From Data to Interactive Visualizations: A Tool for Modeling and Forecasting Longevity Across U.S. Subpopulations

Rosalia Hernandez

Department of Statistics and Applied Probability, University of California, Santa Barbara

Abstract

Longevity analysis provides valuable public health facts that can influence public policy, business decision-making, or new academic research directions. We must explore and interpret mortality data to gain these insights. This paper introduces a Longevity Forecasting Tool that we created with Shiny, an R package that facilitates interactive dashboard development. This tool showcases the use of Gaussian process regression for modeling mortality data. We use publicly available detailed mortality data from the Centers for Disease Control Wide-ranging Online Data for Epidemiologic Research (CDC WONDER). This tool uses interactive data visualizations to engage users to better understand the mortality experiences across several U.S. groups.

Introduction

Data science merges statistics and computer science methods with some domain knowledge [4]. We use data science to interpret, understand, and communicate our data, processes, and/or results to those interested in gaining insights from them. Analytical dashboards are an excellent option to communicate our scientific findings. They provide interactive graphical illustrations of data that allows users to identify and share valuable trends and insights on the spot. They can also be used to facilitate data exploration for those with little to no programming knowledge, whether it be students, researchers, decision-makers, or anyone interested in learning from data. This project aims to create an interactive dashboard that statistically models and forecasts mortality trends among racial groups in the United States. What makes this dashboard different from other mortality dashboards is the modeling methodology, which considers the year over year improvements of mortality by cause of death.

One of the motivations for this project stems from a retrospective study conducted by renowned economists Anne Case and Angus Deaton in 2015, which revealed that mortality improvements in our top two causes of death, cardiovascular disease and cancer mortality, masked an increase among “deaths of despair” [8]. This is the term they coined in their 2017 study to refer to deaths grouped by suicide, drug overdose, and alcohol-related liver disease. Case and Deaton note that an increase in these deaths was most aggressive among the White, middle-aged group, consequently increasing this group’s all-cause mortality rates [7]. When compared to other groups by race within the United States and other rich countries, the White, middle-aged group was the only group showing an increase in all-cause mortality [7]. Case and Deaton’s findings became front-page news and a best-selling book, leading to numerous further studies implementing various analytical methods and grouping of the data in different ways [2], [5], [6], [15]. Inspired to dig further, we employed a machine learning method to model the mortality data used.

The benefits of modeling the data are to quickly smooth out observation noise to identify patterns and make predictions on future longevity. Noise in data refers to extra information that does not add any value to our analysis, often seen as very sharp lines in a plot. Smoothing techniques then allow us to forecast future scenarios. Predictions in mortality are especially valuable since they could offset the 2 to 3-year lag of compiling this data and let us know what is likely happening right now. We used the Gaussian Process (G.P.) regression modeling approach that was introduced to model mortality data in 2018 by Michael Ludkovski, Jimmy Risk, and Howard Zail [13]. The G.P. modeling approach provides simultaneous modeling of mortality rates and improvement factors, uses smoothing techniques that eliminate random observation noise in the data, makes predictions for any age and year combination, and automatically quantifies associated uncertainties [11], [13]. This modeling method would be of particular interest to actuaries and demographers alike.

An actuary specializes in quantifying and managing risk and has a deep understanding of implications. In mortality studies, they ask age and population-specific questions, take a more individualistic approach to grouping, and use statistical modeling to analyze mortality rates and improvement factors. Improvement factors quantify the change in mortality year after year and are too specific to be used in a demographer's mortality analysis approach. Population-specific questions from an actuarial perspective do not consider many factors that a demographer does. There is no need for data on fertility, emigration, or infant mortality when performing longevity analysis. Actuaries analyze data, identify which patterns are systemic, which are noise, measure uncertainty, and make short term projections to manage associated risks of specific groups.

Case and Deaton use raw death rates for their analysis and primarily take on a demographer's approach [7], [8]. One key component they use to indicate this type of approach is age aggregation (grouping ages 50-54, for example) [8]. This is standard of a demographer whose questions are related to the general makeup of a population, asking about its size, distribution, and spread [3]. They consider population attributes such as fertility, infant mortality, emigration, life expectancy, and morbidity. Their goal is to see a large-scale picture of a population, make projections that may go decades into the future, and see trends in raw data [3]. For this project, we used a more actuarial inclined approach.

