*" library which imputes missing values using Random Forest algorithm.

Once the algorithm is implemented in all of the remaining variables with missing values, there are no more NA's left.

1.3 Countries with no region

Previously, it was said that the world population is divided into 6 regions:

1. Sub Saharan Africa
2. Europe and Central Asia
3. Latin America and The Caribbean
4. Middle East and North Africa
5. East Asia and the Pacific
6. North America

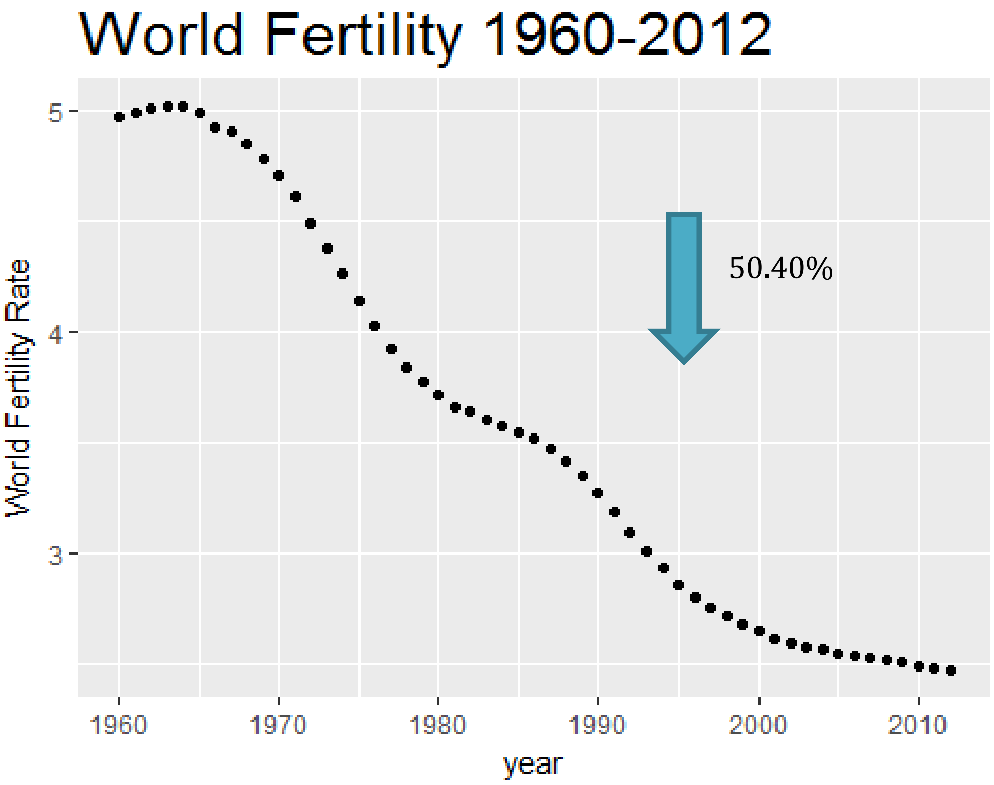
Doing some research *South Asia* region was missing, so it was added. Now the data is divided into 7 regions. However, there are still variables that doesn't belong to a region, so they are excluded for the moment.

2. Exploratory Analysis

2.1 Fertility Analysis

Fertility rate is the average number of children that would be born to a woman over her lifetime.

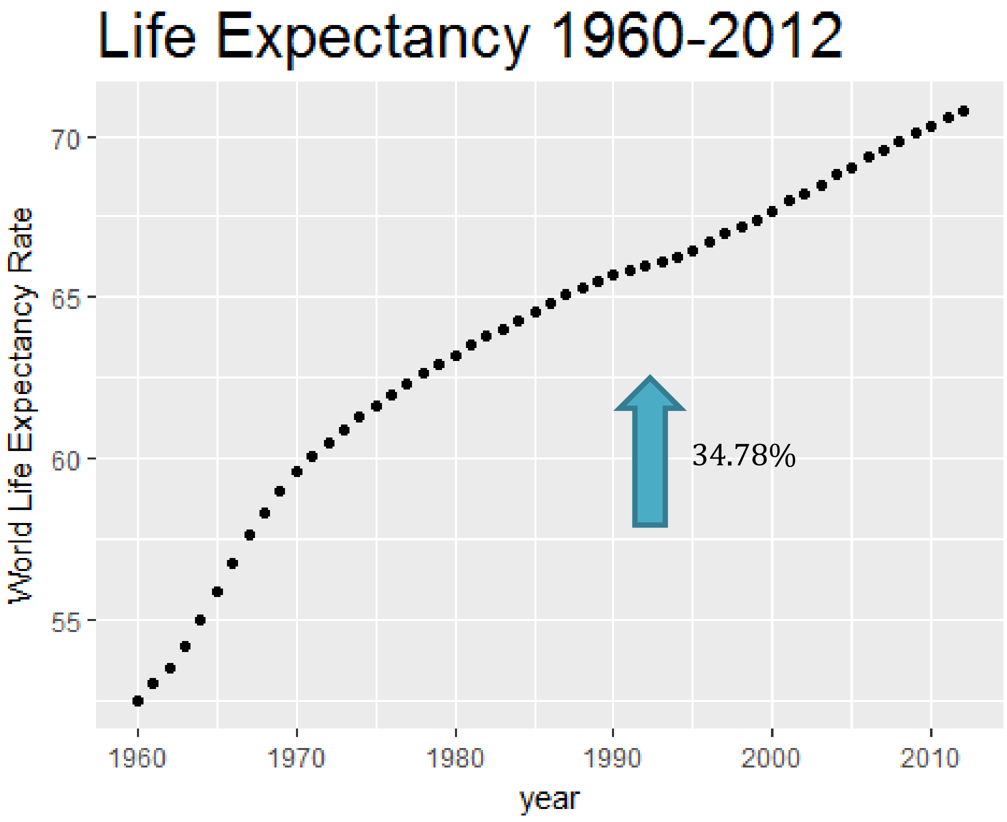
World's fertility rate has been **decreasing** through the years. It has fallen 50.40% since 1960 from 4.98 to 2.47



2.2 Life Expectancy Analysis

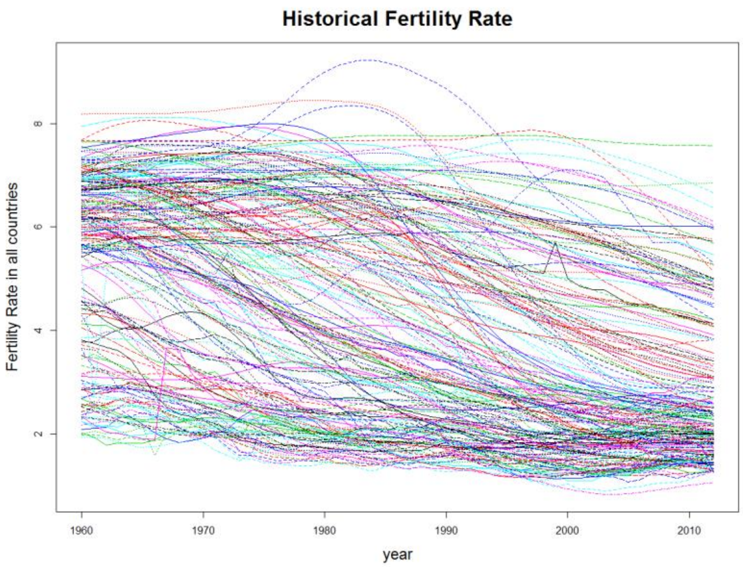
Life expectancy is the average time a person is expected to live.

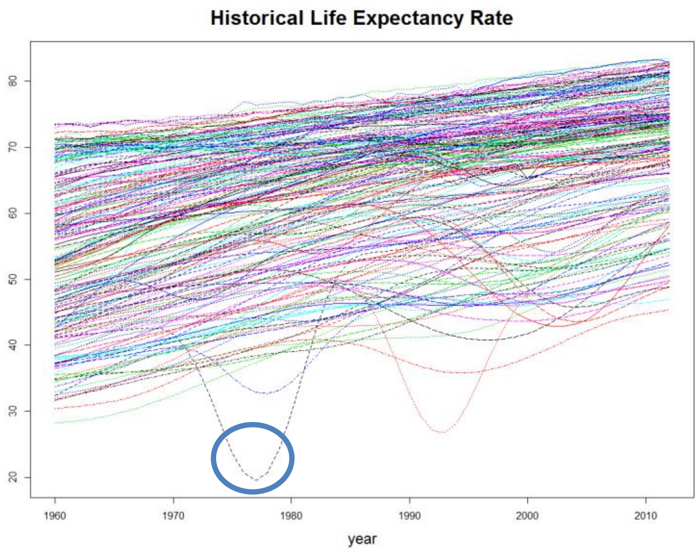
World's life expectancy has **increased** in a 34.78% through the years, from 52.48 to 70.78 years.



