Discrete Stochastic Approach to Modeling Population Age-Structure Change in South Korea

Chad Olsen

August 2024

1 Abstract

This article aims to Model the nation of South Korea's population demography using a stochastic approach using Leslie matrices. The matrix is comprised of random variables and the model uses their expected values to find the the stable population distribution. The problem with only using one year's birth rate in a Leslie matrix is that it allows for no variance, which is unrealistic. Utilizing a distribution that closely simulates birth and death rates, and finding the parameters for such distribution using Maximum Likelihood Estimators from the data over the past years, will allow the model to more closely predict the stable population. A stable population growth rate and structure are obtained by use of Gumbel Distributions for birth rates and Beta distributions for survival rates. From the data used it is shown that the stable population of South Korea at the current trend of birth rates will approach a population structure heavily skewed to older generations. It was also found that the structure of the nation's population will move towards a higher average age quite quickly. The model would work great to estimate population distribution in other cases, and if minor modifications are applied, predict population size.

2 Introduction

South Korea currently has the world's lowest fertility rate at a meager 0.72 babies per woman, which fell from 0.78 in the previous year. This would mean that the population will certainly decrease, likely at a steep rate, and the average age over the next few decades will become much higher. This poses a massive problem for the nation, such a low birth rate almost certainly means that a population collapse is imminent. For reference, a nation needs to maintain a birthrate of about 2.1 to simply maintain the current population. And with the majority of the population being close to or in retirement, the decreasing population of working-age individuals will likely not be enough to support them. This would cause a massive economic strain on the nation. An optimal age structure is a very important part of keeping a society intact, and deviations from this can cause the population to decrease even faster. The goal of this model is to show the stable population distribution based on data provided over the past few years on birth and death rates and how fast the population would converge to it.

To answer that question this model must be age-structured. Much of the recent work in demography comes in the form of stochastic models. These models utilize Leslie matrices, with birth and death rates being random variables. The population can be represented by a vector N consisting of $n_0, n_1, \ldots, n_m$ age-specific populations, and the change in population in time Δt can be represented by,

$$n(t+1) = A(N,t)n(t) \tag{1}$$

where the matrix A is the Leslie matrix. As shown in [2] the Leslie matrix can have time-specific perturbations, where each element of A can be defined as,

$$A_{i,j}(N,t) = a_{i,j}(g(n)) + \epsilon(t)i,j \tag{2}$$

where $g(n)$ is some arbitrary density-dependant function. Since these models utilize random variables for birth and survival rates, in [3] it was shown that equation 1 can be written as,

$$n_m = E(X_{t+1}|X_t = x_t) \tag{3}$$

which is representing that the population vector will be represented as its expected value. With that, we will use both equations 2 and 3 to come up with the Leslie matrix that uses the expected values of the random variables.

The previous work mostly used binomial and normal distributions for the birth and survival rates [4]. It is certain that these may be easier to use, and it is plausible that birth and survival rates could follow them. However, there may be better distributions to model these. My model will use structures similar to those of previous models, but it will aim to use more appropriate distributions. In a comparison of birth rate data and statistical distributions in Turkey, it was found that the Gumbel Distribution fits better than other common distributions [1]. Then for survival rates, this approach is going to use the beta distribution to allow the data of death rates to form the shape of the distribution. Using these distributions along with the data from previous years' birth and death rates, will allow the model to more accurately predict stable population. This is because, if the distributions are better for these specific variables, their expected value would be much more accurate than say a mean value or a value from one year.

The parameters of these distributions will be found using Maximum Likelihood Functions along with data provided by South Korea's Statistical Information Service. The parameters of the distribution will be formulated from the data from 2020-2022 as these more accurately represent the current state off the birth rates. Once the parameters are found for each age group, they then will then be used to fill in a Leslie matrix. And using the properties of the matrix's eigenvalues and eigenvectors, the nation's stable population distribution will be attainable. Using the matrix itself, the rate at which the population converges to stable population distribution will also be shown.

3 The Model

The model will utilize a Leslie matrix along with its eigenvalue properties. This will build on the previous work using Leslie matrices along with stochastic methods in

demography. Let $N_0(t), N_1(t), ..., N_m(t)$ be the populations for each age group $0 - m$ at a time t. And let $b_0, b_1, ..., b_n$ and $S_0, S_1, ..., S_n$ be the birth and survival rates for each age group respectively. From what was found earlier the best distribution to model birth rates is a Gumbel distribution, so we can define the birth rates for each age group as $b_0, b_1, ..., b_n \sim Gumbel(\mu, \beta)$. Where μ is the location variable, and β is the scale variable. Then a good distribution for survival rates in an aging population is the Beta distribution so $S_0, S_1, ..., S_n \sim Beta(\alpha, \beta)$ where α is the shape parameter or in this case, the hazard rate for a specific age group, and β would be the scale parameter or in this case the age of the characteristic life scale.

