An Age-Period-Cohort model for gender gap in youth mortality

Giacomo Lanfiuti Baldi*[1] and Andrea Nigri[2]

[1]*Department of Statistics, Sapienza University of Rome, Rome, Italy.*
[2]*Department of Economics, Management and Territory, University of Foggia, Foggia, Italy*

October 30, 2023

Abstract

In this paper, we introduce a novel framework in longevity study, operating on the statistical approach of the Age–Period–Cohort framework by leveraging the skew-normal distribution and Bayesian estimation. We propose a specific application to gender gap analysis and forecasting. By employing mortality data from the Human Mortality Database in the USA, our study contributes a two-fold advancement. First, we present a novel perspective on gender gap analysis and forecasting, improving the current literature. Second, we contribute an improvement to the statistical framework for Age–Period–Cohort analysis. The proposed model offers invaluable insights applicable to healthcare planning and public interventions, providing a comprehensive snapshot of the gender gap across the population, and indispensable information for devising healthcare strategies.

Keywords: Age-Period-Cohort Model; Skew-Normal; Forecasting; Death rates

*corresponding author, email: giacomo.lanfiutibaldi@uniroma1.it

1 Introduction

Longevity gains are one of the most fascinating achievements of the modern era. How these achievements evolved, are evolving and how they are expected to change in the near future is a matter of continuing debate. Life expectancy increases have nourished optimistic views about the maximum human life expectancy ([34, 45, 3, 51]) and disproving pessimistic speculations on an impending ceiling on life expectancy for humans ([17, 36, 37, 38, 35, 13]). The steady rise in life expectancy at birth has posed challenges to all of governments, public health systems, and pension plans, giving pivotal roles to longevity analysis. These needs necessitate sophisticated statistical frameworks.

Challenges may include how longevity dynamics are developing over time, considering the heterogeneity and thus longevity gaps among different populations. Specifically, how females outlive males is the most discussed gap in population studies. The mortality risk due to gender results from a potentially complex combination of single-risk factors and, despite growing interest in modelling overall mortality (such as death counts or mortality rates), nothing has been done to advance modelling of the gender gap. Working directly on the ratio provides relevant benefits. Indeed, one of the main concerns in longevity modelling is the simultaneous dynamics between males and females, which should be included in order to guarantee reliable estimations through coherent forecasting ([6, 8, 39, 40, 44]). As mentioned by Li and Lee (2005) [25], forecasting the mortalities of two populations separately tends to produce a greater difference, even when using similar methods. Directly exploiting the gender ratio as a unique variable will undoubtedly provide more insightful information reducing the complexity/parsimony–interpretability trade-off, in statistical modelling. Since the sex ratio appears to be less sensitive to mortality level ([5, 28, 14]), it offers a better picture of the disparities by age than the absolute sex differences of the death rates.

In this paper, we use a sex-ratio approach to study gender differences: we analyse the logarithm of the ratio of age-specific mortality rates between males ($m^{\mathrm{M}}_{x,t}$) and females ($m^{\mathrm{F}}_{x,t}$) over time.

$$\mathrm{SR}_{x,t} = \log\left(\frac{m^{\mathrm{M}}_{x,t}}{m^{\mathrm{F}}_{x,t}}\right) \tag{1.1}$$

This measure in Eq. 1.1 is useful for several reasons: it allows an implicitly

gender-consistent model to be defined, it is less sensitive to the general level of mortality than the absolute difference in deaths [7], and, finally, it has a well-defined and known shape [28].

Generally, the sex ratio for age-related mortality is characterised by a peak and a hump. The peak, which is the highest and most concentrated, coincides with youth and is generally attributed to young males engaging in riskier behaviours. The hump occurs during adult ages and is due primarily to excess male mortality from cancer [7]. According to Meslé (2004) [28] we set the threshold age, between the peak and the hump, at 45 years. We will focus only on the peak, i.e. on mortality differences between the sexes at younger ages.

The disparities in infant mortality between males and females are minimal across all periods, with differences beginning to increase during adolescence (see Figure 1). In many cases, male mortality peaks more than three times higher than that of females and this trend remains consistent over the years without any significant shift in the age at which the peak occurs.

The Lexis surface is useful to visualize changes in a phenomenon with respect to ages (vertical lines), periods (horizontal lines) and cohorts (diagonal lines). We report here the Lexis surface of the sex ratio for the data that we will later use to test and apply the proposed model (US, 1960–2020).

The most interesting period is between the second half of the 1970s and the second half of the 1990s. During these years, the gap widened, only to narrow sharply after 1995. The increase in differences mainly affected the upper age groups and cohorts born between the 40's and 50's. The causes of these changes can be found in the literature [47, 20, 42].

Through the proposed model, we aim at identifying each point in the Lexis surface as the sum of an intercept, an age effect, a period effect and a cohort effect.
To do so, we propose a Bayesian approach for modelling and forecasting the changes in the sex ratio in three components, i.e. Age, Period, and Cohort. In particular, we propose the Age–Period–Cohort modelling combined with a probabilistic assumption on the sex ratio. Specifically, we consider the sex-ratio variable to follow a skew-normal distribution.

The skew-normal distribution has recently been used in the study of longevity and in mortality modelling (see [27, 1, 50]) because of its convenient shape to fit the age-at-death distribution at certain ages. We instead propose to use the skew-normal to leverage shape and scale parameters to obtain accurate estimates of the sex ratio. We apply and test the proposed

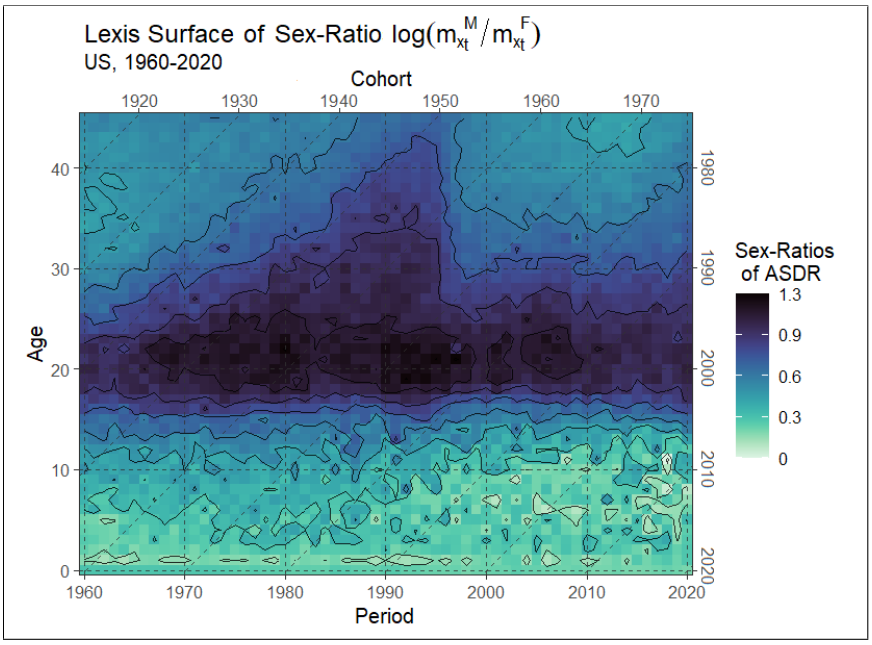

Figure 1: Lexis Surface of the Sex Ratio of the Age-Specific Mortality Rates in US between 1960 and 2020. Data source: HMD

