

# Bayesian Modeling the U.S. Mortality Dataset

Shuo Yang

A thesis submitted in partial fulfillment

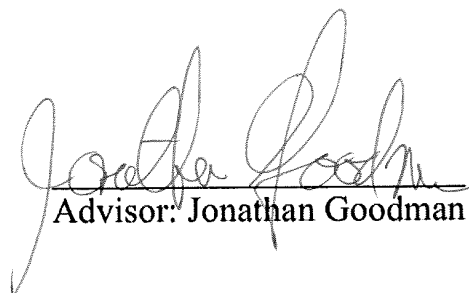
Of the requirements for the degree of

Master of Science

Courant Institute of Mathematical Sciences

New York University

January 2017



Advisor: Jonathan Goodman

# 1 Introduction

Mortality forecasts are of significant importance in the fields of life insurance and pension fund. They serve as the foundation for annuity calculations. The stochastic mortality model uses the historical death rates as its input and produces the future mortality projections.

Death rates are estimated from mortality databases. This is done by dividing death counts by exposures-at-risk. In the following,  $t$  represents a calendar year, and  $x$  stands for the age of the underlying group. Exposures-at-risk,  $E(t, x)$ , is defined by the number of people whose deaths would be counted in  $D(t, x)$ . Then, the death rate,  $m(t, x)$ , is equal to  $D(t, x)/E(t, x)$ . In the majority of mortality databases, exposures are estimated by averaging the beginning and end of the year populations or by taking the mid-year population. Using these estimation methods leads to errors whenever there is a dramatic increase or decrease in the size of populations due to birth or migration.

The dataset under study in this paper is the U.S. male mortality dataset published by the Human Mortality Database (HMD). According to the Methods Protocol of the HMD, for  $0 < x < 80$ ,

$$E(t, x) = \frac{1}{2} [P(t, x) + P(t + 1, x)] + D_{adj}(t, x)$$

where  $P(t, x)$  is the number of the persons who aged  $x$  last birthday on January 1st of year  $t$  and  $D_{adj}(t, x)$  is a small correction that reflects the timing of deaths during year  $t$  [1]. Due to the surge of birth rate in the U.S. in 1946 [2], the calculation for exposures-at-risk is unable to provide accurate values for the exposures-at-risk of the cohort born in 1946.

We adopt the methods proposed by Cairns et al. [3] to study the U.S. male mortality dataset published by the HMD. As our goal is to better understand the longevity risk, we are particularly interested in the datasets containing  $E(t, x)$  and  $D(t, x)$ , where  $x$  ranges from 41 to 100 and  $t$  ranges from 1960 to 2014<sup>1</sup>. In section 2, we present the plots generated by a graphical diagnostic method. This diagnosis enables us to detect the uneven pattern within the mortality dataset. The plots indicate the mortality data of the cohorts born in 1919-1920 and 1943-1947 exhibit irregular behaviors. Historically, the 1918 flu pandemic and the end of World War II caused the unsteady birth rate in the U.S. In section 3, a Bayesian model is developed

---

<sup>1</sup>As of January 2017, the U.S mortality data in the HMD is updated through 2014.

to quantify the errors within the exposures-at-risk. The population data of time intervals of less than one year is not always available and the resulting errors within the exposures, which may involve various complexities, cannot be simply removed through the current methods that create them. The Bayesian model designed using the techniques of time series analysis and posterior simulation is able to overcome these challenges and output the numerical values of the errors within exposures. In section 4, analysis of the impact of the modifications of exposures on mortality forecasts and annuity calculations is given. The proportion of survivors of the cohort born in 1947 over a period of 30 years is approximately 10% higher by using the modified exposures.

## 2 Graphical Diagnostics

The Gompertz law, the simplest version of Gompertz-Makeham law of mortality, provides a fairly good fit to mortality data over some age ranges, particularly from middle age to early old age [4]. Under Gompertz law, the human death rate at age  $x$ , denoted by  $m_x$ , is equal to  $Bc^x$ , where  $B$  and  $c$  are constants such that  $0 < B < 1$  and  $c > 1$  [4]. Thus,  $\log m_x$  is a linear function of age  $x$ . We are able to verify this linearity through various plots. The deviations from the expected structure suggest the underlying uneven patterns.

Assume within a calendar year the underlying log death rates of consecutive ages are approximately linear. We begin by studying the observations of three consecutive ages within a calendar year. Mathematically,

$$C(t, x) = \log m(t, x) - \frac{1}{2}(\log m(t, x - 1) + \log m(t, x + 1)) \approx 0$$

A couple of green, blue, or mixed-colored diagonal lines appear in the 2-Dimensional level-plot of  $C(t, x)$  (Figure 1). Since each diagonal line in the plot reveals the information of the unique cohort it represents (coordinate  $(t + k, x + k)$  gives the information of  $t - x$  Cohort, the cohort born in Year  $t - x$ ), we can detect that the data information of 1919 - 1920 Cohorts and 1943 - 1947 Cohorts stand out. Additionally, mixed-colored horizontal lines at Age 50, 60, 70 and 80 are noticeable, too. That happens because ages (if unknown) are likely to be rounded at death or immigration registration. Figure 2 lists individual concavity plots of some cohorts. The dots in 1925 Cohort disperse randomly and are close to the horizontal line stands for value zero. Nonetheless, this desired phenomenon does not appear in the rest of displays - the dots are either generally above the zero line or below.

The uneven pattern occurs due to the anomalies contained in the mortality dataset. Similar to the fact computing  $E(t, 0)$  by averaging  $P(t, 0)$  and  $P(t + 1, 0)$  will cause the loss of accuracy (Figure 3), errors creep into the exposures during the calculation. Mathematically, more subintervals are in need to estimate  $E(t, x) = \int_0^1 P(t + s, x)ds$ . Collecting quarterly or monthly population data and/or implementation of compulsory birth/death registration system will improve the situation.

