doi: 10.36689/uhk/hed/2023-01-065

Mortality Forecasting – Which Arima Model to Choose for Vector K_T Projection in Lee-Carter Model?

Ondřej ŠIMPACH

Prague University of Economics and Business, Prague, Czech Republic; ondrej.simpach@vse.cz

Abstract: The aim of the paper is to compare two possible methods of projections of the vector k_t in the model and to forecast the age and sex-specific mortality rates in the Czech Republic for next 20 years based on stochastic modelling method. Particularly, we used Lee-Carter model that is based on Principal Component Analysis and is estimated by Singular Value Decomposition method. It estimates age-dependent parameters a_x and b_x and time-dependent vector k_t. Projection of the model's vector k_t can be done by the best ARIMA model or by average ARIMA model based on the smoothed Akaike information criterions. Both k_t are used for forecasting of the mortality rates for period 2021–2040. Forecasting was done in R package demography. The parameters a_x and b_x had expected development. However, best ARIMA model chosen by package forecast did not project k_t realistically. External calculation in EViews software chose average ARIMA model that created better k_t projections. This is reflected in the results of the forecasts of the mortality rates $m_{x,t}$. It can be concluded that the forecasts of mortality rates based on the k_t parameter projected in software EViews is smoother and more realistic.

Keywords: forecast; Lee-Carter model; mortality rates

JEL Classification: J11; C38; C18

1. Introduction

Mortality is one of the demographic processes that are an integral part of the population projections. Projection of the population state is needed for the decision-making purposes. For example, Fiala, Langhamrová, and Průša (2011) projected the human capital (approximated by the education of the population) of the Czech Republic and its regions to year 2050.

There are two approaches towards the demographic projections. However, "nowadays it is not enough to construct the demographic projections based on the deterministic models." (Šimpach & Langhamrová, 2014). Therefore, stochastic model that accounts for random errors are used despite that they are more computationally intensive than deterministic models.

The first stochastic model elaborated by Lee and Carter (1992) has been used since that on many applications and many extensions that improves its forecasting functions has been added. Its advantage is relatively easy computation based on Singular Value Decomposition method. "Among the forecasting advantages are the minimal subjective judgement required (and the relative accuracy of forecasts compared to those based on methods incorporating greater judgement), and the production of probabilistic prediction intervals." (de Jong et al., 2020)

On the other hand, Lee-Carter model is based on historical mortality data and project them into the future. It extrapolates historical trends and forecasts probability distributions of age-specific death rates using standard time-series procedure. (Li & Lee, 2005). This imply that when there is an exogenous shock, it can be reflected in the forecasts (they can be distorted). Despite that the oldest history has the lowest weight in the prediction model it can quite be important even with a little weight, because mortality is a long-term process that has for each population its long-term trend. (Booth et al., 2005).

Another disadvantage of the Lee-Carter model is that the age-specific set of the b_x parameter is estimated on the basis of historical data and does not develop in time. Parameter b_x indicates which rates decline rapidly and which rates decline slowly in response to changes in k_t (time-varying index of the level of mortality for all ages). However, mortality in countries with low infant mortality (Czech Republic is one of them) declines faster in older ages than in young ages (so-called ageing of mortality decline). The assumption of the Lee-Carter model that the pattern of change in mortality is fixed over time is too strong. The shortcomings of the model were evaluated e.g. by Lee and Miller (2001).

Original model was applied by Lee and Carter (1992) on the U.S. population data from year 1933 to 1987 and forecasted mortality rates up to year 2065. de Jong et al. (2020) enlarged the time series of U.S. mortality rates on years 1933–2017 and extended the model by normalization of the parameters, so they have a direct and intuitive interpretation, comparable across populations. They also introduced “needed-exposure” which is “the number required in order to get one expected death and is closely related to the “needed-to-treat” measure used to communicate risks and benefits of medical treatments” (de Jong et al., 2020).

Disadvantage of b_x parameter was solved by Li et al. (2013) who suggested its rotation that is the phenomenon that in developed countries, mortality decline is decelerating at younger ages and accelerating at old ages. They used mortality rates from Japan and U.S. and forecasted the mortality rates up to year 2098.

