STOCHASTIC MODELLING OF FERTILITY RATES – SEARCHING FOR OPTIMAL MODEL

Ondřej Šimpach

Abstract

Modelling of the demographic processes and their projection is important as the results serve to the policymakers for planning. We focus on stochastic Lee-Carter model that consist of time-independent parameters $\mathbf{a}_x$ (age-specific component, fertility profile), $\mathbf{b}_x$ (the additional age-specific components that determine how much each age group changes when parameter $\mathbf{k}_t$ changes) and time dependent index $\mathbf{k}_t$. $\mathbf{a}_x$ is a simple arithmetic mean of fertility rates across time for each age group. Also, $\mathbf{b}_x$ is similar across time for each age group. Therefore, the index $\mathbf{k}_t$ is the only element of the model that must be modelled and projected. The aim of the paper is to find optimal model for $\mathbf{k}_t$ index that ensures that realistic fertility rates are projected. Based on comparison of models it was found that optimal for modelling of $\mathbf{k}_t$ is ARIMA(1,2,1). However, the projection to the future by this model was not optimal, so there must be ways searched how to smooth the $\mathbf{k}_t$ in problematic points and how to make reasonable projection to the future. The results are compared to the assumptions of the Czech Statistical Office about the development of total fertility rates. Our total fertility rates are higher than its optimistic assumptions.

Key words: fertility rates, Lee-Carter model, time dependent index $\mathbf{k}_t$

JEL Code: J13, C53

Introduction

There are two types of demographic projection possibilities – deterministic and stochastic methods. Deterministic methods such as cohort-component method are based on the scenarios of the future development and expert judgements. In the cohort-component method, the components of population change (fertility, mortality, and net migration) are projected separately for each birth cohort (persons born in a given year). Subjective element of the projections can be problematic, especially in changing conditions in the society. Besides, demographic processes are influenced by many unpredictable circumstances.

The method is widely used. For example, Colvin, McLaughlin & Richmond (2022) derived the population estimates by the cohort component method for the population in Ireland in years 1911–1920. Rees et al. (2012) projected the United Kingdom's ethnic group populations for year 2001–2051 in 4 scenarios (benchmark, two trends' projections and projection with certain migration assumptions). Czech Statistical Office (2023a) projects the state of population in middle, low and high variant.

On the other hand, stochastic methods are based on persistent long-term trends, include random component and provide confidence intervals for the projections (e. g. the boundaries in which the values will be with a certain probability).

Lee and Carter (1992) introduced a method for forecasting of the mortality to the future „consisting of a base model of age-specific death rates with a dominant time component and a fixed relative age component and a time series model ARIMA of the time component (Booth, Maindonad & Smith, 2002). The model can be applied not only to mortality rates but also to fertility rates of particular population. A forecast is done through a standard time-series model on the time component, while considering the same structure of the age-specific mortality / fertility level over time. (Rabbi & Mazzuco, 2021).

