Homework II: SVD analysis & Life Tables

Inès Dardouri, Mohamed Boudokhane and Lycia Fezzoua

2020-05-26

---

1 Introduction and objectives

This research work consists in making a demographic study on the mortality rates of several population of the world,describing different countries of Europe and the United States, based on different factors, such as gender and age. As we aim to make a comparison between Europe and USA, we chose to concentrate our study on one particular European country all along the work, to be consistent. Our choice fell on the Netherlands because it is a country that has been affected by the war, but which is relatively small (17.28 million inhabitants), rich and industrialized. We want to explore its future and compare it to the USA's. The study mainly focuses on the "mortality quotient" at a certain age, and according to different years. The mortality quotient is interpreted by the probability for people surviving at this age, to die before the following age.

The objective of this work is to: *Make demographic visualizations for mortality data over time, according to age and gender. This allows us to understand and explore the way in which mortality indicators evolve by age group and to deduce the factors that impact them.* Use the PCA (Principal Component Analysis) and CCA(Canonical-Correspondence-Analysis) methods for the exploration of multivariate datasets and the study of link structures on all variables.

*Use of a mathematical model "Lee Carter" which gives us the possibility to make predictions and adjustments of the mortality rates over time according to age.

2 Life tables data (ETL)

We investigate life tables describing countries from Western Europe (France, Great Britain –actually England and Wales–, Italy, the Netherlands, Spain, and Sweden) and the United States.

We load the one-year life tables for female, male and whole population for the different countries.

The meaning of the different columns:

`mx` : Central death rate between ages x and x+n where n=1, 4, 5, or ∞ (open age interval)

`qx` : Probability of death between ages x and x+n

`ax` : Average length of survival between ages x and x+n for persons dying in the interval

`lx` : Number of survivors at exact age x, assuming l(0) = 100,000

`dx` : Number of deaths between ages x and x+n

`ex` : Life expectancy at exact age x (in years)

But some of the columns need retyping:

- `Year` : should be integer
- `Age` : needs some cleaning, after cleaning it should be typed as integer
- `Lx` : should be integer
- `Tx` : should be integer
- Other columns should be considered as floating point numbers (`numeric`)

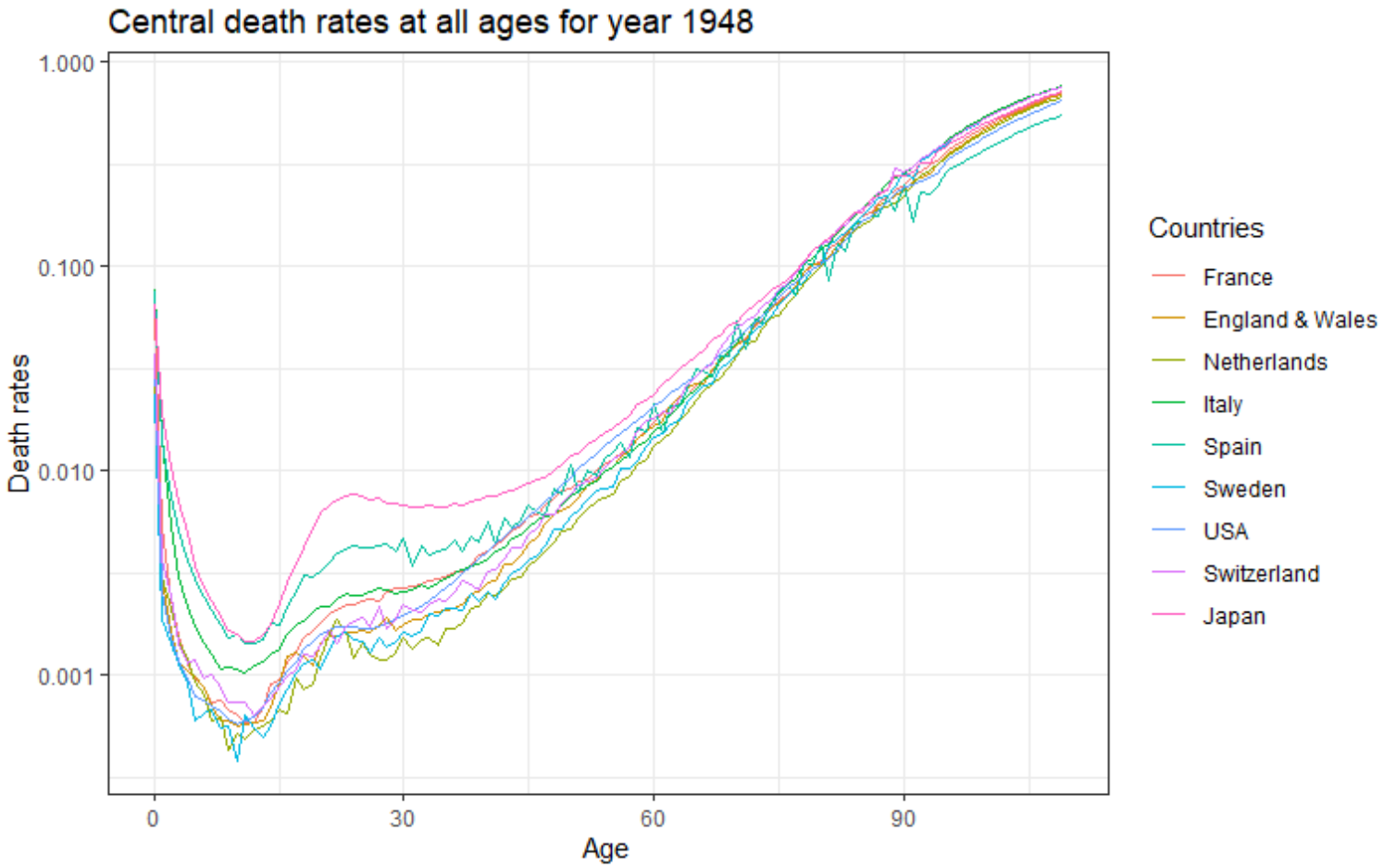
- After ETL processing, we obtain a universal table encompassing all data available in the files located in `LIFE_TABLES` directory. Henceforth, the universal table is named `life_table` , its schema is the following.

Column Name	Column Type
Year	integer
Age	integer
mx	double
qx	double
ax	double
lx	integer
dx	integer
Lx	integer
Tx	integer
ex	double
Country	factor
Gender	factor

Coercion introduces a substantial number of NA warnings. Preliminary inspection of the data suggests that coercion problems originate from column `Age` : `110+` cannot be coerced to an integer value. We discard corresponding rows using `tidyr::drop_na(Age)` .

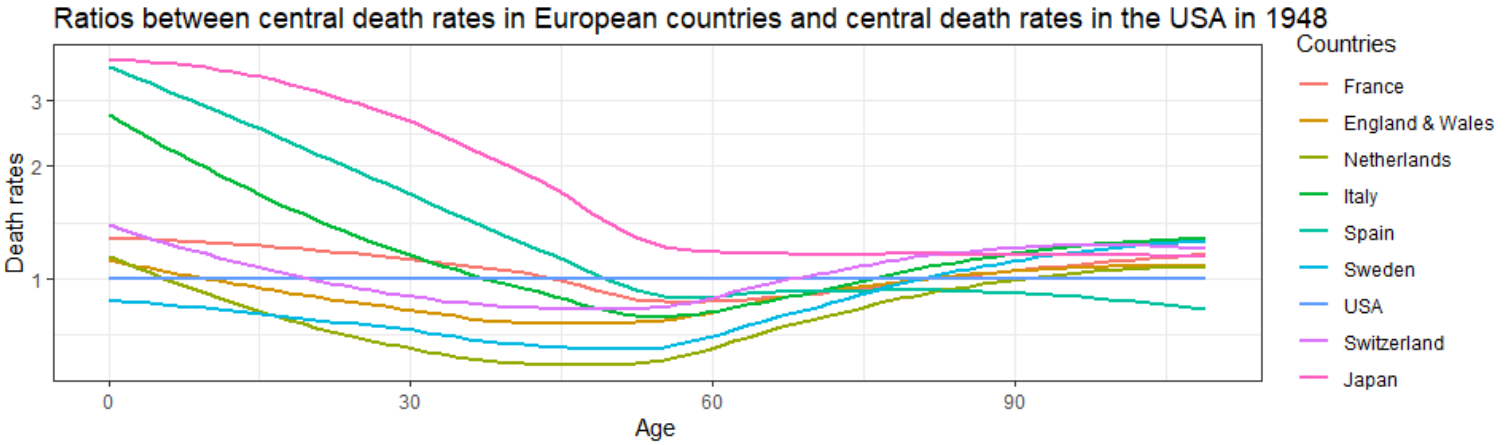
3 Western countries in 1948

- **Visualization of the central death rates of all Countries at all ages for year 1948 :**



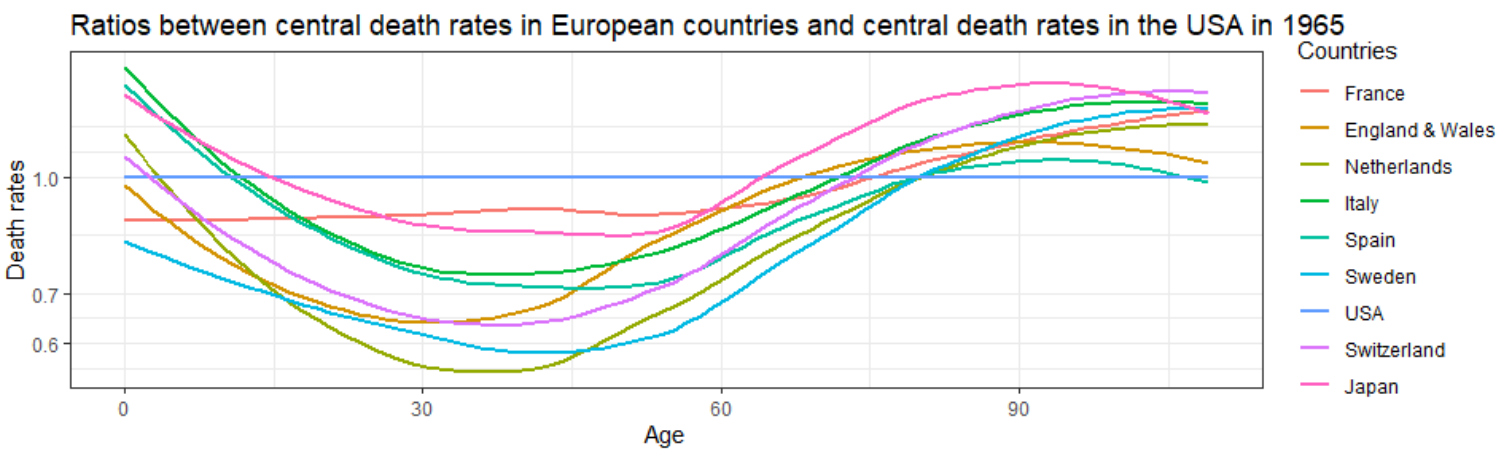
We notice that the death rates for newborns are much higher for Italy and Spain than for the rest of the european countries and for the USA. This difference is still noticable for infant mortality. But for the adults, the death rates are pretty much the same for all countries. The difference for young people's mortality could be explained by the different economic and health conditions at that time between the different countries.

