A Replication for the Paper: General Age-Specific Mortality Model With an Example Indexed by Child Mortality or Both Child and Adult Mortality*

Yiran Huo

12/22/2020

Abstract

Population estimation could be viewed as an essential foundation for one country's strategy building. Most countries in Africa and almost one-third of all countries lack a mortality record; thus, they strongly rely on morality models to conduct population estimations and other demography and epidemiology tasks. In this paper, I will replicate Samuel J. Clark's mortality model in his paper 'A General Age-Specific Mortality Model with An Example Indexed by Child or Both Child and Adult Mortality' (2019), where Clark developed he singular value decomposition component model (SVD-Comp Model) based on previous researches' models. After building the model, I will use some countries as examples to discuss the SVD-Comp Model's performance with the log-quaternion model and conclude that the SVD-Comp Model is a better choice for the mortality model.

keywords: Mortality; SVD; HMD; SVD-Comp; Log-Quad

*Codes for this paper are available at: **github/selinahuo/STA3-4-PS5**. Codes for the original paper is available at **Code for A General Age-Specific Mortality Model with An Example Indexed by Child or Both Child and Adult** .

1. Introduction

1.1 Background

Population estimation is essential for all countries; it provides government policymakers with information to make future policy trends. Possible discussion, including forecast future demographic characteristics, make economic development strategy and health care projects. Having an accurate population estimation is essential. The United Nations Population Division biennially produced the World Population Prospects (WPP), which is generally considered the reference population indicators and is widely used by other domestic and international agencies as a critical reference. This prospect contains estimates of "time, sex, and age-specific mortality, fertility, and population size from 1950 to the present and forecasts of the same quantities to 2100 for all countries of the world" (Clark, 2019). Therefore, WPP updates specific age mortality across the period 1950 to 2100 (Clark, 2019). If one lacks civil registration and vital statistics, WPP is one's best option for discovering accurate fertility or mortality rate (Clark, 2019). Focusing on mortality, "50 countries around the world with a total population of nearly 1 billion people have no information on adult mortality, with the bulk of those in Africa—33 countries with a total population of 666 million people" (Clark, 2019). Consequently, mortality models are used to solve this problem and produce mortality schedules across all age groups.

From Clark's research, some countries in the world, particularly in Africa, lack information on child mortality or adult mortality. Covert this information into numbers; there are about 50 countries, nearly 1 billion population globally, and the majorities are from Africa, 33 countries with a total population of 666 billion people, that is lack of this information. How do these countries predict their populations? Under the absence of a full morality record in these developing countries, the Morality models are therefore brought to view. Most African countries require mortality models for both, and 38.6 % of countries globally require a morality model and 32.6 % for age-specific mortality (Clark, 2019). The standard approach to generating a successful mortality model for countries with insufficient data is based on the advantage, knowing data on child mortality to predict the overall mortality. By looking at the facts, the mortality model is useful and essential for predict on mortality model.

1.2 Paper Outline

Since this is a paper replication project, I will use the same data set as the original paper used, the Human Mortality Database Life Tables, and apply the same method as the original paper, the SVD-Comp Model. My paper will follow 4 sections. 1) In the introduction, I will provide the statistical background information about the SVD-Comp Model. 2) In the Methodology, I will explain the data, where I download it, how I clean it, and how I will apply this data in my model building. I will also explain the linear algebra theory behind the SVD-Comp Model. Then I will discuss the actual model and how we will examine the validity of this model. Lastly, I will use data sets from Mexico and South Africa for both SVD-Comp Model and Log-Quad Model and compare the result. 3) As a result, we will discuss the result of this model. We will also look at the validation of this model and the SVD-Comp Model's performance, and the Log-Quad Model in Mexico and South Africa data set. 4) In the discussion, I will analyze the result and discuss why SVD-Comp Model could be viewed as the "best" morality model. I will also discuss how this

1.3 Model Background

History of Mortality Model Traditional model life tables, as well as some fertility models, take an inductive and empirically driven approach to identify the regularity of mortality based on age across an extensive collection of high-quality life tables. Alternatively, some parametric or functional-form mortality models take a deductive approach (Gompertz 1825; Heligman and Pollard 1980; Li and Anderson 2009; Makeham 1860). More specifically, these models are defining "age-specific measures of mortality in an analytical form, sometimes with interpretable parameters" (Clark, 2019). More recently, Wilmoth et al. (2012) developed the Log-Quad model that combines empirical and functional approaches to mortality models.