Booth et al. (2002) fitted the Lee-Carter model to Australian data for 1907–1999 and found that the “universal pattern” of constant mortality decline as represented by linear k_t did not hold over that fitting period. Therefore, they modified in the later study the method to adjust the time component to reproduce the age distribution of deaths, rather than total deaths, and to determine the optimal fitting period in order to address non-linearity in the time component. (see Booth et al., 2010).

In the Czech Republic, Šimpach and Langhamrová (2014) modelled age-specific mortality rates in the period 1920–2012 by individual random walk with drift and by Lee-Carter model based on Principal Component Analysis method. They concluded that results obtained by Lee-Carter model reflect better the expectation. The random walk models can be used only for the female population. In our article stochastic approach to forecast the mortality rates was used.

The Czech Republic belongs to the developed countries with relatively low mortality rates. As can be seen in Figure 1, the development since the year 1960 was not favourable, as the total (crude) mortality rate had been increasing up to year 1980 when it was the highest (almost 132 deaths per 1,000 inhabitants). Then, it started to decline and reached its minimum in year 2014 (100 deaths per 1,000 inhabitants). Year 2020 noted sudden increase of the crude mortality rate due to the Covid-19 pandemic. The trend of decreasing mortality (expressed by 5-years moving average) was broken.

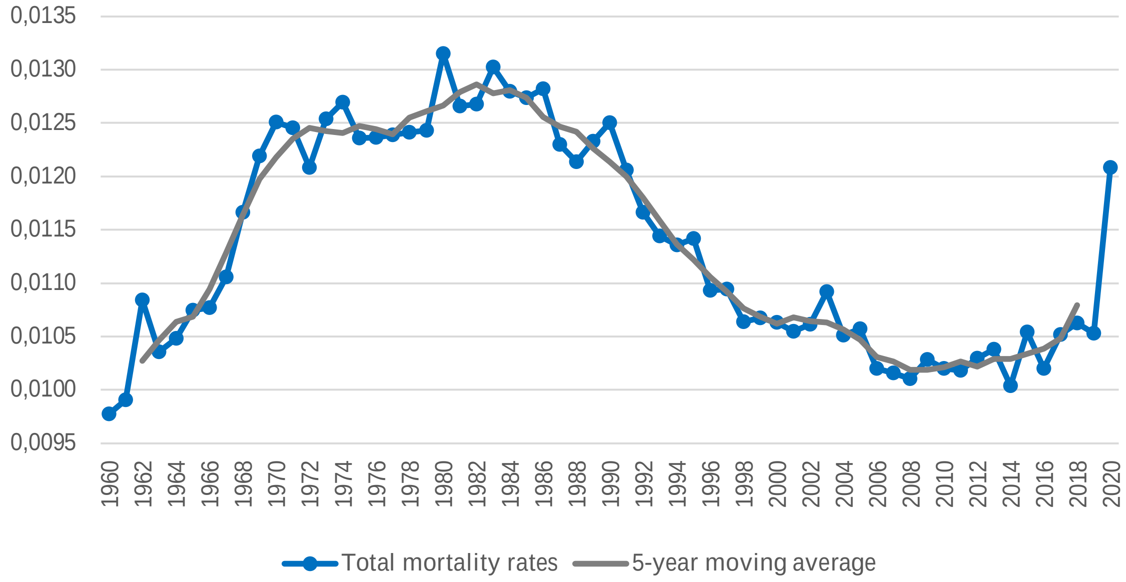

Figure 1. Development of total (crude) mortality rates and trend expressed by 5-year moving average in the Czech Republic in 1960–2020 (Source: Eurostat (2022), own elaboration)

An important factor that must be taken into account when assessing mortality is the age composition of the relevant population. Differences in mortality are also evident between the sexes. Therefore, age and sex-specific mortality rates are taken into account for forecasts. Usually one or five-years age categories are considered and the forecasts are done separately for males, females and total population.