We used the Underlying Cause of Death: Detailed Mortality data from the Centers for Disease Control Wide-ranging Online Data for Epidemiologic Research (CDC WONDER). The CDC WONDER database is a free, publicly available data query system with various public health datasets. It provided us with one of the same data sources that Case and Deaton used in their analysis [8]. We used racial groups categorized by Hispanic ethnicity, gender, and cause-of-death. Cause-of-death options include all-cause, cardiovascular disease, cancer, stroke, and external causes. The external causes option most closely corresponds to Case and Deaton's "deaths of despair." The objective is to allow users to explore and analyze the mortality experiences among different United States (U.S.) population segments and show the Gaussian Process modeling's implementation on this data. With this dashboard, we can examine longevity inequalities and diverging or converging mortality experiences. We can also identify cohort effects, compare mortality improvements among specific populations, and quantify associated uncertainties. Cohort effects are common mortality experiences by specific age group trends over time (like a particular generation).

We create this dashboard completely in RStudio, a free, open-source statistical programming environment [14], using the Shiny package, a popular module package for interactive dashboard development [10]. Shiny provides app developers the flexibility needed to build and stream interactive web apps to communicate complex results engagingly. An analytical dashboard developed with Shiny is widely known as a Shiny App. Since thousands of Shiny Apps exist, Shiny is considered a leading ecosystem of interactive dashboards in the scientific community. Anything that can be done in RStudio can be translated into Shiny. It is easy to deploy from the RStudio console to quickly be made accessible to anyone with an internet connection and also reduces the need for a full-fledged web developer [10]. The Shiny app for this project is publicly available at the following URL: https://rosalia1010.shinyapps.io/Longevity_Forecasting_Tool/

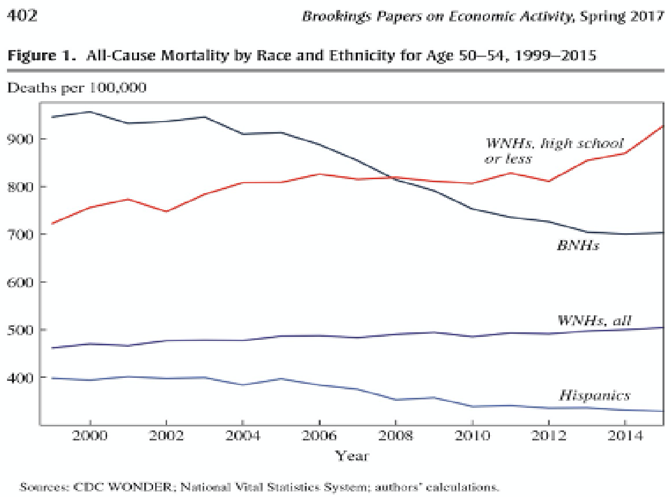
This paper is organized as follows: (2) General Overview: contains definitions of concepts and terminology used to understand longevity analysis, (3) Data: provides detailed information on the data used for this project, (4) Introduction to our Forecasting Tool: introduces the forecasting tool and its components. The output graphs include: (4.1) Smoothed Mortality, (4.2) Forecasting Mortality, (4.3) Mortality Improvement Factors, and (4.4) Mortality Improvement Factors Heat Maps.

General Overview

Case and Deaton's (2017) findings focus primarily on middle-aged mortality experiences and include populations grouped by education level. Since aggregating education level requires incorporating data from different sources, we will focus only on the populations as grouped by in the CDC WONDER Detailed Mortality Database. We will not be able to group by education level. Figure 1 shows two all-cause mortality plots: (a) one by Case and Deaton (2017) and one using our data from the CDC WONDER database. Figure 1(b) shows

mortality for age 50 only, representing a subset of the age-group used in 1(a) Case and Deaton's graphs. Notice that the curves are somewhat pointy; this reflects the use of raw mortality data, which means no statistical modeling.