- **Visualization of ratios between central death rates in European countries and central death rates in the USA in 1948 :**



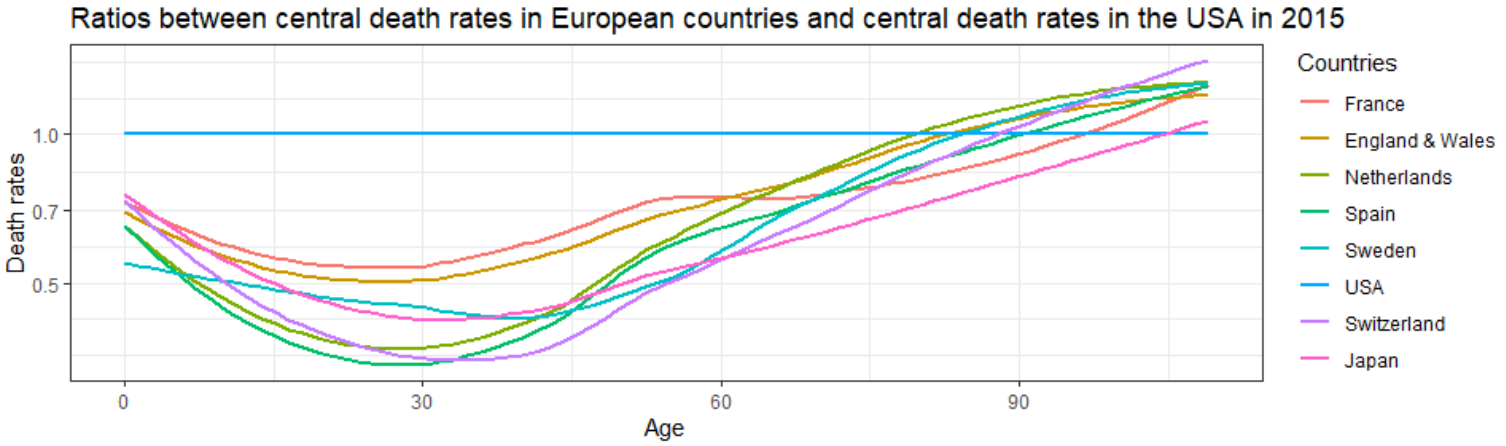
We can see that the ratio between central death rate in Netherlands and central death rate in the USA is less than 1 for almost all the ages except the oldest ones, which means that the central death rate in Netherlands is lower than central death rate in the USA in 1948. But we can also see that this ratio is greater than 1 for almost all the other European countries (except Sweden), which means that the central death rate in the majority of the European countries is higher than the central death rate in the USA in 1948, especially for France and Spain. This difference could be explained by the health conditions and the financial situation of the two continents at that time.

- **Visualization of ratios between central death rates in European countries and central death rates in the USA in 1965 :**



When we plot the same ratio but almost 20 years later, we can see the evolution of European countries' central death rates. Indeed, the ratio becomes lower than 1 for the middle ages, which means central death rates of European countries are approching USA's central death rates, even if it's still not the case at the extremities' ages. So, in 1965 Europe started regaining the health gap between her and the USA, comparing to what it was 20 years earlier

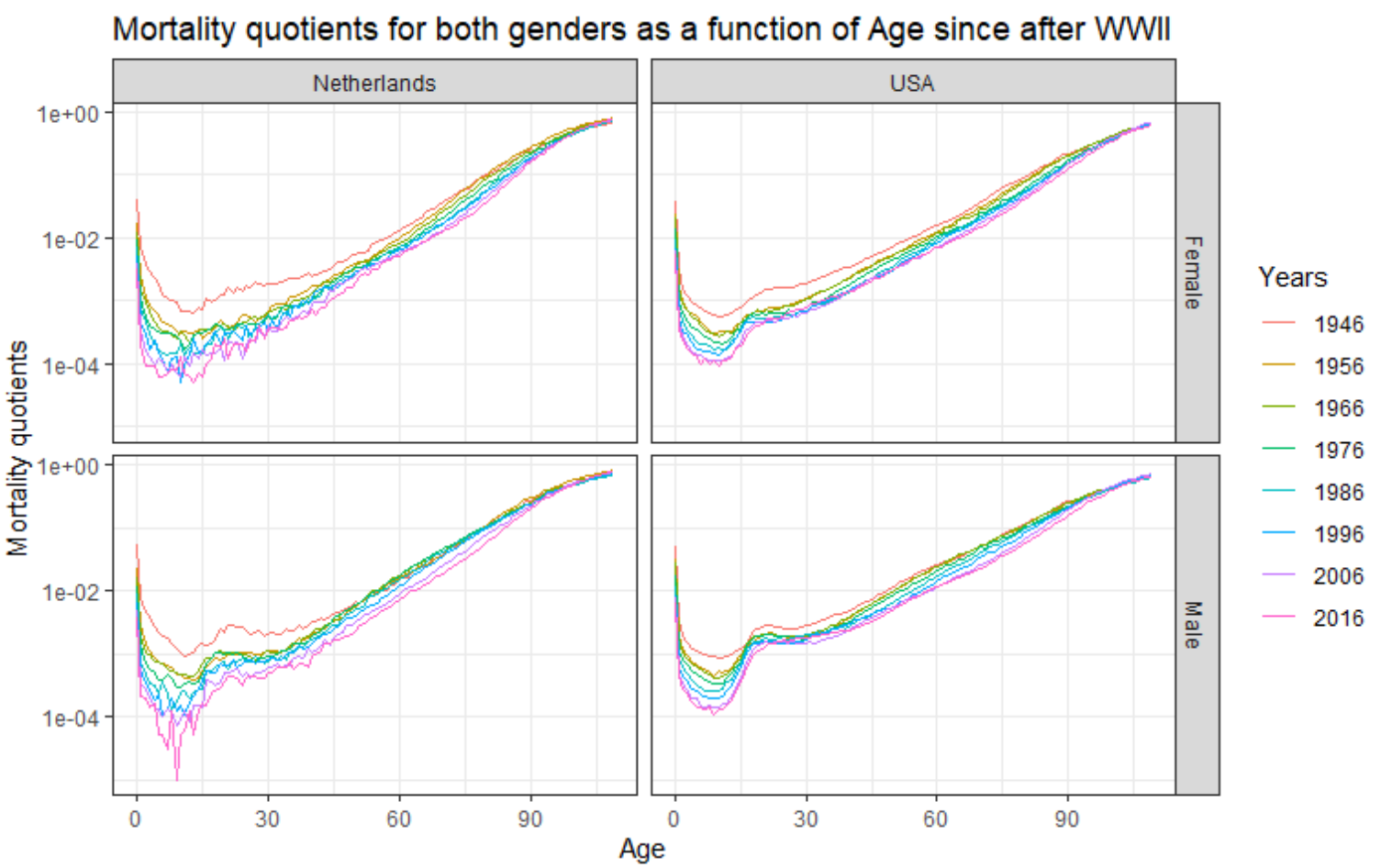
- **Visualization of ratios between central death rates in European countries and central death rates in the USA in 2015 :**



And if we study again the same ratio but for the year 2015, we clearly see an inversion of the 1948's curve for most of the European countries. Europe's death rates are now much lower than USA death rates for ages from 0 to almost 75. In fact, the economical situation of the Europe was remarkably increasing since the end of the WWII and therefore its health situation. We also have to underline the fact that each country has its own health politics which can be more or less in favor of the population. But we also notice that the USA have always had lower death rates for the oldest ages than European countries, since the end of WWII. This may be explained by the fact that the USA care more about their elderly than about the rest of the population, or simply because more Europeans reach old ages then die quickly.

4 Death rates evolution since WW II

- **Visualization of mortality quotients (column `mx`) for both genders as a function of `Age` for years `1946, 1956, ...` up to `2016`.**



- Concerning the plot below:

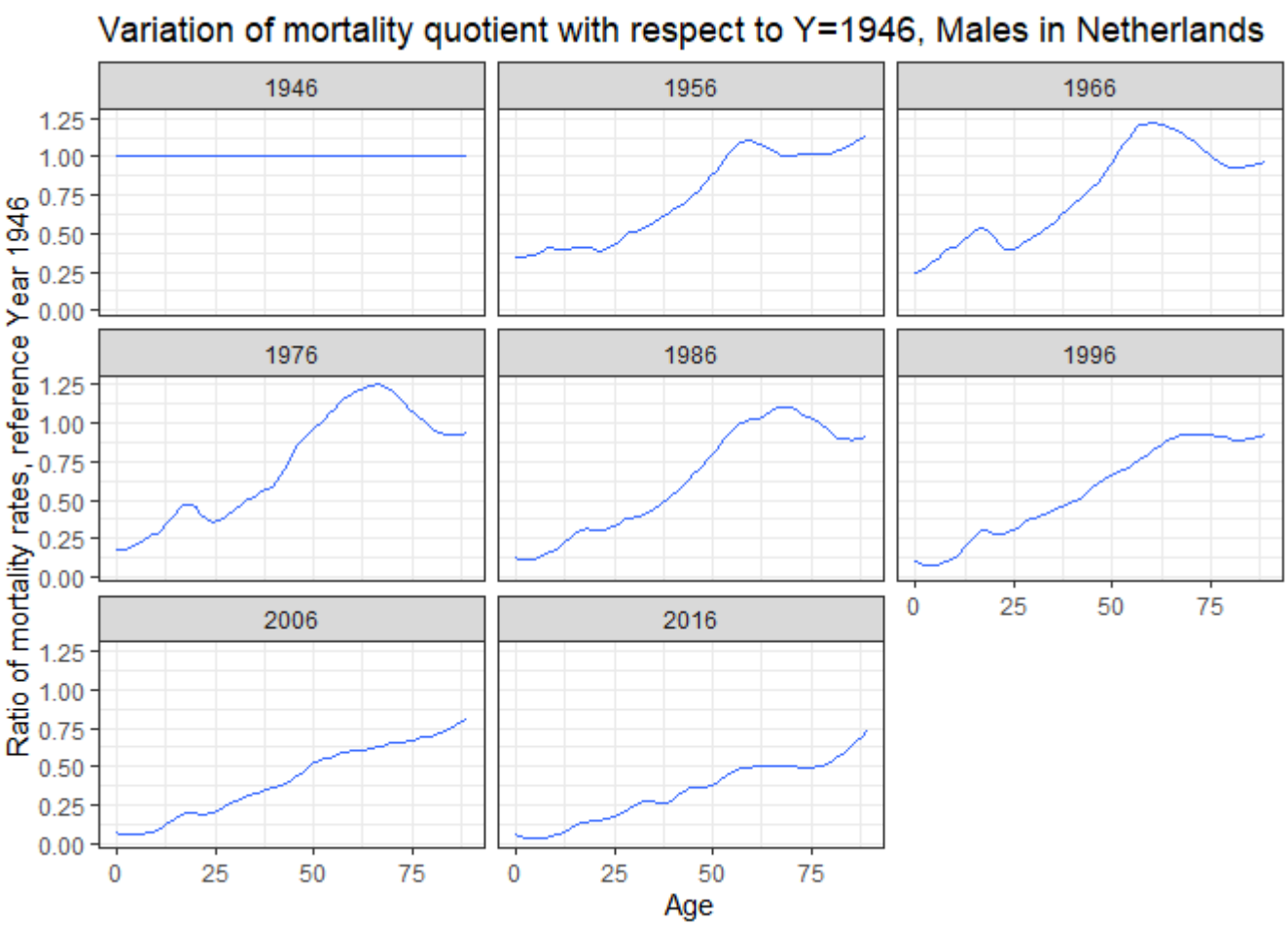
We notice that the mortality quotients of young people in 1946 is smaller in the USA than in all the European countries. This is certainly due to the fact that the USA didn't suffer a lot from human loss during the WWII, unlike the European countries.