3.1 Finding $E[b_i]$

First looking at the random variable b_i the expected value is,

$$E[b_i] = \mu_i + \beta_i \gamma \qquad\qquad \gamma = 0.57721... \tag{4}$$

but we need to define μ_i and β_i for each age group from the birth rate data found. The process of doing this is actually quite simple. We can use a Maximum Likelihood Estimator(MLE), which will estimate the parameters of the distribution given the data.

To find the MLE for the Gumbel Distribution, the likelihood function is,

$$L(\mu, \beta) = \prod_{i=0}^{n} \frac{1}{\beta} e^{-(\frac{b_i - \mu}{\beta} + e^{-(\frac{b_i - \mu}{\beta})})} \tag{5}$$

which gives us the log-likelihood function,

$$l(\mu, \beta) = \ln(L(\mu, \beta)) = \sum_{i=0}^{n} \ln(\beta) - (\frac{b_i - \mu}{\beta} + e^{-(\frac{b_i - \mu}{\beta})}) \tag{6}$$

Then to find the maximum likelihood function we will find the derivatives with respect to both μ and β of the log-likelihood function in equation 6 and find the values of μ and β that satisfy the equations,

$$0 = \frac{\partial}{\partial \mu} \sum_{i=0}^{n} \ln(\beta) - (\frac{b_i - \mu}{\beta} + e^{-(\frac{b_i - \mu}{\beta})}) \tag{7}$$

$$0 = \frac{\partial}{\partial \beta} \sum_{i=0}^{n} \ln(\beta) - (\frac{b_i - \mu}{\beta} + e^{-(\frac{b_i - \mu}{\beta})}) \tag{8}$$

Then let μ_{MLE} and β_{MLE} be the values that make both of these equations true, and thus the values for the parameters that best fit previous data. These are solved for numerically by using the minimize function within the SciPy Library. Using these will provide the most accurate value for $E(b_i)$

3.2 Finding $E[S_i]$

To find the survival rate we will use a similar approach. Since $S_i \sim Beta(\alpha, \beta)$ we will need to use a MLE with the data of death rates for each age group to find the parameters k and λ. Then once those are found, we can find the estimated value by

$$E[S_i] = \frac{\alpha}{\alpha + \beta} \tag{9}$$

To find the MLE for the Beta distribution the likelihood function is,

$$L(\alpha, \beta) = \prod_{i=0}^{n} \frac{S_i^{\alpha-1}(1 - S_i)^{\beta-1}}{B(\alpha, \beta)} \tag{10}$$

where $B(\alpha, \beta)$ is the beta function defined as

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)} \tag{11}$$

which gives us the log-likelihood function,

$$l(\alpha, \beta) = \ln(L(\alpha, \beta)) = \sum_{i=0}^{n} \ln(\frac{S_i^{\alpha-1}(1 - S_i)^{\beta-1}}{B(\alpha, \beta)}) \tag{12}$$

Then to find the maximum likelihood function we will find the derivatives with respect to both α and β of the log-likelihood function in equation 11 and find the values of α and β that satisfy the equations,

$$0 = \frac{\partial}{\partial \alpha} \sum_{i=0}^{n} \ln(\frac{S_i^{\alpha-1}(1 - S_i)^{\beta-1}}{B(\alpha, \beta)}) \tag{13}$$

$$0 = \frac{\partial}{\partial \beta} \sum_{i=0}^{n} \ln(\frac{S_i^{\alpha-1}(1 - S_i)^{\beta-1}}{B(\alpha, \beta)}) \tag{14}$$

Then let α_{MLE} and β_{MLE} be the values that make both of these equations true, and thus the values for the parameters that best fit previous data. These are solved for numerically by using the minimize function within the SciPy Library. Using these will provide the most accurate value for $E(S_i)$

3.3 The Leslie matrix

As stated before the recurrence relation for each age group used in previous works can be represented by $n_m = E(x_{t+1}|X_t = x_t)$. In this case, this relation can be simplified to the expected death(or birth) rate multiplied by the population size. When implementing this relation in the Leslie matrix model similar to that of equation 2, the elements of the matrix will all need to be expected values of their respective random variable. As said before in the elements of the Leslie matrix can be defined as, $A_{i,j}(N, t) = a_{i,j}(g(n)) + \epsilon(t)_{i,j}$, but since ϵ is very small,

$$E[A_{i,j}(N, t)] = E[a_{i,j}(g(n))] + E[\epsilon(t)] = E[a_{i,j}(g(n))] + 0 \tag{15}$$