2. Methodology

The data about mortality rates were gathered and calculated and Lee-Carter model was applied on them to forecast the mortality rates in the Czech Republic for years 2021–2040.

2.1. Data

The dataset of mortality rates $m_{x,t}$ is needed for the calculation. Eurostat provides data about mortality rates from year 1960 up to year 2020 in *Life table by age and sex*. However, the last category is only 85 year or over that is not sufficient for the analysis of mortality in the highest ages. Therefore, the death rates (central mortality rate in time *t*, $m_{x,t}$) were calculated as division of number of deaths in time *t*, $D_{x,t}$, and mid-year population state in time *t*, $P_{x,t.}$ Data about *Population state at 1st January* and *Deaths by age and sex and population on 1st January* were taken from Eurostat (2022). By the calculation was gained the dataset with 1-year age categories from less than 1 year up to 99+ for total population, males and females for years 1960–2020. The projection period was chosen to be 20 years: 2021–2040.

2.2. Lee-Carter Model

The future mortality rates were forecasted by Lee-Carter model developed by Lee and Carter (1992). The model is based on Principal Component Analysis and is estimated by Singular Value Decomposition method. There are 3 parameters of the model that have to be estimated: age specific term which represents the general mortality shape across age (a_x), age-specific profiles which rates decline rapidly and which rates decline slowly in response to changes in k_t (b_x) and time-varying index of the level of mortality for all ages (k_t). k_t indexes the intensity of mortality.

The model is defined to fit the matrix of mortality rates in a form of exponential function that can be linearized by the natural logarithm (1).

$$\boldsymbol{m}_{x,t} = e^{\boldsymbol{a}_x + \boldsymbol{b}_x \boldsymbol{k}_t + \boldsymbol{e}_{x,t}} \quad or \quad ln(\boldsymbol{m}_{x,t}) = \boldsymbol{a}_x + \boldsymbol{b}_x \boldsymbol{k}_t + \boldsymbol{e}_{x,t} \tag{1}$$

A vector of column indices of mortality rates m_x is used to estimate the a_x parameter. “As the model written in this way is over parametrized, the two additional constraints are introduced in order to identify the model.” (Danesi et al., 2015): $\sum_{x=1}^{N} \mathbf{b}_x = 1$ and $\sum_{t=1}^{T} \mathbf{k}_t = 0$. Using these constrains, the least squares estimator for $\mathbf{a}_x$ can be obtained by (2):

$$\hat{\mathbf{a}}_x = \frac{\sum_{x=1}^{N} \log(\mathbf{m}_{x,t})}{N}. \tag{2}$$

Under this normalization, b_x is the proportion of the change in overall log mortality attributable to age *x*. Both parameters a_x and b_x are time-invariant that is one of their disadvantages.

Vector k_t changes in time, so it is projected to the future. The development of the variable k_t is usually projected by ARIMA (an autoregressive integrated moving average) models elaborated by Box and Jenkins (1970). For example, Lee and Carter (1992) used random walk with drift. Russolillo (2017) used the ARIMA (0,1,0) model to forecast the index of mortality k_t for next 25 years. Šimpach and Dotlačilová (2016) used ARIMA (1,1,0) with drift.

AR (autoregressive) process of the model reflects the development of the dependent variable in time. MA (moving average) process means that the residuum is dependent on its own lags. When only autoregressive and moving average part is present, then we talk about ARMA model that can be used only when the time series is stationary. If the time series is not stationary, its difference of d^{th} order must be done. Than the model is ARIMA (*p*, *d*, *q*), where *p* is the order of AR term, *d* is the number of non-seasonal differences and *q* is the order of MA term (3).

$$Y_t = c + \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} \tag{3}$$

where φ and θ are parameters of the lagged explained variable (Y_{t-i}) and lagged stochastic term (ε_{t-j}), respectively.

Diagnostic of the type of the model is done by Autocorrelation function (ACF) and Partial Autocorrelation function (PACF) that are plotted in order to determine the order *p* of AR process and order *q* of MA process. Correlograms of ACF and PACF are simply the plots of ACF and PACF against the lag length (Wang and Zhao, 2009). The order of lags is determined based on Akaike information criterion.

