Journal of Statistics and Computer Science
Vol. 2, Issue 1, 2023, pp. 17-27
https://doi.org/10.47509/JSCS.2023.v02i01.02

Application of Machine Learning to Mortality Modelings during the Pandemic in the U.S.A

Rui Gong* [a] and Jinwei Liu[b]

[a] *Department of Informatics and Mathematics, Mercer University, Macon, GA, USA*
[b] *Department of Computer and Information Sciences, Florida A&M University , Tallahassee, FL, USA*

Abstract

Various stochastic models were developed to predict mortality rates over the past two decade. Because of the COVID-19 Pandemic starting in 2019, the prediction accuracy by each model can be influenced. In this paper the Poisson splitting method is implemented to calibrate parameters in the Lee-Carter(LC) Model and the Poisson Lee-Carter (PLC) Model respectively. The methodology is applied to U.S. mortality data. Mortality rates forecasts are formed for the period 2019-2020 based on data from 2000-2018. These forecasts are compared to the actual observed values to investigate the implementation of the methodology and the quality of such mortality models during the Pandemic.

Keywords: Lee-Carter Model, Poisson Lee-Carter Model, Mortality Rates Forecasts, SBS Poisson Regression Tree.

*Corresponding authors: gong_r@mercer.edu

Article History
Received : 12 December 2022; Revised : 1 May 2023, 14 May 2023; Accepted : 16 May 2023; Published : 30 June 2023

To cite this paper
Rui Gong & Jinwei Liu (2023). Application of Machine Learning to Mortality Modelings during the Pandemic in the U.S.A.. *Journal of Statistics and Computer Science.* 2(1), 17-27.

1 Introduction

Government agencies use mortality forecasts to decide the allocation of funds for government services, plan and develop health policy. Private industries, such as insurance companies and life insurers, also need information from these forecasts to plan their future programs and manage longevity risk.

One popular stochastic mortality model is introduced by Lee and Carter (1992). The Lee-Carter (LC) model is widely used in the world because of its robustness among diverse mortality models proposed in the literature. Singular value decomposition (SVD) is applied to the log-force of mortality to estimate the parameters in the original model. By using the first principal component of the log-mortality matrix, the estimation is presented. The time component is predicted using a random walk model with drift in their original paper. In practice, one of the most commonly adopted approaches to predict mortality rates is the autoregressive integrated moving average (ARIMA) model.

Brillinger (1986) proposed that the number of deaths is a counting number so it reasonably follows Poisson distribution. This leads Brouhns et al. (2002a,b) to extend the LC model to the Poisson LC (PLC) model. To predict the mortality rate in the PLC model, two different models are used. In one model, time is interpreted as a factor (Brouhns et al., 2002a) and in the other one, Renshaw and Haberman (2003) modelled time as a known covariate. Since the effect of calendar time is unknown ex ante to represent in some functional form, the former model is believed to be preferable (Czado et al., 2005). Thus ARIMA models are used most commonly in the PLC model as in the LC model.

Brouhns et al. (2002a,b) assumed a Poisson distribution for deaths and calculated the parameters by log-likelihood maximization. They also implemented a bootstrap procedure for a Poisson log-bilinear formulation of the Lee-Carter model. Non-linear regression and generalized linear model (GLM) is involved in recent approaches in the LC model and the PLC model. However, these computations of prediction error for the mortality forecasts does not account for the estimation error of the parameters involved in both models. In this paper, we focus on one popular Machine Learning algorithm introduced by Deprez et al. (2017) to model and forecast U.S. mortality. Poisson splitting method is used as a complement to standard LC model and PLC model, rather than a substitute. We initialize the mortality rates in Poisson splitting method with the estimation rates obtained from the LC model and PLC model respectively, then apply this Machine Learning approach on the data from 2000 to 2020. We use the data from 2000-2018 as the training set and the data from 2019-2020 as the test set to compare the actual observed values with the forecasts.

