Final Project

Master’s degree in Industrial Technology

Evaluation and Prediction of World's Population Dynamics

MEMORY

Autor: Marc NOGUÉ PUJIUILA

Supervisor: Prof. Andrea RINALDO

Co-supervisors: Post Doc. Damiano PASETTO

Post Doc. Paolo BENETTIN

Date: July 2018

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Contents

List of figures

List of tables

1. Introduction

By 2100, the world could be home of 11 billion people[1]. Nowadays, it is well known that 60% of the world population lives in Asia[2] but, will this trend change in the following years? How fast is the population increasing today compared to in the past? Until the 19th century the world population was less than 1 billion people. Today is more than 7 billion. The 20th century experienced an unprecedented growth in the world population. However, will it reach a steady state? This are some of the questions that can be solved by applying demographic theories. Figure 1 shows the well-known graph of the exponential growth in the world population.

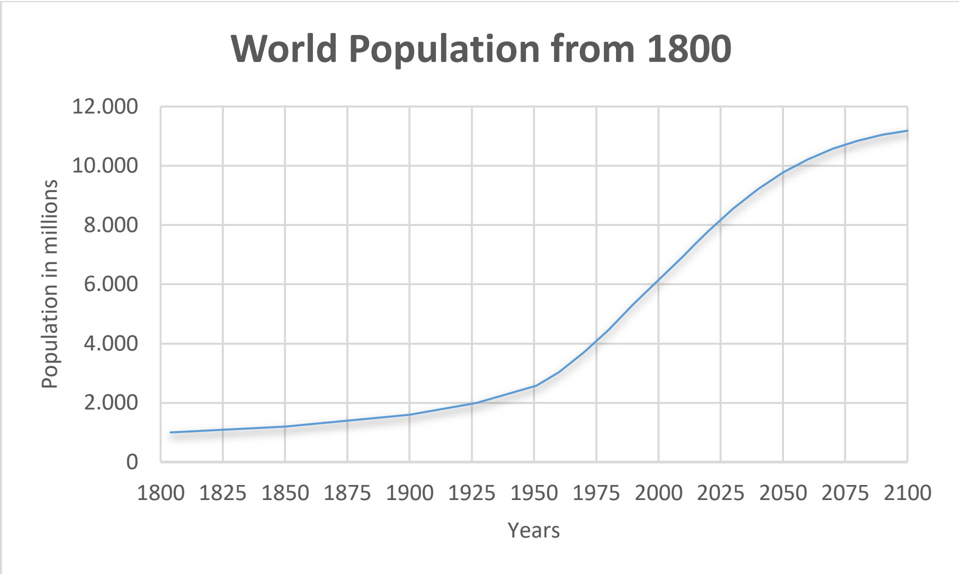


Figure 1: World population from 1800

Population trends are closely related to economic development. Many developing countries are passing through the stage of demographic transition. Understand these changes in population is really important in order to know how many houses, schools, job opportunities, food supply and medical health care amongst others are to be increased. As a result of this transition the net volume of imports and exports will also change and this means that foreign investment might be needed.

Population is also very important for the environment. Human activities have a huge impact towards natural resources consumption, pollution and land use. The larger the population is, the bigger is the effect on the environment. But not only is the size the variable that depicts the impact of one population, the level of wealth is also a very important parameter to take into account. For example, in Europe, food consumption is much higher than in Africa[3]. The study of population dynamics is very useful in order to better project the impact of all the mentioned variables.

This project aim is to understand how the population change over time focusing on the age distribution and also predict the patterns of the future demography. Is near future population expected to be older? Or does the new born mortality keep decreasing over time? If not, which is the lowest level? These are some of the questions that are answered in this report.

It studies the feasibility of a whole new perspective in terms of population dynamics. By applying a model used in hydrology it predicts the future population. It also models the mortality function using a statistical distribution.

In terms of population projections/forecasts there are lots of resources available. However, for this project in particular, it has been chosen the United Nations data base and in some cases government resources.

1.1 Driving variables: Fertility, Mortality and Migration

There are several demographic variables that take part in the population dynamics however there are 3 that play a central role in the study of human population: fertility (birth rate), mortality (death rate) and migration.

Fertility is the ability to reproduce. As a measure, the total fertility rate (TFR) can be used. This variable tells the average number of children that would be born to a woman over her lifetime if:

- She was to experience the exact current age-specific fertility rates (ASFRs) through her lifetime.
- She was to survive from birth through the end of her reproductive life.

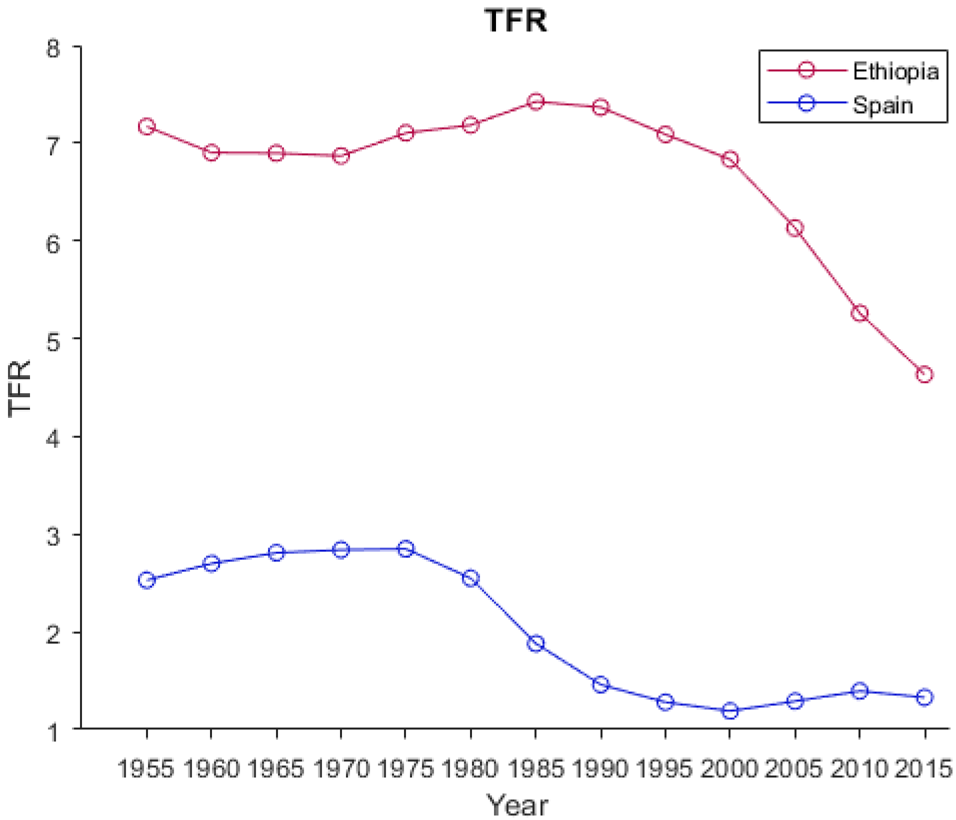


Figure 2: TFR in Ethiopia and Spain

As Figure 2 shows, the differences between developed countries and less developed countries are very remarkable. Although through years this differences are meant to decrease they're still very significant.

The mortality rate is the number of deaths in a particular population. It considers both general and specific causes of death. A very usual plot used in order to better understand the mortality of a population is shown in Figure 3.

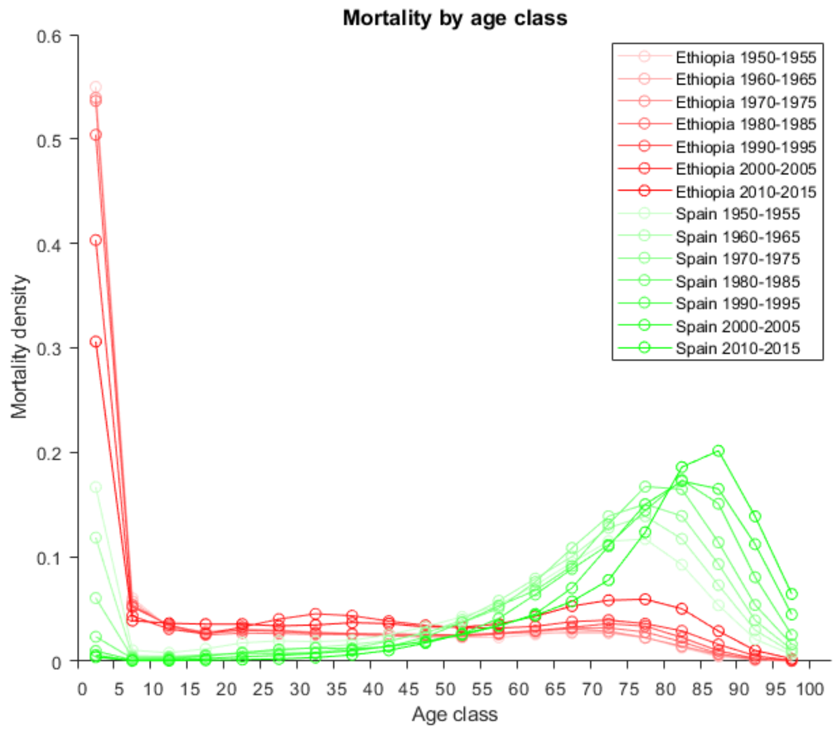


Figure 3: Ethiopia and Spain mortality rate by age

This plot tells us how different the mortality trends in this two countries are. In Ethiopia for example there is a high birth mortality rate whereas in Spain it is much lower. Also in Ethiopia the population is younger than in Spain. The peak in the last age classes in Spain confirms that the percentage of population that die in the last age classes is bigger than in Ethiopia. Meaning that the median age of the population is bigger.

Last but not least there is the migration variable. It is defined as *the movement by people from one place to another with the intentions of settling, permanently or temporarily in a new location.* Migration in many ways generates a positive impact for the host countries. They are enriched by cultural diversity and job vacancies can be filled. However, there is also a negative part. Immigrants might be exploited and unemployment may rise if there are no restrictions for incomers.

Without question these three factors are the main components that drive the changes of populations. However, some events, such as a war or a natural disaster, may have a huge impact in one population. In Serbia[4], as a result of the World War 1, it has been estimated that 15-28% of the country total population were dead. In 1970 a tropical cyclone named Bhola cyclone[5] killed 300.000 people in Bangladesh. It has been considered the world's all-time deadliest weather event. Figure 4 shows the devastating effect of this cyclone in terms of mortality over population.

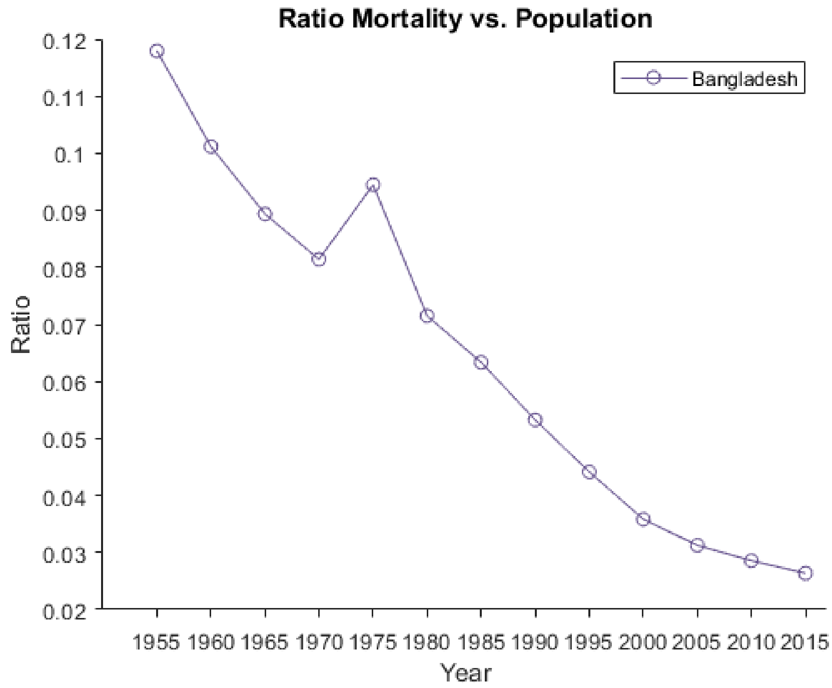


Figure 4 : Effect of Bhola Cyclone in Bangladesh

1.2 State of the art

The state of the art in population projection models can be explained through the uncertainty of the projections. Nowadays there are three different approaches: the classical cohort-component, the frequentist probabilistic model and the Bayesian paradigm [6].

The cohort component[7] in broad terms can be understood by applying the demographic balancing equation:

$$P_{i+1} = Survived\ population + births + net\ migration \qquad \textbf{(1.1)}$$

Where P_{i+1} is the population one period of time after. Basically this equation says that in order to project the population is needed to compute the survived population in a certain time period plus the newborns and the total net number of migration of the same time period. The first task to perform is project the total fertility and mortality rates based on assumptions, usually by applying age-specific rates. Then the net migration is needed to be estimated based on empirical regularities and judgments. The deterministic approach to the cohort-component model is the one more used in demography as of today. The uncertainty is introduced by the demographer through his judgment of the most likely set of elements that will result in future changes in the above mentioned variables. In order to allow some degree of uncertainty the United Nations has developed various scenarios to see the effect of different fertility levels.

The second approach tries to fix the flaws of the first one by providing consistent measures of uncertainty to demographic estimates. The lack of probabilistic meaning for the mortality projection was solved by Lee-Carter[8] in the early 1990s. This model allows the derivation of long-term forecasts of the level and age pattern of mortality and fertility and is based on matrix algebra. Despite having an appealing characteristic such as being relatively simple with a small number of parameters it has its limitations. In the short run it behaves well and the results are really coherent, however in the long run forecasts may provide inconsistent results.

The third approach, the Bayesian projection, is *an interpretation of the concept of probability, in which, instead of frequency or propensity of some phenomenon, probability is interpreted as reasonable expectation representing a state of knowledge or as quantification of a personal belief.* Bayesian probability theory provides a mathematical framework for performing inference, or reasoning, using probability [9].

The Bayesian theory says that for an unknown quantity θ and a vector x with sample information the likelihood function $L(x|\theta)$ provides empirical information on θ. There's also a distribution $\pi(\theta)$, named prior distribution that represents the initial uncertainty on θ. The Bayesian paradigm is a very powerful methodology in order to solve the uncertainty issue in population forecasts when too many subjective assumptions are done. All these beliefs, in the Bayesian approach, can be gathered in the prior distribution.

1.3 Ecohydrological model

The main idea of this project is to adapt a model being used in hydrology in order to understand the world population dynamics.

In hydrology a drainage basin or watershed is considered as storage. A drainage basin is an area of land where precipitation collects and drains off into a common outlet. An outlet might be a sea, bay or any other water reservoir. See Figure 5.

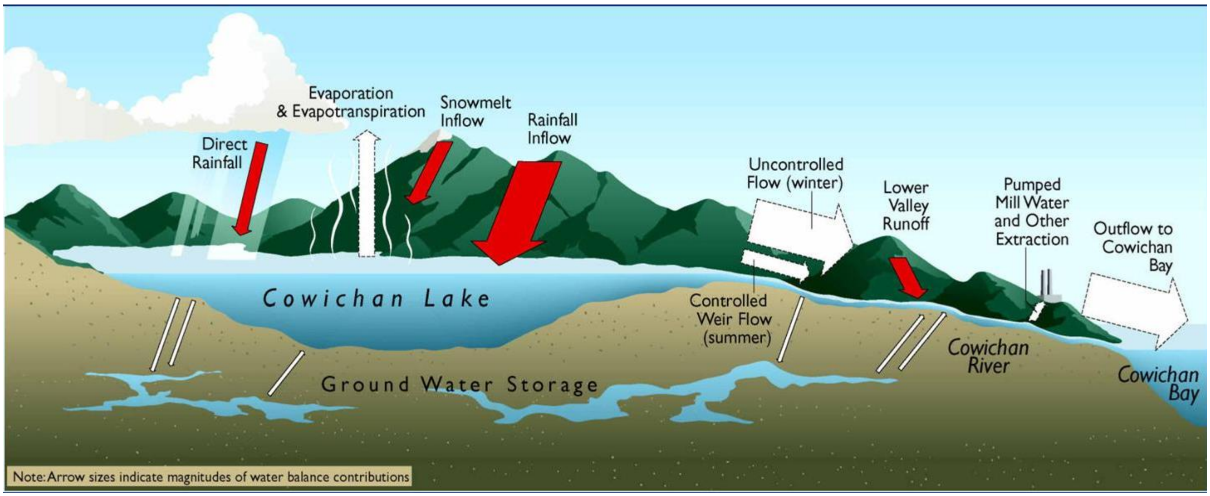


Figure 5: Drainage basin fluxes

In the model proposed in this project the population is interpreted as the storage. The storage is increased due to inflows (newborns & immigration) and decreases through outflows (death & emigration). In the same way a storage, such a reservoir, over time can change its level due to transpiration or evaporation population over time can change its volume too due to the migratory fluxes. This variation might be positive or negative: if the immigration exceeds the emigration will be positive and if the emigration exceeds the immigration will be negative. In other words, if there is an increase in the population the migration will be positive. The following diagram (Figure 6) presents the analogy between the model used in hydrology and the model proposed in this project for population projections.

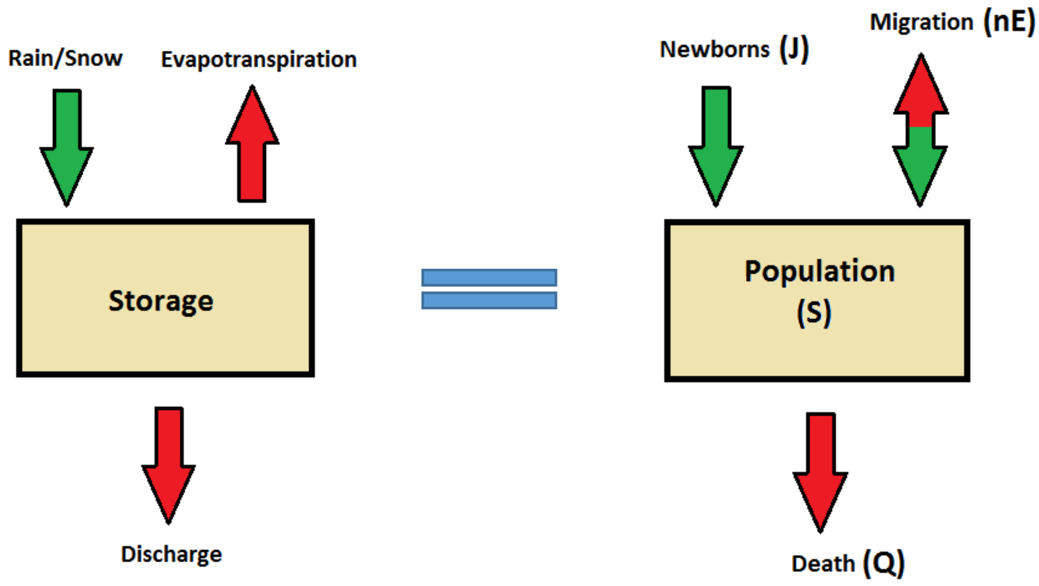


Figure 6: Model comparison

2. Data

2.1 UN role in the population dynamics

The United Nations[10] is a unique organization formed by independent countries that decided to join in order to work for the world peace and social development. UN officially began to exist on October 24, 1945, with 51 countries considered as founding members. By the end of 2008, the number of Member States of the United Nations was 192 countries. There are currently 193 Member States, after the entry of South Sudan on July 14, 2011.