2.3. Calculation

There are two possibilities, how to project variable k_t to the future. First, the best ARIMA model (according to the Akaike information criterion – AIC) can be used. This can be calculated by package `forecast` in software R that was elaborated by Hyndman (2022) – command `auto.arima`.

Second possibility is to use EViews software with automatic ARIMA forecasting method. This method utilizes an averaging technic based on smooth AIC weights. Algorithm calculates all possible types (combination) of ARIMA models from ARIMA (0,0,0) to ARIMA (4,2,4). Then it selects the best 25 models based on Akaike criterion and calculates the average development of the forecast.

Lee-Carter model itself is estimated and mortality rates are forecasted by Hyndman (2022) package `demography` in R software. Results of forecasted mortality rates by both types of k_t parameters are compared.

3. Results

First, the parameters $a_{x,}$ $b_{x,}$ and k_t of the Lee-Carter model were calculated. Then the vector k_t was projected using package `forecast` in R and ARIMA in EViews. Finally, the mortality rates are projected for period 2021–2040.

3.1. Lee-Carter Model

Figure 2 shows the development of the estimated parameters of Lee-Carter model calculated in package `demography` (Hyndman, 2022). Parameter a_x for general mortality shape across age shows the fact that the mortality is high in the first year of the people's life and then it decreases sharply. It starts to increase around the age of 10 again with slow down between ages 20 to 30. It only grows since that. The biggest changes of mortality patterns appear at approximately at age 60 years and above due to low numbers of living at the highest ages that leads to small numbers of deaths. (Šimpach et al., 2014). A kink at the end in the highest ages is caused by the fact that there are only few people living at this age and almost all are dying, hence the data are not accurate. The mortality rates can be smoothed by various methods (see e.g. Šimpach et al., 2014).

Parameter b_x is an age-specific profile that tells which rates decline rapidly and which rates decline slowly in response to changes in $k_{t.}$ It has expected development. It decreases rapidly until the age of 40 and is almost stabilized at certain level until the age of 60. Then it increases up to 80 years and decline afterwards.

However, parameter k_t does not corresponds to the usual development of the time-varying index of the level of mortality for all ages. Its development does not have a decreasing trend but is rather a white noise. In original study of Lee and Carter (1992), k_t declined at a roughly constant rate and had roughly constant variability. It implies that its projection to the future can be difficult.