The outline of this paper is as follows. Section 2 presents a brief description of the LC model, the PLC model and the Poisson splitting method. In Section 3, we apply the original LC and PLC methodology, then implement the Poisson splitting method to U.S. males and females separately. Section 4 provides concluding remarks.

2 Methodology

Lee and Carter (1992) proposed the LC model for the log-mortality rates $log(\mu_{x,t})$:

$$log(\mu_{x,t}) = \alpha_x + \beta_x \kappa_t + \epsilon_{x,t}, \quad \epsilon_{x,t} \overset{iid}{\sim} N(0, \sigma_\epsilon^2), \tag{2.1}$$

We use $D_{x,t}$ as the number of deaths recorded at age group x during year group t, and $E_{x,t}$ as the corresponding number of persons exposed to risk, where $\mu_{x,t} = D_{x,t}/E_{x,t}$ is the observed mortality rate of age group x during year group t, and $x = 1, 2, \ldots, M-1, M$ and $t = 1, 2, \ldots, T-1, T$ represent a set of M different age groups and T year groups, respectively.

But parameters in (2.1) are invariant to the following linear transformation:

$$\beta_x \leftarrow\!\!\shortmid \frac{\beta_x}{d}, \quad \kappa_t \leftarrow\!\!\shortmid (\kappa_t - c)d, \quad \alpha_x \leftarrow\!\!\shortmid \alpha_x + \beta_x c. \tag{2.2}$$

for any $d \in \mathbb{R} \backslash \{0\}$ and $c \in \mathbb{R}$. In actuarial literature the following two constraints are imposed to overcome the identification issue:

$$\sum_x \beta_x = 1, \quad \sum_t \kappa_t = 0. \tag{2.3}$$

The main advantage of applying (2.3) on (2.1) is that we can use the average log-mortality rate over time for age group x and the first principal component of the log-mortality matrix to estimate α_x, β_x and κ_t. Thus SVD is applied to the log-force of mortality rate to estimate the parameters in the original model.

Various adjustments of $\hat{\kappa_t}$ is proposed to reduce the difference between estimated and observed log-mortality rates, it is shown that the random walk model with drift works for most data set and it expresses as:

$$\kappa_t = \kappa_{t-1} + \theta + \omega_t, \quad \omega_t \overset{iid}{\sim} N(0, \sigma_\omega^2). \tag{2.4}$$

where θ is the drift parameter which models a linear trend and ω_t is an error term. So for the lth step-ahead forecast, the mean forecast value of κ_{t+l} is as follows:

$$\hat{\kappa}_{t+l} = \hat{\kappa}_t + l\hat{\theta},$$

And the mean forecast of the log-mortality rate for year $t+l$ is

$$log(\hat{\mu}_{x,t+l}) = \hat{\alpha}_x + \hat{\beta}_x \hat{\kappa}_{t+l}.$$

In the literature, two assumptions made commonly are that $D_{x,t}$ are independent, and each $D_{x,t}$ has a Poisson distribution with a parameter proportional to $E_{x,t}$. The force of mortality stays constant over each period $(t, t+1]$ is also assumed. Brouhns et al. (2002a,b) kept the Lee-Carter log-bilinear form with the same constraints (2.3) and replace (2.1) with:

$$D_{x,t} \sim Poisson(E_{x,t}\mu_{x,t}) \;\; with \quad log(\mu_{x,t}) = \alpha_x + \beta_x \kappa_t. \tag{2.5}$$

Based on the PLC model (2.5), instead of applying the SVD method, they maximized the log-likelihood function to estimate parameters $\alpha_x, \beta_x, \text{and}\, \kappa_t$. Because of the bilinear term $\beta_x \kappa_t$ in this model, Newton's iterative updating scheme and weighted least squares are used to get the MLE (Brouhns et al., 2002b); Renshaw and Haberman, 2003).

Each individual person is identified by its gender, its age, and the calendar year. We assign each individual person p to a feature space $= (g, x, t) \in \chi = G \times X \times T^*$ and feature components are $G = \{female, male\}$, $X = \{0, \dots, M\}$ and $T^* = \{1, \dots, T\}$. Here $x \in X$ represents the age in years of the person, $M \in \mathbb{N}$ denotes the maximal possible age the person can reach. The component $T \subset \mathbb{N}_0$ describes the calendar years considered. This feature space could be extended by further feature components.