There are many different departments that constitute the UN, however for this project the one of interest is the Department of Economic and Social Affairs (DESA) [11]. The work of this division is to achieve a social, economic and environmental sustainable development. Inside UN DESA , the Population Division was established in 1946. Since the early days it has played an active role in the intergovernmental dialogue on population and development, producing constantly updated demographic estimates and projections for all countries. It studies population dynamics and monitors demographic trends and policies worldwide. Population estimates and projections prepared by the Division for all countries – on fertility, mortality, international migration, urbanization, and population size and structure – are widely used by various entities.

The UN official population estimates and projections releases a new version every 2 years. The one being used in this report was released on 21 June 2017. The next one is due in the first half of 2019. This revision provides population estimates for the period 1950-2015 for a total of 233 countries and areas. In fact there are only 201 countries. The UN also added 32 categories of territories or areas with at least 90.000 inhabitants in 2017. Estimates of the population are presented for five-year periods, starting with 1950-1955 and ending with 2010-2015. Stock estimates refer to 1 July of the year in question. Period estimates may be assumed to refer to the mid-point of the period concerned (e.g. the mid-point of the period 1 July 1970 to 1 July 1975 is the 1 January 1973). Data presented by single calendar years or single groups of age are derived by interpolation.

UN also gives projections from 2015-2100. Calculations are carried out by five-year periods using data classified by five-year age group. Projections of stocks are presented for every year that is a multiple of five from 2020 to 2100. Projections of the components of population change are presented for five-year periods, starting with 2015-2020 and ending with 2095-2100. Stock estimates refer to 1 July of the year in question. Data presented by single calendar years or single groups of age are derived by interpolation.

About half of the 233 countries and areas that the World Population proportionate estimates do not report official demographic statistics with the detail necessary for the preparation of cohort-component population projections. The Population Division from the United Nation undertakes its estimation work in order to close those gaps. The availability of data gathered by major survey programs, such as the Demographic and Health Surveys or the Multiple-Indicator Cluster Surveys, are useful in generating some of the data that is not currently being produced by official statistics.

2.2 Main UN data sets

The UN presents 3 major types of projections/estimates[12]. The standard projections, probabilistic projections and special aggregates. Figure 7 shows a quick overview of the different kinds of data available for the standard projections since this one has been the one used for this report. In the diagram there are only the ones used, however in the UN database there are several more.

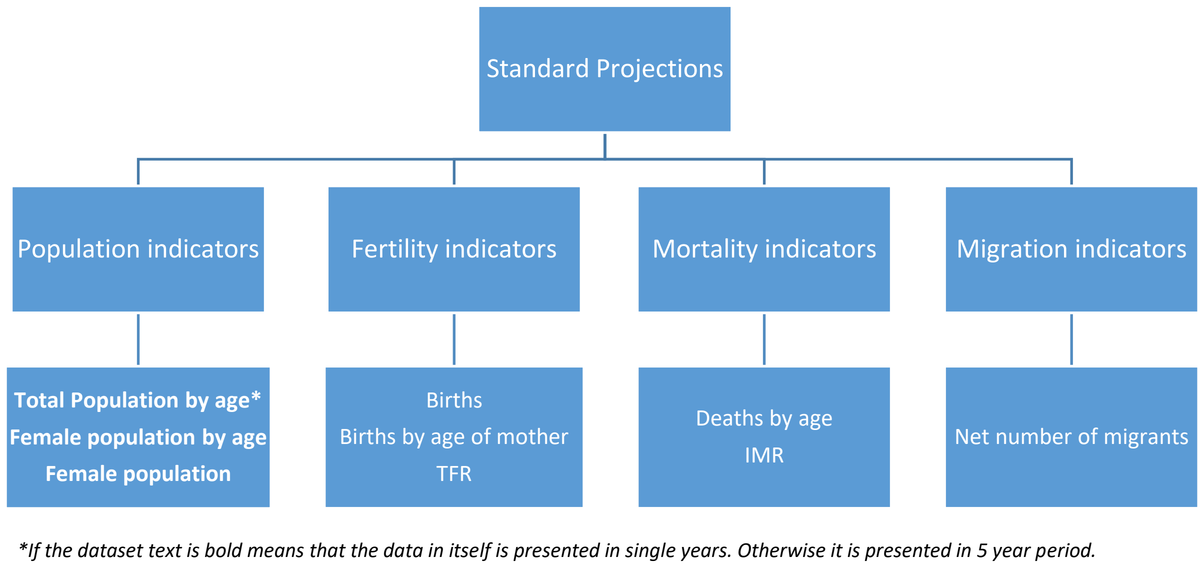


Figure 7: UN data sets used in this project

In order to give a more detailed view of the different data sets a brief explanation of the most important of them is given below.

Total population (both sexes combined) by five-year age group, region, subregion and country, 1950-2100 (thousands)

This data set contains information regarding the world population. It is presented by group of ages (5 years each) starting from 0-4 until 100+. It presents the data of the region, subregion or country every 5 years from 1950 to 2100. There are listed only those countries with more than 90.000 inhabitants in 2017. So Andorra for example is not there.

The population data set is structured as shows the Figure 8. The first column specifies the country /region, the second the year and another one the code of the country/region (not shown in the figure). All the other columns refer to the estimate of individuals in the different group of ages.
In all the countries from 1950 to 1985 (included) the population is registered from 0 years to 80+. No further separation (80 to 100) is done due to lack of consistent data.

Most of the data sets that UN presents have more or less the same structure. The main difference that they have is the way the data is presented. Either by age classes of the different years or just by years.

Region, subregion, country or area *	Reference date (as of 1 July)	Total population, both sexes combined, by five-year age group (thousands)										
		0-4	5-9	20-24	70-74	75-79	80+	80-84	85-89	90-94	95-99	100+
WORLD	1950	338387.265	269870.197	222813.664	37431.362	21923.240	14264.753	...	...	...	...	...
WORLD	1955	405612.155	315829.192	232598.746	40480.240	23763.081	16092.680	...	...	...	...	...
WORLD	1960	432642.795	382144.568	248821.446	44225.916	25977.020	17957.773	...	...	...	...	...
WORLD	1965	480601.050	409585.332	252299.729	47761.883	29549.871	20973.787	...	...	...	...	...
WORLD	1970	523700.776	460552.249	298535.489	55628.158	33116.843	25340.812	...	...	...	...	...
WORLD	1975	543301.991	504504.740	363876.090	65555.904	39243.523	29843.483	...	...	...	...	...
WORLD	1980	548178.631	525955.325	394175.563	75512.963	47618.743	35820.568	...	...	...	...	...
WORLD	1985	592792.343	532911.778	445706.555	84815.892	55190.301	44344.783	...	...	...	...	...
WORLD	1990	644867.279	578146.374	489977.618	86417.436	63046.358	...	34851.456	14466.809	4033.885	770.805	90.055
WORLD	1995	623061.131	631533.678	510557.774	103627.739	64804.836	...	40651.988	17832.387	5389.093	987.378	117.901
WORLD	2000	609348.362	611810.802	518116.336	118770.299	79104.481	...	42485.413	21323.121	6826.818	1348.600	150.362
WORLD	2005	623820.375	600035.988	564207.361	129980.883	91606.737	...	53373.437	23061.075	8524.613	1784.753	212.583
WORLD	2010	651602.453	615729.550	616945.974	145308.797	101812.375	...	63804.385	30343.546	9740.413	2385.612	301.437
WORLD	2015	673649.680	644475.619	600866.814	153908.373	115531.676	...	71830.934	37295.064	13475.354	2916.508	433.635
More developed regions	1950	82923.299	67364.331	71190.753	18186.369	11249.608	8074.107	...	...	...	...	...
More developed regions	1955	87629.332	81827.430	67220.743	20229.376	12947.834	9689.888	...	...	...	...	...
More developed regions	1960	89504.796	87189.637	71836.486	22306.499	14652.409	11459.264	...	...	...	...	...
More developed regions	1965	88011.978	89562.262	66875.683	24370.229	16566.353	13557.146	...	...	...	...	...
More developed regions	1970	83005.031	88600.678	81165.008	27566.035	17980.430	15958.656	...	...	...	...	...
More developed regions	1975	81526.666	83162.048	86986.072	32118.632	20692.812	18370.459	...	...	...	...	...

Figure 8: Example of data set structure (Some age classes have been hided).

In order to estimate the distribution of the 80+ class in the missing age-classes (80-100+) an exponential distribution function has been proposed. The study of how this distribution could fit for this situation is explained in detail in the next section.

Births (both sexes combined) by region, subregion and country, 1950-2100(thousands)

This data set contains information regarding the fertility. Number of births over a given period. Refers to five-year periods running from 1 July to 30 June of the initial and final years. Data are presented in thousands.

Deaths (both sexes combined) by five-year age group, region, subregion and country, 1950-2100 (thousands)

This data set contains information regarding the mortality. Number of deaths by five-year age groups. Data are presented in thousands.

Female population by five-year age group, region, subregion and country, 1950-2100 (thousands)

Population by Five-Year Age Groups - Female. De facto population as of 1 July of the year is classified by five-year age groups (0-4, 5-9, 10-14, ..., 95-99, 100+). Data are presented in thousands.

Births by five-year age group of mother, region, subregion and country, 1950-2100 (thousands)

Number of births over a given period classified by age group of mother (15-19, 20-24, 25-29, 30-34 … 45-49). Refers to five-year periods running from 1 July to 30 June of the initial and final years. Data are presented in thousands.

Total fertility by region, subregion and country, 1950-2100 (live births per woman)

The average number of live births a hypothetical cohort of women would have at the end of their reproductive period if they were subject during their whole lives to the fertility rates of a given period and if they were not subject to mortality. It is expressed as live births per woman.

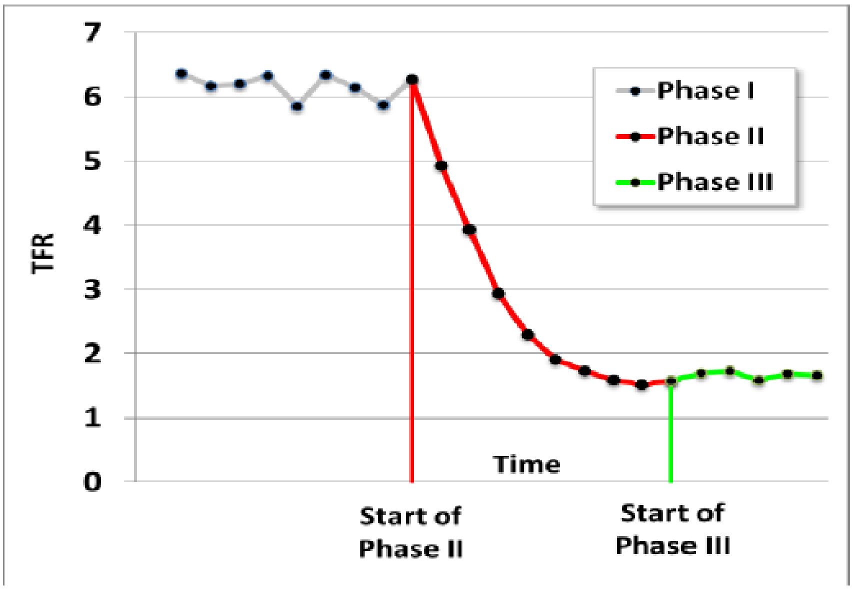
3. Methods

3.1 Elements of demographic theory

Along with the economic development, tendencies of birth-rate and death rate are different. The demographic transition theory [13] is the basis for projections of future country-specific fertility levels. The theory was firstly proposed by Warren Thompson, an American demographer, in 1929.
Demographic transition refers to a population cycle that begins with a fall in the death rate, continues with a phase of rapid population growth and concludes with a decline in the birth rate

The stages of this theory are:

- Phase 0 is characterized by high birth rate, death rate and low rate of population growth.
- Phase I is characterized by high and stationary birth rate, rapidly declining death rate and very rapid increase in population.
- Phase II is characterized by a falling birth rate, low and stationary death rate and rapidly rising population.
- Phase III is characterized by low birth rate and low death rate with stationary population at a low level.



Source: Alkema et al. (2011).

Figure 9: Schematic phases of the TFR transition (in live births per woman)

The start of Phase II was deemed to have occurred before 1950 for countries where the maximum births per woman was less than 5.5, such as Germany, Argentina or Australia. Its end was defined as the midpoint of the time periods when the first two successive increases were observed, after the level of total fertility had fallen below 2 births per woman.
The projected level of total fertility has been allowed to fall below that threshold, reflecting uncertainty with regard to the historic minimum level of fertility (at the end of Phase II) before the start of a recovery (in Phase III).
Based on the most recent population and demographic data available, the United Nations Department of Economic and Social Affairs determined that all countries had begun or already completed their fertility transition, being in either Phase II or Phase III. Thus, fertility transition in these two phases were modelled separately, while Phase I was not modelled in the 2017 UN Revision.

Phase 2 can be subdivided in two phases since the decline is fast from onset to the intermediate of the Phase II and becomes slower before reaching the Phase III.

- Phase II.1, where the countries have witnessed a recent and intense fertility decline.
- Phase II.2, where the countries have already reached a fertility level over 50% from their pre-transition level and are on a slower decline of the fertility close to a TFR of 2.

Table 1 informs of some countries stage in regards of the theory:

Phase III	Phase II.1	Phase II.2
Germany, United Kingdom, Italy, Japan, Spain, USA, France, Ireland, Romania, Poland, Canada and others.	Guatemala, Egypt, Zimbabwe, Sudan, Niger, Afghanistan.	India, Mexico, Nicaragua, Costa Rica, Saudi Arabia, Morocco, Tunisia, Turkey.

Table 1: Countries stage [14]

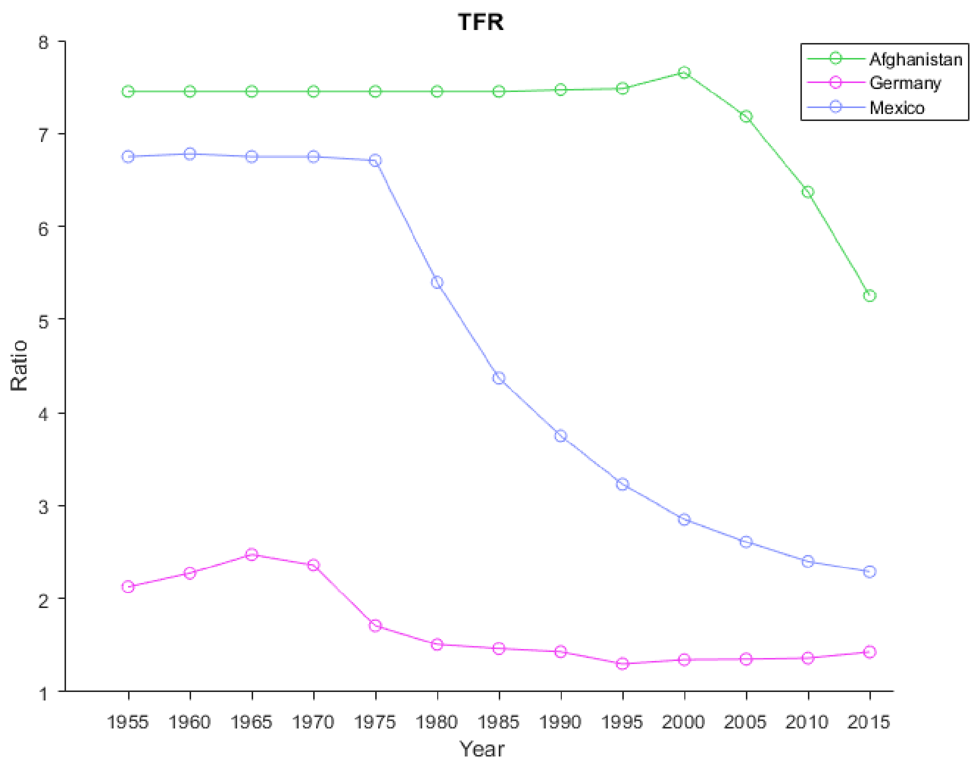


Figure 10: Transition of Mexico, Germany and Afghanistan from 1950

Figure 10 gives an example of the different three possible stages of transition theory a country can be facing right now. Germany has already reached the Phase III with a TFR of 1.5. Then there is Mexico, a country that has faced the whole demographic transition over the last 40 years. Starting with a TFR of 6.8 in the period 1950-1955 it has transitioned to a TFR of 2.2 in 2016. The last one is Afghanistan a country still facing the Phase 2.1 showing a strong and continuous decrease in the TFR. In 2016 the TFR was 4.6. It is expected that by 2050 this country will have achieved the 2.0 TFR level.

Mortality, the second variable that determines the dynamics of a population. It has been defined by the United Nations as the life expectancy at birth by sex. In the 2017 Revision two models have been used. The first one describes the gradual increase through time in female expectancy at birth. This model presents two phases in regards of the transition from high to low levels of mortality. The first phase consists of the initial slow growth in life expectancy associated with the diffusion of improved hygiene and nutrition, followed by a period of accelerated improvements, especially in the mortality of infants and children, associated with social and economic development accompanied by interventions in public health and basic medical care (for example, infant feeding, water and sanitation, and childhood immunization programs). The second phase begins once the easiest gains, mainly against infectious diseases that often strike in childhood, have been achieved. The second phase is characterized by continuing gains against infectious diseases across the age range and also against non-communicable diseases that strike primarily at older ages.[13]

International migration is the most difficult component to project in terms of population dynamics. It is a very volatile process due to the fact that the movement of people from one country to another is very often a response to rapidly changing economic, social, political and environmental factors. One of those factors that makes this an extremely challenging component to predict is the fact that some countries that historically have been countries of origin have now become countries of destination, and vice versa. In the cases that migration flows have historically been small and have had little net impact in terms of population the assumption that migration will remain constant over the time period is usually acceptable.

International migration in terms of the UN is understood as net migration. Net migration is the difference between the number of immigrant and the number of emigrants for a particular country and period of time. When preparing assumptions about international migration, although this is not the purpose of this report, one must take into account several pieces of information:

1. Information on net international migration or in immigration and emigration
2. Data on labor migration flows
3. Estimates of undocumented migration
4. Estimates on refugee movements in recent periods

3.2 Main balances and equations

The main equation used in this project is a mass balance. A material balance is an application of conservation of mass to the analysis of physical systems.

$$Accumulation = Input - Output \qquad \textbf{(3.1)}$$

Before introducing the model parameters it is important to explain the notation used in this report. For the unit of time, either being single years or year steps, there is the index i. Refer to Figure 7 in order to see what i stands for in the different datasets.

Year (i)

$0 \rightarrow 1950$
$1 \rightarrow 1955$
$2 \rightarrow 1960$
...
$year\ i = 1950 + 5 \cdot i$

Year step(i)
$1 \rightarrow 1950 \leftrightarrow 1955$
$2 \rightarrow 1955 \leftrightarrow 1960$
$3 \rightarrow 1960 \leftrightarrow 1965$
...
Year step $i = 1950 + 5 \cdot (i-1) \leftrightarrow 1950 + 5 \cdot i$

Age class(j)
$1 \rightarrow [0:5)$
$2 \rightarrow [5:10)$
$3 \rightarrow [10:15)$
…
$age\ j \rightarrow [5 \cdot (j-1):5 \cdot j)$

In this project the mass balance consists in human individuals. The model variables are the ones shown in Figure 11 :

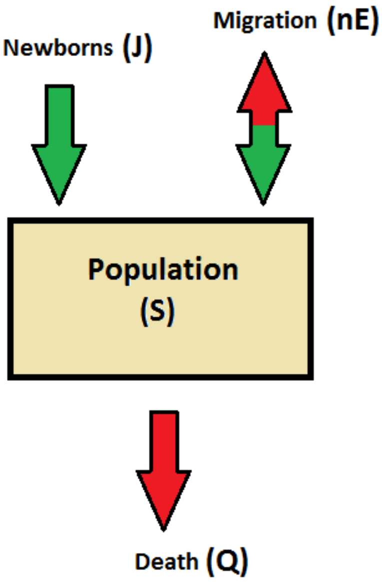


Figure 11: Model variables

Where,
$S_{i,j}$: Population in the age class j in year i.
$Q_{i,j}$: Deaths during year step i, associated to age class j.
J_i : New borns during year step i.
$NE_{i,j}$: Net migration. Immigration/emigration balance during year step i, associated to age class j.

