Hierarchical Mortality Forecasting with EVT Tails: An Application to Solvency Capital Requirement

By

Han Li and Hua Chen

March 30, 2022

Han Li
Department of Economics
Faculty of Business and Economics
University of Melbourne
Phone: +61 (3) 8344 5618
Email: han.li@unimelb.edu.au

Hua Chen
Department of Finance
Shidler College of Business
University of Hawaii at Manoa
Phone: +1 (808) 956-8063
Email: huachen@hawaii.edu

Hierarchical Mortality Forecasting with EVT Tails: An Application to Solvency Capital Requirement

March 30, 2022

Abstract

In this paper, we propose a new framework to coherently produce probabilistic mortality forecasts by exploiting techniques in seasonal time-series models, extreme value theory (EVT), and hierarchical forecast reconciliation. Coherent forecasts implies that forecasts can add up in a manner consistent with the aggregation structure of the collection of time series. We are amongst the first to model and analyze U.S. monthly death counts data during the period of 1968–2019 to explore the seasonality and the age-gender dependence structure of mortality. Our results indicate that incorporating EVT and hierarchical forecast reconciliation greatly improves the overall forecast accuracy, which has important implications for life insurers in terms of rate making, reserve setting, and capital adequacy compliance. Using the solvency capital requirement (SCR) under Solvency II as an example, we show that the SCR calculated by our approach is considerably higher than those calculated by alternative models, suggesting that failing to account for extreme mortality risk and mortality dependence in the hierarchy can result in significant underfunded problems for life insurers. We find that our model can yield death forecasts that are largely consistent with actual data in most of months in 2021 when death tolls surged due to COVID-19. This provides additional evidence of the effectiveness of our model for practical uses.

Keywords: Seasonal time-series model; Extreme value theory; Hierarchical forecast reconciliation; Solvency capital requirement; the COVID-19 pandemic.

1 Introduction

Solvency is the cornerstone of insurance regulation. It requires insurance companies to reserve sufficient capital to operate and meet their financial obligations to policyholders and other claimants, while taking into account relevant risks. In the report on the Quantitative Impact Study 5 (QIS5), the European Insurance and Occupational Pensions Authority (EIOPA) identifies life underwriting risk as the second most material risk (behind market risk) for life insurers. Mortality risk thus becomes increasingly important in calculating solvency capital for life insurers.

To the best of our knowledge, there are only a few studies that quantitatively examine solvency capital for life insurers (see, e.g., Sharara *et al.*, 2010; Silverman and Simpson, 2011; Zhou *et al.*, 2014; Boonen, 2017). The modeling approaches adopted in these studies are based on annual mortality data. Therefore, strong assumptions have to be made regarding the time of death or the distribution of deaths in a year, in order to determine when insurance benefits are paid. Otherwise, it is often assumed that benefits are disbursed at the end of each year. The timing of benefit payments is important for insurance rate making, reserve setting, and solvency capital requirement because all of these policies are based on the calculation of the present value of insurance liabilities which need to be discounted from the payment dates. If we assume payments are made at the end of each year, it will cause an underestimation of the present value of insurance liabilities. Assuming any death distribution in a year without the support of a more granular level of data seems arbitrary and leads to imprecise estimation. Nevertheless, even for some OECD countries high-quality mortality data at a granular level was not available for periods prior to the 1960s.

In this paper, we model monthly death counts data in the U.S. from 1968-2019 and capture the seasonality of mortality. Large-scale population studies have shown seasonal variations in mortality from various causes in different parts of the world (see, e.g., Donaldson and Keatinge, 2002; Armstrong *et al.*, 2011; Rau *et al.*, 2017). For example, influenza is one of the most common winter-specific seasonal causes of death, followed by other respiratory-related causes. External deaths such as suicide, accidents, and homicides tend to be higher in the summer, especially for certain younger age groups. A thorough understanding of mortality seasonality is crucial for life insurers for pricing their products, setting reserves, and calculating solvency capital requirements.

We propose a new framework to produce coherent mortality forecasts in a hierarchical structure while taking into account mortality seasonality. Coherent forecasts implies that forecasts can add up in a manner consistent with the aggregation structure of the collection of time series. The proposed framework consists of three steps. In the first step, we use seasonal time-series models to fit monthly death counts data in the hierarchy. In the second step, we model the standardized residuals obtained from the first step via extreme value theory (EVT). In the third step, we produce mortality forecasts for each time series independently, and then adopt a trace minimization (MinT) approach (Wickramasuriya *et al.*, 2019) to reconcile these probabilistic forecasts according to the dependence structure of forecasting errors and the underlying aggregation constraints. To illustrate the effectiveness of our model, we calculate the solvency capital requirement (SCR) under Solvency II for a hypothetical insurance portfolio. Our results show that ignoring extreme mortality risk or mortality dependence in the hierarchy would lead to underestimated SCR for insurers.

The contribution of this paper is three-fold. First, we are among the first in insurance studies to explore the seasonal patterns in monthly mortality data and its impact on insurance liability and solvency capital. Due to data limitation among other challenges, modeling and forecasting monthly mortality are still in its infancy. More accurate forecasts will enable insurance companies to better prepare for seasonal claim experience and thus achieve better financial results. This greater knowledge could also result in an enhanced understanding of how each individual company's claim experience differs, based on their policyholders' exposure to certain climate-sensitive diseases or causes of death. This is an area of great interest to the global insurance industry, and will undoubtedly receive more attention and research in the future.

Our second contribution lies in developing a model framework which nicely integrates seasonal time-series model, EVT, and hierarchical forecast reconciliation. In addition to the pressing need of understanding mortality seasonality, another challenge for life insurers in calculating solvency capital is how to appropriately account for mortality dependence among insured population. Modeling mortality dependence has been studied in depth in recent years (see, e.g., Li and Lee, 2005; Cairns *et al.*, 2011; Li and Hardy, 2011; Chen *et al.*, 2015, 2017). Most of these existing studies examine mortality dependence at either the bottom level (i.e., age-gender-specific mortality) or at the top level (i.e., aggregate mortality). One drawback is that these models do not take into account available information at each hierarchical level, so the modelled dependence structure is not complete. Additionally, if mortality forecasts are produced for individual series at the same level, it is very unlikely that they will add up in the same hierarchical structure as the original data since aggregation constraints are not imposed in the forecasting process.

In this paper, we exploit a MinT hierarchical forecast reconciliation approach to address these issues. Forecast reconciliation is an emerging cutting-edge methodology which allows one to model univariate time series at each level and reconcile forecasts using all available information in the hierarchy. It has gained popularity in supply chain management, renewable energy, tourism modeling, and macroeconomics, but its application to mortality modeling is rather limited (see, e.g., Shang and Hyndman, 2017; Li and Tang, 2019; Li and Hyndman, 2021). Our paper adds to this strand of literature by providing new empirical evidence about the effectiveness of hierarchical forecast reconciliation. Our results indicate that hierarchical forecast reconciliation has a large impact on SCR calculation compared to the unreconciled approach (independently forecasting each data series in the hierarchy) or the bottom-up approach (independently forecasting each data series at the bottom level of the hierarchy and adding these forecasts to form forecasts at higher levels).

We also integrate EVT in the model framework to take into account extreme mortality experience in order to set up sufficient solvency capital.[1]. Such extreme mortality events occur with a very small probability, thus we can hardly learn from the past. EVT offers an appealing solution to this problem. We refer interested readers to Reiss *et al.* (1997) and Embrechts *et al.* (2013) for detailed discussions of EVT and its application to insurance and finance. Beelders and Colarossi (2004) and Chen and Cummins (2010) have used EVT in mortality securitization

[1]There is no formal definition of "extreme mortality experience" in the literature. Usually it refers to some unexpected shocks that can cause a large-scale loss of life, such as widespread pandemics (e.g., 1918 Spanish flu, COVID-19 pademic), man-made disasters (e.g., World War I and II), or natural catastrophes (e.g., hurricane or earthquakes). In this paper, it refers to the 99.5th percentile in the estimated mortality distribution.

models. In a similar vein, we characterize the tail distribution of standardized residuals from the seasonal time-series models using the "Peaks Over Threshold" (POT) method. Thus we can model the mortality data within samples and also rationally extrapolate more extreme, out-of-sample forecasts. Our results show that the EVT approach can improve interval forecast accuracy and thus help insurers set up solvency capital for extreme mortality risk.

Third, our research sheds some light on the significant death tolls caused by the COVID-19 pandemic in the U.S. Using the most recent mortality data until 2020, we forecast death counts and compare them with actual data in 2021 when death tolls surged because of COVID-19. We find that our model with EVT and hierarchical forecast reconciliation can generate a 99.5th percentile forecast that can cover actual death toll for most of the months in 2021, providing additional evidence of the effectiveness of our model for practical uses. Our model also performs better than the unreconciled approach which is more likely to result in an overfunded issue.

The rest of the paper is organized as follows. Section 2 describes and visualizes the mortality data, followed by univariate time series modeling based on seasonal ARIMA models. Section 3 briefly discusses EVT and explains how the POT method is incorporated in our framework. We elaborate the MinT reconciliation approach in Section 4, and illustrate the steps to reconcile probabilistic forecasts on monthly death counts. Section 5 is devoted to the empirical studies and implementation of the proposed framework. Section 6 concludes the paper.

2 Mortality Data

2.1 Hierarchical Monthly Death Counts Data

By convention, most existing studies on mortality modeling consider mortality rate as an annualized probability (see *e.g.* Lee and Carter, 1992; Cairns *et al.*, 2006; Li *et al.*, 2015). More frequent mortality data which contains richer information has not been extensively studied in the past. In this research, we consider U.S. male and female monthly death counts for age groups 35–64 over the period 1968–2019. The data is collected from the National Center for Health Statistics (NCHS), which provides micro-level data on individual deaths since 1959.[2] Information including age, gender, and month of death are extracted from death certificates filed in vital statistics offices in the U.S.