In order to check the goodness of the Poisson splitting method, we initialize model assumptions with the rates μ_p obtained from the LC model and the PLC model, respectively. We consider for $p \in \chi$ and rewrite (2.5) as follows:

$$D_p \sim Poisson(\psi(p) d_p), \quad with\, \psi(p) \equiv 1 \; and \; d_p = E_p \mu_p. \tag{2.6}$$

Note that d_p represents the expected number of deaths. If the estimated mortality rate μ_p underestimates the crude rate D_p/E_p the factor $\psi(p)$ should be increased and vice versa. As such, we need to calibrate the factor $\psi(p)$ based on the chosen features $p \in \chi$. In addition, we also include birth cohorts, so we extend the feature space χ to the feature space $\bar{\chi} = \{(g, x, t, c = t - x) \mid g \in G, x \in X, t \in T^*\}$, where $c = t - x$ provides the cohort.

We follow the steps of Poisson regression developed by Therneau et al. (2022) to apply the Poisson splitting method with the splitting criterion $D_{parent} - (D_{leftson} + D_{rightson})$. The explicit choice of each split is based on an optimal improvement of a given loss function.

3 Data Analysis

First, we use the LC model and the PLC model to estimate mortality rates based on the U.S.A mortality data from 2000 to 2018. Then we apply Poisson splitting method to the estimators of mortality rates by the LC model and the PLC model respectively to obtain the new estimators. Second, we compare those estimators with the real U.S.A mortality rates through the root mean squared logarithmic error ($RMSLE$) and the root mean squared error ($RMSE$).

3.1 Data

The data employed in this study consists of USA mortality data obtained from the Human Mortality Database. The exposures $(E_p)_{p \in \chi}$ and the number of death $(D_p)_{p \in \chi}$ are the data we consider. The following results are based on the feature space $\bar{\chi} = G \times X \times T^* \times C$ with feature components $G = \{female, male\}$, $X = \{0, 1, \dots, 110\}$, $T = \{2000, 2001, \dots, 2018\}$ and $C = \{1890, 1891, \dots, 2018\}$.

Maximal age 110 here represents ages of at least 110, and the set T consists of 19 years of observations. This results in 4218 data points in those 19 years of observations.

3.2 Results

We run the LC model and the PCL model to estimate the U.S.A mortality rates for the female and the male respectively. Then these estimators are updated by Poisson splitting method and are improved by classifying on the new feature space $\tilde{\chi}$. Note that between the space χ and the new space $\tilde{\chi}$ we have a one-to-one correspondence.

Figure 1 and Figure 2 show the Poisson splitting tree which provides the calibration to estimate $\psi(p), p \in \tilde{\chi}$. The birth cohorts added in the new feature space $\tilde{\chi}$ requires diagonal splits. It is clear that the estimators by the PLC model are classified into further subgroups by Poisson splitting method.