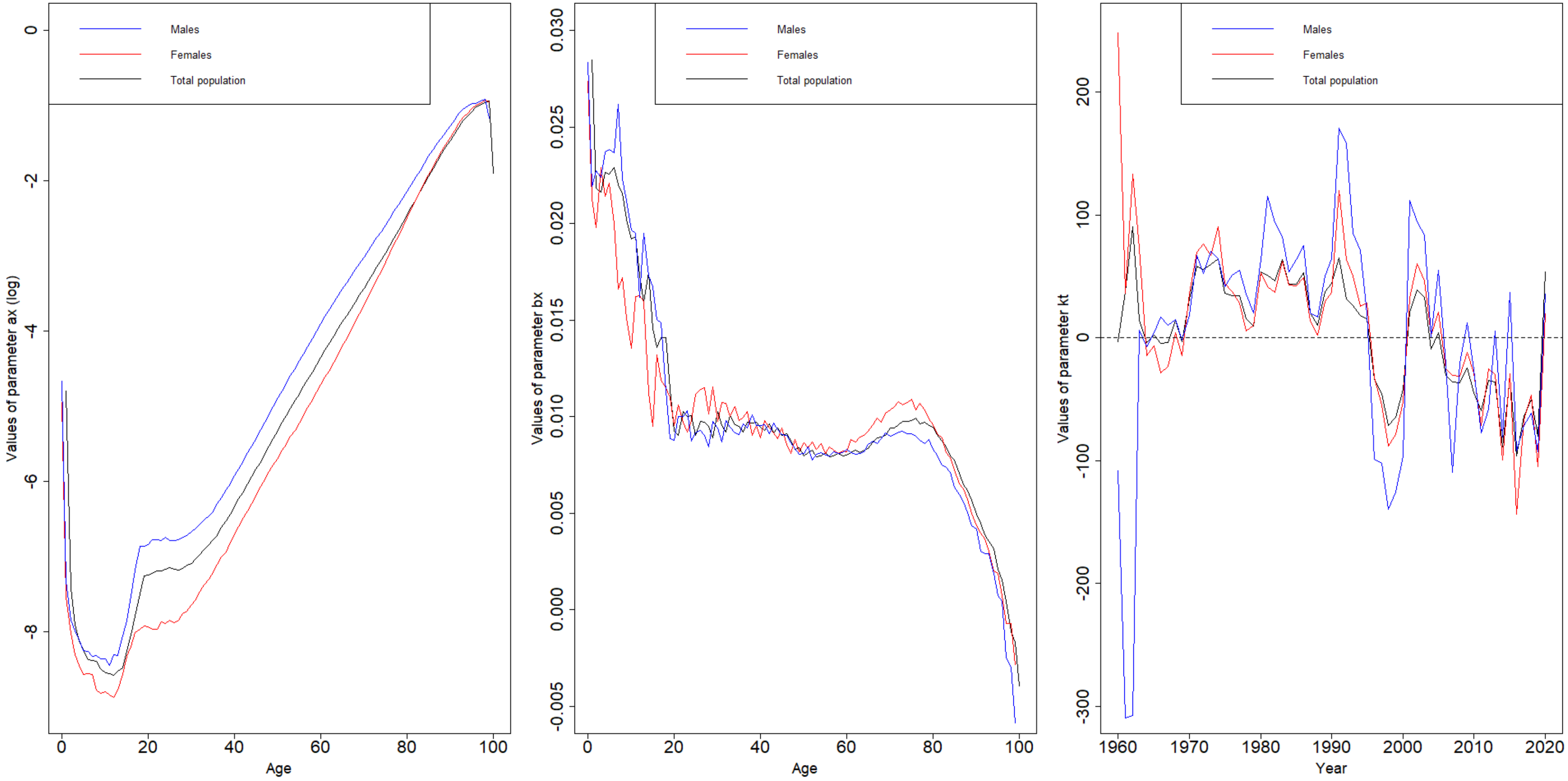

Figure 2. Parameters of Lee-Carter model – ax (left), bx (middle), kt (right) (own elaboration in package demography; (Hyndman, 2022))

