College of Science
ILLINOIS INSTITUTE OF TECHNOLOGY

IIT - CSP 571 - Spring 2018
Data Preparation & Analysis

Project Report

Ignacio Diez	CWID: A20410853
Cristina Fernandez	CWID: A20410992
Pablo Medrano	CWID: A20410758
Alvaro Gomez	CWID: A20410966

I. ABSTRACT

It has been a challenge to deal with the complexity of this data frame because of several facts. It is extremely huge and it has a great number of indicators. We have made a long research to collect the indicators that could resolve each of the four questions.

Furthermore, the time series was a completely new field for the team so we have spent most of the time researching about this topic to understand this kind of data. In spite of these and other difficulties, it has been a nice challenge and very gratifying at the end. We have learned to deal with time series and also we have understood better the tendency of the countries of study over the years.

For the future, it would be better to work with a data frame with less missing data and more records to compute better and more reliable results.

II. OVERVIEW

The data set chosen for the project is the World Development Indicators available in Kaggle (https://www.kaggle.com/worldbank/world-development-indicators). It gathers time series data from 247 countries from 1960 to 2015, concretely it includes more than 5.5 millions of observables and 1344 indicators. In this report, we focus on three different topics: mortality, life expectancy, gender equality and household consumption. We aim to find an inference model for each feature that explains their relationship with other indicators throughout the time.

III. DATA PROCESSING

First, we filter the countries we want to study: China, France, India, Spain and USA and we created a data set with all the indicators of all the countries we want to study called *our_indicators.* Then, from this data set create a data set for each country(5 in total) containing all the indicators for that country called *COUNTRY_indicators* where COUNTRY corresponds to the abbreviation of each country (CHN, FRA, IND, SPA, USA).

This is common for all the question we are trying to address.

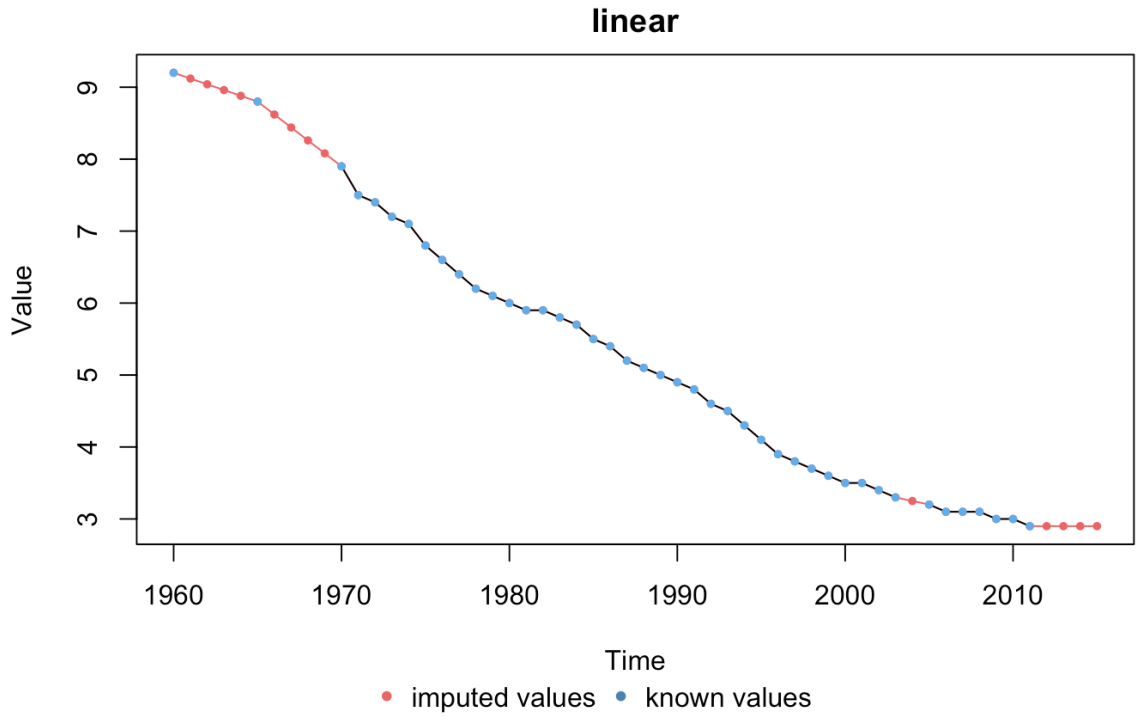
Once we have a data set for each country with all the indicators, we have to select which ones we think that are relevant to answer each question. Finally we will create the final data set for each country containing the values for these indicators we are going to use as predictors. Once the indicators to be used as predictors are selected. We looked for NA values in them. Almost all of them had some NA values and some had an important number of them that would prevent us from using them as predictors as we are going to see next.

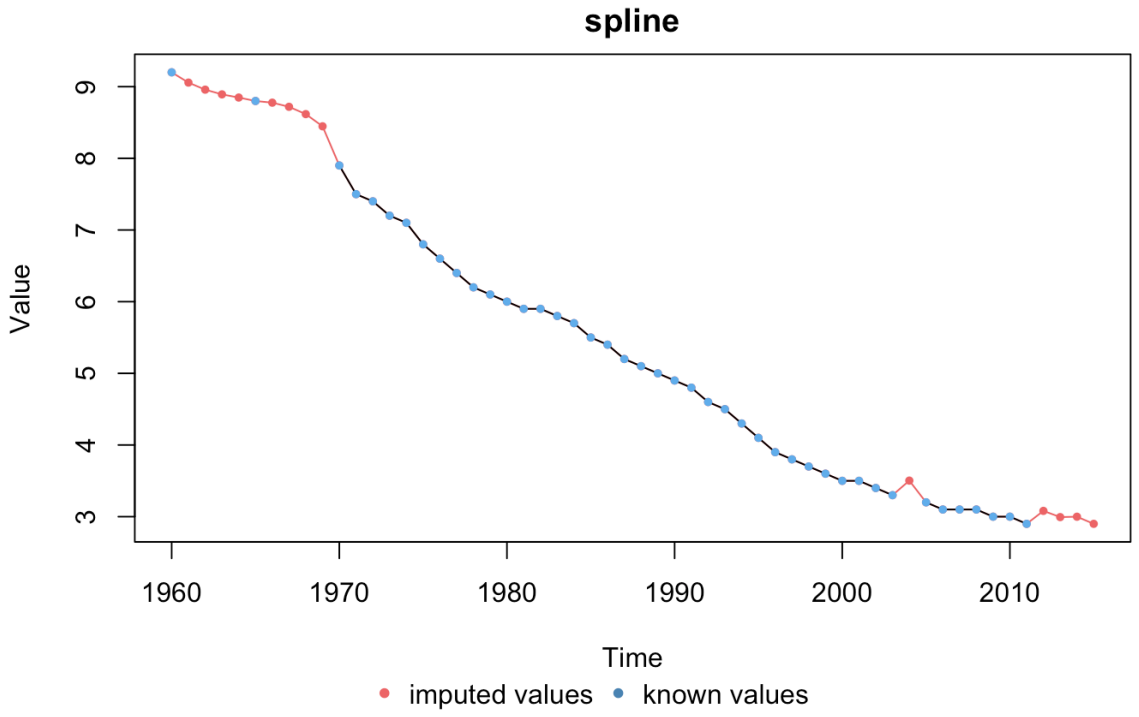
To impute the NA values we considered three options using functions from the imputeTS library:

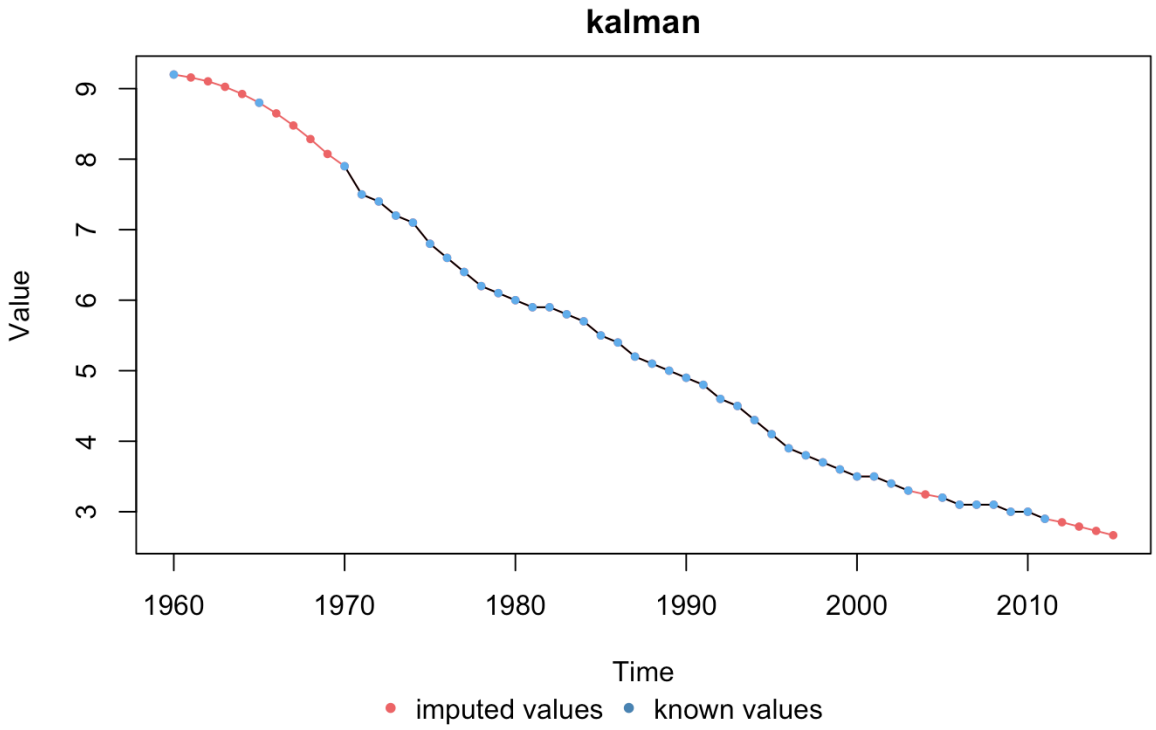
1. Linear interpolation.
2. Spline interpolation.
3. Kalman.

We decided to use Kalman because it seemed to work better when there were a conseriderarbly greater number of values to impute. When the values to impute were smaller they seemed to perform similarly. Next examples of how these three different methods of imputation worked are shown. The first three plots are the imputations done for the Hospital beds indicator and following three are the imputations performed for the Health expenditure indicator, both of them for Spain.

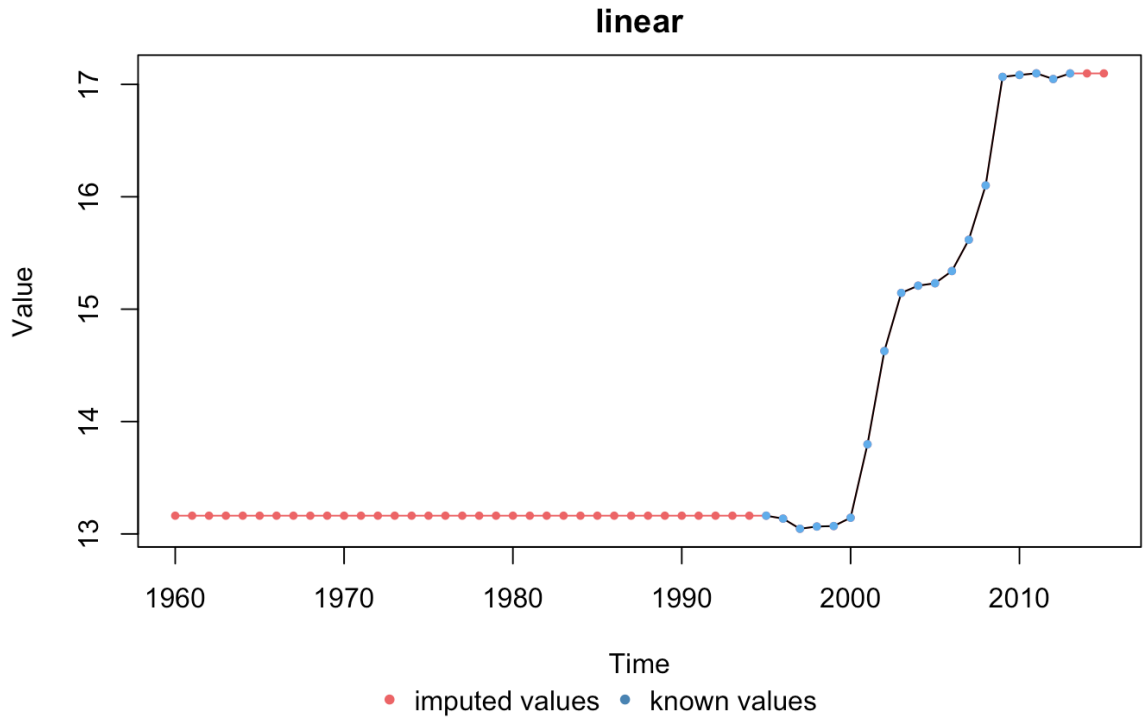
Hospital beds imputations:

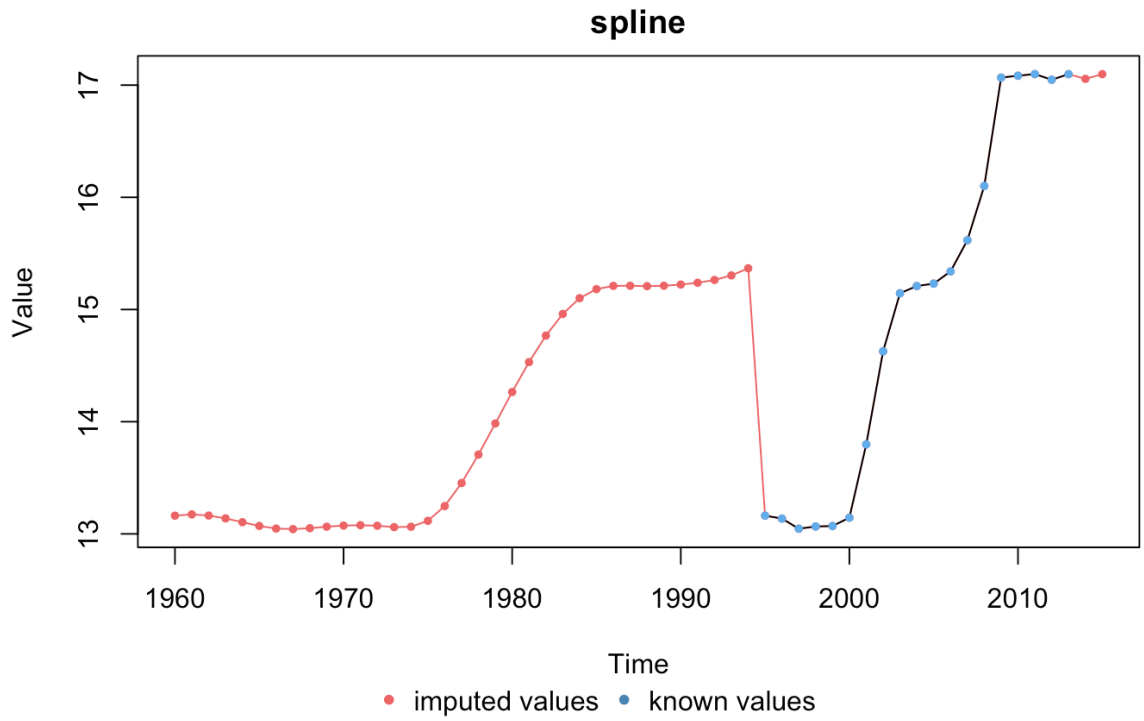

linear
Value
3
4
5
6
7
8
9
1960
1970
1980
1990
2000
2010
Time
imputed values
known values

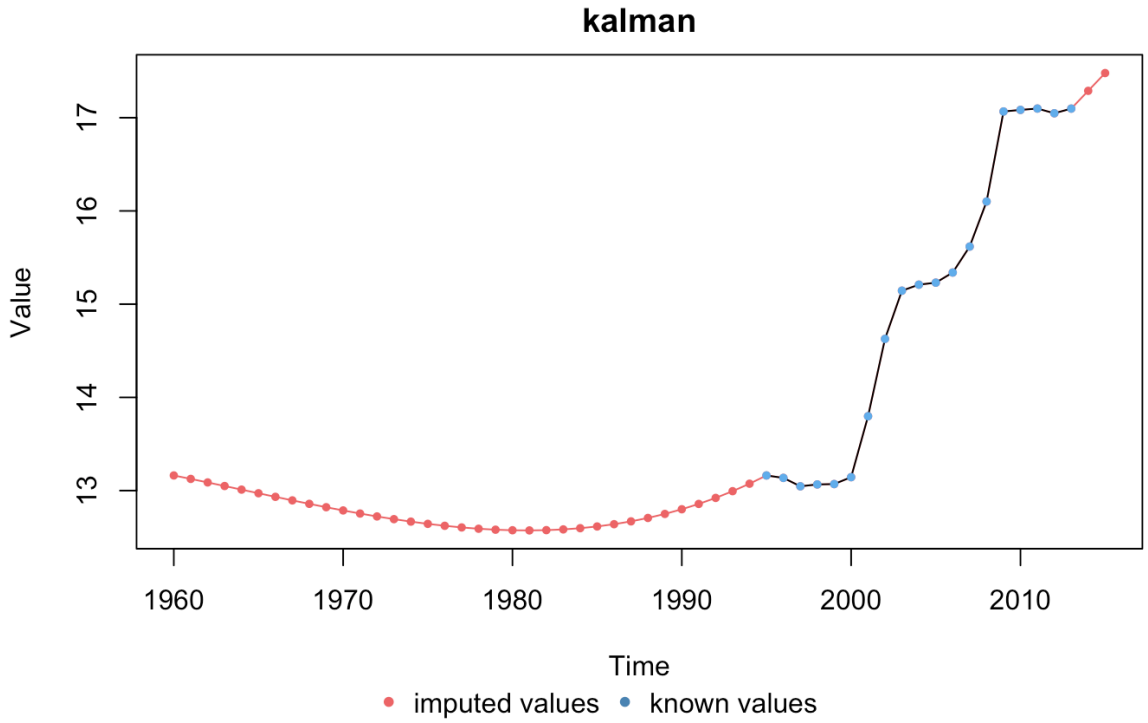

spline
Value
3
4
5
6
7
8
9
1960
1970
1980
1990
2000
2010
Time
imputed values
known values



Health expenditure imputations:







These were the worst cases where the most imputations were done. The imputations done by kalman seemed the smoothest and it is the one used for the rest of the predictors but there were not that many imputations in the rest of predictors as we are going to show in the next plots.

If a predictor had more imputed values than real values they were discarded to use as predictors.

This methodology was commonly followed to answer all the questions we had.

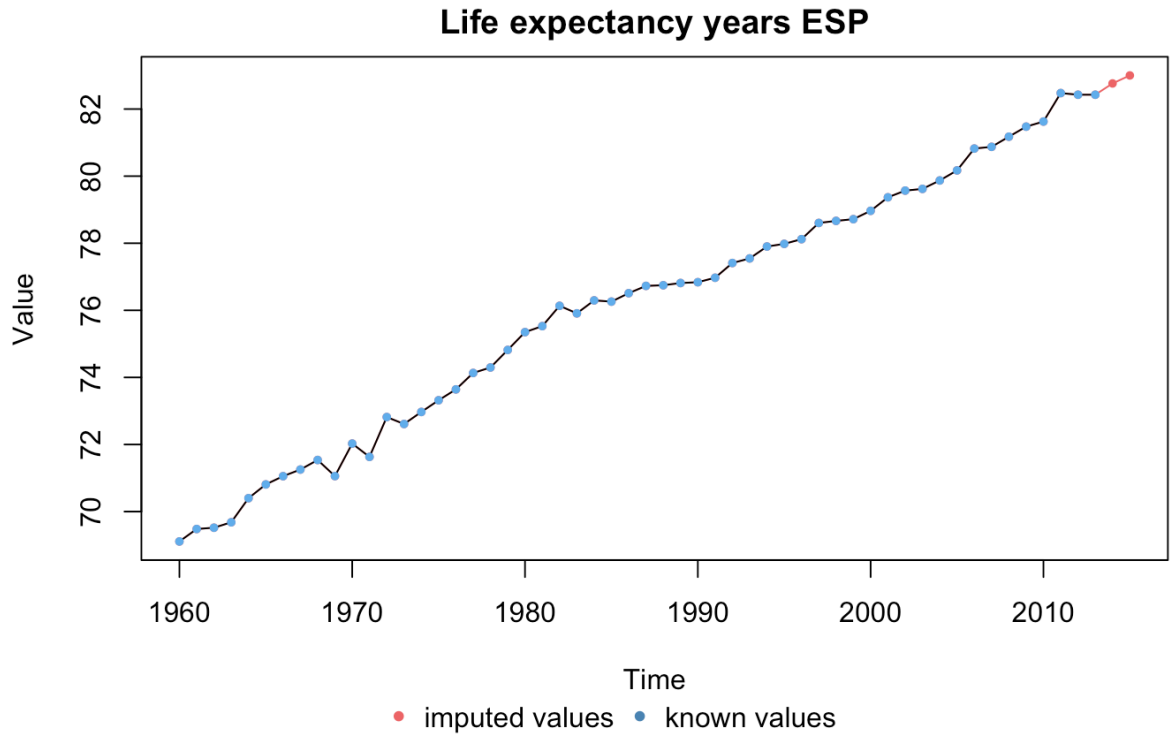
Now, we will show the predictors chosen to answer each question and how the process previously described was followed.