The advantage of working with mortality rates rather than death counts is that mortality rate takes exposure information into account and thus it is more interpretable. However, since population exposure is often not available at a frequency higher than annual, modeling death counts becomes a more direct approach to analyze mortality experience (Rau, 2006). In addition, death counts information is sufficient and more suitable for our application because the number of death is directly related to a life insurer's liability and solvency capital calculation. We therefore focus on monthly death counts data for modelling and forecasting purposes.

Naturally, age-gender-specific death counts can be presented in a hierarchy, as illustrated in Figure 2.1. For age x_n and time t, we denote male and female death counts by $D^M_{x_n,t}$ and $D^F_{x_n,t}$, respectively. We further denote D^M_t and D^F_t as the total male and female death counts for selected age groups at time t, and D_t as the total death counts for these age groups at time

[2]Due to data quality concerns we focus on data from 1968 onwards. See details at: http://www.nber.org/data/vital-statistics-mortality-data-multiple-cause-of-death.html

t. We are not only interested in the age-gender-specific death counts, but also the total death counts from a certain population portfolio.

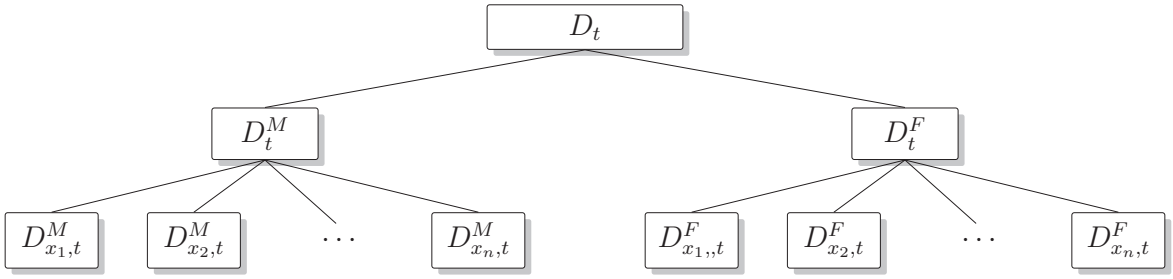


Figure 2.1: 3-level hierarchical tree for death counts

As illustrated in the hierarchy, at any given t, we have the following aggregation constraints:

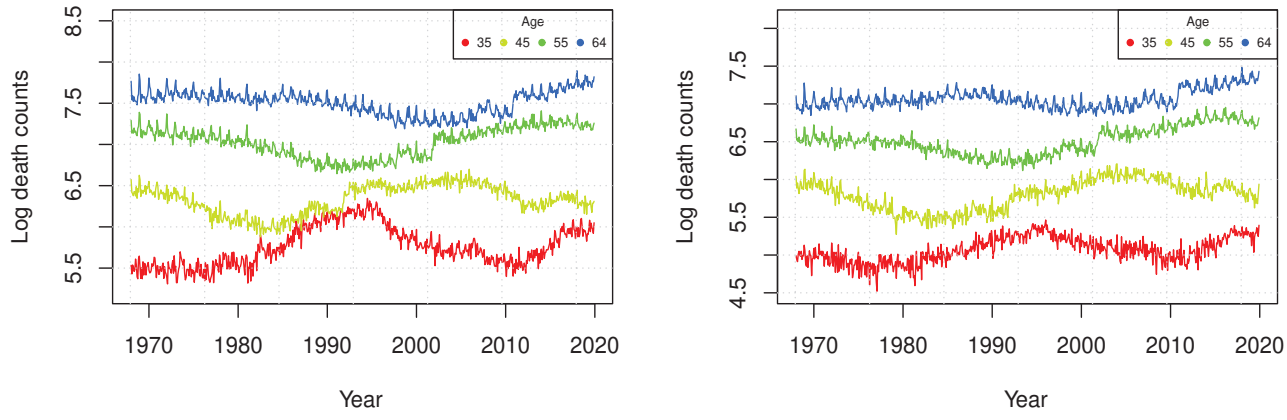
$$\sum_{x=x_1}^{x_n} D_{x,t}^M = D_t^M, \tag{1}$$

$$\sum_{x=x_1}^{x_n} D_{x,t}^F = D_t^F, \tag{2}$$

$$D_t^M + D_t^F = D_t. \tag{3}$$

2.2 Visualization of Seasonality

We plot the U.S. male and female logged death counts for selected ages in Figure 2.2. For logged death counts of the full age range 35–64, please refer to the rainbow plots in Figures A.1 of the Appendix. It can be seen that the overall trend in monthly logged death counts are very similar for both genders.

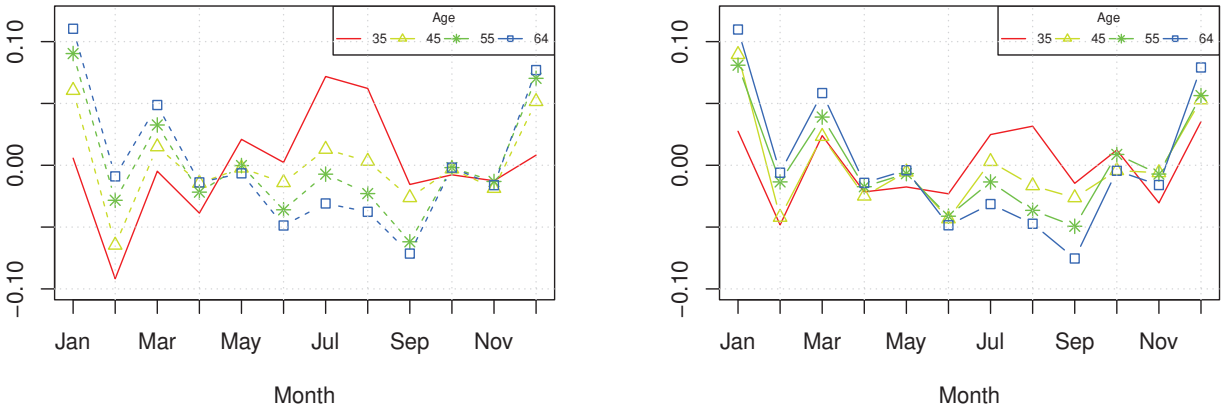


(a) Male

(b) Female

Figure 2.2: Logged monthly death counts for selected ages in the U.S. where ages 35, 45, 55, and 64 are plotted from bottom to the top.

To take a closer look at the seasonality of monthly logged death counts, We use classical seasonal decomposition by moving averages to extract the seasonal component and plot it in Figure 2.3.[3] We see that seasonal patterns across different age groups are not exactly the same. For ages 45, 55, and 64, their pattern are similar with peaks generally observed in winter months (December and January). We also note death peaks in summer months for age 35, but this phenomenon is more pronounced for male than female. This observation is consistent with findings in the literature (Feinstein, 2002; Rau *et al.*, 2017) and can be explained by the increasing number of injuries, especially from road traffic crashes, in the summer (Liu *et al.*, 2005), which are more common to men (Parks *et al.*, 2018).



(a) Male

(b) Female

Figure 2.3: Seasonality of logged death counts for selected ages in the U.S.

2.3 Seasonal ARIMA Models

In this work, we adopt a seasonal ARIMA model, which incorporates both non-seasonal and seasonal factors, to fit the monthly logged death counts series. The model is expressed as follows:

$$\Phi_P(B^s)\phi_p(B)\nabla_s^D\nabla^d d_t = \theta_q(B)\Theta_Q(B^s)\epsilon_t, \tag{4}$$

where

- ϕ_p, ∇^d, and θ_q denote the non-seasonal AR operator of order p, differencing operator of order d, and MA operator of order q, respectively,
- Φ_P, ∇_s^D, and Θ_Q denote the seasonal AR operator of order P, differencing operator of order D, and MA operator of order Q, respectively,
- s is the time span of repeating seasonal pattern, which is set to 12 in this study, and
- d_t is the death count time series under investigation, and ϵ_t is white noise process.

[3]This method decomposes a time series into trend, seasonal, and noise components using moving averages. The decomposition is performed by function "decompose" in R. The function first determines the trend component using a moving average, and removes it from the time series. The seasonal component is found by grouping the results by month and averaging them. Finally, the noise component is determined by removing trend and seasonal components from the original time series.

In total, we work on 63 logged death counts time series, including 60 age-gender-specific death counts (age 35-64 for male and female), total male, total female, and total death counts for these selected age groups. The seasonal ARIMA model for each time series is selected based on Akaike information criterion (Akaike, 1974). The selected ARIMA orders are shown in Table A.1 of the Appendix.[4]

3 An EVT Approach to Error Distribution

In the previous section, we use seasonal ARIMA model to fit the monthly logged death counts series and obtain residuals ϵ_t. Figure 3.1 compares the empirical CDF of the standardized residuals for total death counts with the CDF of a standard normal distribution. It is clear that the distribution of the standardized residuals for total death counts is fat tailed. A similar pattern of fat tail is observed for female and male total death counts residual series, as shown in Figure 3.2. For 60 age-gender-specific residual series, we conduct the Shapiro-Wilk test of normality and report the results in Table A.2 in the Appendix. We cannot reject the null hypothesis of normality for 17 age-gender-specific residual series, thus we assume a standard normal distribution for them. For the remaining 43 residual series, as well as for female, male, and total death counts residuals, we follow Chen and Cummins (2010) to model small variations using a standard normal distribution and capture extreme mortality risk using EVT.