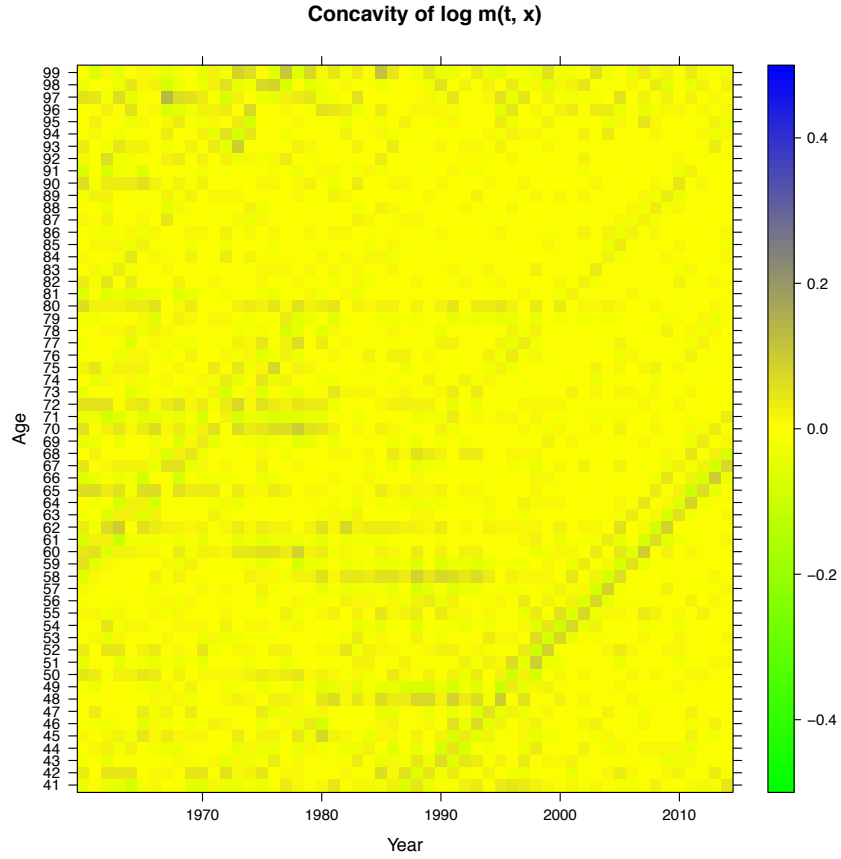


Figure 1: 2-Dimensional Level-plot of  $C(t, x)$

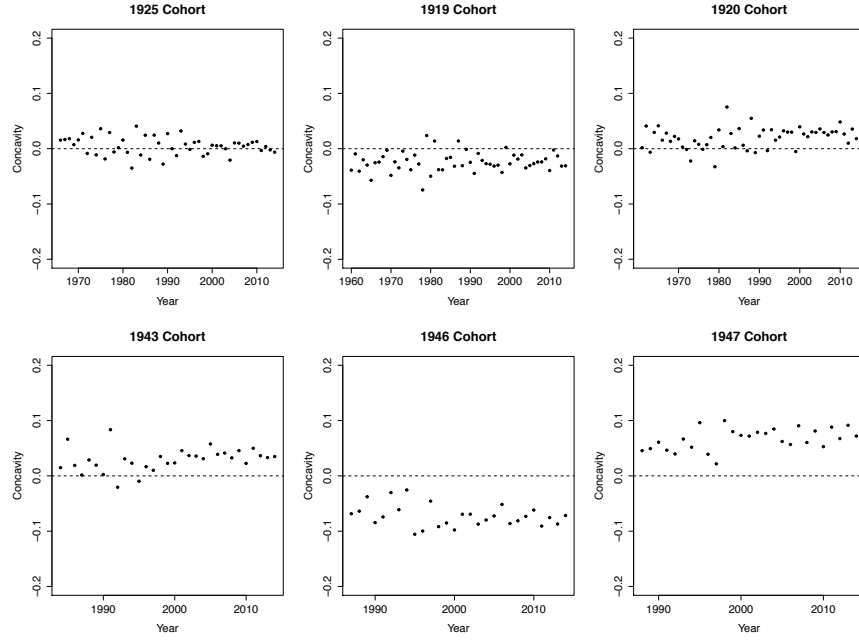


Figure 2:  $C(t, x)$  by Cohort

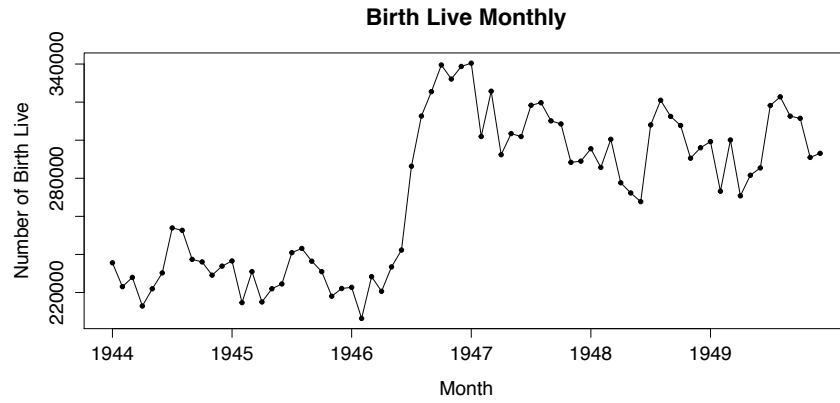


Figure 3: Monthly birth data extracted from Vital Statistics of the United States. Waving (increasing, decreasing, and convex) population invalidates the exposure estimation.

### 3 Quantification of Errors in Exposures

Given the published death counts and exposures, we aim to find the errors in exposures so that the “true” death rates, each of which is calculated by the modified exposure dividing the corresponding published death count, follow a desired pattern. In pursuit of this goal, a posterior density function is successfully constructed, and the Markov chain Monte Carlo (MCMC) method enables us to generate samples from this posterior probability.

#### 3.1 Notation

- Let  $\hat{E}(t, x)$  denote the true exposure. Assume  $D(t, x)$  is accurate. Then,  $\hat{m}(t, x) = D(t, x)/\hat{E}(t, x)$  is the true death rate.
- Define  $\epsilon(t, x) = \log E(t, x)$ ,  $\hat{\epsilon}(t, x) = \log \hat{E}(t, x)$ , and  $\phi(t, x) = \hat{\epsilon}(t, x) - \epsilon(t, x)$ .
- Define  $d(t, x) = \log D(t, x)$  and  $Y(t, x) = \log \hat{m}(t, x)$ .
- Age  $x$  ranges from  $x_0$  to  $x_n$ , and Year  $t$  ranges from  $t_0$  to  $t_n$ .
- Define  $\phi(t, x) = [\phi(t, x_0), \phi(t, x_0 + 1), \dots, \phi(t, x_n)]'$ .

#### 3.2 Assumptions and the Corresponding Models

**Assumption 1:** Errors in exposures within each individual cohort propagate through time.  $\phi(t)$  is thus assumed to follow a vector autoregressive process of order 1,  $VAR(1)$ .

For  $t = t_0 + 1, \dots, t_n$  and  $x = x_0 + 1, \dots, x_n$ ,

$$\phi(t, x) = \theta\phi(t-1, x-1) + \sigma_\phi Z_\phi(t, x)$$

where  $0 < \theta < 1$ ,  $\sigma_\phi > 0$ , and  $Z_\phi(t, x) \sim N(0, 1)$ .