As mentioned before, population forecasting can be estimated by using age-specific mortality and fertility data. In population forecasting, since there are various dimensions of one problem, dimension-reduction

or data-compression techniques are required. More specifically, there is a need to maintain the necessary parameters representing age schedules and their dynamics. The first researchers that apply this strategy to summarize age-specific morality and generate model life tables are Ledermann and Breas (1959) by using PCA. They could be viewed as the founders of this approach, and later, many researchers follow their methods. If we take a closer look at the PCA method, we see that SVD has been widely using for building the population forecasting model based on mortality and fertility.

The Lee-Carter approach (Lee 1993; Lee and Carter 1992) has been widely used in demography. The model as presented in Lee and Carter (1992) is

$$ln(m_{age\times time}) = a_{age} + b_{age}(\text{a column vector})^{'} + \epsilon_{age\times timet} \tag{1}$$

where a is the time constant of mean over m columns; b is the time constant first left singular vector from SVD decomposition. This model could be viewed as a simplified version of the more complex age-period-cohort mortality model of Wilmoth. The purpose of Wilmoth's model is to separate and identify age, period, and cohort effects in an age and time matrix of mortality rates. The basic structure is the following : log(m_x) = (mean model) + (residual model).

In 21st centuries, another method has been brought to attention: the log-quadratic (Log-Quad) model.

$$log(m_x) = a_{age} + b_{age}h + c_{age}h^2 + v_{age}k_i \tag{2}$$

Where a, b, c are constant age-specific coefficients for the quadratic mean model; h is the input value of $log({}_5q_0)$; v is an age-specific correction factor; k is a coefficient for v. The correlation factor, v_x, is "identified by calculating the SVD of the matrix of residuals that remain after the quadratic portion of the model is subtracted from life tables that are part of the Human Mortality Database (HMD) ... and using the resulting first left singular vector as a starting point" (Clark, 2019). The Log-Quad model could be viewed as an innovative new model with maintaining all the advantages of the empirically observed relationship between child mortality and other ages mortality. More specifically, in this method, the Log-Quad model has a familiar mean from the original Wilmoth model, and the structure of the residual model based on the SVD form by Wilmoth. This model is welly performed and accurately represent across all age of life tables (Wilmoth et al. 2012).

Clark's model Other matrix-summary approaches also exist to characterize the variability in mortality rates, but the Lee-Carter model is the most widely used. Based on these various discoveries in matrix factorization methods and the fast Fourier transforms in image compression, Clark developed a new model (Clark, 2001). "The component model is a simple linear sum of independent, age-varying vectors (components) that, when combined with appropriate weights, can closely approximate age-specific mortality schedules"(Clark, 2019).

This model is similar to Ledermann's factor analysis method: building a system based on factors resulting from a PCA decomposition of a matrix of age-specific mortality rates (Ledermann, 1969) and the PCA-based model from the U.N. model life tables (United Nations, 1982). This approach is a perfect combination of a simple linear model with PCA and SVD, which concentrate information along a few dimensions (Clark, 2015).

$$m = \sum_{i=1}^{c} w_i u_i + r \tag{3}$$

where m is a vector of age-specific mortality rates; u_i is a set of c vectors containing different age values; w_i is weight; r is a residual vector. The parameter, u_1, effectively means age-specific mortality, and its weight reflects the overall level of mortality. It follows directly from the properties of the SVD and an interpretation of singular vectors when applied to demographic age schedules (Clark, 2015). In addition, various parameters can be used in different models: "clustered using objective clustering methods to identify groups of similar age schedules, estimation using either traditional or Bayesian methods, or predicted from covariates that vary systematically with age schedules, as this article demonstrates" (Clark, 2019).