Both graphs show similar trends of the Black non-Hispanics (BNH), White non-Hispanics (WNH), and Hispanics (HISP) subpopulations. The mortality rates (y-axis) in figure 1(a) are higher than for those in figure 1(b) because 1(b) illustrates a subset of the group used in 1(a). There is a substantial decrease in mortality for the BNH population (top curve) over time (x-axis), indicating drastic mortality improvements experienced by this group. This improvement shows a reducing gap in mortality disparities illustrated by the narrowing white space between the lines. Both figures 1(a) and 1(b) show this trend. The Hispanic population (bottom curve) appears to be slightly decreasing over time in both graphs as well. The White non-Hispanic group (lower middle curve) is the only population to experience a non-decreasing trend in all-cause mortality during the same time frame. Since both figures show this, it supports Case and Deaton's (2017) findings that the White non-Hispanic population has experienced an increase in all-cause mortality.



402 *Brookings Papers on Economic Activity*, Spring 2017

Figure 1. All-Cause Mortality by Race and Ethnicity for Age 50–54, 1999–2015

Sources: CDC WONDER; National Vital Statistics System; authors' calculations.

Mortality has been strictly decreasing since around the 1950s because of medical and technological advancements and improvements in healthcare [1]. This is why an increase in all-cause mortality is alarming. There are some terms and concepts that are important to understand for mortality analysis. The following sections provide a general overview of Gaussian process regression, the modeling method we use, and its outputs: mortality rates and improvement factors, the metrics used for analysis.

Gaussian Process Regression: Gaussian Process regression is a probabilistic machine learning tool. Its unified modeling of mortality rates and improvement factors produces a smoothed probability surface [13]. A probability surface is a 3-dimensional representation of a 4-space result. In our case, the outputs are mortality rates and improvement factors over age and years. The forecasting tool illustrates different slices of the surface the G.P. model creates to isolate and better understand some key components of longevity across the various groups. G.P.'s main application is to fit a function to historical data and quantify the uncertainty around it. Historical data is data that is observed and collected (also known as in-sample data). In this sense, the approach is entirely data-driven. The credible bands illustrate the uncertainty and are depicted as shaded ribbons on graphs. They highlight the scenarios of projections that are most likely to occur. The ribbons can also be thought of as the scope of our model's uncertainty. The further ahead in time we want to project, the wider these credible bands become because there is more uncertainty.

Uncertainty comes from various places, but in general can stem from 3 categories: (1) uncertainty through unexpected events (like a pandemic), (2) through structural changes (like drug overdoses which may take a long time to identify), or (3) through data estimation. The last group (data estimation) is the most relevant here since getting every person's information from a whole population is impossible. Imagine trying to gather data on □ 300 million people (the approximate number of people in the United States according to the U.S. Census Bureau). Instead, we do our best to use data that are representative of the whole population in calculating the best estimates of the true population. The uncertainty comes with how accurate those data and estimates are of the true population and is accounted for in models as error terms and confidence intervals.

As a setup to G.P. regression, we take in a pair of input variables x (age and year) linked to output variable y (log-mortality) through a function f and error term. The collection of f's for each age and year combination creates the probability surface and is the Gaussian process realization [11], [13].

f(x) ~ GP(mean = m(x), covariance = C(x,x))

The mean and covariance are essential components of the G.P. We initially tried using a constant mean function on age and year, but it made future projections of mortality rates rise unreasonably, which did not make sense to our data. The constant mean function in our model tended to average mortality over the years. Since this is not a real-life characteristic of mortality over time, we decided to use a linear mean function, which considers age (i.e., a person increases in age linearly) and time (i.e., years increase linearly as well) linearly. The covariance function, also known as the G.P.'s kernel, controls how smooth the probability surface is. There are many kernel functions to use and many careers that focus on studying kernel functions. For simplicity and reproducibility, we will only state which commonly used kernel function our models use: squared-exponential kernel.

Mortality Rates vs. Improvement Factors: A mortality rate is the fraction of the number of deaths in a population over the total number of people in that population per time unit. The time unit used here is in years. A mortality improvement factor is the amount of improvement when comparing the current year's mortality rate to the previous one. Essentially, a mortality improvement factor is like the slope of a mortality rate curve. If the mortality rate stays the same for two years in a row, there is no mortality improvement (mortality improvement factor = 0). Alternatively, if the mortality rate decreases from the previous year, there is a positive mortality improvement. If the mortality rate has increased from the previous year, then there is a negative mortality improvement. Examining mortality rates is a general rule of thumb for both demographers and actuaries but examining mortality improvement factors is actuarial. Mortality improvement factors provide a yearly account into the direction mortality rates are moving. In general, mortality improvement factors and mortality rates are defined as follows:

$$\text{Mortality Rate} = \frac{\text{\# of People who died in a population segment (per year)}}{\text{Total \# of people in that population segment}}$$

$$\text{Mortality Improvement} = 1 - \frac{\text{Mortality rate at a specific age for a specific year}}{\text{Mortality rate for the same specific age but for the previous year}}$$

Data

The data used is collected by the National Center for Health Statistics (NCHS) and provided to the public by the CDC WONDER's public online databases. The specific database used is the Underlying Cause of Death: Detailed Mortality database.