For  $t = t_0 + 1, \dots, t_n$  and  $x = x_0$ , assume  $\phi(t, x_0) = \theta\phi(t-1, x_0) + \sigma_\phi Z_\phi(t, x_0)$ .

Thus, for  $t = t_0 + 1, \dots, t_n$ ,  $\phi(t)$  has a multivariate normal distribution with

mean  $A\phi(t-1)$  and covariance matrix  $V_\phi$ , where

$$A = \begin{bmatrix} \theta & 0 & 0 & \dots & 0 \\ \theta & 0 & \dots & & \\ 0 & \theta & 0 & \dots & \\ 0 & 0 & \theta & 0 & \dots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & 0 & \theta & 0 \end{bmatrix}$$

and

$$V_\phi = \begin{bmatrix} \sigma_\phi^2 & 0 & \dots & 0 \\ 0 & \sigma_\phi^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \sigma_\phi^2 \end{bmatrix}$$

Therefore, for  $t = t_0 + 1, \dots, t_n$ ,

$$f_1(\phi(t)) = \log p(\phi(t)) = -\frac{1}{2}(\phi(t) - A\phi(t-1))' V_\phi^{-1} (\phi(t) - A\phi(t-1)) + \text{a constant}$$

$\phi(t_0)$  is assumed to be the stationary distribution of  $\phi(t)$ . Therefore,  $\phi(t_0)$  has a multivariate normal distribution with mean vector 0 [5] and covariance matrix  $V_0$ . Let  $v_{ij}$  denote the elements of  $V_0$ . Then,  $v_{ii} = \sigma_\phi^2 / (1 - \theta^2)$  [5] and  $v_{ij} = v_{ii} \cdot \text{cor}(\phi(1, i), \phi(1, j))$ , where  $\text{cor}(\phi(1, i), \phi(1, j)) = \theta^{2[\max(i, j) - 1]}$ . Hence, for  $t = t_0$ ,

$$f_1(\phi(t_0)) = \log p(\phi(t_0)) = -\frac{1}{2}\phi(t_0)' V_0^{-1} \phi(t_0) + \text{a constant}$$

**Assumption 2:** The curves of underlying log death rate within each individual year are smooth. For a fixed  $t$ , define  $y'_x = Y(t, x) - Y(t, x-1)$ ,  $y''_x = y'_x - y'_{x-1}$ , and  $y'''_x = y''_x - y''_{x-1}$ . By the assumption,  $y'''_x$  only slightly deviates from 0. Equivalently, for a given  $t$ ,  $Y(t, x)$  follows an ARIMA(0, 3, 0) process.

$$Y(t, x-3) - 3Y(t, x-2) + 3Y(t, x-1) - Y(t, x) = \sigma_Y Z_Y(t, x)$$

where  $\sigma_Y > 0$  and  $Z_Y(t, x) \sim N(0, 1)$ .

Define

$$\Delta = \begin{bmatrix} 1 & -3 & 3 & -1 & 0 & \dots & \dots \\ 0 & 1 & -3 & 3 & -1 & 0 & \dots \\ \vdots & 0 & 1 & -3 & 3 & -1 & 0 & \dots \\ \vdots & & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\ & & & 0 & 1 & -3 & 3 & -1 \end{bmatrix}$$

Then,

$$f_2(Y(t)) = \log p(Y(t)) = -\frac{1}{2\sigma_Y^2}(\Delta Y(t))'(\Delta Y(t)) + \text{a constant}$$

**Assumption 3:** Death counts,  $D(t, x)$ , are accurate. Thus,  $\phi(t, x) + \epsilon(t, x) + Y(t, x)$  needs to be the proximation of  $d(t, x)$ .

For computational purposes, assume  $d(t, x) \sim N(\mu, \sigma^2)$ , for each  $t$  and  $x$ . Then,  $D(t, x) \sim \ln N(\mu, \sigma^2)$ , and

$$E(D(t, x)) = \exp(\mu + \frac{\sigma^2}{2}) \quad (1)$$

$$Var(D(t, x)) = \exp(2\mu + \sigma^2)[\exp(\sigma^2) - 1] \quad (2)$$

Alternatively, suppose  $D(t, x) \sim \text{Pois}(\lambda)$ . Then,

$$E(D(t, x)) = \lambda \quad (3)$$

$$Var(D(t, x)) = \lambda \quad (4)$$

Equating (1) to (2) and (3) to (4),

$$\exp(\sigma^2) - 1 = \lambda^{-1}$$

By Taylor expansion,

$$\sigma^2 \approx \lambda^{-1}$$

Additionally, given  $\phi$ ,  $\epsilon$ , and  $Y$ ,

$$\phi(t, x) + \epsilon(t, x) + Y(t, x) = \widehat{\epsilon}(t, x) + Y(t, x) = \log \widehat{E}(t, x) + \log m(t, x) = \log D(t, x) = d(t, x)$$

Therefore, we set  $\mu = \phi(t, x) + \epsilon(t, x) + Y(t, x)$  and  $\lambda = \exp[\phi(t, x) + \epsilon(t, x) + Y(t, x)]$ . Then,  $\sigma^2 = 1 / \exp[(\phi(t, x) + \epsilon(t, x) + Y(t, x))]$ . Since our goal is to make  $\exp[\phi(t, x) + \epsilon(t, x) + Y(t, x)]$  ultimately close to  $D(t, x)$ , we are able to fix  $\sigma^2$  by setting  $\sigma^2 = 1/D(t, x)$ .

$$\begin{aligned} f_3(d(t)) &= \log p(d(t) | \phi(t), \epsilon(t), Y(t)) \\ &= -\frac{1}{2}(d(t) - \phi(t) - \epsilon(t) - Y(t))' V_D(t)^{-1} (d(t) - \phi(t) - \epsilon(t) - Y(t)) + \text{a constant} \end{aligned}$$

where

$$V_D(t) = \begin{bmatrix} D(t, x_0)^{-1} & 0 & \dots & 0 \\ 0 & D(t, x_0 + 1)^{-1} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & D(t, x_n)^{-1} \end{bmatrix}$$



### 3.3 The Bayesian Model and Posterior Sampling

Given  $d$  and  $\epsilon$ , we aim to find  $\phi$  and  $Y$  which satisfy the assumptions above.

$$p(\phi, Y|d, \epsilon) \propto p(\phi) \cdot p(Y) \cdot p(d|\phi, \epsilon, Y)$$

$$\log p(\phi, Y|d, \epsilon) = f_1(\phi(t_0)) + \sum_{t=t_0+1}^{t_n} f_1(\phi(t)) + \sum_{t=t_0}^{t_n} f_2(Y(t)) + \sum_{t=t_0}^{t_n} f_3(d(t)) + \text{a constant}$$