modelling to study the gender gap in youth mortality in the United States.

In the US, more than 30% of deaths at young ages are due to external causes (unintentional injuries) [19] and the issues of road accidents and violent and risky behaviour among young people are often at the centre of public debate.

The remaining part of this work is organized as follows. In Section 2, we describe the Age–Period–Cohort models and offer a review of methods for solving the identification problem. Section 3 is devoted to the extensive explanation of the model for estimating and forecasting the sex ratio in mortality in a Bayesian framework, and Section 4 contains the results from that model. Finally, in Section 5 we discuss the results and conclude.

2 Age-Period-Cohort Model

The Age-Period-Cohort model (APC) is used to study a time-specific phenomenon distinguishing the effects that the variations in age, years, and cohorts have on the phenomenon analysed [49].

An APC model describes the target variable Y (the phenomenon studied) as a function of the I ages, the J periods and the K cohorts, where $K = I + J - 1$.
In the OLS regression framework, we can write the APC model in matrix notation treating ages, periods and cohorts as categorical variables [15][32], as follows:

$$Y = X\beta + \epsilon \tag{2.1}$$

Where Y is a $(I \times J) \times 1$ outcome vector; X is the design matrix consisting of *"dummy variable"* column vectors of dimension $(I \times J) \times (I + J + K - 2)$; β is the $(I + J + K - 2) \times 1$ vector of coefficients (extended form in Equation 2.2) and ϵ is a $(I \times J) \times 1$ vector of residuals.

$$\beta = (\alpha, \beta_1^A, \ldots, \beta_I^A, \beta_1^P, \ldots, \beta_J^P, \beta_1^C, \ldots, \beta_K^C) \tag{2.2}$$

This means that the value y_{ij}, relative to age i-th and year j-th is given by a multiple linear regression:

$$y_{ij} = \alpha + \beta_i^A + \beta_j^P + \beta_k^C + \epsilon_{ij} \tag{2.3}$$

Where: α is the intercept, β_i^A is the i-th age effect ($i = 1, \ldots, I$), β_j^P is the j-th period effect ($j = 1, \ldots, J$), β_k^C is the k-th cohort effect ($k = 1, \ldots, I + J - 1$) and ϵ_{ij} represents the residuals.

By treating the variables age, period and cohort as categorical to allow for nonlinear effects and including the intercept, we have to solve the problem of over-parametrization of the model. In accordance with the literature ([10, 48, 15]), we deal with this problem by using the first category for each of the three variables as a reference, fixing the first set of parameters equal to zero:

$$\beta_1^A = \beta_1^P = \beta_1^C = 0 \tag{2.4}$$

However, the model suffers from a lack of identifiability, due to the linear dependency among the three components: Cohort = Period-Age. In Eq.2.1,

the design matrix is singular with one less than full column rank, therefore it is not invertible. This implies that the Ordinary Least Squares (OLS) estimator of the matrix regression model in Eq. 2.1 is not uniquely defined, there are infinite solutions to the equation:

$$\hat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y \tag{2.5}$$

All of the infinite solutions fit the APC model equally well and correspond to the best fit [11], moreover, according to Holford (2014) [22] the predicted values of the outcome from a linear extrapolation are identifiable and this allows forecasting of the phenomenon.

As is known in the literature [10, 48, 15], it is the linear trend of the three effects that is not identified, while the non-linear part is identified. This means that: taking any of the infinite solutions of the Eq. 2.5, the slope of the three effects will be different, but the deviations from the trend will be the same no matter which solution is chosen [32]. Moreover, Clayton and Schifflers (1987) [12] show that in the model, the drift is also identified. The drift is the sum of the period and cohort trend, i.e. the overall linear trend of the phenomenon. It's not possible to decompose the drift between period and cohort.

To better understand the problem of non-identifiability, in literature ([32, 33, 15, 41]), it is often defined that the infinite number of best-fitting solutions to Eq. 2.5 all lie on the *"line of solutions"*. This means that given any solution vector $\hat{\beta}_c$ obtained under the constraint in Eq. 2.4, any other best-fitting solution can be obtained as:

$$\hat{\beta}'_c = \hat{\beta}_c + sv, \tag{2.6}$$

where v is the null vector for X and s is a scalar.

There are several solutions to the problem of identifiability, the most used is imposing additional constraints on the effects in order to provide a unique solution.
Following the review made by Fosse and Winship (2019) [15], we can divide the solving methods into two macro groups: *Without Measured Causes* and *With Measured Causes*, where the Measured Causes are measures of the causes associated with the three variables of the Age-Period-Cohort model.

We ignore methods using these auxiliary measures since we propose modelling based only on one observed variable.
In the remaining methods, they are divided into two groups *Explicit Constraints* and *Mechanical Constraints*. The former are those that have been used more over time and since the first APC modeling proposals. There are various types of explicit constraints: the most basic are the so-called *Drop One-Variable*, which simply consists of removing one (or more) of the three dimensions (Age, Period and Cohort). Excluding a variable is equal to imposing strict constraints that its linear and nonlinear effects are zero, which is often an overly stringent assumption. Moreover, this clearly results in the loss of potentially interesting information on one of the three dimensions.
Another example of explicit constraints are the *Equality Constraints*. The idea is to add a constraint to reduce the rank of the design matrix X, make it non-singular and thus invertible so as to obtain a single OLS $\hat{\beta}$ estimator in Eq. 2.5. This corresponds to selecting only one of the solutions in the 'solution line' in Eq. 2.6 [32]. The constraint consists to fix two of the effects of one of the three components equal to each other (example: $\beta_i^A = \beta_{i+1}^A$). The main problem of equating two groups is essentially that it is equivalent to assigning specific values to each of the unknown linear effects. This may turn out to be too strong an assumption and it must be based on theoretical hypotheses and knowledge of the phenomenon. There are different variations of Equality Constraints based on the constraint(s) one decides to use. For example, it is possible to set the first and last values of one of the effects of the three dimensions equal to each other, which corresponds with setting the linear effect of that component equal to zero.