When j refers to the first age class (hence $j = 1$) the value of $S_{i-1,j-1}$ is equal to J_i in order to match with the balance of mass. So the balance can be written as,

$$S_i = S_{i-1} - Q_i + nE_i + J_i \qquad \textbf{(3.2)}$$

The initial condition, S_0, refers to the starting population in 1950.

However, the calculations done in this report are done by age classes. Then the equation can be reformulated as,

$$S_{i,j} = S_{i-1,j-1} - Q_{i,j} + nE_{i,j} \qquad j > 1 \qquad \textbf{(3.3)}$$

The UN does not provide a net migration data set by ages so it has been needed to compute it by isolating the net migration (nE) from the equation 3.3.
Once the main equation has been introduced, it can be applied to the available UN datasets for the purpose of this project.

3.2.1 Tail Approximation

In order to find a way to distribute accurately the population of 80 years and above from 1950 until 1985 for the total population data set the exponential distribution has been used.

Exponential approximation:

The data represented in Figure 12 makes clear that the distribution of the last age classes can be approximated as an exponential distribution (note that the population density is normalized):

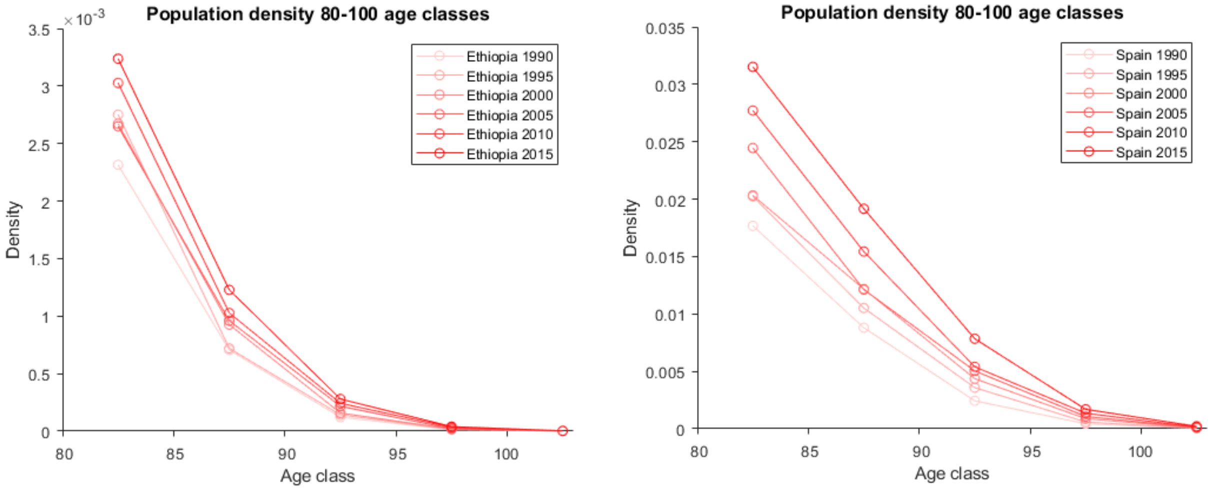


Figure 12: Ethiopia and Spain population

To use the exponential function to estimate the missing values it has been imposed that the area below the distribution must be equal to the total population of 80+ age class.

The starting point, then, is the population of 75-79 y.o.

$$f(x) = \begin{cases} S_{i,75-79} \cdot e^{-\lambda_i x} & if\ x\ \in\ \mathbb{R}_x \\ 0 & if\ x\ \not\ni\ \mathbb{R}_x \end{cases} \tag{3.4}$$

In order to obtain the rate parameter λ, an equation must be solved.

$$\int_0^{\infty} S_{i,75-79,} \cdot e^{-\lambda_i \cdot x} = S_{i,80+} \tag{3.5}$$

Where,

x: Age class starting from 0 → 75-79 y.o.

λ: Rate parameter

$S_{i,75-79}$: Total population of year *i* from 75 to 79 y.o.

$S_{i,80+}$: Total population of year *i* equal or older than 80 y.o.

Once solved, the expression for λ is (note that λ depends on the year):

$$\lambda_i = \frac{S_{i,75-79}}{S_{i,80+}} \tag{3.6}$$

Now that the rate parameter is known, in order to calculate the population for the different age classes a definite integral must be computed:

$$\int_0^1 S_{i,75-79} \cdot e^{-\lambda_i \cdot x} \,;\, \int_1^2 S_{i,75-79} \cdot e^{-\lambda_i \cdot x} \,;\, \int_2^3 S_{i,75-79} \cdot e^{-\lambda_i \cdot x} \,;\, \int_3^4 S_{i,75-79} \cdot e^{-\lambda_i \cdot x}$$

Special attention is needed for the last class, which gathers all the population that has 100 years old or more:

$$\int_4^\infty S_{i,75-79} \cdot e^{-\lambda_i \cdot x}$$

3.2.3 Net Migration

The only unknown term in equation 3.3 (balance equation) is the net migration E . As all other terms are known (from 1950 to 2015), nE can be computed by isolating the nE term from the balance:

$$nE_{i,j} = S_{i,j} - S_{i-1,j-1} + Q_{i,j} \qquad \textbf{(3.7)}$$

It is worth noting that to compute the net migration for the first age class ($nE_{i,1}$) one need to consider the newborns (J_i) as $S_{i-1,j-1}$. This boundary condition is imposed in order to match with the balance of mass.

Finally, and importantly, note that this computation of nE implies that every error in the estimates of S and Q is absorbed in the term E . So, nE in this project must be understood as nE_true + estimate errors.

In order to make the equations more simple a new variable has been introduced ΔS. ΔS matrix is the difference in terms of total population between two age classes in 5 years time. This means that it shows how a generation from a certain country/region have changed in terms of total individuals over a 5 year period. In order to achieve that it has been needed to subtract from each group of age the one of the period before. This allows to always work with the same group of population through the years. The following expression exemplifies the above explained.

$$\Delta S_{ij} = S_{i,j} - S_{i-1,j-1} \qquad \textbf{(3.8)}$$

And the calculation of migration is reformulated as,

$$nE_{i,j} = \Delta S_{i,j} + Q_{i,j} \qquad \textbf{(3.9)}$$

It's very important to remark that the balance here considers a five year period (it is a flux like mortality or newborns) not single years, as the population.

One more issue is that the data set regarding the mortality ranges from 0 to 95+, having one age class less than the population data set and so the ΔS matrix. This means that it is not possible to operate directly with the ΔS matrix.
In order to solve this problem, one easy and accurate way is to sum the last two columns of the ΔS, which represent the classes 95-100 and 100+, so they became just one age class, 95+. Another way to match the matrixes dimensions would be to estimate the deaths in the age class 100+ for the Q matrix. However, this alternative has been ruled out in order to prevent an increase on the accumulated error since it has been already used and exponential approximation in the previous calculations.

Another thing that must be understood is that a 3-year-old boy in 1950 in 1953 would be 6 y.o. However if he dies at 1953 there's a mismatch between the ΔS matrix and the mortality one. In the ΔS it would be considered in the age class from 0 to 5 y.o for the period 1950-1955. However in the mortality matrix it would appear in the age class from 5 to 10 y.o for the same period. This is incorrect however this difference is compensated and believed to be constant through the years so it does not affect in the calculations.

Once the main variables of the model have been studied and computed accordingly to the mass balance the next step is to find a way to model those variables. For this project the Storage Age Selections functions have been picked in order to fulfil that objective.

3.3 SAS functions

StorAge Selection functions (SAS) are defined to represent the way catchment storage supplies the outflows with water of different ages. In the ecohydrology field the variable that determines when the water leaves the system is the age. This means that the water that has been the longest in the catchment storage will mostly escape from it first. However, there's some paths from where the younger water can flow too and so join the old water. This means that in most cases, but not always, is the old water that leaks away from the reservoir first. Another example in which SAS functions are used is in the migratory field in order to study the migratory fluxes from birds from one place to another. It has been demonstrated that the birds that first came into one place are in most cases the ones that first leave as well.
This different approaches of SAS functions can be understood for this project purpose in several ways. The first one is assuming that the out-fluxes are the people that die and the catchment storage can be understood as the population. The second approach is to study the migration fluxes in humans.

3.3.1 Calculating the SAS functions

From the perspective of this project the SAS functions informs about the chances of dying an individual has regarding his age class. Since the SAS functions are the core of this project a brief explanation of how are they computes is given below.

The first thing to be done is normalize the data from the population data set and also from the mortality one. The reason why the data normalization is done is in order to find relations.
Then the population data set must be adapted. The population data set provides data by single years starting from 1950 whilst the mortality one provides the data in periods of 5 years each starting from 1950-1955. In order to overcome this mismatch the population data set will be considered as of 1955 for the mortality during the period 1950-1955. And the same pattern is followed for the next years and periods. To solve the problem of the age classes the population data set must be shortened by one age class adding to the 95-100 y.o one the 100+ population.

The population normalized function can be understood as $p_S(T,t)$ where T refers to the age class and t the time expressed in years.

$$p_S(T,t) \rightarrow \frac{S_{i,j}}{\sum_{j=1}^{20} S_{i,j}} = \frac{S_{i,j}}{S_i} \qquad \textbf{(3.10)}$$

And the mortality normalized function can be expressed as $P_Q(T,t)$.

$$p_Q(T,t) \rightarrow \frac{Q_{i,j}}{\sum_{j=1}^{20} Q_{i,j}} = \frac{Q_{i,j}}{Q_i} \qquad \textbf{(3.11)}$$

The next step is to calculate the aSAS function where '*a*' stands for absolute. Meaning that in the x axis will appear the age classes. The absolute StorAge Selection functions can be expressed as:

$$aSAS = \omega(T,t) = \frac{p_Q(T,t)}{p_s(T,t)} \tag{3.12}$$

This function cannot be considered a probability density function. It is a non-dimensional value that express the ratio between the mortality and the population according to the age class.

Figure 13 shows the aSAS for a random country. More countries will be introduced further in this report however now is important to differentiate the three main parts of any SAS function.

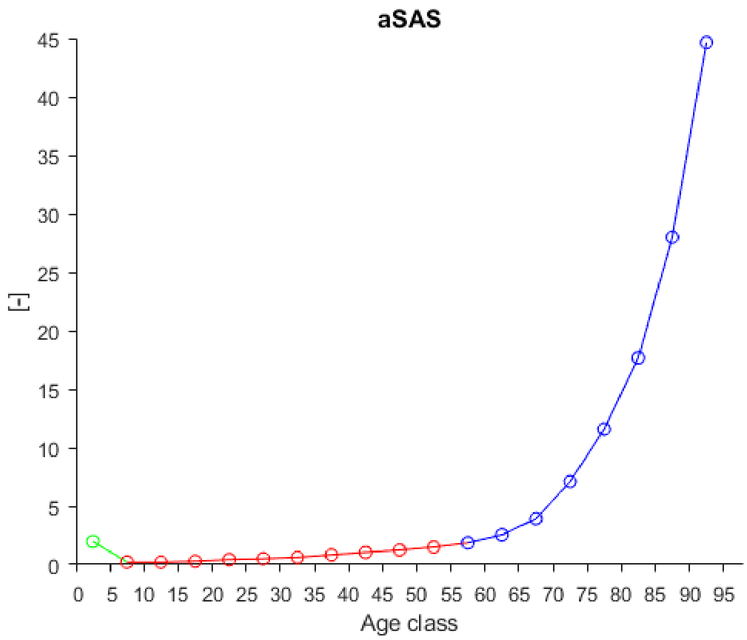


Figure 13: Main parts in a SAS function

The green segment represents the new born mortality. The red one represent the population at a low probability to die. And the blue one is associated with the mortality of the older population

The next objective is to calculate the fractional StorAge Selection function fSAS. The problem that presents the aSAS is that it is not a probability density function (PDF), so in the modelling phase it is difficult to approximate. To solve this problem a change in the x variable is needed so the SAS function can be considered a PDF.
The fSAS functions, like the aSAS, give information regarding the probability of dying. However in this case instead of giving it in terms of age class the numbers are mean to be interpreted in terms of percentage of the total population.
In order to do that the first thing to do is calculate the cumulative distribution function (CDF) of the normalized population p_s. The cumulative function will be the x axis for the fractional SAS function.

$$P_s(T,t) = \int_0^T p_s(\tau,t)d\tau \tag{3.13}$$

Where,
τ: Age class integral variable

With the probability cumulative function and the already calculated absolute StorAge Selection function it can be obtained the fractional StorAge Selection, $\omega(P_s, t)$, by resetting the x axe in the absolute one. This means that each age value must be replaced by its corresponding value from the Probability Cumulative Function.

3.4 Fertility rates

It has already been defined what the TFR is. However in this report a deeper study of the fertility trends has been carried out. TFR in a very simplified way is the sum of a rate named Age Specific Fertility Rate (ASFR). The ASFR *measures the annual number of births to women of a specified age or age group per 1,000 women in that age group*.[15] The equation in order to obtain this balance is:

$$ASFR_{i,j} = \frac{J_{i,j}}{W_{i,j}} \cdot 1000 \qquad \textbf{(3.14)}$$

Where
$W_{i,j}$: Women population in age class j during the period i.
$J_{i,j}$: Births of women in age class j during the period .

Usually there are 7 age groups considered in which a women is able to give birth, from 15 to 49 years old (15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49).

For the considered UN dataset, the TFR, assuming that the ASFR has been computed in single years (not during year steps), can be computed as follows,

$$TFR_i = \sum_{15-19}^{45-49} ASFR_{i,j} \cdot \frac{5}{1000} \qquad \textbf{(3.15)}$$

Note that the United Nations computes the number of women in the different bearing age classes by calculating the mean of the two limit years of the time period. Hence, and as an example, for the TFR in the period 1950-1955 the UN have used the mean value between the women population in 1950 ($W_{1950,j}$) and 1955 ($W_{1955,j}$) when computing the ASFR.

3.5 Model

This section introduces the model used to estimate the future population by modelling the mortality. The model has been structured in phases. Each of them presents an improvement, in terms of less dependence of real data, regarding the previous one. The variables that have been modelled are, S_i ,Q_i ,$S_{i,j}$,$Q_{i,j}$ and $aSAS_j$. All the implemented equations and models have been tested using Matlab.

The starting balance is the one that uses the SAS function to estimate the loss function Q_i. So,

$$Q_{i,j} = Q_i \cdot p_{Q,i,j} = Q_i \cdot aSAS_j \cdot p_{S_{i,j}} = Q_i \cdot aSAS_j \cdot \frac{S_{ij}}{S_i} \qquad \textbf{(3.16)}$$

It has been computed the mean from the absolute SAS function. This means that for every age class it has been taken the value of all the years available and computed the average so the SAS function does not depend in the year. It just depends on the age class.

The S_i and Q_i values represent the total population and the total mortality in the year i respectively.

Once the mortality has been estimated the main balance of the project can be used to obtain the population.

If one replaces the $Q_{i,j}$ value from equation 3.3 with the equation 3.16, $S_{i,j}$ can be expressed as:

$$S_{i,j} = S_{i-1,j} - Q_i \cdot aSAS_j \cdot \frac{S_{ij}}{S_i} + nE_{i,j} \qquad j > 1 \tag{3.17}$$

And solving the equation for $S_{i,j}$;

$$S_{i,j} = \frac{S_{i-1,j-1} + nE}{1 + \frac{Q_i}{S_i} \cdot aSAS_j} \qquad j > 1 \tag{3.18}$$

Again when j referts to the first age class ($j = 1$) the term, $S_{i-1,j-1}$, in the equation has been adopted as the newborns J_i.

Herewith are presented the different steps carried out in order to obtain the latest version of the model in which the use of real data is minimum. It is divided by cases. The highlighted in red parameters are the one being modelled in each case.

Case 1 (M1)

This case proves that using the mean of the aSAS instead of the single year value does not affect much the result. The mean aSAS value has been computed for each age class. It also computes the S_i parameter by using the previous projected time period assuming that the change will be minimum. So the two parameters from the equation 3.18 that have been modelled/computed are.

- aSAS mean value.
- S_i comes from the model using equation 3.2 (mass balance)

$$S_{i,j} = \frac{S_{i-1,j-1} + nE}{1 + \frac{Q_i}{S_i} \cdot aSAS_j}$$

Case 2 (M2)

In this case the number of newborns have been computed by using the Age Specific Fertility Rate. The fertility is modelled by computing the newborns for the next year step using the ASFR.
The ASFR has been obtained using the newborns and female population of the previous year step according to the equation 3.14. For the first year, 1955, the female population was the one from 1950 and the newborns were from the period 1950-1955.

Case 3 (M3)

In this third case the loss function is being modelled. In order to understand how this function behaves and see if it could be used through the SAS function the first step has been to plot the ratio between the total mortality Q_i and the total population S_i.

$$\frac{Q_i}{S_i} = Q\_S_i \tag{3.19}$$

So, by replacing the above ratio in the equation 3.18 the balance becomes,

$$S_{i,j} = \frac{S_{i-1,j-1} + nE}{1 + Q\_S \cdot aSAS_j}$$

Case 4 (M4)

This is the final and last modification in the model. Basically it computes the mean aSAS function but in a different domain in order to later obtain the different single year step SAS function. This process can either be achieved by using the aSAS function or by working with the cumulative SAS for later differentiating it and obtaining back again the aSAS. This second approach is more accurate for numerical reasons. For example, when using the discretized aSAS function, the area below the curve is not equal to 1. However, using the cumulative one it is. Detailed information regarding the domain transformation can be found in Appendix 1. Figure 14 shows in a very simplified way the steps followed to obtain the aSAS function from a different domain.

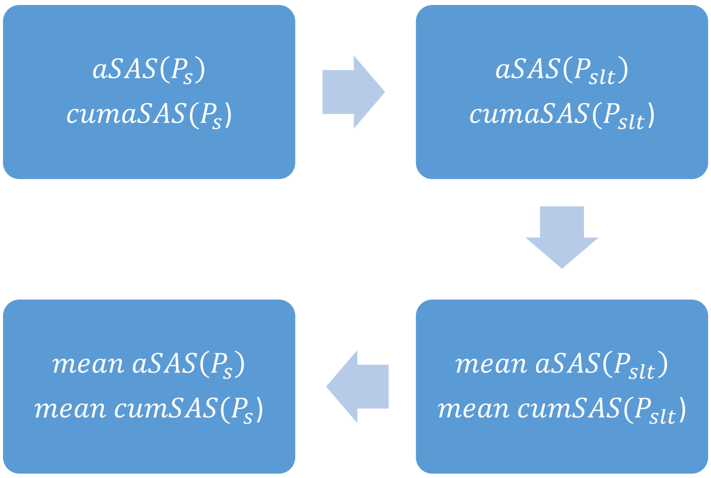


Figure 14: Procedure followed to obtain aSAS from a different domain

4. Results

4.1 Data analysis

4.1.1 Fertility

Fertility by all means is strictly related by the number of women in age of bearing. Starting from this point it is interesting to take a look at three different plots. The first one is the ASFR, showing how the likelihood of having a child varies by age. Figure 15 shows different countries ASFR over different years.