We modify our dataframe so it has the following schema:

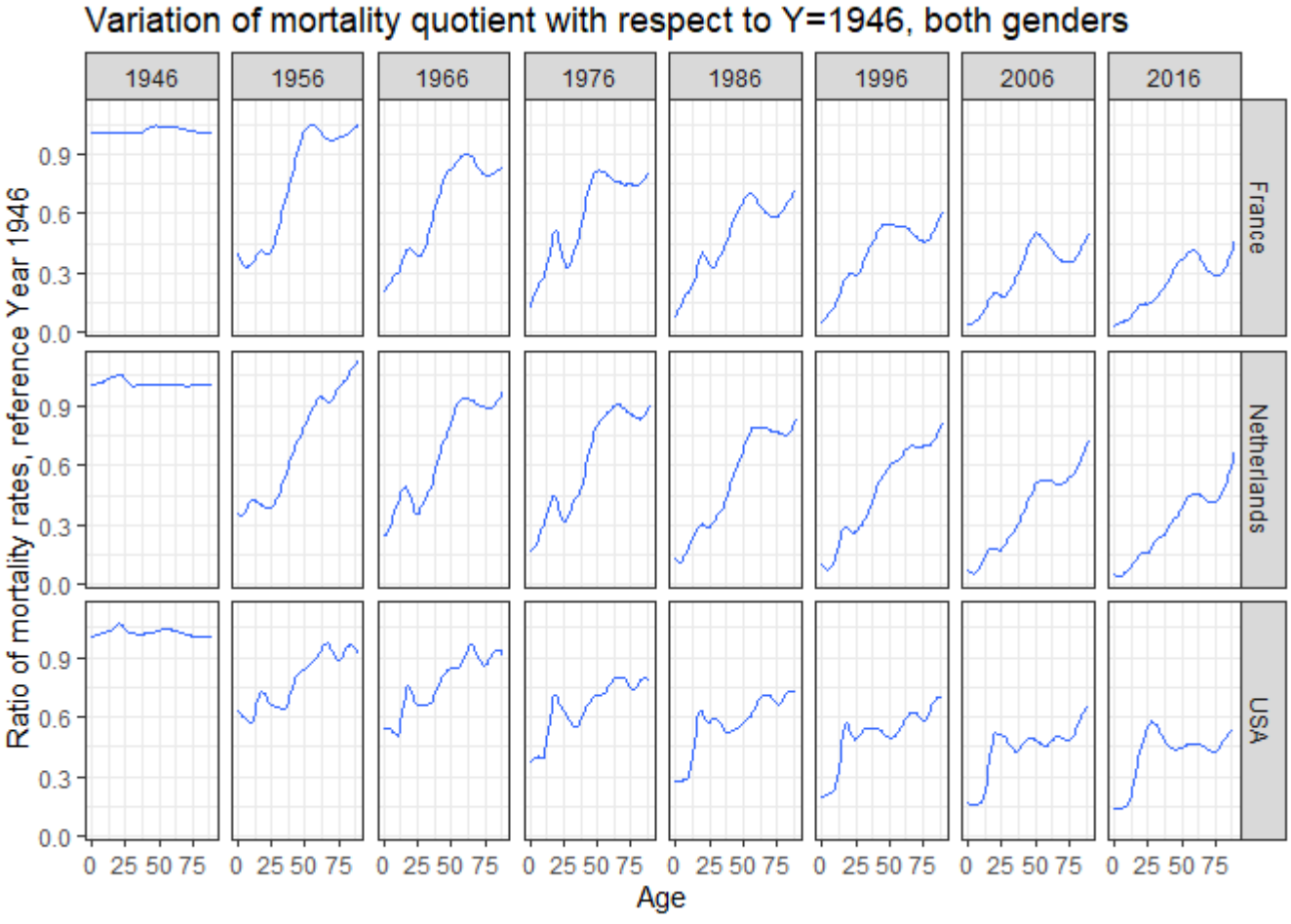
Column Name	Column Type
Year	integer
Age	integer
mx	double
mx.ref_year	double
Country	factor
Gender	factor

where `(Country, Year, Age, Gender)` serves as a *primary key*, `mx` denotes the central death rate at `Age` for `Year` and `Gender` in `Country` whereas `mx_ref_year` denotes central death rate at `Age` for argument `reference_year` in `Country` for `Gender`.

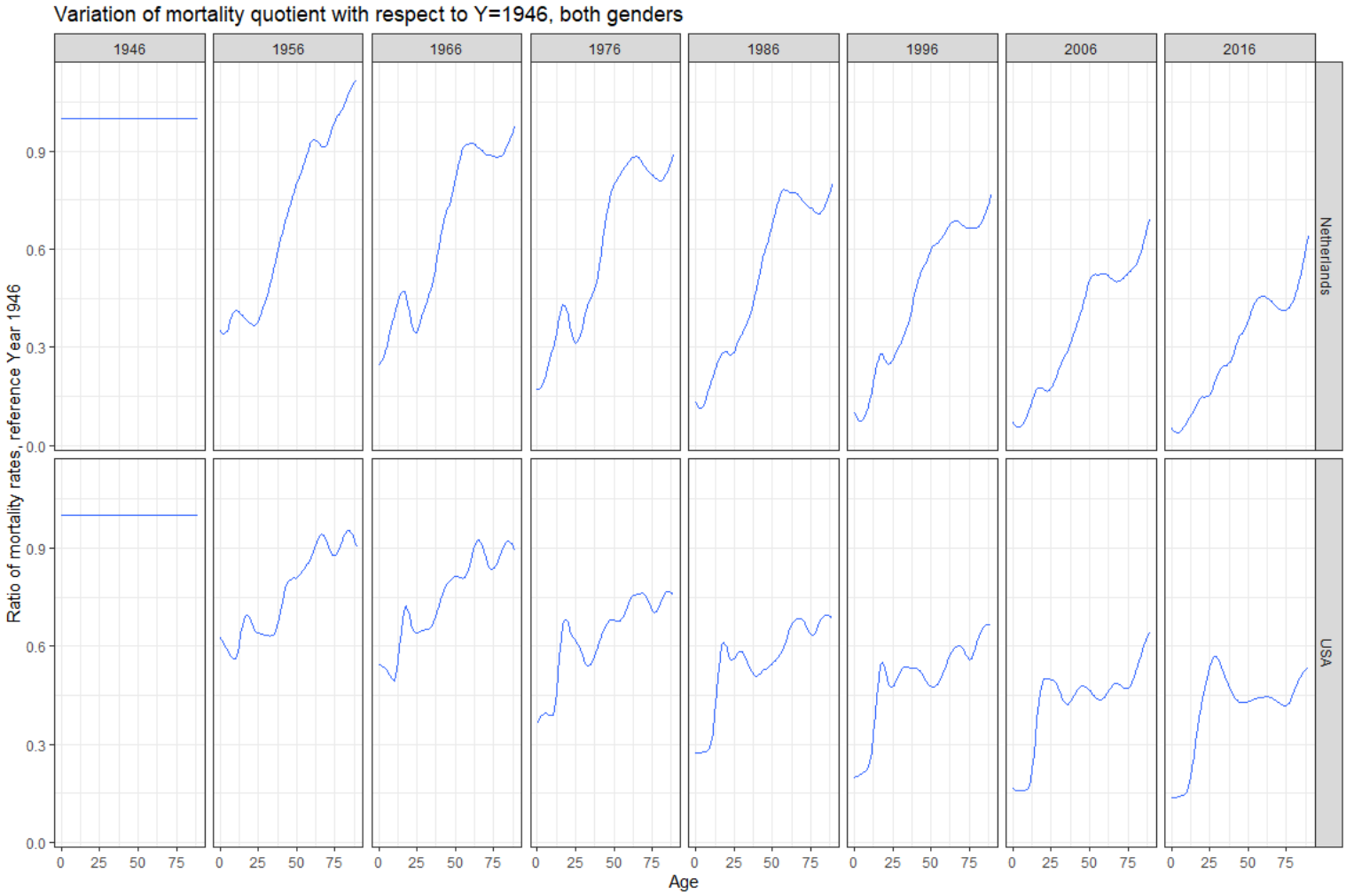
- **Visualization of the ratio $m_{x,t}/m_{x,1946}$ for ages $x \in 1, \ldots, 90$ and year t for $t \in 1946, \ldots, 2016$ where $m_{x,t}$ is the central death rate at age x during year t :**



- We handle both genders and countries `Spain`, `Italy`, `France`, `England & Wales`, `USA`, `Sweden`, `Netherlands`.



But as we did since the beginning, we concentrate on the comparison between te USA and the Netherlands.



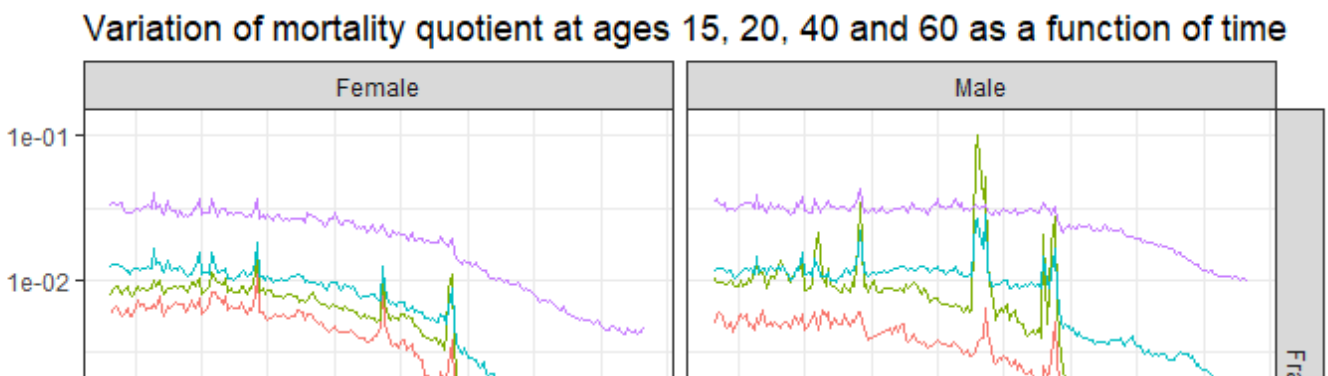
In the USA, the ratio of mortality rates between all the years after 1946 and the year 1946 has always been under 1 for all ages, which meens that since 1946 people die less in the USA than in 1946. Whereas in the Netherlands, this ratio has been higher than in the USA for all years and especially for the older ages. We also notice a difference for the new borns between the two countries : the ratio is twice higher for the USA in 1956 than in the Netherlands, which means that the mortality rate in 1946 was much higher in the Netherlands compared to the other years, whereas the difference is smaller for the USA between 1946 and the other years. The ratio becomes higher in the USA than in the Netherlands for the age 25 since 2006.