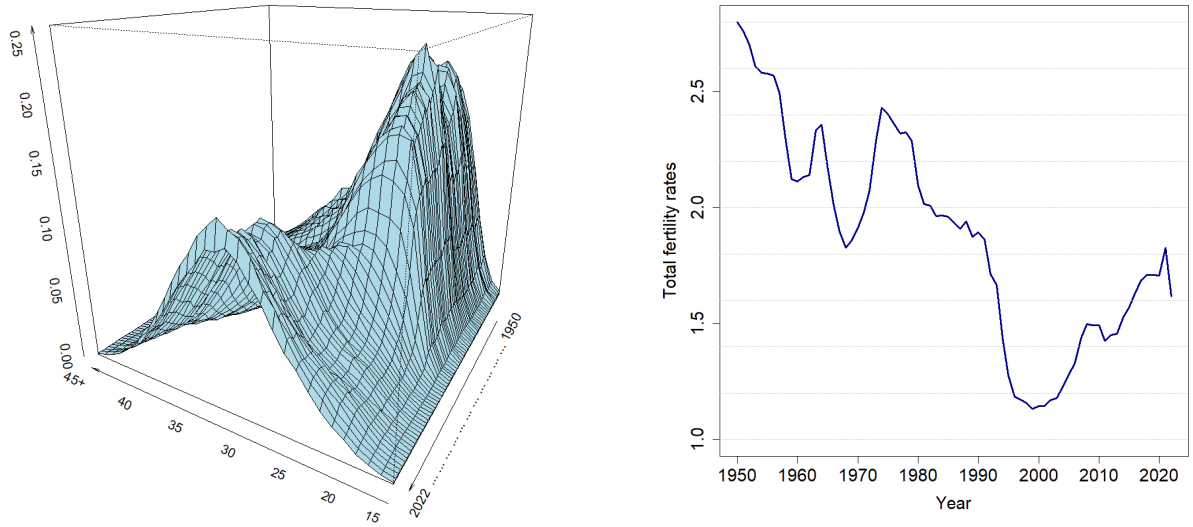
However, there is a disadvantage that the Lee-Carter method of mortality forecasting assumes an invariant age component and most applications have adopted a linear time component. (Booth, Maindonad & Smith, 2002). This is unrealistic especially in developed countries with mortality decline over time. Also, when this model is applied on fertility, there are certain limitations. "Fertility has proved difficult to forecast due to structural change and estimates of uncertainty are highly dependent on the particular model." (Hyndman & Booth, 2008). Many modifications to the original Lee-Carter model have been proposed. For example, Lee & Miller (2001) re-estimated the time component of mortality model according to the observed life expectancy at birth. We use original form of the model, but search for optimal time component using various ARIMA models.

1 Data and Methods

Empirical data about age-specific fertility rates with annual frequency and 1-year age interval were taken from the Czech Demographic Handbook (CZSO, 2023b). The data were available for ages from 15 to 45 years where the marginal categories included also babies born to younger or older women. The longest available time period from 1950 to 2022 is taken. From Fig. 1 of age-specific fertility rates can be seen the decline of fertility in time and also clear

shift of age where women give birth to the child from lower ages in the past to higher ages in the present. Fig. 1 of total fertility rates shows the development in time where especially the decline between 1990–2000 is significant.

Fig. 1: Empirical data of age-specific fertility rates in the Czech Republic 1950–2022 (left) and of total fertility rates (right)



Source: Own elaboration based on CZSO (2023b) data

The modelling and projections of fertility rates are done by original Lee-Carter model. The approach is based on the latent class analysis (LCA) and decompose empirical values of age-specific fertility rates as (1).

$$f_{x,t} = a_x + b_x k_t + \varepsilon_{x,t}, \tag{1}$$

where $f_{x,t}$ denotes the age-specific fertility rates, x stays for 15–49 years of life, t is the time period $t = 1, 2, ..., T$. Parameters a_x and b_x are time independent. Age specific component a_x is calculated as arithmetic mean of age-specific fertility rates $a_{x,t} = \frac{\sum_{t=1}^{T} f_{x,t}}{T}$. b_x are the additional age-specific components that determine how much each age group changes when parameter k_t changes. Total fertility indexes k_t are the time-varying parameters and therefore have to be projected to the future in separate model. $\varepsilon_{x,t}$ is the error term with characteristics of the white noise. Under the Lee-Carter method, the errors are expected to have the same variance over all ages (homoscedastic) which does not have to be true in many cases. (Koissi, Shapiro & Högnäs, 2006).

The estimation of b_x and k_t is based on Singular Value Decomposition of matrix of age-specific fertility rates (Hyndman & Booth, 2008). The identification of the model is

ensured by the condition that the sum of all age-specific components is equal to one ($\sum_{x=15}^{45} b_x = 1$) and that the sum of all time-varying indexes is equal to zero ($\sum_{t=1}^{73} k_t = 0$). Fertility rates are included to the model in natural logarithms.