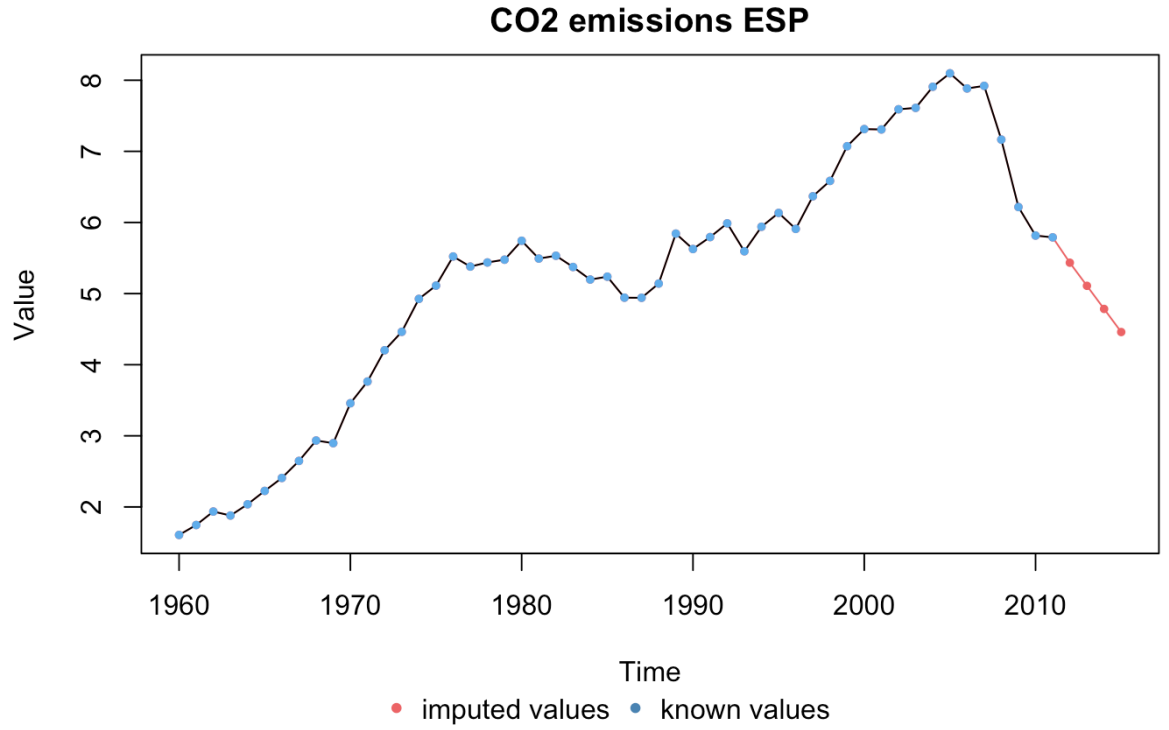
III.I. Life expectancy

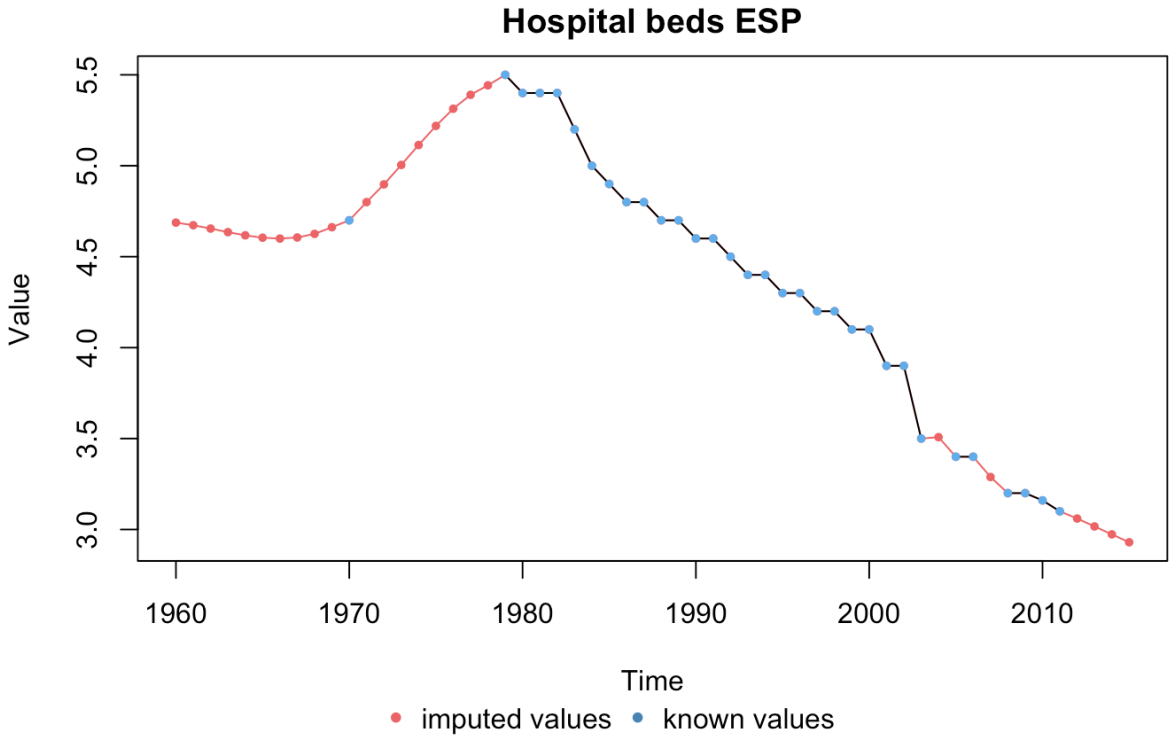
The indicators selected as relevant to be related to life expectancy are the following:

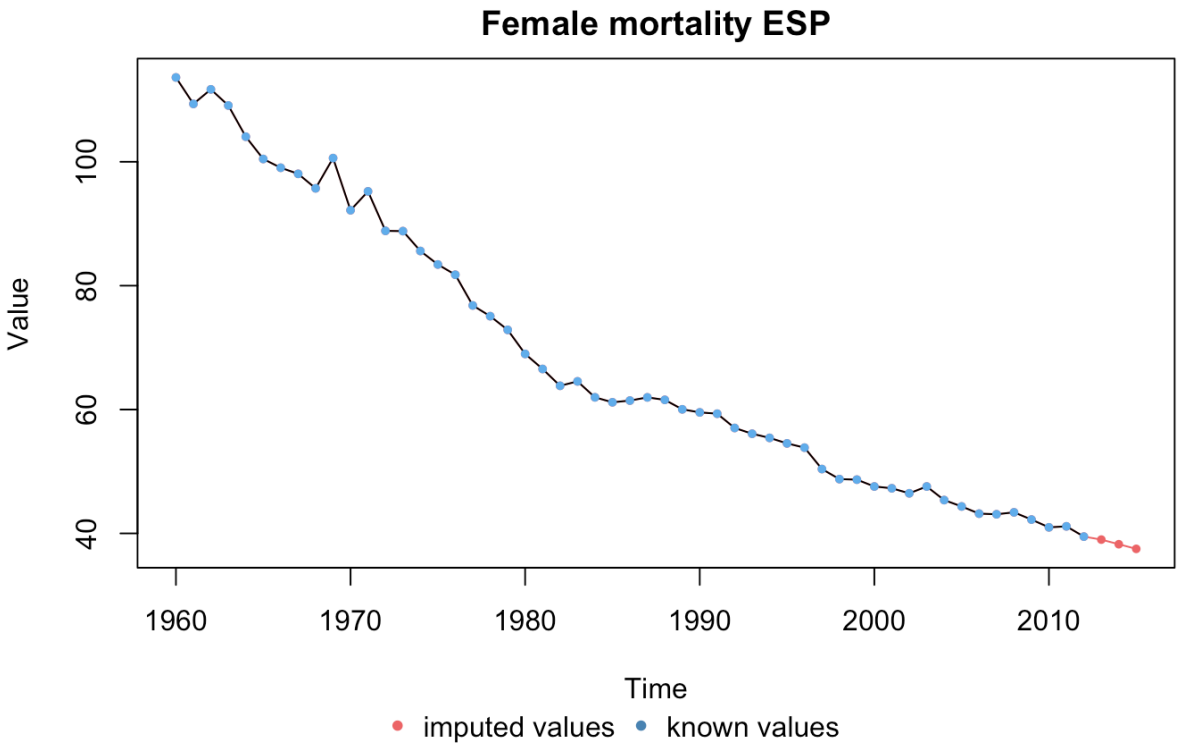
- CO2 emissions (metric tons per capita)
- Hospital beds (per 1,000 people)
- Mortality rate, adult, female (per 1,000 female adults)
- Mortality rate, adult, male (per 1,000 male adults)
- Mortality rate, infant (per 1,000 live births)
- Population ages 65 and above (% of total)
- Central government debt, total (% of GDP)
- Death rate, crude (per 1,000 people)
- GDP per capita (constant 2005 US$)
- Health expenditure, total (% of GDP)

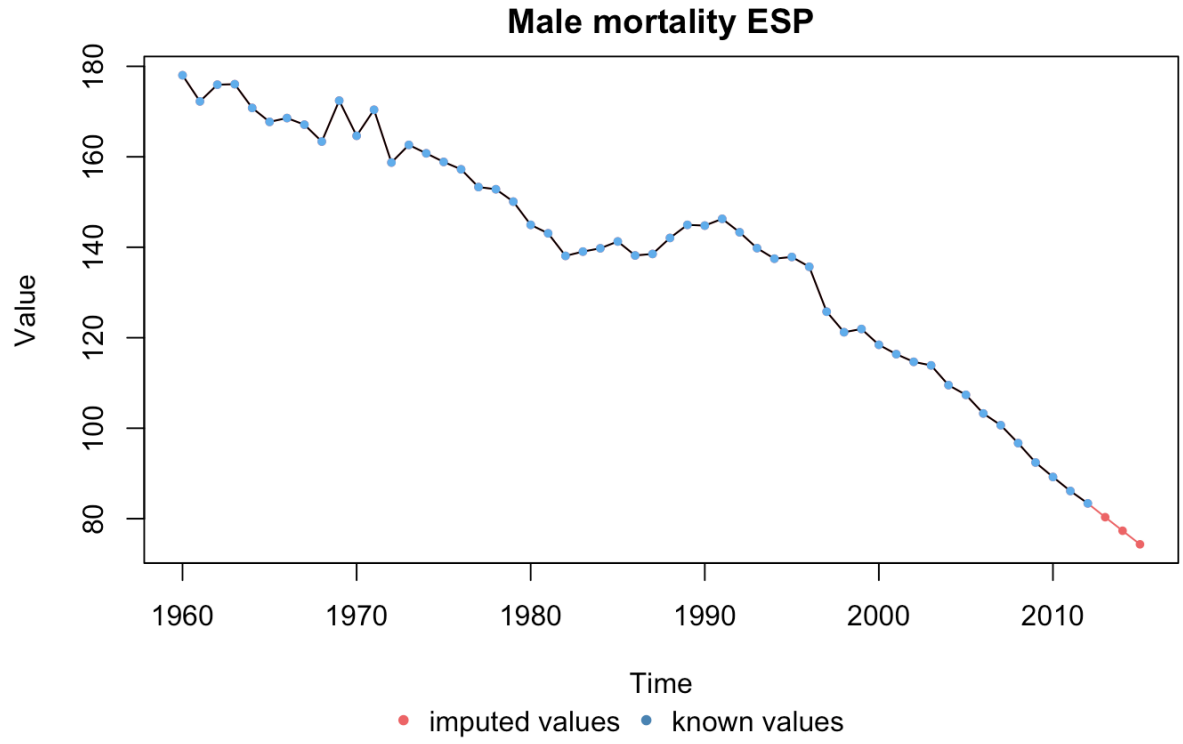
In the following plots the imputations done in each indicator are shown:

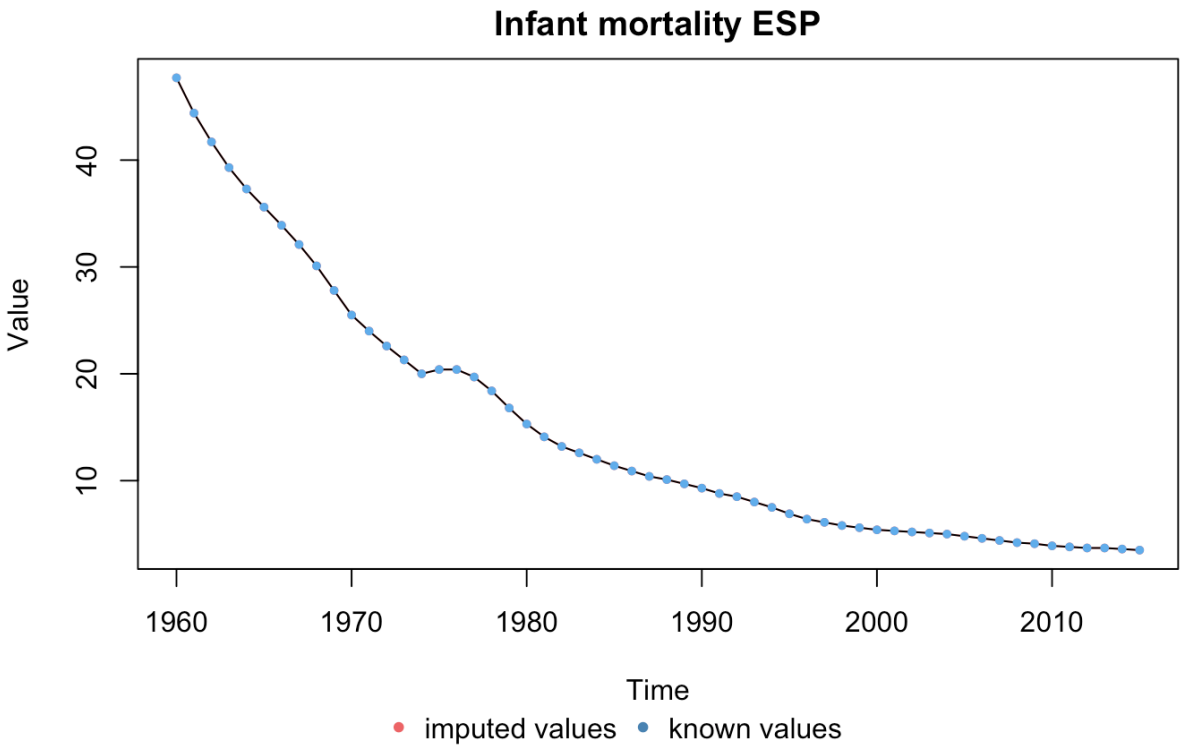


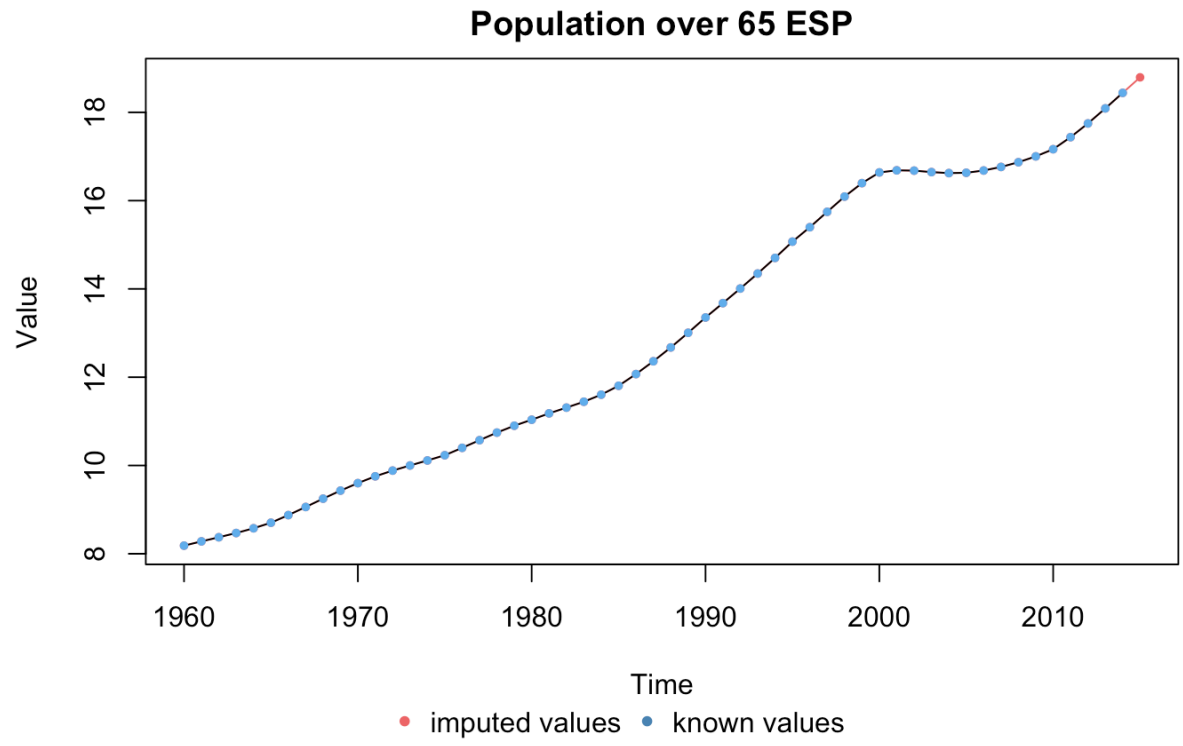

CO2 emissions ESP
Value
2
3
4
5
6
7
8
1960
1970
1980
1990
2000
2010
Time
imputed values
known values

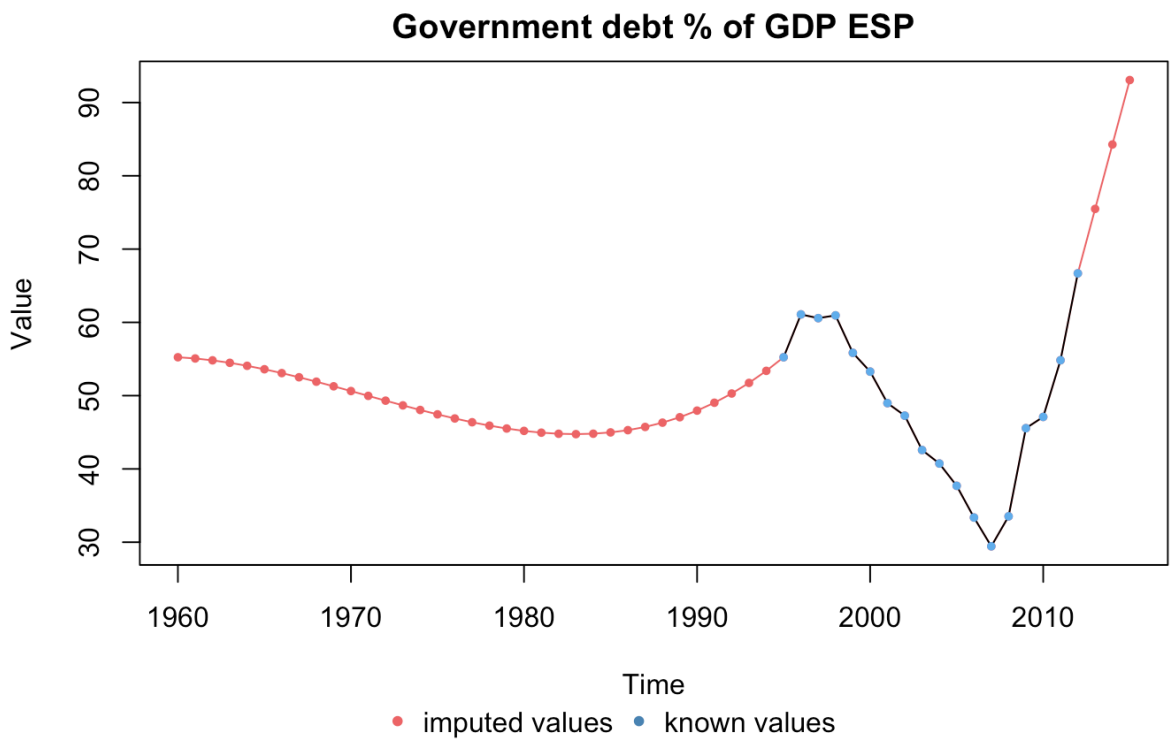

Hospital beds ESP
Value
3.0
3.5
4.0
4.5
5.0
5.5
1960
1970
1980
1990
2000
2010
Time
imputed values
known values

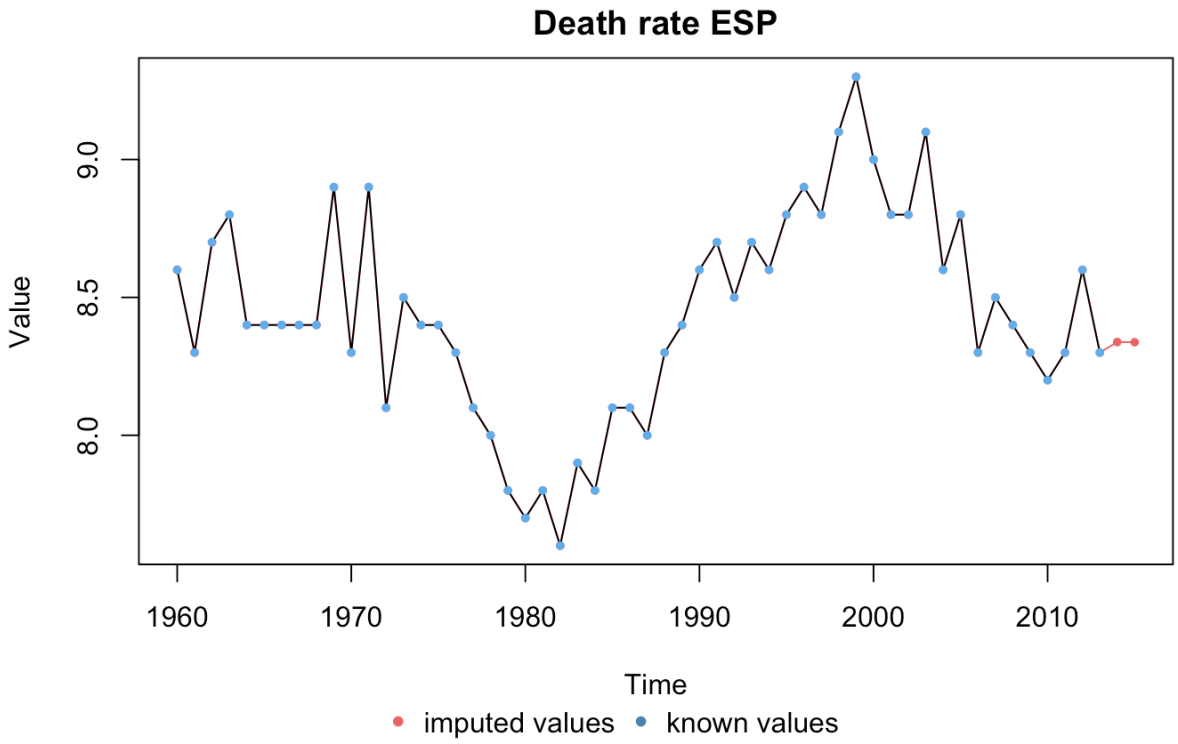

Female mortality ESP
Value
40
60
80
100
1960
1970
1980
1990
2000
2010
Time
imputed values
known values