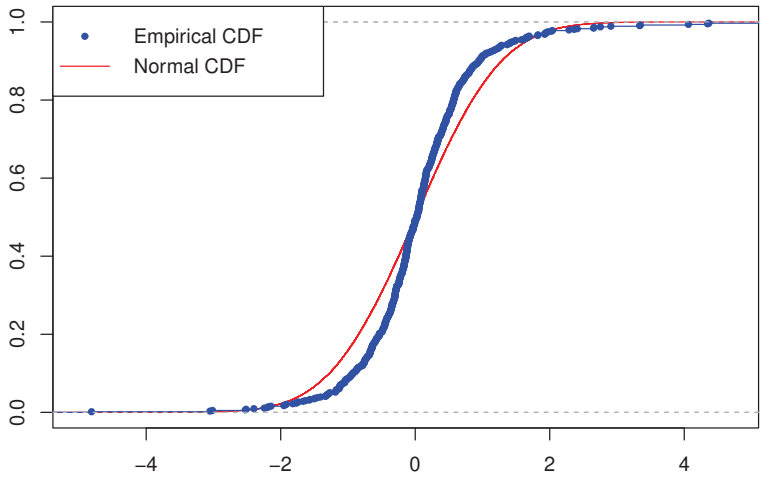
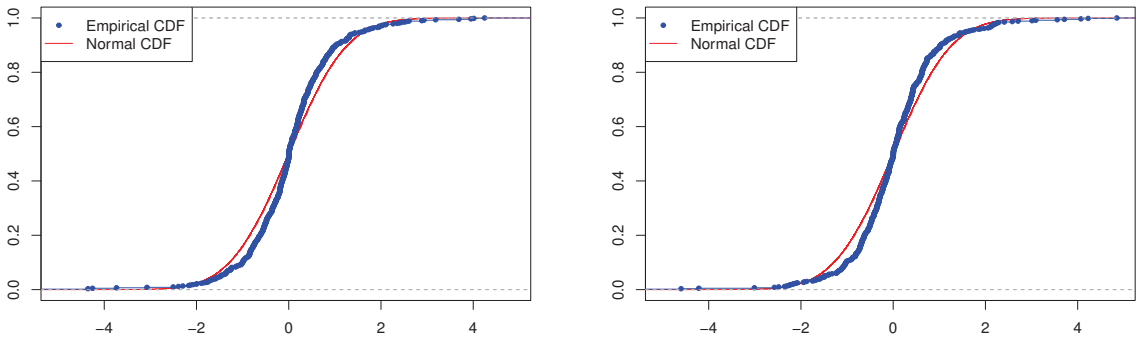


Figure 3.1: Empirical CDF of standardized residuals for total deaths

.

Denote X the standardized residuals. We partition the distribution of X by a high threshold μ, where X follows a standard normal distributions when $X < \mu$ and a distribution driven by EVT when $X \geq \mu$. The threshold μ is chosen based on the trade-off between bias and variance. On the one hand, a high threshold is necessary because it ensures the asymptotic property of EVT and reduces the bias as a result. On the other hand, if we set the threshold too high, we would not have enough data to estimate the parameters which leads to an increase in variance. A typical choice of μ is around the 90th to 95th percentile (Sanders, 2005). In this study, we choose μ to be the 95th percentile.

[4] `R` package `forecast` (Hyndman and Khandakar, 2008) is used to select the optimal models and conduct residual tests.



(a) Male

(b) Female

Figure 3.2: Empirical CDF of standardized residuals for total male and female deaths

Specifically, for $x < \mu$ we assume that,

$$F_1(x) = P\{X \leq x\} = \Phi(x), \tag{5}$$

where $\Phi(x)$ is the distribution function for a standard normal random variable.

For $x \geq \mu$, we employ an EVT approach. There are two main methods in EVT, namely "Block Maxima" (BM) and "Peaks Over Threshold" (POT). The BM approach partitions time series of data into equal blocks of the same size, and for each block gets single maximum value, i.e., block maxima. The resulting BM time series are then used to model extreme value behavior. The POT approach chooses a threshold and selects values higher than this threshold to model the tail distribution. We select the POT method in this paper mainly for two reasons. First, since POT uses all large observations over a threshold while BM estimators may miss some large observations falling into the same block, it is of general consensus that the POT method produces more efficient estimators than the BM method (see Ferreira and De Haan, 2015, and references therein). Second, we are interested in estimating the solvency capital requirement based on Value-at-Risk(VaR), so we choose the POT method which by nature is more preferable for quantile estimation.

Let x_0 be the finite or infinite right endpoint of the distribution $F(x)$. We define the conditional distribution of the exceedances over a high threshold μ by

$$F_\mu(x) = P\{X - \mu \leq x | X > \mu\} = \frac{F(x+\mu) - F(\mu)}{1 - F(\mu)}, \tag{6}$$

for $0 \leq x < x_0 - \mu$.

According to the Pickands-Balkema-de Haan Theorem (Balkema and de Haan, 1974; Pickands, 1975), for a sufficiently high threshold μ, the excess distribution function $F_\mu(x)$ may be approximated by the generalized Pareto distribution (GPD), $G_{\gamma,\sigma}(x)$, for some value of γ and σ. Here $G_{\gamma,\sigma}(x)$ is defined as

$$G_{\gamma,\sigma}(x) = \begin{cases} 1 - \left(1 + \gamma\frac{x}{\sigma}\right)^{-1/\gamma} & \text{if } \gamma \neq 0 \\ 1 - \exp\left(-\frac{x}{\sigma}\right) & \text{if } \gamma = 0 \end{cases} \tag{7}$$

where γ is the shape parameter of the distribution and $\sigma > 0$ is an additional scaling parameter. When $\gamma > 0$ we have a reparameterized version of the ordinary Pareto distribution, $\gamma = 0$ corresponds to the exponential distribution, and $\gamma < 0$ is known as a type II Pareto distribution.

For $x \geq \mu$, the distribution function can be written as

$$F_2(x) = P\{X \leq x\} = (1 - F(\mu))F_\mu(x - \mu) + F(\mu). \tag{8}$$

We can approximate $F_\mu(x-\mu)$ by the generalized Pareto distribution $G_{\gamma,\sigma}(x-\mu)$. We can also approximate $F(\mu)$ by the empirical distribution $F_n(\mu)$. This means when $x \geq \mu$ we can use the tail estimate

$$\hat{F}_2(x) \approx (1 - F_n(\mu))G_{\gamma,\sigma}(x - \mu) + F_n(\mu) = 1 - (1 - F_n(\mu)) \left(1 + \gamma\frac{x-\mu}{\sigma}\right)^{-1/\gamma} \tag{9}$$

to approximate the distribution function $F(x)$.

For a given threshold μ, parameters can be estimated by maximum likelihood estimation. We add a constraint $F_1(\mu) = F_2(\mu) = F_n(\mu)$ to ensure the distribution function is continuous at the threshold μ.

4 MinT hierarchical Forecast Reconciliation

After we forecast each mortality time series in the hierarchy using the seasonal ARIMA model with the EVT error distribution, we need to reconcile these forecasts by imposing aggregation constraints (1)–(3).

While these aggregation constraints are satisfied by observations of death counts shown in Figure 2.1, it is highly unlikely that the forecasts of these time series will also add up in the same way. If we employ a "bottom-up" or "top-down" approach in forecasting these hierarchical time series, which only looks at data at the bottom level or the top level, we end up ignoring important information from other levels in the hierarchy.[5] Therefore, in this research we consider a forecast reconciliation approach which adjusts hierarchical time series forecasts across all levels such that the underlying aggregation constraints are met.

Hierarchical forecast reconciliation is a fast growing research area in recent decades (see, e.g., Zellner and Tobias, 2000; Athanasopoulos *et al.*, 2009; Hyndman *et al.*, 2011; Wickramasuriya *et al.*, 2019; Panagiotelis *et al.*, 2021). It has been applied in a wide range of fields including economics, engineering, and supply-chain management (see, e.g., Capistrán *et al.*, 2010; Borges *et al.*, 2013; Syntetos *et al.*, 2016), and has recently started to see applications in mortality forecasts, including regional-level mortality forecast (Shang and Hyndman, 2017; Shang and

[5]The top-down method generates the top-level forecasts and then produce the bottom-level forecasts based on disaggregation proportions, which dictate how the forecasts of the top-level series are to be distributed to obtain forecasts for each series at the bottom-level of the structure. The disaggregation proportions can be determined by either historical proportions of the data or the proportions of forecasts. Once the bottom-level forecasts have been generated, these are aggregated to generate coherent forecasts for the middle levels in the hierarchy. The bottom-up approach involves first generating forecasts for each series at the bottom-level, and then summing these to produce forecasts for all the series in the structure. See Hyndman and Athanasopoulos (2018), Chapter 10 for details.

Haberman, 2017; Li and Hyndman, 2021), cause-of-death mortality forecast (Li *et al.*, 2019), and mortality improvement rates forecast (Li and Tang, 2019).

As reconciliation process ensures coherency across forecasts, it aids aligned decision making. Moreover, forecast reconciliation provides an alternative way to handle dependence structure in forecasts across large numbers of time series. Basically, one can first obtain independent forecasts for each time series in a hierarchical structure, and then reconcile those forecasts according to aggregate constraints and the dependence structure of forecasting errors. This allows us to circumvent the need for heavily parameterized and possibly misspecified joint models subject to the curse of dimensionality. Finally, by fully utilizing all available information in the hierarchy, improvements in overall forecast accuracy can be achieved via the reconciliation process. Studies have shown that forecast reconciliation can improve overall prediction accuracy and can be applied irrespective of how base forecasts are produced.

In this work, we adopt the trace minimization (MinT) reconciliation method introduced by Wickramasuriya *et al.* (2019). The key idea of the MinT approach is to minimize the trace of the variance-covariance matrix of the in-sample forecast errors. Wickramasuriya *et al.* (2019) improve on the OLS reconciliation method proposed by Hyndman *et al.* (2011) and demonstrate superior forecasting performance in various applications. Before introducing the MinT reconciliation method, we first express aggregation constraints (1)–(3) in a matrix form and define the following notations which will be used in this section:

- Define $\boldsymbol{y}_t = (D_t,\, D_t^M,\, D_t^F,\, D_{x_1,t}^M,\, \ldots,\, D_{x_n,t}^M,\, D_{x_1,t}^F,\, \ldots,\, D_{x_n,t}^F)'$ as a vector that contains observations at all levels in the hierarchy;
- Define $\boldsymbol{b}_t = (D_{x_1,t}^M,\, \ldots,\, D_{x_n,t}^M,\, D_{x_1,t}^F,\, \ldots,\, D_{x_n,t}^F)'$ as a vector that contains observations at the bottom level only.