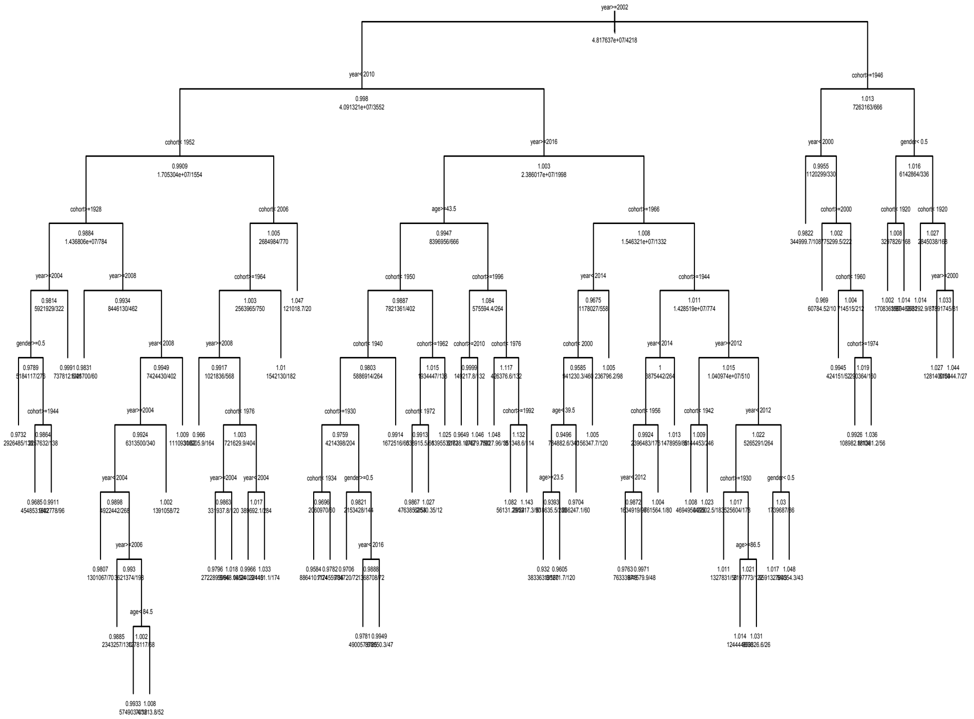

Figure 1: Poisson Splitting Trees of Estimators by the LC model for U.S.A 2000-2018 Mortality Rates

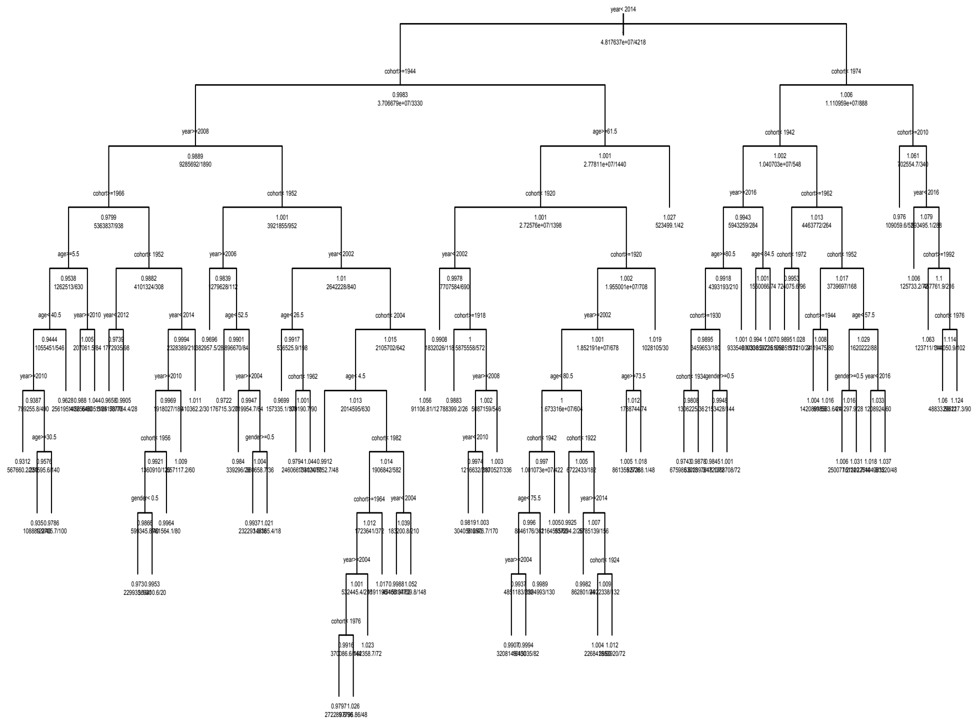

Figure 2: Poisson Splitting Trees of Estimators by the PLC model for U.S.A 2000-2018 Mortality Rates