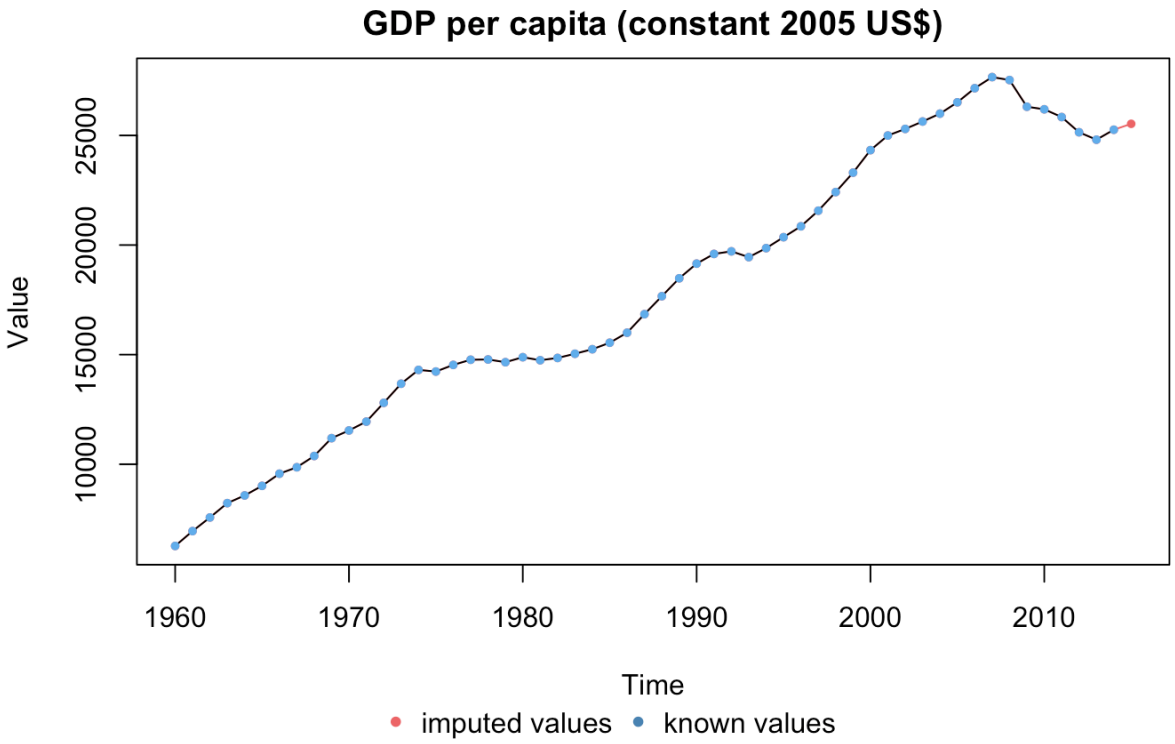

Male mortality ESP
Value
80
100
120
140
160
180
1960
1970
1980
1990
2000
2010
Time
imputed values
known values

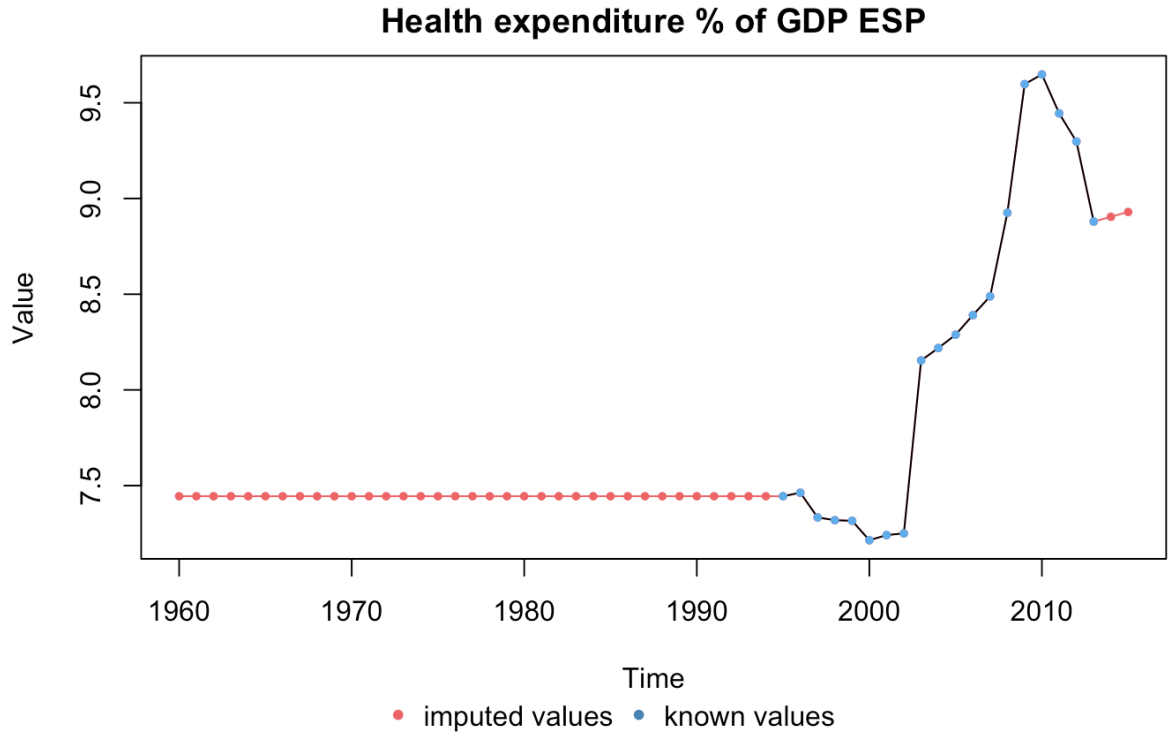

Infant mortality ESP
Value
10
20
30
40
1960
1970
1980
1990
2000
2010
Time
imputed values
known values


Population over 65 ESP
Value
8
10
12
14
16
18
1960
1970
1980
1990
2000
2010
Time
imputed values
known values


Government debt % of GDP ESP
Value
30
40
50
60
70
80
90
1960
1970
1980
1990
2000
2010
Time
imputed values
known values


Death rate ESP
Value
8.0
8.5
9.0
1960
1970
1980
1990
2000
2010
Time
imputed values
known values





These plots are for the indicators for Spain but these same plots for the indicators of each country are in the code, we won't show them all since they are all similar.

From these plot we can see that there are some indicators that have even more imputed values than real ones. We decided not to use some predictors that have too many imputed values.

The indicators discarded for each country were the following:

Spain: Governemnt debt(% of GDP), Health expenditure(% of GDP) and Hospital beds.

China: Governemnt debt(% of GDP) and Health expenditure(% of GDP).

India: Hospital beds, Government debt(% of GDP) and Health expenditure(% of GDP).

France: Government debt(% of GDP) and Health expenditure(% of GDP).

USA: Government debt(% of GDP) and Health expenditure(% of GDP).

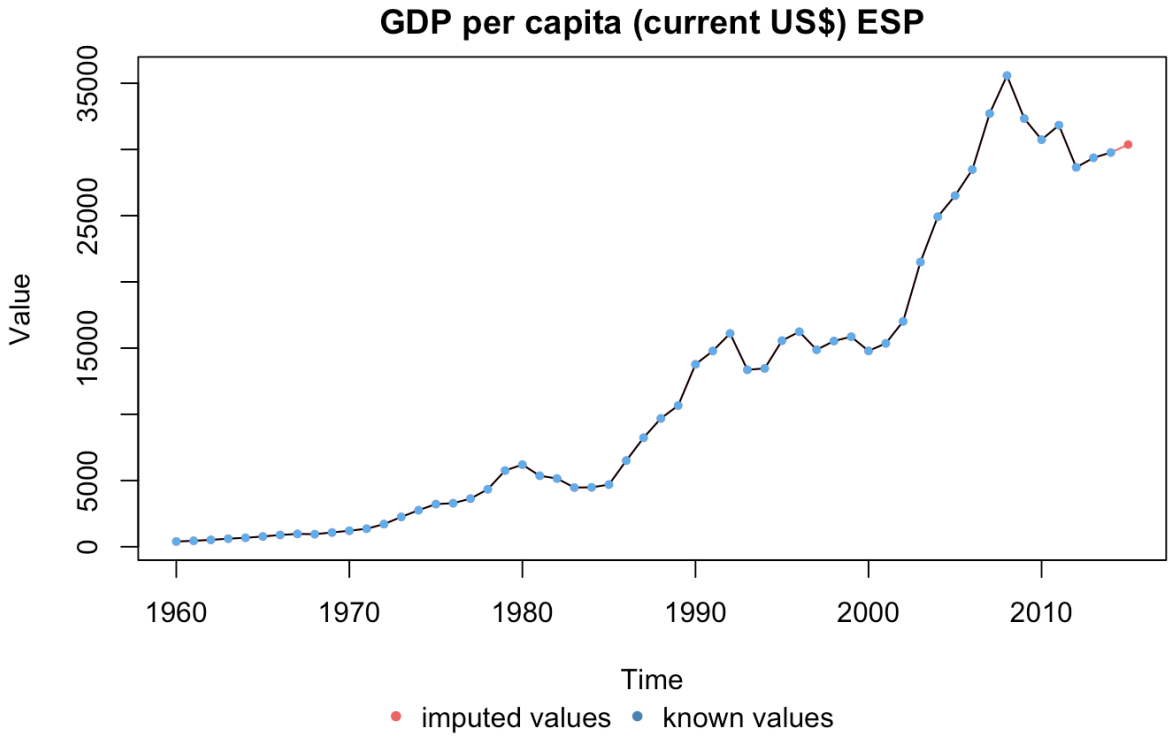
These predictors were deleted from the data sets of the countries to have the final data sets that will be used for the next part of the project.

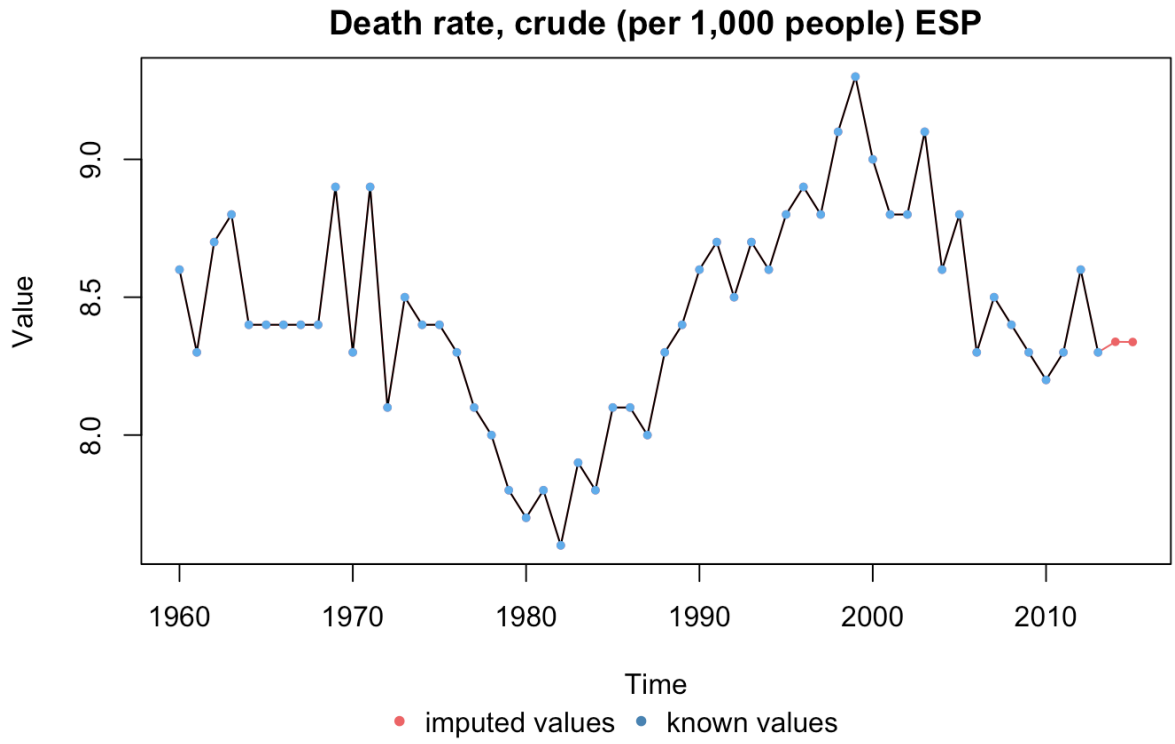
III.II. CO2 related to mortality

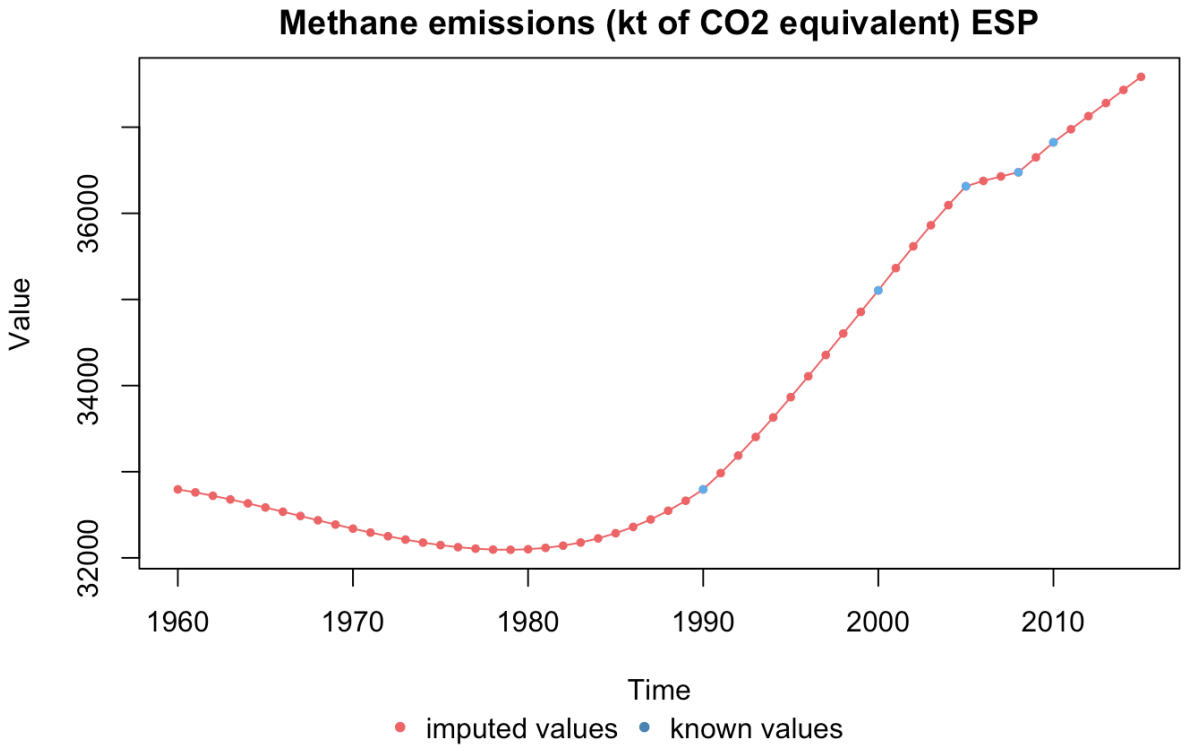
The indicators selected as relevant to be related to the mortality and the emissions are the following:

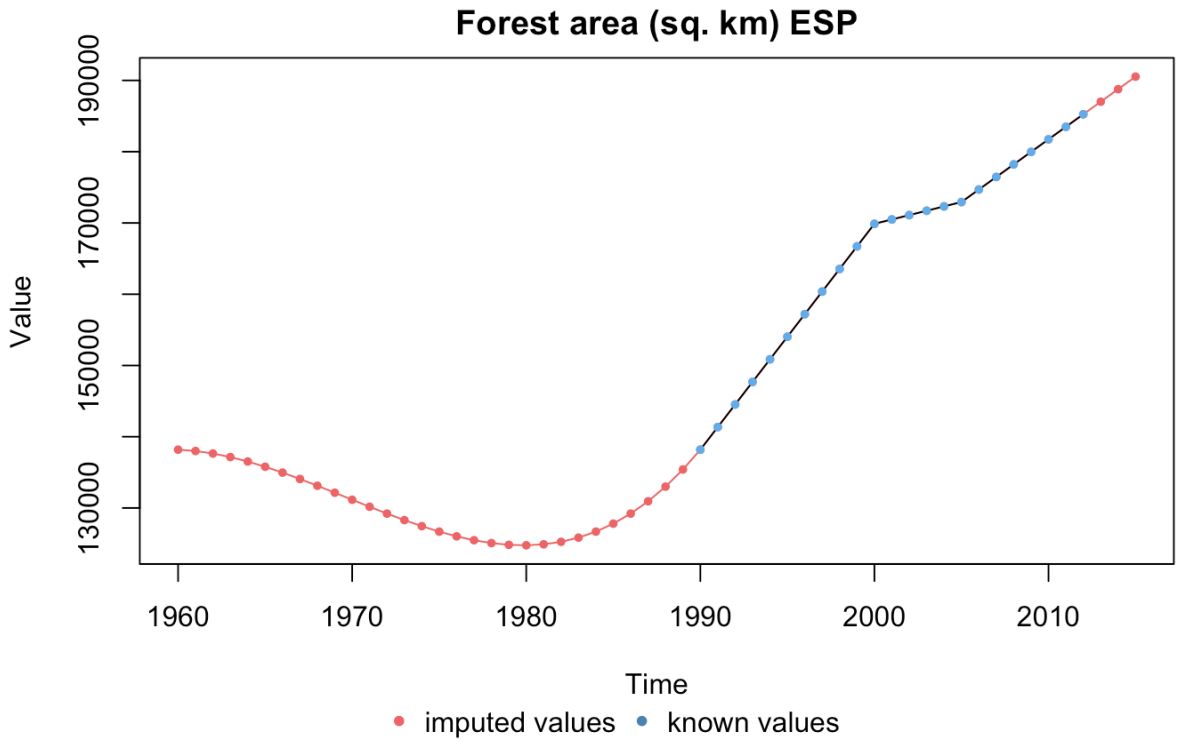
- CO2 emissions (metric tons per capita)
- Mortality rate, adult, female (per 1,000 female adults)
- Mortality rate, adult, male (per 1,000 male adults)
- Death rate, crude (per 1,000 people)
- GDP per capita (current US$)
- Forest area (sq. km)
- Methane emissions (kt of CO2 equivalent)

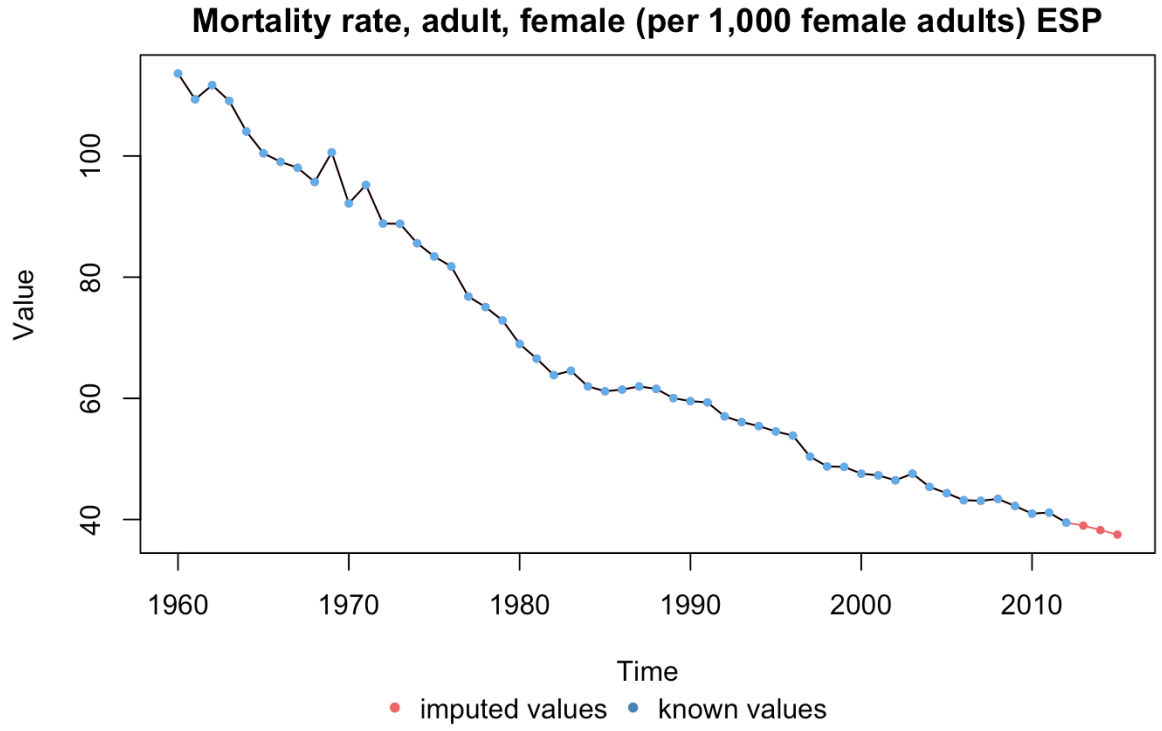
In the following plots the imputations done in each indicator are shown:

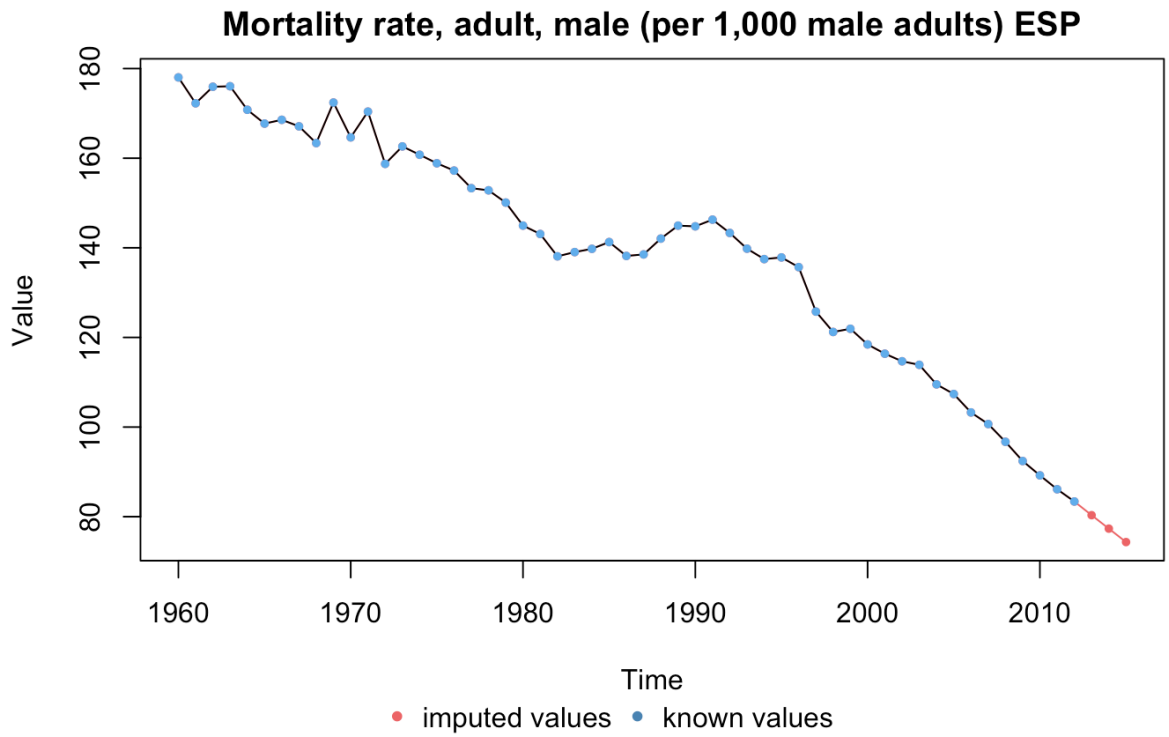


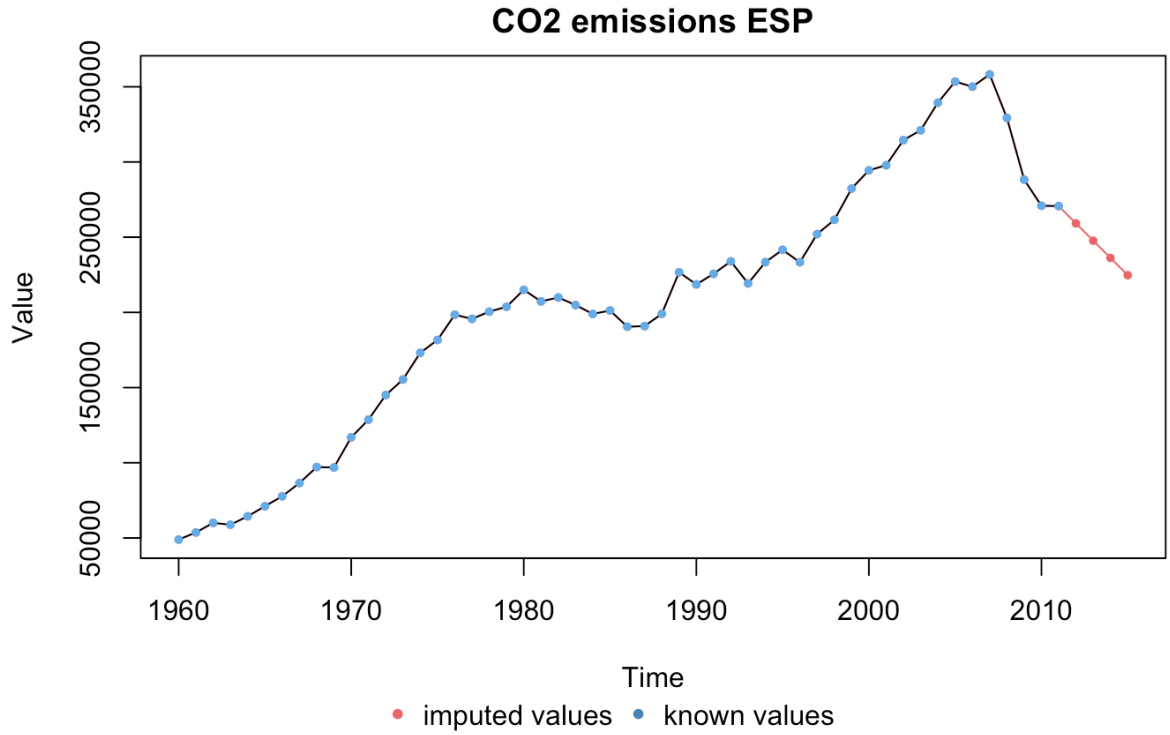

Death rate, crude (per 1,000 people) ESP
Value
8.0
8.5
9.0
1960
1970
1980
1990
2000
2010
Time
imputed values
known values


Methane emissions (kt of CO2 equivalent) ESP
Value
32000
34000
36000
1960
1970
1980
1990
2000
2010
Time
imputed values
known values


Forest area (sq. km) ESP
Value
130000
150000
170000
190000
1960
1970
1980
1990
2000
2010
Time
imputed values
known values


Mortality rate, adult, female (per 1,000 female adults) ESP
Value
40
60
80
100
1960
1970
1980
1990
2000
2010
Time
imputed values
known values





These plots are for the indicators for Spain but as well as with life expectancy we provide in the code the same plots for the indicators of each country.