, meaning that we can ignore the perturbations in the birth rate for the sake of this model. [2] So for each element of the matrix we are going to define the birth and death rates with the expected values of their random variables. Or in mathematical terms,

$$E[a_{i,j}(g(n))] = E[b_i], \qquad j = 0, 0 \leq i \leq m \tag{16}$$
$$E[a_{i,j}(g(n))] = E[S_i], \qquad i - 1 = j \tag{17}$$

Now that the variables are thoroughly defined, the step function is as follows,

$$N(t + \Delta t) = AN(t) = \begin{bmatrix} E[b_0] & E[b_1] & \dots & E[b_m] \\ E[S_0] & 0 & \dots & 0 \\ 0 & E[S_1] & \dots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & E[S_{m-1}] & 0 \end{bmatrix} \begin{bmatrix} N_0(t) \\ N_1(t) \\ \vdots \\ N_m(t) \end{bmatrix} \tag{18}$$

Each expected value in the matrix A will need to be calculated using the methods referenced in sections 3.1 and 3.2.

4 Results

The data being used for this model will all come from South Korea's Statistical Information Service(KOSIS) [5]. Very detailed birth and death rates were given by age groups with intervals of 5 years. For this implementation, I am using data from 2020, 2021, and 2022 to try and represent current birth rates. This model is supposed to answer what the stable age distribution of the nation is and how fast will the current age distribution converges to it. For the sake of simplicity and the question that the model is supposed to answer, the age groups will only go up to 80 years old. This won't make too much of a difference if we are trying to show changes in age structure, but if one were to use this model to simulate change in population size, more age groups past 80 would be necessary, since people live longer than 80 years old. When calculating parameters, values in the data below 0.0001 were considered as 0.

The first step is to find the MLEs for each parameter of b_i and S_i. This process will need to be completed so that the distribution parameters for b_i and S_i have MLEs defined for every age group. As said in sections 3.1 and 3.2, these are numerical approximations. Each parameter's MLE is given in table 1.

Then using the MLEs as their respective distribution parameters, equations 4 and 9 can produce the expected value from each distribution for its respective age group. Then once we have the all of the values of $E[b_i]$ and $E[S_i]$ we can construct the Leslie matrix shown in figure 1.

Table 1: Gumbel and Beta distribution parameters

Age group	μ_{MLE}	β_{MLE}	α_{MLE}	β_{MLE}
0 years old	0	0	2873	7.3358
1 - 4 years old	0	0	2569	0.0046
5 - 9 years old	0	0	2569	0.0046
10 - 14 years old	0	0	2569	0.0046
15 - 19 years old	0	0	6556.8	1.4087
20 - 24 years old	0.0023	0.0003	7035.6	3.0270
25 - 29 years old	0.0123	0.0010	861.22	0.6220
30 - 34 years old	0.0353	0.0012	814.59	0.6538
35 - 39 years old	0.0214	0.0004	720.03	0.7209
40 - 44 years old	0.0035	0.0002	10638	12.31
45 - 49 years old	0	0	461.43	0.8600
50 - 54 years old	0	0	2780.0	8.2324
55 - 59 years old	0	0	2537.4	9.8310
60 - 64 years old	0	0	2766.4	15.21
65 - 69 years old	0	0	3691.3	31.37
70 - 74 years old	0	0	4040.3	57.42
75 - 79 years old	0	0	3507.8	87.50
80 Years or more	0	0	1781.3	161.3

$$\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & .0125 & .0645 & .1800 & .0216 & 0.108 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
.9822 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & .9991 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & .9995 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & .9991 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & .9987 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & .9976 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & .9971 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & .9969 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .9950 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .9940 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .9917 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .9886 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .9826 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .9732 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .9297 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .8648 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .7422 & 0
\end{bmatrix}$$

Figure 1: Leslie Matrix

Now that we have the Leslie matrix, we can now deduce the information needed to answer questions about the stable population structure. Firstly, for reference this matrix has the dominant eigenvalue $\lambda = 0.8907$, which is defined by the properties of the Leslie matrix as the stable population growth rate [4]. As you can guess since λ is less than one, the population is predicted to decrease at a very high rate. The dominant eigenvalue of a Leslie matrix being this low is very alarming fact in itself for a population.