We apply the Poisson splitting method to estimate $\psi(p)$. The estimators by the LC model and the PLC model are used to define the new updated mortality rates respectively as follows:

$$\hat{\mu}_{new}(p) = \hat{\psi}(p)\hat{\mu}_{LC}(\omega) \; or \; \hat{\mu}_{new}(p) = \hat{\psi}(p)\hat{\mu}_{PLC}(\omega), \quad with \; \omega \in \chi \; and \; p \in \tilde{\chi}. \qquad (3.1)$$

The tree algorithm improves the initialized mortality rates $\hat{\mu}_{LC}(\omega)$ and $\hat{\mu}_{PLC}(\omega)$ in (3.1). We consider the relative changes $\triangle\hat{\mu}$ to analyze the improvements, the relative changes being measured as follows:

$$\triangle\hat{\mu} = \hat{\mu}_{new}(p) - 1, \quad with\, p \in \tilde{\chi}.$$

So the relative change is the difference between the estimator of $\mu(p)$ and 1. In Figure 3 and Figure 4, we provide the relative changes by the LC model and the PLC model for the U.S.A female mortality rate and the U.S.A male mortality rate from 2000 to 2018, respectively. Dark blue implies that the relative changes are negative and the estimators by the LC model or PLC model overestimate the mortality rate in this period; light blue implies that the relative changes are positive and the estimators by either model underestimates the mortality rate. Figure 3 and Figure 4 show that the overestimation happens when age increases for the female and the male, the underestimation happens when age is small for both.