It is important to remove the indicators that have more imputed than real values in order to create a more realistic model.

The indicators discarded for each country were the following:

Spain: Forest Area and Methane emissions.

China: Forest Area and Methane emissions.

India: Forest Area and Methane emissions.

France: Forest Area and Methane emissions.

USA: Forest Area and Methane emissions.

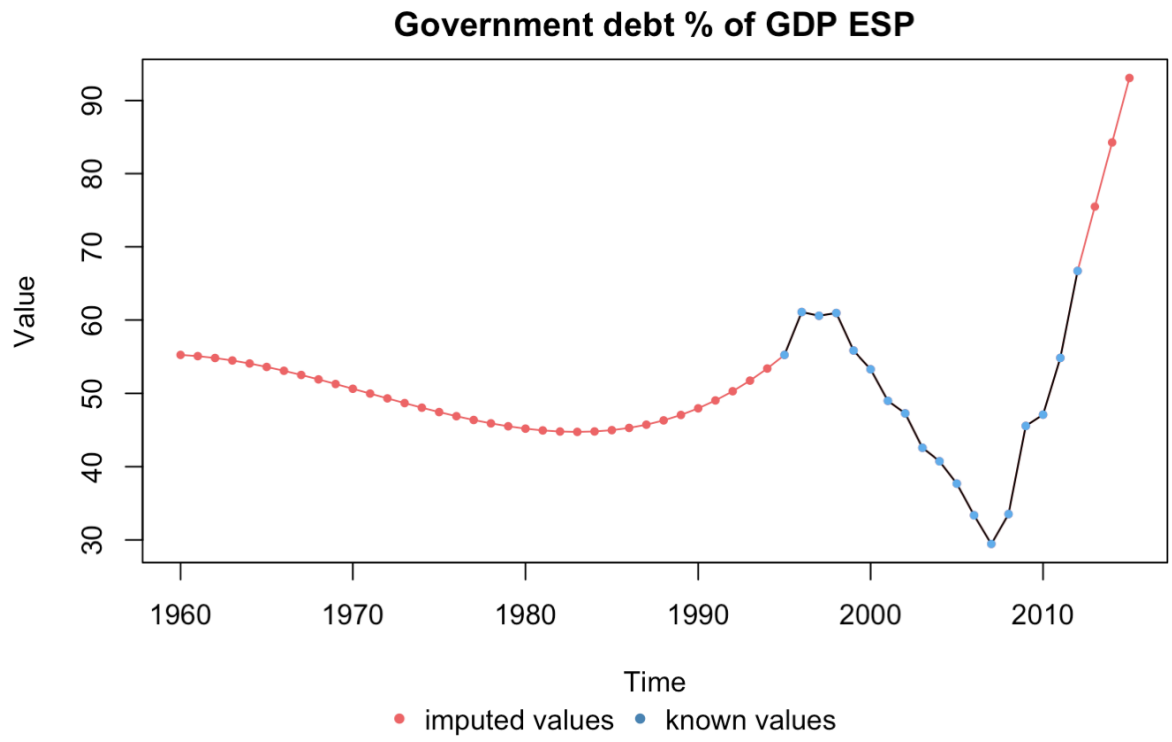
As we have said these predictors should be deleted from the data sets of the countries for the next steps of the project.

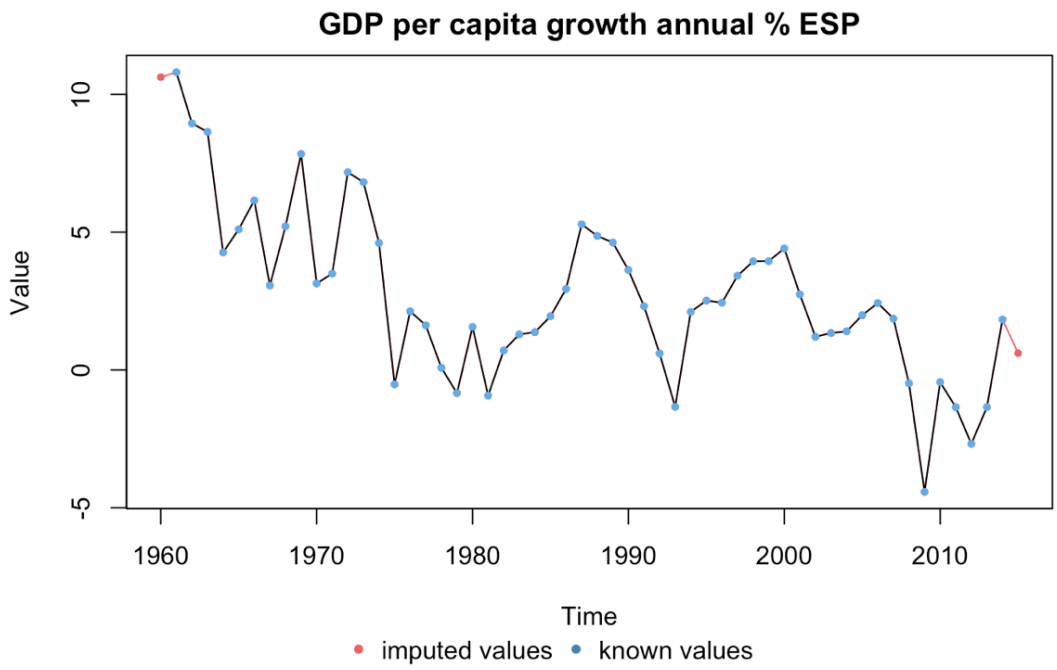
III.III. Household Consumption

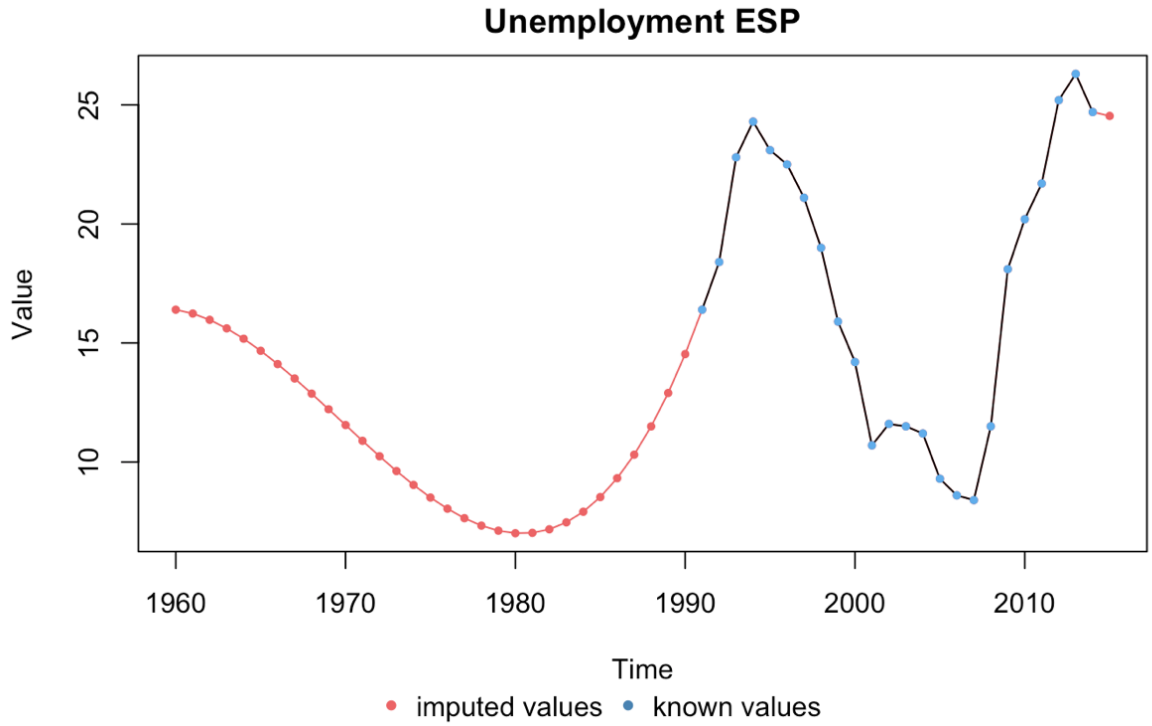
The indicators selected as relevant to be related to life expectancy are the following:

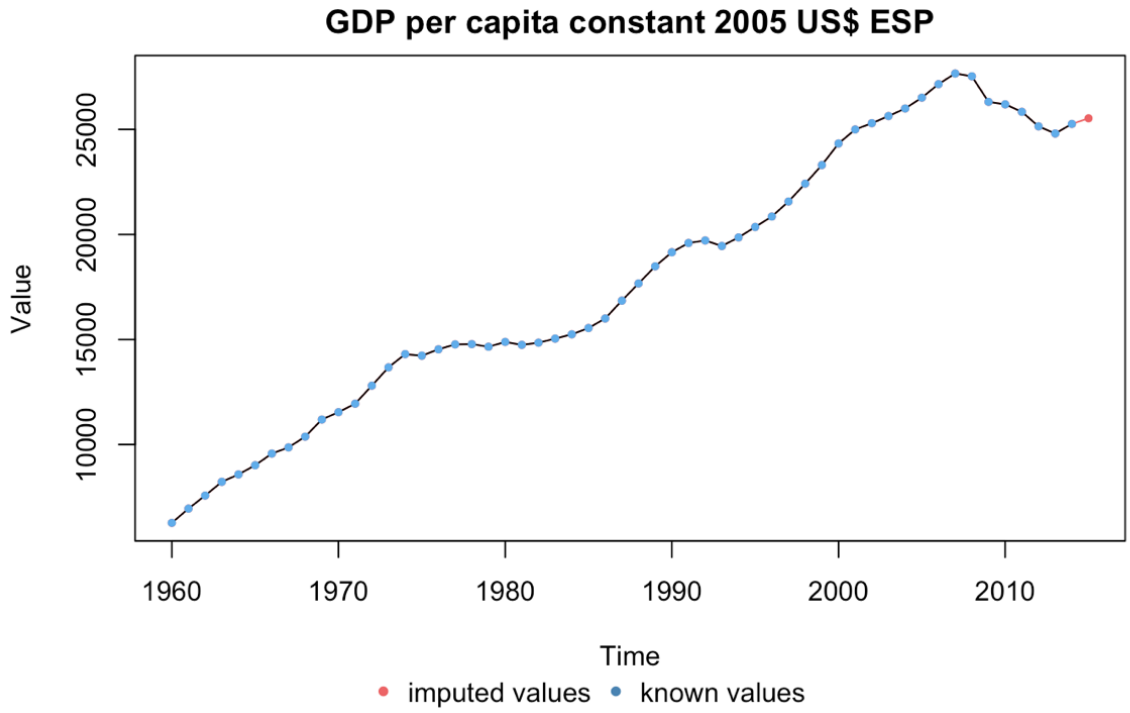
- GDP per capita constant 2005 US$
- Inflation, consumer prices (annual %)
- Unemployment, total (% of total labor)
- Government debt, total % of GDP
- GDP growth (annual %)
- Inflation, consumer prices (annual %)

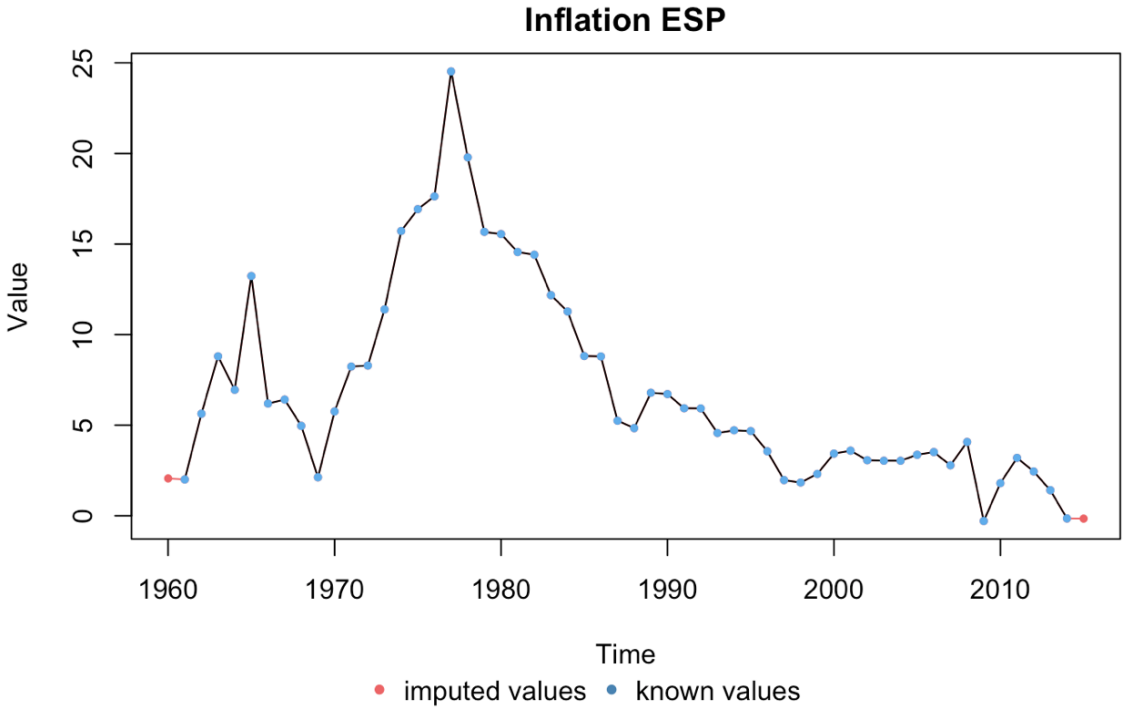
In the following plots the imputations done in each indicator are shown:

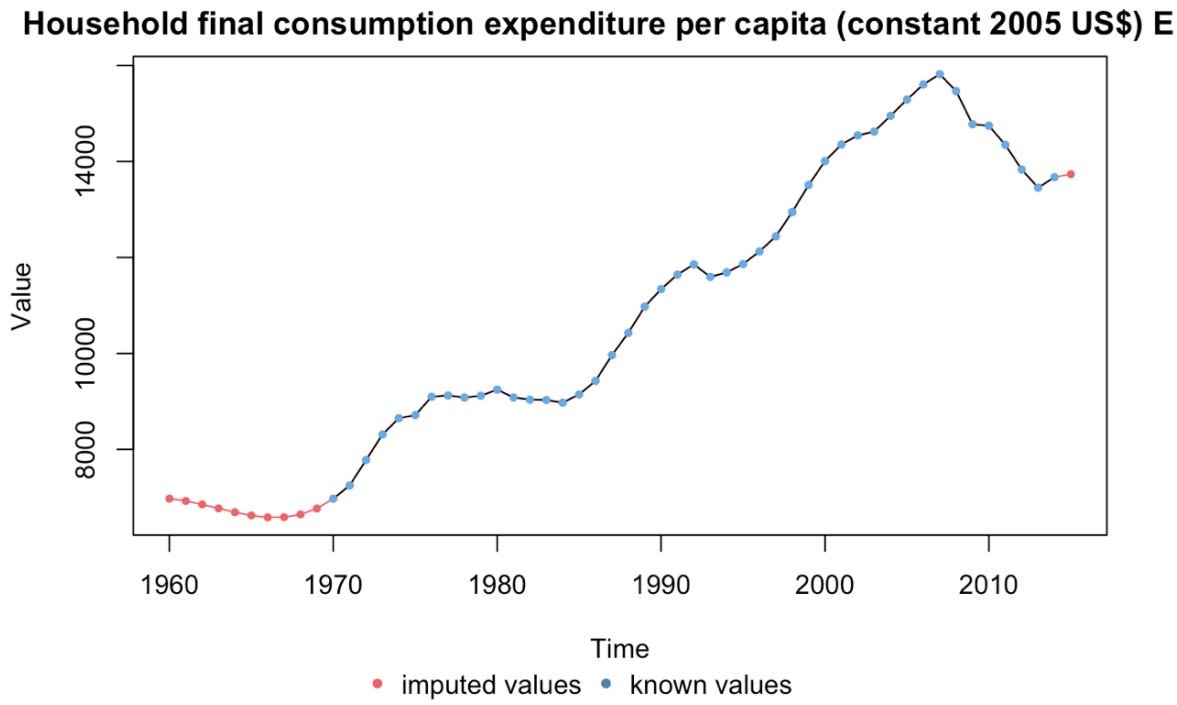



GDP per capita growth annual % ESP
Value
10
5
0
-5
1960
1970
1980
1990
2000
2010
Time
imputed values
known values


Unemployment ESP
Value
25
20
15
10
1960
1970
1980
1990
2000
2010
Time
imputed values
known values


GDP per capita constant 2005 US$ ESP
Value
10000
15000
20000
25000
1960
1970
1980
1990
2000
2010
Time
imputed values
known values


Inflation ESP
Value
0
5
10
15
20
25
1960
1970
1980
1990
2000
2010
Time
imputed values
known values



The indicators discarded for each country were the following:

Spain: Unemployment and Government debt.

China: Unemployment, Government debt and Inflation.