Define the target matrix

$$M = \begin{bmatrix} \begin{array}{c} | \\ Y(t_0) \\ | \end{array} & \begin{array}{c} | \\ Y(t_0+1) \\ | \end{array} & \dots & \begin{array}{c} | \\ Y(t_n) \\ | \end{array} & \begin{array}{c} | \\ \phi(t_0) \\ | \end{array} & \begin{array}{c} | \\ \phi(t_0+1) \\ | \end{array} & \dots & \begin{array}{c} | \\ \phi(t_n) \\ | \end{array} \end{bmatrix}$$

Let  $M_j$  denote the  $j^{th}$  column of  $M$  and  $M_{-j}$  denote the set  $\{M_k | k = 1, 2, \dots, j-1, j+1, \dots, 2t_n - 2t_0 + 2\}$ . Since  $p(\phi, Y|d, \epsilon)$  has a multivariate normal distribution,  $M_j | M_{-j} \sim N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ , for each  $j$ , and we are able to infer  $\boldsymbol{\mu}_j$  and  $\boldsymbol{\Sigma}_j$ .

Note

$$M_j = \begin{cases} Y(j+t_0-1) & j = 1, \dots, t_n - t_0 + 1 \\ \phi(j+2t_0-t_n-2) & j = t_n - t_0 + 2, \dots, 2t_n - 2t_0 + 2 \end{cases}$$

For  $j = 1, \dots, t_n - t_0 + 1$ ,

$$\begin{aligned} \boldsymbol{\Sigma}_j &= V_D(j+t_0-1)^{-1} + \frac{1}{\sigma_Y^2} \Delta' \Delta \\ \boldsymbol{\mu}_j &= \boldsymbol{\Sigma}_j^{-1} V_D(j+t_0-1)^{-1} [d(j+t_0-1) - M_{j+t_n-t_0+1} - \epsilon(j+t_0-1)] \end{aligned}$$

For  $j = t_n - t_0 + 2$ ,

$$\begin{aligned} \boldsymbol{\Sigma}_j &= V_D(t_0)^{-1} + V_0^{-1} + A' V_\phi^{-1} A \\ \boldsymbol{\mu}_j &= \boldsymbol{\Sigma}_j^{-1} [V_D(t_0)^{-1} (d(t_0) - \epsilon(t_0) - M_1) + A' V_\phi^{-1} M_{t_n-t_0+3}] \end{aligned}$$

For  $j = t_n - t_0 + 3, \dots, 2t_n - 2t_0 + 1$ ,

$$\begin{aligned} \boldsymbol{\Sigma}_j &= V_D(j+2t_0-t_n-2)^{-1} + V_\phi^{-1} + A' V_\phi^{-1} A \\ \boldsymbol{\mu}_j &= \boldsymbol{\Sigma}_j^{-1} [V_D(j+2t_0-t_n-2)^{-1} l(j) + V_\phi^{-1} A M_{j-1} + A' V_\phi^{-1} M_{j+1}], \text{ where} \\ l(j) &= d(j+2t_0-t_n-2) - \epsilon(j+2t_0-t_n-2) - M_{j-(t_n-t_0+1)} \end{aligned}$$

For  $j = 2t_n - 2t_0 + 2$ ,

$$\begin{aligned}\Sigma_j &= V_D(t_n)^{-1} + V_\phi^{-1} \\ \mu_j &= \Sigma_j^{-1} [V_D(t_n)^{-1} (d(t_n) - \epsilon(t_n) - M_{t_n-t_0+1}) + V_\phi^{-1} A M_{2t_n-2t_0+1}]\end{aligned}$$

The algorithm of the posterior simulation follows.

---

**Algorithm 1** Gibbs Sampling

---

```

 $M^{(0)} \leftarrow \lceil \log \mathbf{m}(t, x) \mathbf{0} \rceil$ 
for  $s = 1$  to  $N$  do
  for  $j = 1$  to  $2t_n - 2t_0 + 2$  do
     $M_j^{(s)} \leftarrow$  a sample from distribution  $p(M_j | M_{-j}) =$ 
     $\{M_1^{(s)}, \dots, M_{j-1}^{(s)}, M_{j+1}^{(s-1)}, \dots, M_{2t_n-2t_0+2}^{(s-1)}\}$ 
  end for
end for

```

---

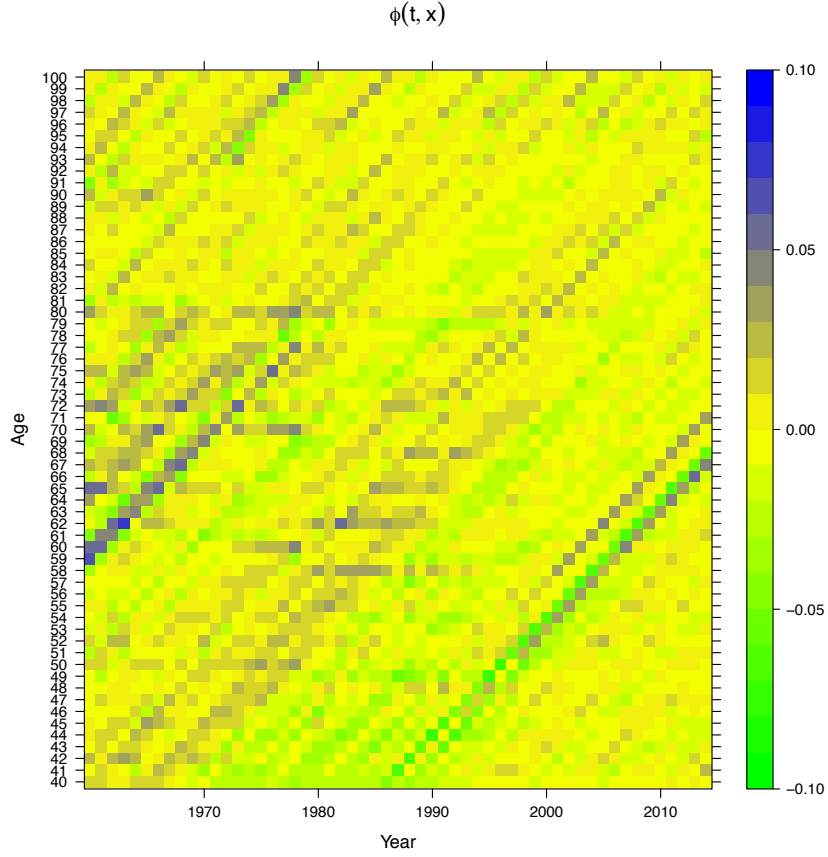


Figure 4: Level-plot of  $\phi(t, x)$

### 3.4 Output and Assumption Examination

By setting  $\theta = 0.9$  and  $\sigma_\phi = 0.02$ , we emphasize the strong persistence of errors in exposures within each individual cohort. Setting  $\sigma_Y = 0.01$  allows us to attain the smoothness in  $Y(t, x)$ . Figure 4 to 7 are generated by a single run of 20000 iterations and the posterior mean of  $M$  ( $Y$  and  $\phi$ ) is calculated by averaging the 7th outcome of every 10-iteration from iteration 10001 to iteration 20000.

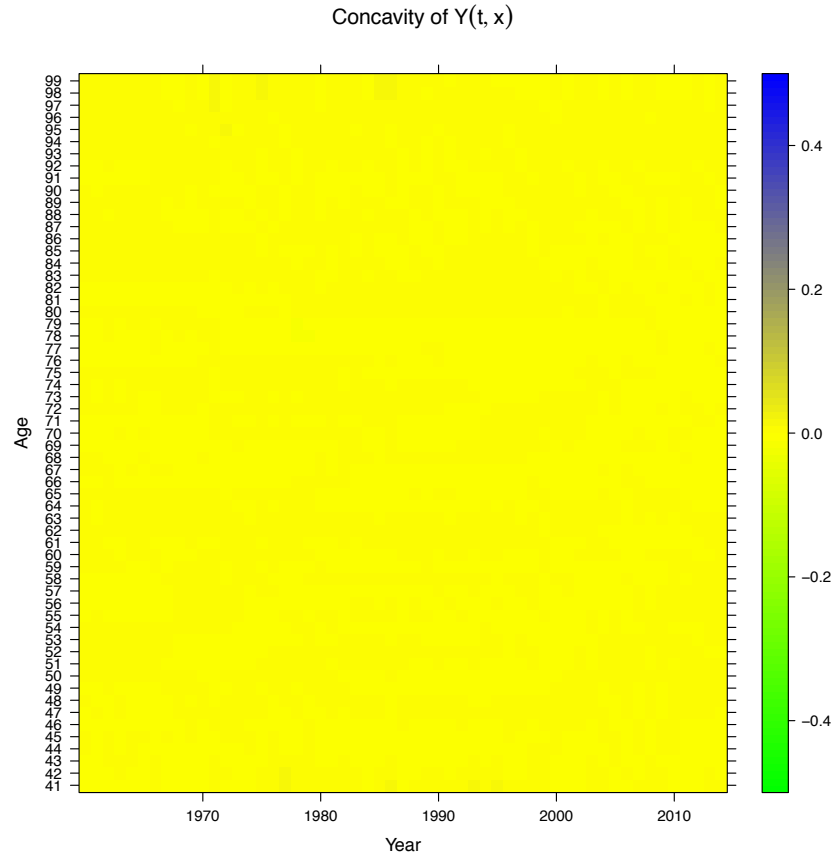


Figure 5: Level-plot of  $C(t, x)$ , after the posterior simulation. This corresponds to Figure 1, as  $Y(t, x) = \log \hat{m}(t, x)$ .  $C(t, x) \approx 0$  for each  $t$  and  $x$  in this graph.

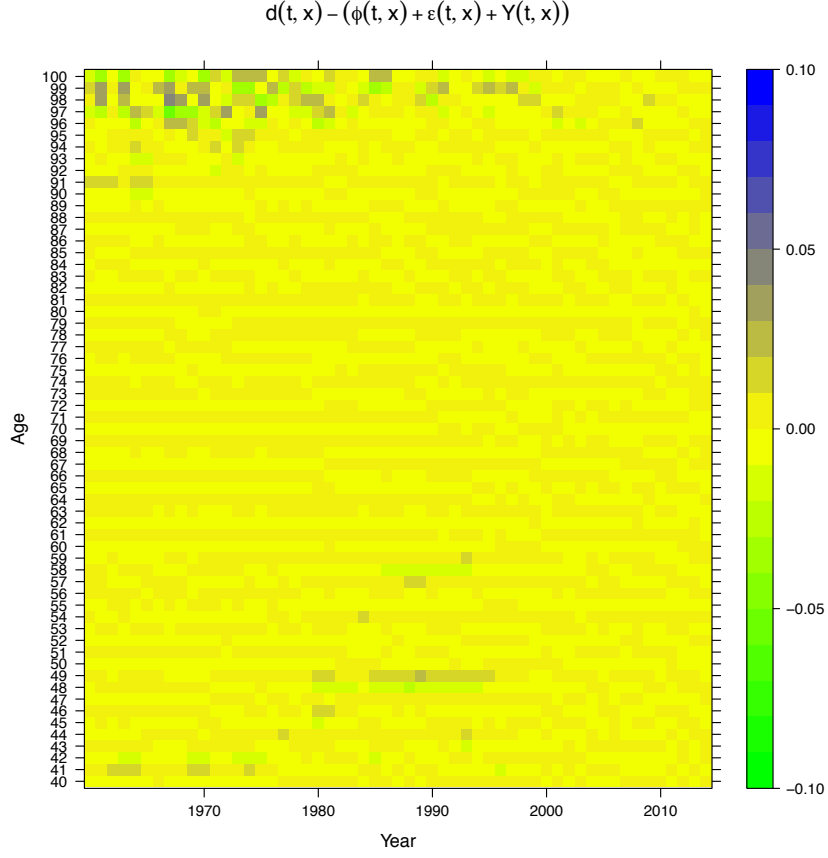


Figure 6: Level-plot of  $d(t, x) - (\phi(t, x) + \epsilon(t, x) + Y(t, x))$ . By Assumption 3, the value of  $d(t, x) - (\phi(t, x) + \epsilon(t, x) + Y(t, x))$  is desired to be close to zero.

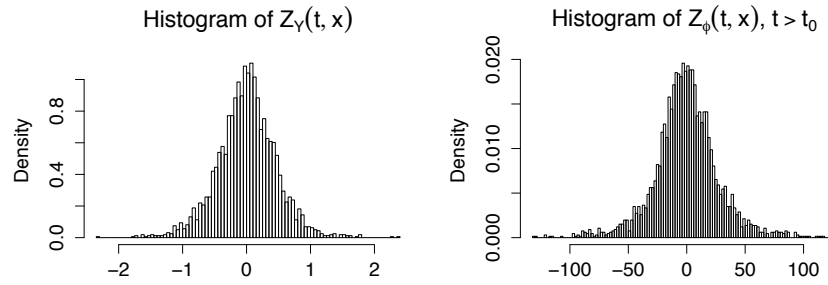


Figure 7: Histograms of residuals. Both are in mount shape.

## 4 Impact of Exposure Modification on Mortality Forecasts and Annuity Calculations

We use M7 for mortality projections, as it's designed for higher ages [6] and it's proved to be a suitable model for the U.S. mortality data [7]. M7 is an extension of the CBD model, where

$$\begin{aligned}\text{logit } q(t, x) &= \kappa_t^{(1)} + \kappa_t^{(2)}(x - \bar{x}) \text{ and} \\ q(t, x) &= 1 - \exp(-m(t, x))\end{aligned}$$

Note

$$\begin{aligned}\text{logit } q(t, x) &= \log \frac{q(t, x)}{1 - q(t, x)} = \log \frac{1 - \exp(-m(t, x))}{\exp(-m(t, x))} \\ &= \log[\exp(m(t, x)) - 1] \approx \log m(t, x), \text{ by Taylor expansion.}\end{aligned}$$

The logic of the CBD model is that, for a given  $t$ , log death rates vs. ages is linear. M7 includes a cohort effect term  $\gamma_{t-x}$  and a quadratic term in mortality projections. The former (obviously by its name) gives the information of cohort effect; the latter is inspired by the possibility of some curvature in the logit  $q(t, x)$  in the U.S. data [6].

$$\text{logit } q(t, x) = \kappa_t^{(1)} + \kappa_t^{(2)}(x - \bar{x}) + \kappa_t^{(3)}[(x - \bar{x})^2 - \sigma_X^2] + \gamma_{t-x}, \text{ where } \sigma_X^2 = N^{-1} \sum_i (x_i - \bar{x})^2$$

The period effects,  $\kappa_t^{(i)}$ , and the cohort effect,  $\gamma_{t-x}^{(4)}$ , follow a multivariate random walk and an AR(1) process (independent of the period effects), respectively.

We fit the model by using the data from year  $t = 1995$  to  $t = 2014$  and from age  $x = 55$  to  $x = 100$ . In Figure 8, the grey curve has significant different behaviors from the black curve in the bottom right panel.  $\gamma_{t-x}$  produced by the M7 model fitting using the original exposures-at-risk has a sharp fall at 1946 and a jump at 1920. However,  $\gamma_{t-x}$  produced by the M7 model fitting using the modified exposures-at-risk is smooth. Illustrations of mortality forecasts for age  $x = 65, 75, 85$ , and  $95$  are given in Figure 9 and 10, where Figure 9 is produced by using the original exposures-at-risk as data input and Figure 10 is produced by using the modified exposures-at-risk as data input. Bumps appear in Figure 9. For instance, the fan chart of age  $x = 75$  becomes lumpy at year  $t = 2021$ . This is because 1946 Cohort is at age  $x = 75$  in year  $t = 2021$ . Nonetheless, the fan charts are smooth in Figure 10.

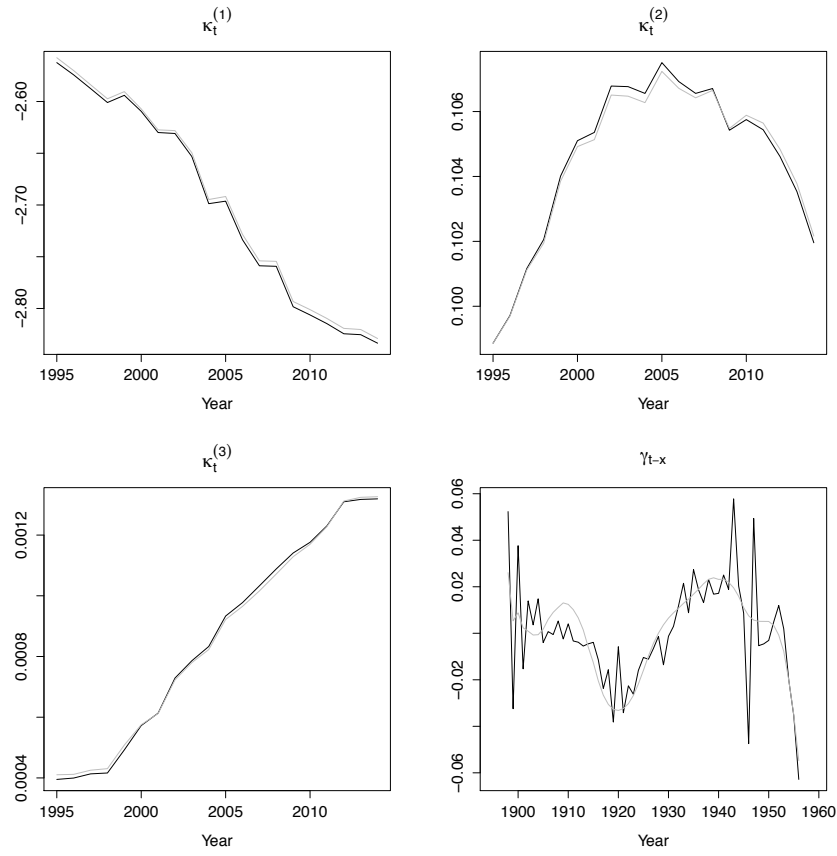


Figure 8: Black lines, results of M7 fitting where the original exposures-at-risk are used as data input; grey lines, results of M7 fitting where the modified exposures-at-risk are used as data input.

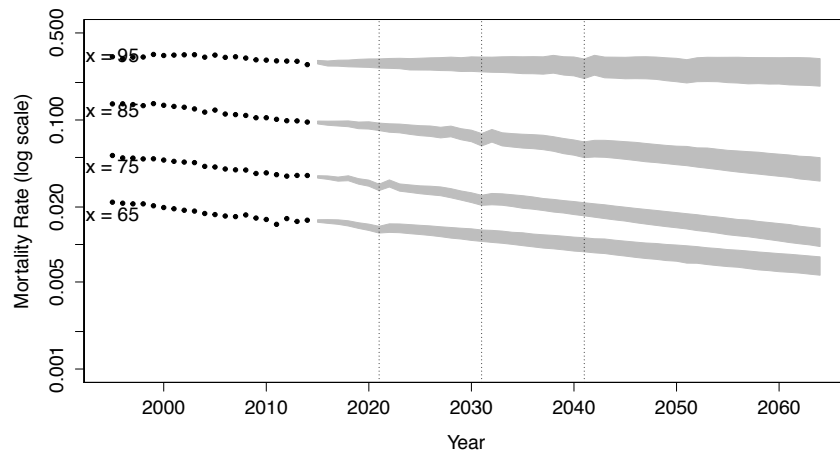


Figure 9: Mortality fan charts produced by using original exposures-at-risk as data input. Black dots, historical data; grey fans, mortality forecasts (90% prediction interval).

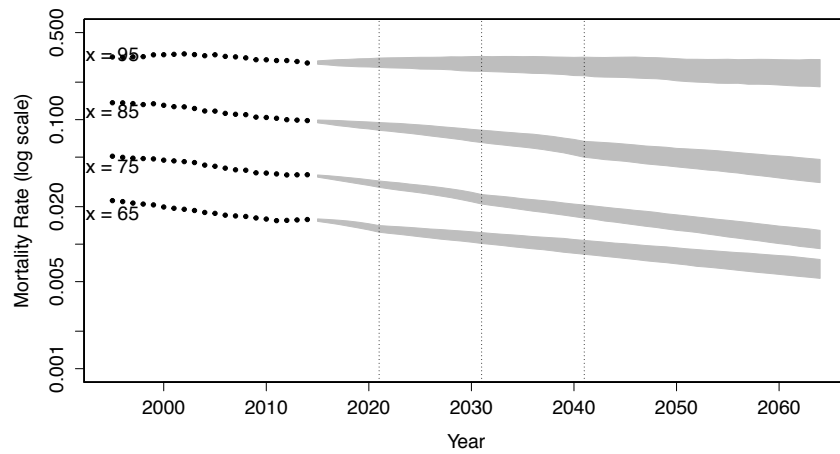


Figure 10: Mortality fan charts produced by using modified exposures-at-risk as data input. Black dots, historical data; grey fans, mortality forecasts (90% prediction interval).



Table 1: Interest rate  $r$  is assumed to be 3%. SI stands for the survivor index at the end of the 30-year period.

Age in 2014	Cohort	PV, original	PV, modified	Difference in PV (%)	SI, original	SI, modified	Difference in SI (%)
70	1944	10.888	10.885	0.02	0.00586	0.00569	2.89
69	1945	11.399	11.370	0.26	0.01284	0.01222	5.07
68	1946	12.086	11.844	2.04	0.02878	0.02331	23.50
67	1947	12.146	12.305	-1.29	0.03627	0.04007	-9.46

Mortality forecasts provide the building blocks of calculating mortality-related quantities.

The survivor rate,

$$S(t, s, x) = \prod_{j=0}^{s-1} [1 - q(t + j, x + j)]$$

indicates the proportion of males at age  $x$  in year  $t$  who survive to year  $t + s$ .

Since the dataset under study contains the historical data till year 2014, our interests lie in finding the present value of an annuity of \$1 payable annually in arrears for a maximum of 30 years starting at year 2015.

$$a(x) = \sum_{s=1}^{30} (1 + r)^{-s} S(2015, s, x)$$

As we used data from year  $t = 1995$  to  $t = 2014$  for the forecasts and Figure 4 indicates the adjustments on exposures of 1946 Cohort are negative in this time interval, the death rates of 1946 Cohort decreased after the modification. Therefore, the survivor index at the end of the 30-year period and the present value of the annuity are less than the ones calculated by using unadjusted exposures.

## 5 Conclusion

With the development of stochastic mortality models, the model itself becomes sensitive to capture the uneven patterns that appeared in a mortality dataset (Figure 9). We anticipate finding the cause of these irregularities. The anomalies that exhibited in the mortality data of the 1946 Cohort is

due to the surge of birth rate in the U.S. in 1946. The current world-widely adopted calculation for exposures-at-risk in mortality databases is unable to accurately present the values of the exposures-at-risk when there exists dramatic changes in the size of populations. The Bayesian model examined in this paper offers a resolution in dealing with this situation and provides the quantitative modifications on the exposures-at-risk. The future work may include advancing the estimation method for exposures-at-risk directly, since the posterior simulation becomes time costing as the mortality matrices expand and the outcome given by the mean of the samples from this simulation may carry errors too and hide the real information of the mortality dataset under study.

## Appendix A Graphic Diagnostic (Revisited)

There are two more graphical diagnostic tools available to exam the mortality dataset.

Hypothesis 1: Death rates by age for adjacent cohorts are similar. That is,  $m(t+k, x)$  is close to  $m(t, x)$ , when  $k$  is close to 0. Figure 11 plots the ratio  $m(t+s+k, x+s)/m(t+s, x+s)$  vs. age  $x$ . The illustration of death rates of 1925 Cohort and its four nearest neighborhood cohorts (top right panel of Figure 11) is an example of good pattern/phenomenon - the five lines in the plot follow a downward trend. In the bottom left panel, 1947 Cohort serves as the base cohort. The position of the black is considered a bit above the rest of four lines. The detailed information revealed in the bottom right panel about 1952 Cohort and its neighborhood cohorts is a little bit counterintuitive - from age 58 to 60, the younger cohort has a higher death rate.

Hypothesis 2: Changes in cohort exposure sizes match closely the number of reported deaths. Mathematically,  $D(t, x) \approx E(t+1, x+1) - E(t, x)$ . The condition  $E(t+1, x+1) > E(t, x)$  is not applicable to this dataset at large. However, the change of mathematical derivation of exposure-at-risk for the groups above age 80 can be observed by this diagnosis (Figure 12).

At the end, Figure 13 gives the level-plot of  $C(t, x)$ , where the data input is the U.S. female mortality dataset published HMD. The two datasets (by different genders) possess strong similarities.