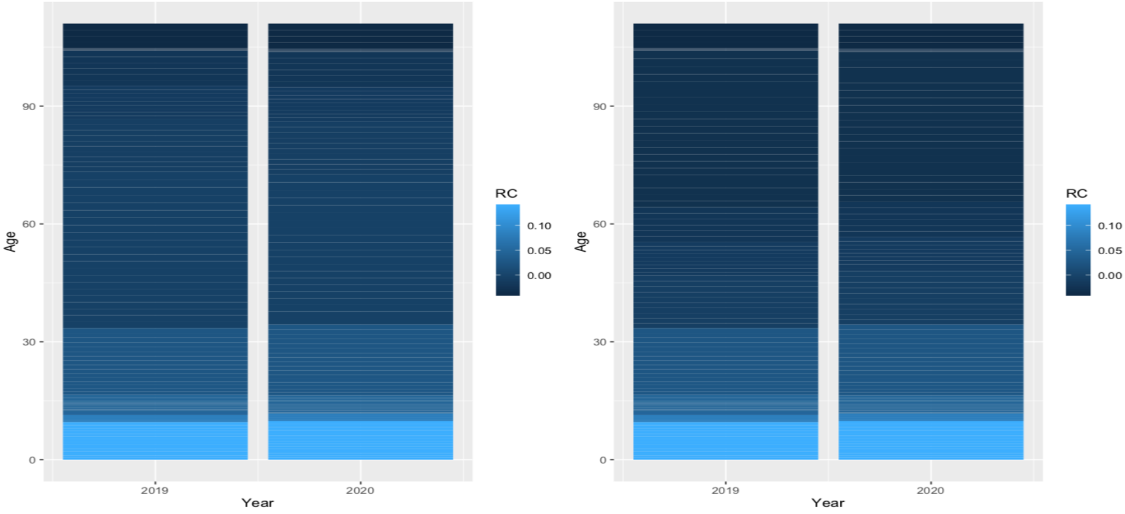

Figure 3: Relative Changes $\triangle\hat{\mu}$ based on the LC model

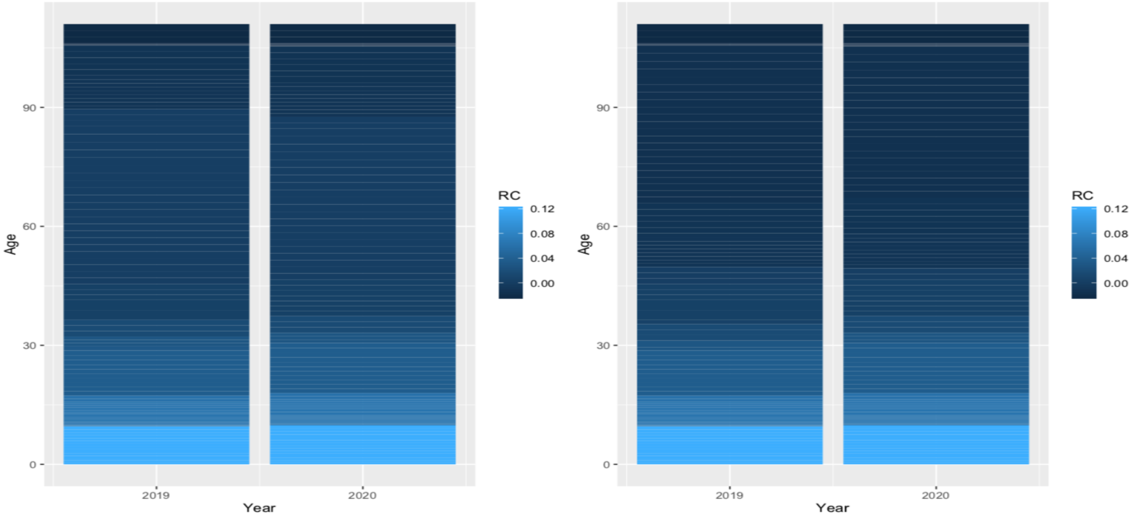

Figure 4: Relative Changes $\triangle\hat{\mu}$ based on the PLC model

Figure 5 and Figure 6 are logarithms of mortality rates for different ages and calendar years 2019 and 2020. The red solid lines illustrate the crude mortality rates of the U.S.A mortality data, the blue dashed lines illustrates the estimators by the LC model or the

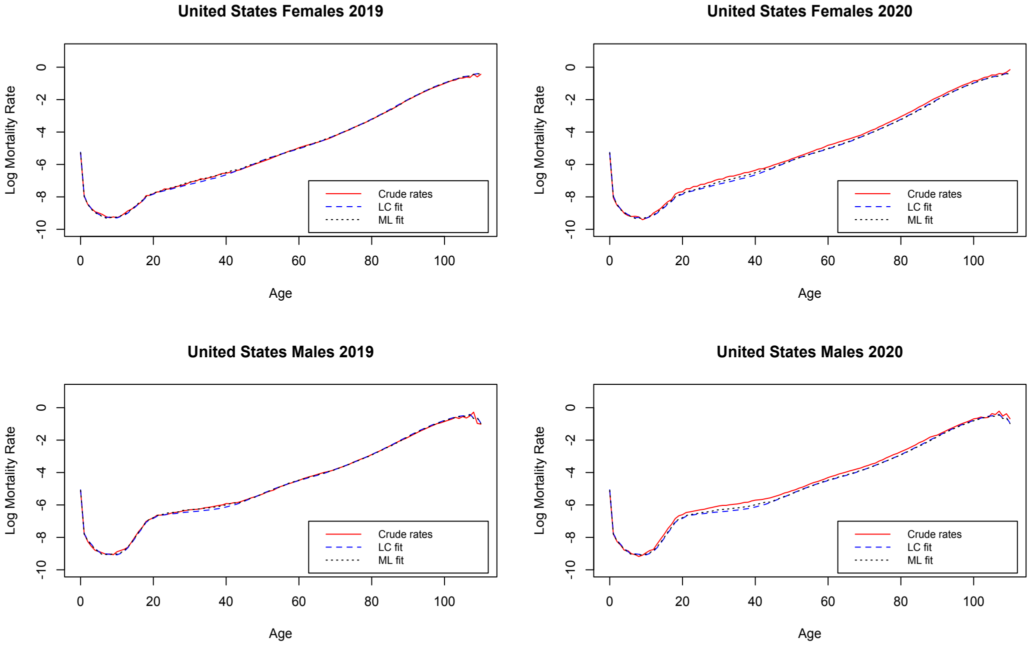

Figure 5: Logarithms of Mortality Rates for Males and Females based on the LC model