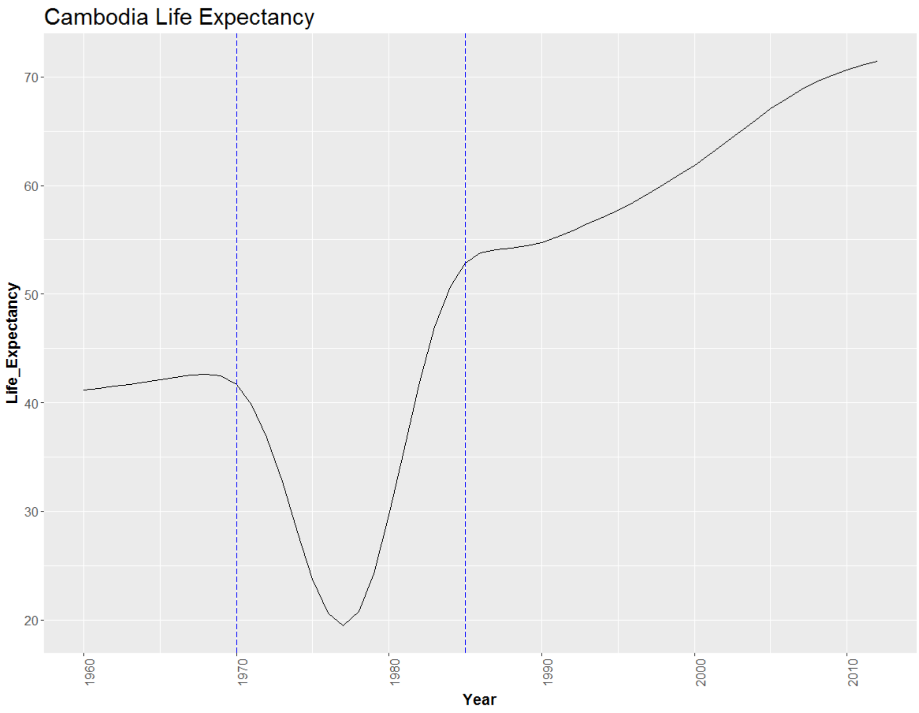
2.3 Comparison

Fertility rates have been diminishing through the years. To the contrary, life expectancy has been increasing.





Looking at the above graphic from the historical life expectancy, it is noticeable that there is a very low value between 1970 and 1980. This is related to the case of Cambodia which between those years it had an important Life Expectancy decrease.



2.4 Cambodia Case

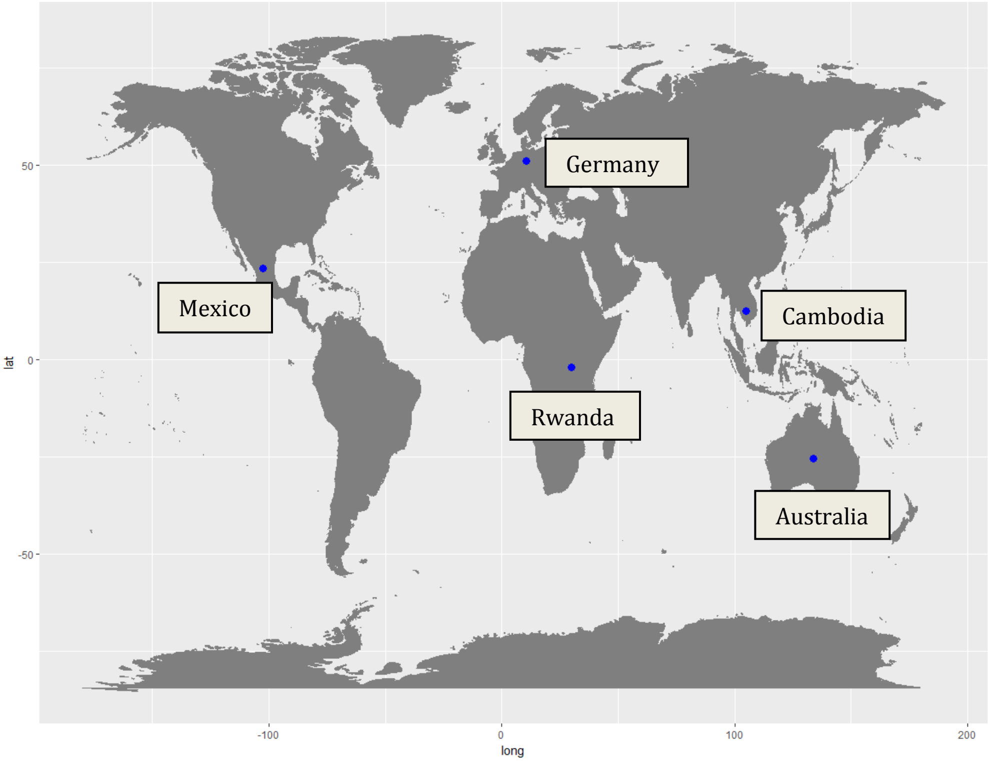
This country belongs to the East Asia and the Pacific region

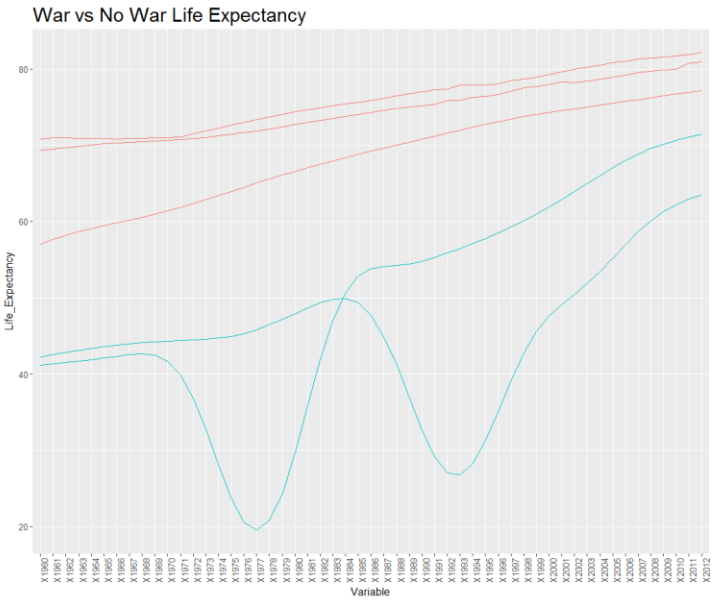
As best as can now be estimated, over two million Cambodians died during the 1970s because of the political events of the decade, the vast majority of them during the mere four years of the âKhmer Rouge â regime. This number of deaths is even more staggering when related to the size of the Cambodian population, then less than eight million. In my estimation, about a third of the 1970 population would have survived to the end of the decade under normal demographic conditions but did not under the circumstances that prevailed. No single factor alone explains the rare intensity of the Cambodian mortality crisis. Instead, the excess mortality pattern reflects one of the worst imaginable mixes of conditions, including war casualties, massive population displacement, ethnic cleansing, health system collapse, and famine.

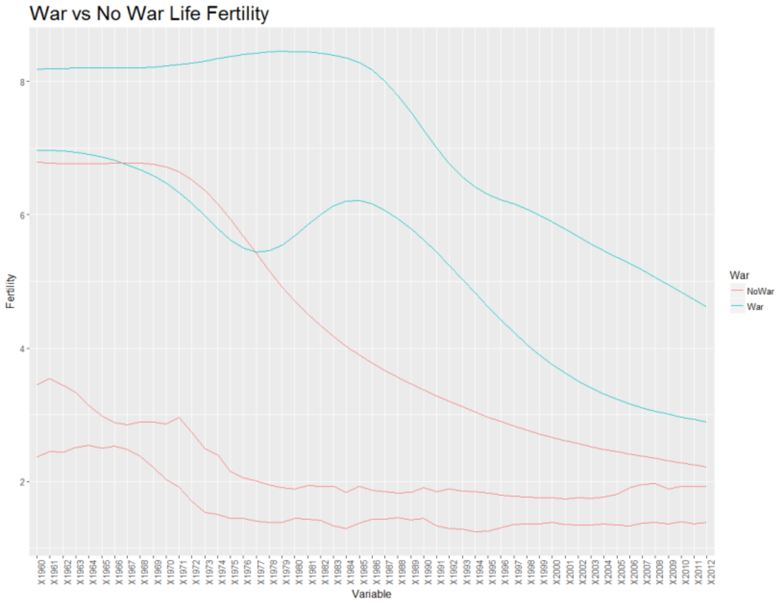
2.5 War vs. No War

Looking at the above case, it would be a good idea to dig into different countries situations from the ones who haven't been on war and the ones who already have been.