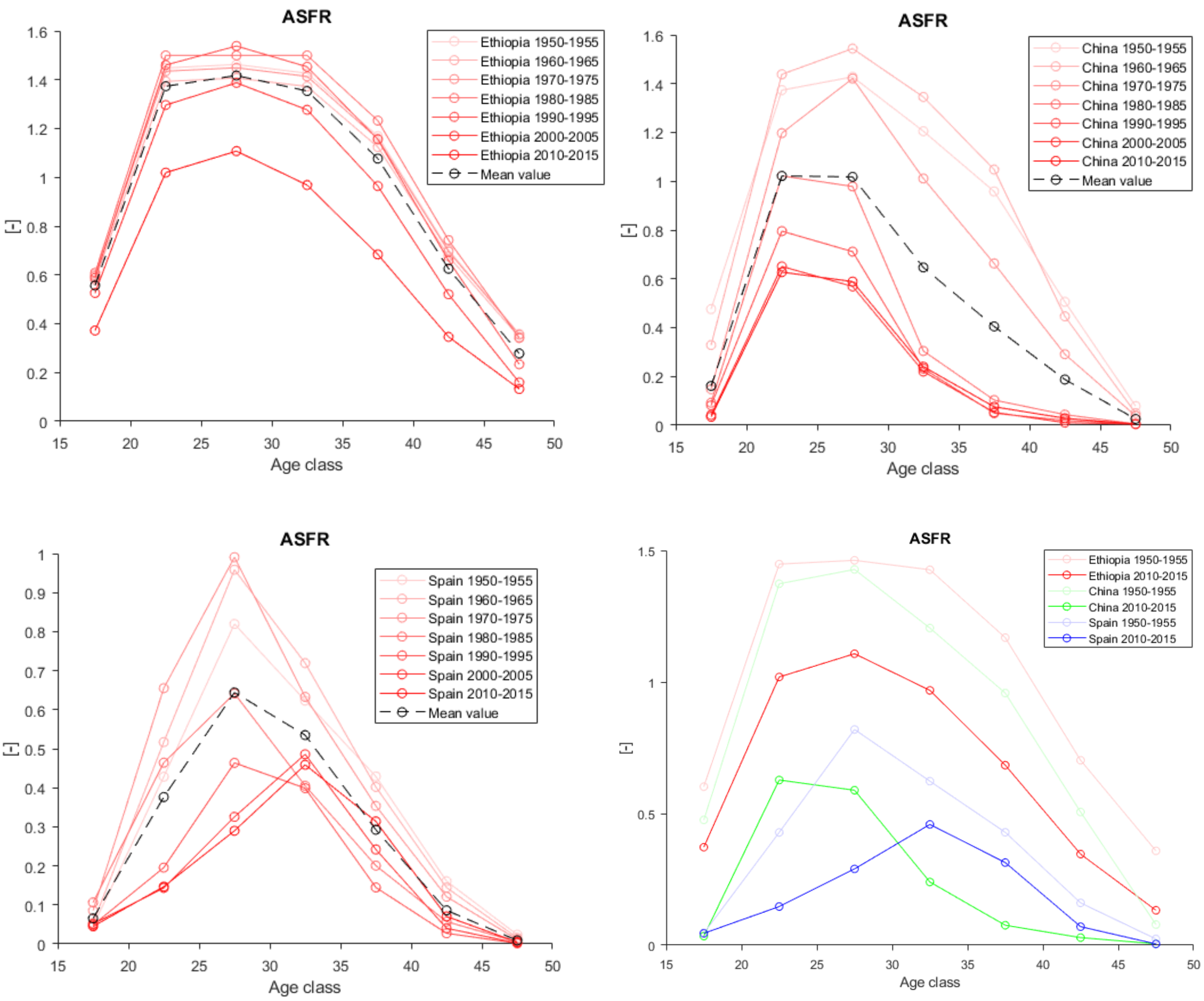


Figure 15: Ethiopia, China and Spain ASFR

The first one, Ethiopia, confirms that less developed countries tend to have a higher number of births per women compared with more developed ones. In Ethiopia during the period 2010-2015 the number of births given for a women of age between 25-30 was slightly over 1 meaning that the probability to give birth was more than 100% at that age class. In the other hand, in Spain, for the same period, a women of the same age class would have around 30% of probability to give birth. Another big difference between Spain and Ethiopia apart from the probability of giving birth is the ASFR distribution regarding the age class. Spain over time presents a transition not only in terms of the number of births but also in the age class of those births. The peak has shifted from the age class 25-30 to the next one, 30-35.

On the contrary, in Ethiopia the peak did decrease but it did not move to another age class. Last but not least China has experienced also a decrease in the number of births but also the peak age class has shifted but to a younger one. A reason to this might be that many women still believe that the younger you are the healthier the child is.

The last plot reflects the different trends an ASFR can present. In all the countries plotted the number of births per women, no matter the age class, has decreased over time. Apart from that, another interesting conclusion from the plots is that the age class with more births can either remain the same (Ethiopia) or shift to another age class over the years (Spain or China). Although the case of China is not common most of the developed countries have shifted their peak to the right meaning that women nowadays tend to give birth at an older age than 30 or 50 years ago.

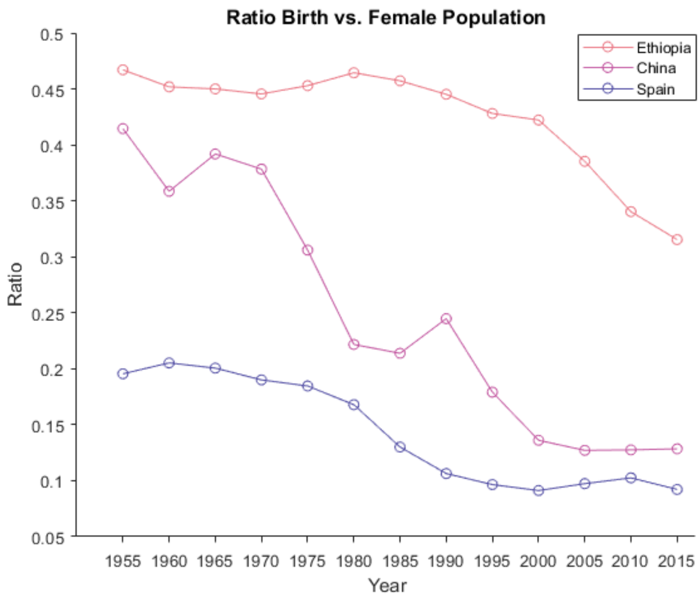


Figure 16: Ratio birth versus female population

Figure 16 corroborates what have been said previously. That the trend over the years is to decrease the number of births per woman. However, it's interesting to observe that Ethiopia until 1995 did not show any significant sign of decreasing the total number of births. Starting from 2000 it has presented a continuous and strong decrease. This means the beginning of a transition into the developed country trend since the most common causes of decreasing child birth are related to first world conditions [16]. Some factors that contribute into decrease the number of births in less developed countries are: Bigger access to contraceptives, higher levels of female education, change in the social structure and religious beliefs, economic prosperity…

It's also interesting to observe that the ratio in China rose in the period 1985-1990, reversing a long-term trend toward low fertility. There are two main reasons that explain this situation:

- In the period 1960-1965 there was a baby boom and in 1985-1990 they reached the reproductive age.
- The introduction of the new marriage law in 1980 that allowed women to marry when they're 20[17].
- The one child policy was established in 1979 and for the first years the policy experienced certain periods of relaxation.

Figure 17 reflects how the number of births changes over time.

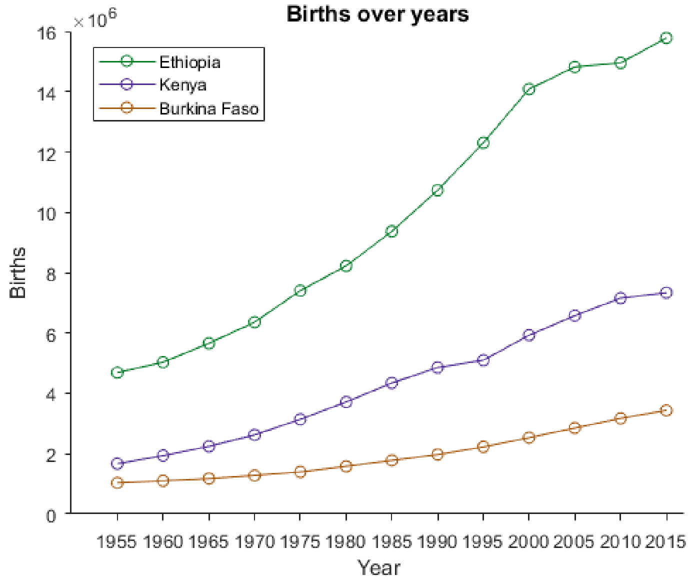


Figure 17: Number of births over time

An increase in the number of birth does not necessary mean an increase in the ASFR ratio since this ratio also depends in the female population. Then, these three countries have experienced an increase in their female population over time and Figure 18 corroborates so.

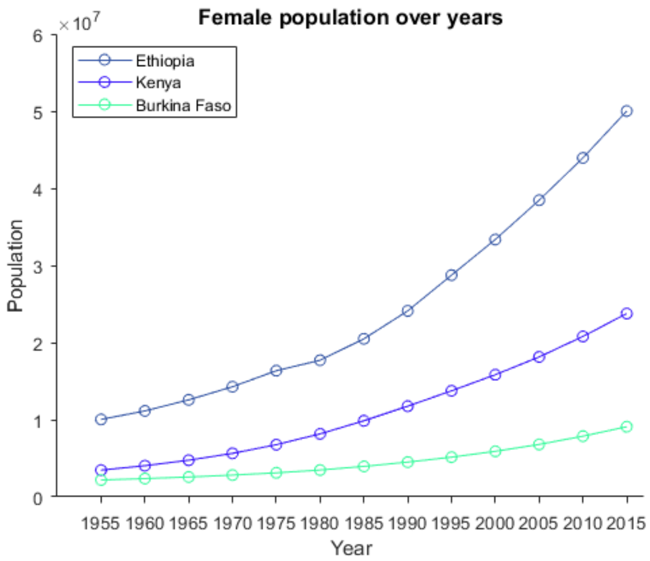


Figure 18: Female population over time

4.1.2 Mortality

Death is affected by many different factors. It may be biological, environmental, physiological… However, from the demographic point of view, mortality is related to age and sex of individuals. Mortality trends has been studied by understanding several different plots. The first one is the ratio between mortality and population (See Figure 19).

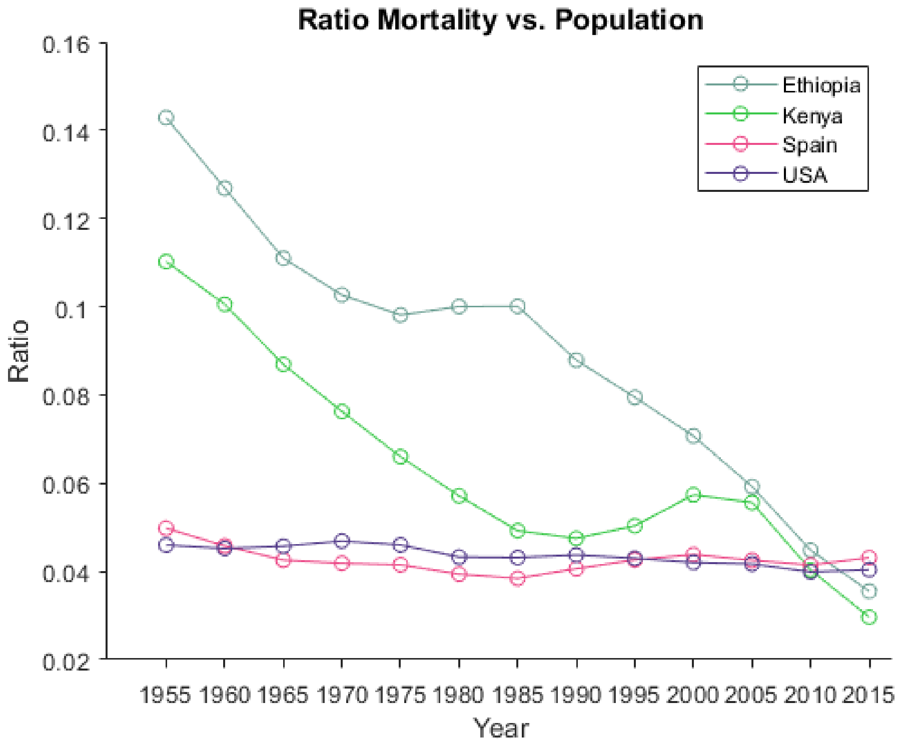


Figure 19: Ratio mortality versus population

In this plot the two main trends are presented. The first one reflects the situation in less developed countries whereas the other one, represented by USA and Spain, represents the more developed countries situation. The situation in countries such as Ethiopia or Kenya, where the ratio presents a continuous decrease over time, is a consequence of three different scenarios. The first one is due to the economic and social development and as a result a better medical care system is available or clothing, transportation, water supply has been improved. The second possibility is due to social policy measures and the third one is that technical changes have reduced the cost of health equipment. [18]

Since 1960 mortality reductions worldwide have been associated with two factors: the prevention of death caused by low birth weight in infants (in the next plots this situation will be discussed) and a better understanding of the cardiovascular disease in the aged population.[19]

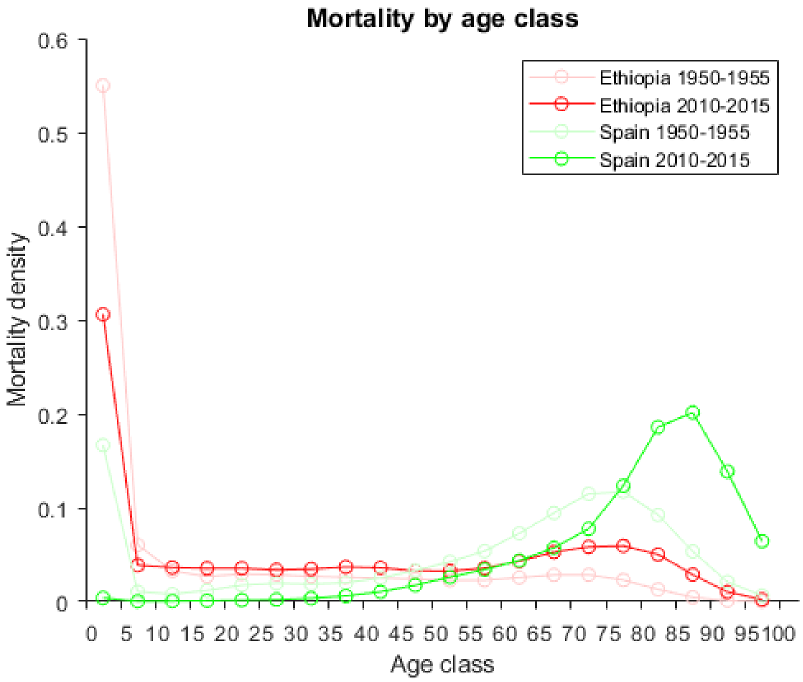


Figure 20: Mortality by age class

Figure 20 shows the mortality by age class. Here it is important to differentiate the curve between less developed countries such as Ethiopia and more developed countries like Spain. In regards of the initial peak of mortality it is due to infant mortality and will be discussed next. What is interesting to observe in this plot is that in Kenya the population is younger than in Spain. This can be concluded by looking the peak in the last ages of Spain and also because the younger mortality in Spain is lower than in Ethiopia. The plot vertical axis is mortality density, meaning that informs in which age class does people most likely die by giving the percentage in terms of total mortality. The higher is this value in the last age classes the older is the population.

Figure 21 corroborates what implicitly tells Figure 20. The median age of the population in less developed countries is much lower than in developed countries.

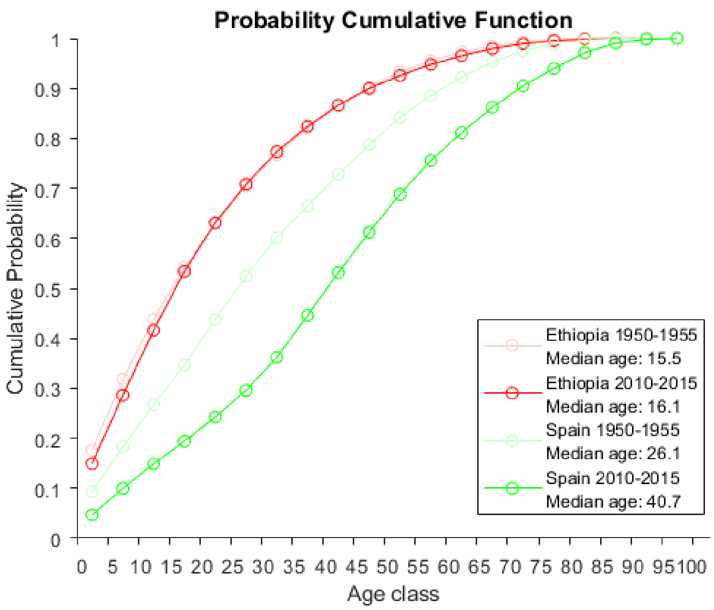


Figure 21: Probability cumulative function of Ethiopia and Spain

In Ethiopia the median age of the population have not increased really much during the past 60 years. In Spain, on the contrary, it has increased almost a 100% meaning that the population nowadays is older than 60 years ago.

Figure 22 informs about the differences in terms of newborn mortality. The peak mentioned before in the first age class is due to new born/infant mortality. Over the years both Ethiopia and Spain have decreased this levels and this trend is reflected in the plot.

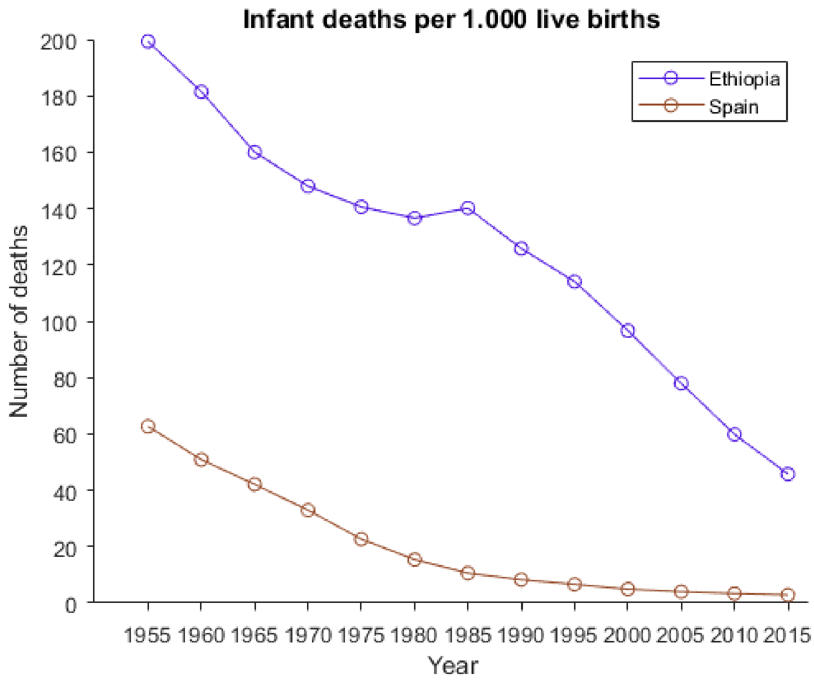


Figure 22: New born mortality in Ethiopia and Spain

The study of mortality has allowed us to better understand the nowadays trends towards this variable and forecast different possible future outcomes. It can be observed that since 1955 the infant mortality has dropped drastically either in developed regions and less developed ones meaning that both of them are entered either the phase II or III of the demographic transition theory. Also this decrease in the infant mortality has turned out in an increase in the population especially notable in the less developed regions in which there have also been a decrease in the population mortality. And last but not least the number of births over the years seemed to increase in the less developed regions contributing in the increase of its population.

4.1.3 Migration

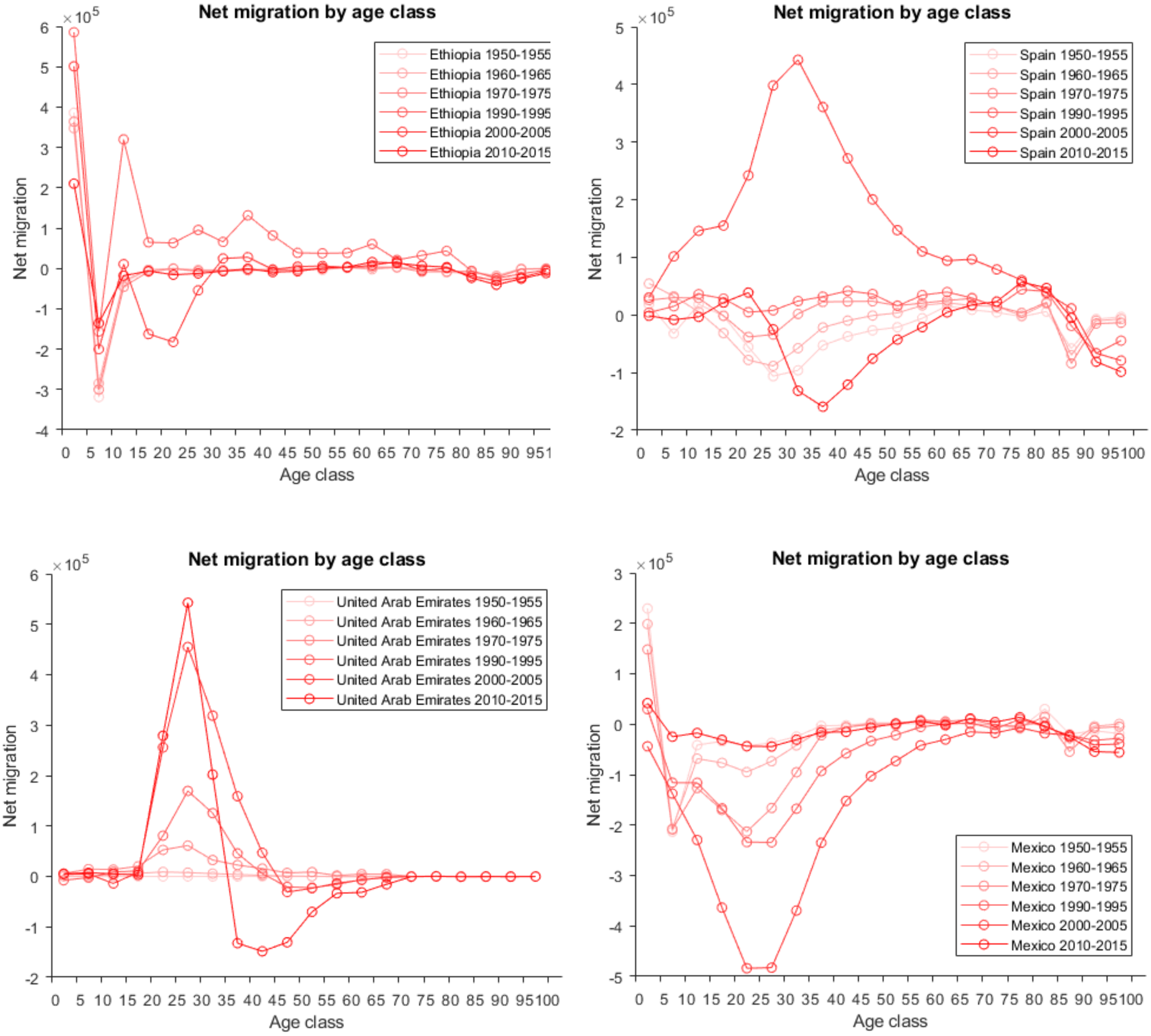


Figure 23: Net migration in Ethiopia, Spain, UAE and Mexico

Figure 23 plots several countries net migrations fluxes. It shows a clear tendency. The more developed countries tend to receive more immigrants whereas in the less developed ones the emigration rates are higher.

The first group of age should presents an almost 0 rate of migration. This anomalous value is due to the lack of accuracy in the number of new borns and the population in the age class of 0-4 years old. The lack of data is especially remarkable for the 1950 to 1980 years and is more accentuated for the less developed countries.

Mexico is the country with most emigration worldwide, reason why there is such a high peak between the ages 20 and 30. On the other hand, there's United Arab Emirates, the country with most immigration. In its plot it is clear that the immigration skyrockets up to 600000 individuals above the emigration. It is a positive value since the net migration balance is done as follows,

$$nE_{i,j} = I_{i,j} - E_{i,j} \qquad \textbf{(4.1)}$$

Trends in migration are difficult to predict and subject to environmental events or wars. Migration movements are also deeply affected by countries economies and legislations. This is one of the reasons why in less developed countries there's a tendency in emigrate much higher than in more developed ones. It is also interesting to take a look at the world migration presented in Figure 24.

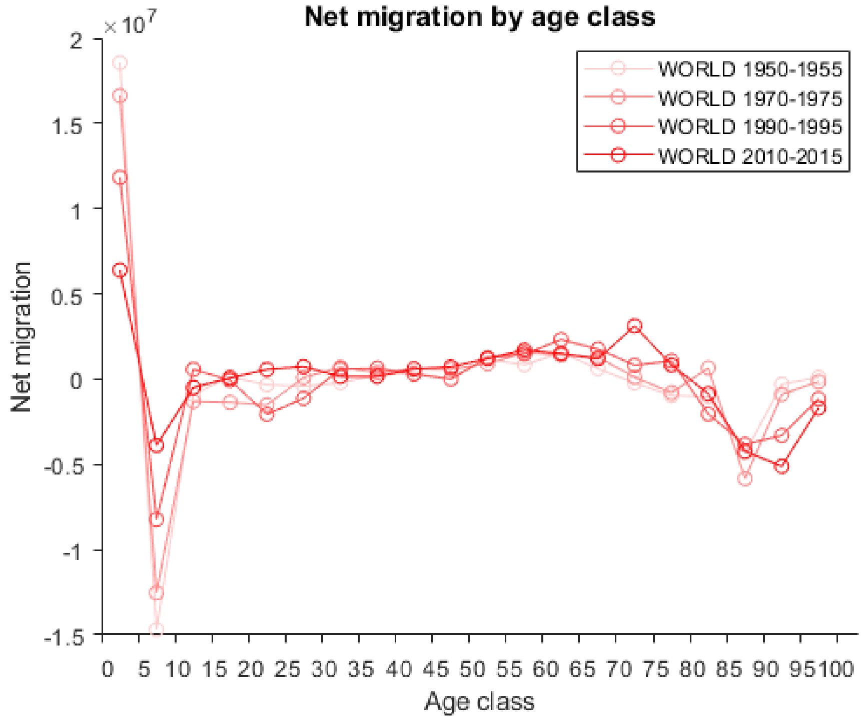


Figure 24: World migration

The data for the first age class should be 0 however it is not. World migration according to our balance should be 0 since people cannot migrate from one world to another. The reason why it does not add to zero is basically a consequence of the following actions:

- The assumption of considering the number of newborns while computing the ΔS matrix.
- Estimating the population of the last age classes using an exponential distribution.
- The UN nations computed the migration by an iterative process until they achieved 0 migration.
- Net migration has been computed from equation 3.3 meaning that it is subject to the other parameters accuracy.

However the error is not significant since the most of the age classes present a value around 0 and if one sums up all the classes the total balance is close to 0.

4.1.4 SAS function

This section explores and gives an insightful view of how SAS functions can be useful in the modeling of population dynamics. The SAS functions express the probability one have, as an individual, to leave the system (die). The SAS function are presented in two different formats, one using the absolute value and the other one using the CDF of the population.

The Figure25 and Figure 26 show the aSAS curve of Spain and Ethiopia.

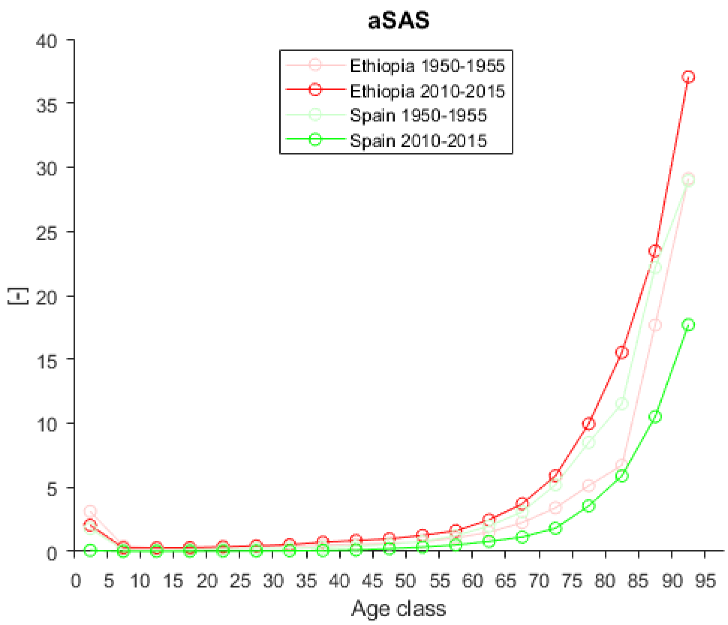


Figure 25: aSAS function

Spain shows the normal trend in any developed country in which over the years the infant mortality presented by the peak in the first age class have notably decreased and also the curve maintains the flat tendency during more age classes meaning that population tend to live longer.

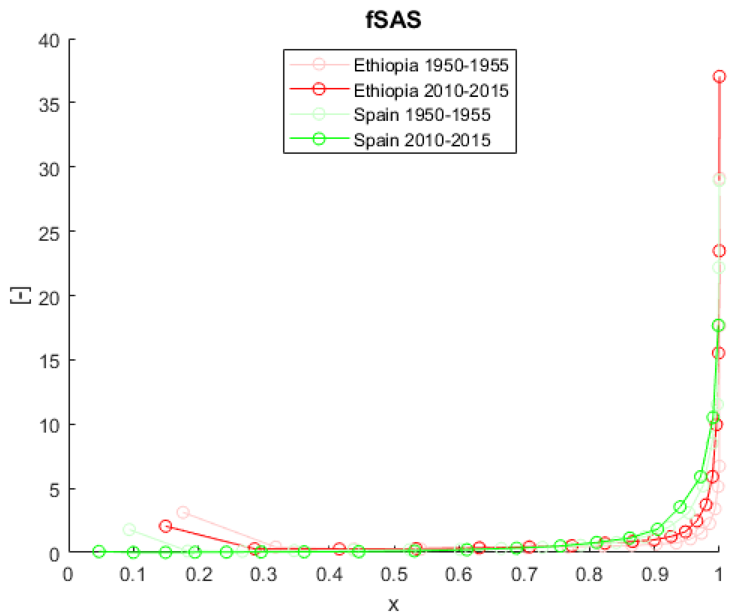


Figure 26: fSAS function

From Figure 26, in 2010-2015 Ethiopia's first age class (0-5) represented almost 15% of the total mortality whereas in Spain in 2010-2015 it only represented the 5%. This is a low value for a country however looking at it in a retrospective way it can be concluded that over the years the tendency is to reduce this value. In 1950-1955 the value in Ethiopia almost represented the 20% whereas in Spain for the same period it was 10%.

In the fSAS, Spain curve is below Ethiopia's. This means that in Ethiopia the population tend to die younger and so the percentage that the last age classes represent in terms of mortality is smaller than the Spain ones.

4.2 Model ability

Figure 27 shows dashed in black the population of different countries and in red the population predicted using the developed model (M4).

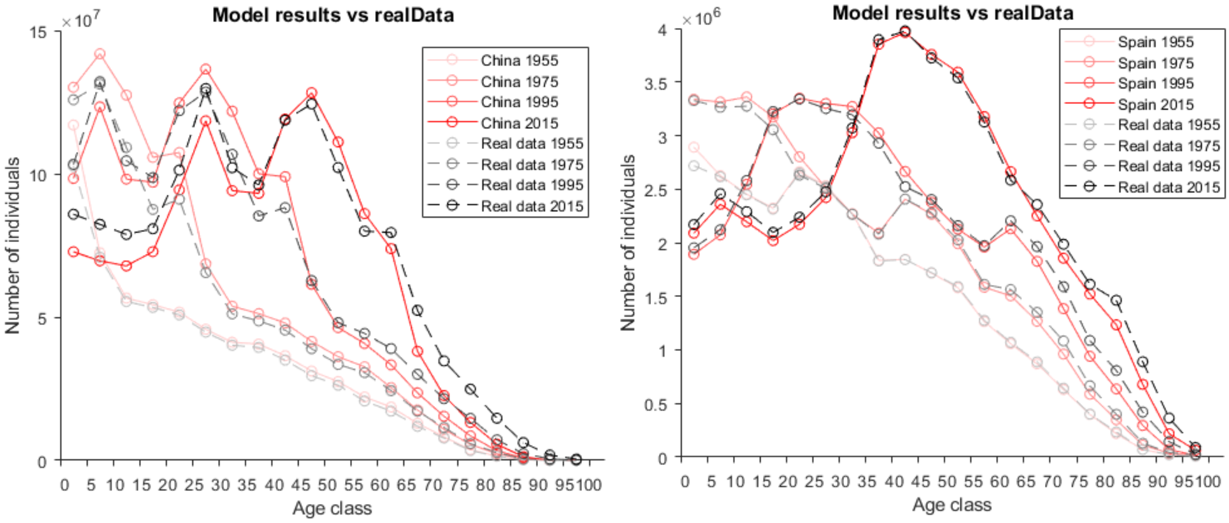


Figure 27: Modelled population vs UN data

By analyzing the previous figure it can be concluded that the model in overall works well. The variables that have been modeled are the population (S_i), the $aSAS$ and the cumulative $aSAS$. However, the error accumulates over the age classes. That means that an anomaly in the first age classes will affect the performance of the model for the older ones.

It is also interesting to observe the increase over the years of the number of births for the less developed countries. And also to pay attention to the sudden rises or decreases in the population due to the migration effect.

5. Discussion

5.1 Levels of model complexity

Herewith are discussed the different phases carried out in order to obtain the latest version of the model.

The next table sums up all the assumptions for the different models:

Case	Parameters modelled
M1	$aSAS$ & S_i
M2	$aSAS$ & S_i & J_i
M3	$aSAS$ & S_i & Q_S_i
M4	$aSAS$ & S_i & cumSAS

Table 2: Parameters modelled

In M1 the focus was on the SAS function. In it the possibility of using the mean aSAS value had been tested. Figure 28 shows the low deviation between the mean aSAS and single years aSAS values for the most points in the function. This difference may seem to increase when reaching the more advanced age classes, however the impact of this deviation in terms of error in the model is not significant since the total amount of people in this age classes tend to be very smaller in comparison with the other ones.

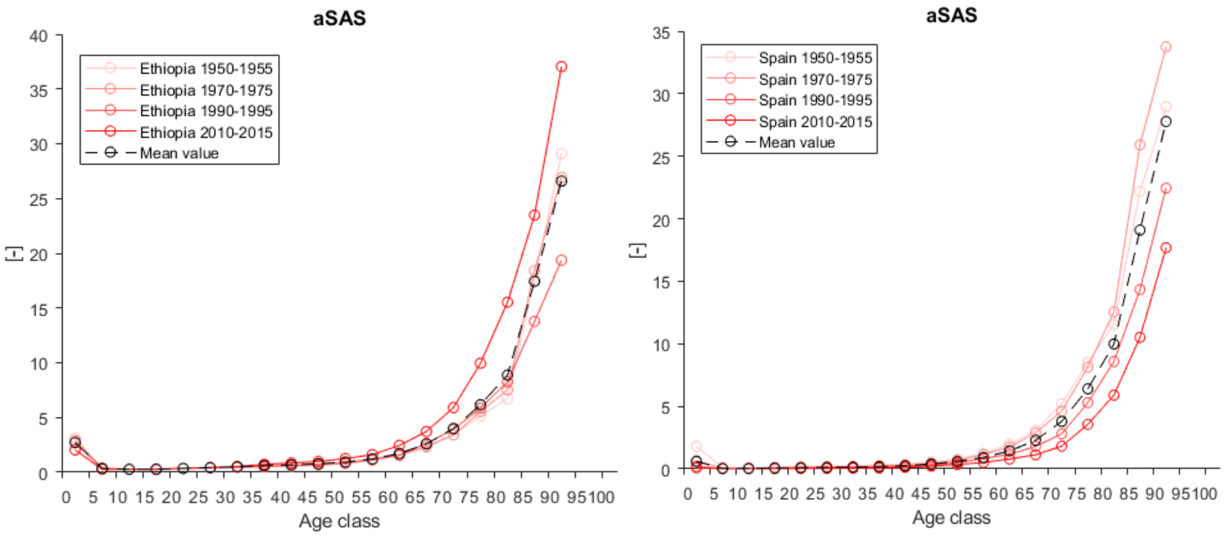


Figure 28: aSAS Ethiopia and Spain

In M3 the loss function is being modelled. In order to understand how this function behaves and see if it can be used through the SAS function the ratio between the total mortality Q_i and the total population S_i has been plotted in Figure 29.

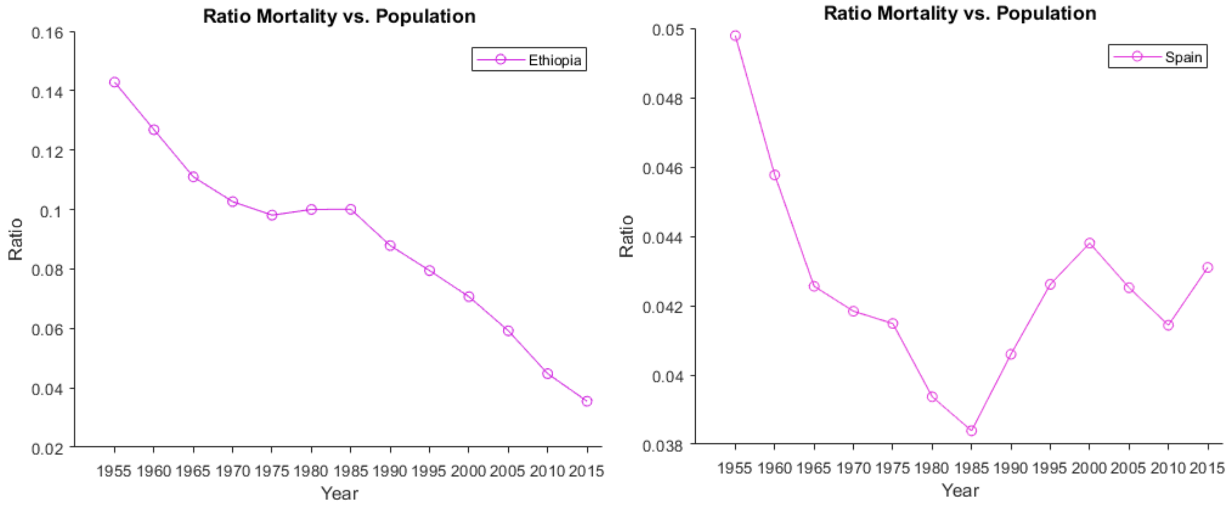


Figure 29: Ratio Mortality vs Population Ethiopia and Spain

As the plots show in some countries this value oscillate whilst in others decrease over time. Since there is no common pattern over the countries this ratio has not been considered for the M4 case of the model.