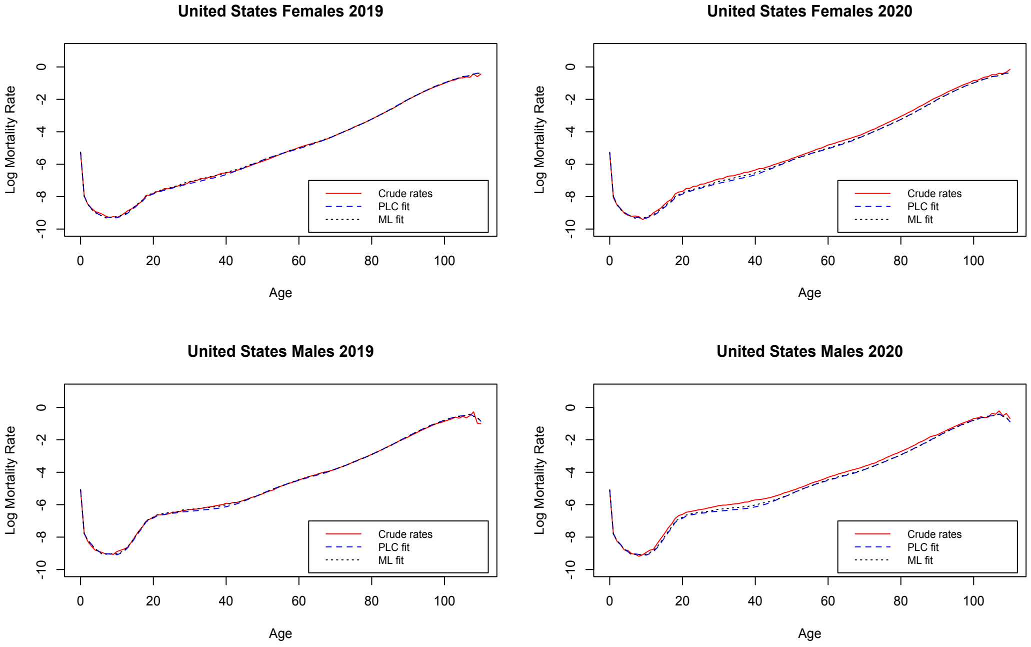

Figure 6: Logarithms of Mortality Rates for Males and Females based on the PLC model

PLC model, and the black dots illustrate the tree updated estimators. We apply the LC model and the PLC model to each gender $g \in G$ separately. We aim at back-testing these fitted mortality rates by using Machine Learning (ML) technique, splitting method to be precise.

By implementing (3.1), we find $\hat{\mu}_{new}(p)$ to predict the logarithms of mortality rates. We compare the predicted mortality rates with the crude mortality rates in 2019 and 2020 for the female and the male in Figure 5 and Figure 6. It can be observed that the PLC model, the LC model and the Poisson splitting method predict the logarithm of mortality rates well because the forecasts are close to the real mortality rates in these two year. However, the estimation by the Poisson splitting method seems closer to the crude mortality rates compared with the estimation by both models. In addition, the estimations by both the models and the Poisson splitting method in 2019 are better than the estimations in 2020 for the female and the male.

To compare the performance of forecasting between either model and the Poisson splitting method, we use two measures: $RMSLE$ and $RMSE$. $RMSLE$ and $RMSE$ are measured as follows:

$$\begin{aligned} RMSLE &= \sqrt{\sum(log(\hat{\mu}_{x,t}) - log(\mu_{x,t}))^2/N}, \\ RMSE &= \sqrt{\sum(\hat{\mu}_{x,t} - \mu_{x,t})^2/N}, \end{aligned}$$

where N is the number of predicted points, and in our data analysis $N = 444$. $RMSLE$ uses the logarithms of mortality rates to provide a relatively large amount of weight to errors at young ages, while $RMSE$ uses mortality rates directly to provide a relatively large amount of weight to errors at older ages.