Five countries will be compared: Cambodia, Rwanda, Mexico, Germany and Australia







With this graphics, we can conclude that war does affect countries life expectancy, it also reduces fertility; but it doesn't mean that countries who live more "peacefully" have both increments in fertility and life expectancy.

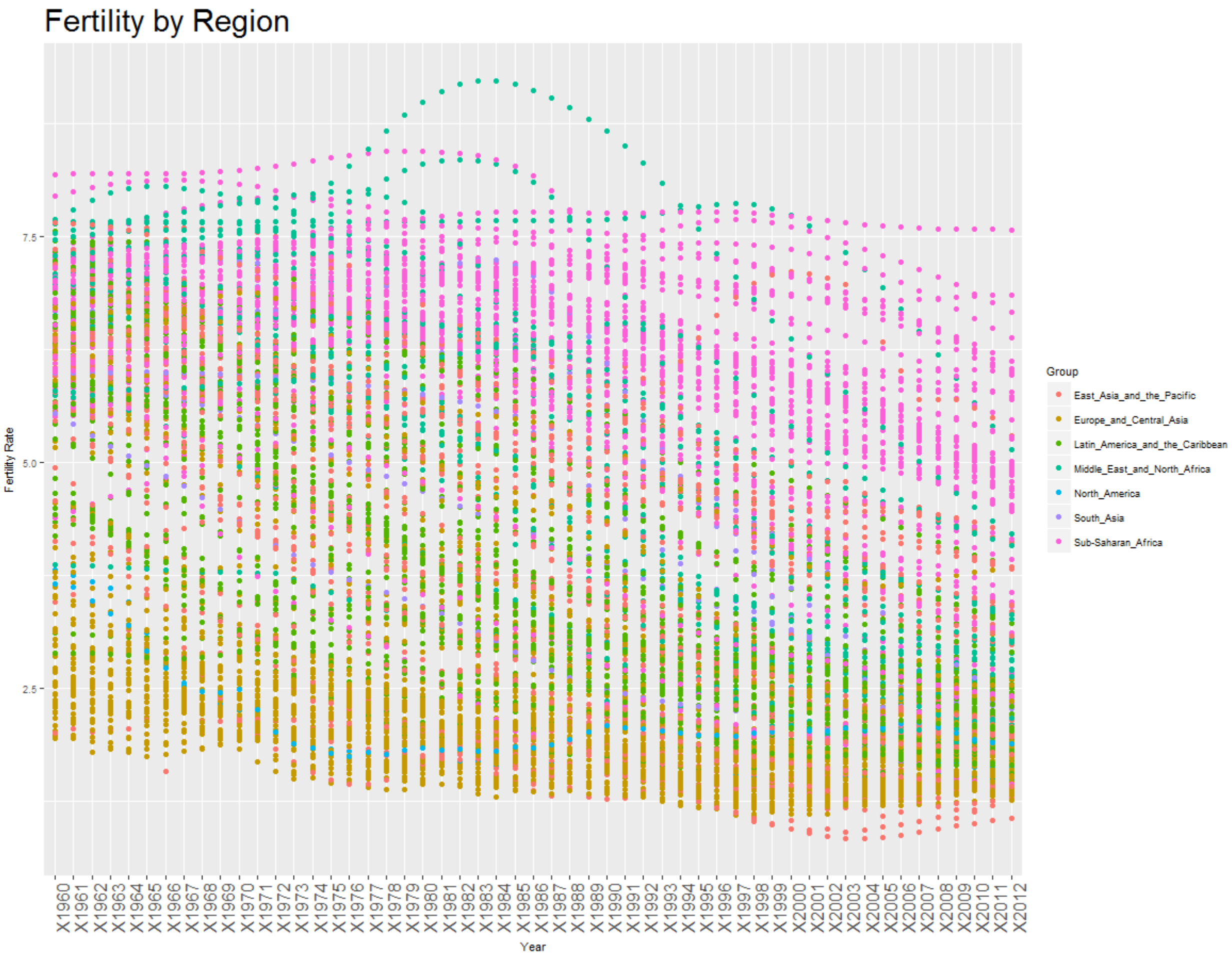
Like the case of Australia, through the years its life expectancy has increased more than the other countries, but its fertility has the lower rates within the countries compared.

Some populations have high fertility although its survival is low. They are characteristic of changing or unstable environments subject to high mortality rates, which offset explosive growth in good times.

2.5 Analysis by region

2.5.1 World Fertility by region

How has the fertility changed by region over the years?

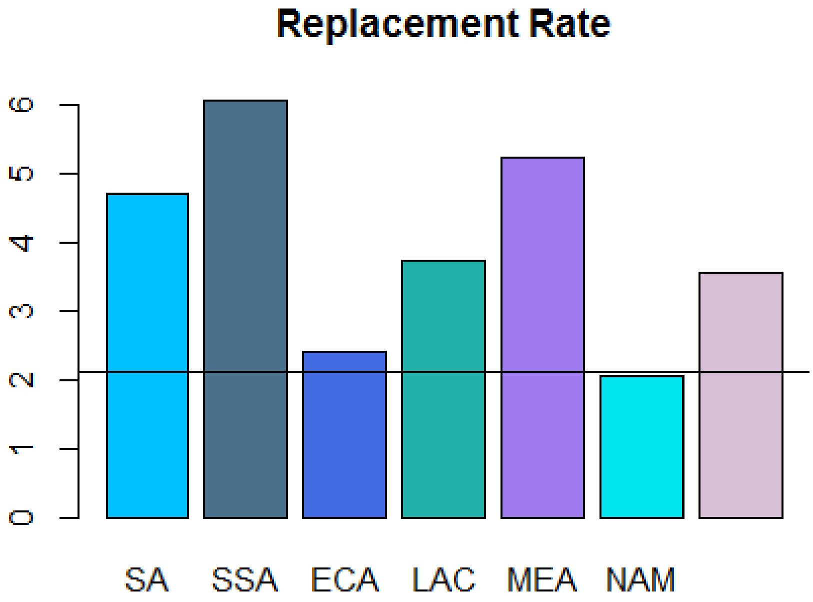


Countries with the **highest fertility rate** are in the **sub-Saharan region.** In this area, are countries like Angola, Ethiopia, Cambodia, Nigeria; African countries with many scarcity and social issues.

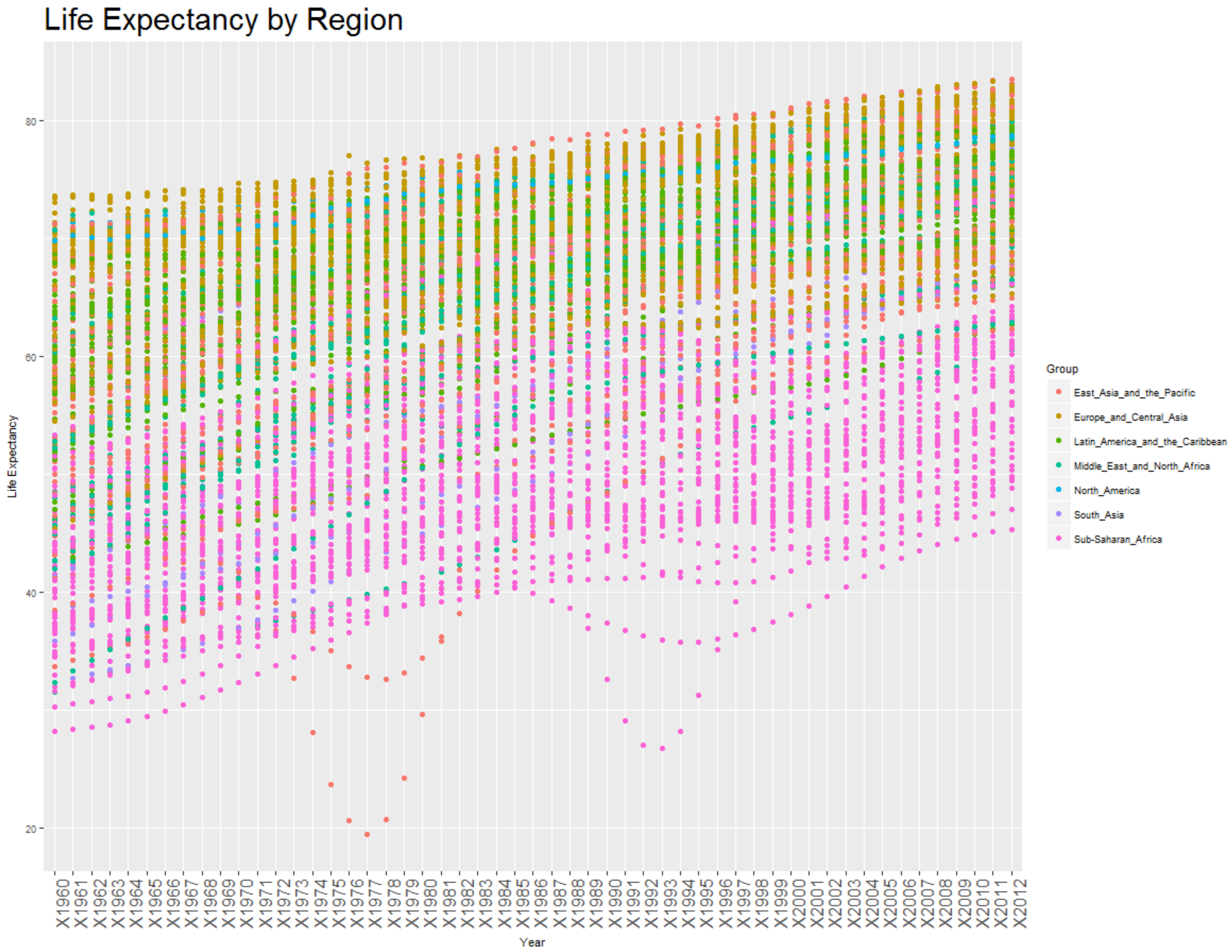
By contrast, in the bottom of the graph is the region **Europe and Central Asia** (first world countries) with the **lowest fertility rates**.