The aim of the paper is to find optimal model for $\mathbf{k}_t$ index that would ensure that realistic fertility rates are projected. First, `auto.arima` command in library `forecast` in R is used. Second, we projected pre-selected ARIMA model using `arima` command. ARIMA(*p*, *q*, *d*) model with drift can be expressed by (2) and consists of constant, moving average process (MA) and autoregression process (AR).

$$Y_t = c + \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} \qquad (2)$$

where Φ and θ are parameters of the lagged explained variable (Y_{t-i}) and lagged stochastic term (ε_{t-j}), respectively. To determine the order *p* of AR process and order *q* of MA process the Autocorrelation function (ACF) and Partial Autocorrelation function (PACF) are plotted. Correlograms of ACF and PACF are simply the plots of ACF and PACF against the lag length (Wang & Zhao, 2009). The order of lags is determined based on Akaike information criterion (AIC).

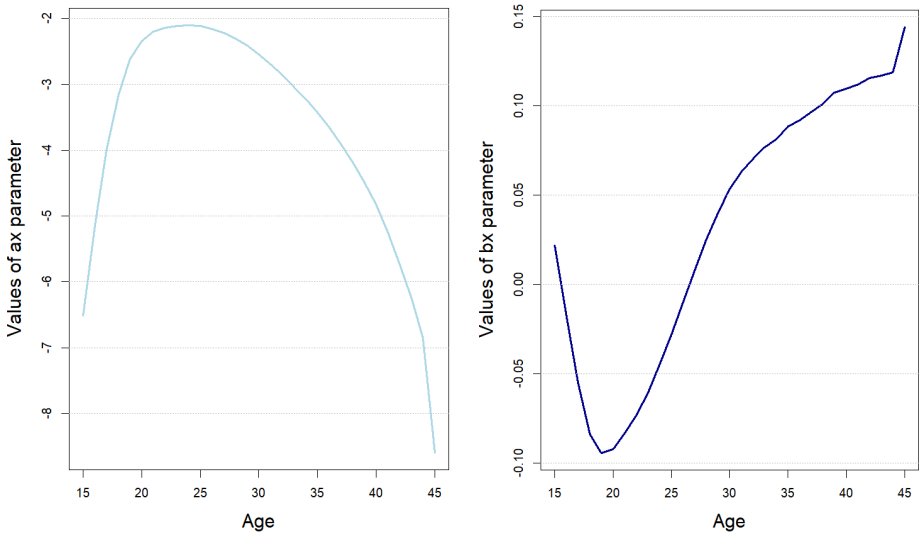
We compare our results with the middle, low and high variant projection of the total fertility rates of the CZSO (2023a) for year 2023 to 2100 which is displayed at Fig. 6. Middle variant of the CZSO (2023a) assumes total fertility rates to be unchanged until 2100 in height 1.50. There is a decline of fertility rates projected in low variant – from 1.45 in 2023 to 1.25 at the end of the projection period. On the other hand, high variant starts with total fertility rate 1.62 and stay constant from 2049 to 2100 on value 1.75.

Age-specific fertility rates are recalculated to total fertility rates as simple the sum of age-specific fertility rates (of women aged 15 to 45 years).

2 Results

The age-specific fertility rates were included into the Lee-Carter model and its parameters were estimated. We obtained vector of age specific fertility rates for each age $\mathbf{a}_x$ (31 x 1), vector of time independent parameter $\mathbf{b}_x$ (31 x 1), and time dependent index $\mathbf{k}_t$ (73 x 1). The development of the time-invariant parameters is displayed at Fig. 2. Parameter $\mathbf{a}_x$ has its typical concave shape with maximum at 24 years and $\mathbf{b}_x$ is increasing with age.

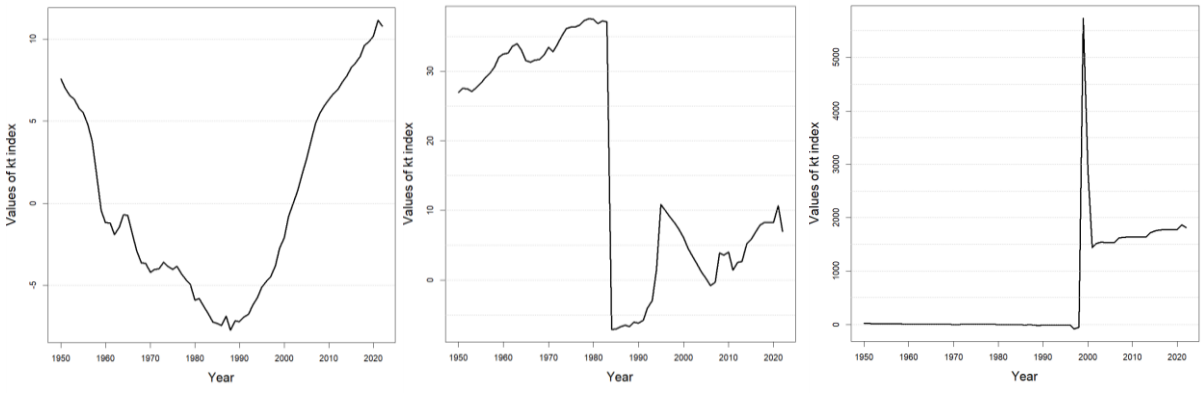
Fig. 2. Development of the parameters a_x – age specific profile (left) and b_x – fertility fluctuations according to the age (right) of Lee-Carter model



Source: Own elaboration

Three types of index $\mathbf{k}_t$ were estimated. First one was original index without any adjustment. Its development is presented at Fig. 3 on the left. Then, there was $\mathbf{k}_t$ index adjusted according to the methodology of Lee and Carter ("`dt`") and using "`e0`" (method based on life expectancy). However, both adjustments are meant for mortality modelling, so the results are not reasonable. Therefore, we fitted the model without any adjustments of $\mathbf{k}_t$.

Fig. 3: Parameter k_t of Lee-Carter model: without adjustment (left), adjusted according to Lee-Carter methodology (middle), adjusted according to e_0 (right)



Source: Own elaboration

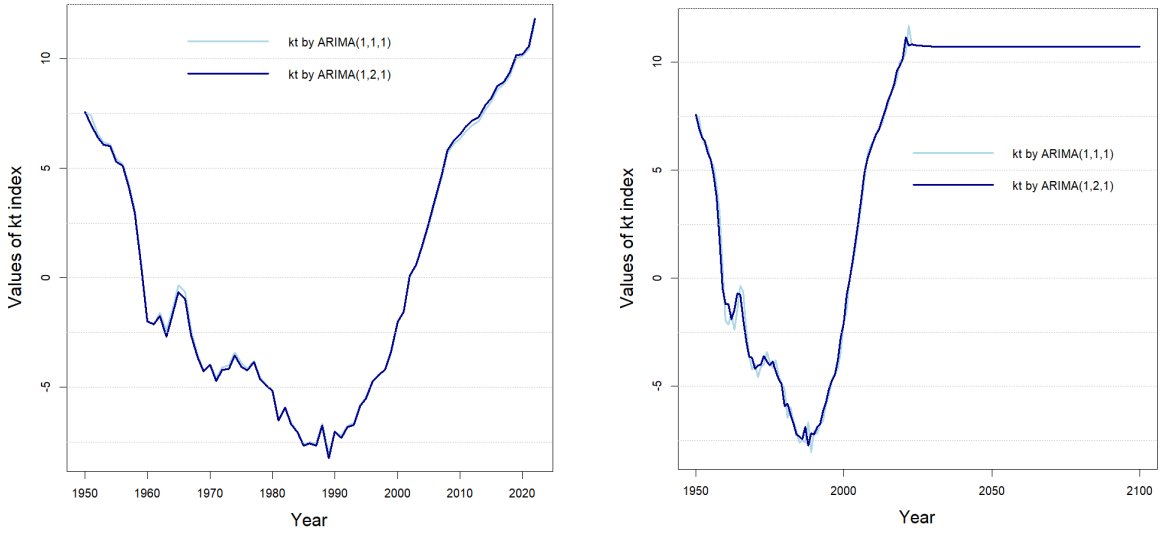
For projection of the index, it is necessary to find suitable ARIMA(p, d, q) model. We used `auto.arima` command in library `forecast` for automatic selection of the

parameters *p*, *d*, *q*. This command chose ARIMA (1,2,1) process. The model was chosen without the drift (without constant). The second difference of the time series must be done before it is stationary. The value of $\mathbf{k}_t$ index depend only on previous value of $\mathbf{k}_t$ and white noise.