India: Unemployment and Government debt

France: Unemployment and Government debt

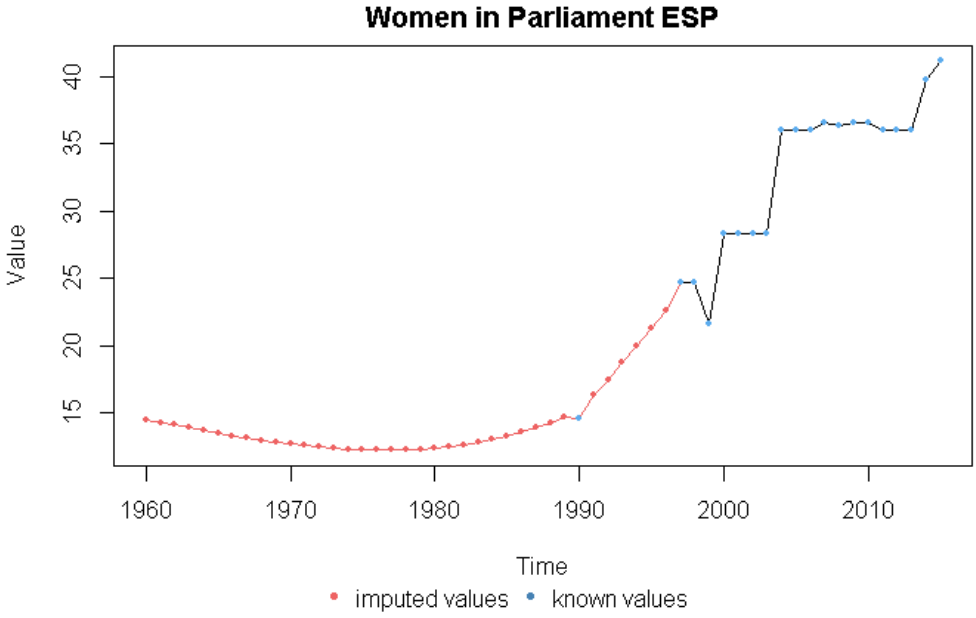
USA: Unemployment and Government debt

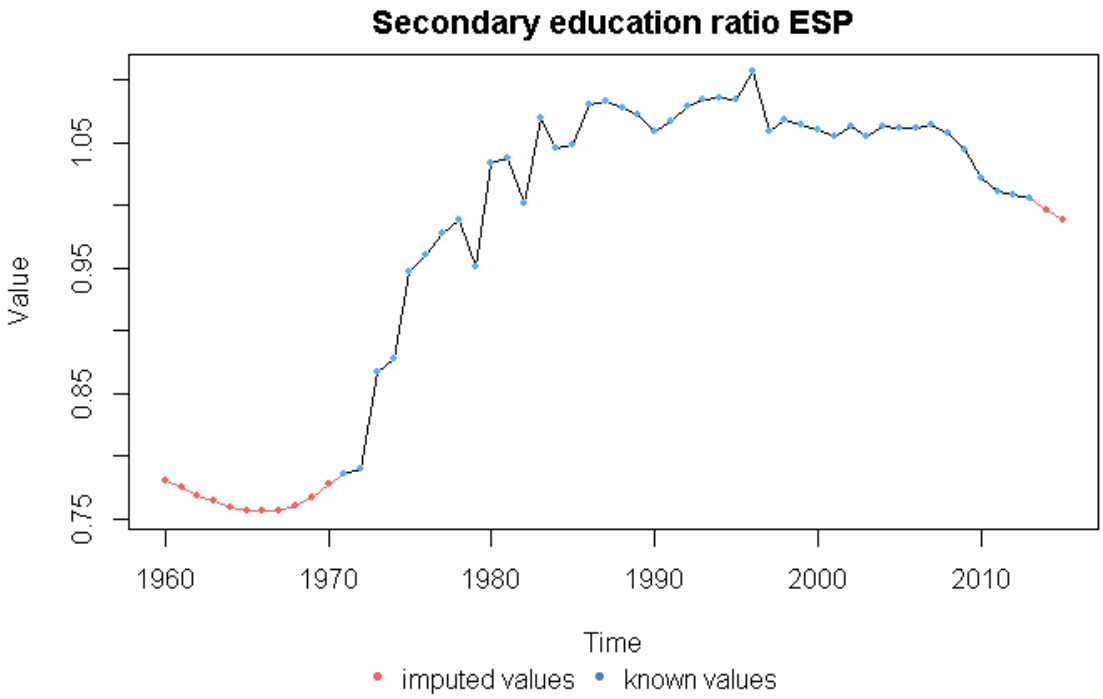
III.IV. Gender equality

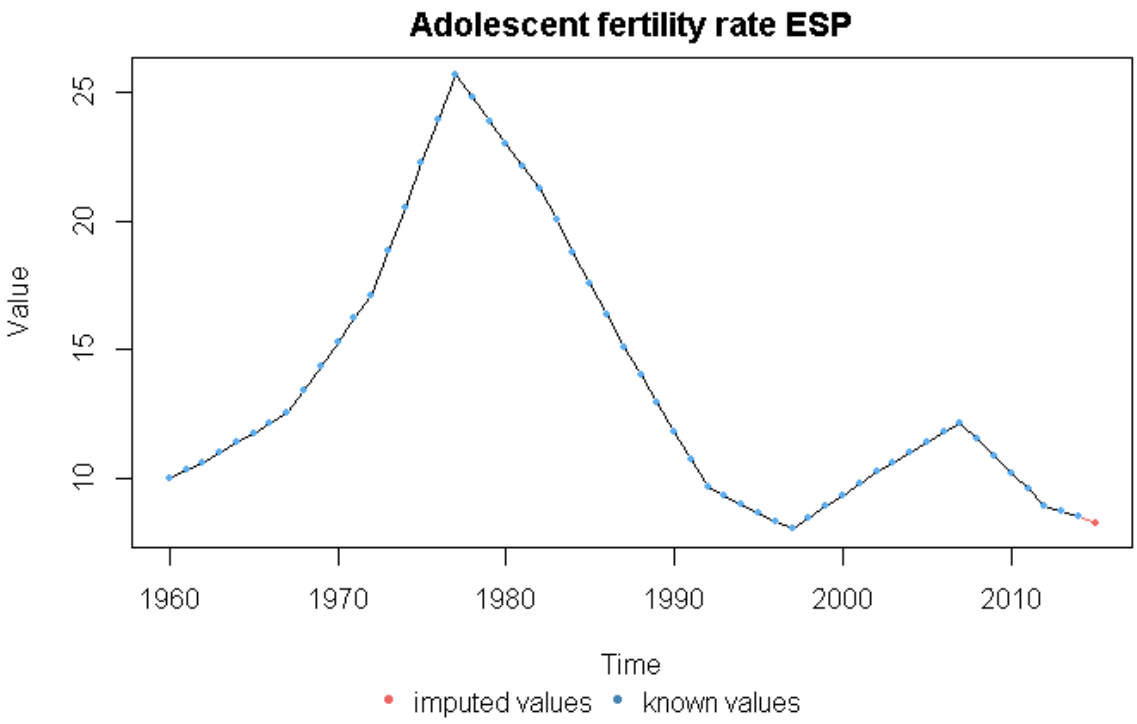
The indicators selected as relevant to be related to gender equality are the following:

- Proportion of seats held by women in national parliaments (%).
- Gross enrolment ratio, secondary, gender parity index (GPI).
- Adolescent fertility rate (births per 1,000 women ages 15-19).

In the following plots the imputations done in each indicator are shown:



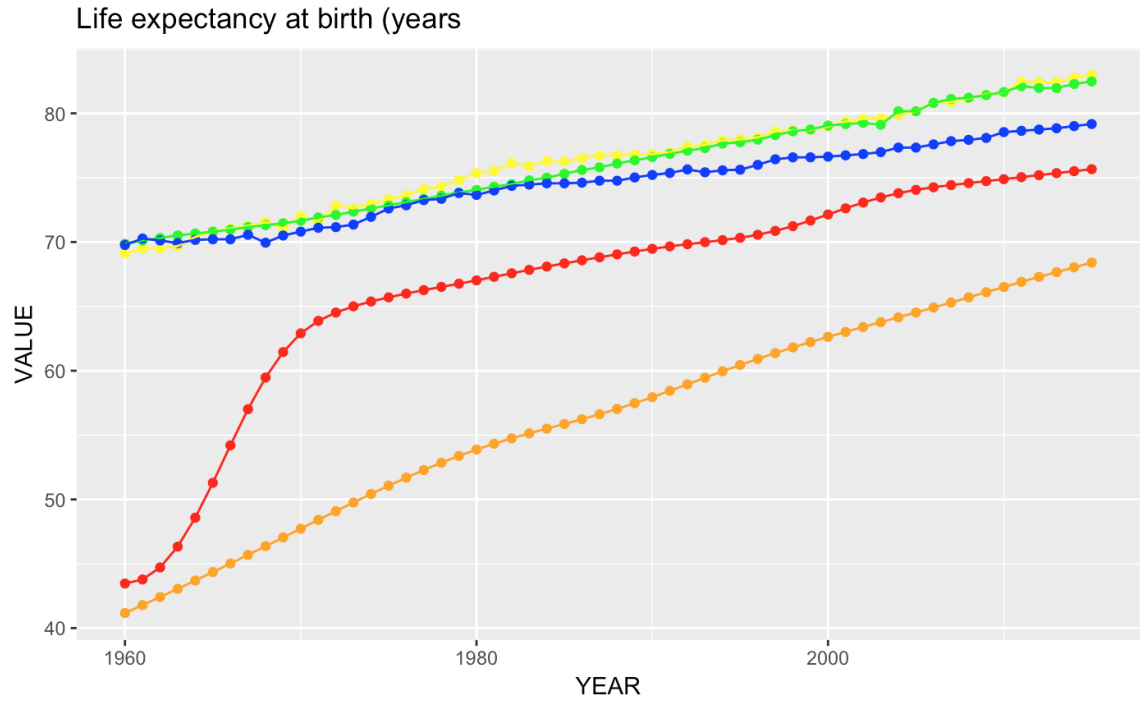



Adolescent fertility rate ESP
Value
10
15
20
25
1960
1970
1980
1990
2000
2010
Time
imputed values
known values

IV. DATA ANALYSIS

To analyze the data we are going to show in a plot the evolution of each indicator, the one we want to study and predict and the rest we want to use as predictors, for the five countries. The point is to find if any of the indicators we want to use as predictors have a trend that can be related to the trend of the indicator we have considered as response and try to confirm if our guess was close to the reality with the models' information.

IV.I. Life expectancy



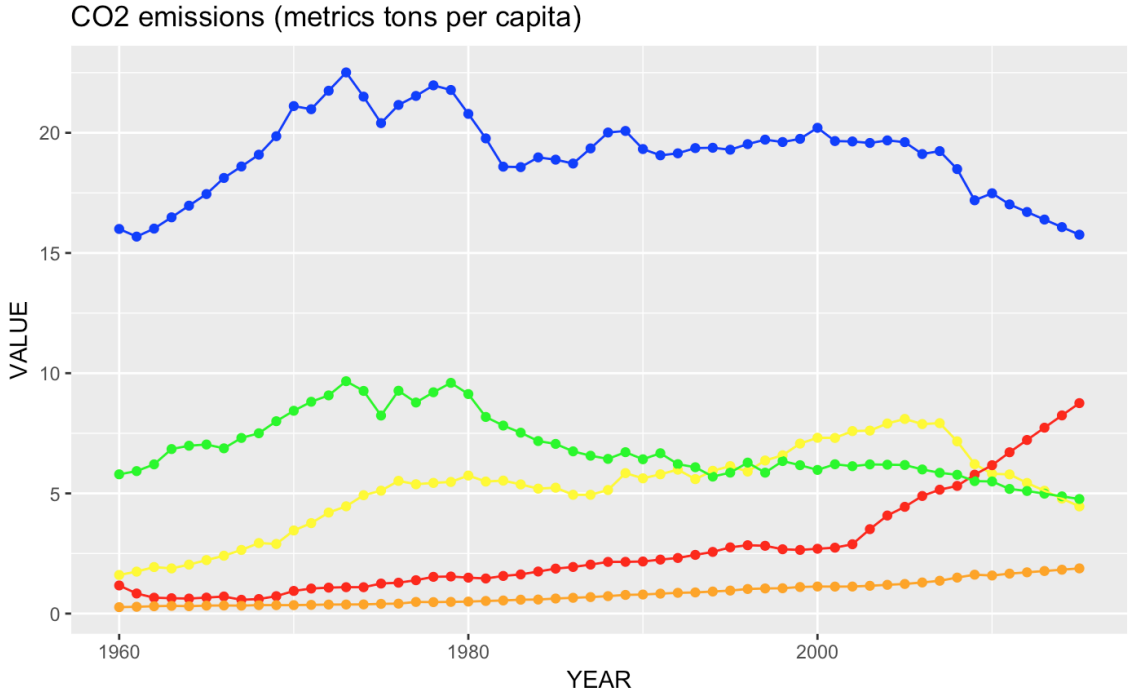
In this plot we can see the trend that life expectancy has had from 1960 to 2015 for all the countries.

The trend for Spain, USA and France are quite similar but life expectancy in Spain and France is over 80 year while in USA it is not.

Then in China, life expectancy grew exponentially in the 1960s, from less than 45 years to more than 60, to follow a linear growing trend after that with a more significant slope what has made life expectancy in China to be over 75 years and get closer to the values for the previous countries.

Finally, India's life expectancy was lower than China's in the 1960s but it didn't had an exponential growth, it has follow a linear growth over all the years and, although its sloper is the biggest, India's life expectancy is the lowest of all the countries we are considering and it is below 70 years.

In the plots for the next indicators, we want to see if these trends in life expectancy are somehow related to the trends of these others indicators as we said before.



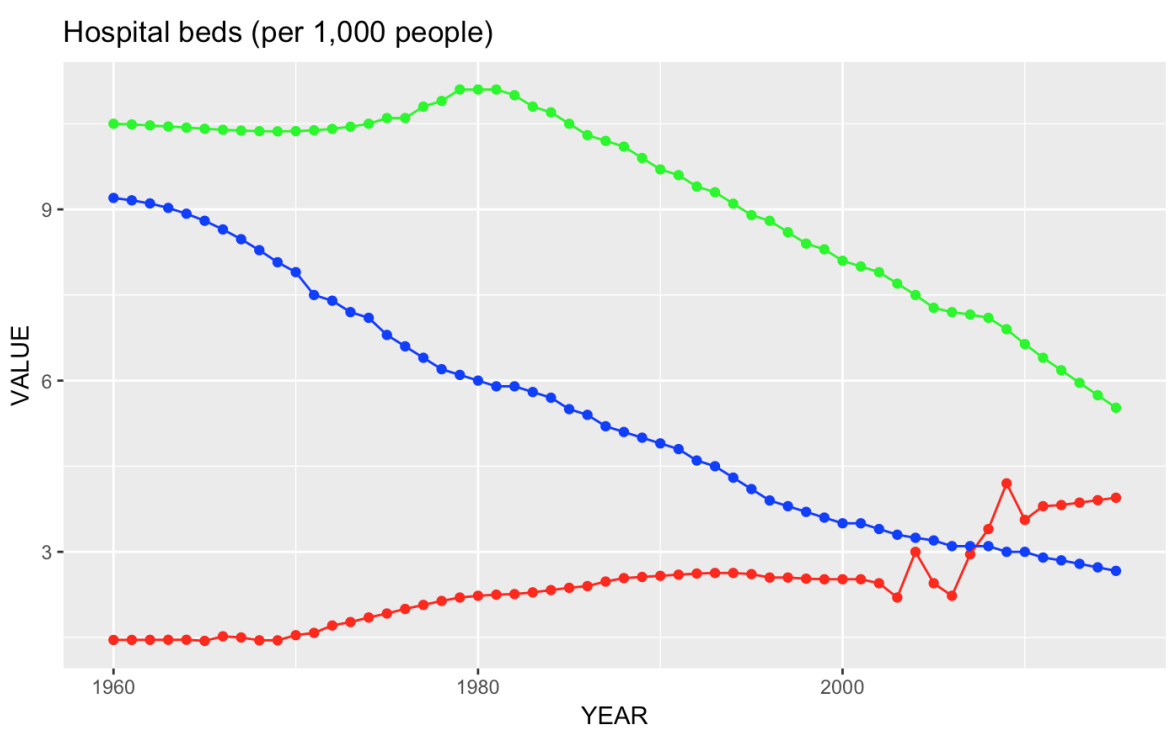
In this plot the CO2 emission (metrics tons per capita) are shown.

USA is by far the country with the most emissions.

France was historically in second place but Spain caught up in the 1990s but they are both quite equal since 2010 and following a decreasing tendency.

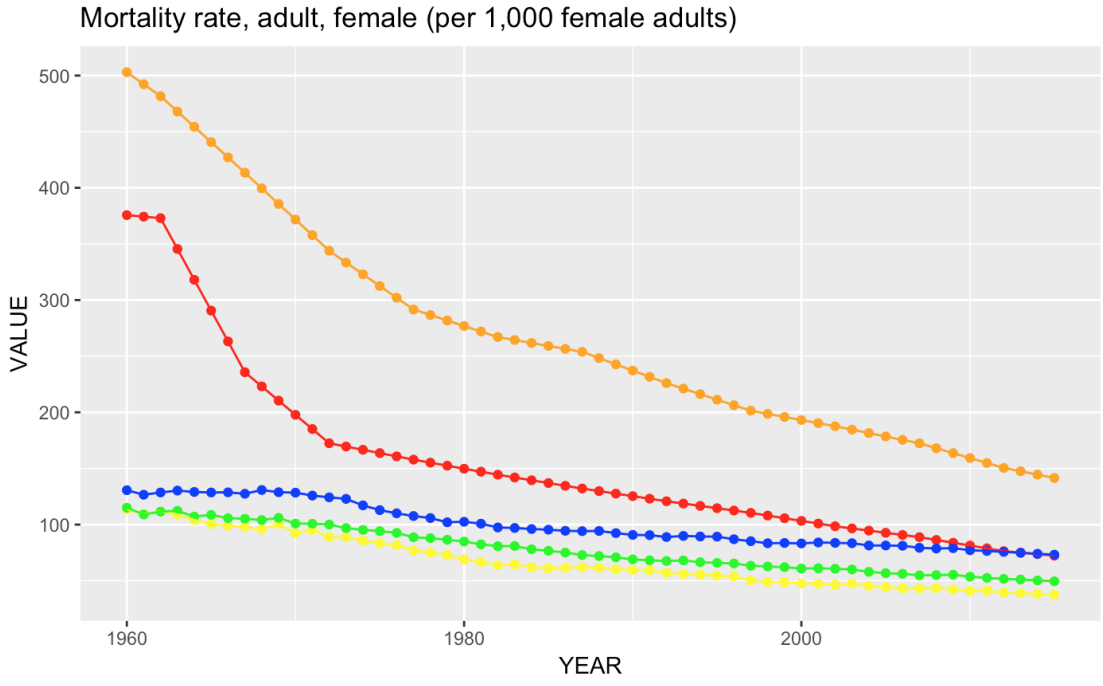
Both China and India had really low values in 1960 and had been following a slowly increasing tendency with China over India. Then, after 2000 China had a significant increment on its tendency having more emissions than France and Spain after 2010 and keeping this tendency.

There doesn't seem to be a connection between these values of CO2 emissions and life expectancy since the two countries with the highest life expectancy, France and Spain, aren't the countries with more emissions or the ones with less emissions.

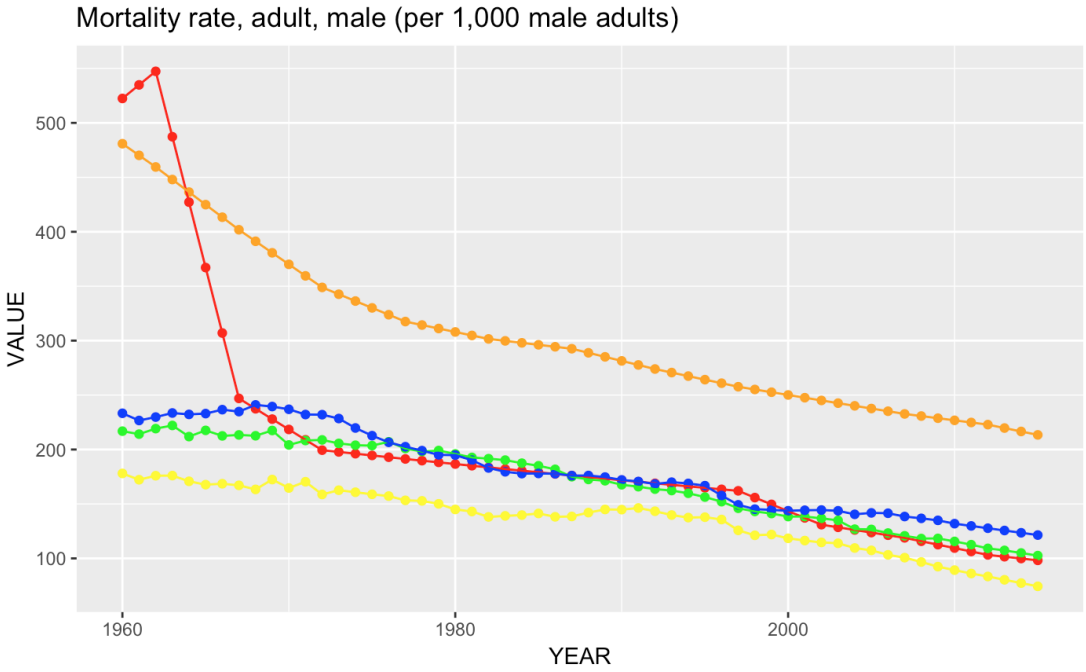


The number of hospital beds (per 1000 people) is shown in this plot. Only USA, France and China are represented because there weren't enough values for Spain and India.

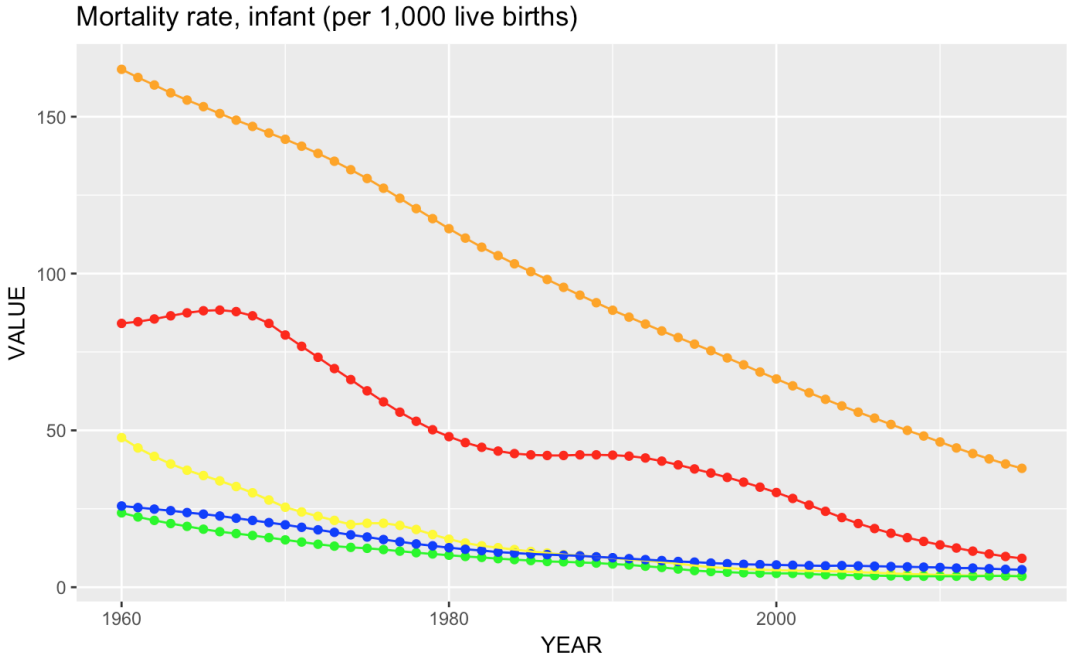
Both France and USA follow a decreasing tendency while, as we saw before, life expectancy was increasing in these countries. However, China has an increasing tendency of hospital beds and his life expectancy was increasing as well. France has the lowest hospital beds per 1000 people but it has the highest life expectancy so we may think that having more hospital beds doesn't imply a higher life expectancy.



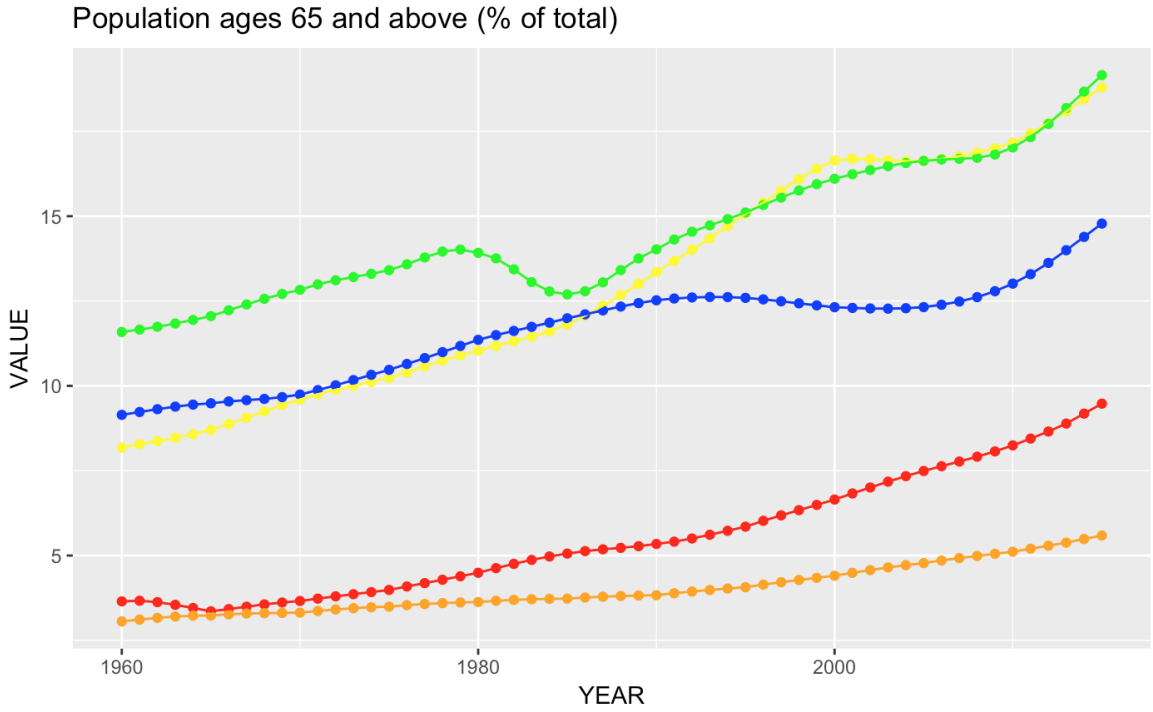
In this plot, the female mortality rate (per 1000 females) is shown. Spain, France and USA follow a similar decreasing trend, China had a huge decrease in the 1960s and then followed a similiar trend to the previous countries and it is at the same values now.
India has had an important decrease as well but there is still a gap between this country and the others. In this case we can relate the mortality rate to the life expectancy since it seems that the countries with a higher life expectancy have a lower female mortality rate.