As the fertility rate measures the children per woman, a rate of 2 would mean that every woman gives birth to (on average) one girl. Therefore, the current generation of women would just replace itself by a younger generation of girls. The fertility rate that would create this scenario is called the **replacement rate**. In reality the replacement rate is higher than 2 because not all girls reach the age to give birth. This means that the replacement rate for populations with low mortality is close to 2 and for populations with a high mortality is considerably over 2. As mortality differs widely between different nations, the replacement rate for different countries ranges from below **2.1 for rich countries** to almost **3.5 for poor countries.**

The region of North America in average, has been below the 2.1 fertility number.



2.5.2 Life Expectancy by region



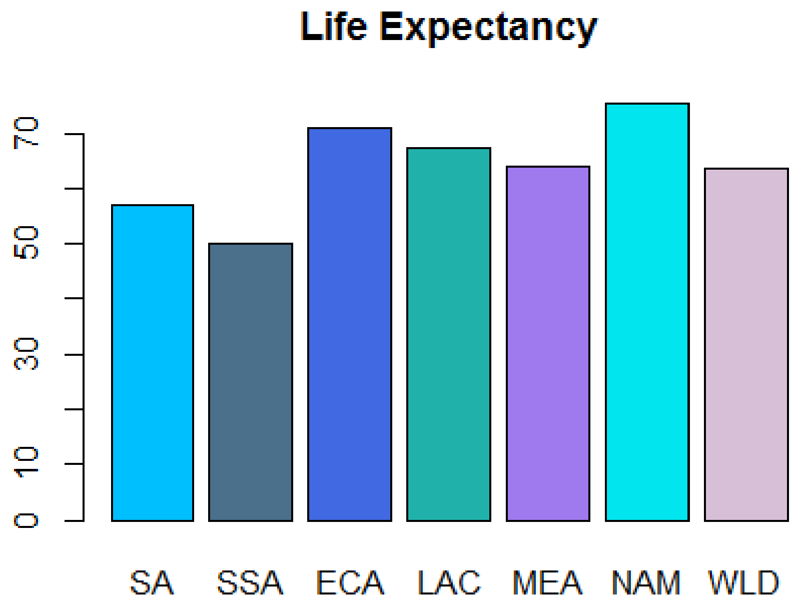
Here is an opposite behavior to what was observed with fertility rates. It is expected that in most developed countries life expectancy is greater, compared to countries with social problems.

Countries with **higher life expectancy** are in the **Europe and Central Asia zone**. Countries with **lower life expectancy** are in the **Sub-Saharan Africa**.

There are great variations in life expectancy between different parts of the world, mostly caused by differences in public health, medical care and diet.

North America region has the higher life expectancy with an average of 75 years, compared to Africa with an average of 50 years.

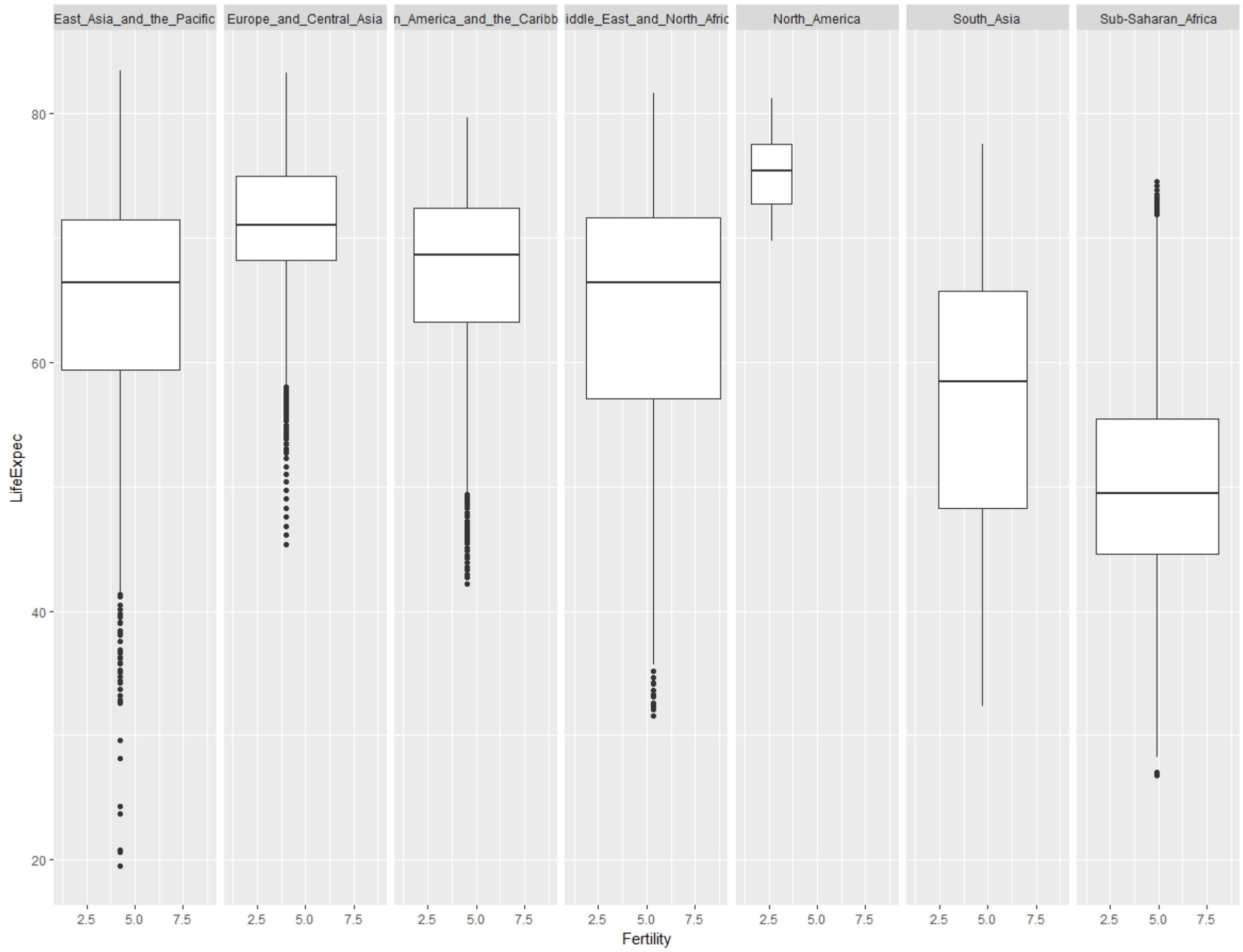
25 years of difference!!



2.5.1 Fertility vs. Life Expectancy

Through the analysis, it has been seen the gradual increase in life expectancy and the corresponding decrease in fertility rates worldwide.

It is very interesting to note, that although they have improved, a large cluster of African countries remain far from the overall trend, with lower life expectancy and higher fertility rates.