Table 1: $RMSLE$ and $RMSE$ by the LC model and the Poisson splitting method(ML)

	RMSLE		RMSE	
Model	Female	Male	Female	Male
LC	0.1313603	0.1647777	0.0214511	0.0292743
ML	0.1054338	0.1336998	0.0218488	0.0307270

Table 2: $RMSLE$ and $RMSE$ by the PLC model and the Poisson splitting method(ML)

	RMSLE		RMSE	
Model	Female	Male	Female	Male
PLC	0.1318182	0.1651776	0.0211829	0.0271793
ML	0.1052516	0.1333952	0.0211801	0.0277720

Table 1 shows the test results for the estimations by the original LC model and the estimations improved by machine learning when $\hat{\mu}_{new}$ is forecasted using the LC framework. By using the Poisson splitting method, the reduction of $RMSLE$ is around 19% for males and 23% for females; while when considering $RMSE$, Poisson splitting method makes the error become larger. Table 2 presents the similar patter for $RMSLE$ and $RMSE$ for the PLC model and the ML approach. So we can see that the Poisson splitting method produces a significant improvement in forecasting with respect to the standard LC model and PLC model at young ages, but this method does not improve the estimations by both models at older ages during the Pandemic in the U.S.A.

4 Conclusions

We introduce the ML estimator and compare the forecasting qualities provided by the LC model, the PLC model and this ML approach during the Pandemic, where the Poisson splitting method is used as a support and not as substitute for those stochastic models. We aim to update the predicted mortality rates provided by the LC model and the PLC model and create a bridge between the ML approach and theory to help find a rational explanation of the results during the Pandemic. All the analysis is carried out on the calibration period: 2000-2018. The forecasting results for 2019-2020 are discussed.

We illustrate how the Poisson splitting method is applied to update forecasting of the LC model and the PLC model. Our work is the extended work of Deprez et al. (2017), which applied a regression tree boosting machine to improve the fitting of the LC model and the Renshaw-Haberman model. We test the improvement in the forecasting quality of the Poisson splitting method based on the LC model and the PLC model. Our results, obtained from a data analysis structured on the U.S.A population during the Pandemic, demonstrate that this ML approach produces significant improvements at young ages, but it does not improve both models at older ages. Those results match the fact that the old people have consistently accounted for a larger share of COVID-19 deaths than the young people during the Pandemic.

Conflict of Interest

The authors confirm that this article content has no conflict of interest.

Acknowledgement

We thank Ms. Christina Green, student ambassador from Mercer university, for collecting data for this article and Mr. Trevor Taylor, graduate student from Florida A&M university, for cleaning data and helping generate graphs for this article. We thank Reviewers for taking the necessary time and effort to review the manuscript.

References

Deprez, P., Shevchenko, P. V. and Wüthrich, M. V. (2017). Machine learning techniques for mortality modeling. *European Actuarial*, 7, 337–352.

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

Brillinger, D. R. (1986). The natural variability of vital rates and associated statistics, *Biometrics.* 42(4), 693–743.

Brouhns, N., Denuit, M., and Vermunt, J. K. (2002). Measuring the Longevity Risk in Mortality Projections. *Psychosomatic Medicine - PSYCHOSOM MED*, 2002, 105–130.

Brouhns, N., Denuit, M., and Vermunt, J. K. (2002). A Poisson log-bilinear regression approach to the construction of projected lifetables. *Mathematics and Economics*, 31(3), 373–393.

Renshaw, A. E., and Haberman, S. (2003). Lee-Carter mortality forecasting with age specific enhancement. *Insurance:Mathematics and Economics*, 33(2), 255–272.

Renshaw, A. E., and Haberman, S. (2005). A Cohort-Based Extension to the Lee-Carter Model for Mortality Reduction Factors. *Insurance:Mathematics and Economics*, 38(3), 556–570.

Czado, C., Delwarde, A., and Denuit, M. (2005). Bayesian Poisson log-bilinear mortality projections. *Insurance:Mathematics and Economics*, 36(3), 260–284.

Therneau, T. M., Atkinson, E. J., and Foundation, Mayo. (2022). An Introduction to Recursive Partitioning Using the RPART Routines. Available at: https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf.