Semester project : Modeling the evolution of world populations through the McKendrick-von Foerster equation

Author : Louise Aubet
Instructor : Dr. Paolo Benettin
May 27, 2020 *EPFL, Switzerland*

Abstract

The McKendrick–von Foerster equation (MKVF) is convenient for describing the evolution in time of the age structure of a population. When using non-linear fluxes for death and birth rates, the equation cannot be solved analytically. Therefore, we build a numerical model to solve the equation. The model is applied to the case of the Philippines, in order to validate it. The aim of this work is to build a modelling framework that can be used to evaluate and predict the evolution of world population. Results show that a good approximation can be achieved and that the MKVF equation is well suited for the modelling of human populations.

1 Introduction

A population of individuals is characterised, among various properties, by its size, i.e. the total number of individuals, and its demography, i.e. how individuals are distributed over age. The evolution of a population is driven by biological and environmental processes such as birth and death rates, immigration and emigration. Depending on the importance of each of these processes, the population goes through a growth, an ageing or a decline phase. For example, Figure 1 displays the dynamics of Swiss population from 1950 to 2010. We can observe that the population is ageing as the curves are progressively shifting to the right. The number of newborns is decreasing but immigration leads to an increase in the age groups between 30 and 50 years old. Deriving a mathematical model that would perfectly describe these dynamics would be extremely complex as they rely on a lot of various processes, from geopolitics to human psychology. The aim is rather to gain insight on the crucial elements that rule the evolution of a population over time, especially the ones that affect the age structure of the population. This would help to plan in advance policies to cover social needs by age (health services, education programs, transportation, labour supply), to study fertility policies to avoid population ageing, and many other issues related with age.

For the problem we are considering, i.e. the human population of a country, there are 4 main types of population flows to take into account : newborns, deaths, emigrants and immigrants. The McKendrick–von Foerster equation describes how a single-sex population with a given age repartition evolve with time. It does not consider space, i.e. the spatial distribution of the population inside the system. It is interesting to note that this equation is very general and can be applied to various systems, including water flowing through a watershed [1], migrating birds [2] and proliferating cell populations [3]. The aim of this project is to build a valid mathematical model to solve the McKendrick–von Foerster equation numerically. Then, this model is tested on human population data provided by the United Nations.

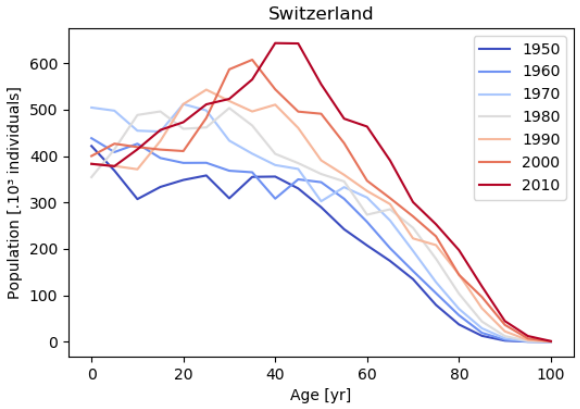

Figure 1: Evolution of the Swiss population from 1950 to 2010.

2 Human Population Data

The data from the United Nations and more precisely from the Department of Economic and Social Affairs [4] has been used for this project. It displays estimates and standard projection variants for each geographical region. Indeed, one can find key demographic indicators for each UN development group, i.e. World Bank income group, geographic region, Sustainable Development Goals region, subregion and country or area, for dates within 1950 and 2100. Data can be downloaded as CSV files encoded in UTF-8 for bulk processing in statistical softwares and databases. Some of the indicators that can be found on this platform and that are used in this model are presented below.

In order to visualise the evolution of the population, Lexis diagrams are often used [5]. They provide an effective language for communicating about and understanding demographic statistics. Given a coordinate plane with axes for chronological time t [yr] and age a [yr], every demographic event is be represented by a point on this plane. The life of a person may be represented by the straight line, connecting the points representing that person's birth and death. Figure 2 gives an example of a Lexis diagram illustrating some life lines. We can observe that the person d is born during the period $[t, t+\mathrm{d}t]$ whereas the persons b and c died between age $[a, a+\mathrm{d}a]$ during the period $[t, t+\mathrm{d}t]$.

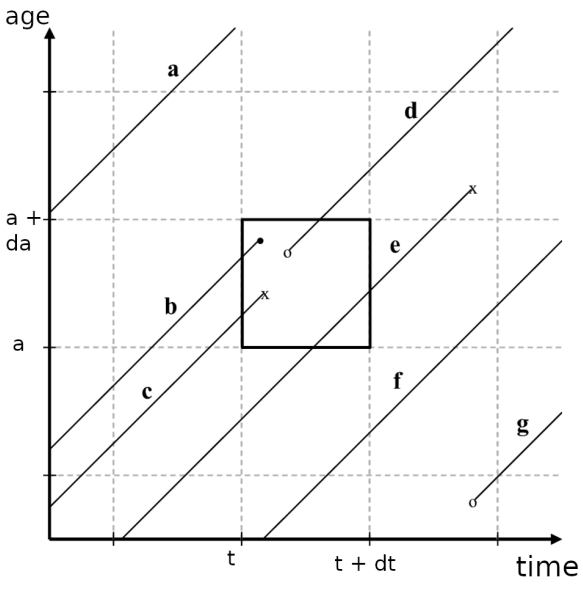

Figure 2: Example of Lexis diagram, taken from [6].

2.1 Population indicators

The population repartition is given by 5-year age group (in thousands of individuals [10^3 indiv.]) every 5 years. The population repartition is separated into women repartition and men repartition. The total population [10^3 indiv.] can also be useful. This data is more precise in time as it is given every year. The net migration rate [10^3 indiv.] is another indicator that is going to be used in this project. It takes into account the cumulative effect of both immigration and emigration. Indeed, this rate is affected positively by immigration and negatively by emigration.

2.2 Fertility indicators

The *Age-Specific Fertility Rate* (ASFR) [ ] is the number of births per woman in a particular age group. Data is given by 5-year age group, from 15 to 45 years old, every 5 years. Below 15 and above 45 years old, the ASFR is considered to be zero. The age-specific fertility pattern has a typical shape common in all human populations through years. Figure 3 illustrates the shape of the average ASFR curves in all the continents. We can observe that in all cases the shape is very similar to a Gaussian. The parameters of this Gaussian function would evolve with time. As it can be seen in Figure 3, the amplitude is decreasing with time and the abscissa of the maximum is shifting to the right. This change is typical to the demographic transition. The *demographic transition* is a phenomenon which refers to the shift from high birth rates and high infant death rates to low birth rates and low death rates. Generally, the society goes from minimal technology, education and economical development to an industrialised economical system with advanced technology and better education. This shift can be seen clearly in Europe, Oceania and Northern America. In Africa, Asia and Latin America, the decrease of birth rates doesn't go along with a shift of fertility towards older age groups. There is still a lot of differences between the continents.