Our data set included the years spanning from 1999 to 2014. We used single age years beginning from age 50 to age 84 since our interest is in modeling longevity. We included sex data of males and females. We grouped Hispanic origin and race into a single demographic option, like in Case and Deaton's (2017) study. For Hispanic origin, we only used the observations that provided Hispanic origin information ("Hispanic or Latino" and "not Hispanic or Latino"). We excluded observations with a "Not Stated" entry under Hispanic origin because these entries do not have corresponding populations calculated [9]. Entries that state "not Hispanic or Latino" are then grouped into a race category. We used the four race options available: American Indian or Alaska Native, Asian or Pacific Islander, Black or African American, and White [9].

The cause-of-death list used is the tenth revision of the International Classification of Diseases (ICD-10 113 Cause List) code list, used after 1999 for classifying deaths [9]. Our dataset has 136 causes of deaths, with 52 "Level 1" causes. Level 1 causes are "rankable" causes of death as indicated by the NCHS [9]. There are 4 remaining levels for causes that are subsets of the Level 1 causes. For instance, all level 2 causes of death add up to the level 1 cause they are a subset of, all the level 3 causes add up to the level 2 cause they are subsets of, and so on. An example would be malignant cancers as a level 1 cause, with breast cancer, prostate cancer, skin cancer, stomach cancer, and more being level 2 causes. The cause of death observations originated from the physician's information on each death certificate [9]. Our longevity forecasting app currently has 6 cause of death groups: all-cause, cardiovascular disease, strokes, cancer (malignant), and external causes. The exposed population entries are between-census estimates from the U.S. census counts [9].

The Underlying Cause of Death database allows for up to 5 groupings. The data of interest included six data variables: age, year, gender, Hispanic origin, race, and cause of death. We used more involved methods for extracting the data with R code and the help of

Mr. Howard Zail, an actuary and partner at Elucidor, LLC.

The data was grouped as follows:

- Sex: Male, Female, Both (3 options)
- Demographic: All, Hispanic Origin, Black non-Hispanic, White non-Hispanic, American Indian or Alaska Native non-Hispanic, Asian or Pacific Islander non-Hispanic (5 options)
- Cause of Death: Aggregated, Cardiovascular Disease, Cancer (Malignant), Stroke, External Causes (5 options)

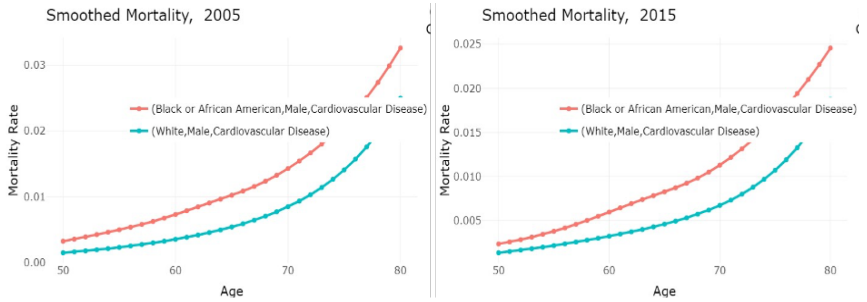
Mortality Modeling and Forecasting Tool

The longevity forecasting tool engages users to explore mortality rates and improvement factors across different U.S. subpopulations. It uses Gaussian process regression models based on which groups are selected. The "Choose Population" column is where we choose the groups of interest. Options include groups by sex, demographic, and cause-of-death, which means that there were 3×5×5 models (75 models). Additional cause of death scenarios can be implemented upon request. More information on making such requests is found in the Longevity Forecasting Dashboard's "Cause of Death Table" tab.

In Gaussian Process regression, the data is smoothed via in-sample predictions [13]. Our graphs visually demonstrate slices of the mortality probability surface produced by the G.P. models. The following are the output graphs: smoothed mortality across ages, smoothed mortality across years, smoothed mortality improvement curves across ages, and improvement factors as heat maps. Each output graph has options to save the graph as a .png file, zoom in and out on specific areas, hover over the graphs to data at a particular point, or compare data at specific points by clicking on the legend to include or exclude chosen populations. Most output graphs also have reactive elements that allow users to change the data they are examining. Some examples include changing the year, age, number of years to forecast, and including or excluding confidence intervals or raw observations.

Smoothed Mortality

We illustrated smoothed mortality over ages in single years from 1999 to 2015. Figure 2 shows the smoothed mortality over a ten-year timeframe for the populations mentioned previously. We identified that the BNH male population experiences higher rates of cardiovascular disease mortality when compared to the WNH male population. Notice that the overall mortality rates have decreased from 2005 to 2015, suggesting an overall decrease in cardiovascular disease mortality for all sub-populations over this ten-year period. We also identified a slight disparity convergence around ages 50-55. The gap between the curves appears smaller in that area for the year 2015 than it does for the year 2005.



(a)2005 (b)2010

Figure 2. Cardiovascular disease mortality rates among Black non-Hispanic and White non-Hispanic males, 10-year difference.

Additionally, it is usual for overall mortality curves to increase as age increases because the risk of dying increases roughly exponentially with age [12]. This pattern was observed and described in 1825 by Benjamin Gompertz as the law of human mortality, a law that still holds true. The Gompertz model has been widely studied and translated into other disciplines. It provides a powerful way to examine mortality patterns [12].

Forecasting Mortality

The next tab allowed us to see changes in smoothed mortality over time and future mortality predictions for single age, as shown in Figure 3. The year 2015 is marked by the vertical dashed line. This is where our in-sample data ends, and projections for out-of-sample data began.

We verified the mortality convergence we saw in the previous graph by looking at the trends in Figure 3. The vertical distance between the two curves in Figure 3(a) is a measurement of the disparity between them. Over time, this disparity among those aged 50 reduces. The convergence is primarily due to the mortality improvements experienced by the BNH group, illustrated by a considerable decrease in this curve. We also saw slightly declining mortality rates for the male WNH group. For the groups aged 80, there appeared to be similar mortality improvement rates, illustrated by paralleling curves. Again, it is evident that the BNH group experiences higher cardiovascular disease mortality rates.

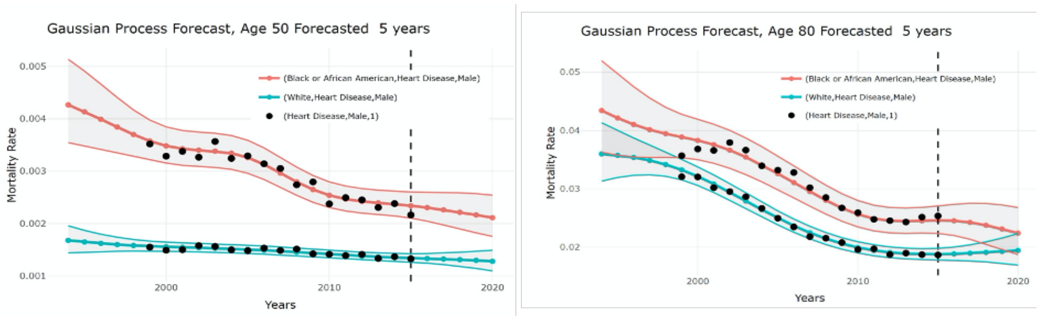


Figure 3. Cardiovascular disease mortality rates among Black non-Hispanic and White non-Hispanic males, 10-year difference.

This Forecasting tab has many reactive options. Users can change the age of interest, the number of years to forecast, view the 95% confidence intervals, and include the actual observations against the fitted model. The confidence intervals, shown as ribbons, highlighted all the possible curves that we can use to fit the data we provide. The option for "in-sample observations" showed black dots on the fitted curves. These black dots represent real raw values, the actual data. They showed us how well the model has fit the in-sample data. Users can also view forecasts of up to 20 years into the future.