I will follow Clark's paper to replicate and explain how the SVD is used for developing a general modelling framework for demographic age schedules.

2. Methodology

2.1 Data

Human Mortality Database Life Tables (HMD) The data set used in this analysis is the HMD data were downloaded on Tuesday, December 15, 2020, from the HMD website (http://www.mortality.org/hmd/zip/all_hmd/hmd_statistics.zip). This data set contains validated information on deaths and "exposure from several mainly developed countries,"where death registration and census data are virtually complete" (Clark, 2019). Since the data is aggregated in various formats, according to this report's objective, building the morality prediction model, the standard life table for each year will be the take a closer look.

Data Cleaning The data includes mortality information for each year across 49 countries. We begin with download the HMD by using a set of functions. The initial data will be a zip file. By unzip the data file, we received three files separated by genders: It_both for female and male, It_female for female, and It_male for male. In detail, the data files are separated by the age period. For example, 1x1 is a single calendar year by a single year of age, and 5x1 is a five-year age group by a single year period. Then we want to separate the files into two separate RData files: 1x1 RData and 5x1 RData. We want to get each year or period of probabilities of dying and life expectancies in the format of age x life table matrices and save it as an RData directory separated by countries. There are 6 columns and 2 data collection methods (by single year or 5 years). Therefore, we received 18 RData files. Then, we will clean the data. We want to remove the 'NA' value in our life table matrices and the same rows. The same country dataset is supposed to be the same length (same number of rows); after cleaning the data, we then compare the number of rows of the female, the male file to confirm the value of the number of rows are matched.

hmd 1x1 data example:

```
print(hmd.1x1.list[[1]]$AUS[2,])
```

```
## # A tibble: 1 x 10
##   period age       mx     qx    ax    lx    dx    Lx      Tx    ex
##   <chr>  <chr>  <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl>   <dbl> <dbl>
## 1 1921   1     0.0121 0.0120   0.5 94250  1130 93685 6221704  66.0
```

hmd 5x1 data example:

```
print(hmd.5x1.list[[1]]$AUS[2,])
```

```
## # A tibble: 1 x 10
##   period age        mx     qx    ax    lx    dx     Lx      Tx    ex
##   <chr>  <chr>   <dbl>  <dbl> <dbl> <dbl> <dbl>  <dbl>   <dbl> <dbl>
## 1 1921   1-4   0.00602 0.0237  1.38 94250  2234 371152 6221704  66.0
```

2.2 Mortality Model

Firstly, I want to discuss the definition and the relationship between principal components analysis (PCA) and singular value decomposition (SVD). Then, I will introduce the background information about the research history in the mortality model, explain why choosing this model, and how it is being "invented". After that, I will discuss different approaches to the mortality model and explain the model chosen by Clark.

Principal Components Analysis (PCA) and Singular Value Decomposition (SVD) PCA is commonly used to extract central themes under the condition by analyzing the dataset columns. Statistically speaking, PCA is using Bayesian models for the matrices and applying estimation. Whereas SVD focuses on simplify data sets, analyzing two different themes, and then by eigenvalue decomposition, this method will produce a characterize result. SVD could be viewed as a tool for getting the result, PCA. There are some characteristics of the SVD. The matrix X in the following function has a particular property.

$$X = USV^T \tag{4}$$

U is a matrix of the left singular vectors (LSV), and V is a matrix of the right singular vectors (RSV). S is a diagonal matrix of singular value (SV). Clark has explained this function: "The LSVs and RSVs are independent and have unit length. If one view X columns as a set of dimensions, then the rows of X locate points defined along those dimensions—the data cloud. The RSVs define a new set of dimensions that line up with the axes of most variation in the data cloud. The first RSV points from the origin to the data cloud, or if the cloud is around the origin, then it points along the line of maximum variation within the cloud. The remaining RSVs are orthogonal to the first and each other and line up with successively less variable dimensions within the cloud. The elements of the LSVs are values that correspond to the projection of each point along the new dimensions defined by the RSVs. The SVs effectively stretch the new dimensions defined by the RSVs, following the cloud variation along each RSV. The numeric value of each SV is the square root of the sum of squared distances from the origin to each point along the corresponding SVD dimension, and their squares sum to the total sum of squared distances from the origin to each point along all of the original dimensions." (Clark, 2019)