3. Forecasting

Following algorithms will be tested:

1. Linear Regression
2. Logistic Regression
3. Principal Component Regression
4. Decision Trees
5. Random Forest

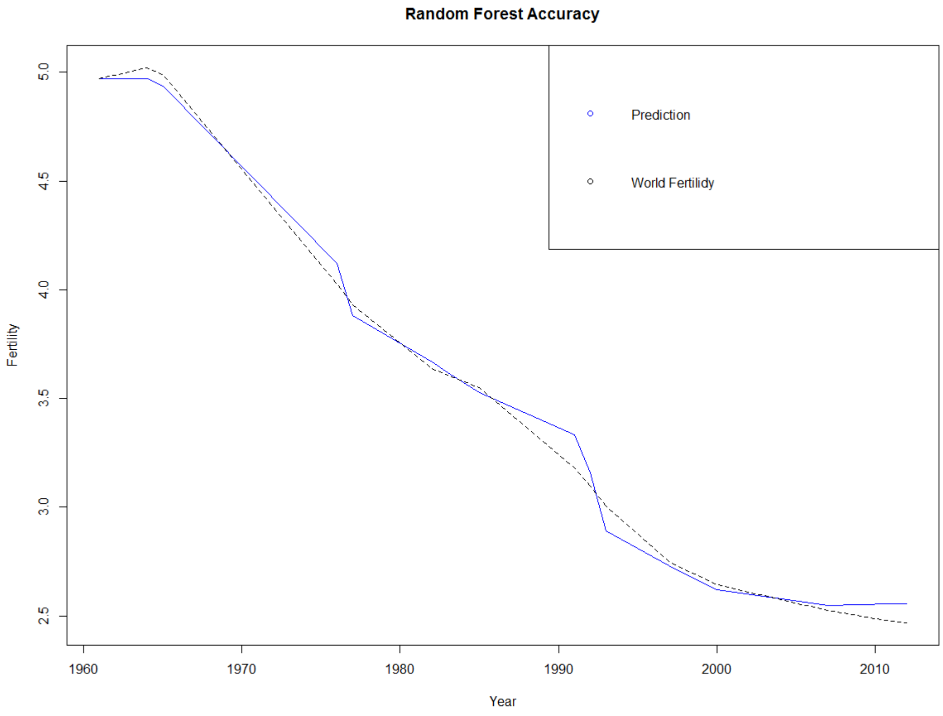
A training set was taken with 70% of the data and with the 30% left, the models were evaluated.

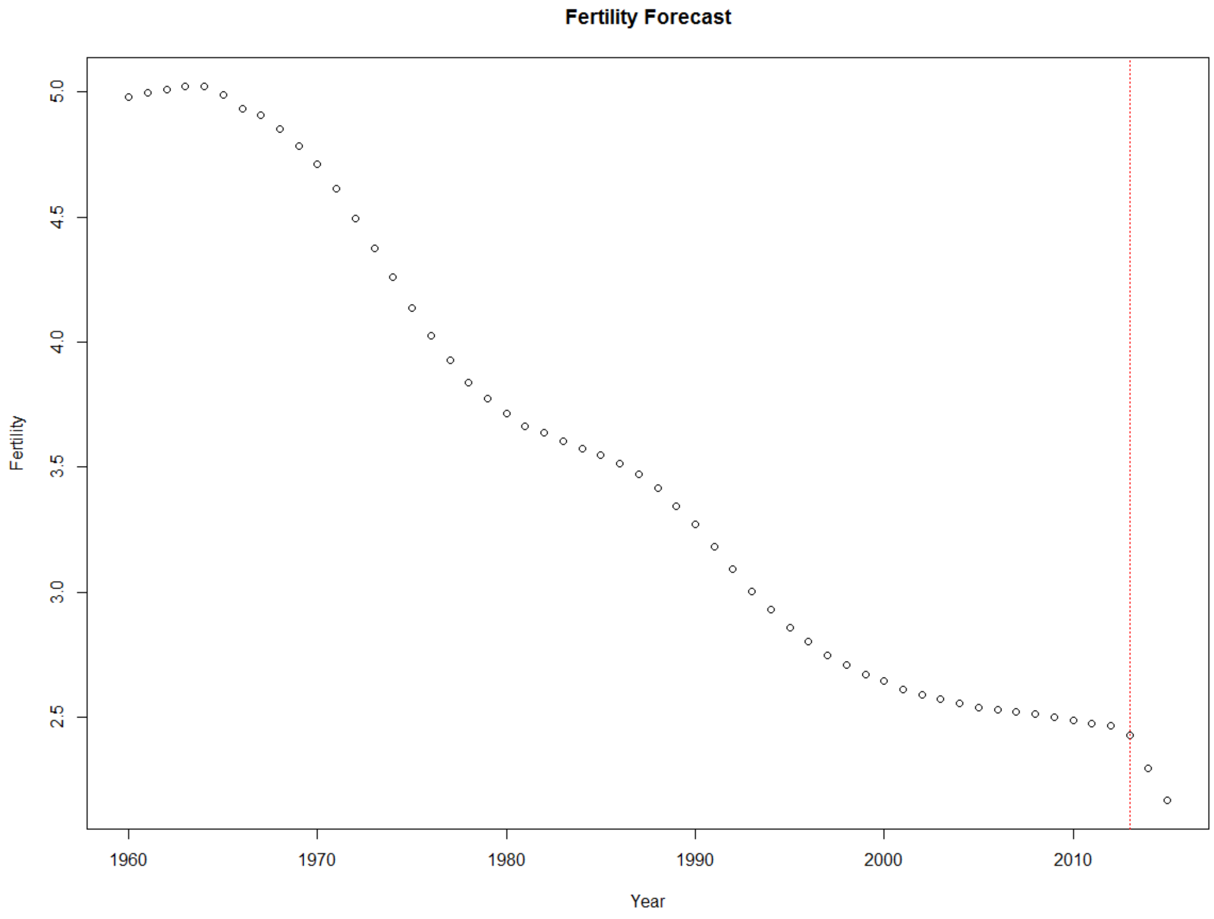
3.1 World Fertility Forecast

The outcome variable will be the Worlds Fertility rate, using as predictors each region average fertility rate.

Linear Regression (MSE)	Logistic Regression (MSE)	Principal Component (MSE)	Decision Tree (MSE)	Random Forest (MSE)
2.177702	2.177290	2.177651	2.256225	**2.177152**

The best model is the Random Forest with the lowest Mean Squared Error (MSE).





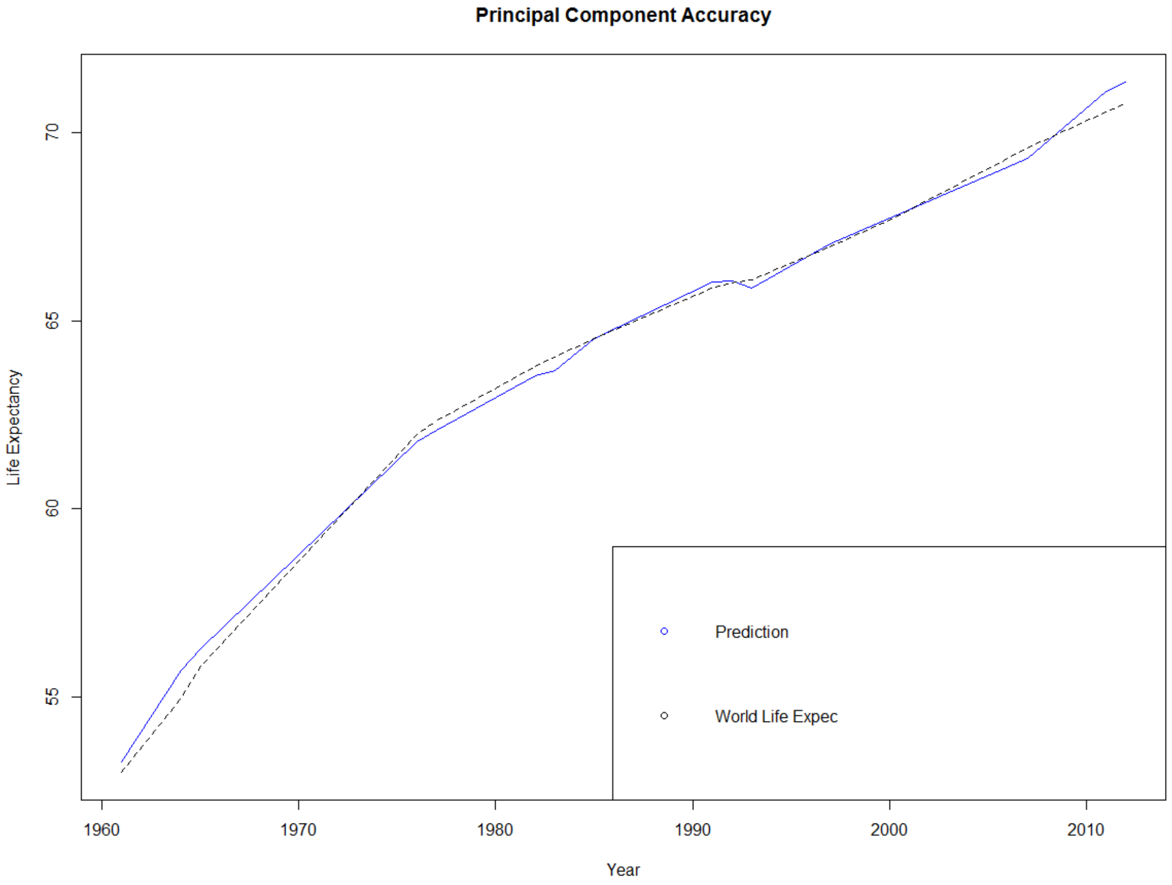
As the in the plot can be seen, world’s fertility will keep decreasing in the following two years.