The hierarchical structure can be presented by linking the two vestors via the following equation

$$\boldsymbol{y}_t = \boldsymbol{S}\,\boldsymbol{b}_t, \tag{10}$$

where $\boldsymbol{S}$ is a "summing matrix" of dimension $(3+2n)\times 2n$, which aggregates age-gender-specific death counts to death counts at higher levels. It is given by

$$\boldsymbol{S} = \begin{pmatrix} 1 & 1 & 1 & \ldots & \ldots & 1 & 1 & 1 \\ 1 & 1 & \ldots & 1 & 0 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 & 1 & 1 & \ldots & 1 \\ & & & & & & & \\ & & & & \boldsymbol{I}_{2n} & & & \end{pmatrix},$$

where $\boldsymbol{I}_{2n}$ denotes an identity matrix of dimension $2n \times 2n$. Therefore, the aggregation constraints (1)–(3) are reflected in the first three rows of the $\boldsymbol{S}$ matrix.

Let $\hat{\boldsymbol{y}}_{T+h}$ be a vector of h-step-ahead independent forecasts of all series in the hierarchy, referred to as base forecasts. We produce these base forecasts using the univariate seasonal ARIMA models described in Section 2.3. Denote $\tilde{\boldsymbol{y}}_{T+h}$ a vector of forecasts for all levels which satisfy the aggregation constraints, referred to as reconciled forecasts. We express the reconciliation process as

$$\tilde{\boldsymbol{y}}_{T+h} = \boldsymbol{S}\boldsymbol{P}\hat{\boldsymbol{y}}_{T+h}, \tag{11}$$

for some matrix $\boldsymbol{P}$ of dimension $2n \times (2n+3)$. It is important to note that the choice of $\boldsymbol{P}$ is not unique. For example, if we simply add up the bottom level forecasts to form forecasts for higher levels, $\boldsymbol{P}$ is chosen as

$$\boldsymbol{P} = (\boldsymbol{0}_{2n\times 3}, \boldsymbol{I}_{2n}), \tag{12}$$

where $\boldsymbol{0}_{2n\times 3}$ is a zero matrix of dimension $2n \times 3$. The hierarchical forecasting reconciliation method becomes simply the bottom-up method.

For the MinT approach, the reconciliation matrix $\boldsymbol{P}$ is given by[6]

$$\boldsymbol{P} = (\boldsymbol{S}'\boldsymbol{W}_h^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{W}_h^{-1}, \tag{13}$$

where $\boldsymbol{W}_h$ is a positive definite covariance matrix of the h-step-ahead base forecast errors defined as

$$\boldsymbol{W}_h = \mathbb{E}[\hat{\boldsymbol{e}}_{T+h}\hat{\boldsymbol{e}}'_{T+h}|\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_t], \tag{14}$$

and $\hat{\boldsymbol{e}}_{T+h} = \boldsymbol{y}_{T+h} - \hat{\boldsymbol{y}}_{T+h}$.

Thus, we obtain reconciled forecasts as

$$\tilde{\boldsymbol{y}}_{T+h} = \boldsymbol{S}(\boldsymbol{S}'\boldsymbol{W}_h^{-1}\boldsymbol{S})^{-1}\boldsymbol{S}'\boldsymbol{W}_h^{-1}\hat{\boldsymbol{y}}_{T+h}, \tag{15}$$

Since $\boldsymbol{SPS} = \boldsymbol{S}$, the reconciled forecasts are shown to be unbiased given that base forecasts are also unbiased (Wickramasuriya *et al.*, 2019).

Estimating $\boldsymbol{W}_h$ is nevertheless challenging for a large collection of time series, especially when $h > 1$. Wickramasuriya *et al.* (2019) discuss a few alternative estimates including a shrinkage type estimator.[7] They illustrate that using the shrinkage estimator generally leads to substantial improvements in forecast accuracy. We therefore use the shrinkage estimator in our paper.

The MinT approach was originally introduced with a focus on reconciling point forecasts. As we want to calculate the SCR for life insurance companies which requires interval forecasts of death counts at the 99.5th percentile (*i.e.* the point at which the probability that a random event exceeds this value is 0.5%), we adopt the "ranked sample" approach introduced by Jeon *et al.* (2019) together with the MinT reconciliation approach. The basic idea of a ranked sample approach for probabilistic forecast reconciliation is to first generate forecasts from the predictive probabilistic distribution, rank these forecasts from the lowest to the highest value (or vice versa), and then reconcile these ranked forecasts. The "ranked sample" approach has demonstrated strong performance particularly when the underlying time series are positively correlated. In our data series, we expect a positive dependence structure for death counts across age and gender.

To illustrate this method, we further define the following terms:

- Let $\boldsymbol{f}(\boldsymbol{y}_{T+h}|\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_T)$ be the predictive probabilistic distribution of h-step-ahead forecasts;

[6]For a detailed proof of this result, please refer to Wickramasuriya *et al.* (2019), Appendix A.1.

[7]Basically, the shrinkage estimator is a weighted combination between the full covariance matrix of the in-sample one-step-ahead base forecast errors and the diagonal matrix comprising the diagonal elements of the full covariance matrix. We refer interested readers to Section 2.4 in Wickramasuriya *et al.* (2019).

- Let $\hat{\boldsymbol{y}}_{T+h}^{k}$ be the k^{th} sample of unreconciled forecasts generated from the predictive probabilistic distribution $\boldsymbol{f}(\boldsymbol{y}_{T+h}|\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_T)$;
- Let $\hat{\boldsymbol{Y}}_{T+h}$ be a sample of size N generated from $\boldsymbol{f}(\boldsymbol{y}_{T+h}|\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_T)$, where $\hat{\boldsymbol{Y}}_{T+h} = (\hat{\boldsymbol{y}}_{T+h}^{1}, \hat{\boldsymbol{y}}_{T+h}^{2}, \ldots, \hat{\boldsymbol{y}}_{T+h}^{N})$.

Similar to MinT reconciliation of point forecasts, we can express the process as

$$\tilde{\boldsymbol{Y}}_{T+h} = \boldsymbol{SP}\hat{\boldsymbol{Y}}_{T+h}, \tag{16}$$

where $\boldsymbol{S}$ and $\boldsymbol{P}$ are as previously defined, and $\tilde{\boldsymbol{Y}}_{T+h}$ denotes the reconciled forecasts for N sample paths.

5 Empirical Study

We present results from the empirical study based on the monthly death counts data in the U.S. described in Section 2. First, we showcase the strong forecasting performance of the MinT approach combined with EVT via a cross-validation exercise. Our training dataset goes from 1968 to 2009 (504 months), and the validation dataset goes from 2010 to 2019 (120 months). We then move on to the SCR calculation, and quantify the impact of EVT and reconciliation on solvency capital requirement. Finally, we apply the proposed approach to project death counts during the COVID-19 pandemic period.

5.1 Point Forecast Accuracy Comparison

We first fit seasonal ARIMA models to the 63 death counts series in our holdout sample from 1968 to 2009 as described in Section 2. We then forecast death counts for each month in the next ten years. We evaluate the h-step-ahead point forecast accuracy using the mean absolute percentage error (MAPE), which is defined as follows

$$\text{MAPE}_{T+h} = \frac{1}{h}\sum_{t=1}^{h}\left|\frac{\hat{D}_{T+t} - D_{T+t}}{D_{T+t}}\right|, \tag{17}$$

where T denotes the end of year 2009 and h increases by month. A smaller MAPE value indicates a better forecast accuracy.

Table 5.1 reports the MAPE based on three methods: the unreconciled (Unrec) approach, the bottom-up (BU) approach, and the MinT reconciliation approach. To save space, we only report MAPE values at the end of each forecasting year (i.e., $h = 12, 24, ..., 120$).[8] As the error distributions in the seasonal ARIMA models do not affect point forecasts, we do not report results from different error distribution assumptions in this table. We consider all three hierarchical levels, including total deaths, male or female total deaths, and age-gender-specific deaths. The smallest MAPE values for each forecast horizon among the three models (indicating the best performing model) are highlighted in bold. We can see that overall MinT outperforms other two methods for point forecasts.

[8]Statistics for monthly forecasts are available from the authors.

Table 5.1: MAPE (%) of out-of-sample point forecasts for period 2010–2019.

	Total			**Male**			**Female**			**Male-Age**			**Female-Age**		
Year	**Unrec**	**BU**	**MinT**	**Unrec**	**BU**	**MinT**	**Unrec**	**BU**	**MinT**	**Unrec**	**BU**	**MinT**	**Unrec**	**BU**	**MinT**
1	1.8	1.8	**1.5**	1.8	**1.5**	1.7	2.0	**1.5**	1.8	**5.6**	**5.6**	5.7	7.3	7.3	**6.4**
2	1.5	1.4	**1.4**	1.6	**1.5**	1.6	1.6	1.7	**1.6**	**6.5**	**6.5**	6.9	7.8	7.8	**7.1**
3	1.7	1.7	**1.4**	1.9	1.9	**1.7**	1.9	**1.7**	1.8	**7.5**	**7.5**	7.9	9.0	9.0	**7.8**
4	2.0	1.9	**1.7**	2.1	2.3	**2.0**	2.2	2.0	**2.0**	**8.6**	**8.6**	9.1	10.1	10.1	**8.4**
5	2.2	2.1	**1.8**	2.4	2.9	**2.2**	2.4	2.4	**2.2**	**9.7**	**9.7**	9.9	10.3	10.3	**9.0**
6	2.2	2.1	**1.7**	2.4	3.2	**2.2**	2.4	3.0	**2.1**	**10.1**	**10.1**	10.3	11.5	11.5	**9.5**
7	2.1	2.1	**1.7**	2.3	3.5	**2.1**	2.4	3.6	**2.1**	**10.7**	**10.7**	11.0	12.0	12.0	**10.0**
8	2.2	2.1	**1.7**	2.3	4.0	**2.1**	2.5	4.1	**2.2**	**11.3**	**11.3**	11.5	12.8	12.8	**10.6**
9	2.6	2.5	**2.1**	2.6	4.8	**2.4**	3.0	4.3	**2.7**	**11.9**	**11.9**	12.4	13.7	13.7	**11.3**
10	3.0	2.9	**2.5**	2.9	5.7	**2.7**	3.6	4.4	**3.3**	**12.6**	**12.6**	13.9	14.0	14.0	**12.0**

Note: Bold numbers indicate the smallest MAPE values for each forecast horizon among the three models.