Second, we projected pre-selected ARIMA model using `arima` command. This time we chose ARIMA(1,1,1). There is only one differentiation done in this model. The fit of both models is relatively similar as it is evident from the Fig. 4 (left). Detailed results of the models are in the Table 1 in the Attachment. Comparison of various models for mortality rates can be found e. g. in Šimpach (2023). The results are very comparable to fertility rates.

Projection to the year 2100 expects that $\mathbf{k}_t$ is almost constant. In year 2023, the first projection year, there is still continuing increase of the previous trend which decreases and stabilizes only next year which can be seen from Fig. 4 (right). The deflection between real data and the projection is problematic and cause the deflection also in the fertility rates. Therefore, the model with lower deflection shall be preferred.

Fig. 4: Total fertility index k_t of Lee-Carter model fitted by ARIMA(1,2,1) and ARIMA(1,1,1) models: time period 1950–2022 (left) and time period 1950–2100 (right)

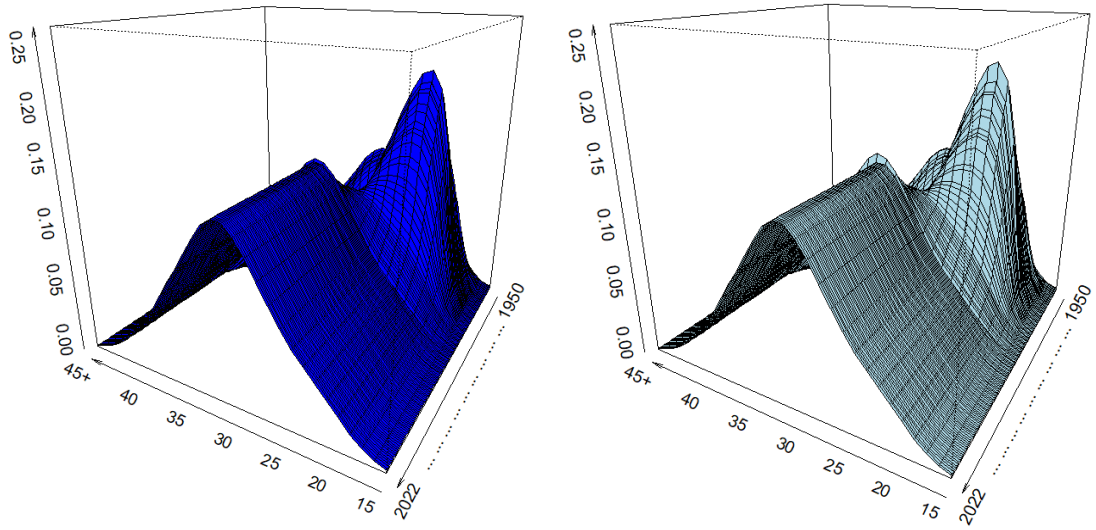


Source: Own elaboration

The consequently projected age-specific fertility rates are displayed at Fig. 5. There is a visible deflection (breakpoint) on the border between fitted and projected values. The difference is lower for ARIMA(1,2,1) model. Hence, the weakest point of the Lee-Carter

method is the projection of $\mathbf{k}_t$ index to the future because even though its fit on historical data is very good, there is a bias when moving into the projection horizon. Also, the projection of the fertility rates is then biased. The highest difference is in age 32 years. Besides, the projection of the $\mathbf{k}_t$ in the projection horizon is close to be straight line. Therefore, there must be ways searched how to smooth the $\mathbf{k}_t$ projection in problematic points and how to make reasonable projection to the future.