The following Figure 30 Figure 31 Figure 32 and Figure 33 show the performance of the model for Spain and Ethiopia in 1955 and in 2015.

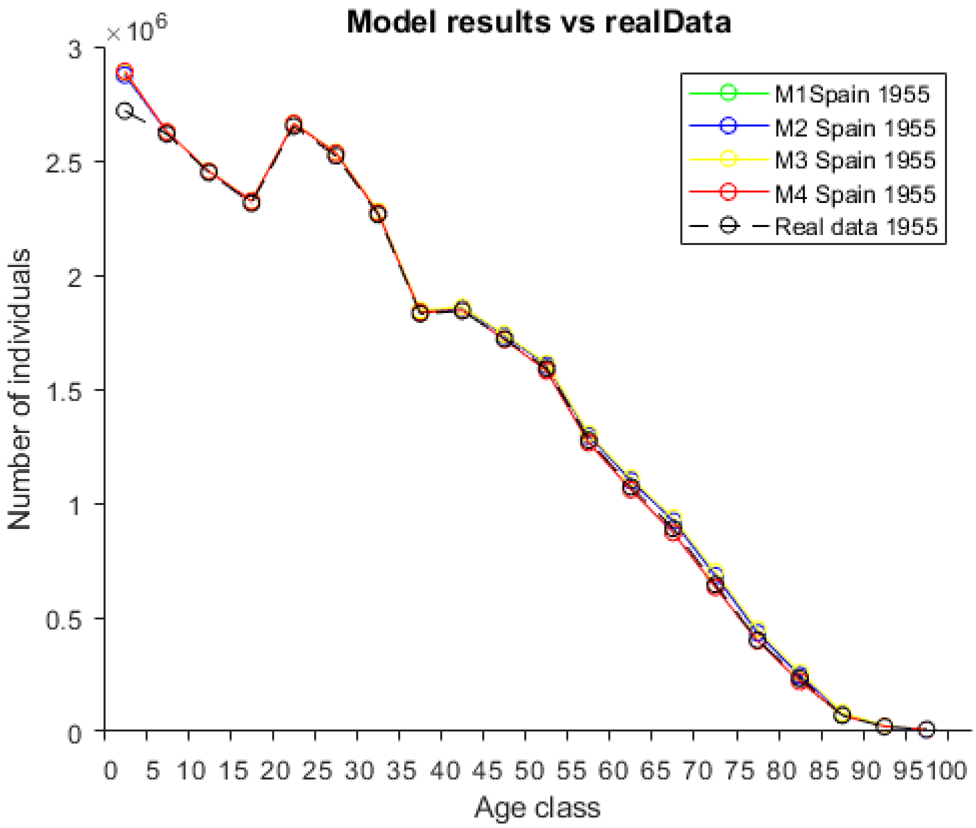


Figure 30: All cases from the model and real data of Spain in 1955

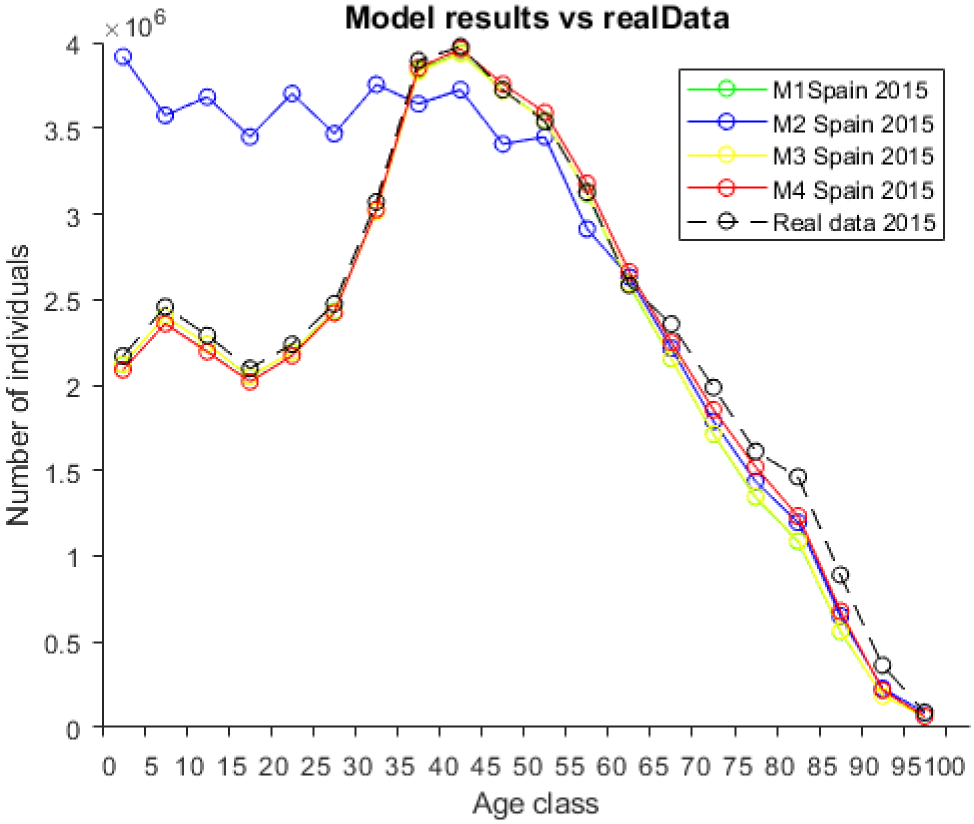


Figure 32: All cases from the model and real data of Spain in 2015

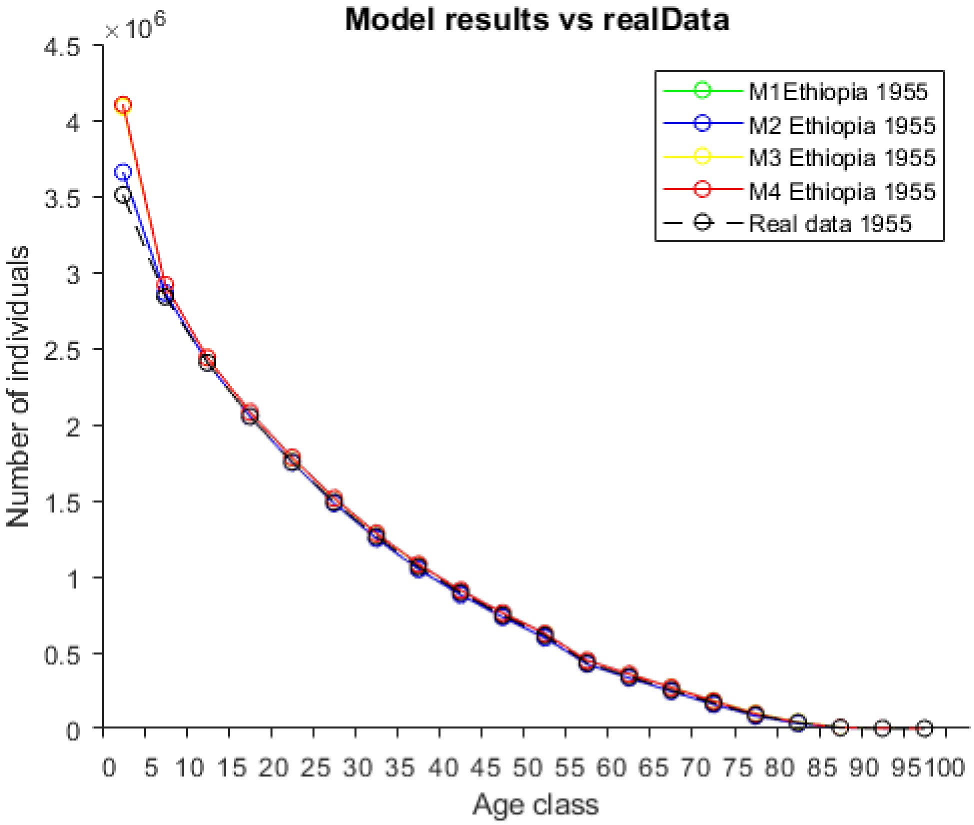


Figure 31: All cases from the model and real data of Ethiopia in 1955

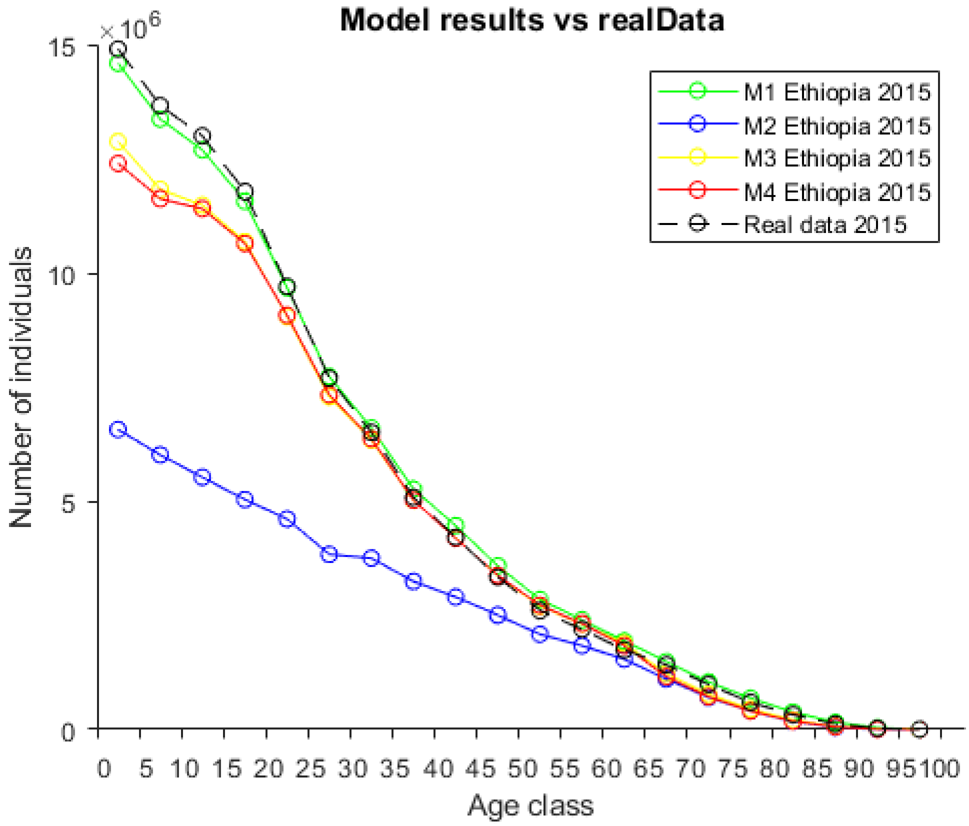


Figure 33: All cases from the model and real data of Ethiopia in 2015

From the previous graphs is understandable that the more parameters you model the bigger is the difference between modelled data and real data. If one focusses in M1 it's almost the same curve as the real data from UN. This is due to the fact that in it only the population has been modelled. The weight of the mean aSAS value does not affect the results of the model severely. Not even the last age classes where the differences between the single year aSAS functions and the mean one present the higher differences are effected. Since the way the model computes the new population is by updating the previous period population via equation 3.19 the error accumulates. This is why in 2015 the curve is a little bit less accurate than the 1955 one.

M2 case presents an error that cannot be permitted. It performs well for the early years starting from 1955. However over time and due to the fact that ASFR values and women population as a percentage of the total population change is not taken into account in M2 the accuracy of the model decreases. In order to obtain a more reasonable result for this model the ASFR should be computed taking into account more variables and not just the newborns and women population of the previous year. Also the women percentage over years change so considering that 50% of the population is female is not enough.

M3 case works well according to the previous plots. In this plot the ratio Q_S has been considered constant over time. The effect of this is negligible as one can observe.

M4 case, the last one and the one from which the probability density functions for the predictions have been obtained presents a decent accuracy. The results are very similar as the ones from M3 however here the aSAS function has been transformed into a different domain and instead of using the mean value for each year, it has been computed its own aSAS. However the Q_S ratio has not been considered in this case.

5.2 Modeling the SAS function

One step further has been carried out in this project and two different probability distributions, the beta and the power-law functions, have been tested in order to give the interpolated data a mathematical form. Both distributions are well suited for this project purpose in the sense that they are defined from 0 to 1 and also they present a shape similar to the cumulative SAS functions one. The shape of this distributions have been computed using the cumulative SAS functions of the High-Income (HI), the Middle-Income(MI) and the Low-Income(LI) countries considered by the UN. The countries that conform this thee groups are defined depending on his gross national income per capita (GNI). Those countries whose GNI is above $12,615 are considered High-Income countries. Countries with less than $1,035 GNI per capita are classified as Low-Income countries and countries whose GNI is in between $1,035 and $12,615 are considered Middle-Income countries. [20] [21] This classification depending on the income has been chosen due to the fact that the cumulative SAS function of all the UN countries seemed to fall in either one of the three previous domains. Figure 34, Figure 35 and Figure 36 reflect the situation.

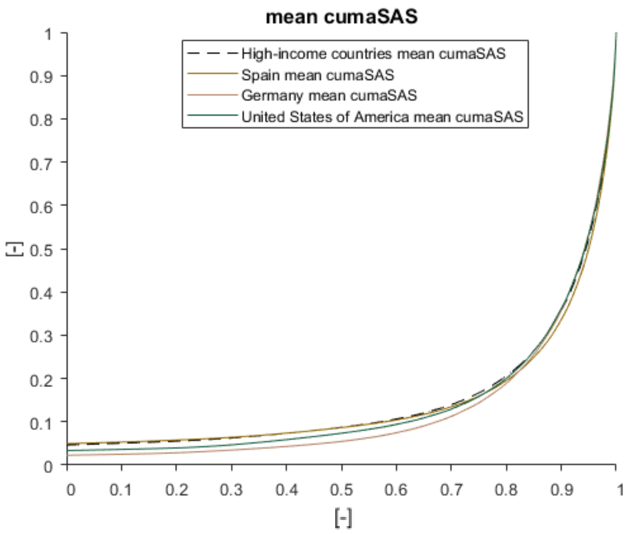


Figure 34:High income countries cumulative SAS

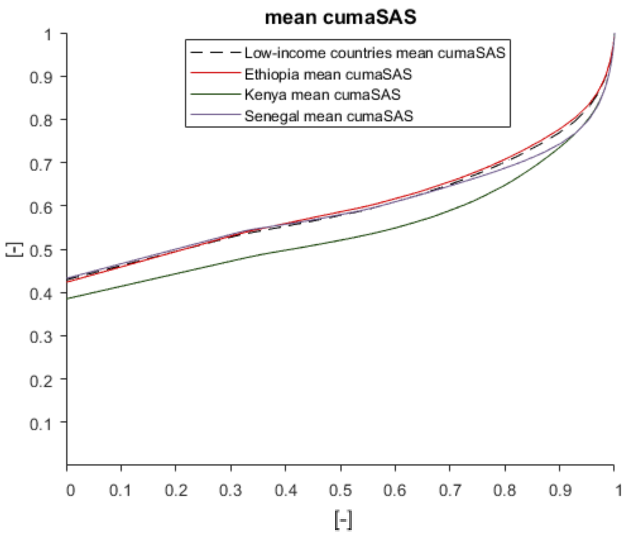


Figure 35: High income countries cumulative SAS

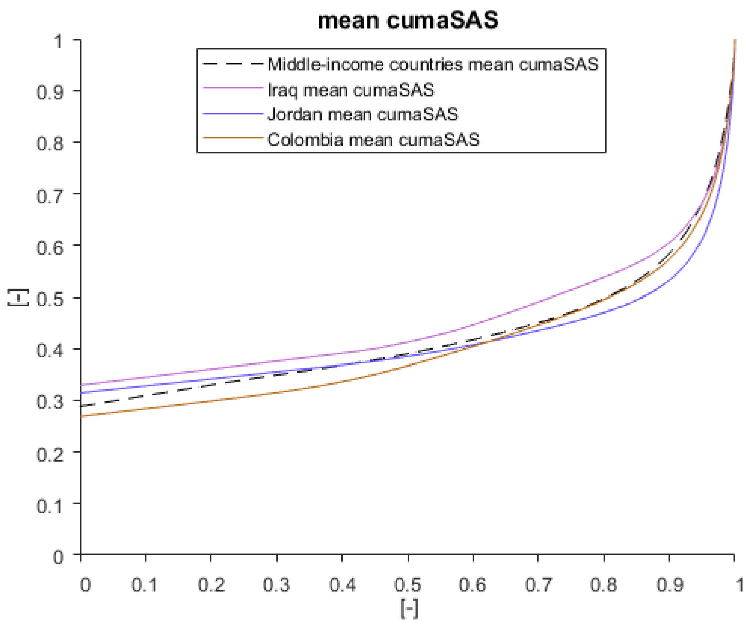


Figure 36: Middle income countries cumulative SAS

The Less developed countries present a very different cumaSAS function than the High-income or the Middle income ones. At the same time Middle income countries also present a different cumaSAS compared to the one of the High income ones.

In the last two figures countries like Ethiopia, Kenya, Senegal considered LIs present a very different shape for the cumaSAS function than HIs like Spain, USA or Germany. This difference is mainly due to the fact that HIs present a high newborn mortality reason why they start in a higher level than the HIs. This is why three different situations H-I, L-I and M-I were needed to be contemplated.

Beta

Beta distribution is defined in terms of a parameter theta, it's a continuous distribution, this is a function of two parameters α and β. The interesting thing about beta distribution is that is only defined for theta from 0 to 1. That's why it is commonly used when talking about probabilities and it's needed to specific what is the prior knowledge about the probability of something occurring. In the range of beliefs it can be specified quite a large range of beliefs by changing the parameters alpha and beta.

The most common form to present this distribution is as follows,

$$P(\theta|\alpha,\beta) = \frac{\theta^{\alpha-1} \cdot (1-\theta)^{\beta-1}}{\mathrm{B}(\alpha,\beta)} \propto \theta^{\alpha-1} \cdot (1-\theta)^{\beta-1} \rightarrow \theta \in [0,1] \qquad \textbf{(5.1)}$$

Figure 37 tries to give a better understanding of the different possibilities and shapes beta distribution can offer.

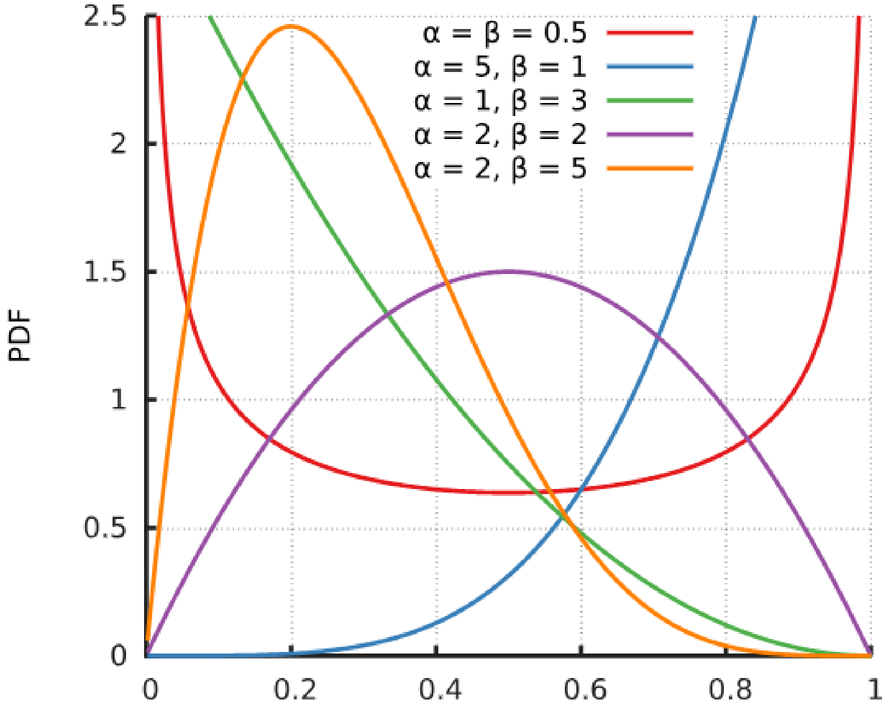


Figure 37: Beta distribution PDF

By using a trial and error algorithm, the best alfa and beta values for each of the three different regions mentioned before can be obtained.

	Alfa	Beta
High-Income countries	1.6	0.24
Middle-Income countries	0.19	0.13
Low-Income countries	0.16	0.23

Table 3: Alfa and Beta values for High/Middle and Low income countries beta distribution.

Figure 38 shows both the mean cumaSAS function and beta distribution for High-income countries.

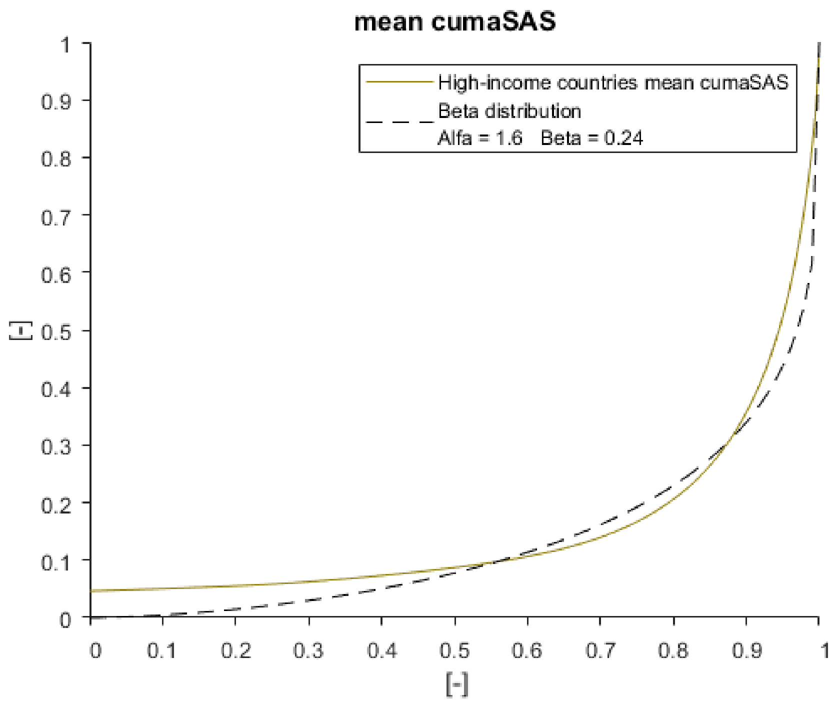


Figure 38: Mean cumulative SAS function for high-income countries and Beta distribution

Since the results were not as good as expected this function was discarded for this project. The most relevant issue that led to not consider this distribution was the deviation in the last part of the curve. The last part of the curve is the most critical part since is where most of the data is concentrated. Another important thing was the beginning, the beta distribution could not provide the required shape since it starts in 0 and then gradually increases. The problem is that this increase happens in a very slow way and should be faster.

Power-law

In statistics, a power law is a functional relationship between two quantities, where a relative change in one quantity results in a proportional relative change in the other quantity, independently of the initial size of those quantities: one quantity varies as a power of another. In order to work with the power-law function has been necessary to integrate them in the right domain, from 0 to 1. This way they can be considered as CDF's,

$$\int_0^{P_s} a \cdot (1-x)^{a-1} dx = 1 - (1-P_s)^a \qquad \textbf{(5.2)}$$

After some try outs the power-law distribution have proved to offer better results than the beta distribution. However, in order to achieve a better accuracy a weighted function has been implemented. This function comprises to power-law functions one in order to model the left part of the cumaSAS and another one to model the right part.

So,

$$F_l = 1 - (1-P_s)^{a_l} \qquad \textbf{(5.3)}$$
$$F_r = P_s{}^{a_r} \qquad \textbf{(5.4)}$$

And the coupled one,

$$y = (1-w) \cdot F_l + w \cdot F_r \qquad \textbf{(5.5)}$$

By applying trial and error methodology the best power-low values for the weighted distribution are presented in Table 4 and Figure 39, Figure 40 and Figure 41 are the beta distributions using those values:

	a_l	a_r	w
High-Income countries	0.28	13.44	0.51
Middle-Income countries	34.65	7.76	0.63
Low-Income countries	1843.10	3.50	0.49

Table 4: a_l, a_r and w values for High/Middle and Low income countries

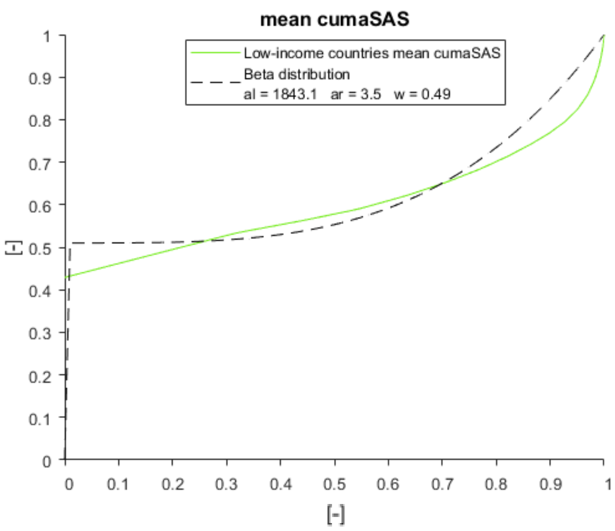


Figure 39: Low income countries beta distribution

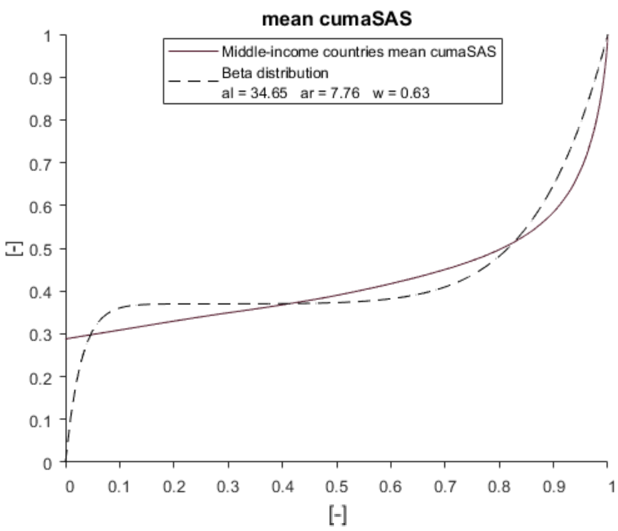


Figure 40: Middle income countries beta distribution

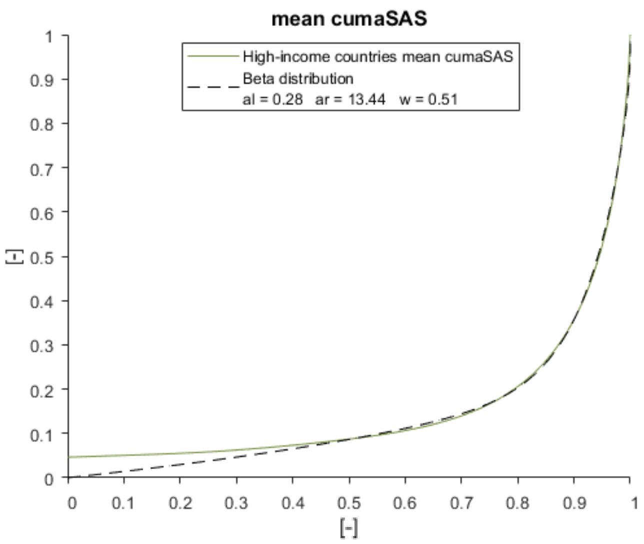


Figure 41: High income countries beta distribution

The weighted distribution for the HI countries gives outstanding results. Regarding the MI and LI countries the weighted distribution provides a better performance than the beta one.

In order to better understand how the weight affects the two functions the following plot, for the High-Income countries, intends to give a better outlook of it.

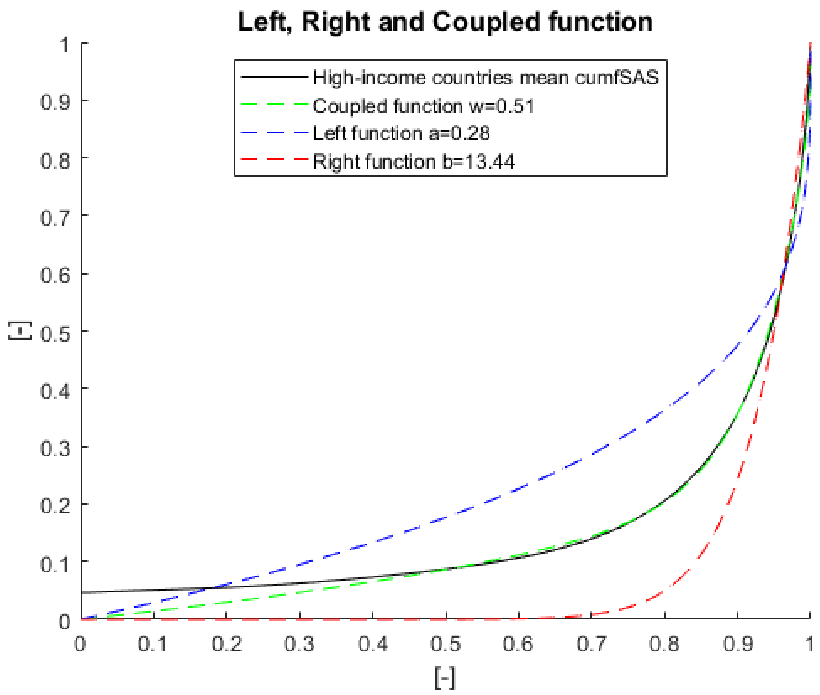


Figure 42: Left, Right and coupled function comparison

The plot reflects the contribution towards the coupled function from the left function and the right function.

The following Figure 43 and Figure 44 show projections using the model but instead of the aSAS function the power-law weighted distribution has been used.

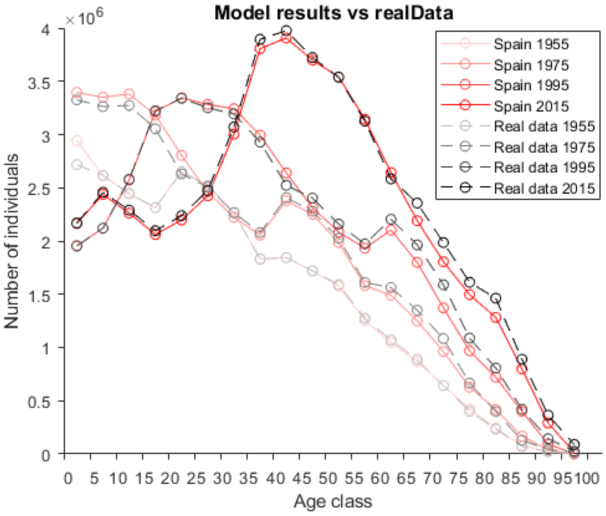


Figure 44: Spain's population using the weighted CDF function

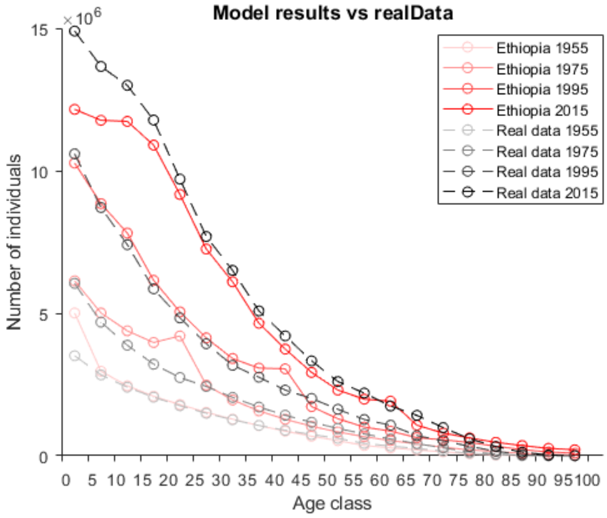


Figure 43: Ethiopia's population using the weighted CDF function

It is interesting to observe how the peaks in population shift at the same pace the years do. This is also important to corroborate that the model works correctly.

5.4 Future predictions 2015-2100.

This section shows and compares the prediction made by using this report model and the ones done by the UN. It has been used the weighted distribution and the results are presented in Figure 45 and Figure 46.

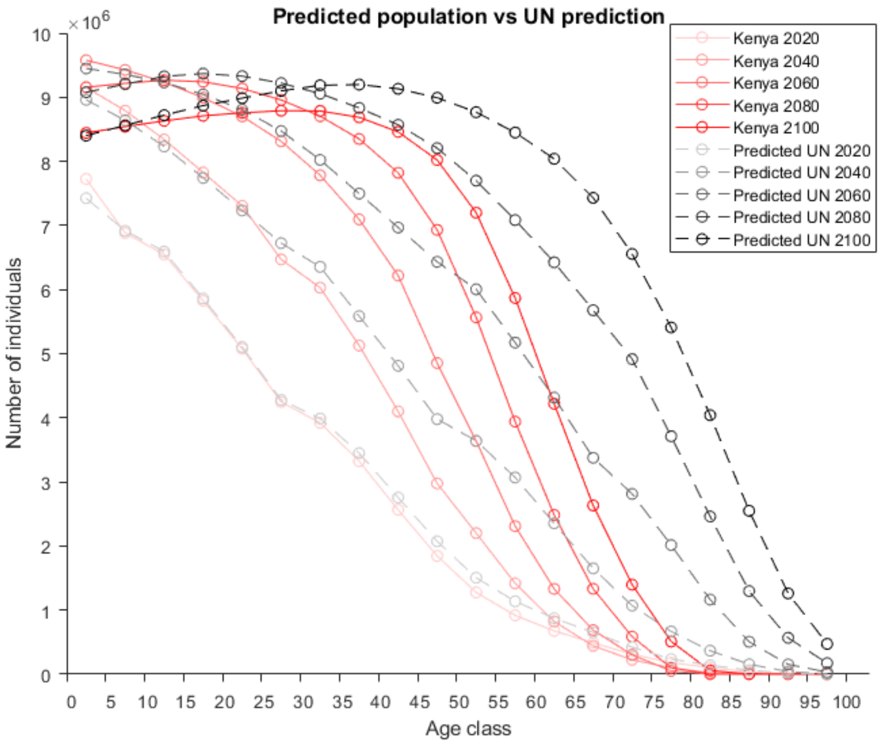


Figure 45: Ethiopia's predicted population using the weighted CDF function

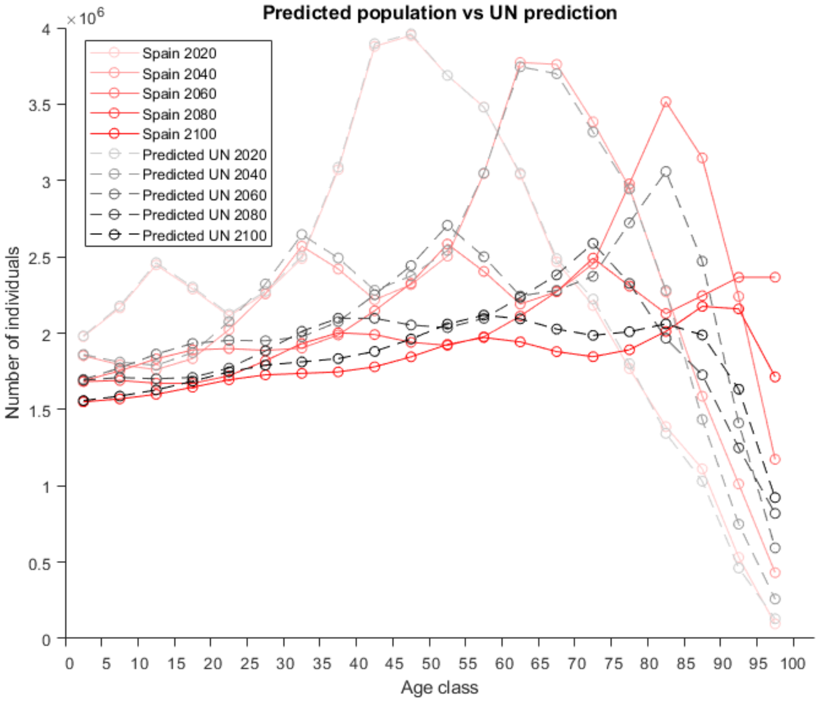


Figure 46: Spain's predicted population using the weighted CDF function

The model presents a good performance for developed countries. However for the less developed ones there seems to be some differences between the UN predictions and the ones from the model.

The trends for future population in basis of the previous plots are:

- General ageing of the population.
- Decrease in the number of births.

World's population is ageing, this means that there is an increasing percentage of older people in the population. This is believed to be one of the most significant social phenomena of the last century. This transformation will affect labor and financial markets, the demand for goods and services and health industry amongst many others. Some solutions in order to sustain economic growth in order to guarantee the well-being of the population can be:

- Educational system. Create better human capital in terms of results and capacity.
- Labor market. The decline in the labor force population means an increased participation of women and older adults in the industry.
- Higher automatization of the industry sector. Less dependency of the human factor.

Figure 47 shows the world predicted population until 2100 and also the resident population since 1955.

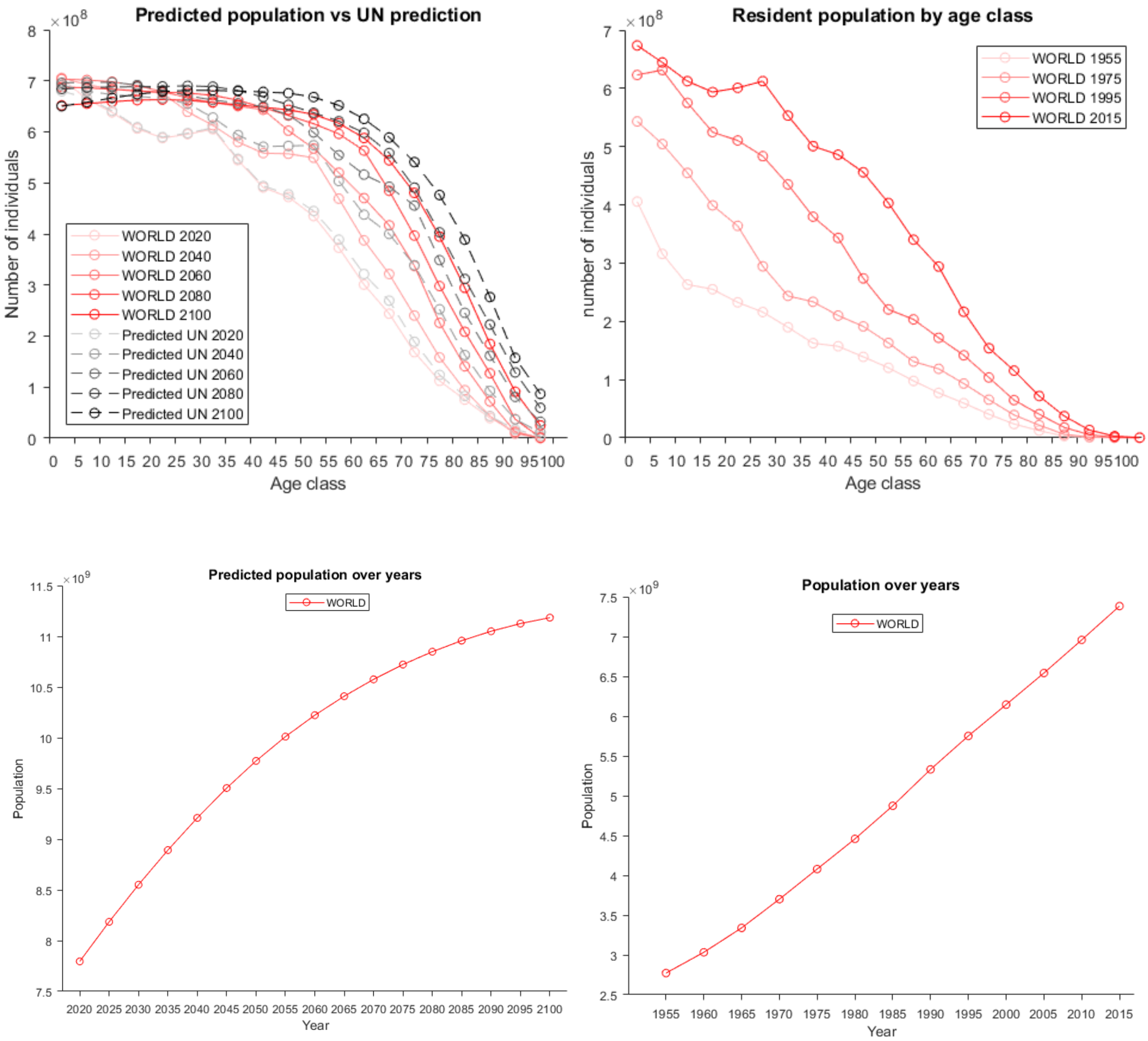


Figure 47: World population since 1955 and predicted world population until 2100

It's clear that boom of new borns have ceased. The last 100 years have foreseen an unprecedented increase in the population. This is mainly because of the medical advances and lower mortality rates. The predicted growth tells us that by 2050 the world population would be 9.7 billion and by 2100 will reach the 11 billion. This does not necessarily mean that population will keep increasing over time although demographers cannot know for sure if today's high-fertility countries will follow historical patterns. What is known is that fertility trends say that the number of newborns is to decline. However, there are several more factors that can impact on this situation. Governments for example by creating policies such as China's one child policy or by promoting economic development can depict the trends of fertility. So it is hard to predict if one can expect peak population in the following decades. In regards to the earth population capacity is safe to say that population growth increase resource consumption and as a consequence this impacts the environment. A study made by a group of biologists estimated that using all the Earth's land, the planet could host a population of 282 billion people. This is an extreme forecast however there is no simple and unique threshold that balances the human population and Earth's resources [22].

5.5 What is next

The are three parameters that determine the dynamics of the population. In this project only the mortality has been modelled. In further revisions several things can be done:

- Model the fertility. The fertility nowadays is modelled using the ASFR, however different approaches may be good too.
- Model migration. Net migration is one of the most difficult parameters to model. It's highly volatile. A high efficient way to do so is by creating a model that contemplates as many migration variables as possible. One less efficient might be create a model that only predicts migration for the next 50 years.
- Find a better distribution for approximate the SAS function. The one used in this project requires one for the High Income countries and another for the Low Income countries (3 parameters change) . Next goal is to find one that by just changing one parameter would offer better results.
- Consider different scenarios. For example what effect would have a war in population or a natural disaster such a tsunami or an earthquake.

6. Conclusions

This project has studied the actual trends in population dynamics and has created a model to predict future population distribution by age. The first lesson that has been learned is that fertility, mortality and migration are the key factors in terms of population changes. Culture and politics can impact a lot in the paths of fertility and mortality. However, the possible outcome when there's a change in the cultural, social, economic and political conditions is not easy to predict due to the absence of a good theory of prediction. One solution to this problem is go out there and collect better data, especially in the less developed countries.

The second thing that has been studied is the different level of fertility and mortality in terms of newborns depending on the country. In broad terms it has been concluded that in less developed countries there is still a high fertility rate and also a high new born mortality rate. However, the data tells us that most of this countries have already started a transition into the more developed countries where there is a lower fertility but the new born mortality is much lower as well.

The third thing discovered is that population is ageing. Especially in the more developed countries where they faced a baby boom several decades ago and now the fertility rate seems to be lower. The population of this countries over the following years will experience a continuous ageing. This occurs for two reasons: fewer people are dying and families tend to have less children. One of the many consequences this might induce, and perhaps the most important, is that the size of the labor force will decrease and this will limit its economic capacity to cover the expenses of the ones who have already retires.

The fourth thing that has been proven in this project is that using the SAS functions as a way to model the mortality is a valid approach. The well understanding of SAS function has allowed to create a model capable to predict future populations based on it with an acceptable level of accuracy.

The last goal achieved is to have found a probabilistic distribution in order to model the shape of the cumulative SAS function. This discovery has permitted to decrease the real data used in the model an increase the percentage of the modelled one. The proposed solution is a weighted power-law distribution that depends on three variables. Give a different value to these three variables and the function can be used in several different environments. In this project the values have been computed for three major regions classification: High-income countries, Middle-income countries and Low-Income countries.

7. Acknowledgements

First, I would like to thank Dr. Damiano Pasetto and Dr.Paolo Benettin. Without their help and support the realization of this project would not have been possible. Thanks for your time and dedication. Thank you for trusting me and supporting me from the moment we met, and giving me the opportunity to work in the laboratory of ecohydrology (ECHO) of the Faculty of environmental Engineering at EPFL. This has enabled me to learn from and his team and to grow both professionally and personally.

Thanks to Professor Andrea Rinaldo for always being willing to help me and to solve all the doubts that have arisen during the development of the project.

Finally I would like to express my gratitude to my friends for being with me in the good and not so good times, and to my parents and my partner for giving me their unconditional support.

8. Bibliography

1. Frequently Asked Questions About Population The Global Population by 2100. (2016).

2. Animation: The World's Population in 2100 by Region. Available at: http://www.visualcapitalist.com/animation-world-population-2100-region/. (Accessed: 22nd June 2018)

3. World Health Organization. 3. Global and regional food consumption patterns and trends. (2018). Available at: http://www.who.int/nutrition/topics/3_foodconsumption/en/index5.html. (Accessed: 22nd June 2018)

4. Ellis, J. *The World War II databook : the essential facts and figures for all the combatants*. (Aurum, 1993).

5. Frank, N. L. & Husain, S. A. The Deadliest Tropical Cyclone in History. *Bull. Am. Meteorol. Soc.* **52,** 438–445 (1971).

6. Rangel De, R. & Guimarães, M. Uncertainty in population projections: the state of the art. *R. bras. Est. Pop* **31,** 277–290 (2014).

7. Lesson 8: The Cohort Component Population Projection Method — MEASURE Evaluation. Available at: https://www.measureevaluation.org/resources/training/online-courses-and-resources/non-certificate-courses-and-mini-tutorials/population-analysis-for-planners/lesson-8. (Accessed: 13th July 2018)

8. Lee–Carter model - Wikipedia. Available at: https://en.wikipedia.org/wiki/Lee–Carter_model. (Accessed: 24th June 2018)

9. von der Linden, W., Dose, V. & von Toussaint, U. Bayesian probability theory. *Bayesian Probab. Theory* **9781107035,** 1–637 (2010).

10. United Nations. Overview | United Nations. *United Nations* (2018). Available at: http://www.un.org/en/sections/about-un/overview/. (Accessed: 24th June 2018)

11. UN DESA. Who we are | UN DESA | United Nations Department of Economic and Social Affairs. Available at: https://www.un.org/development/desa/en/about/who-we-are.html. (Accessed: 24th June 2018)

12. World Population Prospects - Population Division - United Nations. Available at: https://esa.un.org/unpd/wpp/Download/Standard/Population/. (Accessed: 12th July 2018)

13. Wang, S., Fang, C., Wang, Y., Huang, Y. & Ma, H. Quantifying the relationship between urban development intensity and carbon dioxide emissions using a panel data analysis. *Ecol. Indic.* **49,** 121–131 (2015).

14. World Bank. Fertility rate, total (births per woman). (2016). Available at: https://data.worldbank.org/indicator/SP.DYN.TFRT.IN. (Accessed: 25th June 2018)

15. Age-Specific Fertility Rate.

16. Nargund, G. Declining birth rate in Developed Countries: A radical policy re-think is required. *Facts, views Vis. ObGyn* **1,** 191–3 (2009).

17. Haupt, A. China's birth rate reported on rise. *Popul. Today* **15,** 3–4 (1987).

18. Preston, S. H. Causes and Consequences of Mortality Declines in Less Developed Countries during the Twentieth Century. **ISBN,** 0–226 (1980).

19. Francis, D. R. Why do Death Rates Decline? *The National Bureau of economic research* (2005). Available at: http://www.nber.org/digest/mar02/w8556.html. (Accessed: 27th June 2018)

20. Dji. Country Classification System. *Dow Jones Indexes* 1–3 (2011). doi:http://web.worldbank.org/WBSITE/EXTERNAL/DATASTATISTICS/0,,contentMDK:20421402~pagePK:64133150~piPK:64133175~theSitePK:239419,00.html#Low_income (Accessed February 5, 2007)

21. Guinea, A. *et al.* 2017 List of Low, Lower-Middle, and Upper-Middle income economies according to the World Bank Low-Income Economies ($1,025 or less) Lower-Middle-Income Economies ($1,026 to $4,035). *38Th Annu. Conf. Int. Soc. Clin. Biostat.* 2 (2017).

22. Franck, S., von Bloh, W., Müller, C., Bondeau, A. & Sakschewski, B. Harvesting the sun: New estimations of the maximum population of planet Earth. *Ecol. Modell.* **222,** 2019–2026 (2011).

Appendix

ASFR study

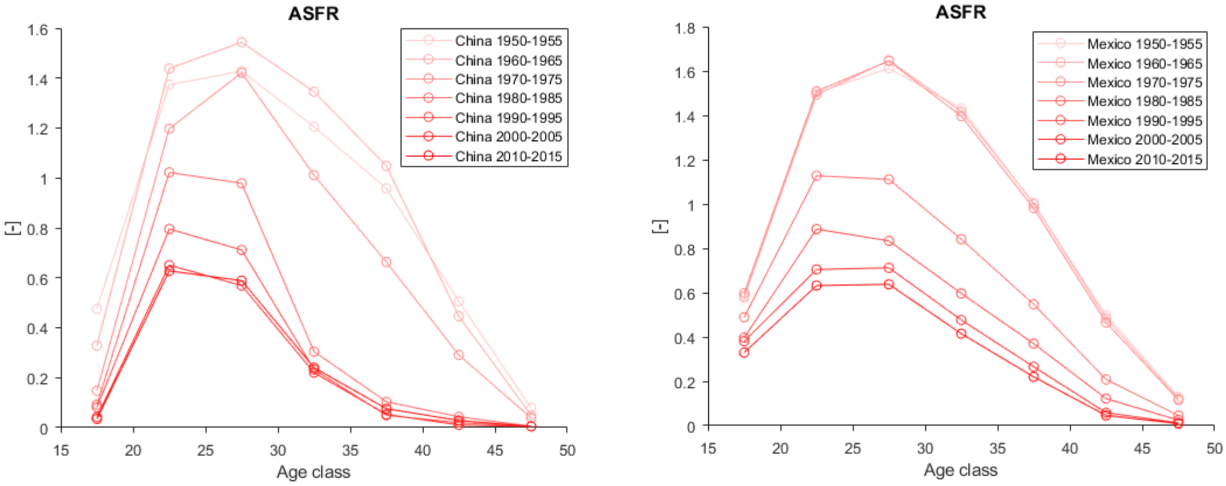
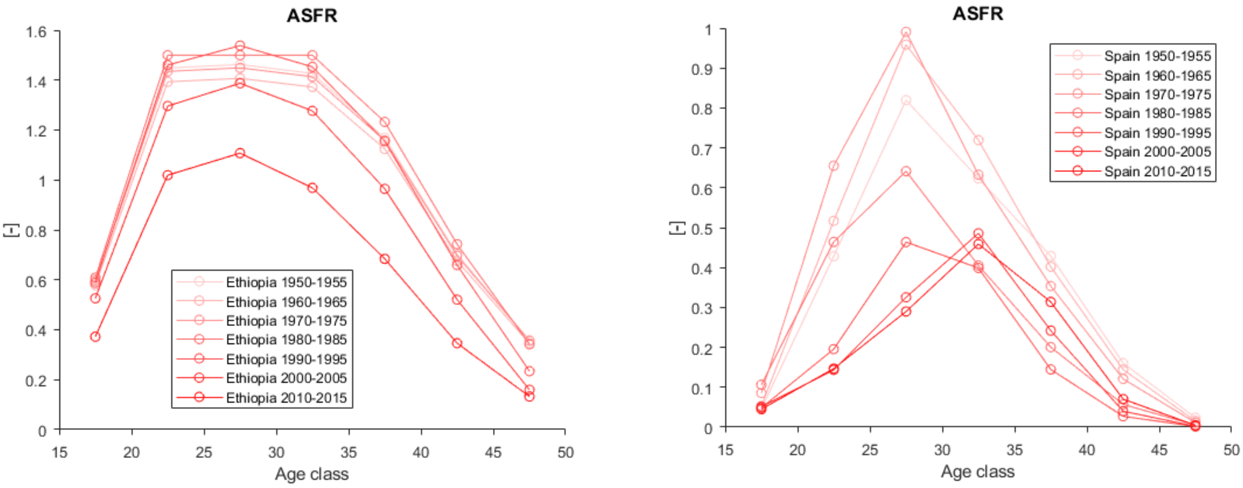


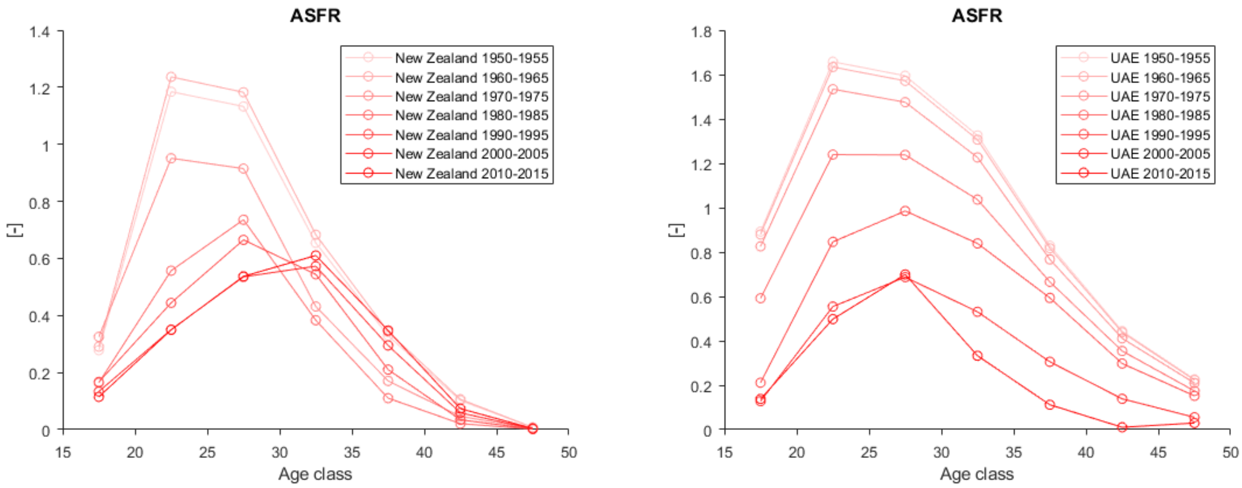
Figure A.1: China & Mexico ASFR over time by age class

Looking at Figure A.1, in China one can observe the effect of 1 child-policy introduced in 1979. In 2015-16 began to be formally phased out and apparently the curve tends to remain the same. Next to it there is Mexico, the plot shows a difference regarding the ASFR function for the early years however for the 2005 and 2015 the difference seems to be much lower.



A.2: Ethiopia & Spain ASFR over time by age class

Ethiopia is in the list of the less developed countries. The information provided in Figure A.2 tells us that the tendency is to reduce the number of births over the years however this country has not yet achieved the phase 3 of the fertility theory in which there's low birth rate and low death rate with stationary population at a low level. Then we have Spain which presents the usual transition for a developed country in the sense that has been a shift to the right regarding the age class with highest ASFR.



A.3: New Zealand & UAE ASFR over time by age class

Figure A.3, confirms that New Zealand curve is the one of a developed country. And also shows that for the last years the ASFR curve has remained more or less the same. UAE in the other hand shows a constant decrease over the years and seems to have reached its lowest point as the ASFR values for 2005 and 2015 doesn't present a really big difference.

The more developed countries present their ASFR max in the age class 30-35 after it gradually shifted over the years. Furthermore, for the most recent years the difference in terms of ASFR have been really small.

The less developed countries the pattern is not as predictable as the one before however, in all of them the ASFR is very high but over the years it has been decreasing. The main fact that characterize a less developed country is that the peak of the ASFR is in a younger age class compared with a developed country. Since not a general pattern could have been concluded from this study the case 2 of the model has not been implemented.

Domain transformation

When interpolating, two different domains can be used. The change in the domain has been used because both the aSAS function and the fSAS function present a small number of points in the beginning whilst in the end the number exponentially increases (see Figure 26). In order to interpolate correctly, the number of points should be increased in the last part of the domain. That's why an exponential expression has been chosen to fulfill this irregularity.

Both domains, the equally spaced and the exponential one, have been tested and the second one is obviously better. This is how the domain has been obtained,

$$P_s\ domain = 0\ to\ 1\ by\ increments\ of\ 0.0001$$

$$P_{slt}\ domain = 0\ to\ 1\ by\ an\ exponential\ exprexion$$

$$dlt = 0\ to - 15\ by\ increments\ of\ 0.01 \rightarrow lt = (1 - e^{(dlt)})$$

Once the $mean\ cumaSAS(P_s)$ is obtained next step is to convert it into the aSAS. Then the operation to do here is difference, the contrary one applied when obtaining the CDF. So,

$$aSAS = \frac{\sum_0^{P_s} mean\ cumaSAS(x_2) - mean\ cumaSAS(x_1)}{\sum_0^{P_s} x_2 - x_1}$$

Where,

x_i: Every single value in the P_s domain.

And here the accuracy obtained is increased in comparison with the one used when using $mean\ aSAS(P_s)$, M1, since here it is divided by p_s. And p_s in this case can be computed every time meaning that now we're not exactly obtaining the mean SAS function but one that the population of that year has an impact on it. That's another reason why the cumulative approach is also better than simply using the SAS function.