Then using the dominant eigenvalue λ, we know by the properties of the Leslie matrix that its corresponding eigenvector V, represents the stable age structure of the population [4]. Note that V must be normalized to get the actual percentages for each age group. After normalizing V, we are given the stable age structure shown in Table 2, and its histogram in Figure 2.

Table 2: Stable age distribution

Age group	Stable Population %
0 years old	2.02
1 - 4 years old	2.23
5 - 9 years old	2.50
10 - 14 years old	2.80
15 - 19 years old	3.14
20 - 24 years old	3.53
25 - 29 years old	3.95
30 - 34 years old	4.42
35 - 39 years old	4.95
40 - 44 years old	5.53
45 - 49 years old	6.17
50 - 54 years old	6.87
55 - 59 years old	7.62
60 - 64 years old	8.41
65 - 69 years old	9.19
70 - 74 years old	9.59
75 - 79 years old	9.32
80 -84 years old	7.76

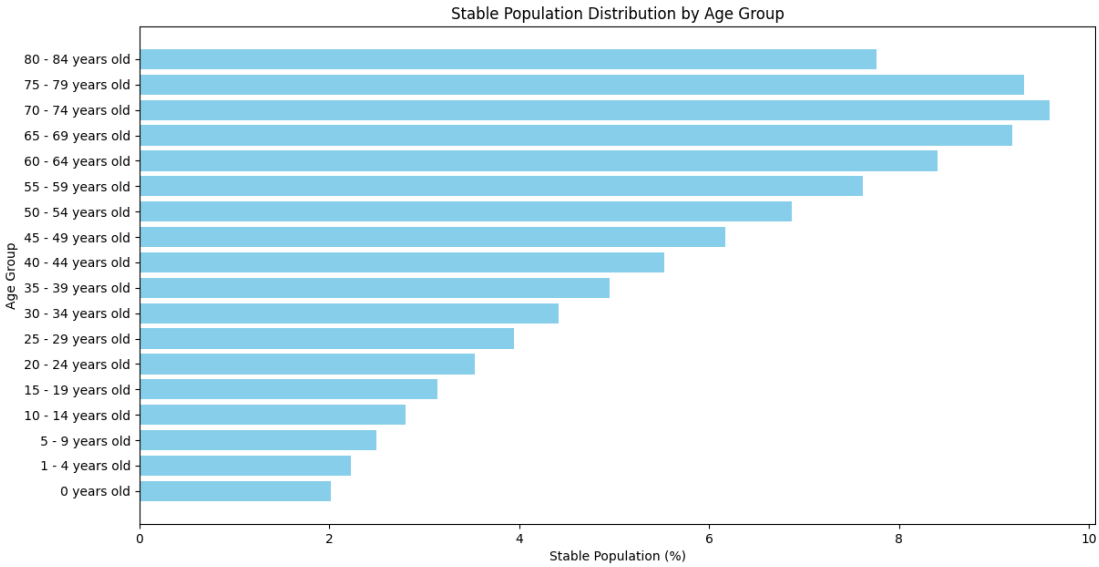

Figure 2: Stable population distribution

This answers the question of what the age distribution would converge to under current birth rate trends. As one can see with the small birth rate, the age distribution of

the population is heavily skewed toward the older generation. It would be expected that the distribution would be skewed given how low the birth rates are. However, the magnitude of the skew is quite extreme. From a social science standpoint, this would affect many factors of the nation. The economy would be under a large strain as the amount of working-age individuals is very small in comparison to those of retirement age. The workforce would likely not be big enough to support the entire population.

The model can also be used to show changes in structure over time rather than just deriving the stable population structure. To answer the question of how fast the population structure will move towards the right skewed structure show in in Figure 2, we can plot the average age in each time step of 5 years. The age structure after t time steps(5 years) can be easily calculated by $A^t N_0$ where A is the Leslie matrix shown in figure 1, and N_0 is the initial age structure, in this case 2022 age structure. Figure 3 graphs the average age using the 2022 population structure as time passes.