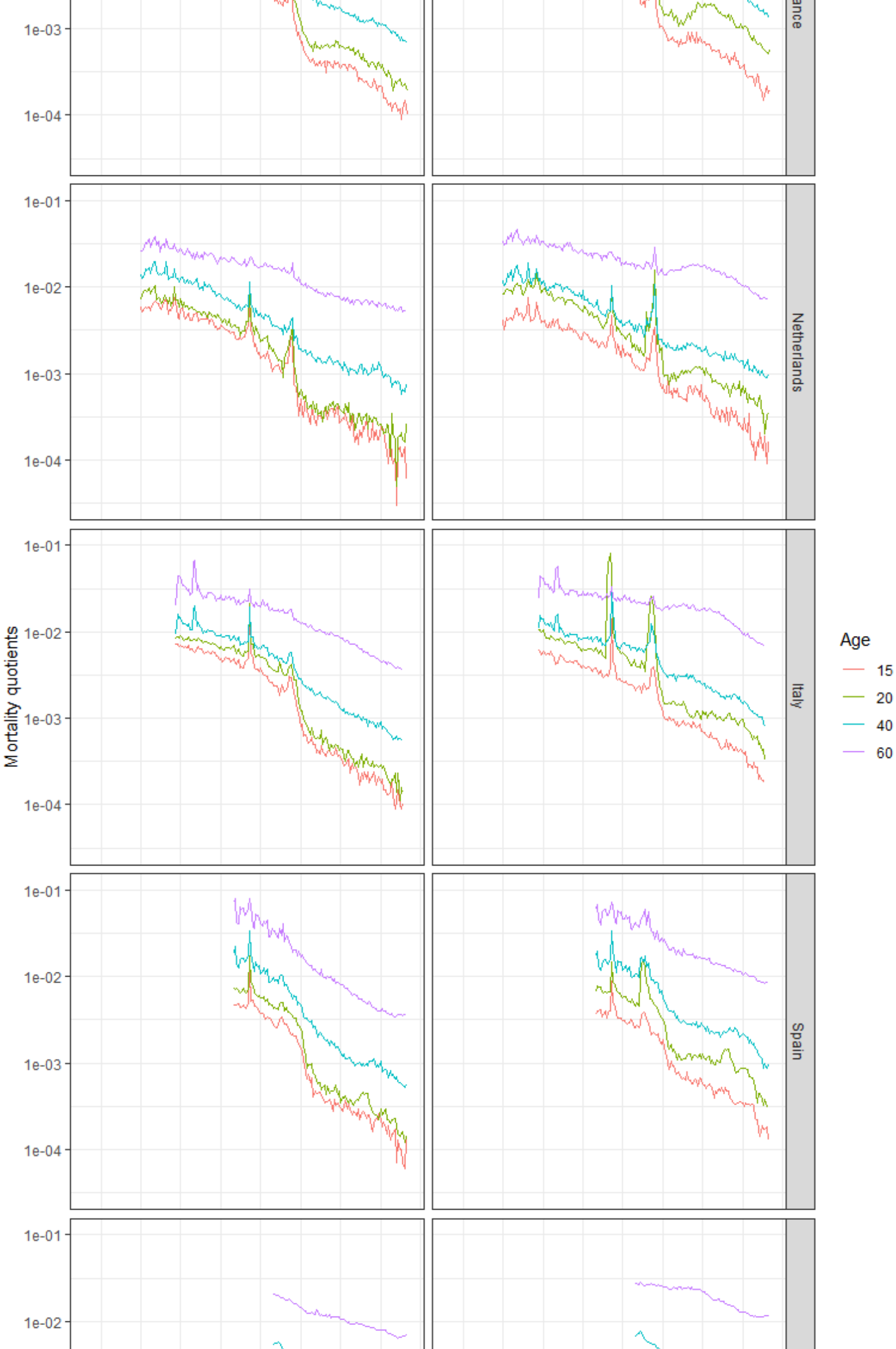
5 Trends

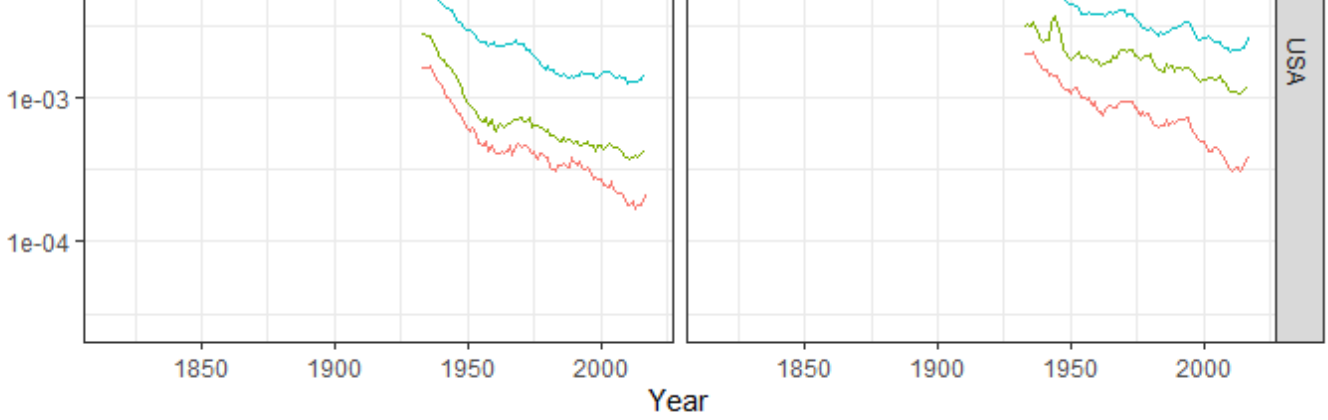
- **Visualization of mortality quotients at ages $0, 1, 5$ as a function of time, facetted by Gender and Country :**

New borns die more than children of ages 1 and 5 for both genders and for all countries, and children of age 5 die the less. We don't notice any difference between the mortality quotients of the two genders, for all the countries. We can see some noticable peaks for the Netherlands corresponding to the two world wars, that don't appear on the US plot, for the 3 different ages. Also, for both the USA and our European country, the mortality quotients were obviously higher 100 years ago than what they are nowadays.

- **Visualization of mortality quotients at ages $15, 20, 40, 60$ as a function of time, facetted by `Gender` and `Country` :**




Mortality quotients
1e-01
1e-02
1e-03
1e-04
Netherlands
Italy
Spain
Age
15
20
40
60



We observe that the mortality quotients for male are higher than for women for ages 15 to 60 for both countries. We can also see that the Netherlands peaks are still observable and they are even sharper, for the same years (WWI and WWII). Also, for both the USA and our European country, the mortality quotients were obviously higher 100 years ago than what they are nowadays.

6 Rearrangement

- From our dataframe `life_table`, we then compute another dataframe called `life_table_pivot` with primary key `Country`, `Gender` and `Year`, with a column for each `Age` from `0` up to `110`. For each age column, the entry should be the central death rate at the age defined by column, for `Country`, `Gender` and `Year` identifying the row.

The resulting schema looks like:

Column Name	Type
Country	factor
Gender	factor
Year	integer
`0`	double
`1`	double
`2`	double
`3`	double
⋮	⋮

- Using this new dataframe, we compute life expectancy at birth for each Country, Gender and Year :

7 Life expectancy

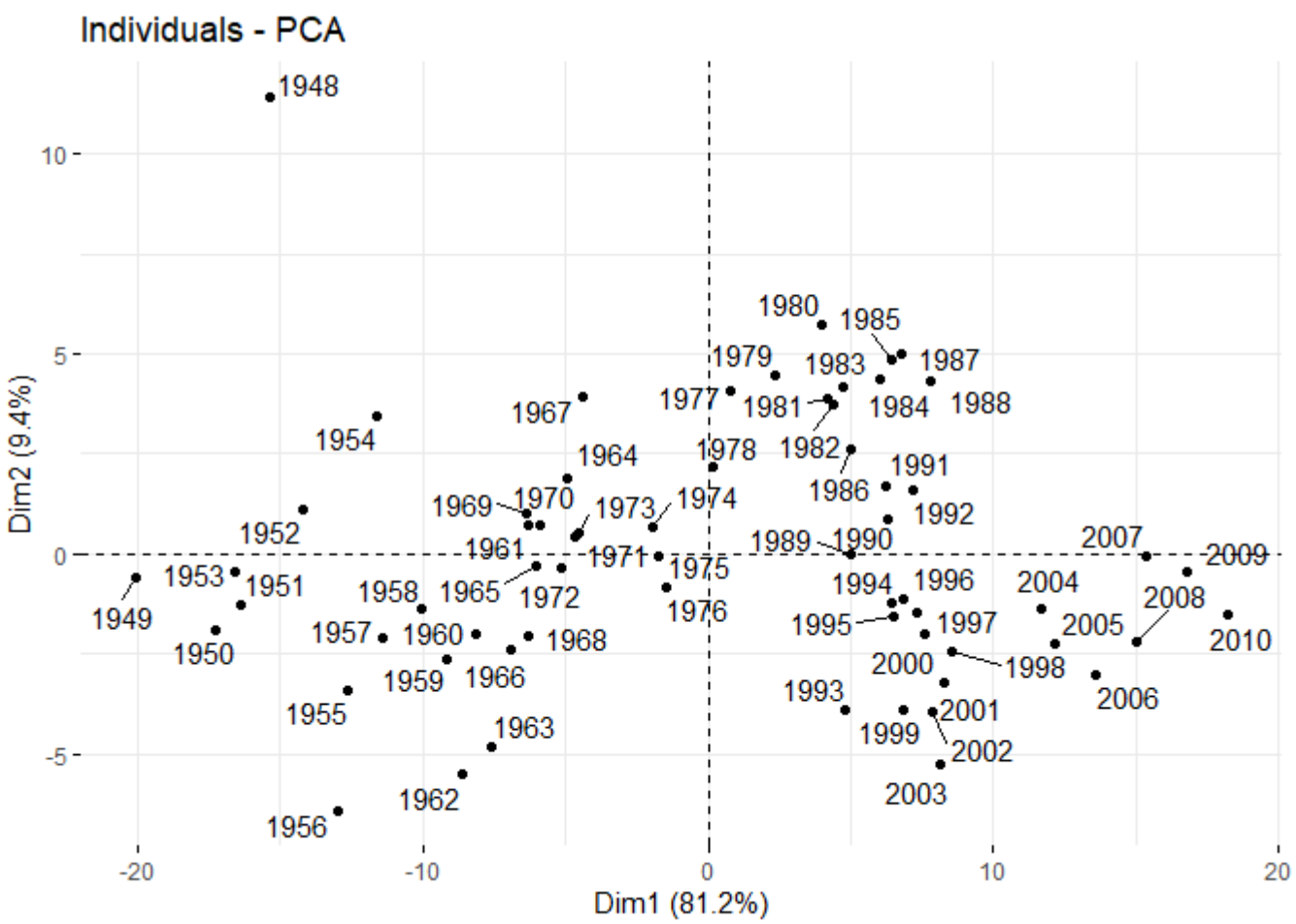
- We then calculate the residual life expectancy corresponding to the vector of mortality quotients and a given age.