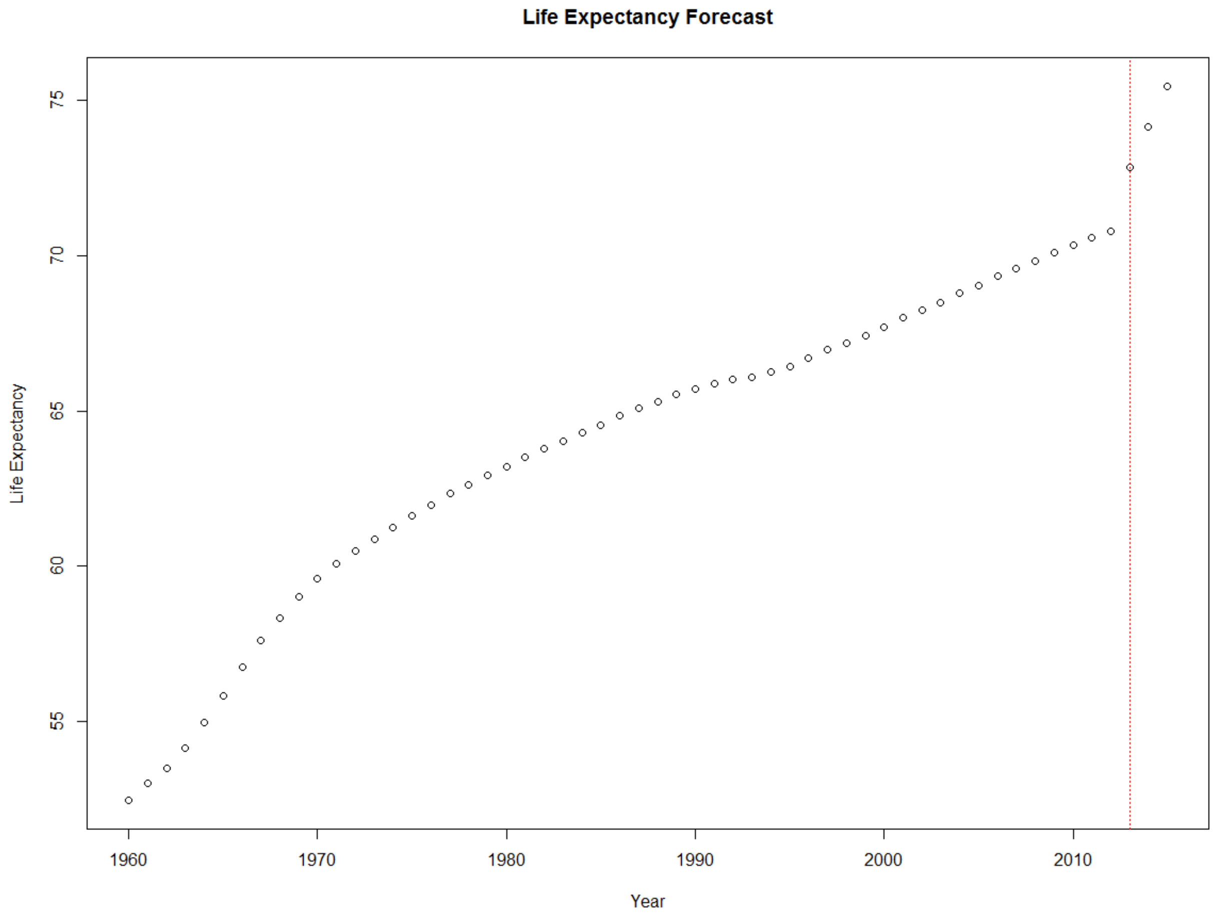
3.2 World Life Expectancy Forecast

The outcome variable will be the world's life expectancy rate, using as predictors each region average life expectancy rate.

Linear Regression (MSE)	Logistic Regression (MSE)	Principal Component (MSE)	Decision Tree (MSE)	Random Forest (MSE)
64.49821	64.49847	64.45062	66.33104	64.83437

The best model is the Principal Component with the lowest Mean Squared Error (MSE).





As the in the plot can be seen, world's life expectancy will keep growing in the following two years.

Conclusions

Changes in life expectancy and fertility rates determine population growth. We would expect that the global improvement in the life expectancy worked to increase the world population, but it is more than offset by the fall in fertility. The global average fertility rate was 5 children per woman until the end of the 1960s and has halved since then.

The replacement rate is a function of life expectancy. Increasing life expectancy means that more daughters reach the age to become mothers themselves

We would expect that most first world regions (North America and Europe/Central Asia) would have higher life expectancy and lower fertility.

On the other hand we would expect the opposite to happen in poorest regions (Sub-Saharan Africa and South Asia) with higher fertility and lower life expectancy. So, wealth is a factor to consider.

As example, a poor women, begging in the streets in India tend to have more children. Her reason is that she now has more hands to collect money. She also have a higher chance of survival even if many of her children die of disease or malnutrition. Some solutions: children should survive and shouldn't be needed for work, women join the labor force and getting education should be accessible.

In a near future fertility rates will keep decreasing to the contrary of life expectancy which will keep increasing.