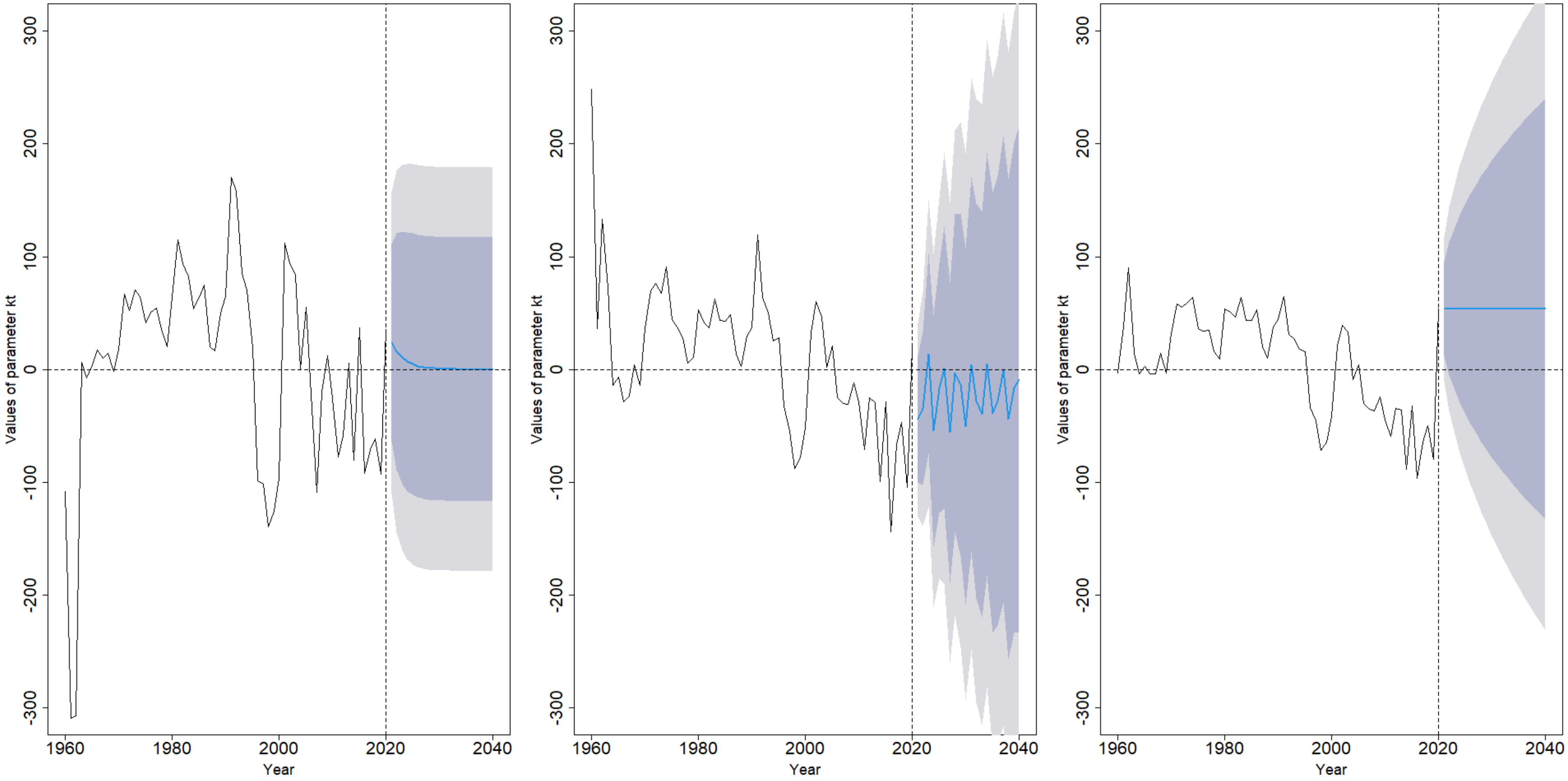

Figure 3. Prediction of vector k_t – males (left), females (middle), total population (right) (own elaboration in package `forecast`; (Hyndman, 2022))

3.2. Projection of Vector k_t

As k_t is only variable that varies across time, its future development is projected. We utilize and compare two approaches. First, a tool called `auto.arima` in R that automatically select one model based on Akaike criterion is used. However, the models seem not to be optimal. As can be seen in Figure 3, the parameter k_t tends to conditional zero mean in case of males, seems to be stationary (or autocorrelated) in case of females and is a constant for total population because there is probably not enough variability (not enough information) in the development of parameter k_t. Usual trend of kt parameter is decreasing – see e.g. Šimpach (2013) where all projections of k_t are straight decreasing line.

Because the results of the projection were not optimal, we tried another tool. The vector of k_t parameters was exported to another software and forecasted by smooth AIC weights. Figure 4 displays possible ARIMA models that were compared with each other. Then the average forecast was chosen by the algorithm incorporated in EViews software. As same as in previous case, the projection of k_t parameter for males tends to 0 value after year 2025. The development is more volatile and below zero in case of females, but less volatile than in previous case. This imply that the forecasting could be easier. Finally, the parameter k_t for total population seems better than previously, because it is not just a straight line. It declines (following the trend of females' population), but slightly consolidates in 2026 and onwards (following the trend of males' population). Those projections correspond to the expected decreasing trend of k_t development (see e.g. Šimpach, 2013).

3.3. Forecast of Mortality Rates

The results of forecasts are displayed in Figure 5. Solid line marks the results of the first approach when k_t was projected in software R (the best model was chosen). Dashed line shows the second approach when k_t is projected in another software and the weighted average ARIMA model is chosen. It can be seen that different projection of k_t parameter gives completely different results.