For total number of deaths, the MinT reconciliation is the best performing method while the unreconciled method gives the worst forecast performance. At the middle level, the MinT reconciliation method still dominates the other two method in most cases, while the bottom-up method provides the best forecast results within two or three years of the forecast horizon. For the gender-age-specific death counts, it is noteworthy that the unreconciled method and the bottom-up method provide the same MAPE values as these two methods generate the bottom-level forecasts in the same way. They are superior to the MinT method for male-age-specific death forecasts, but the MinT method perform better for female-age-specific death forecasts. This is not surprising as the MinT approach achieves optimality by minimizing the trace of the covariance matrix of forecast errors for the entire hierarchy, not each individual level of the hierarchy (Panagiotelis *et al.*, 2021).

5.2 Interval Forecast Accuracy Comparison

In this section we compare the interval forecast accuracy of the proposed model with results from alternative models. To achieve this goal, we utilize the Winkler score for prediction intervals (Winkler, 1972). The Winkler score is calculated as

$$W_{T+h} = (u_{T+h} - l_{T+h}) + \frac{2}{\alpha}(l_{T+h} - D_{T+h})\mathbb{1}(D_{T+h} < l_{T+h}) + \frac{2}{\alpha}(D_{T+h} - u_{T+h})\mathbb{1}(D_{T+h} > u_{T+h}). \tag{18}$$

where $[l_{T+h}, u_{T+h}]$ is the 100(1-α)% prediction interval for h-step-ahead forecast, D_{T+h} is the actual observed value, and $\mathbb{1}$ denotes the indicator function. In our experiment, we choose α to be 1%, meaning that we compare forecast accuracy of 99% prediction intervals (to be consistent with the 99.5th percentile we will use in Section 5.3). Table 5.2 presents the Winkler score based on the unreconciled forecasts, bottom-up forecasts, and MinT reconciled forecasts. We present the results under both the empirical error distribution and the EVT error distribution. The smallest Winkler scores for each forecasting horizon across different models and different error assumptions are highlighted in bold.

First, it can be seen that the incorporation of EVT improves the overall forecast accuracy except for a few cases, compared to using the empirical error distribution. Combined with the improvement achieved by EVT, we find that the MinT approach gives the smallest Winkler scores most of the time for total, male (and female) total, and female age-specific forecasts. For male age-specific forecasts, the unreconciled and bottom-up forecasts perform slightly better, consistent with our results from point forecasts. To sum up, the combination of EVT errors and MinT reconciliation greatly improves accuracy of the interval forecasts.

5.3 SCR Calculation

Capital requirements constitute the foundation for the financial regulation of insurance companies as in the banking industry. Insurers that fail to meet the solvency standards are subject to regulatory intervention. Informal actions may include regulatory inquiries, meetings with an insurer's management teams, and corrective action plans. Formal interventions often necessitates regulators seizing control of a company and instituting conservation, rehabilitation, and liquidation actions depending on the condition of the insurer and its prospects (Klein, 2012).

In the U.S., insurers are subject to fixed minimum capital requirements set by each state as well as risk-based capital (RBC) standards promulgated by the National Association of Insurance Commissioners (NAIC) (Klein, 2012). RBC is an extensive factor-based approach, where

Table 5.2: Winkler score of out-of-sample forecasts for period 2010–2019.

	Year	Unrec		BU		MinT	
		Empirical	**EVT**	**Empirical**	**EVT**	**Empirical**	**EVT**
Total	1	6711	**6062**	15081	14255	6491	6304
	2	7346	**6685**	19855	16786	6979	6763
	3	7976	7492	24251	19379	7512	**7258**
	4	8646	8354	29642	22040	8093	**7787**
	5	9373	9286	30955	24747	8736	**8369**
	6	10134	10255	33337	27504	9443	**9009**
	7	10894	11248	38038	30322	10171	**9648**
	8	11712	12235	42408	33193	10928	**10289**
	9	12539	13272	44893	36121	11742	**10967**
	10	13381	14313	50199	39105	12589	**11648**
Male	1	3942	**3659**	7241	7228	3965	3681
	2	4275	4008	8505	8207	4383	**4003**
	3	4625	4399	11088	8975	4813	**4327**
	4	5018	4805	14180	9778	5269	**4679**
	5	5447	5259	17763	11488	5754	**5063**
	6	5910	5738	21913	13501	6273	**5482**
	7	6405	6241	26738	15437	6797	**5905**
	8	6948	6776	32425	19548	7346	**6344**
	9	7527	7350	39187	20711	7900	**6804**
	10	8128	7941	45373	21913	8472	**7264**
Female	1	**1375**	2720	6840	6023	1832	1719
	2	2258	1996	8350	7076	1989	**1838**
	3	2635	**1850**	10163	8236	2318	2009
	4	2936	2258	12461	9510	2459	**2217**
	5	3270	5655	15192	10858	2627	**2448**
	6	3756	5144	18424	12269	2813	**2696**
	7	4181	4846	22300	13741	3017	**2955**
	8	4667	4660	26983	15269	3238	**3221**
	9	5159	5134	32705	16854	3494	**3471**
	10	5562	5036	39827	18491	3773	**3722**
Male-Age	1	2056	**1132**	2056	**1132**	1375	1312
	2	2480	**1935**	2480	**1935**	2363	2258
	3	2508	**2232**	2508	**2232**	2856	2635
	4	2634	**2485**	2634	**2485**	3265	2936
	5	2834	**2772**	2834	**2772**	3340	3270
	6	**2989**	3214	**2989**	3214	3756	3613
	7	**3264**	3628	**3264**	3628	4181	4089
	8	**3682**	4086	**3682**	4086	4667	4242
	9	**4147**	4540	**4147**	4540	5159	4628
	10	**4678**	4933	**4678**	4933	5562	4835
Female-Age	1	592	517	592	517	657	**340**
	2	667	641	667	641	709	**450**
	3	713	673	713	673	802	**456**
	4	715	707	715	707	947	**458**
	5	808	737	808	737	1028	**475**
	6	889	738	889	738	1047	**497**
	7	986	739	986	739	1136	**533**
	8	1053	740	1053	740	1261	**578**
	9	1072	746	1072	746	1286	**629**
	10	1174	788	1174	788	1395	**685**

Note: Bold numbers indicate the smallest Winkler scores for each forecast horizon among the three models under two error distribution assumptions.

selected factors are multiplied by various accounting values (e.g., assets, liabilities, or premiums) to produce RBC charges for each item, which are summed into several "baskets" with a covariance adjustment. Separate formulas were developed for life insurers, property-casualty insurers, and health insurers, but life RBC formula is the only one incorporating some models-based components, largely driven by the recognition that "life insurers were developing products with increasingly complex guarantees and such risks were not captured by the basic factor-based capital requirements" (Vaughan, 2009).

In the EU, the required capital level, called the solvency capital requirement (SCR), is determined by Solvency II which came in force from 2016. Solvency II is based on a three-pillar approach similar to the banking industry's Basel II, including (1) quantitative requirements for measuring capital adequacy, (2) a supervisory review process, and (3) increased transparency and reporting requirements. The SCR considers all risk categories an insurer faces, including market risk, credit risk, non-life, life, and health underwriting risk, as well as operational risk. It is derived by using a value-at-risk (VaR) calibration at a 99.5 percent level over a one-year time horizon (or a 1-in-200-year event). The company may calculate the SCR by either standard formulas or with a regulator-approved internal model. In exchange for simplicity, the standard formulas may implicitly contain substantial margins which can result in higher capital requirements than those calculated using internal models (Silverman and Simpson, 2011).

In this section, we estimate the capital requirement for a hypothetical life insurer based on the SCR prescribed by internal models under Solvency II. One reason for us to use U.S. data to estimate the SCR under Solvency II is because we do not have access to a granular mortality data with monthly frequency for any of the EU countries. There are, however, other compelling reasons for us to believe that it is critical for U.S. insurers to comprehend Solvency II capital requirement standard and compare their own capital levels to the SCR.

On the one hand, though the solvency regulation in the U.S. has evolved towards internal models in life insurance, it has been incremental and supplemental (Vaughan, 2009). Some scholars (see, e.g., Holzmüller, 2009; Cummins and Phillips, 2009) have criticized the reliance on static factor-based formulas in the U.S. system and its failure to make more extensive use of stochastic modeling and scenario testing. Cummins and Phillips (2009) also contend that U.S. RBC takes a "one-size-fits-all" approach contrary to Solvency II that can be geared to individual company characteristics.