This plot is similar to the previous one but it represents the male mortality rate now. The trends are quite similar to the previous plot and we can see that the values of male mortality were higher in the past in all countries but they are quite the same nowadays.



The mortality rate in infants (per 1000 births) is shown in this plot. All countries follow a decreasing trend with all but India having really low values nowadays although India is following a more decreasing tendency.

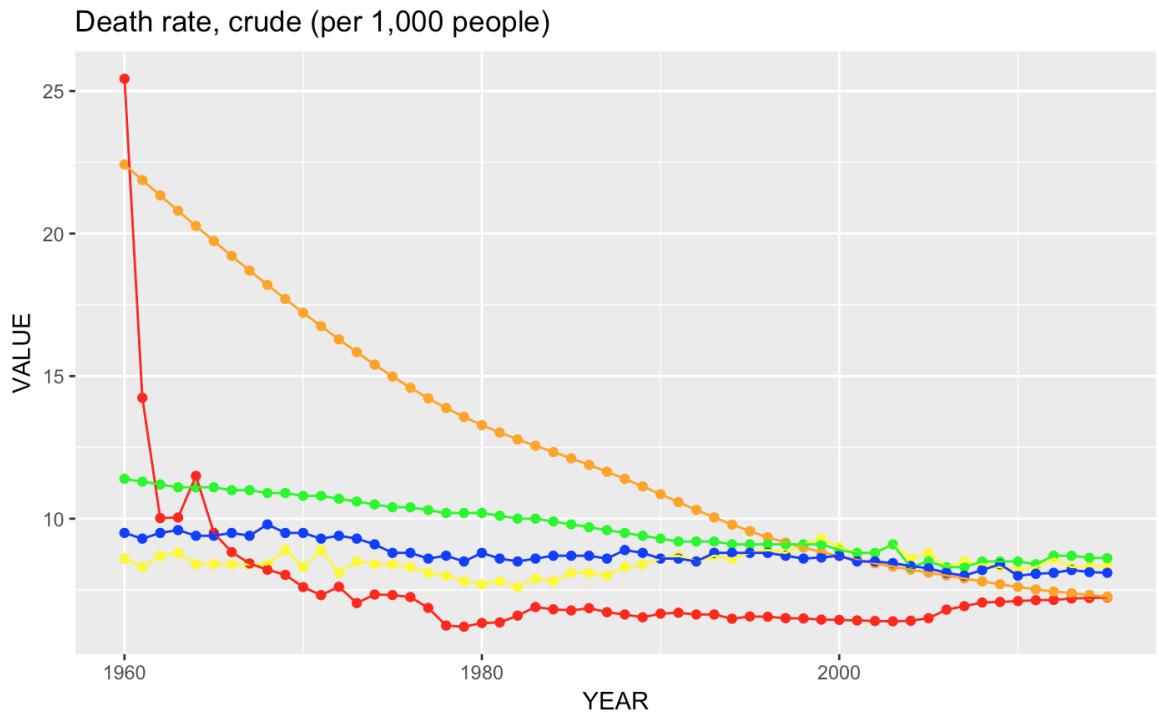


The population with ages 65 and above(% of total) is represented in this plot.

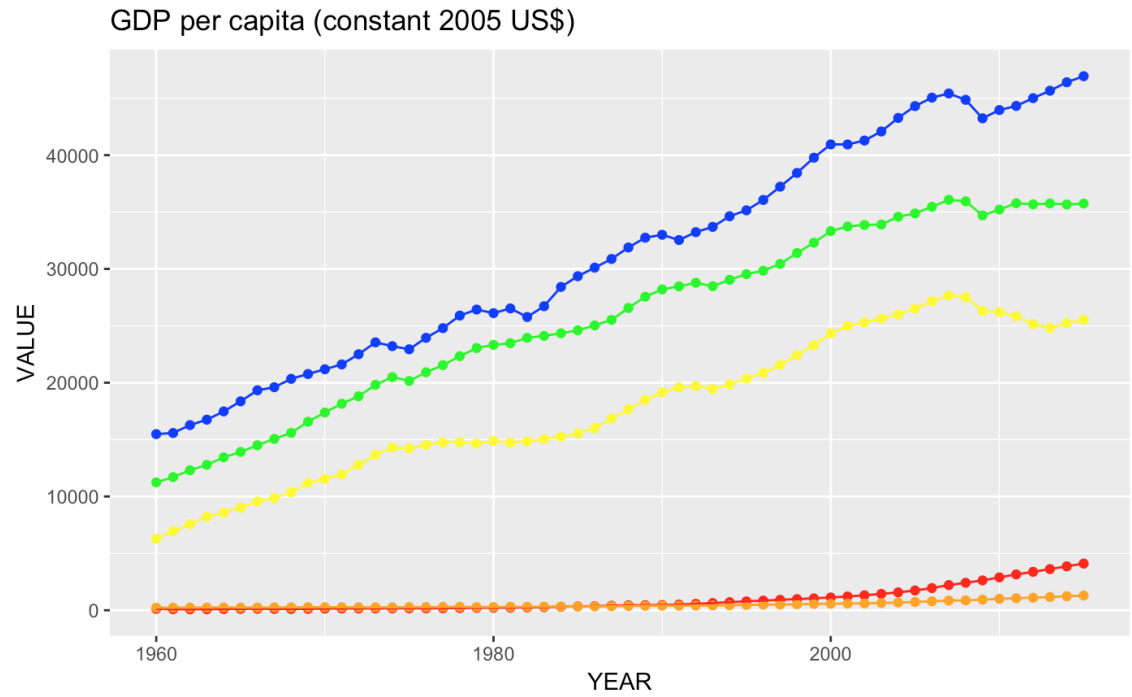
Spain and France are the countries with the highest values what makes sense since they have the highest life expectancy but USA is a bit far from them even if it has a similar life expectancy.

Then China and India are far away from the others and the follow an increasing tendency as well but with a smaller slope.

This makes sense because even if their life expectancy is nowadays closer their population is still younger.



The death rate (per 1000 people) shown in this plot has been very similar for all countries except India since mid 1960s and India joined the same trend in the mid 1990s.

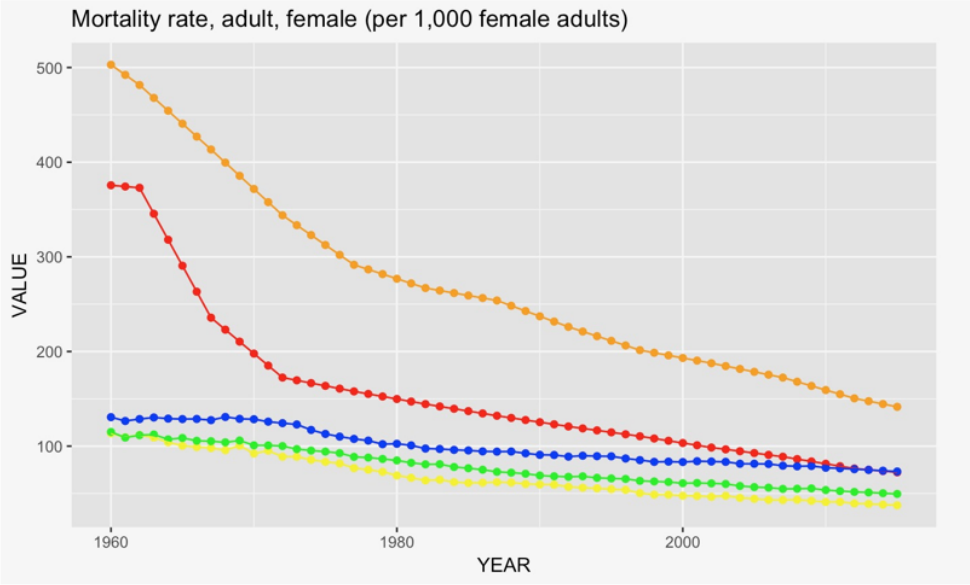


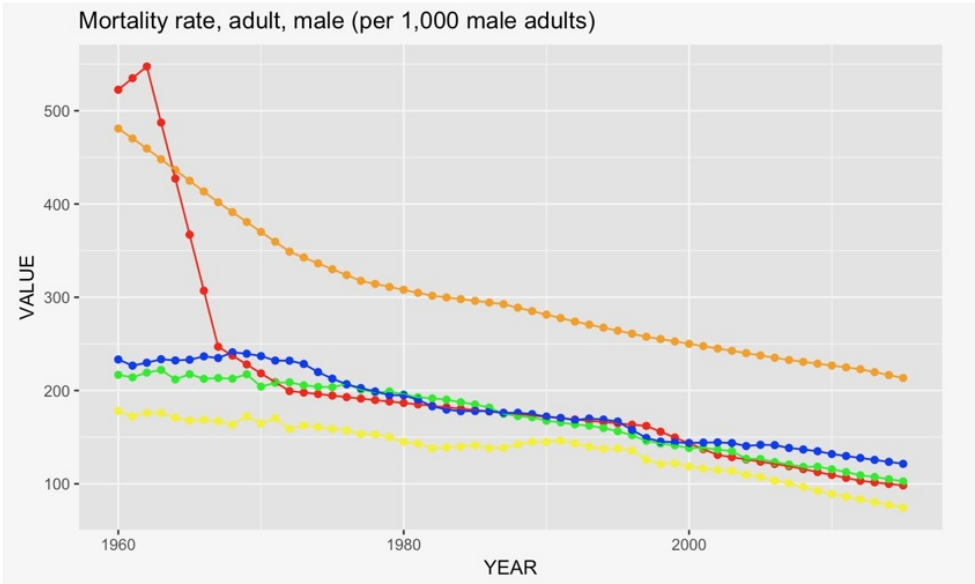
A priori, we could think that the more GDP per capita a country has the greater life expectancy it may have since people in the country could afford a better life. However, it does not seem the case since USA is the country with the greatest GDP per capita and France is the second but Spain has a greater life expectancy than the previous two and France greater than USA as well.

Also, the huge difference of GDP per capita between these three countries and China and India does not correspond with the difference in the life expectancy that is considerably lower at least with China.

IV.II CO2 related to mortality

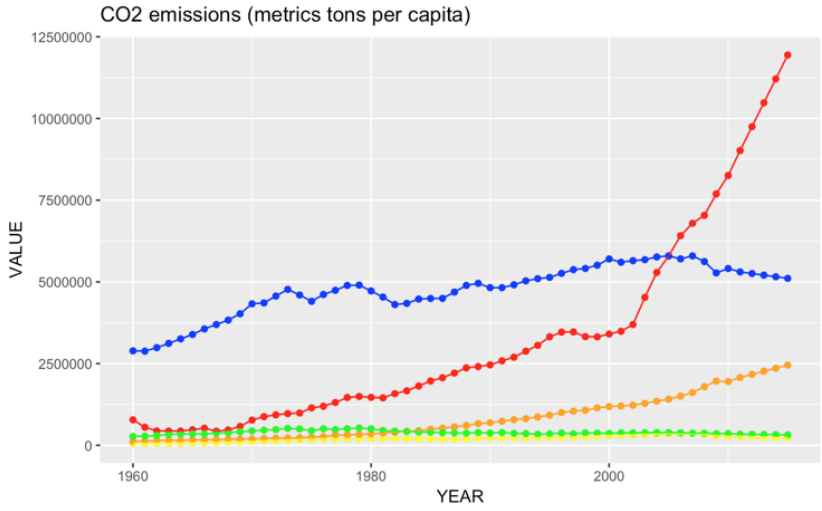
Taking a general vision of the first two plots we can see Spain, France and USA follow a similar decreasing trend of the mortality in both men a women. China had a huge decrease in the 1960s in both sex achieving the same values nowadays.
India has had an important decrease as well but there is still a gap between this country and the other four.

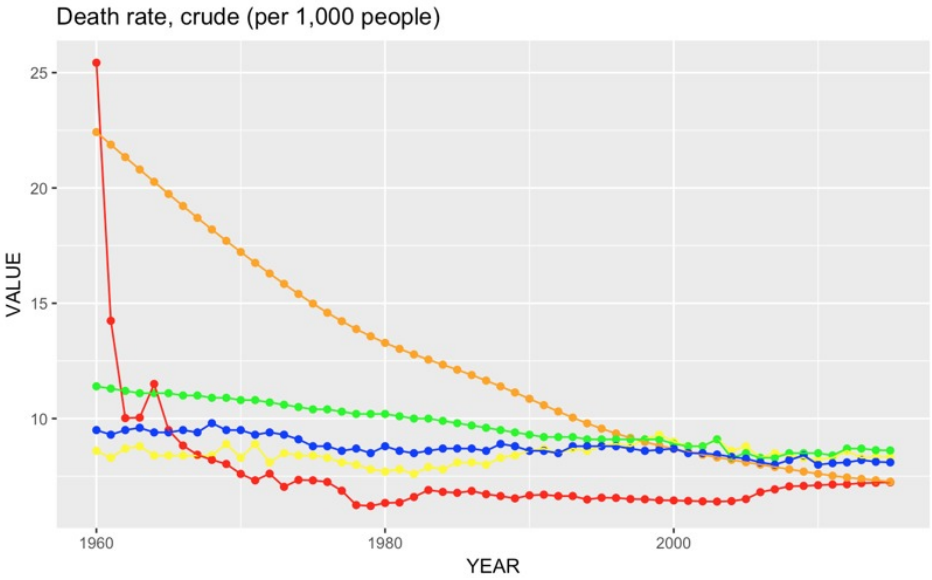


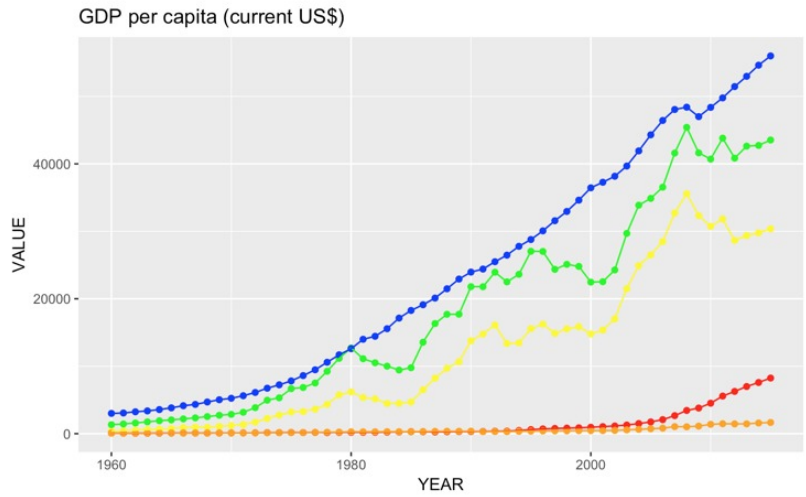


From the rest of the plots we can see that there not seem to be a direct relation between the mortlity and the C02 emission because while the mortality has decreased between 1960 and 2015 the emissions are more numerous nowadays than 15 years ago.

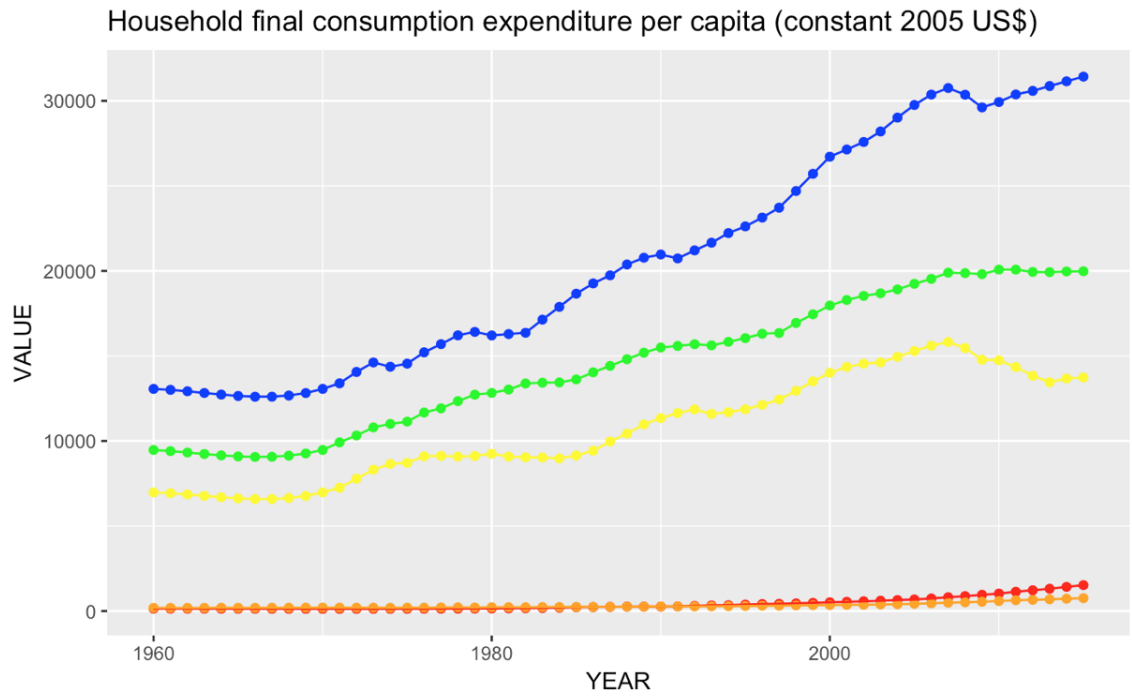
On the one hand if we have a look to the GDP plot we can appreciate a huge increase since in Spain, France and USA and the mortality has decreased, so this could mean that with more money in the country people take care a little bit more about their health. On the other hand the GDP of India and China has remain in the same values and their mortality is the one that has decreased more, so there not seem to be a direct relationship in every country.


CO2 emissions (metrics tons per capita)
12500000
10000000
7500000
5000000
2500000
0
VALUE
1960
1980
2000
YEAR