Attachments

R Code

```r
#Set working directory
setwd("C:/Users/rodsara/Desktop ")

#load files
fertility <- read.csv("./fertility.csv", header=TRUE)
lifeExpec <- read.csv("./life_expectancy.csv", header=TRUE)
lifeExpec$Group <- fertility$Group  #Add region to life expectancy data

##Load Libraries
library(ggplot2)
library(missForest)
library(reshape2)
library(mapdata)
library(ggmap)
library(maptools)
library(dplyr)
library(plyr)
library(randomForest)
library(gbm)
library(pls)
library(data.table)
library(tree)

##Missing Values By year
na <- 0
na1 <- 0
for (i in 3:56){ #count NA in each data set
  na[i-2]<- sum(is.na(fertility[,i]))
  na1[i-2]<- sum(is.na(lifeExpec[,i]))
}

nas <- as.data.frame(cbind(names(fertility)[3:56],na,na1)) #create data frame
names(nas) <- c("Year", "NA's Fertility","NA's Life Expec")

fertility <- fertility[,-56] #Eliminate Year 2013
lifeExpec <- lifeExpec[,-56] #Eliminate Year 2013


nas1 <- subset(nas, nas$`NA's Fertility`==258 & nas$`NA's Life Expec`==258) #
select year 2013

##Missing Values By country
MV <- 0
MV1 <- 0
fertility$Country_Name <- as.character(fertility$Country_Name)
for (i in 1:258){
  MV[i]<- sum(is.na(fertility[i,3:55])) #COUNT NA in each data set
  MV1[i]<- sum(is.na(lifeExpec[i,3:55]))
}
namesFer<- fertility[,1]
MVS<- as.data.frame(cbind(namesFer,MV,MV1))#create data frame
names(MVS) <- c("Country", "NA_Fertility","NA_Life_Expec")

##Find variables with 100% NA
MaxMissingFer <-  subset(MVS,NA_Fertility==53,select=Country)
MaxMissingLife <- subset(MVS,NA_Life_Expec==53,select=Country)

##Country comparation 100% NA by Fertility and Life Expectancy
Missing <- cbind(MaxMissingFer,MaxMissingLife)
names(Missing) <- c("Fertility Country","Life Expec Country")

#Exclude those variables for each data set and the Country_Names column
fertility <- fertility[-c(2,6,11,34,37,39,45,46,53,57,82,108,134,148,149,164,
204,227,238), -1]
lifeExpec <- lifeExpec[-c(2,6,11,34,37,39,45,46,53,57,82,108,134,148,149,164,
204,227,238), -1]

###Impute NA in Fertility data set with MissForest
f <- missForest(fertility[2:54],verbose=FALSE)

f <- as.data.frame(f$ximp)
fertility <- cbind(fertility[1],f,fertility[55])

###Impute NA in Life Expectancy data set with MissForest
g <- missForest(lifeExpec[2:54],verbose=FALSE)

g <- as.data.frame(g$ximp)
lifeExpec <- cbind(lifeExpec[1],g,lifeExpec[55])

#Find variables with no region
noRegion <- subset(fertility, fertility$Group==0)
noRegionCol <- fertility$Group   #Column with the group names
noRegion1 <- subset(lifeExpec, lifeExpec$Group==0)

##Exlude variables with no region
fertility <- subset(fertility, fertility$Group!=0)
fertility <- fertility[,-56]
lifeExpec <- subset(lifeExpec, lifeExpec$Group!=0)
lifeExpec <- lifeExpec[,-56]

#Plot World's Fertility rate over the years
year <- 1960:2012
Fer <- fertility[,-c(1,55)]
Fer <- as.data.frame(t(Fer))
WLD2 <- subset(noRegion,Country_Code=="WLD")
WLD <- t(WLD2[,-c(1,55,56)])
```

```r
qplot(x=year,y=WLD,geom="point", main="World Fertility 1960-2012", ylab="World Fertility Rate")+ theme(plot.title = element_text(size=22))#+geom_vline(xintercept = c(1964,1979,1995),colour="blue", linetype = "longdash")

#Plot World's Life Expectancy over the years
Life <- lifeExpec[,-c(1,55)]
Life <- as.data.frame(t(Life))
WLD1 <- subset(noRegion1,Country_Code=="WLD")
WLD <- t(WLD1[,-c(1,55,56)])
qplot(x=year,y=WLD,geom="point",main="Life Expectancy 1960-2012", ylab="World Life Expectancy Rate")+ theme(plot.title = element_text(size=22))

#Plot Historical Fertility and Life Expectancy in all countries
par(mfrow=c(1,2))
matplot(year, Fer , type="l", main="Historical Fertility Rate", cex.main=2, cex.axis=1, cex.lab=1.5, ylab="Fertility Rate in all countries")
matplot(year,Life, type="l", main="Historical Life Expectancy Rate", cex.main=2, cex.axis=1, cex.lab=1.5, ylab="Life Expectancy in all countries")

#Using GGPLOT, plot the Base World Map
mp <- NULL
mapWorld <- borders("world", colour="gray50", fill="gray50") # create a layer of borders
mp <- ggplot() +   mapWorld
visited <- c("Cambodia", "Mexico", "Australia", "Rwanda","Germany")  #Countries compared
ll.visited <- geocode(visited)
visit.x <- ll.visited$lon
visit.y <- ll.visited$lat
#Now Layer the cities on top
mp <- mp+ geom_point(aes(x=visit.x, y=visit.y) ,color="blue", size=3)


###Divide selected countries by war/nowar
CAM1 <- subset(lifeExpec,Country_Code=="KHM")
CAM2 <- subset(fertility,Country_Code=="KHM")
CAM1$War <- "War"
CAM2$War <- "War"
GER1 <- subset(lifeExpec,Country_Code=="DEU")
GER2 <- subset(fertility,Country_Code=="DEU")
GER1$War <- "NoWar"
GER2$War <- "NoWar"
MX1 <- subset(lifeExpec,Country_Code=="MEX")
MX2 <- subset(fertility,Country_Code=="MEX")
MX1$War <- "NoWar"
MX2$War <- "NoWar"
RWA1 <- subset(lifeExpec,Country_Code=="RWA")
RWA2 <- subset(fertility,Country_Code=="RWA")
RWA1$War <- "War"
RWA2$War <- "War"
```

```
AUS1 <- subset(lifeExpec,Country_Code=="AUS")
AUS2 <- subset(fertility,Country_Code=="AUS")
AUS1$War <- "NoWar"
AUS2$War <- "NoWar"
year <- 1960:2012

#create data frame with the above selection
compLife <- rbind(CAM1,GER1,MX1,RWA1,AUS1)
row.names(compLife)<-c("CAM","GER","MX","RWA","AUS")
compLife <-compLife[,-55]
compFer <- rbind(CAM2,GER2,MX2,RWA2,AUS2)
row.names(compFer)<-c("CAM","GER","MX","RWA","AUS")
compFer <-compFer[,-55]

#Plot War vs No war cases
compLife1 <- melt(compLife, id.var = c("Country_Code","War"))
colnames(compLife1)<- c("Country_Code","War", "Variable","Life_Expectancy")
compFer1 <- melt(compFer, id.var = c("Country_Code","War"))
colnames(compFer1)<- c("Country_Code","War", "Variable","Fertility")
ggplot(compLife1, aes(x=Variable, y=Life_Expectancy, group=Country_Code,colou
r = War)) + geom_path(alpha = 1) + theme(text = element_text(size=12),
        axis.text.x = element_text(angle=90, vjust=1), plot.title=element_tex
t(size=22)) + ggtitle("War vs No War Life Expectancy")

ggplot(compFer1, aes(x=Variable, y=Fertility, group = Country_Code, colour =
War)) + geom_path(alpha = 1) + theme(text = element_text(size=12),
        axis.text.x = element_text(angle=90, vjust=1), plot.title=element_tex
t(size=22)) + ggtitle("War vs No War Life Fertility")

#plot only Cambodia Life Expectancy
dev.off()

## null device
##           1

cambodiaLIFE <- CAM1[,-c(55,56)]
row.names(cambodiaLIFE)<-"CAM"
cambodiaLife1 <- melt(cambodiaLIFE, id.var = "Country_Code")
cambodiaLife1 <- cbind(cambodiaLife1,1960:2012)
colnames(cambodiaLife1)<- c("Country_Code", "Variable","Life_Expectancy","Yea
r")
ggplot(cambodiaLife1, aes(x=Year, y=Life_Expectancy)) + geom_path(alpha = 1)
+ theme(text = element_text(size=8),
        axis.text.x = element_text(angle=90, vjust=1)) + ggtitle("Cambodia Li
fe Expectancy")+geom_vline(xintercept = c(1970,1985),colour="blue", linetype
= "longdash")+theme(plot.title = element_text(size=22), axis.text=element_tex
t(size=12),
        axis.title=element_text(size=14,face="bold"))

#plot fertility rate by region
```

```r
fertilityM <- melt(fertility, id.var = c("Country_Code","Group"))
qplot(variable,value,data=fertilityM,colour=Group)+theme(text = element_text(
size=8),
       axis.text.x = element_text(angle=90, vjust=1,size=12)) + ggtitle("Fert
ility by Region")+theme(plot.title = element_text(size=22))+labs(x="Year",y="
Fertility Rate")

#plot Life expectancy by region
LifeM <- melt(lifeExpec, id.var = c("Country_Code","Group"))
qplot(variable,value,data=LifeM,colour=Group)+theme(text = element_text(size=
8),
       axis.text.x = element_text(angle=90, vjust=1,size=12)) + ggtitle("Life
Expectancy by Region")+theme(plot.title = element_text(size=22))+labs(x="Year
",y="Life Expectancy")

#create data frame ordered by country code and group
data <- melt(lifeExpec, id=c("Country_Code","Group"))
data <- setorder(data,Country_Code)
data1 <- melt(fertility,id=c("Country_Code","Group"))
data1 <- setorder(data1,Country_Code)
data <- cbind(data,data1$value)
names(data) <- c("Country_Code","Group","Year","LifeExpec","Fertility")

#boxplot of each region comparing Fertility and Life Expectancy
ggplot(data,aes(x=Fertility,y=LifeExpec))+geom_boxplot()+facet_grid(.~Group)

#subset each region by fertility
SA <- subset(fertility,Group=="South_Asia")
meanSA <- apply(SA[2:54],2,mean)

SSA <- subset(fertility, Group=="Sub-Saharan_Africa")