The Mechanical Constraints group contains various solutions proposed in recent years, the two most widely used techniques being *Intrinsic Estimators (IE)* and *Hierarchical age-period-cohort models (HAPC)*.
The Intrinsic Estimators are part of the *Moore-Penrose (MP)* estimators. This technique works in two stages: least-squares and norm-minimum. The first stage is simply the calculation of the least-squares solution set in Eq. 2.5. The second stage selects the particular set of solutions such that the square root of the sum of the squared estimates is as small as possible. The main problem of the MP estimators is that the solution defined depends on the design matrix used in Eq. 2.1. For example, by changing the grouping into classes of the three variables (i.e. 1-year or 5-year age classes) or mixing their order in the model (ACP). Fosse and Winship [16, 15] report some qualities

of this estimator: clearly it produces a unique solution, it has minimum sampling variance among all possible estimators based on the same design matrix and it is unbiased. This technique is also much criticised for the lack of robustness of the estimation with respect to certain aspects: for example, with respect to the number of categories used for the three components (age, period and cohort) and with respect to the choice of reference category.
Lastly, Hierarchical age-period-cohort models use a mixed-effects approach [32]. This method produces a unique identified set of effects without the need for theoretical assumptions and the imposition of explicit constraints. On the contrary, several authors stated that this method introduces implicitly some *nonobviuos* constraints [16].

See Table 1 for a schematic summary of constraints in APC.

	Explicit Constraints	
	Drop-one variable	**Equality constraints**
Pros	Easiest technique to use Most widely used technique	Knowledge-based and arbitrary Transparent research decision Easy to use
Cons	Assumption stronger than necessary Loss of potentially interesting information	Very strong assumption
	Mechanical Constraints	
	Intrinsic Estimator (MP)	**Hierarchical APC**
Pros	Unbiased estimator Minimum sampling variance	No need of theoretical assumption
Cons	Lack of robustness	Strong implicit constraints

Table 1: Selection of techniques to solve the problem of non-identifiability of Age-Period-Cohort models based on review by Fosse and Winship (2019).

3 Model

In this section, we describe the methodological background of our proposal, useful for modeling the gender gap in longevity. Age–Period–Cohort (APC) models can be expressed as a least-squares regression, where normality distribution is assumed along with homoscedasticity of the errors [48]. We propose an APC model assuming that the errors are distributed skew-normal, allowing control and analysis of the error skewness. Moreover, Age–Period data can exhibit unstable variability, possibly represented by increases in variance with age [26]. Since the gender gap in mortality is a complex phenomenon, distributed with respect to both age and period, assuming a constant variability appears very stringent.

Therefore, we relax the assumption of normality and homoscedasticity (used in the OLS estimation), and leverage the skew-normal distribution while also modelling the variance and the skewness.

As defined by Azzalini in 1985 [2], Z has a skew-normal distribution with parameter $\lambda \in \mathbb{R}$ if:

$$f(z|\lambda) = 2\phi(z)\Phi(\lambda z),\ z \in \mathbb{R} \tag{3.1}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the $N(0,1)$ probability density function and cumulative distribution function, respectively.

If $Z \sim (\lambda)$, then the random variable $Y = \mu + \sigma^2 Z$ still has a skew-normal distribution with a *location* parameter $\mu \in \mathbb{R}$, a *scale* parameter $\sigma^2 \in \mathbb{R}^+$ and a *shape* parameter λ. A transformation of the λ parameter is the *skewness* parameter: $\tau = \lambda/\sqrt{1+\lambda^2}$, $-1 < \tau < 1$.

The probability density function of $Y \sim SN(\mu, \sigma^2, \lambda)$ is given by:

$$f(y;\theta) = f(y;\mu,\sigma^2,\lambda) = \frac{2}{\sigma^2}\phi\left(\frac{y-\mu}{\sigma^2}\right)\Phi\left(\lambda\frac{y-\mu}{\sigma^2}\right). \tag{3.2}$$

It is easy to see in Eq. 3.1 and in Eq. 3.2 that, for $\lambda = 0$, the skew-normal reduces to the normal distribution.

The mean and variance of Y are:

$$\mathbb{E}(Y) = \mu + \sqrt{\frac{2}{\pi}}\sigma^2\tau \tag{3.3}$$

$$Var(Y) = \sigma^2 \left(1 - \frac{2\lambda^2}{\pi}\right). \tag{3.4}$$

We assume that the sex ratio of the gender-age-specific mortality rates ($\mathrm{SR}_{x,t} = \log\left(\frac{m^{\mathrm{M}}_{x,t}}{m^{\mathrm{F}}_{x,t}}\right)$), follows a skew-normal distribution for each age x and year t:

$$\mathrm{SR}_{xt} \sim SN(\mu_{xt}, \sigma_x^2, \lambda_x). \tag{3.5}$$

By leveraging the properties of the skew-normal distribution, we can model the sex ratio, assuming that $\mu_{xt}, \sigma_x^2, \lambda_x$ follow the regression structure given by:

$$\mu_{xt} = \alpha + \beta_x^A + \beta_t^P + \beta_{t-x}^C \tag{3.6}$$

where α is the intercept, β_x^A is the xth age effect ($x = 1, \ldots, X$), β_t^P is the tth period effect ($t = 1, \ldots, T$), and β_{t-x}^C is the $(t-x)$th cohort effect.

$$f(\sigma_x^2) = \gamma X \tag{3.7}$$

$$h(\lambda_x) = \omega X \tag{3.8}$$

where X is the design matrix for the explanatory variable (i.e. Age) for the scale and shape parameters.

Note that, in Eq. 3.6, we choose to model the location parameter based on the APC framework, exploiting the categorical coding for the Age, Period and Cohort variables. As a result, after the estimation procedure, we obtain many parameters as the cardinality of the variables.

On the contrary, to ease the interpretation and improve the parsimony of parameter estimation, for equations 3.8 and 3.7, we choose a canonical regression framework in which the estimated coefficient is a scalar and $f(\cdot)$ and $h(\cdot)$ are the logarithmic link function[1] and identity function, respectively.