This function could be expressed into new expressions:

$$x = \sum_{i=1}^{Q} s_i v_{li} v_i^T \tag{5}$$

$$x_l = \sum_{i=1}^{Q} s_i v_{li} u_i \tag{6}$$

"Because the RSVs define successively less variable dimensions in the data cloud, the first term in the functions contains the most information, and subsequent terms contain less and less (Golub et al. 1987). Including all ρ terms replicates the original data matrix X or any of its columns x_l exactly while including only the first few terms provides a good approximation." (Clark, 2019)

Model Scales We are having two datasets: a single year dataset and five-years dataset. Our purpose is to conduct an analysis baed on life table probabilities of dying. For single year prabilities will be denote as ${}_1q_x$ and the prabability for who survive will directly taken from the original dataset.Five year probabilities will be denote as${}_5q_x$ representing the child mortality and calculate as ${}_5q_x = 1 - \prod_{a=x}^{x+4}(1 -_1 q_a)$, and ${}_{45}q_{15}$ is the adult mortality and is calculated as ${}_{45}q_{15} = 1 - \prod_{a=15}^{59}(1 -_1 q_a)$.

Since we are using the real time data, logit transformation for keeping the data ontime is needed. We applied a logit function

$$logit(x) = \ln\left(\frac{x}{1-x}\right) \tag{7}$$

to make sure the probabilities of dying 'q' can occupy the full real time. Also, we want 'q' could be transformed back too in a probability sclae of [0,1]. So we also have a expit function, which is an inverse of the

$$logit : expit(x) = \frac{\mathrm{e}^x}{1+\mathrm{e}^x} \tag{8}$$

2.3 Method

SVD Component Model (SVD-Comp Model) The methodology involves building the models and checking the validity, which in total consists of five steps: 1. Calculate/estimate an SVD-Comp Model by using age-specific ${}_nq_x$ as inputs 2. Calculate/estimate a smoothed SVD-Comp model by using age-specific ${}_nq_x$ as inputs 3. Randomly sample age-specific ${}_nq_x$ as inputs, and calculate/estimate a smoothed/un-smoothed SVD-Comp model that predicts ${}_nq_x$ for the not-sampled age-specific ${}_nq_x$ 4. Summarize the prediction errors in step 3 5. Repeat step 1 to step 4 for n times (where $n \in N$)

After these five steps, we will receive a return object containing very detailed results for everything requested.

As previously mentioned, we will use the two functions of SVD. In a $A \times L$ matrix, Q, using the resulting factors as in

$$x_l = \sum_{i=1}^{Q} s_i v_{li} u_i$$

what we received is that each A-element mortality schedule, q_{zl}, is approximated as the c-term sum,

$$q_{zl} \approx \sum_{i=1}^{c} v_{zli} \times s_{zi} u_{zi} \tag{9}$$

Where A is the number of age groups and rows in Q_z; L is the number of life tables and columns in Q_z; $z \in (female, male)$; $c \leq p$, the rank of Q_z; $l \in (1...L)$.

Validation We will check the validity in two ways: compare the result with the Log-Quad Model, and check the produced errors. 1. We will compare the performance of the SVD-Comp Model and the Log-Quad Model based on the code that is published and provided by Wilmoth et al. (2012) to see if our model predicts adequately. Since Log-Quand Model only predicts mortality in the five-years age group, I will transform the singular year inputs to five-years values. From there, I will make a comparison. 2. We will summarize the overall errors produced by each model across every mortality schedules in HMD. More specifically, we will look at the absolute value of each year-, sex-, and age-specific error. After that, we will sum the errors across ages and years for each sex and conclude the result as a single number. This single number,e_o, indicates the overall differences between the predicted and actual values for all years and ages.