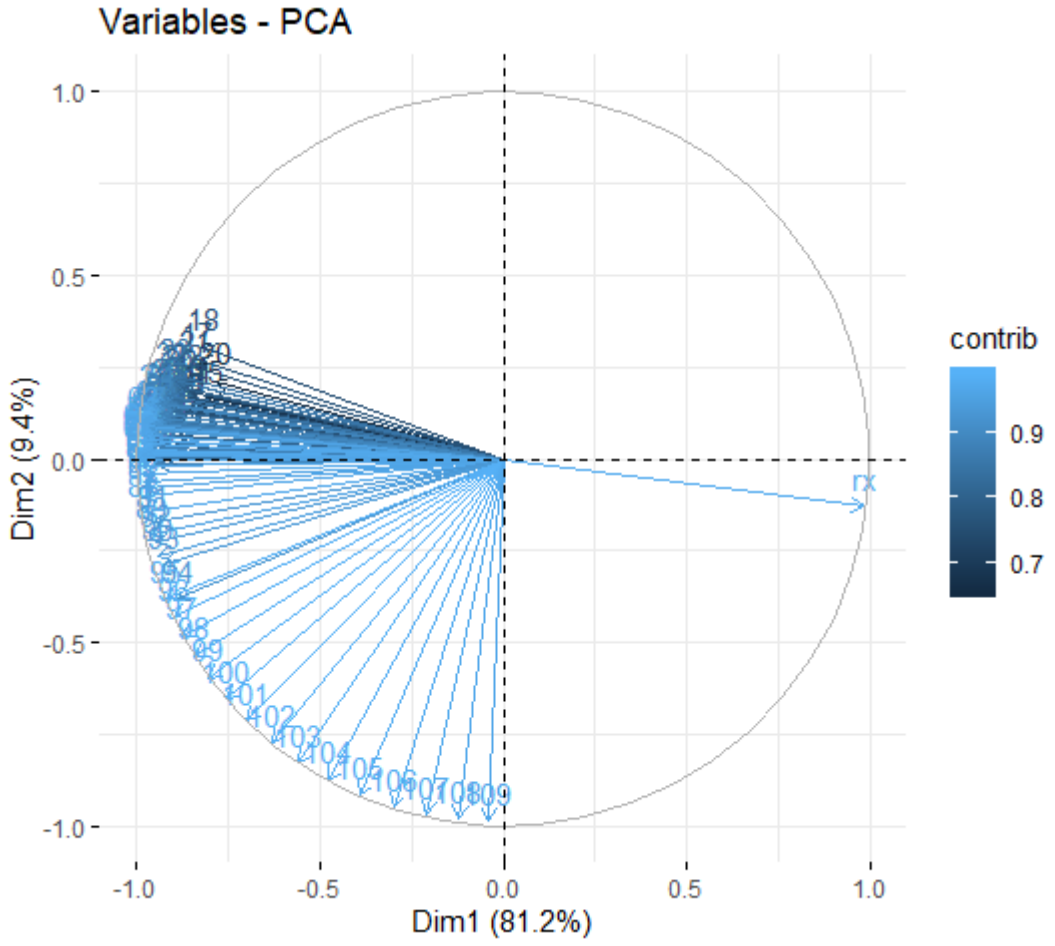
$$ex = \sum \prod 1 - mx$$

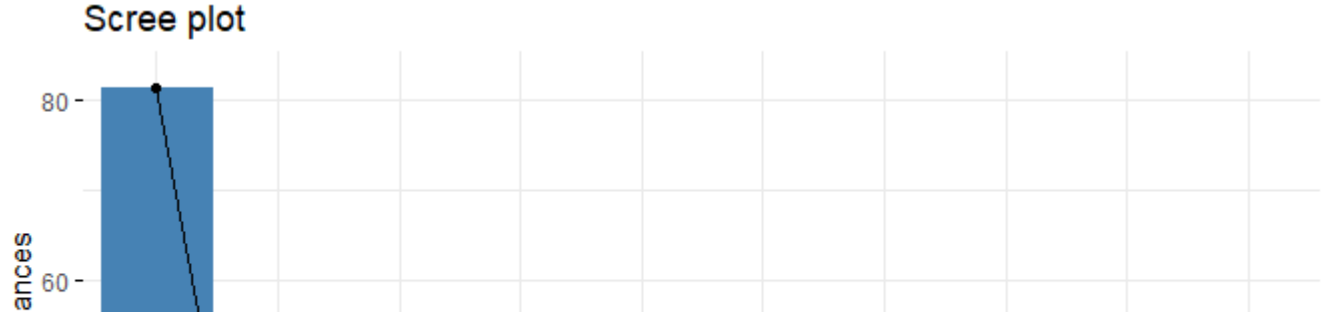
- **Visualization of residual life expectancy as a function of `Year` at ages 60 and 65, facetted by `Gender` and `Country` :**

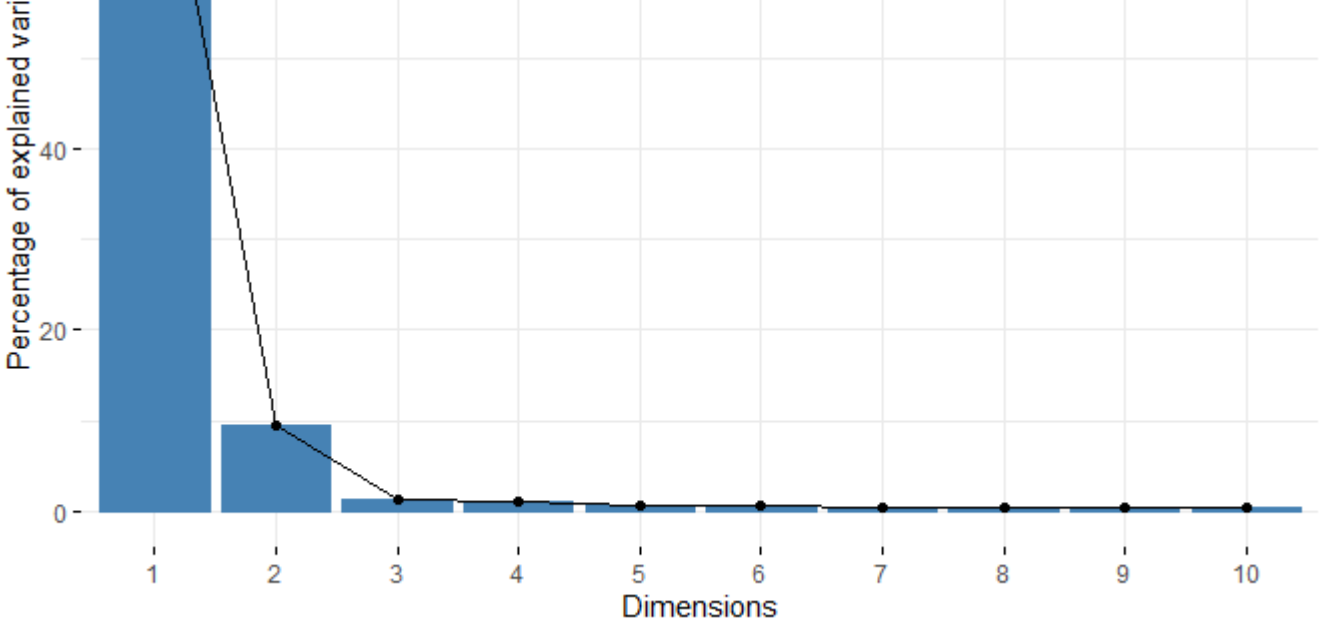
8 PCA and SVD over log-mortality tables

- As we chose to study on the Netherlands as our European country, we also now pick a range of years `1948:2010`. Then we extract the corresponding lines from `life_table_pivot`, with taking logarithms of central death rates. Once we did all that, we perform principal component analysis :


Individuals - PCA
1948
1980
1985
1979
1983
1987
1977
1981
1984
1988
1967
1954
1964
1978
1982
1991
1969
1970
1973
1974
1986
1992
1952
1961
1971
1989
1990
2007
2009
1953
1951
1975
1994
1996
2004
1958
1965
1972
1976
2008
1949
1957
1960
1968
1995
1997
2005
2010
1950
2000
1998
1959
1966
1993
2006
1955
2001
1963
1999
2002
1962
2003
1956
Dim2 (9.4%)
Dim1 (81.2%)
10
5
0
-5
-20
-10
0
10
20


Variables - PCA
rx
contrib
0.9
0.8
0.7
Dim2 (9.4%)
Dim1 (81.2%)
1.0
0.5
0.0
-0.5
-1.0


Scree plot
80
60
ances



- Concerning the screeplot :

Our scree plot displays how much variation each principal component captures from the data. Since our scree plot is a steep curve that bends quickly and flattens out, the first two PCs are sufficient to describe the essence of the data. So we can say that PCA works well on our data.

- Concerning the correlation circle :

We see on the correlation circle that the infant mortality is inversely correlated with life expectancy. Indeed, all the advanced ages are tending down whereas the younger ages are tending up, on the left side of the circle. And the mx arrow is going to the right side of the circle. But we have to consider the fact that the oldest ages reprensent a small percentage of the total population.

- Concerning the biplot :

We see that the recent years are more distributed on the right side of the biplot, which means that they follow the direction of mx on the correlation circle. So the PCA allows us to conclude that the life expectancy is getting higher as time goes by

9 Canonical Correlation Analysis

- We perform a Canonical Correspondance Analysis of Z :
- ☐ Perform a Canonical Correspondance Analysis of Z

10 Lee-Carter model for US mortality

During the last century, in the USA and in western Europe, central death rates at all ages have exhibited a general decreasing trend. This decreasing trend has not always been homogeneous across ages. Governments and many stakeholders in the health and insurance sectors were intreseted in understanding these trends in order to more precisely model and predict the evolution of the mortality rate.
One of the biggest hurdles was to correctly predict the age specific trends and variations of mortality.