Therefore, linking to Eq 3.3, the point estimate for the sex ratio is given by:

$$\begin{aligned}\widehat{\mathrm{SR}_{xt}} = \mathbb{E}(\mathrm{SR}_{xt}) = \mu_{xt} + \sqrt{\frac{2}{\pi}}\sigma_x^2\tau_x = \\ = \alpha + \beta_x^A + \beta_t^P + \beta_{t-x}^C + \sqrt{\frac{2}{\pi}}e^{\gamma x}\frac{\omega x}{\sqrt{1+(\omega x)^2}}.\end{aligned} \tag{3.9}$$

[1] $\ln(\sigma_x^2) = \gamma X \Rightarrow \sigma_x^2 = \exp(\gamma X)$

The total set of parameters to be estimated is:

$$\theta = (\alpha, \beta_1^A, \ldots, \beta_X^A, \beta_1^P, \ldots, \beta_T^P, \beta_1^C, \ldots, \beta_{T-X}^C, \gamma, \omega) \tag{3.10}$$

To tackle the problem of non-identifiability of the model, it is necessary to include external constraints (see Section 2). In the specific case analysed, the change in the mortality gender gap at younger ages ($x < 45$) in the United States between 1960 and 2020, we propose to use an *Explicit Constraint* (see Section 2). By analysing this phenomenon, we can calculate the overall linear trend (drift) of the sex ratio over these 60 years. Perhaps because of the nature of the study variable, the drift is small (about +0.12%), so we use an *equality constraint*, setting the first and last period effects equal to each other and, more specifically, equal to zero (Eq. 3.11):

$$\beta_{1960}^P = \beta_{2020}^P = 0. \tag{3.11}$$

With this explicit constraint, the period effect is de-trended (i.e. it has zero slope) and the unidentifiable linear effect is completely absorbed by the cohort effect [22].

At the stage of interpreting the results, it should be kept in mind that the estimated effect values are related to the imposed constraint, while the non-linear trend can be interpreted independently of the constraint.

3.1 Implementation

Samples from the posterior distributions of the parameters were drawn by using Hamiltonian Monte Carlo sampling and specifically using the `stan` software package [43]. Stan and its interface in the R programming language (R Core Team, 2017) allows the construction of a Hamiltonian Monte Carlo sampling 'no U-turns sampler' [21] from a simple user specification of the Bayesian model to be estimated. Hamiltonian Monte Carlo simulates movement through the parameter space by analogy to a physical system where the potential energy is equal to negative log-posterior [29], and it is a special case of the more general Metropolis–Hastings algorithm for Markov chain Monte Carlo sampling. Four parallel chains were constructed and used to assess convergence to the better posterior distribution, each with 2000 samples, and the first half of each chain was used as a warm-up period. Gelman and Rubin [18] split $\hat{r}$ diagnostics are below the suggested 1.05 threshold

and examination of trace-plots indicate sampler convergence to the target distribution [18] [43].

Priors for the model hyper-parameters are: $\alpha \sim t_3(0.6, 2.5)$ for the intercept and flat priors for all the APC parameters $(\beta^A, \beta^P, \beta^C)$ and the regression coefficients on shape and scale parameters.

We adopt weakly informative priors, but one of the potentialities of the proposed framework is to be able to use a priori knowledge of the phenomenon to provide more accurate estimates.

3.2 Forecasting

In this section, we leverage the framework in section 3 (excluding the constraint in Eq. 3.11) to focus on the temporal component, represented by the period and cohort sets of parameters. The proposed forecasting method falls into the context of the three-factor forecasting models that have been emerging in recent years [9].

The idea behind the forecasting method is to fix the age effect estimated by the APC method on the baseline time interval and to forecast the period and cohort effects.

Suppose that we forecast the sex ratio h periods in the future for the same set of age groups:

$$\mu_{x,T+h} = \alpha + \beta_x^A + \beta_{T+h}^P + \beta_{T+h-x}^C. \tag{3.12}$$

If $T - x + h < X$ (highest age), we use the observed cohort effects from the data, otherwise we project new cohort effects.

We project the time series of period and cohort effects from the data, so the forecasting functions need to be defined such that:

$$\beta_{T+h}^P = f_P\left(\beta_1^P, \ldots, \beta_T^P\right), \text{ and } \beta_{T+h-x}^C = f_C\left(\beta_1^C, \ldots, \beta_{T+X}^C\right). \tag{3.13}$$

We use the best ARIMA model for the projection of the two time series. To select of the best ARIMA, we use the `auto.arima` function in the *R* software [23]. This evaluates, via the Aikake Information Criterion (AIC), which ARIMA model best serves for a specific time series.

The ARIMA model powerfully captures various patterns and trends in time series data, making it a popular choice for forecasting future values in fields like finance, economics, and demographics [4]. A brief description of

the ARIMA models and the best ARIMA estimated with their parameters and time series plots are shown in the Appendix.

Lastly, since, we assume heteroscedasticity (Eq. 3.9), we forecast the sex ratio at age x in time $T+h$ ($\widehat{\mathrm{SR}_{x,T+h}}$), leveraging the setting in Eq. 3.9.

4 Results

In this section, we show the results obtained by applying the model and provide a preliminary interpretation of the estimated APC effects.
In order to provide model estimates, we use high-quality 1×1 life tables from the Human Mortality Database (HMD, 2021) [46], categorized by sex, for the United States population due to its peculiar longevity behaviour – characterized by stagnations and periods of slow improvements in life expectancy, with interesting consequences for the cohort profile ([30, 31]).

The results for 1960–2020 are shown in Figure 2. It is important to remember that this is only one of the possible estimates of the three effects on the line of solutions; specifically, it is the solution in which the first and last period effects are equal and equal to zero. The non-linear trend does not depend on the chosen solution, so it can be interpreted from a demographic point of view. It must be kept in mind that, by using age, period and cohort as categorical variables in which the first category is used as a reference, each effect must be interpreted as an addition to (or subtraction from) the base level of the sex ratio, that is given by the intercept.

The intercept is 0.64, which means that the base level of the sex ratio, when the three effects are zero, is equal to $\alpha = 0.64 \Rightarrow \mathrm{SR} = \exp(0.64) = 1.89$. This means that, on average, males die almost twice as often as females between the ages of 0 and 45.

Exploiting the benefit of the proposed framework allows us to provide and discuss also the scale and shape estimation: the point estimate obtained with the APC model is corrected with respect to the spread and the skewness according to the relation in Eq. 3.3.