Application to Mexico and South Africa We will also use both models, SVD-Comp and Log-Quad, to predict age-specific mortality rates for Mexico in 1983-1985 and South Africa in 2005 based on inputs: both child and adult mortality. Data for Mexico is collected from the Human Life Table Database (Max Planck Institute for Demographic Research et al.); data for South Africa is collected from the World Health Organization's Global Health Observatory data repository (World Health Organization). For the modelling and the result, we will discuss the limitation and possible solutions in the following section, Discussion.

3. Result

We will look at the methodology and discuss our findings. In the figure of age (years) versus ${}_1q_x$ (logit scale), we use ${}_1\mathrm{q}_x$ on the logit scale with ${}_5\mathrm{q}_0$ as inputs. Overall, we could see that Sweden in 1751 consists of a higher mortality rate than Austria in 1990. There exists an extremely high death rate for infants in both countries, and after age 5, the death rate drops and becomes steadily increasing. However, in Austria 1990, the death rate is deeply increasing throughout the teenage, and in the age group of youth, especially in the age range of 15-20, the death rate becomes stability increasing in a positive linear relationship.

In the figure of age (years) versus Scaled LSV Values (logit scale), we could view that every s_1u_1 is negative, which implies that all $s1_u1$ captures the underlying average shape of the mortality profile with age. This figure displays a smooth version of the scaled LSVs; by looking at the figures, we could see that fewer variables provide a smoother graph. The more variable one graph consists of, the more "rough" it will be.

Result of Validation

Then, we will look at the errors as we discussed before. In our error distribution section, we could see that the differences between the total absolute errors for the two models in both models (Log-Quad and SVD-Comp) are in a propositional form: $((Log - Quad) - (SVD - Comp)) \div SVD - Comp$, which implies that the SVD-Comp model provides fewer errors, which can be closer to the HMD.

Result of Application to Mexico and South Africa

In our application to Mexico, we conclude that both Log-Quad and SVD-Comp Models successfully produce Mexico's same predictions. However, the situation for South Africa is different. Since South Africa's data does not consider the case of HIV, the prediction for South Africa is not accurate. Both of the model failed the predictions for South Africa.

4. Discussion

4.1 Summary

Various conditions or parameters could influence mortality models mortality; researchers always want to find a "perfect" model covering all conditions. Generally speaking, Clark's paper introduces the history of research in the mortality model, the fundamental linear algebra method of SVD, the SVD-Comp model, the model application in Mexico and South Africa, and the advantages as well as limitations for this model. By providing the background information and discussing the SVD-Comp model, we can conclude that this model's performance is as expected.

4.2 Conclusion

Based on our previous findings and our knowledge of SVD, we conclude that the SVD-Comp model consists of 4 advantages:

1. A Simple Linear Structure
2. A General Interface
3. An Ability to Handle Arbitrarily Age Groups Without Alter the Fundamental Structure of the Model
4. An Inherent Constraint that Ensures Mortality

More detailed, the SVD-Comp model generally "allows all-age mortality schedules... to be predicted from any covariates that are related to age-specific mortality" (Clark, 2019). SVD-Comp Model could be viewed as the most accurate, well-rounded, and successful mortality model in population estimation. This model could fully replace the Log-Quad Model, "which would give it the ability to model changing levels and age patterns of mortality independently and generally be more flexible" (Clark, 2019). The general relationship in this model is simple: just age and weights. When the variable weight is calibrated using the relationship between the empirical weights and the child mortality, under this condition, this SVD-Comp model could be viewed as the same as the Log-Quad Model. By comparing these two models, we know that the SVD-Comp model performs better. The weight is a public interface. Inputs can affect the age pattern of mortality, but not the weight. By looking at our model result of Mexico and South Africa, we observed that the Log-Quad Model could not reproduce the bulge, whereas the SVD-Comp model does.