Mortality Improvement Factors

Mortality improvement factors can be slightly less intuitive at first. Figure 4 shows the mortality improvement factors for the two groups in 2012. The bold line at mortality improvement = 0 means that compared to the previous year, the current year has not changed in its mortality rate. The farther away the curve is from this line, the larger the increase or decrease in mortality rate experienced. If the curve falls below this line, the mortality improvement factor is negative. Thus, the population of interest has experienced an increase in mortality. If the mortality improvement curve is above this line, that means that the population has experienced a decrease in the mortality rate. Thus, the mortality improvement factor is positive. The reason we look at mortality improvement factors is that year after year mortality rates change slowly. By examining mortality improvements, we have a better visual of the severity of change in mortality rates year after year. For decades, these curves have been strictly positive.

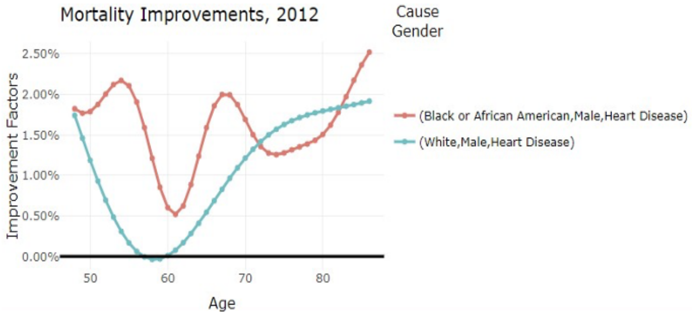


Figure 4. Mortality improvements for 2012.

Figure 4 shows the mortality improvement factors of the two groups we have compared for the year 2012. The improvement curves of both groups get close around ages 45, 72, and 84. This suggests similar improvement rates at that age for both populations. The WNH subpopulations improvement curve at the bold line (where the improvement factor = 0) tells us there is no mortality improvement for this subpopulation at age 58. Overall, the improvement factor curve for the BNH subpopulation is higher, which means this population has experienced higher mortality improvements than the WNH population, supporting our previous note about the BNH subpopulation experiencing larger decreases in mortality rates relative to the WNH population. This supports Case and Deaton's (2017) findings that show that the all-cause mortality gap between the BNH group and other racial groups are decreasing. They indicate that it is mainly due to improvements in our top cause-of-death contributors.

Mortality Improvement Factors Heatmap

Mortality improvement factors can be illustrated across ages and years through heat maps. The color bar on the right indicates the improvement factor level by hue. The lighter green colors indicate positive mortality improvement, and the darker colors indicate negative improvements. We can read heat maps vertically, horizontally, and diagonally.

Figure 5 shows a heat map of the improvement factors of both groups. A single row illustrates mortality improvements over age, a single column illustrates mortality improvements over a year, and a diagonal line illustrates mortality improvements by cohort. The arrows on the heat maps in Figure 5 highlight the cohort effects. Cohort effects show us generational trends as certain generations experience mortality. In 2000, the baby boomer generation was between 36-54 years old. This generation shows a cohort effect of lower mortality improvements illustrated by the darker shades along the arrows. We see reduced improvement rates by the change in colors in the area consistent through many older ages.

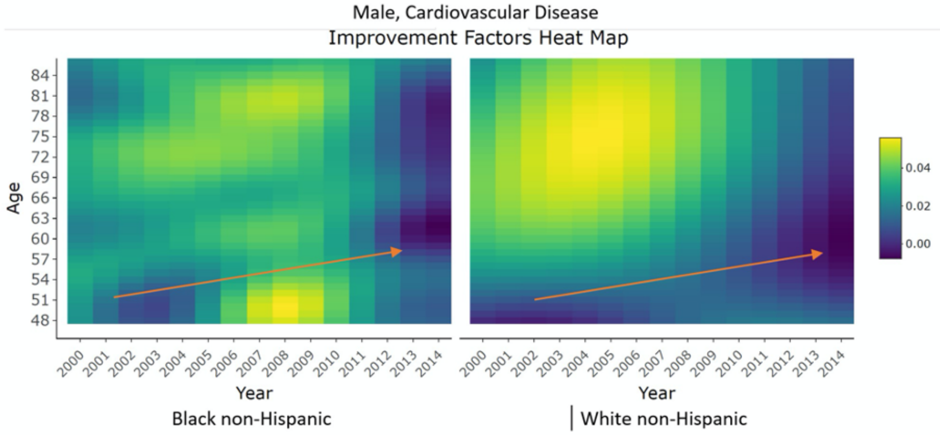


Figure 5. Heatmaps for white non-Hispanic males with cardiovascular disease and for Black non-Hispanic males with cardiovascular disease.