The regression models imposed on the scale and shape parameters indicate that the two parameters behave differently with respect to the age variable:

$$\sigma_x^2 = e^{-0.1 \cdot x}$$

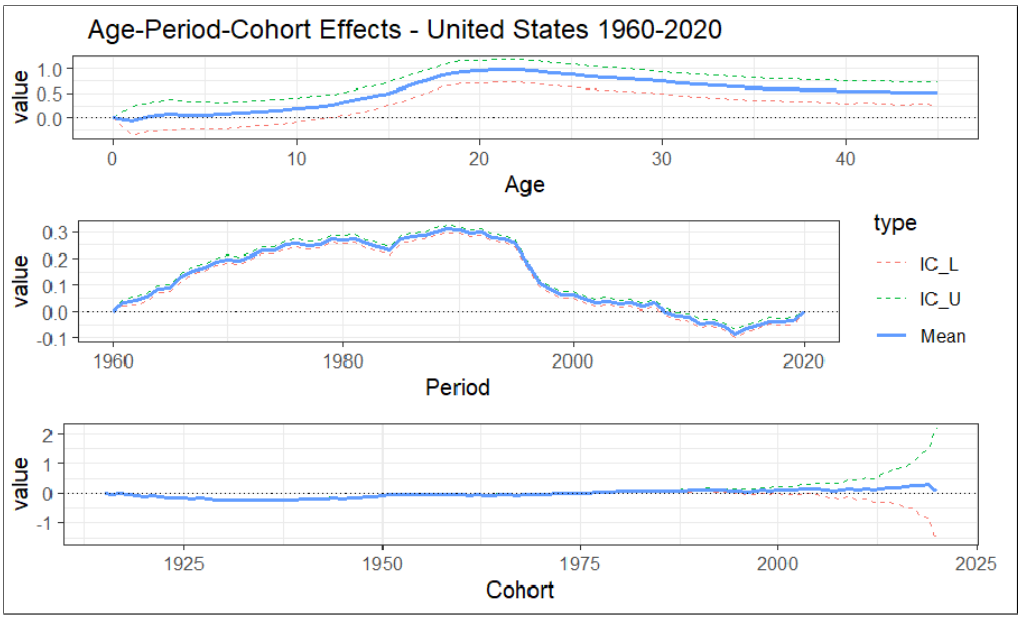

Figure 2: From top to bottom: Age effects, reference category is age 0 ($x = 0$), Period de-trended effects, reference categories are year 1960 and 2020 ($p = 1960$), Cohort effects and drift, reference category is year 1915 ($k = t - x = 1915$). Dotted lines represent Bayesian credibility intervals for the estimates. Data source: HMD.

$$\lambda_x = 0.06 \cdot x \Rightarrow \tau_x = \frac{0.06 \cdot x}{\sqrt{1 + (0.06 \cdot x)^2}}$$

The scale parameter decreases exponentially with respect to the ages, whereas the shape increases slightly. Given the relationship between shape and skewness, we can observe the skewness trend with respect to age: it increases between 0 and 18 years and then decreases.

It is possible to interpret the effects in Figure 2 as the contribution of each component on the gender gap. The age effect represents the average profile of the sex ratio with respect to age over the time interval considered. During childhood, the differences are small; around age 3, there is a small drop in the sex ratio compared to at age 0. At later ages, the gender gap increases, peaking at age 23, after which it decreases, but remaining always higher than the initial level. Thus, compared to the base level of mortality differences between the sexes, increasing age increases the gap, in agreement with the

Baseline period	**Intercept α**	**Scale γ**	**Shape ω**
1960–1990	0.806	− 0.089	0.313
1970–2000	0.703	−0.099	0.099
1980–2010	0.69	−0.095	4.829

Table 2: Estimation of the intercept and regression coefficients for the parameters of the model on the baselines.

literature . In the solution obtained under the constraint in Eq. 3.11, the period between 1960 and 1995 shows an increase in gender differences, with no major effect afterward. In the literature, specifically concerning the United States, many explanations can be found for the large period of variability in the gender gap observed in the second half of the last century.

Analysing cohort effects via the chosen solutions shows that the first cohorts contributed negatively to gender differences, thus indicating a convergence in mortality between men and women. Cohorts born between the years 45' and 55', however, show an increase in the gender gap, suggesting that these cohorts experienced social conditions under which men were more likely to die than women.

Out of sample To assess the robustness of our method to forecast purposes, we performed an out-of-sample test over three-time windows (1960–1990, 1970–2000, and 1980–2010) with a 30-year baseline period to estimate Age, Period, and Cohort effects, as well as the intercept and regression coefficients for the shape and scale parameters. Thus, the subsequent years of each time window (1991–2000, 2001–2010, and 2011–2020 respectively) have been used as the out-of-sample set.

The base levels of the sex ratio in the baseline periods are similar to that estimated over the entire period 1960–2020, with a reduction in the intercept over time. The coefficient δ is negative in each of the three baseline periods, indicating that the variance decreases exponentially as the ages increase. Coefficient ω provides different trends in the three periods: in periods 1960–1990 and 1980–2010, $\omega > 1$, whereas, in the period 1970–2000, the shape parameter increases slightly with ages.

To evaluate the goodness of the prediction, we compute the Root Mean Square Error (RMSE) and the Mean Absolute Error (MAE), respectively

defined as:

$$\text{RMSE} = \sqrt{\frac{1}{XT}\sum_{x=1}^{X}\sum_{t=1}^{T}(\text{SR}_{xt} - \widehat{\text{SR}}_{xt})^2},$$

$$\text{MAE} = \frac{1}{XT}\sum_{x=1}^{X}\sum_{t=1}^{T}|\,\text{SR}_{xt} - \widehat{\text{SR}}_{xt}|.$$

Period	**MAE**	**RMSE**
1991–2000	0.195	0.141
2001–2010	0.094	0.068
2011–2020	0.096	0.072

Table 3: RMSE and MAE for the three forecast windows.

As shown by both RMSE and MAE (see Table 3), the model is more accurate in more recent forecasting windows. This can be attributed to the years of greater variability in the sex ratio observed between the 1960s and 2000s and the rapid decline observed after 1995, which is hard to predict (Figure 4 in the Appendix shows the forecast for period and cohort effects). In the proposed framework, forecasts are mostly affected by period forecasts, while cohort forecasts affect only younger cohorts.

We also analyse the model performances with respect to different ages and years, using the relative differences ($\Delta_{x,t}$) defined as follows:

$$\Delta_{x,t} = \frac{\left(\widehat{\text{SR}}_{x,t}\right) - (\text{SR}_{x,t})}{(\text{SR}_{x,t})}.$$

The models appear to overestimate gender differences at younger ages (Figure 3), due to the uncertainty added by the prediction of cohort effects.