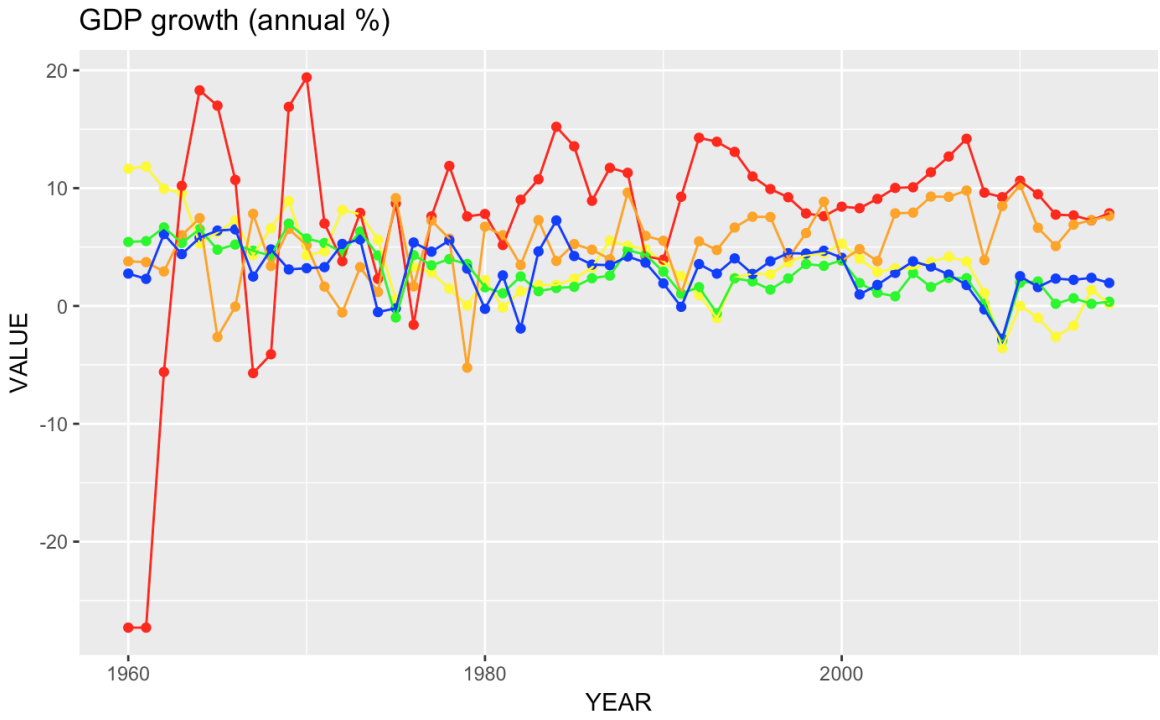

Death rate, crude (per 1,000 people)
25
20
15
10
VALUE
1960
1980
2000
YEAR


GDP per capita (current US$)
40000
20000
0
VALUE
1960
1980
2000
YEAR

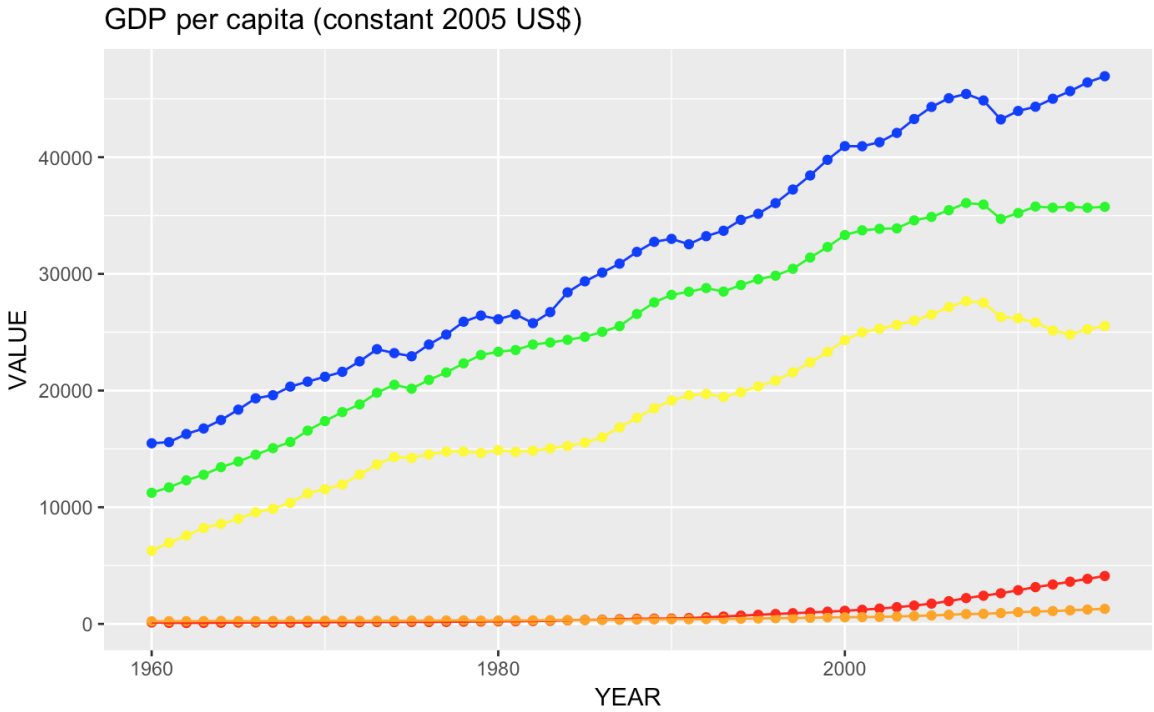
IV.III. Household consumption



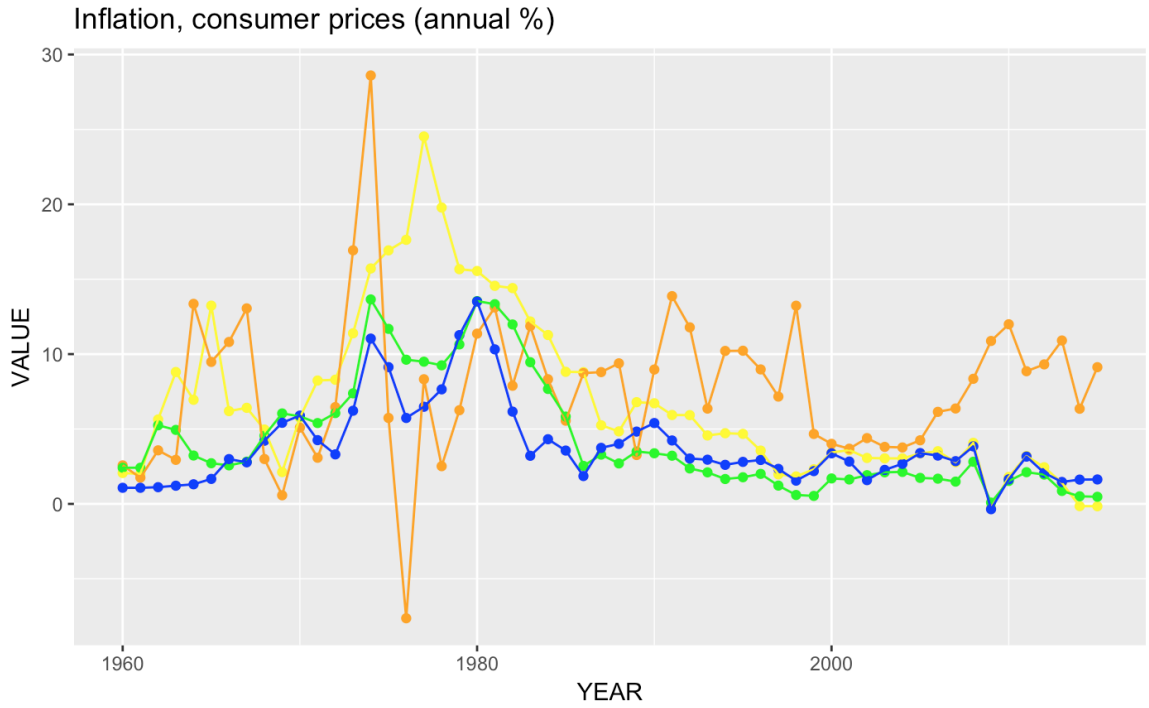
This graph show the evolution of household consumption. We can observe that Spain, USA, and France have a steady growth until 2008 and after only USA has continue to grow. On the other hand, China and India had a flatter GDP per capita but in the recent years the is a small but steady growth.



In the graph above, we can see how the GDP growth has evolved. All the countries seem to have a stationary GDP growth with no clear trend.



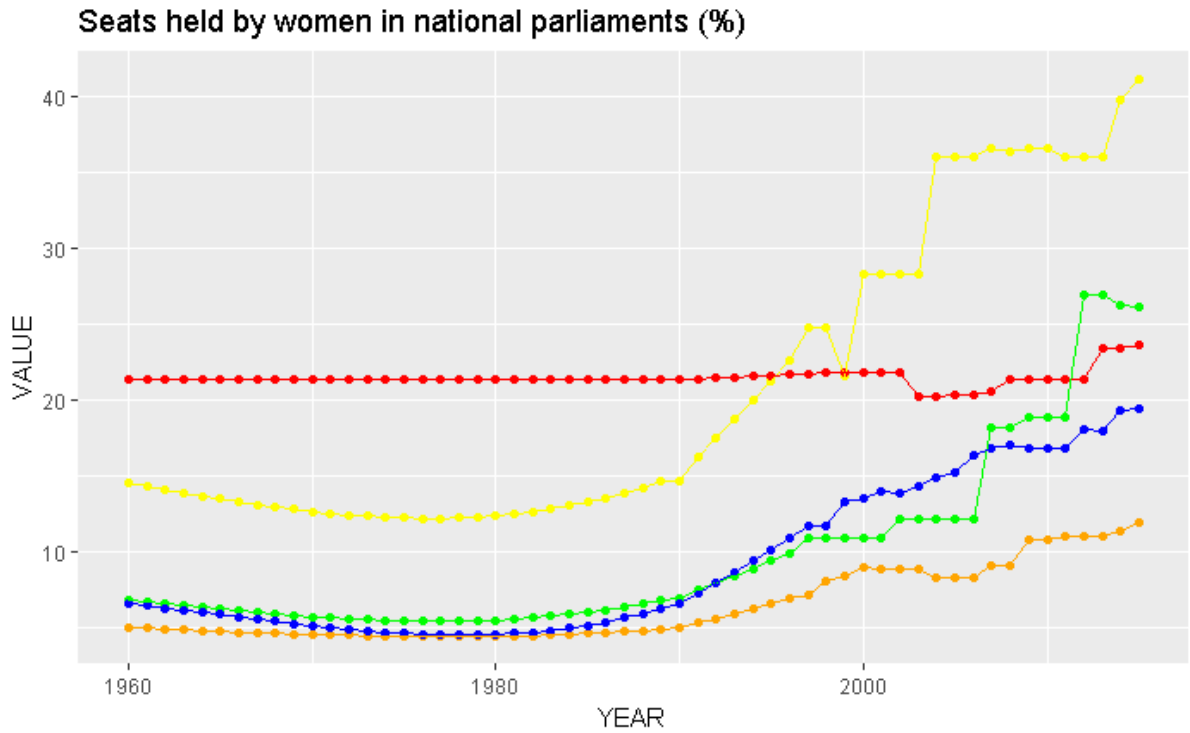
This graph depicts the GDP per capita. We can observe that Spain, USA, and France have a steady growth until 2008 and after only USA has continue to grow, whereas China and India had a flatter GDP per capita but in the recent years the is a small but steady growth. This behaviour is very similar to the household consumption and therefore it is a possibility a good predictor.



The Inflation graph shows a pattern that seems stationary for countries India until 1998. This is not the case of the other three countries that show a growing trend before 1980 and after a decreasing one.

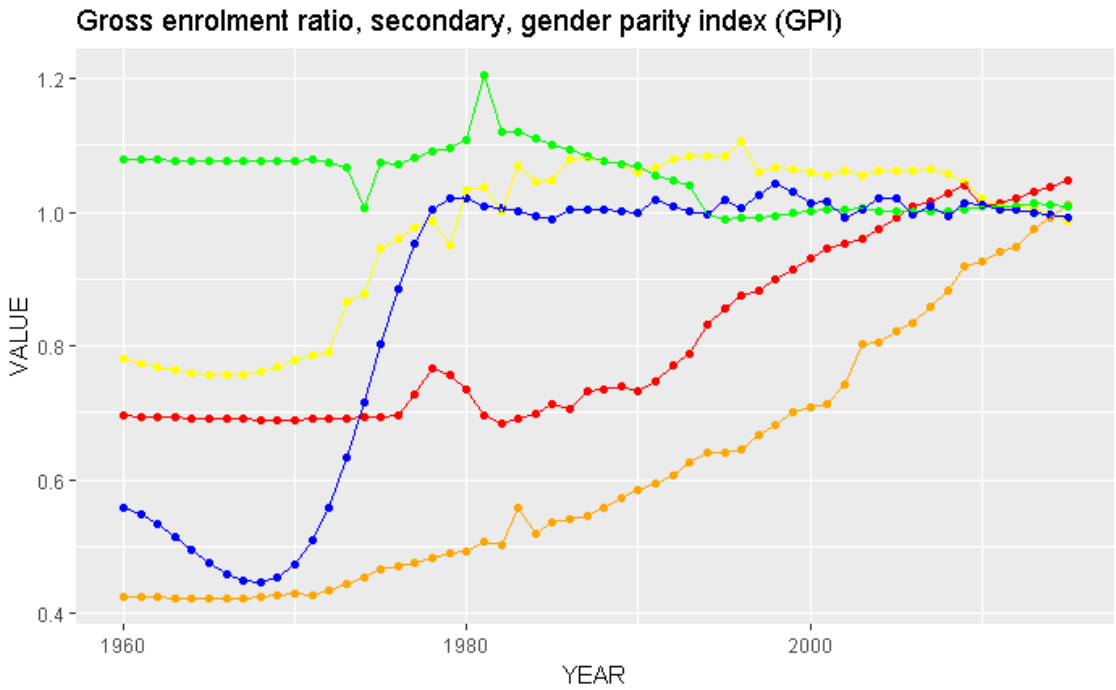
IV.IV. Gender equality

This first graph shows the tendency along the years of the number of seats held by women in national parliaments. Although all the countries have a growing tendency (except China, which is the red line). We can see that Spain (yellow line) has had a way bigger number since the 90s. This fact is explained because democracy came in Spain after the end of the dictator Francisco Franco in the 80s and it allowed women fight successfully for the equality gender.

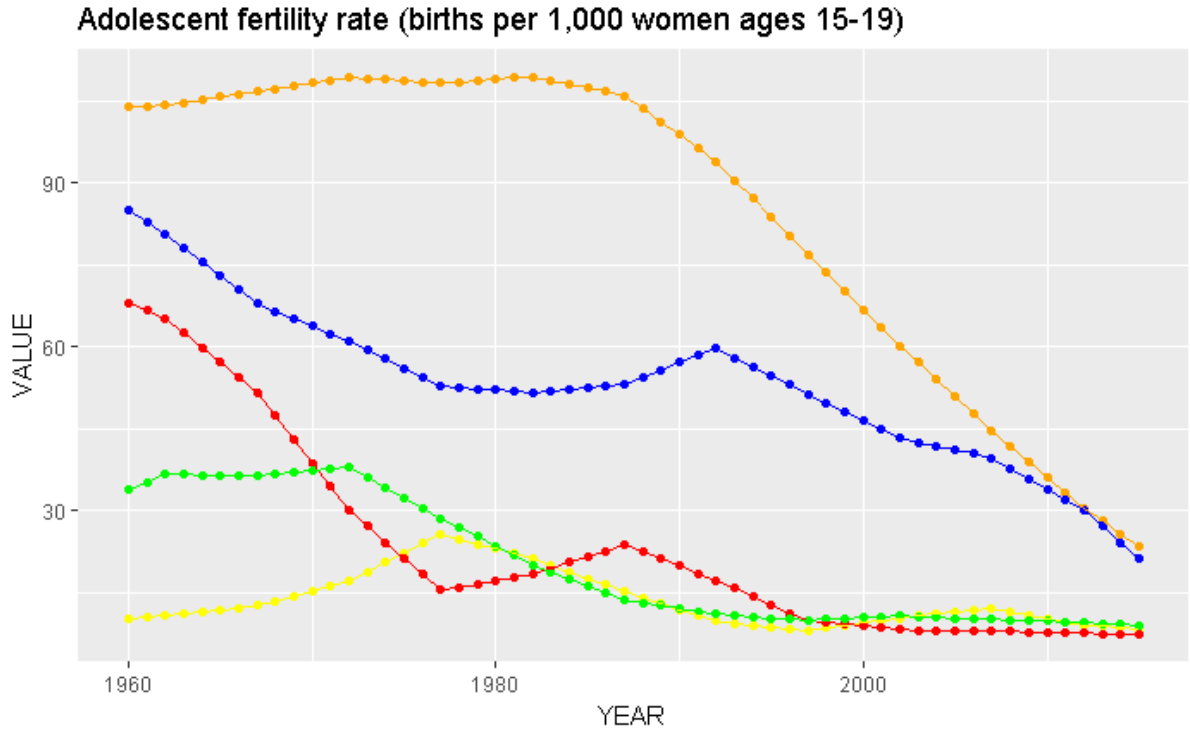


Regarding the second graph which shows the tendency along the years of the gross enrolment ratio (secondary, gender parity index (GPI)). This indicator is the ratio of girls to boys enrolled at secondary level in public and private schools. A value of less than 1 suggests girls are more disadvantaged than boys in learning opportunities and a GPI of greater than 1 suggests the other way around. Eliminating gender disparities in education would help increase the status and capabilities of women.

The most remarkable fact is that all the countries seem to be making an effort to provide the same education to men and women in the current years (value equals to 1 in the graph).



In the last graph, we can see the tendency along the years of the adolescent fertility rate (births per 1,000 women ages 15-19). This indicator is the number of births per 1,000 women ages 15-19. Adolescent fertility rates are based on data on registered live births from vital registration systems or, in the absence of such systems, from censuses or sample surveys. For countries, without vital registration systems, fertility rates are generally based on extrapolations from trends observed in censuses or surveys from earlier years.



V. MODELLING

Regarding modelling, we decided to try to fit some AR and MA univariable models first to use them as our baseline to beat when performing multivariate analysis.

Since we did not know anything about time series we started to work together with only one response for one of the questions we had.

To fit an AR or MA model we had extract the lag to use from the ACF and PACF plots.

We got some results we show next but it is not easy to get a correct lag form these two plots, it is an iterative process. After you choose a lag and fit the model you have to look at the residuals and see if there is a pattern that repeats every certain number what can indicate that there is a different lag you did not take into account.

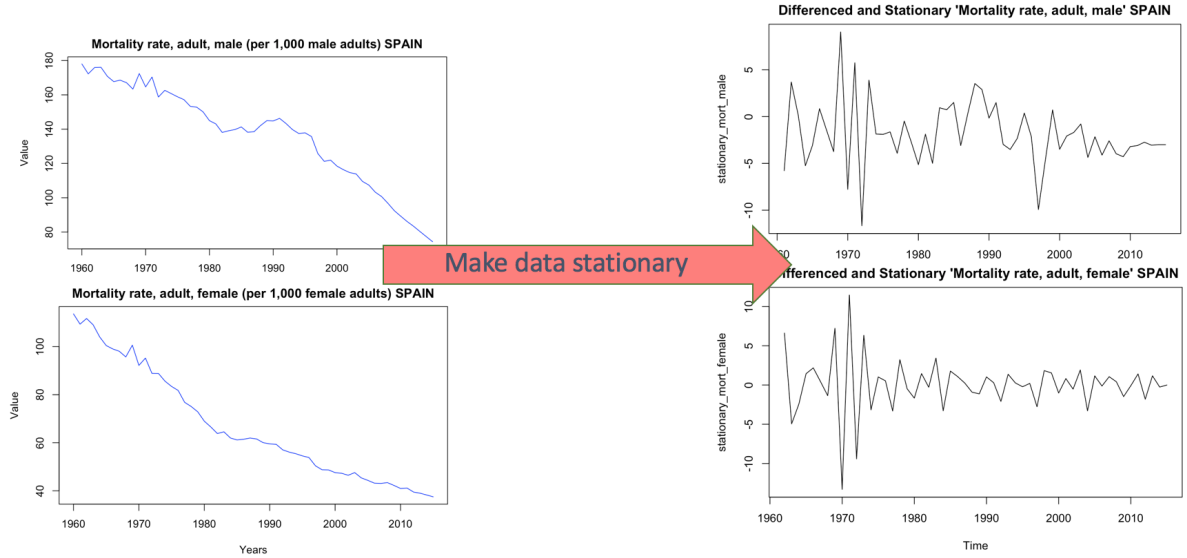
Since we did not have previous knowledge about time series as we said, we decided to try the auto.arima function as well. This function chooses the parameters for the AR, I and MA parts by itself and you don't have to select them from the ACF and PACF plots.

For this part we have used two different features that were the ones that we wanted to predict:

- Mortality rate, adult, female (per 1,000 female adults)
- Mortality rate, adult, male (per 1,000 male adults)

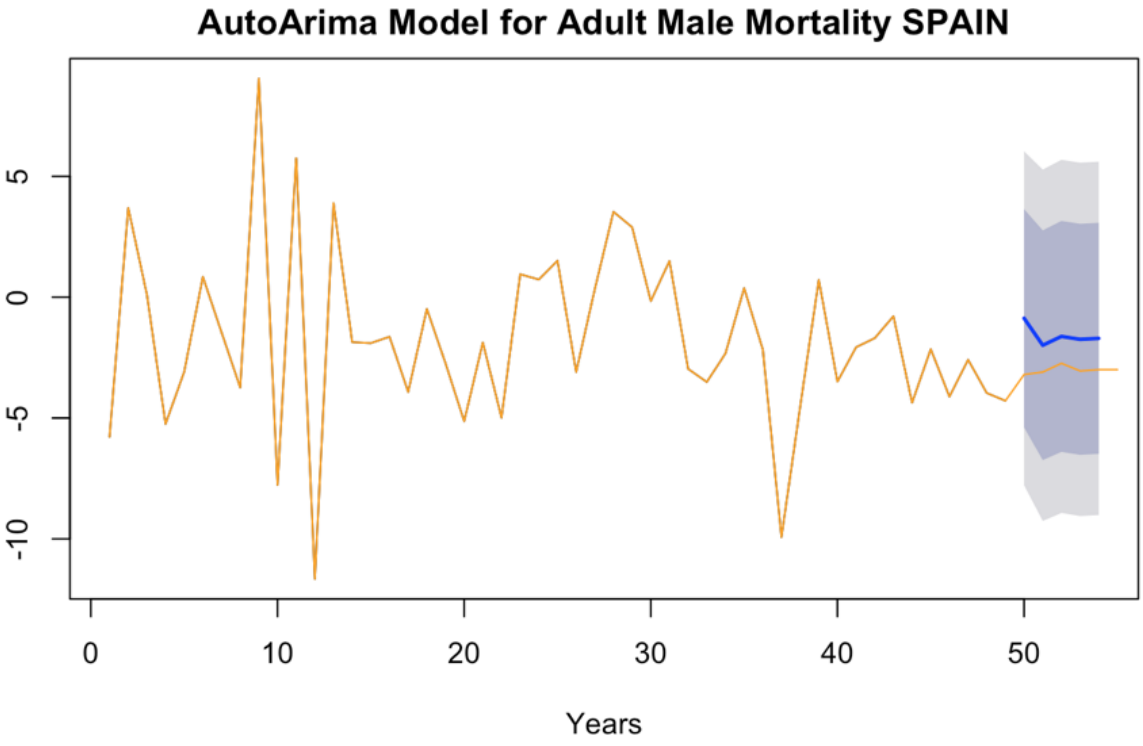
To fit an AR, MA or ARIMA model the data has to be stationary, so the first thing to do here is to check if the data is stationary or not. To do this we ran a Dicker-Fuller test through the adf.test function. The null hypothesis is that the data is not stationary and the alternative hypothesis is that it is stationary using a threshold of 0.05.

We are going to show this transformation in the Spain set of data:



We applied to each model AR, MA and ARIMA transformations that can be seen in the *stationary_mortality.Rmd,* in this report we are going to show just the ones that perform better results in each country.

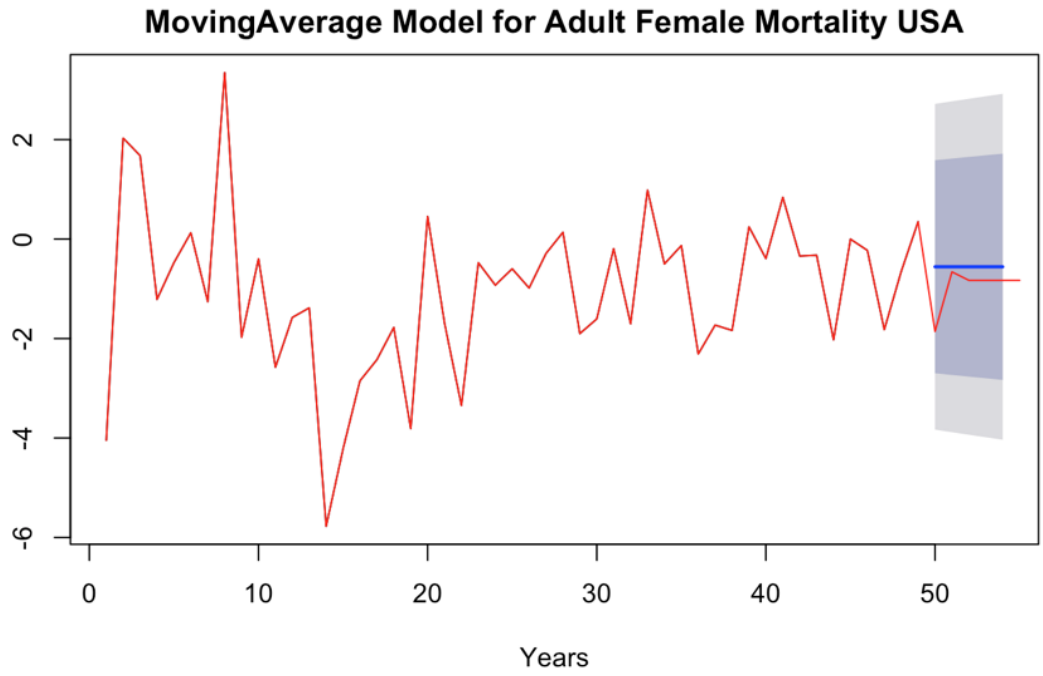
SPAIN:



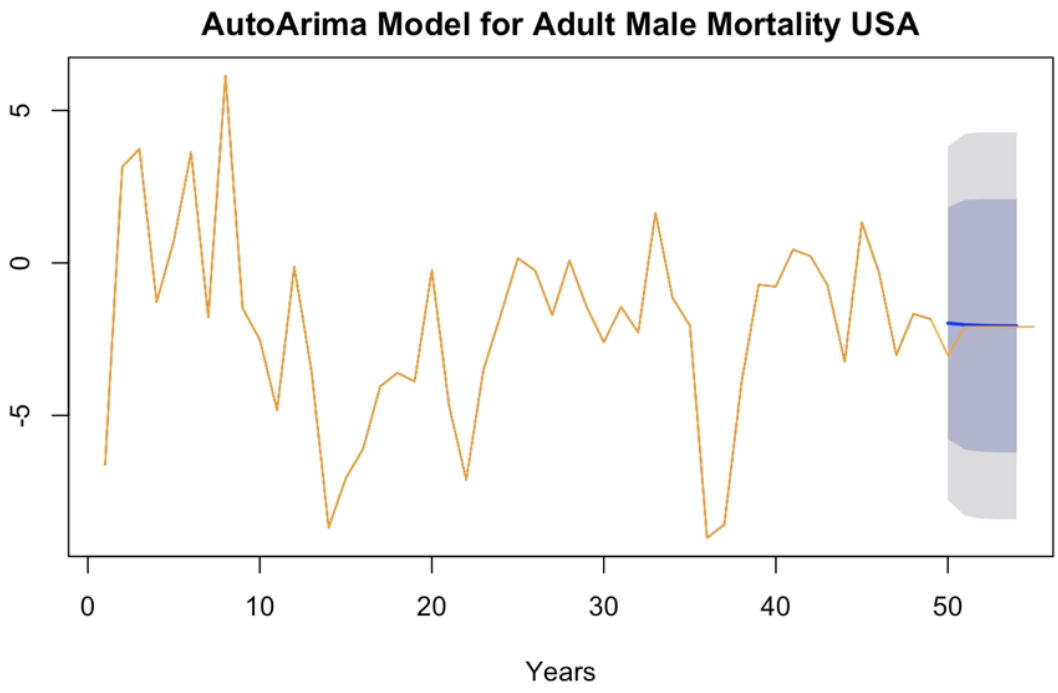
In Spain for the Adult Male Mortality the model that performs better was the Auto Arima model, despite we can see that it doesn't adjust really well to the real data. On the other

hand, for the Adult Female Mortality was the AutoRegressive model the one that performs better. In this model we can see that even if the predict data does not match perfect with the real data, the peaks of the predicted data coincide more or less with the ones that exists in the real data.