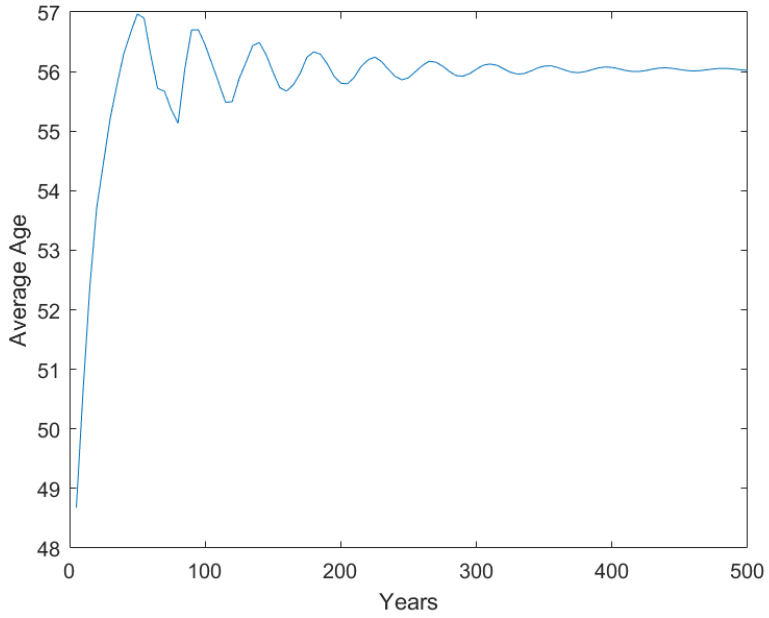

Figure 3: Average Age Change

If you look at the left side of the graph you notice how fast the average age increases. This would suggest that the current age structure will become much older in a short amount of time under the current birthrates. This means that if birth rates in the nation stay the same for much longer, the nation will be facing population collapse. As rebuilding a population with this age structure would be almost impossible. This highlights the danger of these birth rates for the nation, and how fast they can cause irreversible change in the population size.

5 Conclusion

The use of sampling distributions can greatly improve the Leslie matrix model. The Gumbel distribution estimates birth rates really nicely. It was found that the beta distribution may not have been the most ideal, but it still provided a fairly accurate simulation for survival rates. While the values attained from the estimated variables may have seemed fairly close to the sample mean of the data collected, if more data is used and more numerically sophisticated methods used to find the MLEs, a stochastic model of this nature will provide a much better simulation than simply using mean values.

Also if the model was implemented in one computer program, there would be no need to round calculations, and thus the actual value given would be much more accurate. It was shown that the Gumbel distribution is very accurate for birth rates, but if recreating this it might be worth considering other distributions for Survival Rates. The computer program that was written to approximate the MLEs, output a very exact number, however, to fit the information within this paper, numbers were rounded. This isn't as important when trying to find the stable age distribution, but if using a similar model to simulate population change, those decimal places are very important.

It should also be noted that future works can use this same method to estimate the actual population size. However, if doing this it is important that the age brackets of the population vector go all the way up to the maximum lifetime for the population. For the sake of this model, it wasn't really relevant to go past 80 years old, so we represented everyone above 80 as one age group. And by the the process of multiplying the Leslie matrix by the population vector, everyone in this group dies in the next time step. Which obviously would cause problems for estimating population size as people live well beyond 80 years old.

In terms of the problem for the Nation of South Korea, as you can see the stable population distribution shown in Figure 2 is heavily skewed to older generations, as well as a dominant eigenvalue less than 0. This means that the population will decrease at a large rate, but the more obvious issue is the structure of age. A very large of the population is in retirement age while the working-age individuals remain very small. This can cause massive economic and social issues for the nation. When a population structure is skewed like this, it suggests that the population will continue to decrease until it is restored to a more optimal one. Figure 3 really shows the severity of the current birth rates. It wont take much time for the average age to rise from 45 years old to 56 years old. This means that if the current trends in birth rates stay the same, the nation's age structure will grow towards one similar to figure 2 at a very fast rate. And once the population reaches this point it is very unlikely that they will be able to reach a positive population growth any time in the near future.

References

[1] Gamze ÖZEL Ceren ÜNAL. A comparison of statistical distributions for the crude birth rate data. *İstanbul Commerce University Journal of Science*, 2023.

[2] Engen S. Sæther B.-E Lande, R. Evolution of stochastic demography with life history tradeoffs in density-dependent age-structured populations. *Proceedings of the National Academy of Sciences - PNAS*, 2017.

[3] Åke Brännström Linnéa Gyllingberg, David J.T. Sumpter b. Finding analytical approximations for discrete, stochastic, individual-based models of ecology. *Mathematical Biosciences*, 2023.

[4] Fabio Milner Mimmo Iannelli. *The Basic Approach to Age-Structured Population Dynamics*. Springer, 2017.

[5] Korean Statistical Information Service. Live births by age group of mother, sex and birth order, 2023. data retrieved from KOSIS, https://kosis.kr/statHtml/statHtml.do?orgId=101tblId=DT$_1$*B*80*A*01*conn*$_p$*ath* = *I*2*language* = *en*.