Summary

Mortality modeling provides us with an avenue to measure public health and manage associated risks. Being able to see trends among different subpopulations and in cause-specific groups, we can identify key factors that contribute to the nation's overall longevity trend. This longevity exploration tool uses a Gaussian process approach to modeling mortality that provides unified modeling of mortality rates and improvement factors which take into consideration the mortality improvements year after year. This approach also quantifies uncertainty for in-sample modeling and future trajectories. This allows us to examine the longevity experience of specific groups defined by sex, ethnicity, and cause of death through a series of mortality rates and mortality improvement factors graphs.

The next step of this project is to refine the tool to include more interactivity. This will enhance user experience and facilitate understanding through interactive data visualizations. We also aim to dive deeper into the data and perform an analysis using this tool and modeling method. We aim to understand the cause-specific mortality experiences of different groups. This project could be enhanced by including additional cause-of-death criteria.

References

[1] Alicia H. Munnell Anqi Chen and Geoffrey T. Sanzenbacher. "What's Happening to

U.S. Mortality Rates?" In: Center for Retirement Research at Boston College 1717 (2017), p. 14. url: https://collections.nlm.nih.gov/master/borndig/ 101716951/IB_17-17.pdf.

[2] Magali Barbieri. "The decrease in life expectancy in the United States since 2014". In: Population & Societies 570 (2019), p. 4. url: https://www.ined.fr/fichier/s_ rubrique/29581/570.population.societies.decrease.life.expectancy.usa. eng.en.pdf.

[3] Assia Billig and Adrian Gallop. Why Actuaries Are Interested in Population Issues and Why Other Organisations Interested in Population Issues Should Talk to Actuaries?

Executive Summary. Tech. rep. International Actuarial Association Population Issues Working Group. url: https://www.actuaires.org/CTTEES_PIWG/Documents/ PIWG_Paper_Why_Actuaries_Interested_Population.pdf.

[4] David M. Blei and Padhraic Smyth. "Science and data science". In: Proceedings of the National Academy of Sciences of the United States of America 114.33 (Aug. 2017), pp. 8689–8692. url: https://www.pnas.org/content/114/33/8689%20https:

//www.pnas.org/content/114/33/8689.

[5] Alexandre Boumezoued et al. Modeling and Forecasting Cause-of-Death Mortality. Tech. rep. Society of Actuaries, 2019, p. 71.

[6] Andrew J G Cairns and Cristian Redondo Loures. "U.S. Mortality: Underlying Trends By Socioeconomic Group and Cause of Death". In: Actuarial Research Conference, August 2019. Indianapolis, 2019. url: www.actuaries.org.uk/arc.

[9] Centers for Disease Control and Prevention. Underlying Cause of Death 1999-2015. url: https://wonder.cdc.gov/wonder/help/ucd.html#.

[10] Winston Chang et al. Shiny: Web Application Framework for R [Computer Software]. 2017.

[11] Nhan Huynh and Mike Ludkovski. "Multi-Output Gaussian Processes for Multi-Population Longevity Modeling". (2020).

[12] Thomas B L Kirkwood. "Deciphering death: a commentary on Gompertz (1825) 'On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies'". (1825).

[13] Mike Ludkovski, Jimmy Risk, and Howard Zail. Gaussian process models for mortality rates and improvement factors. Vol. 48. 3. Cambridge University Press, Sept. 2018, pp. 1307–1347.

[14] RStudio Team. RStudio: Integrated Development for R. Boston, MA, 2020. url: http: //www.rstudio.com.

[15] Steven H. Woolf and Heidi Schoomaker. "Life Expectancy and Mortality Rates in the United States, 1959-2017". In: JAMA - Journal of the American Medical Association 322.20 (Nov. 2019), pp. 1996–2016.

About the Author

Rosalia Hernandez is a fourth-year actuarial science major at the University of California Santa Barbara (UCSB). She enjoys learning about emerging technologies and methods in the data science realm and engaging in conversations about their applications, impacts, and implications. As a first-generation non-traditional student, she is passionate about making data insights accessible to everyone and strives to empower and encourage minoritized students to pursue research. Her goal is to continue her education in a graduate program in statistics and ultimately seek a professorship position.