The last heat map (1980–2010) is on a different scale and is strongly affected by the dynamic exhibited in the year 2018. Indeed, looking at the observed data, 2018 is a single case in which male mortality is lower than female mortality.

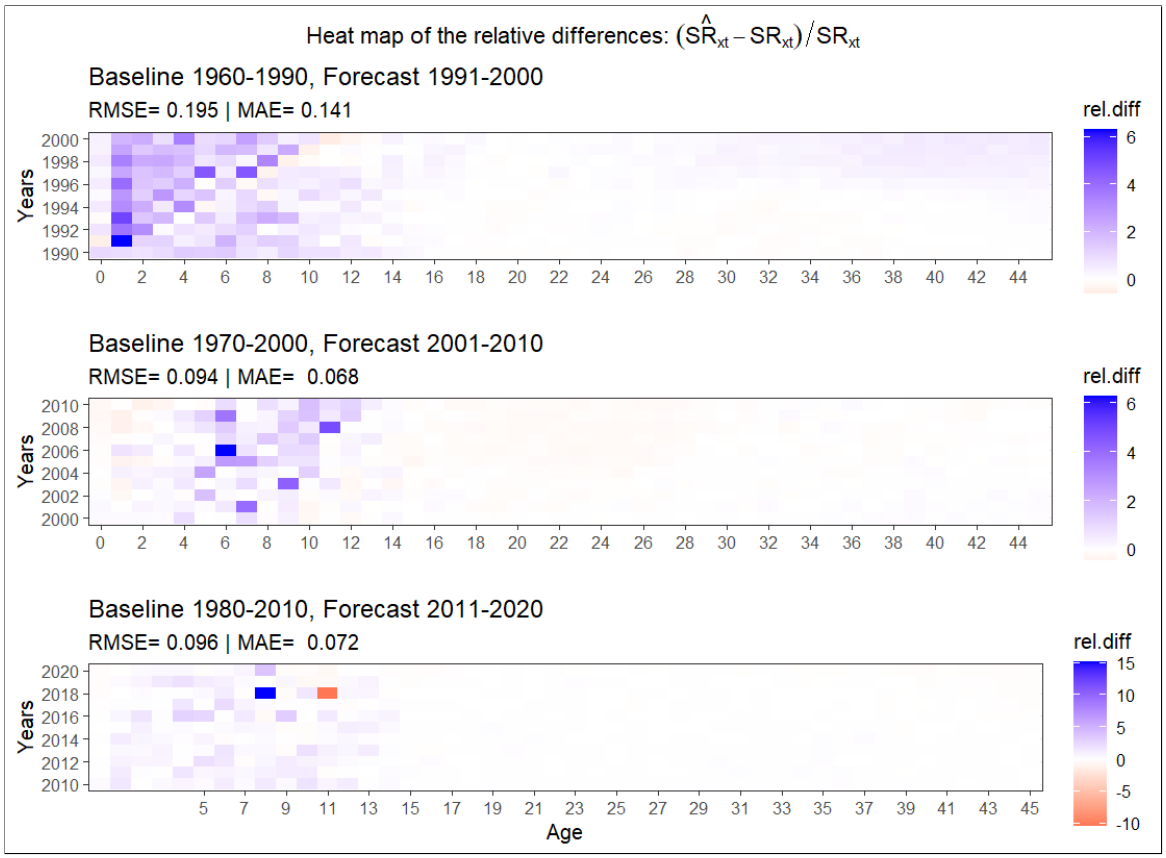

Figure 3: Relative differences $\Delta_{x,t}$ between forecasted and the observed sex ratio by age and year for each time interval of forecasting. Red hues indicate that the model underestimates the gender gap, while blue hues indicate overestimation. RMSE and MAE for each out-of-sample test are reported. Different scales are used in each heat map.

5 Discussion and Conclusion

This paper extends and improves the statistical approach for Age–Period–Cohort frameworks by introducing, for the first time, the modelling of sex ratio in longevity using the skew-normal distribution in a Bayesian framework. The adoption and the modelling of the scale and shape parameters provide explicit accounting for the potential asymmetries and variability in sex ratio mortality. Using data on USA mortality from the HMD database, we describe the statistical details of our method introducing the Bayesian

approach and the forecasting procedure, and we test the model accuracy in the out-of-sample exercise over three contiguous but non-overlapping time windows. Our investigation provides twofold insight into the related literature: i) We introduce an innovative statistical framework in the study of APC aimed to model the sex ratio in a given population, ii) the study leverages the double lens of the model accuracy evaluation and provides an assessment of the demographic significance of our approach for the gender gap in longevity. The suggested model provides valuable information that can be used to improve healthcare and public policy. Indeed, healthcare planning can benefit from our approach, which provides a useful snapshot of the magnitude of gender differences. Introducing the cohort effect, we can also speculate that the gender difference which, in the recent past, was considered a natural public health target because of its size, might not be deemed a significant problem today. In this regard, we have stressed the novelties and implications of our proposal. Our model provides significant insights to further discussion in population studies. In particular, we consider that the cohort effect is a huge aspect to consider in gender gap analysis and forecasting. Indeed, working on the sex ratio approach to implementing the cohort effect is not straightforward because of the unique relationship between age, time, and cohort. In the existing literature, we draw particular attention to the work of Bergeron et al. 2018 ([6]) who preferred to work in a Lee–Carter framework. None of the models considered include cohort effects to account for specific survival in some cohorts. Nevertheless, it is widely believed to be very likely that cohort components could improve the fit and forecast in some populations (see e.g., Kjærgaard et al. [24]). In light of these aspects, we do not deem it appropriate to directly compare our model with others such as the Lee–Carter, which is clearly out of the scope of this proposal.

References

[1] Emanuele Aliverti, Stefano Mazzuco, and Bruno Scarpa. "Dynamic modelling of mortality via mixtures of skewed distribution functions". In: *Journal of the Royal Statistical Society Series A: Statistics in Society* 185.3 (2022), pp. 1030–1048.

[2] Adelchi Azzalini. "A class of distributions which includes the normal ones". In: *Scandinavian journal of statistics* (1985), pp. 171–178.

[3] Elisabetta Barbi et al. "The plateau of human mortality: Demography of longevity pioneers". In: *Science* 360.6396 (2018), pp. 1459–1461.

[4] Francesco Battaglia. *Metodi di previsione statistica*. Springer Science & Business Media, 2007.