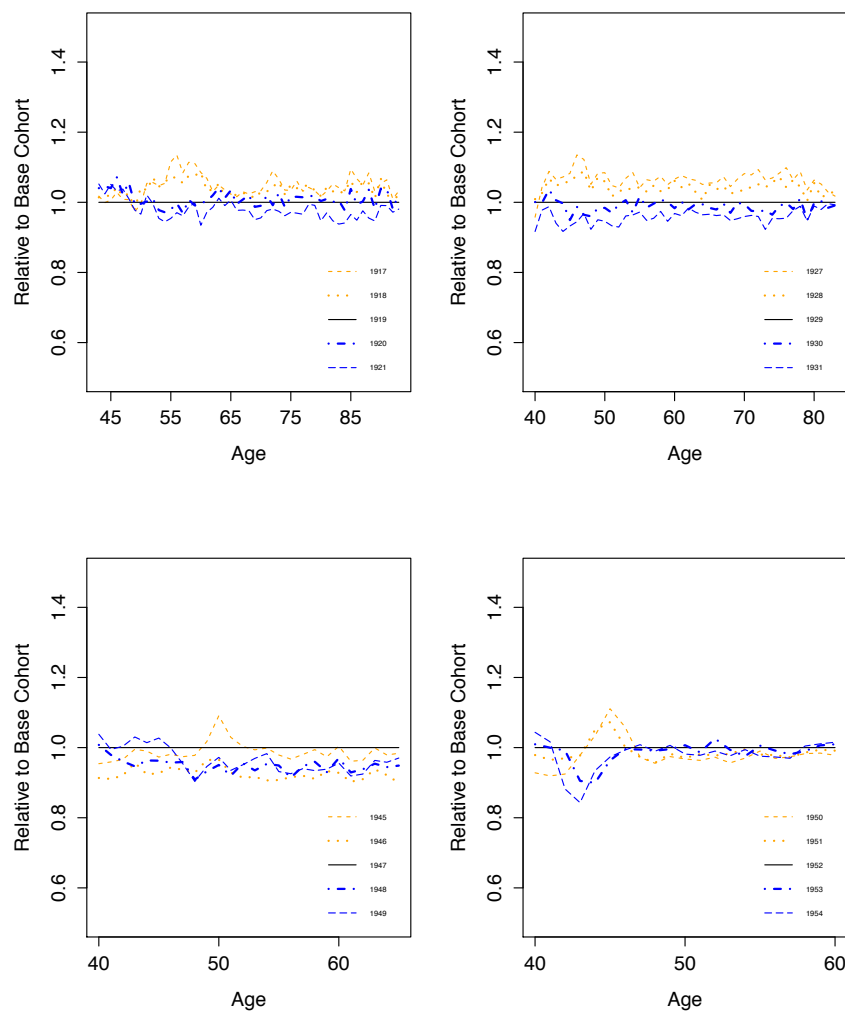


Figure 11: Base cohorts are 1919 (top left), 1929(top right), 1947(bottom left), and 1952(bottom right).

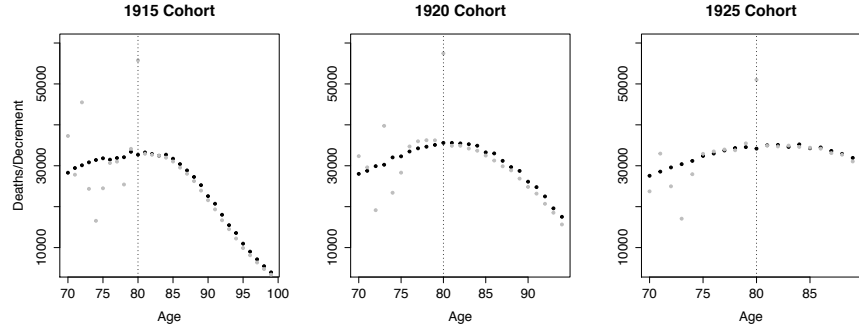


Figure 12: Black dots, death counts; Grey dots, exposure decrement.

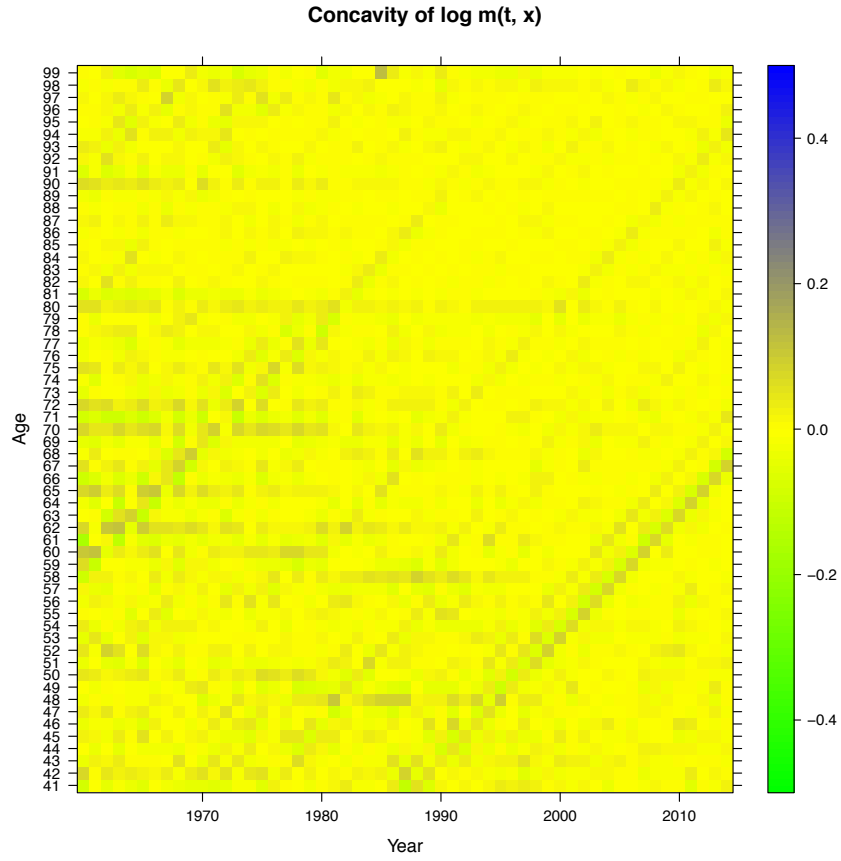


Figure 13: Level-plot of  $C(t, x)$ . Data: the U.S. female mortality dataset published by HMD.

## References

- [1] J. R. Wilmoth, K. Andreev, D. Jdanov, and D. A. Gleijer, “Methods protocol for the human mortality database,” 2007. (Available at <http://www.mortality.org>).
- [2] National Office of Vital Statistics, *Vital Statistics of the United States 1946*.
- [3] A. J. G. Cairns, D. Blake, K. Dowd, and A. R. Kessler, “Phantoms never die: living with unreliable population data,” *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 179, no. 4, pp. 975–1005, 2016.
- [4] D. C. Dickson, M. R. Hardy, and H. R. Waters, *Actuarial Mathematics for Life Contingent Risks*. Cambridge University Press, second ed., 2013.
- [5] G. P. Nason, “Stationary and non-stationary time series,” in *Statistics in Volcanology*, ch. 11, University of South Florida, 2010.
- [6] A. J. Cairns, D. Blake, K. Dowd, G. D. Coughlan, D. Epstein, A. Ong, and I. Balevich, “A quantitative comparison of stochastic mortality models using data from england and wales and the united states,” *North American Actuarial Journal*, March 2007.
- [7] A. J. Cairns, D. Blake, K. Dowd, G. D. Coughlan, D. Epstein, and M. Khalaf-Allah, “Mortality density forecasts: An analysis of six stochastic mortality models,” *Insurance: Mathematics and Economics*, vol. 48, no. 3, pp. 355 – 367, 2011.
- [8] University of California, Berkeley (USA) and Max Planck Institute for Demographic Research (Germany), *Human Mortality Database*.
- [9] National Office of Vital Statistics, *Vital Statistics of the United States 1944*.
- [10] National Office of Vital Statistics, *Vital Statistics of the United States 1945*.
- [11] National Office of Vital Statistics, *Vital Statistics of the United States 1947*.
- [12] National Office of Vital Statistics, *Vital Statistics of the United States 1948*.
- [13] National Office of Vital Statistics, *Vital Statistics of the United States 1949*.