The SVD-Comp Model and The Log-Quad Model

The SVD-Comp model makes predictions in a single-year age group, whereas the Log-Quad Model is based on the five-year age groups. The complexity and model building cannot be compared since they build differently. The significant difference between these two models is: the Log-Quad Model does not directly constrain the relationship of mortality for a specific age; the SVD-Comp model uses a sing-year age group and manipulates a linear combination of vectors in its model building, which constrained the relationship between ages.

4.3 Weakness & Next Steps

As we mentioned before, both of the models, the Log-Quad model and the has failed the prediction for South Africa due to the influence of HIV. This failure reveals a fundamental limitation of all empirically based mortality models: all models cannot represent mortality age profiles that are fundamentally different from those contained in the data used to create them. Potential solutions to this problem could be: having specific life tables representing the age profiles and using that new data files for model building.

In addition, we select Mexico and South Africa as examples because we know that Mexico will perform as expected, and South Africa contains other uncontrollable situations, HIV. In the South Africa case, even we believe the SVD-Comp model could replace the Log-Quad Model, we still observe that both models cannot reproduce the HIV related mortality bulge at adult ages. More detailed, the SVD-Comp model produced plausible mortality schedules for all sexes that were as close as possible to South Africa's, given that it could not reproduce the bulge. In contrast, Log-Quad Model produced a plausible mortality schedule for males but a nonsensical schedule for females. These results reveal an urgent need to increase the diversity of mortality schedules available in freely accessible archives, such as HMD, and in particular, a critical need to compile

much better mortality data for Africa and other developing world regions where age schedules of mortality are different from what has been observed in the developed world. Further works could be research into how to avoid bulge like the HIV situation in South Africa? Moreover, we could generalize the limitation to phrase a broader question: Are there any situations that this model cannot cover? However, since models are built based on databases since the data are limited, inaccuracy is reasonable and will be expected. By having high-quality empirical mortality data sets, models may reveal a more accurate result.

5. References

Bell, William R. 1997. "Comparing and Assessing Time Series Methods for Forecasting Age-Specific Fertility and Mortality Rates." Journal of Official Statistics 13. s Brass, William. 1971. "On the Scale of Mortality." In Biological Aspects of Demography, edited by William Brass, 69–110. Taylor; Francis: London, UK.

Clark, S. J. (2001). An investigation into the impact of HIV on population dynamics in Africa (Doctoral dissertation). University of Pennsylvania, Philadelphia, PA. Retrieved from https://repository.upenn.edu/dissertations/AAI3031652, and available at http://www.samclark.net/pdfs/clark-2001_phd_an-investigation-into-the-impact-of-hiv-on-population-dynamics-in-africa.pdf

Heligman, L., & Pollard, J. H. (1980). The age pattern of mortality. Journal of the Institute of Actuaries, 107, 49–80.

Hlavac, Marek (2015). stargazer: Well-Formatted Regression and Summary Statistics Tables.

Lee, R. D. (1993). Modeling and forecasting the time series of U.S. fertility: Age distribution, range, and ultimate level. International Journal of Forecasting, 9, 187–202.

Lee, R. D., & Carter, L. R. (1992). Modeling and forecasting U.S. mortality. Journal of the American Statistical Association, 87, 659–671.

Ledermann, S. (1969). Nouvelles tables-types de mortalité [New standard mortality tables] (Travaux et Documents No. 53, Institut national d'études démographiques). Paris: Presses Universitaires de France.

Ledermann, S., & Breas, J. (1959). Les dimensions de la mortalité [The dimensions of mortality]. Population (French ed.), 14, 637–682.

Makeham, W. M. (1860). On the law of mortality and the construction of annuity tables. Assurance Magazine, and Journal of the Institute of Actuaries, 8, 301–310.

Max Planck Institute for Demographic Research, University of California, Berkeley, & Institut d'Études Démographiques (INED). Human life table database [Data set]. Retrieved from https://www.lifetable.de/data/hld.zip

World Health Organization. Global Health Observatory data repository [Data set]. Retrieved from http://apps.who.int/gho/data/?theme=main&vid=61540