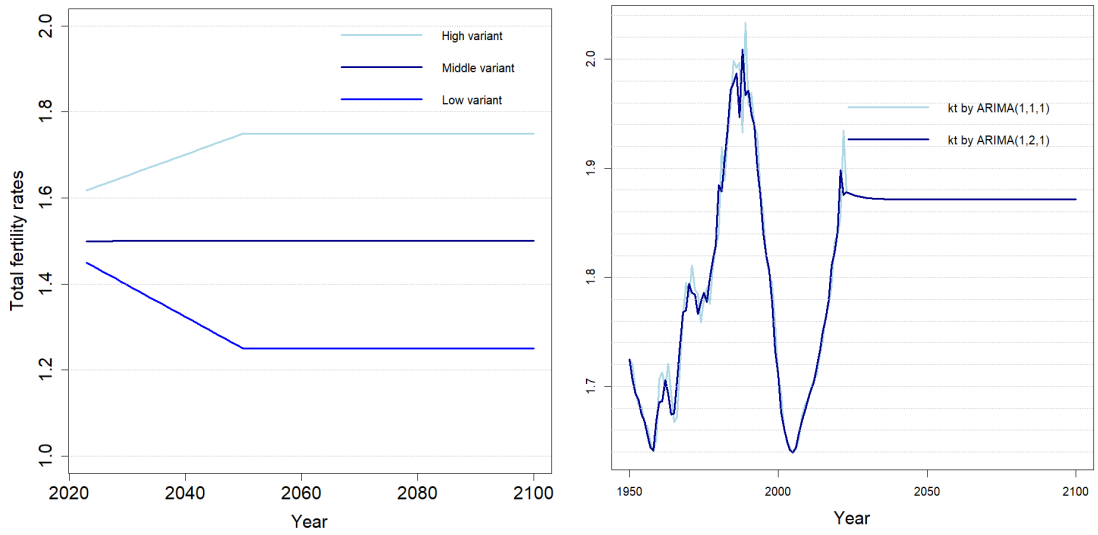
Fig. 5: Fitted and projected age-specific fertility rates with total fertility index k_t fitted and projected by ARIMA(1,2,1) model (left) and ARIMA(1,1,1) model (right)



Source: Own elaboration

Then the total fertility rates were calculated as the sum across all ages. It converges to 1.87 in the future which is the closest to the high assumption of the CZSO and too optimistic considering the last development in the Czech Republic. It is still following the growing trend since 2000, but the pace of increase is slowing down. The year 2022 experienced a decrease in fertility rates which could be expected also in 2023. The 2023 generation was the least numerous in the last 22-year-period. The number of live births decreased year-on-year by 10.2 thousand, or 10%, to 91.1 thousand. A similar decline could be expected for the total fertility rate. (CZSO, 2024).

Fig. 6: Assumptions of development of total fertility rates according to CZSO (left) and fertility rates fitted and projected by ARIMA(1,2,1) and ARIMA(1,1,1) models (right)



Source: Own elaboration based on CZSO (2023a) data (left), own elaboration (right)

Conclusion

The aim of the paper was to find optimal ARIMA model for $\mathbf{k}_t$ index that would ensure that realistic fertility rates are projected by standard Lee-Carter model. First, `auto.arima` command in library `forecast` in R was used. Second, we projected pre-selected ARIMA model using `arima` command.

Automatic algorithm selected ARIMA(1,2,1) model which fitted the historical data better than pre-selected ARIMA(1,1,1). However, both models did not perform well in the breakpoint of historical data and projected data. The last historical value is higher or lower than the first projected value. A deviation arises at the point where the projection does not exactly follow the historical data. It would therefore might be necessary to replace the last historical value with, for example, the arithmetic average of two values: 1) the last but one value of the historical period and 2) the first value of the projection period. Another possibility is to scale (shift) the projection of future values to follow the last historical value.

Some adjustments can be made also to the $\mathbf{a}_x$ parameter (shortening of the period from which it is calculated – see e. g. Šimpach & Arltová, 2016) or to the $\mathbf{b}_x$ parameter (its rotation in case of mortality rates – see e. g. Li, Lee & Gerland, 2013). There is still place for improvements and adjustments of the Lee-Carter model which is also our challenge for the future research.

Acknowledgment

The paper was supported by long-term institutional support of research activities by the Faculty of Informatics and Statistics of Prague University of Economics and Business.