[5] Hiram Beltrán-Sánchez, Caleb E Finch, and Eileen M Crimmins. "Twentieth century surge of excess adult male mortality". In: *Proceedings of the National Academy of Sciences* 112.29 (2015), pp. 8993–8998.

[6] Marie-Pier Bergeron-Boucher et al. "Coherent forecasts of mortality with compositional data analysis". In: *Demographic Research* 37 (2017), pp. 527–566.

[7] Marie-Pier Bergeron-Boucher et al. "Modeling and forecasting sex differences in mortality: a sex-ratio approach". In: *Genus* 74 (2018), pp. 1–28.

[8] Christina Bohk-Ewald, Marcus Ebeling, and Roland Rau. "Lifespan disparity as an additional indicator for evaluating mortality forecasts". In: *Demography* 54.4 (2017), pp. 1559–1577.

[9] Heather Booth and Leonie Tickle. "Mortality modelling and forecasting: A review of methods". In: *Annals of actuarial science* 3.1-2 (2008), pp. 3–43.

[10] Bendix Carstensen. "Age–period–cohort models for the Lexis diagram". In: *Statistics in medicine* 26.15 (2007), pp. 3018–3045.

[11] Bendix Carstensen and Niels Keiding. "Age-Period-Cohort models: Statistical inference in the Lexis diagram". In: *Lecture Notes, Department of Biostatistics, University of Copenhagen. http://www. heart-intl. net/HEART/011507/AgePeriodCohort. pdf* (2005).

[12] David Clayton and E Schifflers. "Models for temporal variation in cancer rates. II: age–period–cohort models". In: *Statistics in medicine* 6.4 (1987), pp. 469–481.

[13] Xiao Dong, Brandon Milholland, and Jan Vijg. "Evidence for a limit to human lifespan". In: *Nature* 538.7624 (2016), pp. 257–259.

[14] Louis I Dublin, Alfred James Lotka, and Mortimer Spiegelman. "Length of life: A study of the life table". In: *(No Title)* (1949).

[15] Ethan Fosse and Christopher Winship. "Analyzing age-period-cohort data: A review and critique". In: *Annual Review of Sociology* 45 (2019), pp. 467–492.

[16] Ethan Fosse and Christopher Winship. "Moore–Penrose estimators of age–period–cohort effects: Their interrelationship and properties". In: *Sociological Science* 5 (2018), p. 304.

[17] JF Fries. *Aging, natural death, and the compression of morbidity. fu: New England Journal ofMedicine 303*. 1980.

[18] A Gelman and D B Rubin. "Inference from iterative simulation using multiple sequences". In: *Statist. Sci* 7 (1992), pp. 457–472.

[19] M. Heron. *Deaths: Leading causes for 2019*. Vol. 70. National Vital Statistics Reports 9. Hyattsville, MD: National Center for Health Statistics, 2021. DOI: 10.15620/cdc:107021.

[20] Patrick Heuveline and Gail B Slap. "Adolescent and young adult mortality by cause: age, gender, and country, 1955 to 1994". In: *Journal of Adolescent Health* 30.1 (2002), pp. 29–34.

[21] MD Hoffman and A Gelman. "The no-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo". In: *J. Mach. Learn. Res.* 15 (2014), pp. 1–31.

[22] Theodore R Holford. "Age–period–cohort analysis". In: *Wiley StatsRef: Statistics Reference Online* (2014), pp. 1–25.

[23] Rob J Hyndman and Yeasmin Khandakar. "Automatic time series forecasting: the forecast package for R". In: *Journal of statistical software* 27 (2008), pp. 1–22.

[24] Søren Kjærgaard et al. "Forecasting causes of death by using compositional data analysis: the case of cancer deaths". In: *Journal of the Royal Statistical Society Series C: Applied Statistics* 68.5 (2019), pp. 1351–1370.

[25] Nan Li and Ronald Lee. "Coherent mortality forecasts for a group of populations: An extension of the Lee-Carter method". In: *Demography* 42 (2005), pp. 575–594.

[26] Ian B MacNeill, Y Mao, and L Xie. "Modeling heteroscedastic age-period-cohort cancer data". In: *Canadian Journal of Statistics* 22.4 (1994), pp. 529–539.

[27] Stefano Mazzuco, Bruno Scarpa, and Lucia Zanotto. "A mortality model based on a mixture distribution function". In: *Population Studies* 72.2 (2018), pp. 191–200.

[28] France Meslé. "Life expectancy: a female advantage under threat". In: *Population and Societies* 402.4 (2004), pp. 1–4.

[29] R Neal. "MCMCusing Hamiltonian Dynamics. In Handbook of Markov Chain Monte Carlo". In: *a Raton: Chapman and Hall–CRC* (2010).

[30] A. Nigri, E. Barbi, and S. Levantesi. "The relationship between longevity and lifespan variation". In: *Statistical Methods and Applications* (2021). DOI: 10.1007/s10260-021-00584-4.

[31] A. Nigri, E. Barbi, and S. Levantesi. "The relay for human longevity: Country-specific contributions to the increase of the best-practice life expectancy". In: *Quality Quantity* (2022). DOI: 10.1007/s11135-021-01298-1.

[32] Robert M O'Brien. "Mixed models, linear dependency, and identification in age-period-cohort models". In: *Statistics in Medicine* 36.16 (2017), pp. 2590–2600.

[33] Robert M O'brien. "Constrained estimators and age-period-cohort models". In: *Sociological Methods & Research* 40.3 (2011), pp. 419–452.

[34] Jim Oeppen and James W Vaupel. *Broken limits to life expectancy.* 2002.

[35] S Jay Olshansky, Bruce A Carnes, and Christine Cassel. "In search of Methuselah: estimating the upper limits to human longevity". In: *Science* 250.4981 (1990), pp. 634–640.

[36] S Jay Olshansky, Bruce A Carnes, and Aline Désesquelles. "Prospects for human longevity". In: *Science* 291.5508 (2001), pp. 1491–1492.

[37] S Jay Olshansky, Leonard Hayflick, and Bruce A Carnes. "Position Statement on Human Aging: This article originally appeared on the Scientific American Web site. Reprinted by permission of the authors." In: *Science of Aging Knowledge Environment* 2002.24 (2002), pe9–pe9.

[38] S Jay Olshansky et al. "A potential decline in life expectancy in the United States in the 21st century". In: *New England Journal of Medicine* 352.11 (2005), pp. 1138–1145.

[39] Marius D Pascariu, Vladimir Canudas-Romo, and James W Vaupel. "The double-gap life expectancy forecasting model". In: *Insurance: Mathematics and Economics* 78 (2018), pp. 339–350.