The Lee-Carter model has been designed to model and forecast the evolution of the log-central death rates for the United States during the XXth century.

The result is a matrix of age specific foracasted mortality rates.

Let $A_{x,t}$ denote the log central death rate at age x during year $t \in T$ for a given population (defined by Gender and Country).

The Lee-Carter model assumes that observed loagrithmic central death rates are sampled according to the following model

$$A_{x,t} \sim_{\text{independent}} a_x + b_x \kappa_t + \epsilon_{x,t}$$

where $(a_x)_x$, $(b_x)_x$ and $(\kappa_t)_t$ are unknown vectors that satisfy

$$a_x = \frac{1}{|T|}\sum_{t\in T} A_{x,t} \qquad \sum_{t\in T}\kappa_t = 0 \qquad \sum_x b_x^2 = 1$$

and $\epsilon_{x,t}$ are i.i.d Gaussian random variables.

To estimate our vectors we have used the SVD decomposition

$$M = UDV^T$$

a_x and $b_x = \frac{V_{x,1}}{\sum_x V_{.,1}}$ are the age dependent elements of our model and will be used for the rest of the modelisation.

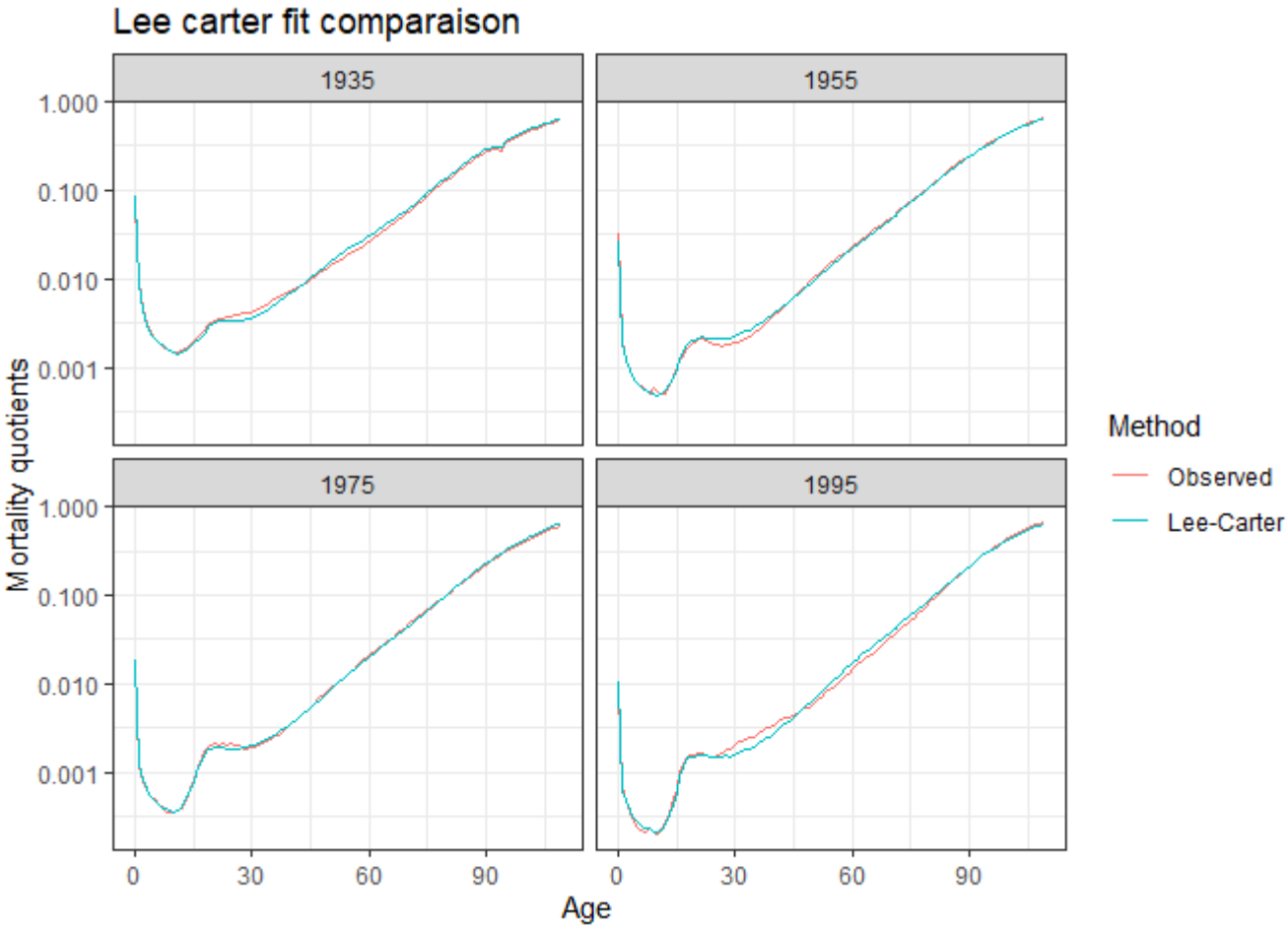
$$k_t = D_1 * U_{1,x} * \sum_x V_{.,1}$$

is the time (Year) dependent element of the Lee-carter model, each value of this vector corresponds to a year.

The following graphs show that our Lee-Carter fit raging from 1933 to 1995 of US mortality data is acceptable and we can proceed with the prediction of our futur k_t values.

10.1 US data

- We fit a Lee-Carter model on the American data (for Male and Female data) training on years `1933` up to `1995`.



- We use the Lee-Carter model to predict the central death rates for years 2000 up to 2015

For the prediction we have used a random walk as mentionned in the 1992 publication by Ronald Lee:

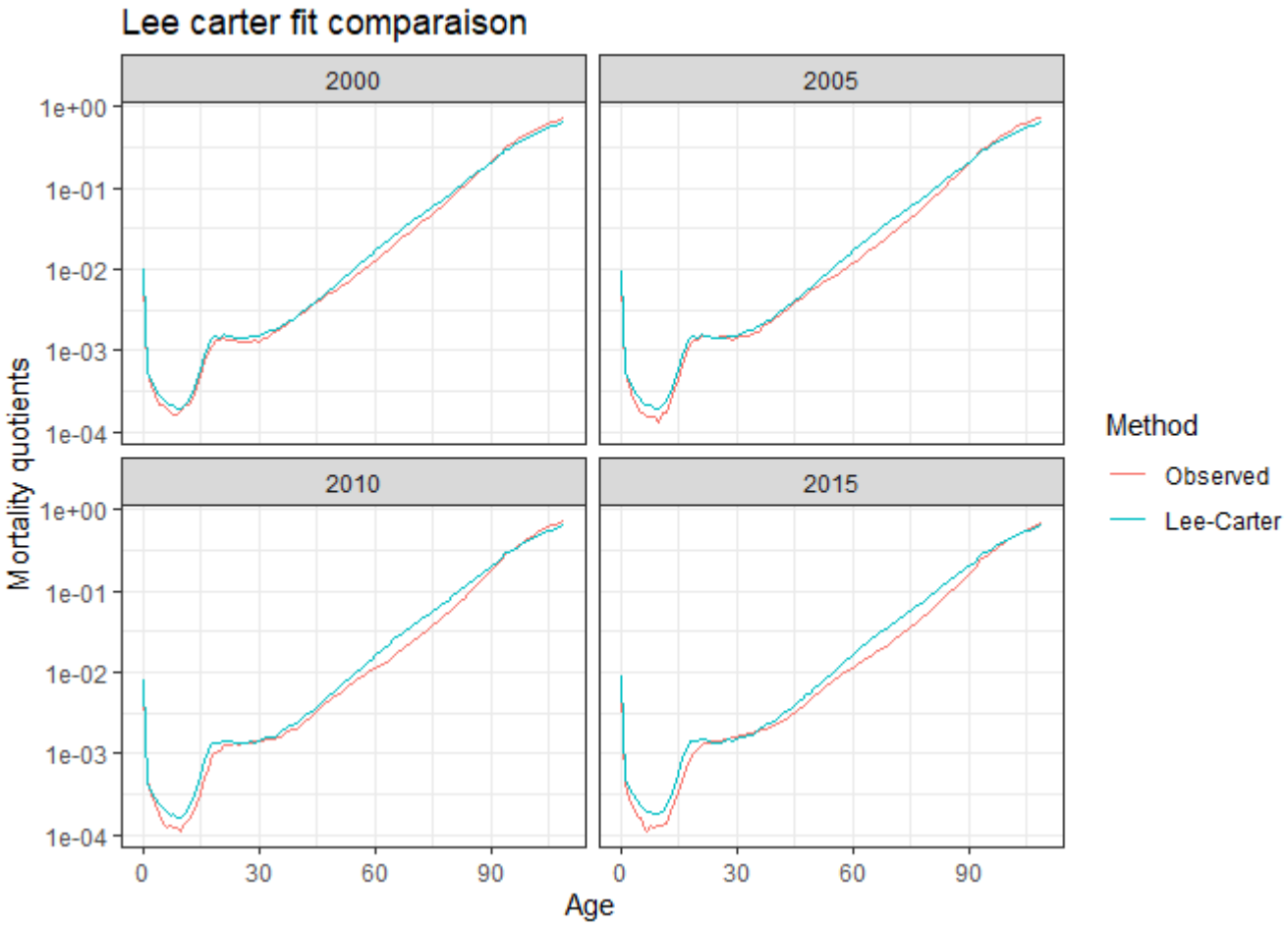
$$k_t = k_{t-1} + d + \epsilon$$

with d being the drift and epsilon a gaussian random variable with the same standard deviation as k_t. the drift is the trickiest part to estimate, it has a massive impact on the trend of ou prediction, we setteled on