References

Booth, H., Maindonald, J., & Smith, L. (2002). Applying Lee-Carter under conditions of variable mortality decline. Population Studies, 56, 325–336. https://doi.org/10.1080/00324720215935

Colvin, C. L., McLaughlin, E., & Richmond, K. J. (2022). Cohort component population estimates for Ireland, 1911–1920: A new county-level dataset for use in historical demography. Research Data Journal for the Humanities and Social Sciences, 7(1), 1–18. https://doi.org/10.1163/24523666-bja10022

Czech Statistical Office (2023a). Projekce obyvatelstva České Republiky 2023–2100. Available from: https://www.czso.cz/documents/10180/191186777/13013923.pdf/2cf4ca35-38fa-46a9-9c37-eeadde18a4ea?version=1.1 (Cit. 5th Apr. 2024)

Czech Statistical Office (2023b). Czech Demographic Handbook – 2022. Available from: https://www.czso.cz/csu/czso/czech-demographic-handbook-2022 (Cit. 23rd March 2024)

Czech Statistical Office (2024). Population change – year 2023. Population of the Czech Republic exceeded 10.9 million. Available from: https://www.czso.cz/csu/czso/ari/population-change-year-2023 (Cit. 30th Apr. 2024)

Hyndman, R. J., & Booth, H. (2008). Stochastic population forecasts using functional data models for mortality, fertility and migration. International Journal of Forecasting, 24(3), 323–342. https://doi.org/10.1016/j.ijforecast.2008.02.009

Koissi, M. C., Shapiro, A. F., & Högnäs, G. (2006). Evaluating and extending the Lee–Carter model for mortality forecasting: Bootstrap confidence interval. Insurance: Mathematics and Economics, 38, 1–20. https://doi.org/10.1016/j.insmatheco.2005.06.008

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

Lee, R. D., & Miller, T. (2001). Evaluating the performance of the Lee–Carter method for forecasting mortality. Demography, 38(4), 537–549. https://doi.org/10.1353/dem.2001.0036

Li, N., Lee, R. D. & Gerland, P. (2013). Extending the Lee-Carter Method to Model the Rotation of Age Patterns of Mortality Decline for Long-Term Projections. Demography, 50, 2037–2051. https://doi.org/10.1007/s13524-013-0232-2

Rabbi, A. M. F, & Mazzuco, S. (2021). Mortality Forecasting with the Lee–Carter Method: Adjusting for Smoothing and Lifespan Disparity. European Journal of Population, 37(1), 97–120. https://doi.org/10.1007/s10680-020-09559-9

Rees, P., Wohland, P., Norman, P., & Boden, P. (2012). Ethnic population projections for the UK, 2001-2051. Journal of Population Research, 29(1), 45–89. https://doi.org/10.1007/s12546-011-9076-z

Šimpach, O., & Arltová, M. (2016). An increasing of prediction power of the Lee-Carter model: The case of Czech and Spanish age-specific fertility rates' forecasting. In: ITISE 2016. Granada: University of Granada, 453–462. ISBN 978-84-16478-93-4.

Šimpach, O. (2023). Mortality Forecasting – Which Arima Model to Choose for Vector $\mathbf{K}_t$ Projection in Lee-Carter Model? In: Hradec Economic Days, 690–702. https://doi.org/10.36689/uhk/hed/2023-01-065

Wang, H. J., & Zhao, W. G. (2009). ARIMA Model Estimated by Particle Swarm Optimization Algorithm for Consumer Price Index Forecasting. Artificial Intelligence and Computational Intelligence, 5855, 48–58.

Attachment

Table 1. Comparison of ARIMA models

Model	**ARIMA (1,2,1)**		**ARIMA (1,2,1)**	
	AR (p = 1)	**MA (q = 1)**	**AR (p = 1)**	**MA (q = 1)**
Coefficients	0.3202	-0.8289	0.8683	-0.4046
Standard error	0.1607	0.0948	N/A	N/A
Models' diagnostics				
σ^2	0.2553		0.2447	
log likelihood	-51.59		-51.85	
AIC	109.19		109.70	
AICc	109.55		N/A	
BIC	115.98		N/A	

Source: Own elaboration

Contact

Ondřej Šimpach

Prague University of Economics and Business, Faculty of Informatics and Statistics, Department of Statistics and Probability

W. Churchill Sq. 1938/4, 130 67 Prague 3, Czech Republic

ondrej.simpach@vse.cz