On the other hand, Solvency II's impact extends well beyond the EU to the U.S. insurance market. First of all, EU subsidiaries of a U.S. insurer are required to comply with Solvency II including solvency capital requirement. Second, Solvency II recognizes the regulatory regimes in other countries if they meet the "equivalence" principles. Without equivalence, third-country subsidiaries will have to comply with Solvency II. The U.S. has been granted provisional equivalence with regard to solvency calculations for 10 years from January 1, 2016. This means the U.S. subsidiaries with EU parents could base their required capital on U.S. RBC model. Meanwhile, it implies that U.S. solvency regime does not completely meet Solvency II equivalence requirements and whether it is renewable after 10 years is questionable. Third, Solvency II is likely to raise the bar for capital requirements and risk management practices for all insurers. This will be fueled by regulators and rating agencies and thus have a bigger market impact.

We assume that a large life insurer writes n-year term-life insurance policies with face value of

$500,000 at the end of our training period. Death benefits are paid at the end of the month when the policyholder dies. The hypothetical portfolio consists of policyholders aged 35-64 at the time of issuance, with the age-gender composition of the portfolio proportional (set to be 6%) to that of the U.S. population.[9] This is a simplifying assumption mainly because our data is population death counts instead of actual mortality experience of an insurer's portfolio. However, our model framework can be readily extended to insurers' own mortality data once such data is available.

For each age x at time T, we simulate the path of death counts at the end of each month $T+h$, denoted by $D^M_{x,T+h}$ or $D^F_{x,T+h}$ for 10,000 times. For each path, we calculate the insurer's liability at $T+h$ $(h = 1, 2, \ldots, 120)$

$$\tilde{L}_{T+h} = 0.06 \times 500000 \sum_{x=35}^{64} (D^M_{x,T+h} + D^F_{x,T+h}). \tag{19}$$

Discounting liabilities back to time T using an appropriate discount rate r, we can calculate the present value of liabilities at time T,

$$\tilde{L}_T = \sum_{h=1}^{12n} \tilde{L}_{T+h}(1+r)^{-h}, \tag{20}$$

and the best estimate liability at time T

$$L^{BE}_T = \sum_{h=1}^{12n} E_T\left[\tilde{L}_{T+h}\right](1+r)^{-h}. \tag{21}$$

Then the SCR for mortality risk is calculated as follows,

$$SCR_T = VaR_{0.995}\left(\tilde{L}_T\right) - L^{BE}_T. \tag{22}$$

Table 5.3 reports the SCR calculation based on different interest rates and different terms of life insurance contracts. We can see that the SCR calculated by using EVT errors and MinT forecasts is the highest in each scenario. Let's take a look at 10-year term life insurance and 1% interest rate, for example. Under the empirical error distribution, the MinT approach raises the SCR by around 3.2% when compared to the unreconciled approach and about 16.7% when compared to the bottom-up approach. Under the EVT error distribution, the SCR grows 16.7% using MinT v.s. the unreconciled approach and about 26.5% using MinT v.s. the bottom-up approach. Another notable finding is that modeling the errors with EVT significantly boosts the SCR since the fat tail of the distribution is taken into consideration. When we focus on the MinT technique but move from the empirical error distribution to the EVT error distribution, the SCR rises from $25.198 bn to $31.974 bn, a 26.9% increase. The implication is that life insurers will significantly underestimate capital requirement related to mortality risk and thus may not be able to meet their financial obligations if they fail to model mortality dependence and the extreme mortality risk appropriately.

[9]According to LIMRA's 2020 Insurance Barometer Study, 54 percent of all people in the United States were covered by some type of life insurance. The total percentage of market penetration for the life insurance industry is relatively stable, though it has been trending downward over the past decade. MetLife is the largest life insurer according to direct premiums written, accounting for 13% of the total market share. This means that MetLife insures about 7% of U.S. population if we assume these numbers are proportional in terms of the age-gender composition. We assume 6% in our example just for illustration purposes.

Table 5.3: SCR ($bn) based on monthly death counts data

3-year term life insurance						
Interest rate	**Empirical error**			**EVT error**		
	Unrec	**BU**	**MinT**	**Unrec**	**BU**	**MinT**
1%	4.424	2.087	4.740	5.231	2.114	**5.280**
3%	4.287	2.015	4.592	5.068	2.040	**5.114**
5%	4.158	1.946	4.452	4.914	1.972	**4.958**
5-year term life insurance						
Interest rate	**Empirical error**			**EVT error**		
	Unrec	**BU**	**MinT**	**Unrec**	**BU**	**MinT**
1%	8.235	5.224	8.898	9.895	5.229	**9.960**
3%	7.806	4.908	8.430	9.373	4.914	**9.435**
5%	7.412	4.619	8.002	8.896	4.626	**8.954**
10-year term life insurance						
Interest rate	**Empirical error**			**EVT error**		
	Unrec	**BU**	**MinT**	**Unrec**	**BU**	**MinT**
1%	24.412	21.600	25.198	26.788	25.285	**27.399**
3%	21.762	19.146	21.983	23.921	22.052	**24.412**
5%	19.271	18.052	19.500	21.471	19.326	**21.863**

Note: Bold numbers indicate the highest SCR value for each interest rate level among the three models under two error distribution assumptions.

Table 5.4: SCR ($bn) based on annual death counts data

3-year term life insurance			
Interest rate	**Unrec**	**BU**	**MinT**
1%	3.143	1.402	**3.198**
3%	3.008	1.343	**3.059**
5%	2.881	1.287	**2.929**
5-year term life insurance			
Interest rate	**Unrec**	**BU**	**MinT**
1%	7.337	2.922	**7.450**
3%	6.843	2.564	**6.948**
5%	6.396	2.273	**6.493**
10-year term life insurance			
Interest rate	**Unrec**	**BU**	**MinT**
1%	22.110	10.072	**23.736**
3%	19.314	8.821	**20.723**
5%	16.957	7.766	**18.183**

Note: Bold numbers indicate the highest SCR value for each interest rate level among the three models.

With monthly death counts data, we are able to model and forecast the number of deaths in each month, achieving a higher precision in insurance pricing, reserve setting, and solvency

capital calculation. If annual data is used, a commonly used approach is to assume the death and thus the payment of death benefits occurring at the end of the year. Otherwise we will have to assume some fractional life time function, e.g., uniform distribution of death or constant force of mortality, neither of which captures the seasonality of mortality. Another advantage of using monthly death counts data is to allow us to have enough data points to estimate the extreme mortality risk using EVT.

In Table 5.4, we report the SCR calculated based on annual death counts data for the purpose of comparison. We assume the death occurs at the end of each year, immediately followed by the payment of death benefits. We sample from the empirical distributions of the residuals from time-series models as we do not have enough data points for EVT estimation. Under these assumptions, the SCR calculated using annual data is always lower than that calculated using monthly data, suggesting an underestimation of the SCR. Continuing with the example of 10-year term life insurance and 1% interest rate, the SCR calculated using the MinT approach is \$23.736 bn with annual data, representing a decrease of 5.8% from \$25.198 bn when we use empirical errors with monthly data and a much larger decrease of 13.4% from \$27.399 bn when we use EVT.

5.4 Excess Deaths During COVID-19

In Sections 5.1 and 5.2, we use 1968–2009 as the training period and find our model performs well in terms of point forecast accuracy and interval forecast accuracy during the period of 2010–2019. In this section, we will look at the performance of our model against actual observations during the COVID-19 period, when the U.S. and the rest of the world were hitting by one of the most devastating global events since World War II, the COVID-19 pandemic. The COVID-19 has resulted in a substantial level of excess mortality for most countries. The United States has recorded more than 1 million "excess deaths" since the start of the pandemic, according to Robert Anderson, chief of the mortality statistics branch of the Centers for Disease Control and Prevention (CDC)'s National Center for Health Statistics.[10]

We collect age-specific death data in 2020 and 2021 from the CDC COVID-19 Death Data and Resources.[11] Now our training period ranges from 1968 to 2020 and we look at the performance of our model against actual observations in 2021.[12] Table 5.5 compares the actual number of deaths in each month of 2021 with the 99.5th percentile forecasts from the three methods under the empirical error distributions and the EVT error distribution, respectively. Bold numbers in Table 5.5 refer to the 99.5th percentile forecasts which are higher than the actual deaths. To better visualize the results, we also plot the 99.5th percentile forecasts and actual death counts in Figure 5.1.

Note: Bold numbers indicate a 99.5th percentile forecast higher than the actual death count in a given month.

[10] see media report at https://www.washingtonpost.com/health/2022/02/15/1-million-excess-deaths-in-pandemic/

[11] More details can be found at: https://www.cdc.gov/nchs/nvss/covid-19.htm.

[12] Another experiment we have conducted is to use data from 1968-2019 to forecast death counts in 2020 and compare our forecasts with the actual observations in 2020. We find that the MinT approach with EVT errors generates the highest numbers for the 99.5th percentile forecast in each month, but it covers the actual death counts in the first three month of 2020 only. This is mainly due to the lack of extreme death events in the training period of 1969-2019.