[40] Adrian E Raftery et al. "Bayesian probabilistic projections of life expectancy for all countries". In: *Demography* 50.3 (2013), pp. 777–801.

[41] Willard L Rodgers. "Estimable functions of age, period, and cohort effects". In: *American sociological review* (1982), pp. 774–787.

[42] Susan B Sorenson. "Gender disparities in injury mortality: consistent, persistent, and larger than you'd think". In: *American journal of public health* 101.S1 (2011), S353–S358.

[43] "Stan Development Team". In: *Stan Modeling Language Users Guide and Reference Manual* 15 (2015).

[44] Tiziana Torri and James W Vaupel. "Forecasting life expectancy in an international context". In: *International Journal of Forecasting* 28.2 (2012), pp. 519–531.

[45] Shripad Tuljapurkar, Nan Li, and Carl Boe. "A universal pattern of mortality decline in the G7 countries". In: *Nature* 405.6788 (2000), pp. 789–792.

[46] Berkeley (USA) University of California and Max Planck Institute for Demographic Research (Germany). *Human Mortality Database*. `http://www.mortality.org`. 2021.

[47] I. Waldron, ed. *Contributions of changing gender differences in behavior and social roles to changing gender differences in mortality*. SAGE Publications, Inc., 1995. DOI: `10.4135/9781452243757`.

[48] Yang Yang, Wenjiang J Fu, and Kenneth C Land. "A methodological comparison of age-period-cohort models: the intrinsic estimator and conventional generalized linear models". In: *Sociological methodology* 34.1 (2004), pp. 75–110.

[49] Yang Yang et al. "The intrinsic estimator for age-period-cohort analysis: what it is and how to use it". In: *American Journal of Sociology* 113.6 (2008), pp. 1697–1736.

[50] Lucia Zanotto, Vladimir Canudas-Romo, and Stefano Mazzuco. "A mixture-function mortality model: illustration of the evolution of premature mortality". In: *European Journal of Population* 37.1 (2021), pp. 1–27.

[51] Wenyun Zuo et al. "Advancing front of old-age human survival". In: *Proceedings of the National Academy of Sciences* 115.44 (2018), pp. 11209–11214.

Appendix

The ARIMA model comprises three parameters, namely p, d, and q, which represent the order of auto-regression, differencing, and moving average, respectively.
The autoregression component (AR) represents the relationship between the current observation and its past values. The p in ARIMA(p,d,q) denotes the order of autoregression, which specifies how many past time steps are used to predict the current value. The integration component (I) refers to the differencing of the time series data. Differencing involves subtracting the previous value from the current value to make the data stationary. The d in ARIMA(p, d, q) indicates the order of differencing needed to achieve stationarity. The moving average component (MA) represents the dependency between the current observation and the error terms from past observations, capturing the short-term dependencies. As usual, The q in ARIMA(p, d, q) specifies the order of the moving average, indicating how many past error terms are considered for predicting the current value.

Mathematically, the ARIMA(p,d,q) for the period effect (and equivalently for the cohort effect) can be represented as:

$$\overset{d}{\nabla}\beta_t^P = \delta + \sum_{i=1}^{p} \phi_i \overset{d}{\nabla}\beta_{t-i}^P + \epsilon_t + \sum_{j=1}^{p} \theta_j \epsilon_{t-j} \tag{5.1}$$

Where $\overset{d}{\nabla}$ represents the differencing operator that calculates the difference between consecutive observations at a lag of d time periods, δ is the drift process, ϕ_i are the autoregressive parameters, ϵ_t are the error terms (normally distributed with zero mean and variance σ^2) and θ_j are the moving average parameters.

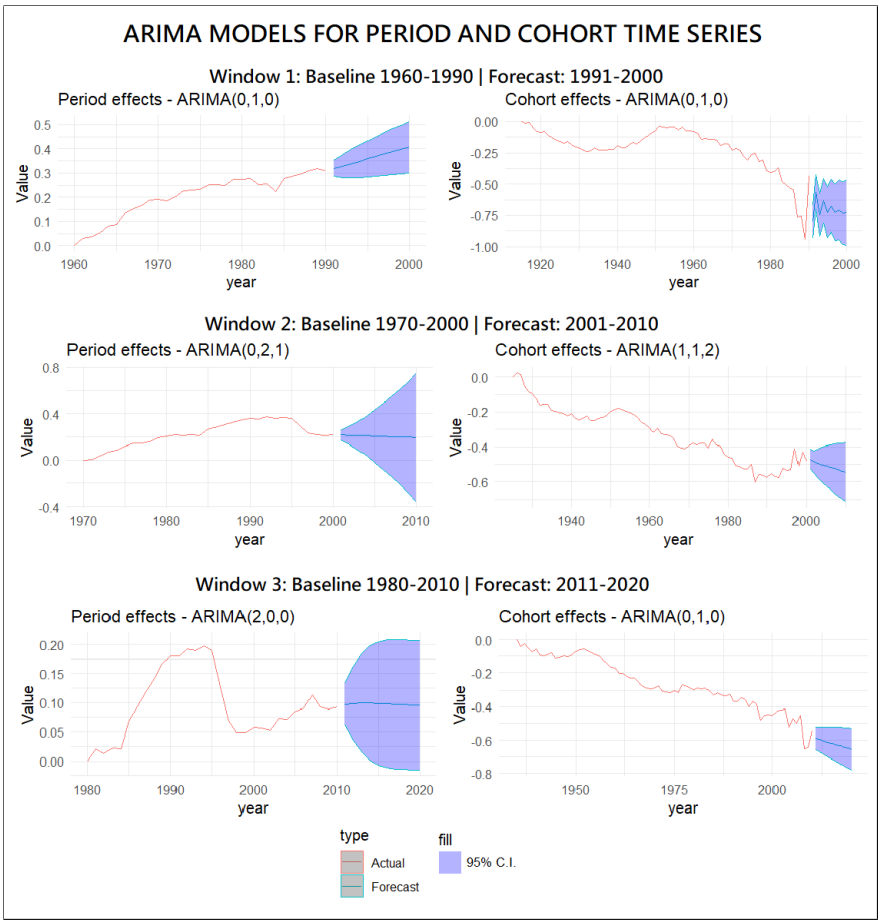

Figure 4: Period and cohort effects projected without any constraint in the three different baseline time intervals and relative 10-year forecast using the best ARIMA model.

Conflicts of Interest Statement

All authors declare that they have no conflicts of interest.