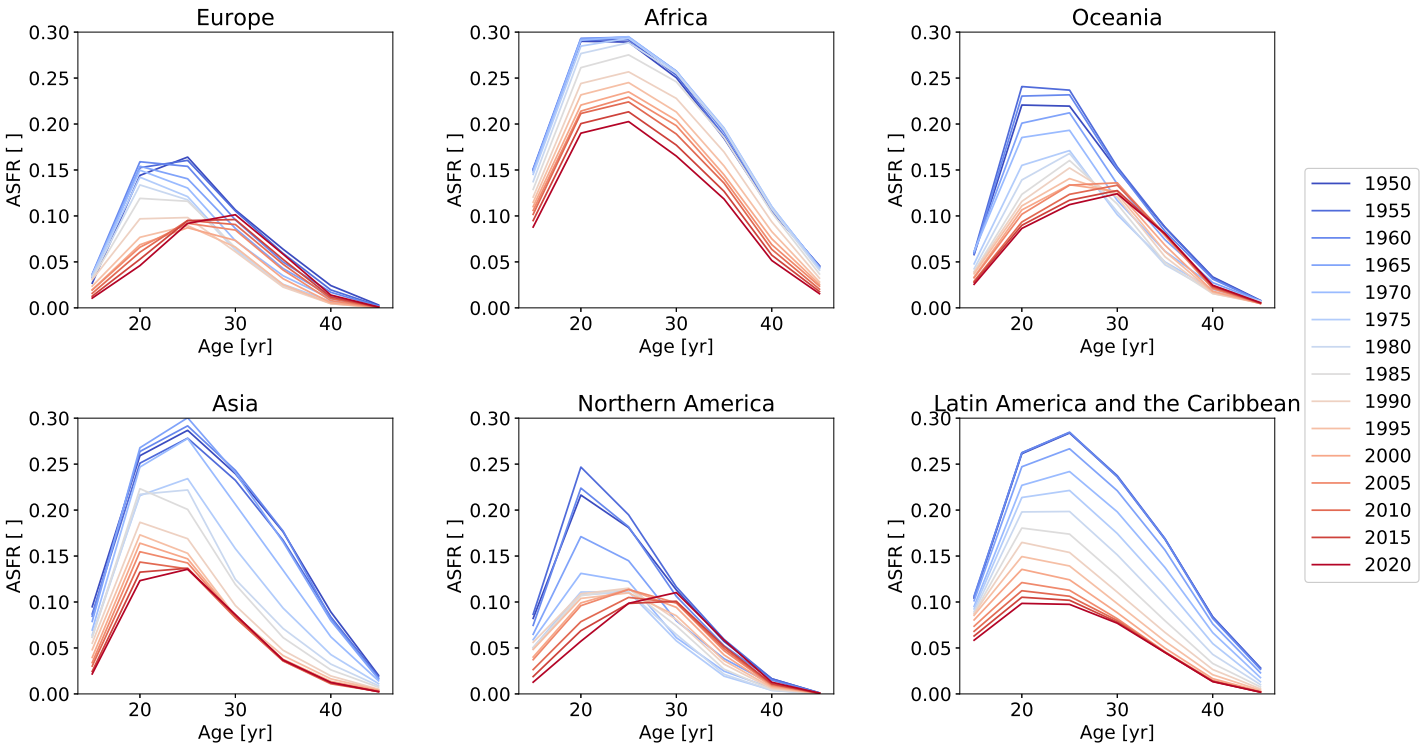

Figure 3: ASFR value in the different continents, as a function of time.

The *Total Fertility Rate* (TFR) [indiv.] is the average number of children a woman would have at the end of her reproductive period. Data is given every 5 years. Figure 4 illustrates TFR values for all the continents as a function of time. A great variety of shapes can be observed. In 2020, almost all continents have a TFR value below 3. In Africa, TFR is decreasing steadily but is still high with approximately 4.5 in 2020.

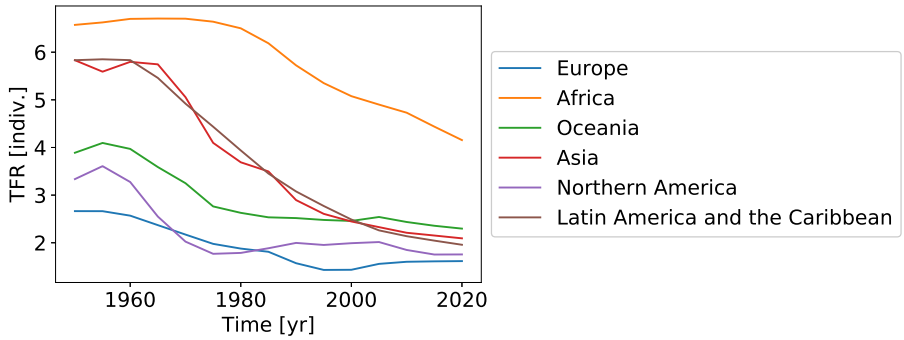

Figure 4: TFR value in the different continents, as a function of time.

Another interesting quantity is the *Replacement Level* (RL). It represents the particular TFR value such that a population exactly replaces itself from one generation to the next. It results in a stable population without it increasing or decreasing. The idea is to look at the population that is decreasing due to death or mobility and estimate the number of births necessary to replace the loss. This number varies greatly across the globe and in time. Nowadays, in European countries like Switzerland, the value of the TFR_{RL} is slightly above 2. The reason is that in a case without early death and no net migration, it would take exactly two children to replace the two parents. In a more realistic system with non zero death and migration rates, the replacement level increases slightly. In Figure 4, we can observe that almost all of the TFR curves are converging towards the line TFR $= 2$ which corresponds approximately to the replacement level in developped countries.

2.3 Death indicators

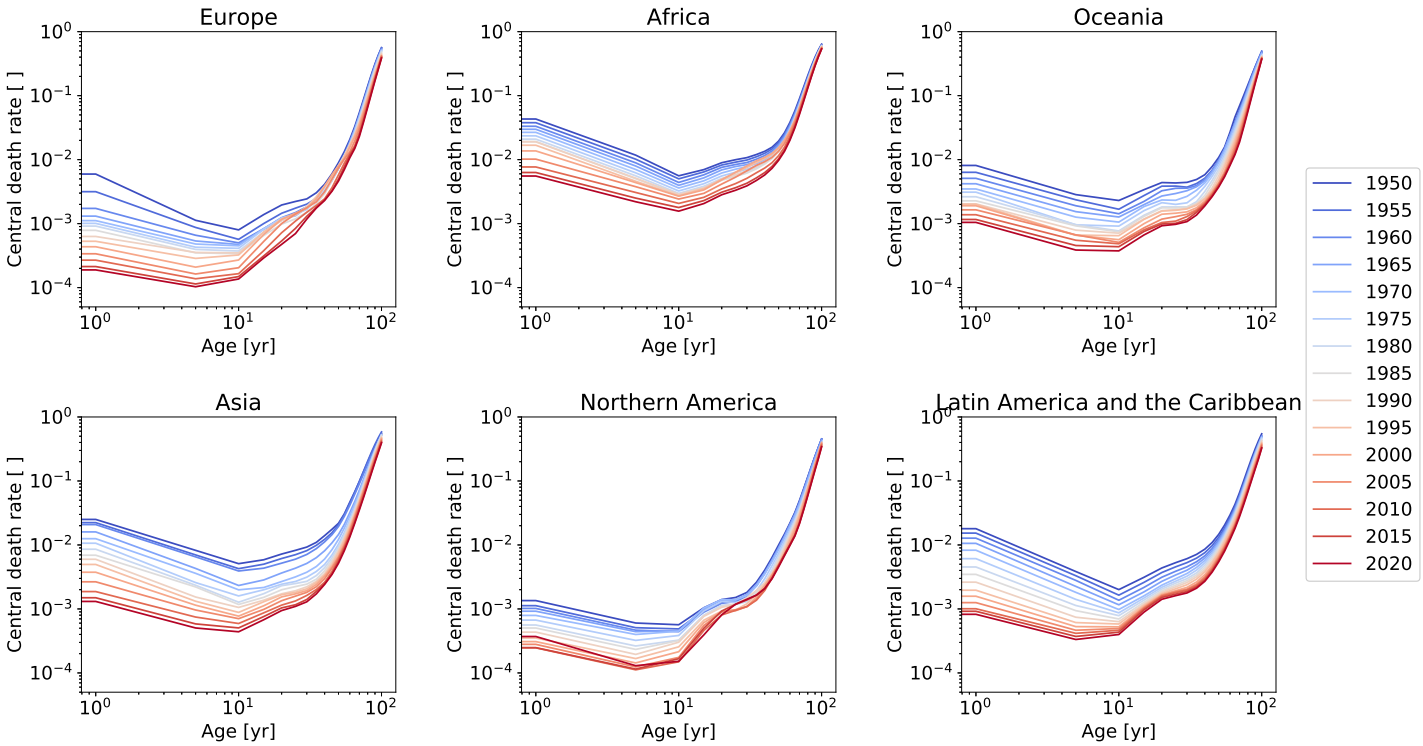

Figure 5: Central death rate in the different continents, as a function of time. Both axis have logarithmic scale.

For a given population, the central death rate at age a during a given period of da is found by dividing the number of people who died during this period while aged a, i.e. after they had reached the exact age a but before reached the exact age $a + \mathrm{d}a$, by the average number of those who were living in that age group during the period. [7] Data is given by 5-year age group, starting from 1 year old, every 5 years. The year 0 is given separately as infant mortality is critical for that period. Figure 5 displays the central death rate in the different continents as a function of time from 1950 to 2020. The logarithmic scale is used in order to better visualise the shape of the curves. One can see that the death rate is approximately an exponential function with respect to age and in addition the infantile mortality. I can be seen that the death rate is decreasing with time. The lowest rate is found at 10 years old and then it increases monotonically. A noticeable fact is the recent increase of death rate in Northern America from 2015 to 2020, going against the tendency of the other curves.

3 Definition of the physical problem

First, we need to select the geographical boundaries of the system, usually a country or a group of countries. Then, the fluxes of population can be seen as inputs and outputs, people that enter or leave the system. This is a mass balance approach : fluxes that enter the system are positive and fluxes that leave are negative. Given some initial population, the system will evolve with time following the governing equations and interact with the exterior through the fluxes. The aim is therefore to create a mathematical model that will help us understand how a population of individuals evolve through time and age.

The considered physical problem and the equations involved need to be defined precisely. The domain of definition Ω is in two dimensions as it needs to describe two dimensions of time : the age $a \in [0, \infty]$ and the chronological time $t \in [0, \infty]$ both measured in years [yr]. The main equation is the McKendrick–von Foerster equation :

$$\begin{cases} \frac{\partial p}{\partial a}(a,t) + \frac{\partial p}{\partial t}(a,t) = p(a,t) \sum_k f_k(a,t) \\ p(0,t) = \int_0^A \beta(a,t) p(a,t) \mathrm{d}a = B(t) \\ p(a,0) = p_0(a) \end{cases} \tag{1}$$

In this system, the second and third equations describe the boundary and initial conditions respectively. $p(a,t)$ [# individuals.yr^{-1}] is the age density of the population, i.e. the number of individuals of age in a [yr] at time t [yr]. $f_k(a,t)$ [yr^{-1}] represents the k-th flow of individuals. $\beta(a,t)$ [yr^{-1}] is the age-specific fertility, characterising the flux of newborns. They are produced inside the system and are taken into account in the boundary condition, so there is no explicit flux f_β for births. The birth rate depends on the fertility of existing individuals in their reproductive age. The age-specific mortality $f_\mu(a,t)$ describes the number of individuals that die. The age distribution has a physiological shape : the vulnerability depends on the age. $f_e(a,t)$ and $f_i(a,t)$ are the fluxes for emigration and immigration. Immigrants are produced outside and then cross the system boundaries. On the opposite, emigrants are individuals who leave the system boundaries. Both emigration and immigration age distributions are complex and depend on many diverse drivers. Therefore, in the beginning, we will only take into account the mortality rate, thus the system becomes :

$$\begin{cases} \frac{\partial p}{\partial a}(a,t) + \frac{\partial p}{\partial t}(a,t) = -p(a,t) f_\mu(a,t) \\ p(0,t) = \int_0^A \beta(a,t) p(a,t) \mathrm{d}a = B(t) \\ p(a,0) = p_0(a) \end{cases} \tag{2}$$

The *survival probability* is defined as :

$$\Pi(a) = \exp\left(-\int_0^a f_\mu(\tau)\mathrm{d}\tau\right) \tag{3}$$

It will play an important role in the understanding of age repartition over time.

4 Numerical Implementation

4.1 Numerical Scheme

First, let's bound the domain Ω : the age a is restricted to the interval $[0, A]$ and the year t to the interval $[0, T]$. We naturally discretises ages and time with the same parameter as Equation (2) is a first order partial differential equation with constant coefficients. Let $h > 0$ be the discretisation step, N is the number of sub-intervals in time and M the number of sub-intervals in age. Then the samples are $t^n = n \cdot h$, $n = 0, \ldots, N$ and $a_i = i \cdot h$, $i = 0, \ldots, M$ for time and age respectively. The solution becomes $p\left(a_i, t^n\right) = P_i^n$. It is important to pay attention to the units and to keep in mind that $p(a, t)$ in a continuous set is a population density with respect to the variable a. With the discretization, P_i^n describes the number of individuals with age in the interval $[a_i, a_{i+1}[$, and has the units of [# individuals]. The implementation is done in python.

Equation (2) can be used directly to build a numerical scheme. The *method of characteristics* is used. When treating the differential operator as an ODE in the characteristic variable t and using backward finite differences, we approximate:

$$\left(\frac{\partial}{\partial t} + \frac{\partial}{\partial a}\right) p\left(a_i, t^n\right) \simeq \frac{P_i^n - P_{i-1}^{n-1}}{h} \tag{4}$$

This approximation can be used to discretize the left-hand side of the MKVF equation. [8] suggests a method to discretize the right-hand side leading to a 2^{nd} order scheme. The boundary condition is approximated using a trapezoidal formula. The following system of equations is obtained :

$$\begin{cases} \frac{1}{h}\left(P_i^n - P_{i-1}^{n-1}\right) = -\frac{1}{2}\left(P_i^n + P_{i-1}^{n-1}\right)(f_\mu)_{i-1/2} \\ P_0^n = h\sum_{i=1}^{M-1} \beta_i P_i^n + \frac{h}{2}\left(\beta_0 P_0^n + \beta_M P_M^n\right) \\ P_i^0 = (p_0)_i \end{cases} \tag{5}$$

which gives the explicit scheme :

$$\begin{cases} P_i^n = P_{i-1}^{n-1}\left(\frac{2-h(f_\mu)_{i-1/2}}{2+h(f_\mu)_{i-1/2}}\right) \\ P_0^n = \frac{h}{(2-h\beta_0)}\left(2\sum_{i=1}^{M-1}\beta_i P_i^n + \beta_M P_M^n\right) \\ P_i^0 = (p_0)_i \end{cases} \tag{6}$$

Populations are always positive numbers, so the following condition needs to be fulfilled : in the main iteration, the numerator needs to be positive, $2 - h(f_\mu)_{i-1/2} > 0$. Otherwise the population will oscillates between positive and negative values. The inequality leads to $h < \frac{2}{\max_{a,t}\{f_\mu(a,t)\}}$. This is an upper bound for the value of h.

4.2 Input functions

The flux of newborns is taken into account in the boundary condition. The birth flux $\beta(a, t)$ [yr^{-1}] is linked to the ASFR and TFR defined in Section 2.2. At first, a simple function is used to test the model : $\beta(a, t) = \text{TFR}(t) \cdot g(a, t)$ where $g(a, t)$ is a shape function characterising the reproductive period. More precisely, g is a step function, between 15 and 45 years old, multiplied by the TFR value.

$$\beta(a, t) = \text{TFR}(t) \cdot \text{rect}\left(\frac{a - 30}{30}\right) \tag{7}$$

This form is useful to make basic tests but it suffers a number of weaknesses. Indeed, function $g(a,t)$ does not depend on time. Also, Section 2.2 illustrated the fact that the fertility rates was following a Gaussian curve rather than a rectangular function. It would be more precise to use $\beta(a,t) = \text{ASFR}(a,t)$. In order to do so, we must transform the set of discrete points from the UN database into an analytical function, defined continuously for every age a and time t. Fitting $\text{ASFR}(a,t_k)$ for a particular t_k is a reasonable task, using some assumptions on the shape of the curve. For example, a good approximation would be a Gaussian curve with two different σ values for the left and right slopes [9] :

$$f(x) = A \cdot \exp\left[-\left(\frac{x-\mu}{\sigma(x)}\right)^2\right] \text{ with } \sigma(x) = \begin{cases} \sigma_1, \text{ if } x < \mu \\ \sigma_2, \text{ if } x \geq \mu \end{cases} \tag{8}$$

Unfortunately, fitting $\text{ASFR}(a_k,t)$ for a particular a_k would be much more complex as the evolution depends a lot on the region of the world. Furthermore, fitting $\text{ASFR}(a,t)$ as a 2D function would require a lot of computational power. A more reasonable solution would be to use local interpolation. The library `scipy` contains a bivariate spline approximation over a rectangular mesh, which can be used for both smoothing and interpolating data [10]. The order of the interpolation can be chosen, it is here set to 1. This first order interpolation results in a continuous 2D function $\beta(a,t)$.

For basic tests on a simplified model, the death rate is set constant over age and time. This approximation is far from reality as we have seen in Section , that the death rate $f_\mu(a,t)$ can be modelled by an exponential curve. Later, for more realistic results, we use the central death rate from the UN database. As before, making a 2D fit from the discrete data set would be difficult. Therefore, a 2D 1$^{\text{st}}$ order interpolation is also the preferred solution. The output is a continuous function $f_\mu(a,t)$ defined at any age a and time t.

In this model, the population is represented by a single-sex population, no difference in made between men and women. However, when defining the birth rate, one needs to distinguish the two as only women can have children. We make the assumption that women represents half of the population. The boundary condition becomes : $p(0,t) = \int_0^A \beta(a,t) p_{\text{women}}(a,t)\text{d}a = \int_0^A \beta(a,t)\frac{1}{2}p(a,t)\text{d}a$. The validity of this hypothesis is checked before applying the model to any country. Immigration and emigration fluxes are not considered in this analysis. Therefore, the hypothesis that the net migration is negligible must be fulfilled in order to obtain realistic simulations. The initial population is given for 5-year age group every 5 years. By supposing a equiprobable repartition over these 5 years, we divide the raw data by 5 to get a population for 1-year age group. Then, a 1D 3$^{\text{rd}}$ order interpolation is applied.

5 Numerical Implementation Tests

5.1 Basic tests

First some basic tests are run in order to check the validity of the numerical model. The system chosen is Switzerland, starting from $y_0 = 1950$, with $A = 100$, $T = 100$ and $h = 1$. The birth rate is modeled using the rectangular window and different values of TFR. The simulated population is plotted as a function of age, for different time steps. The first case is the simplest one, zero newborn and zero death, and is displayed in Figure 6. We can observe a progressive shift to the right of the population repartition as time passes. The population is zero on the left as there is no newborn. Only the initial population is ageing. At $t = 100$, there is no population anymore.

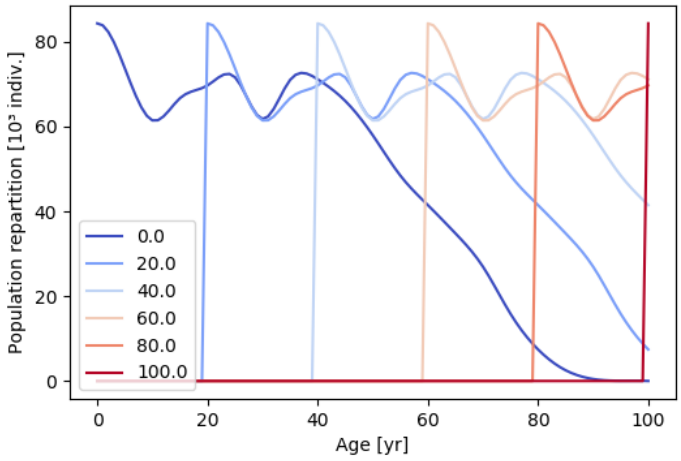

Figure 6: Simulated population in a simplified model with no death and no newborn.

Then, we set a non zero birth rate. Using TFR= $1, 2$ and 3, we get the results displayed in Figures 19 to 9. We can still observe the shift with time. The right part of the graph is not varying in amplitude as there is no mortality. Only the left part is changing, depending on the TFR value. If the TFR value is small (i.e. below the replacement level) as in Figure 19, we can see that the number of newborns decreases with time as there are less and less women in their fertility period. The total population goes to zero with time. When the TFR value is high (TFR $= 3$ as Figure 8), the number of newborns increases every year. The total population increases exponentially. It is possible to notice a step from the first iteration, and that is propagating with time. This is due to the fact that at the first iteration, the number of newborns is much higher than the number of newborns in the initial population.

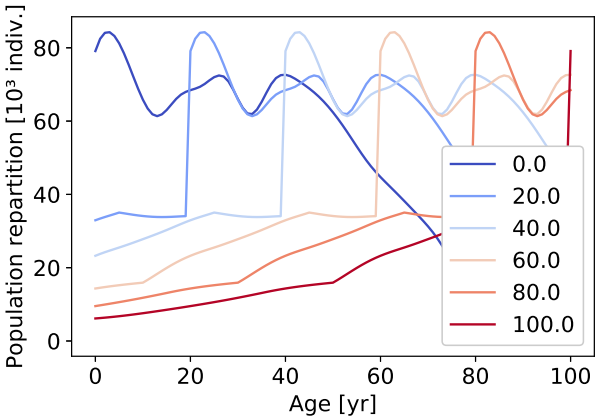

Figure 7: Simulated population in a simplified model with TFR $= 1$ and no death.

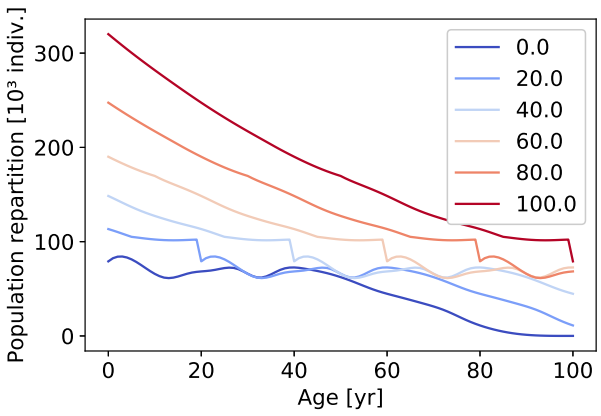

Figure 8: Simulated population in a simplified model with TFR $= 3$ and no death.

Also let's study an interesting case. Without death, the replacement level is exactly 2. To keep the population constant, every family of two parents would hypothetically need to have 2 children to replace them. The result of the simulation with $\mathrm{TFR}_{\mathrm{RL}} = 2$ is illustrated in Figure 9. The right part of the curve is shifting to the right as before, and we can observe that the left part is almost constant. It is not perfectly constant, some slight iterations can be seen, due to the fact that the initial population is not equally partitioned over age. Also, the total population should remain constant. Indeed, we can see that the population stabilises itself after approximately 50 years, in Figure 10. It corresponds to the moment at which the newborns in 1950 are approximately 45 or 50 years old, so they are not in the fertility period anymore. From that time, the number of women in the fertility period stays constant so the total population stays constant.

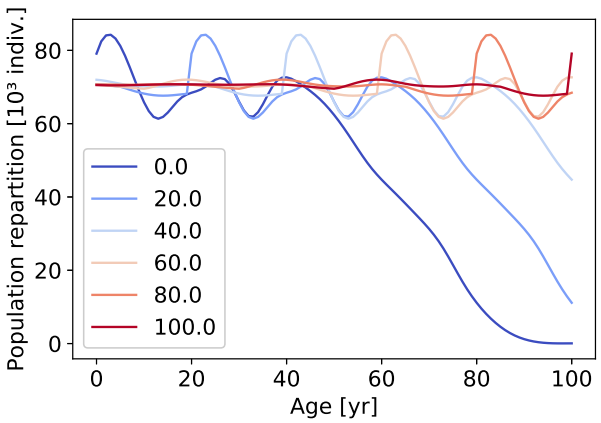

Figure 9: Simulated population in a simplified model with TFR = 2 and no death.

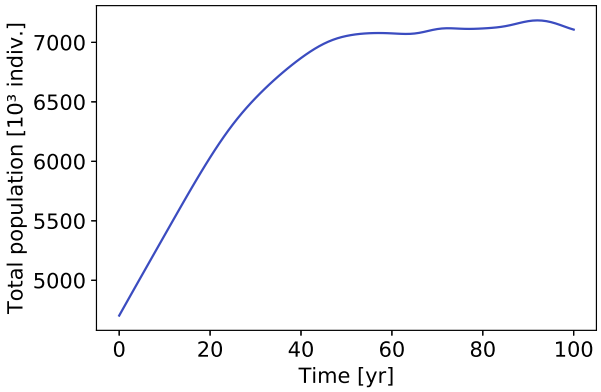

Figure 10: Simulated total population in a simplified model with TFR = 2 and no death.

Figures 11 shows the result for a simulation with no birth and a typical death rate for Switzerland in 1950. We can observe that the population goes progressively to zero. Here, the survival probability defined by Equation (3) plays an important role. It is plotted in black in the graph. The curves seem to tend towards the shape of the survival probability, while progressively decreasing. Finally, typical birth and death functions for Switzerland are put together in a simulation. The result is shown in Figure 12. As before, the survival probability is drawn in black. We can still observe that the survival probability curve determines the shape of the population repartition. The curves of population repartition get closer and closer to the survival probability. The birth rate is sufficiently high but not extremely high so that the population is only slowly increasing and the mortality function is prominent in the shape of the population repartition. In this case, the survival probability seems to be a steady state solution.

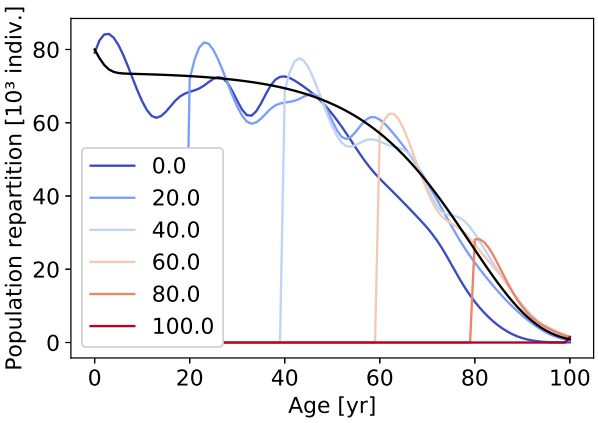

Figure 11: Simulated population with TFR = 0 and a typical mortality function. The survival probability is drawn is black.

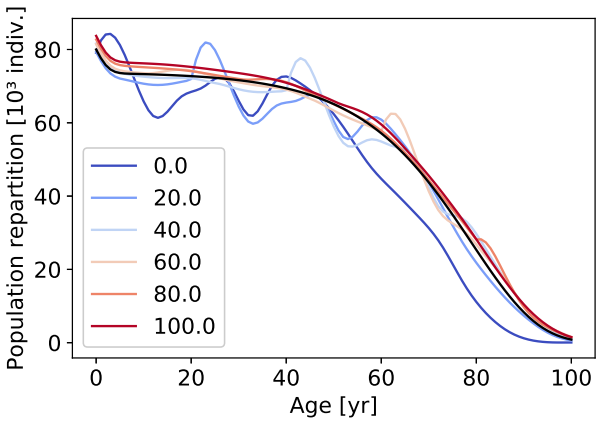

Figure 12: Simulated population with TFR = 2.31 and a typical mortality function. The survival probability is drawn is black.

5.2 Comparison with analytical solution

Let's further test this model and compare the numerical approximation with the analytical solution. The chosen values are $f_\mu(a,t) = 3$, $\beta(a,t) = 2$, $p_0(a) = 10^5 \cdot e^{-a}$. They are very basic input functions but doing so allows us to find an analytical solution :

$$p_{\text{th}}(a,t) = 10^5 \cdot e^{-2a-t} \tag{9}$$

The aim is to study the convergence of the numerical method. The approximation error between numerical and analytical solution for a time step h_k is defined as :

$$E^k(t_{\text{end}}) = \max_{a\in[0,A]}\{|\frac{p_{\text{th}}(a,t_{\text{end}}) - p_{\text{num}}(a,t_{\text{end}})|}{|p_{\text{th}}(a,t_{\text{end}})|}\} \tag{10}$$

where t_{end} is time step corresponding to the end of the simulation. In other words, the error for a given h_k is computed at the last time step and is defined as the maximum normalised difference over the age. Figure 13 displays the error as a function of the size of the time step h in a logarithmic scale. We can see that the error is decreasing when h is decreasing. More precisely, the curve is parallel to the curve h which means the error converges with order 3. The scheme is supposed to be a second order so the error should be at least convergent with order two. It is coherent. However, when h is big, the error curve is much more flat. Therefore, we must be careful to choose a value of h that is small enough so the model can convergence fast enough.

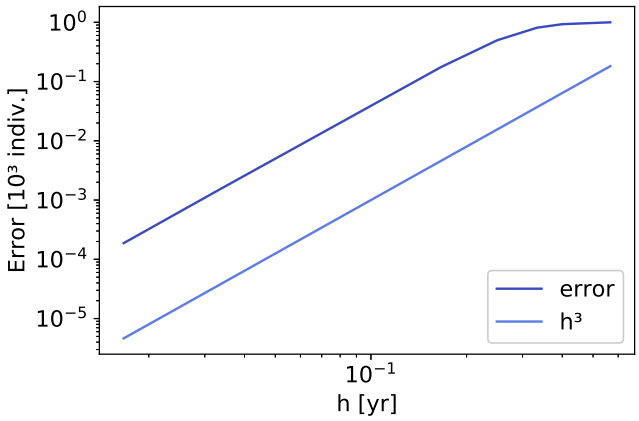

Figure 13: Error study between analytical and theoretical solution.

6 Application: Philippines population from 1950 to 2020

Now let's apply the numerical model to a specific country. The Philippines are chosen because it is known as not to have high emigration and immigration rates. This hypothesis will be further tested. Therefore, our simplified model without emigration and immigration has chances to be realistic compared to another country in which the effects of these fluxes would have been non negligible. In order for the model to be the most realistic, the input functions are taken as interpolation of UN data. The numerical method is run starting from $y_0 = 1950$ and ends at $y_f = 2020$, with time step $h = 1$ yr.

Before applying the model, the hypotheses are verified. Figure 14 displays the net migration rate normalised over the total population from 1950 to 2020 in the Philippines. The data concerning the net migration rate and the total population are coming from the UN database. We can see that the net migration rate represents at most 0.35% of the total population. Therefore we can consider that the net migration is negligible and this flux can be omitted in the equations. This particularity validates the choice of the Philippines for the first application. Figure 15 exhibits the fraction of women in the total population from 1950 to 2020. The data concerning the women population and the total population are coming from the UN database. We can observe that this ratio varies between 0.494 and 0.502, which forms very small oscillations around 0.5. Therefore, the hypothesis that approximately half of the population are women is validated. When applying the birth function, we can apply it to half of the total population.

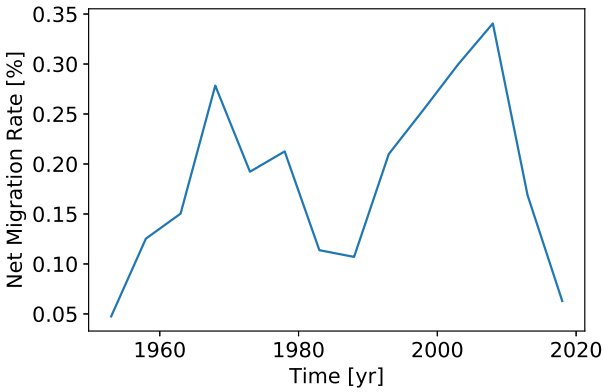

Figure 14: Net migration rate as a percentage of the total population in the Philippines.

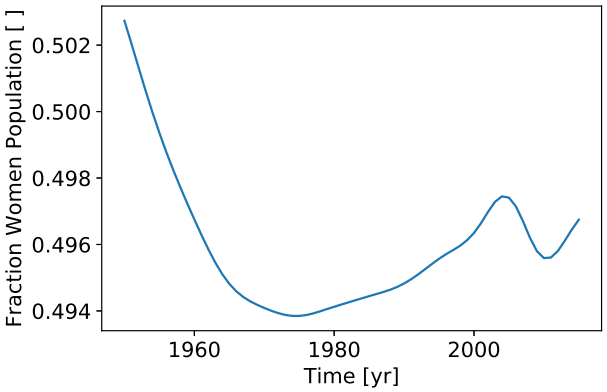

Figure 15: Fraction of women in the total population in the Philippines as a function of time.

The result of the numerical simulations is exhibited in Figure 16. The population repartition is computed every year but is plotted every 10 years only, for visibility reason. The dark blue curve at $t = 0$ corresponds to the initial population in 1950, as given in the UN database. The dark red curve at $t = 70$ corresponds to the final result of the model, that is the approximated population in the Philippines in 2020. It can be observed that the population is continuously increasing. This graph alone can give a lot of informations about the evolution of the population over time. There are lots of newborns and their number increases every 10 years by approximately $500 \cdot 10^3$ in average. This is due to the very high fertility rates especially in the first years of the simulation. To give an example, in 1950 the TFR was equal to 7.42 and it progressively decreases up to 2.45 in 2020. This explains the slow-down of the births increase in the last 30 years. Indeed, at time step 60, i.e. in 2010, the slope for the low age group is less steep. At $t = 70$, the number of births is slightly smaller than at $t = 60$ and the repartition in age groups between 0 and 20 years old is quite constant. Also, a decrease of the infant mortality due to better sanitary conditions can participate to this fast increase of the number of births. This graph characterises a very young population. The integral under the curve represents the total population so it can be seen that the majority of the population is below 50 years old. Also the curves are shifting to the right, which means people live longer and life expectancy is increasing with time. Here, the curves are very different from the shape of the survival probability. We are not in the "steady case" that has been observed above. This can be explained by the birth rate that is really high and drives the evolution of the population.

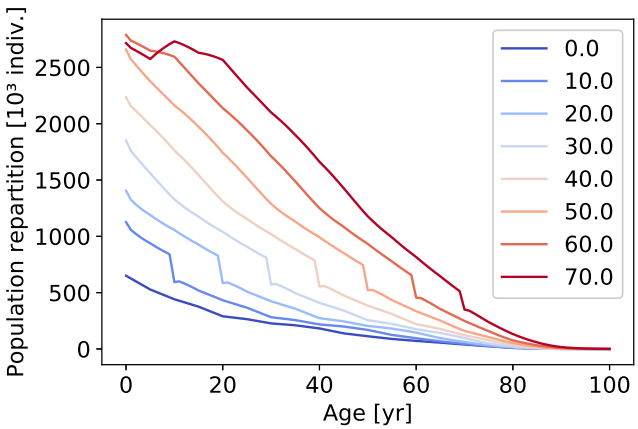

Figure 16: Evolution of the population repartition over time in the Philippines, from 1950 to 2020.

The total population as a function of time is plotted in Figure 17. It is compared to the total population given by the UN, in black in the graph. We can see that the numerical approximation is really good up to time step 25. Then they start to diverge more and more because the model overestimates the total population. At the last time step, the overestimation is of about $20000 \cdot 10^3$ individuals. To understand what causes this discrepancy, let's do more precise comparisons between the numerical approximation and the UN data. Figure 18 compares the output of the model and the UN data for the population repartition in 2020. The first interesting fact is that the numerical approximation almost perfectly overlap the measured values for ages above 70 years old. 70 is not a random number, it corresponds to the actual time step. It means the population older than 70 was alive at the beginning of the simulation and they have been affected by the mortality rate only. The population younger than 70 was born during the simulation and was affected by the birth rate and the mortality rate. For this part of the population, there is a gap between the measures and the numerical approximation. Even so the curves have similar shape : the slopes are approximately the same, there is a gap and this gap is bigger for lower age groups. This discrepancies would mean that the birth rate is overestimated. This error is accumulated from one time step to another so that the gap between the approximated number of newborns and the actual one is increasing with time. The non-negligible discrepancy for low age groups may be exacerbated by round-off errors. Indeed, the numerical model is dealing with big numbers and approximations errors are unavoidable.

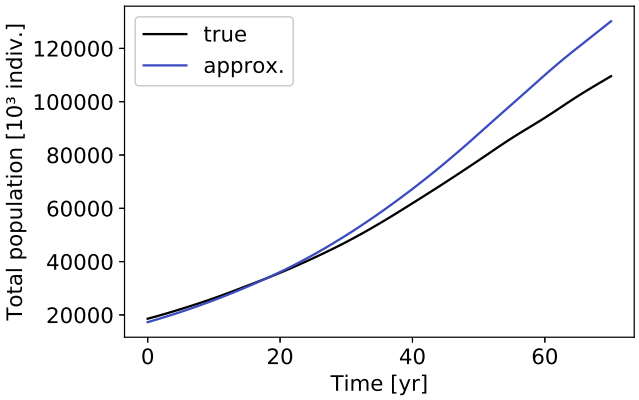

Figure 17: Approximation of the total population over time in the Philippines compared to the population measured by the UN.

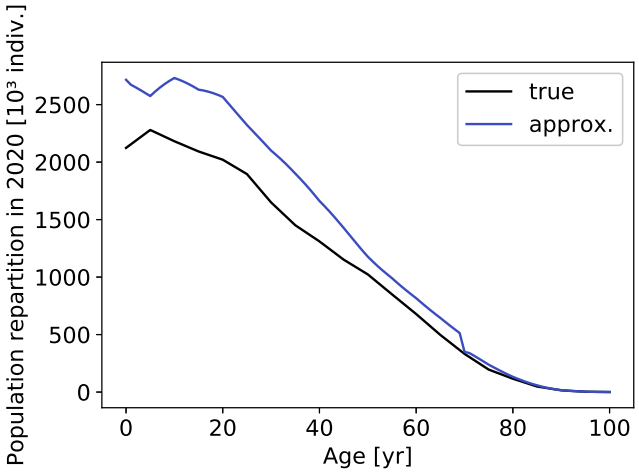

Figure 18: Comparison between the approximation of the population and the UN data, in 2020 in the Philippines.

An anomaly can be noticed on Figure 16. At time step $t = 10$, a discontinuity appears and is then propagated with time. Let's study in more details the first steps of the simulation in order to understand the formation of this anomaly. Figures 19 and 20 compare the approximation of the population and the UN data, in 1951 and 1955. At the first time step, we can observe that the number of newborns is suddenly very high compared to the one in the initial population, due to the very high fertility rate. It changes the slope of the curve, which becomes steeper. This discontinuity is exacerbated by a relatively high infant death rate in the next steps. To summarise, the combination of high natality and relatively high infant mortality creates this anomaly. It is due to the input functions and not the model in itself. Indeed, when decreasing artificially the fertility rate and removing the infant mortality, this discontinuity disappears. Figure 21 illustrates the output of the model when the input functions are modified : the ASFR values are decreased and the infant mortality is suppressed from the mortality rate. The curves are smoother and no discontinuity is visible. However, we can notice now a clear underestimation of the population, as a consequence of the decrease of fertility rate.

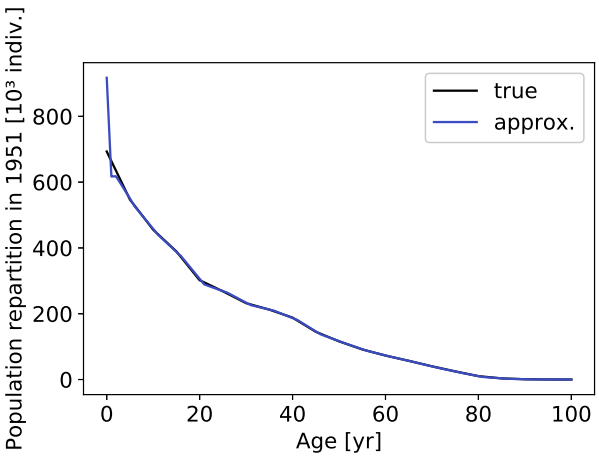

Figure 19: Comparison between the approximation of the population and the UN data, in 1951 in the Philippines.

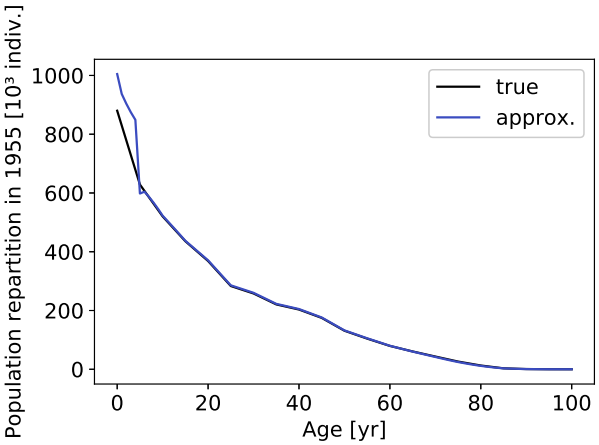

Figure 20: Comparison between the approximation of the population and the UN data, in 1952 in the Philippines.

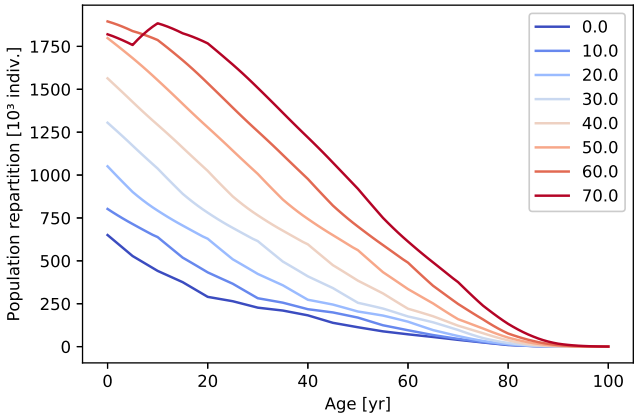

Figure 21: Evolution of the population repartition over time in the Philippines, from 1950 to 2020, with an artificially low birth rate and no infant mortality.

7 Alternative Model Implementation

As seen above, round-off errors might lead to an increase of discrepancies between the numerical approximation and the real data. In order to limit computational errors, it is convenient to define the normalised age profile $v(a,t) = \frac{p(a,t)}{P(t)}$ where $P(t) = \int_0^A p(a,t)\mathrm{d}a$ [# individuals] is the total population at time t. By replacing this change of variable in Equation (1), and applying the following hypothesis on the evolution of the total population :

$$\begin{cases} \frac{\mathrm{d}}{\mathrm{d}t}P(t) = \alpha(t)P(t) \\ \alpha(t) = \int_0^A \left[\beta(a,t) - f_\mu(a,t)\right]\mathrm{d}a \\ P(0) = P_0 \end{cases} \tag{11}$$

we gets :

$$\begin{cases} \frac{\partial v}{\partial t}(a,t) + \frac{\partial v}{\partial a}(a,t) = -v(a,t)A(a,t) \\ \text{with} \quad A(a,t) = f_\mu(a,t) + \alpha(t) \\ v(0,t) = \int_0^A \beta(a,t)v(a,t)\mathrm{d}a \\ \int_0^A v(a,t)\mathrm{d}a = 1 \\ v(a,0) = v_0(a) \end{cases} \tag{12}$$

We can find the solution $v(a,t)$ of Equation (12) and $P(t)$ of Equation (11) then multiply them obtaining $p(a,t)$, the solution of MKVF equation.

7.1 Numerical implementation

Equation (12) can be solved using a 2^{nd} order Runge-Kutta method as described in [11] :

$$\begin{cases} V_{i+1}^{n+1} = V_i^n + K_1 + K_2, \quad i = 0, \dots, M-1 \\ K_1 = -hV_i^n A_{i+1/2}^{n+1/2} \\ K_2 = -\frac{h}{2}A_i^n K_1 \end{cases} \tag{13}$$

The boundary condition is approximated with a trapezoidal formula, as before :

$$V_0^{n+1} = \frac{h}{(2 - h\beta_0)}\left[2\sum_{k=1}^{M-1}\beta_k V_k^{n+1} + \beta_M V_M^{n+1}\right] \tag{14}$$

The coefficient A must be approximated in two different ways depending if we need A_i^n or $A_{i+1/2}^{n+1/2}$. First, A_i^n is approximated using a trapezoidal formula :

$$A_i^n = (f_\mu)_i^n + \alpha^n = (f_\mu)_i^n + \frac{h}{2}\left[\left[\beta_0 - (f_\mu)_0^n\right]V_0^n + 2\sum_{k=1}^{M-1}\left[\beta_k - (f_\mu)_k^n\right]V_k^n + \left[\beta_M - (f_\mu)_M^n\right]V_M^n\right] \tag{15}$$

Computing $A_{i+1/2}^{n+1/2}$ with this same method would require a value for $V_{i+1/2}^{n+1/2}$ which is not known. Another way is to approximate $V_{i+1/2}^{n+1/2}$ with half a step of Euler method which gives $V_{i+1/2}^{n+1/2} \simeq (1 - \frac{h}{2}A_i^n)V_i^n$. By using the midpoint method for the integral, we obtain :

$$A_{i+1/2}^{n+1/2} = (f_\mu)_{i+1/2}^{n+1/2} + h\sum_k \left(\beta_{k+1/2}^{n+1/2} - (f_\mu)_{k+1/2}^{n+1/2}\right)\left(1 - \frac{h}{2}A_k^n\right)V_k^n \tag{16}$$

And we need also to solve Equation (11) in order to obtain $p(a,t)$ at the end. Let's use a simple forward finite difference method :

$$\frac{P^{n+1} - P^n}{h} = \alpha^n P^n \implies P^{n+1} = (1 + h\alpha^n)P^n \tag{17}$$

These two numerical schemes can be run separately and then their results can be combined at the end.

7.2 Limitations

When testing this method however, we don't get coherent results. Let's take the simplified example with no death and no birth to highlight the main problem. The expected result for $p(a,t)$ is a simple shift to the right with time. We can indeed observe this result in Figure 22. As there is no death rate and no birth rate, the initial population should simply age with time and the individuals reaching 100 years should die, so the population is decreasing with time. Also we are expecting the normalised population repartition would increase for old age groups as the population gets older and is not replaced. In Figure 23 we can observe that the total population is actually constant and in Figure 24 that the normalised repartition doesn't increase. The problem comes from the evolution of the total population. Indeed, by looking at Equation (11), we see that in our case with $\beta(a,t) = 0$ and $f_\mu(a,t) = 0\ \forall a,t$, then the coefficient $\alpha(t)$ is zero and the total population stays constant. In other words, the result of the numerical model is coherent with the initial equations but this is not realistic. The hypothesis detailled in Equation (11) needs to be revised. A possible solution may be to define a boundary condition at the upper boundary of age $f_\mu(A,t) = 1, \forall t$.

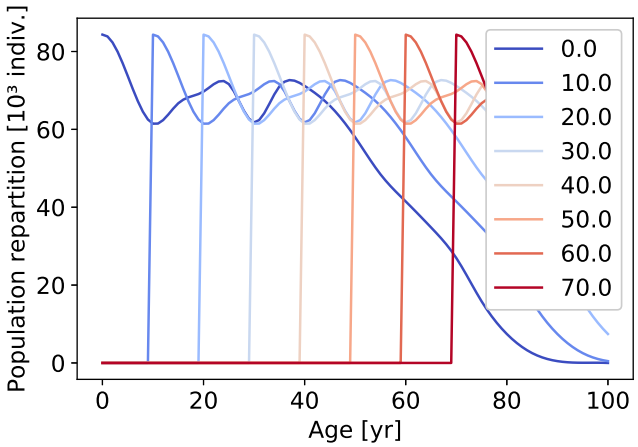

Figure 22: Population repartition in the simplified case with no death and no birth, simulated with the RK2 model.

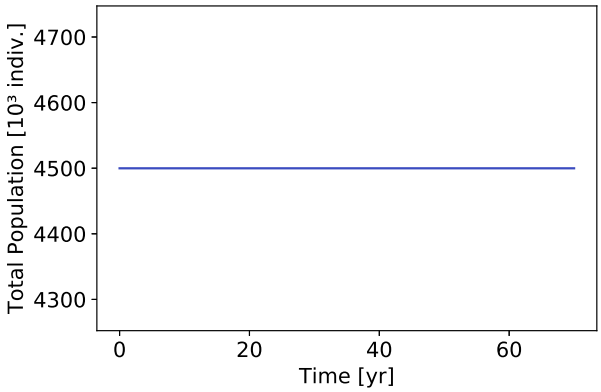

Figure 23: Total population in the case with no death and no birth, simulated with the RK2 model.

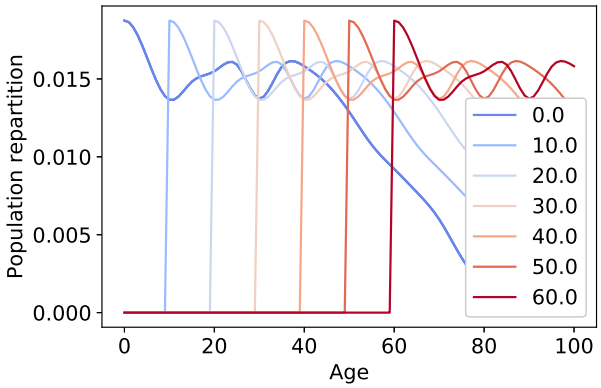

Figure 24: Normalised population repartition in the case with no death and no birth, simulated with the RK2 model.

8 Discussion and Conclusion

This work has shown me the importance of defining precisely the different indicators and their units. First, we must be careful when going from a continuous to a discrete space. The units may "change" or be interpreted differently. Especially when the data is separated into groups of a few years (5-year age group and 5-years time period), we must distinguish absolute from relative values. For example, population data is given in 5-years age groups so if we need the number of individuals of a specific age, we need to divide this number by 5. The problem is not encountered for ASFR or central death rate as these values are ratio, therefore they don't depend on the size of the time interval. Also, there can be some anomalies in the computations due to the discretization. For example, the computation of the death rate from the number of deaths and the number of individuals in a certain age group may lead to some values that are superior to 1. This result doesn't make sense mathematically but it can be partially explained by the fact that the data from the UN are actually estimations, they are not perfect. Another complexity is the exact definition of the different indicators. It was not always possible to find definitions of the indicators, even with the United Nations documentation. For example the difference between the ASFR and the PASFR (Percentage age-specific fertility rate) couldn't be figured out. Also, the central death rate seems to be an appropriate death indicators as we can see that the curves in Figure 18 are almost overlapping. However, there are discrepancies for younger age groups which could mean that the ASFR is not the best suited indicator for modelling the birth rate.

An idea of extension of this project would be to add the immigration and emigration fluxes. This would allow the model to be even more realistic. It could then be apply to any country, without the constraining hypothesis of negligible net migration. Another prolongation could be to make the alternative implementation work. Equation (11) should be revised in order to get coherent results. Finally, the main limitation concerns the input fluxes. Indeed, we need to know precisely the fertility and death rates in order to make good predictions. This represents a consequent amount of information in order to be able to simulate the evolution of the population. An amelioration would be to be able to guess the behaviour of these fluxes in the future, from past data. Fortunately, this task is easier to do than predicting the evolution of the population.

This project have shown that a simple model can already give significant results concerning the evolution of the population. Applying the model to the Philippines allows us to obtain realistic results, even without the modelling of emigration and immigration fluxes. The aim is not to perfectly describe these dynamics but rather to gain insight on their behaviour. This must be the first stone into a more comprehensive model. We can see that these approximations are good enough to allow us to predict the main tendencies of the population repartition.

References

Sources verified on May 27, 2020:

[1] A. Porporato and S. Calabrese, *On the probabilistic structure of water age*, Duke University, Durham (USA), 2015

[2] M. C. Drever and M. Hrachowitz, *Migration as flow : using hydrological concepts to estimate the residence time of migrating birds from the daily counts*, Methods in Ecology and Evolution (Vol.8, Issue 9), 2017

[3] P.L. Chen, D.J. Brenner, R.K. Sachs, *Ionizing radiation damage to cells: Effects of cell cycle redistribution*, Mathematical Biosciences (Vol.126, Issue 2), 1995

[4] *2019 Revision of World Population Prospects*, United Nations, `https://population.un.org/wpp/`

[5] *Handbook on the Collection of Fertility and Mortality Data*, Department of Economic and Social Affairs, United Nations, 2004

[6] J.R. Wilmoth, K. Andreev, D. Jdanov, and D.A. Glei, *Methods Protocol for the Human Mortality Database*, 2019

[7] A.R. Thatcher, V. Kannisto, and J. W. Vaupel, *The Force of Mortality at Ages 80 to 120*, Odense University Press, 1998

[8] Fabio A. Milner and Guglieimo Rabbiolo, *Rapidly converging numerical algorithms for models of population dynamics*, Università di Roma, Italy, 1990

[9] Paraskevi Peristera and Anastasia Kostaki, *Modeling fertility in modern populations*, Demographic Research (Vol. 16), 2007

[10] *Scipy documentation*, `https://docs.scipy.org/doc/scipy/reference/generated/scipy.interpolate.RectBivariateSpline.html`, 2019

[11] Galena Pelovska, Mimmo Iannelli, *Numerical methods for the Lotka–McKendrick's equation*, Universita degli Studi di Trento, 2005