USA: In the case of the Adult Female Mortality in the USA neither of the methods perform really well, the one that do it better is the MA although it miss an important raise in the year 50.

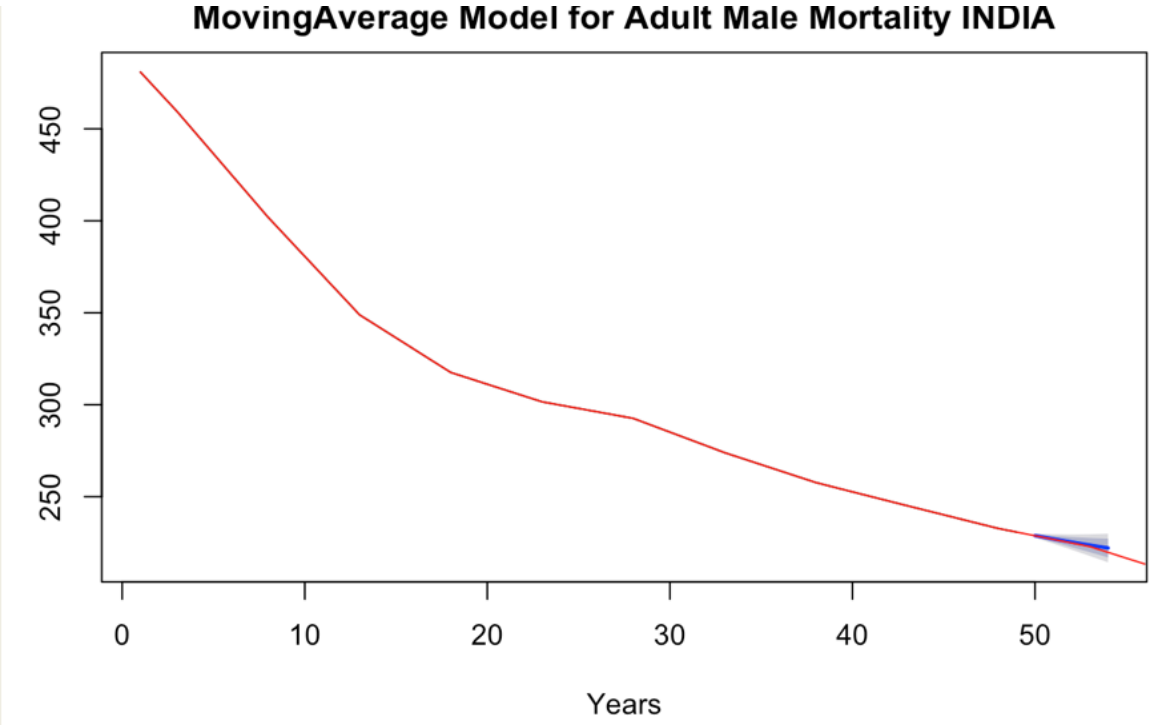


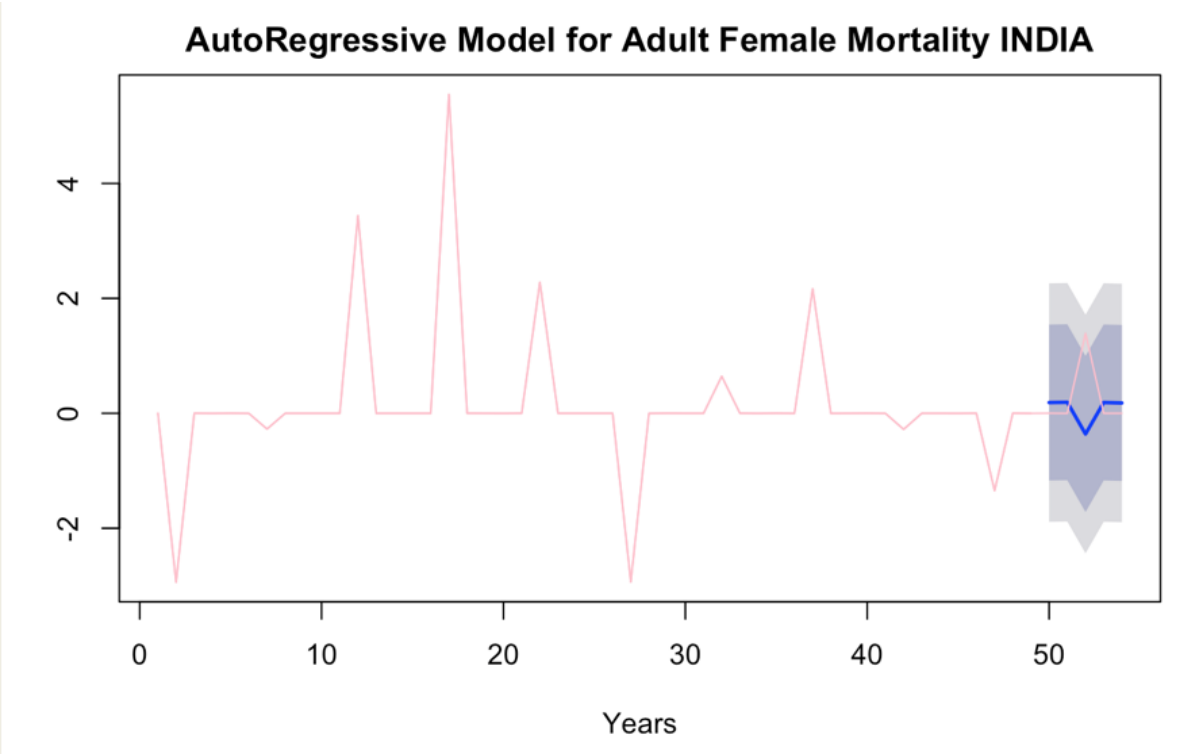
For the case of Adult Male Mortality is the AutoArima the one that provide a really good estimation for the data.



INDIA:

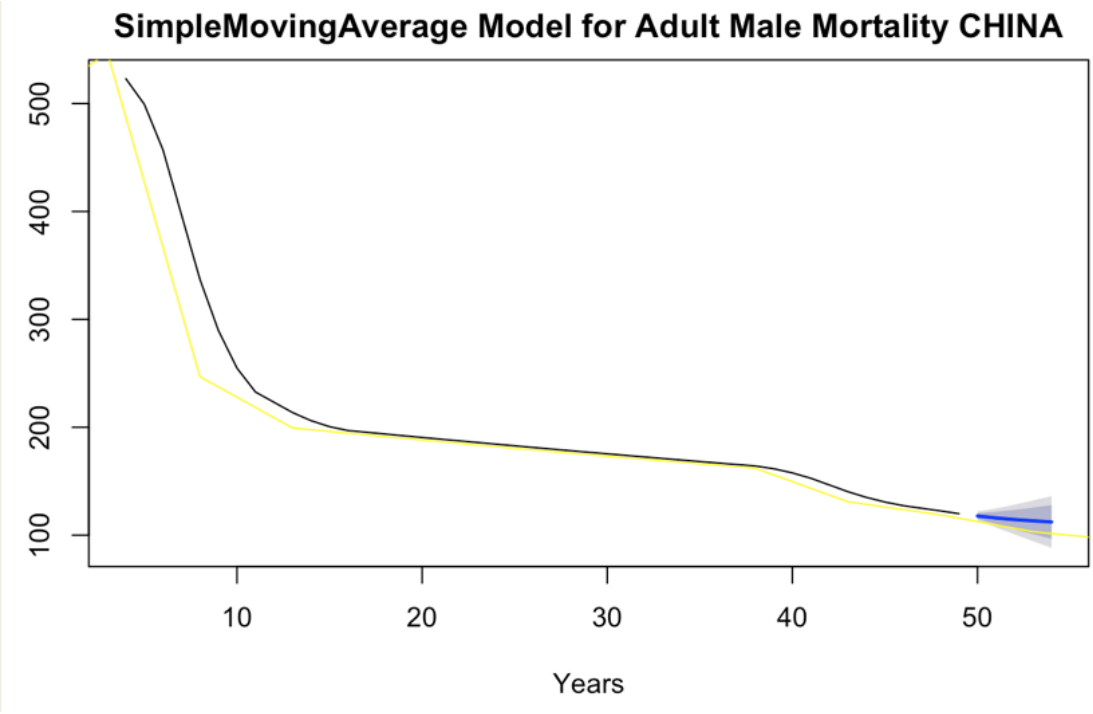
In India, the only one that was correctly estimated was the Adult Male Mortality, the Adult Female Mortality has nothing to do with the real data, probably because of the peaks that appeared once I made the data stationary and as we don't have many years of information it did not made an accurate prediction based on the past years.

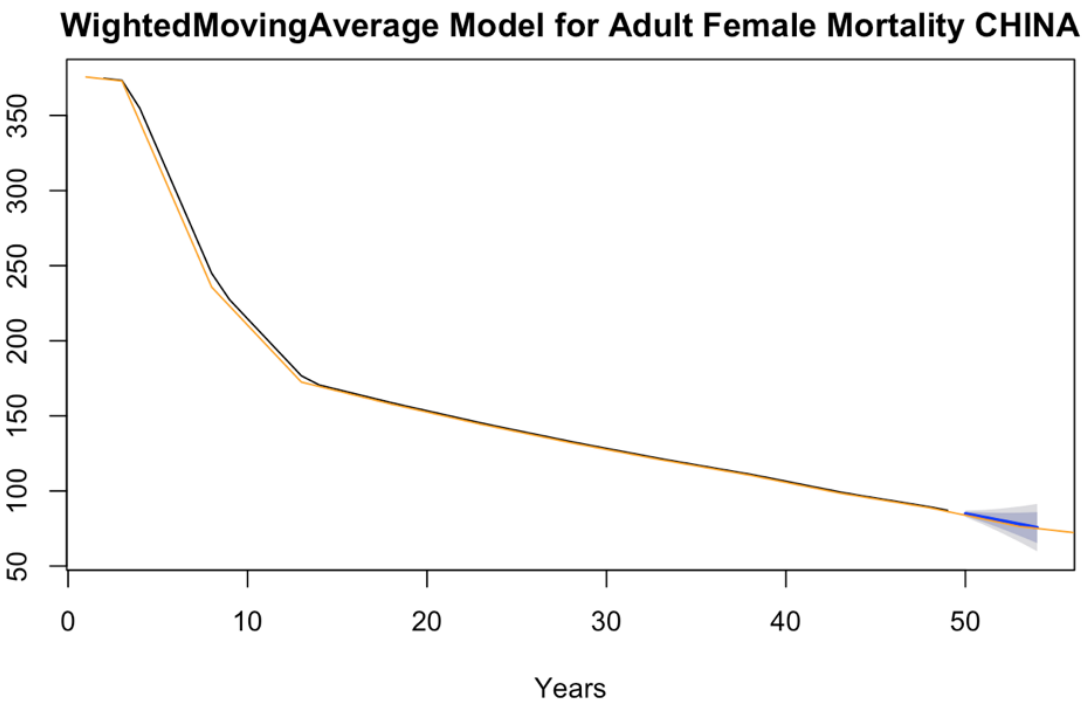




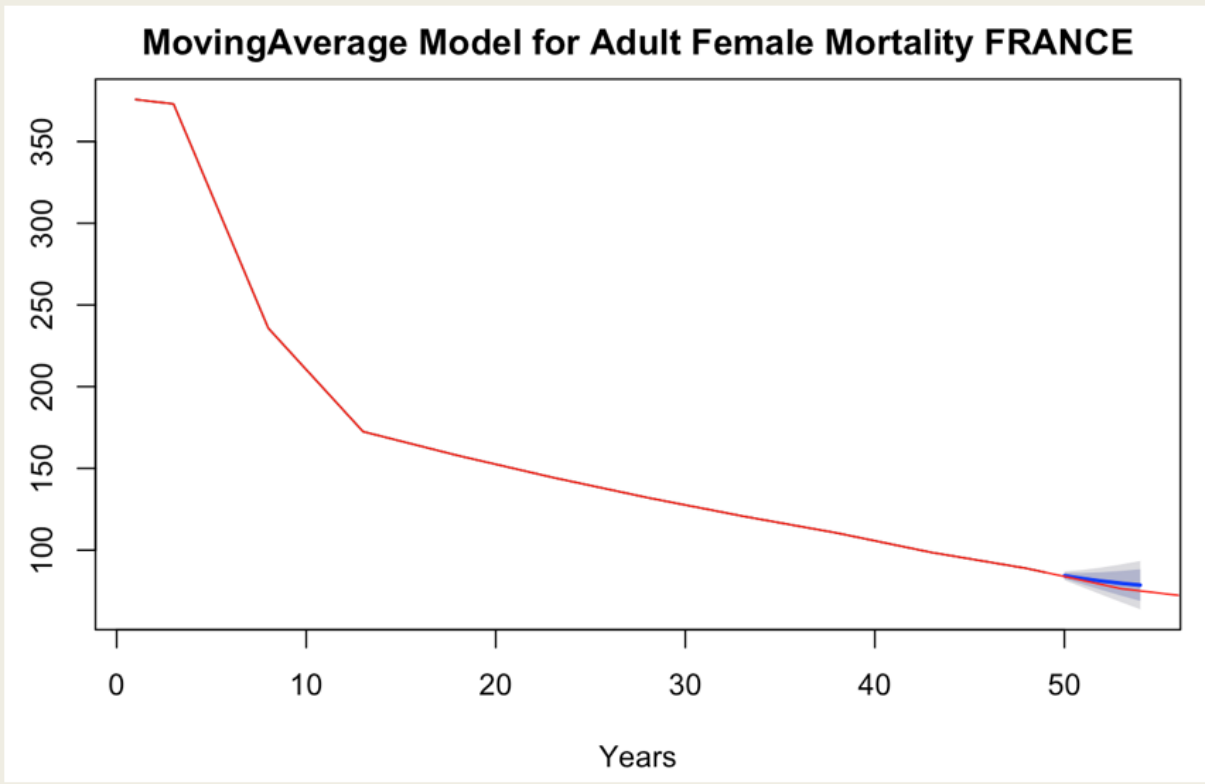
CHINA:

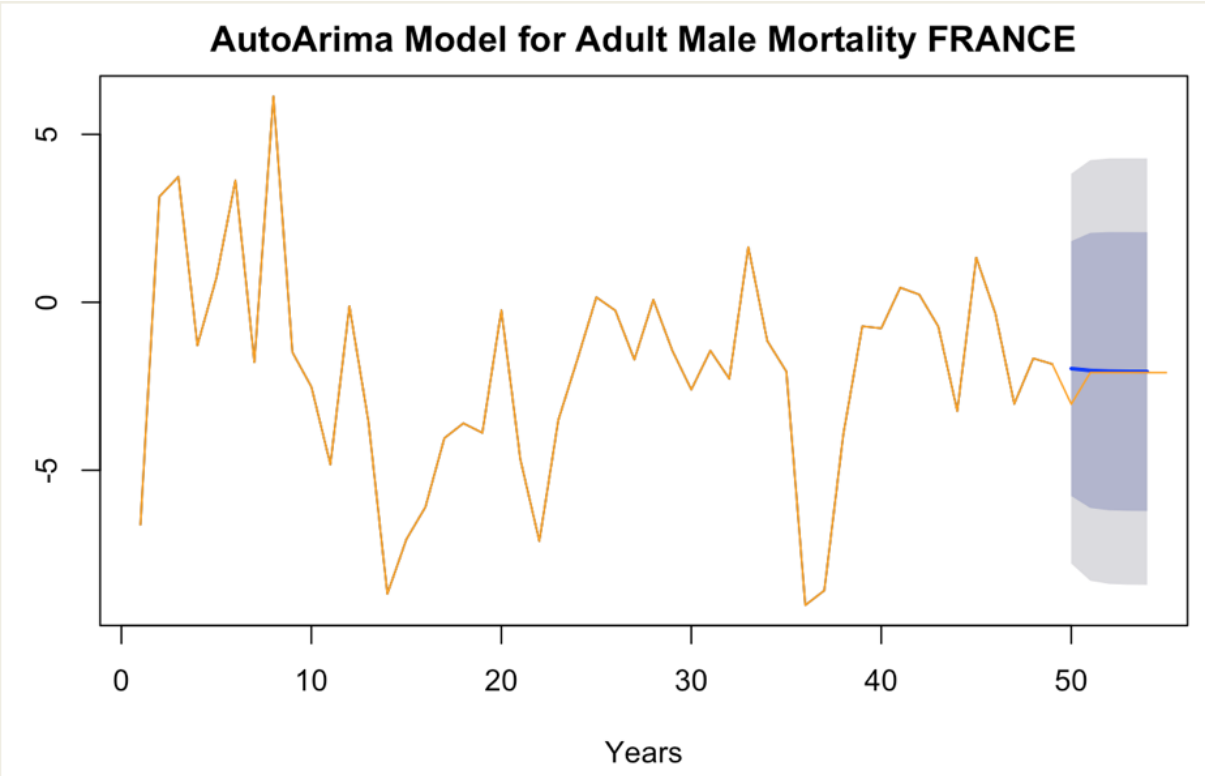
In China the Adult Male Mortality and the Adult Female Mortality have a very similar shape. So that's the reason why both plots look similar. In this case I applied two different versions of the MA algorithm, the Simple MA and the Weighted MA and both performed well in the data as we can see in the plot below.




WightedMovingAverage Model for Adult Female Mortality CHINA
50
100
150
200
250
300
350
0
10
20
30
40
50
Years

FRANCE:





The auto.arima models were created for the indicators selected to be related with life expectancy as well and they can be seen in the life_expectancy.rmd file, we don't show them here because the previous plots are already representative of the methodology.

The mse was computed for the models to have a metric to compare them.

Once we had some univariable models we tried to fit some multivariable models.

First we tried to fit a multivariate VAR model but then we realized that the auto.arima function that we used before had a parameter to include predictors so we ended up using this function instead.

We were able to fit a model and do predictions but since the model results don't provide a p-value or a clear measure to say if a predictor is relevant or not we could not decide on this topic that is what we were looking for.

Furthermore, we were already in the last week when we got to this point so we could not get deeper in this final topic of multivariate time series analysis that was the last but quite important step of the project.

VI. CONCLUSION

This project has been a great opportunity to introduce ourselves to time series analysis. Moreover, the data preparation has challenge us with a difficult original data set and we had to face a lot of missing values that we imputed with kalman as well. On the other hand, our main goal of this project was not prediction but inference. Despite we have been able to create various models to fit the data and predict it(AR, MA and ARIMA), those only used previous lags. Afterwards, we tried to create a model including the predictors that were chose in the previous stages of the project but we were not able to interpret the output in R and fit more models using different sets of indicators as predictors to compare between them. This would have been the best scenario, but anyways we are satisfied with the things we have learned during the project.

VII. DATA SOURCES AND BIBLIOGRAPHY

An Intro to Statistical Learning: http://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf
R for Data Science: http://r4ds.had.co.nz/
Data source: https://www.kaggle.com/worldbank/world-development-indicators
Data source: https://data.worldbank.org/products/wdi
Time Series Objects: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/ts.html
Handle Missing Values in Time Series For Beginners: https://www.kaggle.com/juejuewang/handle-missing-values-in-time-series-for-beginners
Welcome to a Little Book of R for Time Series: https://a-little-book-of-r-for-time-series.readthedocs.io/en/latest/
Analysis of multivariate timeseries using the MARSS package: https://cran.r-project.org/web/packages/MARSS/vignettes/UserGuide.pdf
imputeTS: Time Series Missing Value Imputation in R: https://cran.r-project.org/web/packages/imputeTS/vignettes/imputeTS-Time-Series-Missing-Value-Imputation-in-R.pdf
Basic Regression Analysis with Time Series Data: http://www.eco.uc3m.es/~jgonzalo/teaching/TecnicasEconometricas/WooldridgeCh10-12.pdf
Econometrics in R: https://cran.r-project.org/doc/contrib/Farnsworth-EconometricsInR.pdf
Forecasting: Principles and Practice: https://otexts.org/fpp2/
Time Series Analysis and Its Applications: http://www.stat.pitt.edu/stoffer/tsa4/
ARIMA Modelling of Time Series: http://stat.ethz.ch/R-manual/R-devel/library/stats/html/arima.html
Package forecast: https://cran.r-project.org/web/packages/forecast/forecast.pdf

VIII. SOURCE CODE

VIII.I. Source code - Life expectancy

Please, see “life_expectamcy.rmd” or “life_expectancy.pdf”.

VIII.II. Source code - CO2 related to mortality

Please, see the file “stationary_mortality.Rmd”.
Please, see the file “dataprep&ana_co2.Rmd”.

VIII.III. Source code - Household consumption

Please, see the file “consumption.Rmd”.

VIII.IV. Source code - Equality gender

Please, see the file “GenderEquality.html”.