$$d = \frac{k_T - k_1}{T}$$

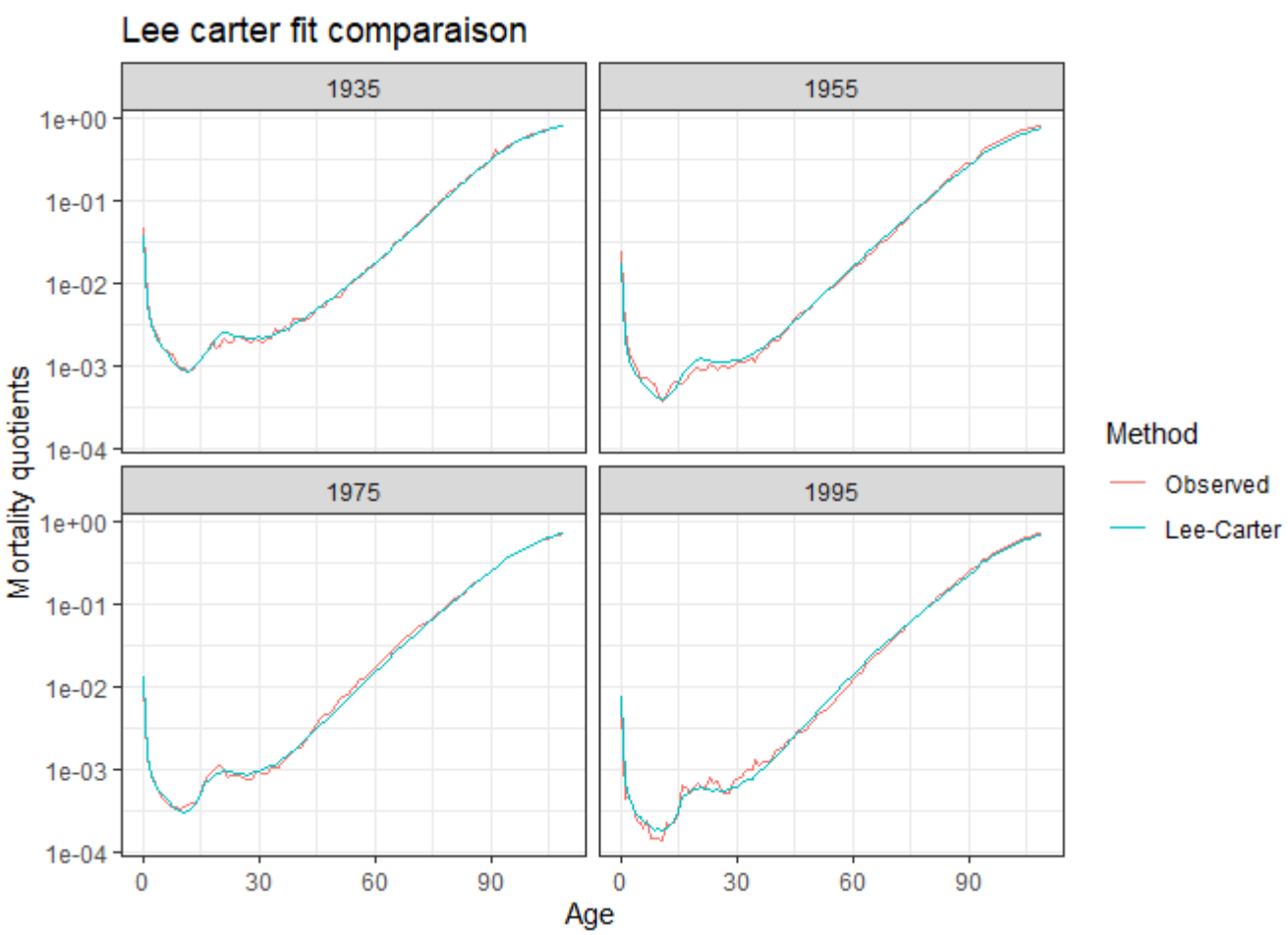
which. We have used the same function as the one used for the fit and just concatenate the new k values to the fitted vector. the resulting matrix will have x lines corresponding to the fitted values and y lines for the predicted years.

- We plot predictions and observations for years $2000, 2005, 2010, 2015$



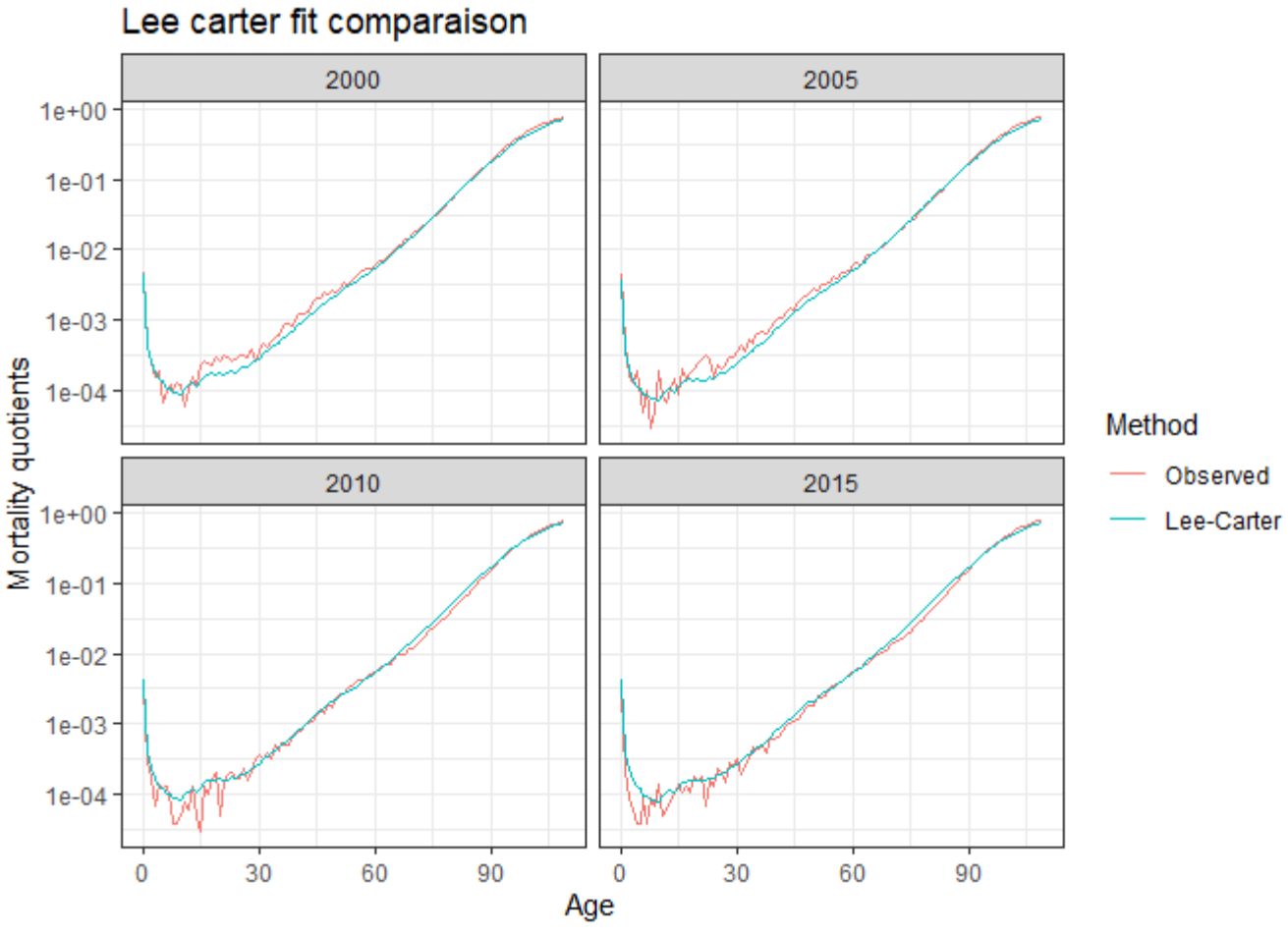
10.2 Application of Lee-Carter model to a European Country

- We fit a Lee-Carter model to a European country



The lee-Carter fit is very close to the observed values. Although it's smoother than the real mortality curve, it shouldn't impact our next prediction.

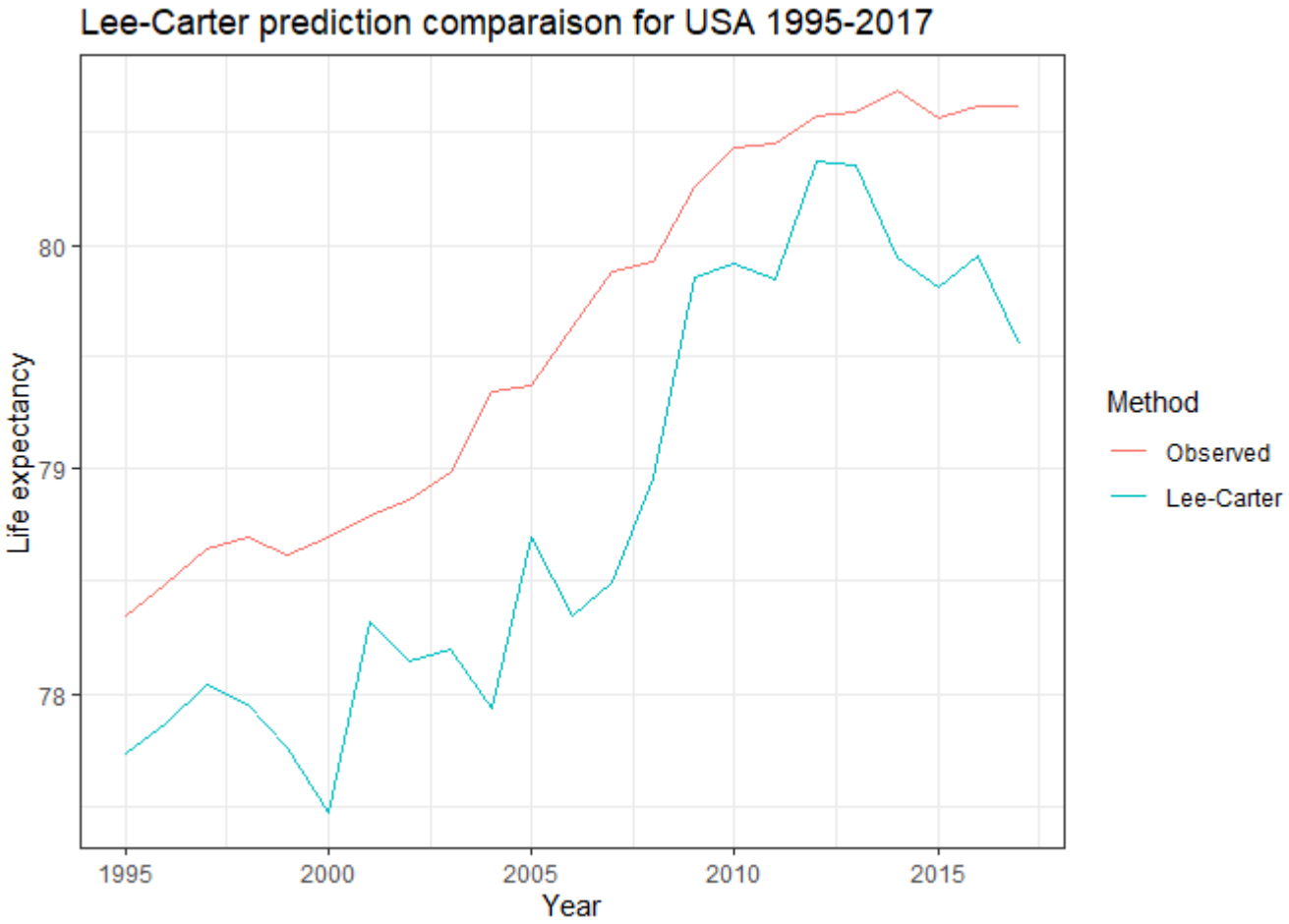
- We compare with rank-2 truncated SVD
- We use the Lee-Carter model to predict the central death rates for years 2000 up to 2015 We Plot predictions and observations for years $2000, 2005, 2010, 2015$