Table 5.5: 99.5th percentile forecasts vs observed number of deaths in 2021

Month	Observed	Empirical error			EVT error		
		Unrec	BU	MinT	Unrec	BU	MinT
Jan	104709	86390	73350	86408	92696	73443	92918
Feb	83632	74753	65150	74598	80304	65226	80101
Mar	84017	80747	71069	80278	**90669**	71182	**91006**
Apr	82350	76915	71105	76163	**83117**	71193	**82413**
May	82822	76249	70648	75462	**89874**	70809	**89777**
Jun	78846	73786	67967	72862	**88741**	68014	**88956**
Jul	83225	77137	71825	76024	**88083**	71977	**87558**
Aug	106355	75856	70621	74839	89036	70637	88887
Sep	111267	73788	67284	72660	85238	67303	84944
Oct	98183	76115	70041	74929	86919	70133	85706
Nov	87836	75922	70871	74640	**89585**	71017	**88934**
Dec	96806	83479	77891	81890	**101213**	77919	**100960**

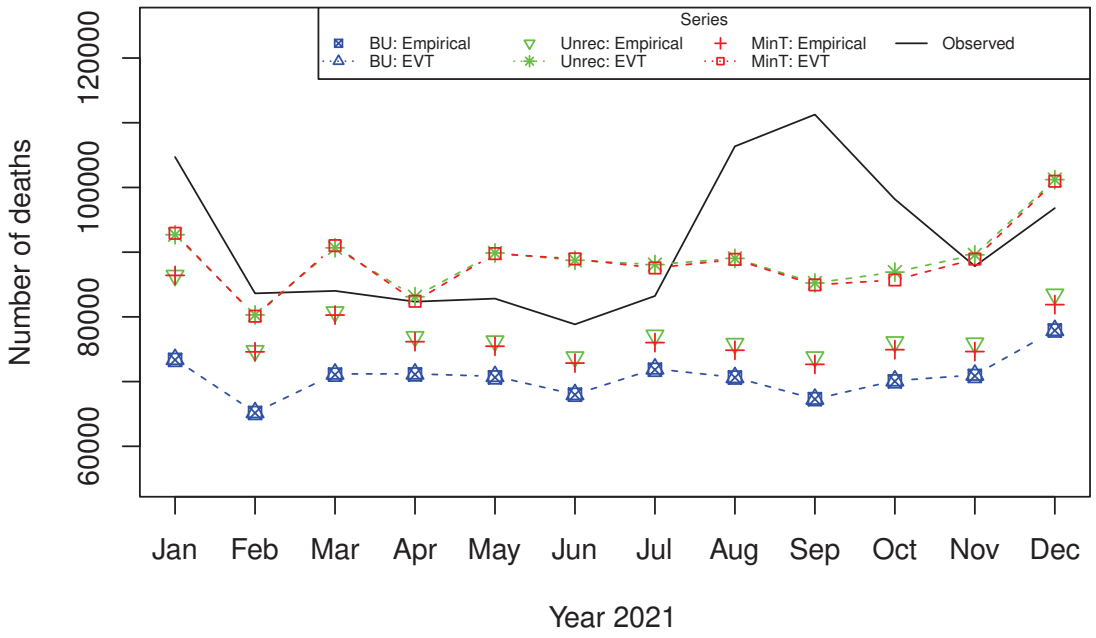


Figure 5.1: 99.5th percentile forecasts and observed number of deaths in 2021

.

First, using the empirical error distribution, none of the three methods can produce a 99.5th percentile forecast that can cover the actual death counts. This is because we can only extract errors from the empirical distribution, but cannot extrapolate more extreme, out of sample forecasts. Second, under the EVT error distribution, there are seven months (March-July and November-December) in 2021 when the observed deaths are within the 99.5th percentile forecasts from the MinT method and the unreconciled method, while in other months the 99.5th percentile forecasts fail to cover the actual deaths. This can be explained by the observed COVID-19 waves in 2021. In fact, a large spike in COVID-19 cases occurred over the winter months of 2020-2021 when people traveled and gathered for the winter holidays. On February 22, 2021, the total death toll from COVID-19 exceeds 500,000 in the U.S. The arrival of FDA-

authorized vaccines helped bring new infection levels and death toll back down in many areas through the spring of 2021. Another surge began in July 2021 as the contagious delta variant began to circulate and eventually become dominant, leading to the high death counts in the period of August-October. Third, in five out of seven months when both the unreconciled and EVT methods perform well, the MinT method generates a lower 99.5th percentile forecast than the unreconciled method. It indicates that the MinT method has a relatively better forecast precision because it takes advantage of all information in the hierarchy but the unreconciled method fails to do so. It also implies that compared to the MinT method, the unreconciled method is more likely to cause an overfunded issue for solvency capital calculation.

6 Conclusion

In this paper, we propose a new framework to coherently produce probabilistic mortality forecasts by exploiting techniques in seasonal ARIMA models, EVT, and hierarchical forecast reconciliation. Our model is built as a response to the pressing needs of life insurers to better understand the seasonal fluctuation of mortality, to account for extreme mortality risk, and to take into account mortality dependence in a hierarchical setting. Our results show that integrating MinT forecast reconciliation with EVT greatly improves the overall forecast accuracy for both point forecasts and probabilistic forecasts for time series in the hierarchy.

The improved probabilistic forecasts have important implications in insurance rate making, reserve setting, and capital adequacy compliance. To illustrate that, we create a hypothetical insurance portfolio and calculate the SCR under Solvency II. We find that MinT reconciliation considerably heightens the SCR compared to the unreconciled approach and the bottom-up approach, other things being equal. Assuming EVT errors also significantly contributes to the increase in the SCR. Therefore, insurers may face significant underfunded issues if they ignore extreme mortality risk or the dependence structure of forecast errors.

In another experiment, we project total deaths for each month in 2021 to see if our model can anticipate the surge in death tolls due to COVID-19. We find that our model can generate a 99.5th percentile forecast that can cover the actual number of deaths for most of months in the year. Thus our model can help life insurers better predict the excess deaths caused by COVID-19 given the previous, and particularly year 2020, mortality experience, and thus financially prepare for such extreme mortality events.

Our results call for the need of recording more accurate mortality data on a more frequent basis in order to assess the scale of short-term mortality elevations and to evaluate the effectiveness of different strategies used to address pandemics. Currently, the Human Mortality Database is establishing a new data resource, i.e., Short-term Mortality Fluctuations (STMF) data series, which collects the weekly death counts for 38 countries. Further research is warranted to better understand the patterns of mortality seasonality and mortality dependence as well as their impacts to life insurers.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, **19**(6), 716–723.

Armstrong, B. G., Chalabi, Z., Fenn, B., Hajat, S., Kovats, S., Milojevic, A., and Wilkinson, P. (2011). Association of mortality with high temperatures in a temperate climate: England and Wales. *Journal of Epidemiology & Community Health*, **65**(4), 340–345.

Athanasopoulos, G., Ahmed, R. A., and Hyndman, R. J. (2009). Hierarchical forecasts for australian domestic tourism. *International Journal of Forecasting*, **25**(1), 146–166.

Balkema, A. A. and de Haan, L. (1974). Residual life time at great age. *The Annals of Probability*, **2**(5), 792–804.

Beelders, O. and Colarossi, D. (2004). Modelling mortality risk with extreme value theory: The case of swiss re's mortality-indexed bond. *Global Association of Risk Professionals*, **4**(July/August), 26–30.

Boonen, T. J. (2017). Solvency ii solvency capital requirement for life insurance companies based on expected shortfall. *European actuarial journal*, **7**(2), 405–434.

Borges, C. E., Penya, Y. K., and Fernandez, I. (2013). Evaluating combined load forecasting in large power systems and smart grids. *IEEE Transactions on Industrial Informatics*, **9**(3), 1570–1577.

Cairns, A. J., Blake, D., and Dowd, K. (2006). A two-factor model for stochastic mortality with parameter uncertainty: theory and calibration. *Journal of Risk and Insurance*, **73**(4), 687–718.

Cairns, A. J., Blake, D., Dowd, K., Coughlan, G. D., and Khalaf-Allah, M. (2011). Bayesian stochastic mortality modelling for two populations. *ASTIN Bulletin: The Journal of the IAA*, **41**(1), 29–59.

Capistrán, C., Constandse, C., and Ramos-Francia, M. (2010). Multi-horizon inflation forecasts using disaggregated data. *Economic Modelling*, **27**(3), 666–677.

Chen, H. and Cummins, J. D. (2010). Longevity bond premiums: The extreme value approach and risk cubic pricing. *Insurance: Mathematics and Economics*, **46**(1), 150–161.

Chen, H., MacMinn, R., and Sun, T. (2015). Multi-population mortality models: A factor copula approach. *Insurance: Mathematics and Economics*, **63**, 135–146.

Chen, H., MacMinn, R. D., and Sun, T. (2017). Mortality dependence and longevity bond pricing: A dynamic factor copula mortality model with the gas structure. *Journal of Risk and Insurance*, **84**(S1), 393–415.

Cummins, J. D. and Phillips, R. D. (2009). Capital adequacy and insurance risk-based capital systems. *Journal of Insurance Regulation*, **28**(1).

Donaldson, G. and Keatinge, W. (2002). Excess winter mortality: influenza or cold stress? Observational study. *BMJ*, **324**(7329), 89–90.

Embrechts, P., Klüppelberg, C., and Mikosch, T. (2013). *Modelling extremal events: for insurance and finance*, volume 33. Springer Science & Business Media.

Feinstein, C. A. (2002). Seasonality of deaths in the us by age and cause. *Demographic Research*, **6**, 469–486.

Ferreira, A. and De Haan, L. (2015). On the block maxima method in extreme value theory: Pwm estimators. *The Annals of statistics*, **43**(1), 276–298.

Holzmüller, I. (2009). The united states rbc standards, solvency ii and the swiss solvency test: a comparative assessment. *The Geneva Papers on Risk and Insurance-Issues and Practice*, **34**(1), 56–77.

Hyndman, R. J. and Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts.

Hyndman, R. J. and Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for r. *Journal of Statistical Software*, **27**(3).

Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G., and Shang, H. L. (2011). Optimal combination forecasts for hierarchical time series. *Computational Statistics & Data Analysis*, **55**(9), 2579–2589.

Jeon, J., Panagiotelis, A., and Petropoulos, F. (2019). Probabilistic forecast reconciliation with applications to wind power and electric load. *European Journal of Operational Research*.

Klein, R. W. (2012). Insurance regulation and the challenge of solvency ii: Modernizing the system of us solvency regulation. *NAMIC (2012). http://www. namic. org/pdf/publicpolicy/insRegSolvII. pdf*.

Lee, R. D. and Carter, L. R. (1992). Modeling and forecasting US mortality. *Journal of the American Statistical Association*, **87**(419), 659–671.

Li, H. and Hyndman, R. (2021). Assessing mortality inequality in the us: What can be said about the future? *Insurance: Mathematics and Economics, Forthcoming*.