meanSSA <- apply(SSA[2:54],2,mean)

ECA <- subset(fertility, Group=="Europe_and_Central_Asia")
meanECA <- apply(ECA[2:54],2,mean)

LAC <- subset(fertility, Group=="Latin_America_and_the_Caribbean")
meanLAC <- apply(LAC[2:54],2,mean)

MEA <- subset(fertility, Group=="Middle_East_and_North_Africa")
meanMEA <- apply(MEA[2:54],2,mean)

NAM <- subset(fertility, Group=="North_America")
meanNAM <- apply(NAM[2:54],2,mean)

meanWLD <- t(WLD2[2:54])

dataFer <- cbind(meanSA,meanSSA,meanECA,meanLAC,meanMEA,meanNAM,meanWLD)
dataFer <- as.data.frame(dataFer)
```

```r
names(dataFer) <- c("SA","SSA","ECA","LAC","MEA","NAM","WORLD")


meanFer <- apply(dataFer,2,mean)
#create bar chart with replacement rate for Fertility by region
barplot(meanFer, col=c("deepskyblue","skyblue4","royalblue","lightseagreen","
mediumpurple2","turquoise2","thistle"), main="Replacement Rate")
abline(h=2.1)

#subset each region by fertility
SA <- subset(lifeExpec,Group=="South_Asia")
meanSA <- apply(SA[2:54],2,mean)

SSA <- subset(lifeExpec, Group=="Sub-Saharan_Africa")
meanSSA <- apply(SSA[2:54],2,mean)

ECA <- subset(lifeExpec, Group=="Europe_and_Central_Asia")
meanECA <- apply(ECA[2:54],2,mean)

LAC <- subset(lifeExpec, Group=="Latin_America_and_the_Caribbean")
meanLAC <- apply(LAC[2:54],2,mean)

MEA <- subset(lifeExpec, Group=="Middle_East_and_North_Africa")
meanMEA <- apply(MEA[2:54],2,mean)

NAM <- subset(lifeExpec, Group=="North_America")
meanNAM <- apply(NAM[2:54],2,mean)

meanWLD <- t(WLD1[2:54])

dataLife <- cbind(meanSA,meanSSA,meanECA,meanLAC,meanMEA,meanNAM,meanWLD)
dataLife <- as.data.frame(dataLife)
names(dataLife) <- c("SA","SSA","ECA","LAC","MEA","NAM","WLD")


meanLife <- apply(dataLife,2,mean)

#create bar chart of life expectancy by region
barplot(meanLife, col=c("deepskyblue","skyblue4","royalblue","lightseagreen",
"mediumpurple2","turquoise2","thistle"), main="Life Expectancy")

#analysis of life expectancy and fertility by income
LowL <- subset(noRegion1,Country_Code=="LIC") #lifeExpectancy
LowF <- subset(noRegion,Country_Code=="LIC") #fertility
MedL <- subset(noRegion1,Country_Code=="LMC") #lifeExpectancy
MedF <- subset(noRegion,Country_Code=="LMC") #fertility
HighL <- subset(noRegion1,Country_Code=="HIC") #lifeExpectancy
HighF <- subset(noRegion,Country_Code=="HIC") #fertility
year<-1960:2012
```

```r
incomeL <- rbind(LowL,MedL,HighL)
incomeL <- incomeL[,-55]
incomeF <- rbind(LowF,MedF,HighF)
incomeF <- incomeF[,-55]
incomeL1 <- melt(incomeL, id.var=c("Country_Code"))
incomeF1 <- melt(incomeF, id.var=c("Country_Code"))

#plot life expectancy by income
qplot(variable,value,data=incomeL1,colour=Country_Code)+theme(text = element_
text(size=8),
        axis.text.x = element_text(angle=90, vjust=1)) + ggtitle("Life Expect
ancy by income")+theme(text = element_text(size=8),
       axis.text.x = element_text(angle=90, vjust=1,size=12)) + ggtitle("Life
Expectancy by income")+theme(plot.title = element_text(size=22))

#plot fertility rate by income
qplot(variable,value,data=incomeF1,colour=Country_Code)+theme(text = element_
text(size=8),
        axis.text.x = element_text(angle=90, vjust=1)) + ggtitle("Fertility b
y income")+theme(text = element_text(size=8),
       axis.text.x = element_text(angle=90, vjust=1,size=12)) + ggtitle("Fert
ility by income")+theme(plot.title = element_text(size=22))

#Forecasting
set.seed(1)
train <- sample(1:nrow(dataFer),nrow(dataFer)*0.7) #training set
data.train <- dataFer[train,]
data.test <- dataFer[-train,] #test set
error.fin <- 1:5
names(error.fin) <- c("Linear Regression", "Logistic Regression","Principal c
omponent","Decision Tree","RandomForest")


##Linear Regression
linear.reg <- lm(WORLD~., data.train)
info.linear <- step(linear.reg)

info.linear <- predict(info.linear, dataFer[-train,])
error.fin[1] <- mean((info.linear - data.test)^2)

###Logistic regression
logistic.data <- glm(WORLD~.,data=data.train,family=gaussian)
info.logistic <- predict(logistic.data,newdata=dataFer[-train,],type="respons
e")
error.fin[2]<- mean((info.logistic - data.test)^2)

###Principal Component Regression
pcr.reg <- pcr(WORLD~., data=data.train, validation="CV")
info.pcr <- predict(pcr.reg, data.test, ncomp=5)
```

```r
error.fin[3] <- mean((info.pcr - data.test)^2)

###Decition Trees
tree.reg <- tree(WORLD~.,data=data.train)
info.tree <- predict(tree.reg,data.test)
error.fin[4]<-mean((info.tree-data.test)^2)

###Random Forest
set.seed(1)
ntrees <- c(100,300)
mtry <- c(1,2,5)
errores.random <- data.frame()
for(i in 1:(length(ntrees))){
  for(j in 1:length(mtry)){
    rf <- randomForest(WORLD~.,data=data.train,mtry=mtry[j],ntree=ntrees[i],i
mportance=TRUE)
    info.rf <- predict(rf,data.test)
    errores.random[i,j]<- mean((info.rf - data.test)^2)
  }
}
optim <- which(errores.random==min(errores.random),arr.ind=TRUE)
mtry[optim[1]]

ntrees[optim[2]]

error.fin[5]<- min(errores.random)

error.fin


year1 <- c(1961,1964,1965,1976,1977,1982,1983,1985,1991,1992,1993,1997,2000,2
007,2011,2012)
WORLD <- c(4.97,5.020,4.99,4.027,3.929,3.637,3.606,3.548,3.182,3.095,3.006,2.
750,2.645,2.524,2.476,2.469)
prediction1 <- cbind(info.rf,WORLD)
prediction1 <- as.data.frame(prediction1)
names(prediction1) <- c("Prediction","WORLD")

#plot real data and accuracy of the model
matplot(year1,prediction1, type = c("l"),pch=1,col = c("blue","black"), main=
"Random Forest Accuracy", xlab="Year", ylab="Fertility")
legend("topright", legend =c("Prediction","World Fertility") , col=c("blue","
black"), pch=1) # optional legend

library(forecast)
f <- forecast(info.rf,h=3)
f <- f$mean
forecastFer <- append(dataFer$WORLD, f)
#plot fertility forecast
plot(x=1960:2015,y=forecastFer, main="Fertility Forecast", xlab="Year", ylab=
"Fertility")
abline(v = 2013, col = "red", lty = 3)


##Training and test set
set.seed(1)
train <- sample(1:nrow(dataLife),nrow(dataLife)*0.7)
data.train <- dataLife[train,]
data.test <- dataLife[-train,]
error.fin <- 1:5
names(error.fin) <- c("Linear Regression", "Logistic Regression","Principal c
omponent","Decision Tree","RandomForest")


##Linear Regression
linear.reg <- lm(WLD~., data.train)
info.linear <- step(linear.reg)

info.linear <- predict(info.linear, dataLife[-train,])
error.fin[1] <- mean((info.linear - data.test)^2)

###Logistic regression
logistic.data <- glm(WLD~.,data=data.train,family=gaussian)
info.logistic <- predict(logistic.data,newdata=dataLife[-train,],type="respon
se")
error.fin[2]<- mean((info.logistic - data.test)^2)

###Principal Component Regression
pcr.reg <- pcr(WLD~., data=data.train, validation="CV")
info.pcr <- predict(pcr.reg, data.test, ncomp=5)
error.fin[3] <- mean((info.pcr - data.test)^2)

###Decition Trees
tree.reg <- tree(WLD~.,data=data.train)
info.tree <- predict(tree.reg,data.test)
error.fin[4]<-mean((info.tree-data.test)^2)

###Random Forest
set.seed(1)
ntrees <- c(100,300)
mtry <- c(1,2,5)
errores.random <- data.frame()
for(i in 1:(length(ntrees))){
  for(j in 1:length(mtry)){
    rf <- randomForest(WLD~.,data=data.train,mtry=mtry[j],ntree=ntrees[i],imp
ortance=TRUE)
    info.rf <- predict(rf,data.test)
    errores.random[i,j]<- mean((info.rf - data.test)^2)
  }
```

```r
}
optim <- which(errores.random==min(errores.random),arr.ind=TRUE)
mtry[optim[1]]

ntrees[optim[2]]

error.fin[5]<- min(errores.random)

error.fin

##The best model is the Principal Component Regression

year <- c(1961,1964,1965,1976,1977,1982,1983,1985,1991,1992,1993,1997,2000,20
07,2011,2012)
WLD <- c(53.006,54.977,55.826,61.974,62.330,63.798,64.033,64.545,65.864,65.99
9,66.086,66.961,67.694,69.591,70.561,70.779)
prediction <- cbind(info.pcr[,1,],WLD)
prediction <- as.data.frame(prediction)
names(prediction) <- c("Prediction","WLD")

#plot real data against accuracy of the model
 matplot(year,prediction, type = c("l"),pch=1,col = c("blue","black"), main="
Principal Component Accuracy", xlab="Year", ylab="Life Expectancy")
 legend("bottomright", legend =c("Prediction","World Life Expec") , col = c("
blue","black"), pch=1) # optional legend

 library(forecast)
f <- forecast(info.pcr,h=3)
f <- f$mean
forecastLife <- append(dataLife$WLD, f)
#plot life expectancy forecast
plot(x=1960:2015,y=forecastLife, main="Life Expectancy Forecast", xlab="Year"
, ylab="Life Expectancy")
abline(v = 2013, col = "red", lty = 3)
```