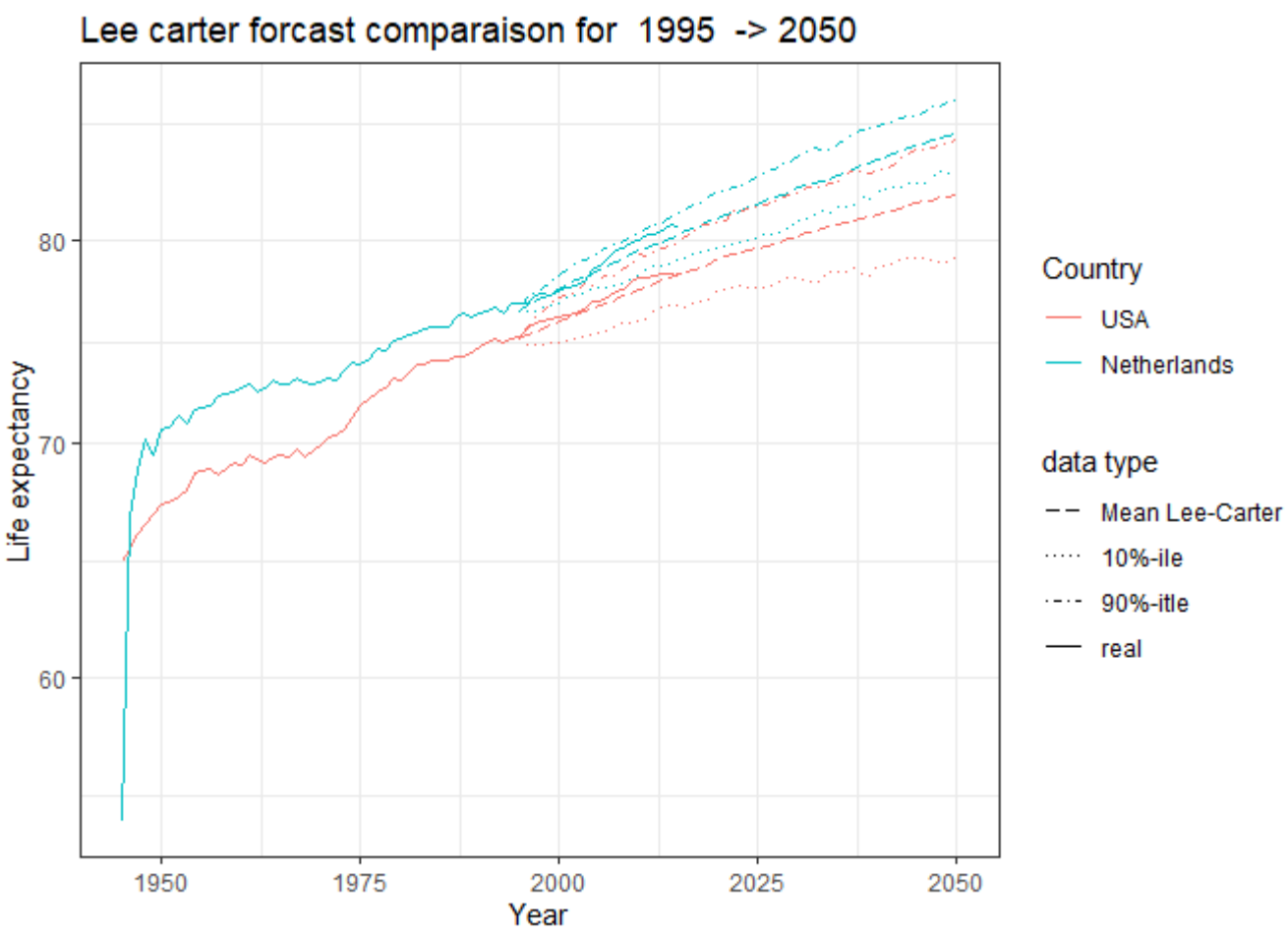
10.3 Predictions of life expectancies at different ages

- We use Lee-Carter approximation to approximate residual life expectations

```
## [1] "Year"        "life_exp_lc"
```

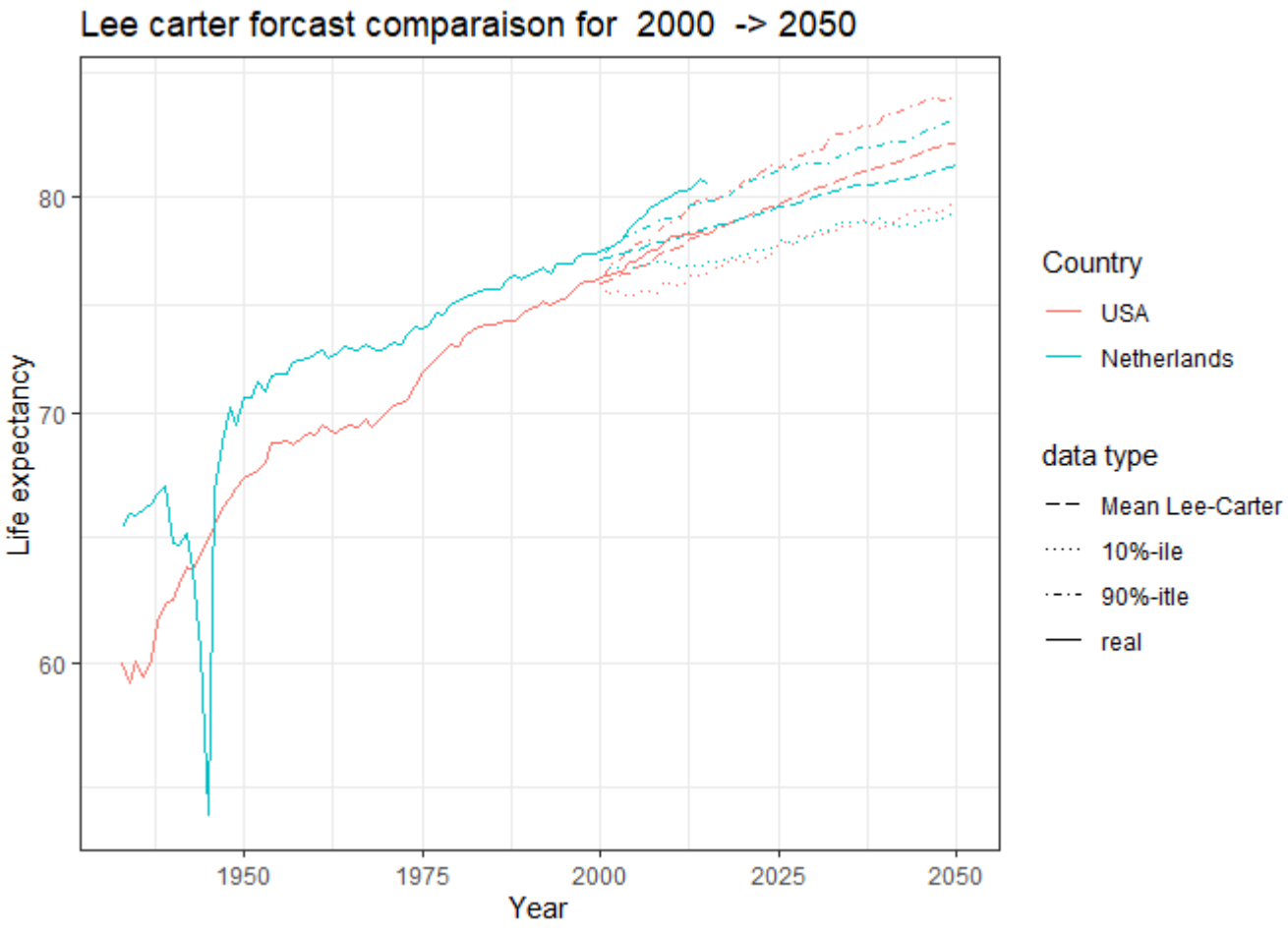


We can se that the prediction is relatively accurate, but we can note that it's very noisy and varies a lot between each iteration. we should try to perform a great number of Lee-Carter forcast and compare the average life expectancy with the observed values.



If we exculde the years of WWII we can get mean life_expectancies very close to real life data, the sudden growth of the post-war era helps the dutch k_t keep it upward trend while it doesn't affect the US Data since the civilian population was not impacted, at least directly, by the war.

While in the plot below we fit from 1933 to 2000 and our drift doesn't account for the sudden drop in the mortality for the war impacted counties and our prediction paint a very dim prospect for the Netherlands with their life expactancy raching the level of the USA in 2016 while in reality they have around a 1.5 year advantage over the USA.



10.4 Issues

Our last attempts af find the right balance for the fit show that the Lee-Carter fit is very sensitive to overfitting, they year choices can be misleading and used to inintetionnally or dishonestly to pain a different picture from reality. The model is also greatly impacted by short but impactful events such as wars or other disaters. It also suggest that current upward trends will continue even if they are due to specific medical advancments or other helpful factors that may not continue in the futur.

Simply put this model while great for predicting the futur mortality based on current trends, it will struggle to account for external factors or time-limited events and sould be used and interpreted with great caution.

Finally we sould explore other ways to get a more accurate k vector, an obvious possibility is a better understanding of the drift and trying other possibilities.

11 References

Life tables and demography

- Human Mortality Database (https://www.mortality.org)
- Tables de mortalité françaises, Jacques Vallin et France Meslé (https://www.lifetable.de/data/FRA/FRA000018061997CY1.pdf)
- [Modeling and Forecasting U.S. Mortality, R.D.Lee and L.R. Carter, JASA 1992]
- [Les dimensions de la mortalité, S. Ledermann, Jean Breas, Population, 1959]

Graphics and reporting

- Interactive web-based data visualization with R, plotly, and shiny (https://plotly-r.com/index.html)
- R for Data Science (https://r4ds.had.co.nz)
- Layered graphics (http://vita.had.co.nz/papers/layered-grammar.pdf)
- Plotly (http://plotly.com/)

Tidyverse

- tidyselect (https://tidyselect.r-lib.org/articles/tidyselect.html)
- dbplyr (https://cran.r-project.org/web/packages/dbplyr/vignettes/dbplyr.html)
- data.table (https://github.com/Rdatatable/data.table)
- DT (https://rstudio.github.io/DT/)

PCA, SVD, CCA

- FactoMineR (http://factominer.free.fr/index_fr.html)
- ade4 (http://pbil.univ-lyon1.fr/ade4/accueil.php)
- FactoInvestigate (http://factominer.free.fr/reporting/index_fr.html)
- PCA and Tidyverse (https://cmdlinetips.com/2019/05/how-to-do-pca-in-tidyverse-framework/)
- tidyprcomp (https://broom.tidyverse.org/reference/tidy.prcomp.html)