Li, H. and Tang, Q. (2019). Analyzing mortality bond indexes via hierarchical forecast reconciliation. *ASTIN Bulletin: The Journal of the IAA*, **49**(3), 823–846.

Li, H., O'Hare, C., and Zhang, X. (2015). A semiparametric panel approach to mortality modeling. *Insurance: Mathematics and Economics*, **61**, 264–270.

Li, H., Li, H., Lu, Y., and Panagiotelis, A. (2019). A forecast reconciliation approach to cause-of-death mortality modeling. *Insurance: Mathematics and Economics*, **86**, 122–133.

Li, J. S.-H. and Hardy, M. R. (2011). Measuring basis risk in longevity hedges. *North American Actuarial Journal*, **15**(2), 177–200.

Li, N. and Lee, R. (2005). Coherent mortality forecasts for a group of populations: An extension of the lee-carter method. *Demography*, **42**(3), 575–594.

Liu, C., Chen, C.-L., and Utter, D. (2005). Trend and pattern analysis of highway crash fatality by month and day. Technical report.

Panagiotelis, A., Athanasopoulos, G., Gamakumara, P., and Hyndman, R. J. (2021). Forecast reconciliation: A geometric view with new insights on bias correction. *International Journal of Forecasting*, **37**(1), 343–359.

Parks, R. M., Bennett, J. E., Foreman, K. J., Toumi, R., and Ezzati, M. (2018). National and regional seasonal dynamics of all-cause and cause-specific mortality in the usa from 1980 to 2016. *Elife*, **7**, e35500.

Pickands, J. (1975). Statistical inference using extreme order statistics. *The Annals of Statistics*, **3**(1), 119–131.

Rau, R. (2006). *Seasonality in human mortality: a demographic approach*. Springer Science & Business Media.

Rau, R., Bohk-Ewald, C., Muszyńska, M. M., and Vaupel, J. W. (2017). *Visualizing Mortality Dynamics in the Lexis Diagram*, volume 44. Springer.

Reiss, R.-D., Thomas, M., and Reiss, R. (1997). *Statistical analysis of extreme values*, volume 2. Springer.

Sanders, D. (2005). The modelling of extreme events. *British Actuarial Journal*, **11**(3), 519–557.

Shang, H. L. and Haberman, S. (2017). Grouped multivariate and functional time series forecasting: An application to annuity pricing. *Insurance: Mathematics and Economics*, **75**, 166–179.

Shang, H. L. and Hyndman, R. J. (2017). Grouped functional time series forecasting: An application to age-specific mortality rates. *Journal of Computational and Graphical Statistics*, **26**(2), 330–343.

Sharara, I., Hardy, M., and Saunder, D. (2010). A comparative analysis of u.s., canadian and solvency ii capital adequacy requirements in life insurance. *Society of Actuaries*.

Silverman, S. and Simpson, P. (2011). *Case Study: Modelling Longevity Risk for Solvency II*. Milliman, United States.

Syntetos, A. A., Babai, Z., Boylan, J. E., Kolassa, S., and Nikolopoulos, K. (2016). Supply chain forecasting: Theory, practice, their gap and the future. *European Journal of Operational Research*, **252**(1), 1–26.

Vaughan, T. M. (2009). The implications of solvency ii for us insurance regulation. *Networks Financial Institute Policy Brief*, (2009-PB), 03.

Wickramasuriya, S. L., Athanasopoulos, G., and Hyndman, R. J. (2019). Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *Journal of the American Statistical Association*, **114**(526), 804–819.

Winkler, R. L. (1972). A decision-theoretic approach to interval estimation. *Journal of the American Statistical Association*, **67**(337), 187–191.

Zellner, A. and Tobias, J. (2000). A note on aggregation, disaggregation and forecasting performance. *Journal of Forecasting*, **19**(5), 457–465.

Zhou, R., Wang, Y., Kaufhold, K., Li, J. S.-H., and Tan, K. S. (2014). Modeling period effects in multi-population mortality models: Applications to solvency ii. *North American Actuarial Journal*, **18**(1), 150–167.

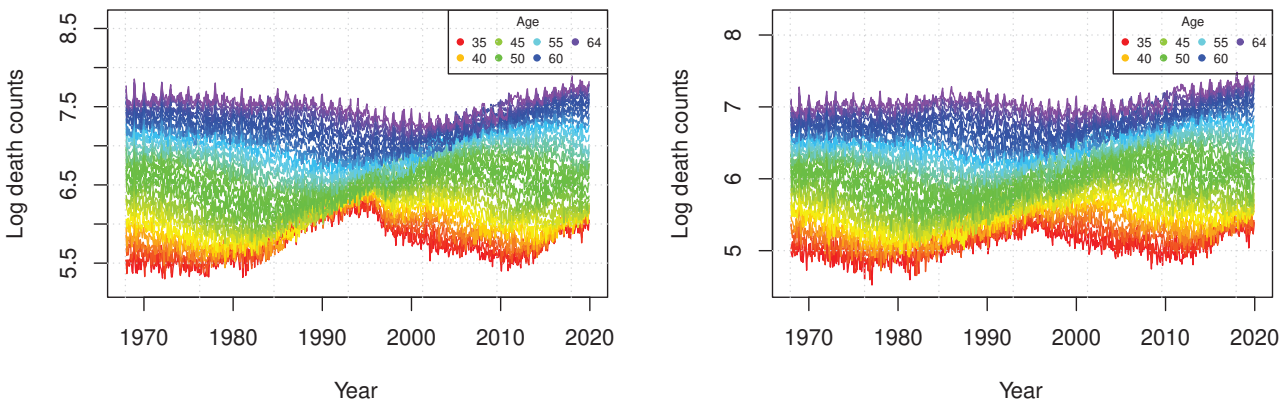
Appendix

Table A.1: Selected ARIMA models for male and female death counts

Series	Non-seasonal			Seasonal			Series	Non-seasonal			Seasonal		
	p	d	q	P	D	Q		p	d	q	P	D	Q
Total	0	1	2	2	1	1							
MTotal	3	1	2	1	1	2	FTotal	0	1	2	2	1	1
M35	0	1	1	2	1	0	F35	5	1	2	2	1	0
M36	0	1	1	2	1	0	F36	4	1	1	2	1	0
M37	0	1	1	2	1	0	F37	5	1	0	2	1	0
M38	0	1	1	2	1	0	F38	0	1	1	2	1	0
M39	0	1	1	2	1	0	F39	5	1	1	2	1	0
M40	0	1	1	2	1	0	F40	1	1	1	2	1	0
M41	0	1	1	2	1	0	F41	4	1	1	2	1	0
M42	0	1	1	2	1	0	F42	3	1	1	2	1	0
M43	2	1	1	2	1	0	F43	0	1	1	2	1	0
M44	0	1	1	2	1	0	F44	4	1	2	2	1	0
M45	2	1	1	2	1	0	F45	4	1	1	2	1	0
M46	0	1	1	2	1	0	F46	3	1	1	2	1	0
M47	0	1	1	2	1	0	F47	2	1	1	2	1	0
M48	0	1	1	2	1	0	F48	3	1	1	2	1	0
M49	0	1	1	2	1	2	F49	5	1	1	2	1	0
M50	0	1	1	2	1	0	F50	4	1	1	2	1	0
M51	1	1	0	2	1	0	F51	5	1	1	2	1	0
M52	3	1	0	2	1	0	F52	3	1	1	2	1	0
M53	0	1	1	2	1	0	F53	4	1	1	2	1	0
M54	0	1	1	2	1	0	F54	5	1	1	2	1	0
M55	1	1	1	2	1	0	F55	2	1	1	2	1	0
M56	0	1	1	2	1	0	F56	3	1	2	2	1	0
M57	0	1	1	2	1	0	F57	2	1	1	2	1	0
M58	0	1	3	2	1	0	F58	2	1	1	2	1	0
M59	0	1	3	2	1	0	F59	1	1	2	2	1	0
M60	0	1	1	2	1	2	F60	5	1	0	2	1	0
M61	0	1	1	1	1	0	F61	4	1	1	2	1	0
M62	0	1	3	2	1	0	F62	5	1	0	2	1	0
M63	0	1	2	2	1	0	F63	2	1	1	2	1	0
M64	0	1	2	2	1	1	F64	3	1	1	2	1	0

Table A.2: Shapiro normality test on age-specific deaths

Male			Female		
Age	p-value	Fat-tailed	Age	p-value	Fat-tailed
35	0.05	N	35	0.02	Y
36	0.00	Y	36	0.07	Y
37	0.11	N	37	0.02	Y
38	0.01	Y	38	0.00	Y
39	0.24	N	39	0.10	N
40	0.13	N	40	0.17	N
41	0.01	Y	41	0.03	Y
42	0.10	N	42	0.13	N
43	0.01	Y	43	0.05	N
44	0.00	Y	44	0.09	N
45	0.00	Y	45	0.08	N
46	0.00	Y	46	0.01	Y
47	0.00	Y	47	0.07	N
48	0.00	Y	48	0.13	N
49	0.00	Y	49	0.09	N
50	0.00	Y	50	0.02	Y
51	0.02	Y	51	0.04	Y
52	0.00	Y	52	0.07	N
53	0.00	Y	53	0.08	N
54	0.00	Y	54	0.09	N
55	0.00	Y	55	0.02	Y
56	0.00	Y	56	0.02	Y
57	0.00	Y	57	0.01	Y
58	0.00	Y	58	0.00	Y
59	0.00	Y	59	0.01	Y
60	0.00	Y	60	0.03	Y
61	0.00	Y	61	0.00	Y
62	0.00	Y	62	0.01	Y
63	0.00	Y	63	0.00	Y
64	0.00	Y	64	0.01	Y



(a) Male (b) Female

Figure A.1: Rainbow plots for U.S. monthly death counts: 1968–2019