The forecasts done by original k_t parameter in R package forecast shows high volatility in case of female population which is then reflected in high volatility of total population. This is true for all selected ages. On the other hand, k_t projected by ARIMA models project smoother lines. Surprisingly the mortality of 0-years old is very high for male's population that affect also the forecast for total population. The most stable age groups are around middle age (20 to 60 years), when the intensity of mortality is not fluctuated as a result of either systematic or random influences.

Both forecasts are realistic in following aspects. The mortality is the highest in the infant age and then in the highest ages. Mortality of men is highest than of women in the Czech Republic that is due to the living conditions and lifestyle of men (more demanding work, riskier behavior, addictions etc.). The trend of mortality is decreasing (due to progress in health care, higher living standards etc.). The decrease of mortality rates is faster and the beginning of the forecast period and slows down later. There can be seen the impact of Covid-19 pandemic as and exogenous shock that increased the mortality in 2020. It can be expected

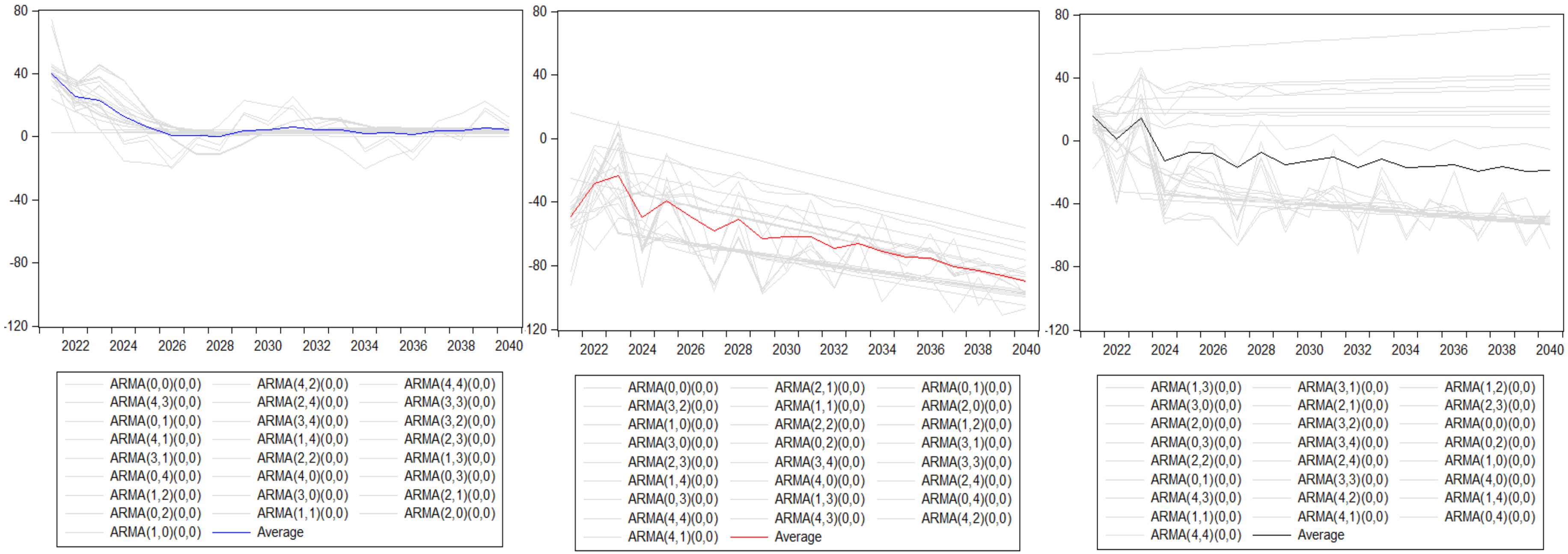

Figure 4. Prediction of vector k_t – forecast comparison graphs – males (left), females (middle), total population (right) (own elaboration in EViews)

Both types of k_t were utilized (and same a_x and b_x parameters) for mortality rates forecast.

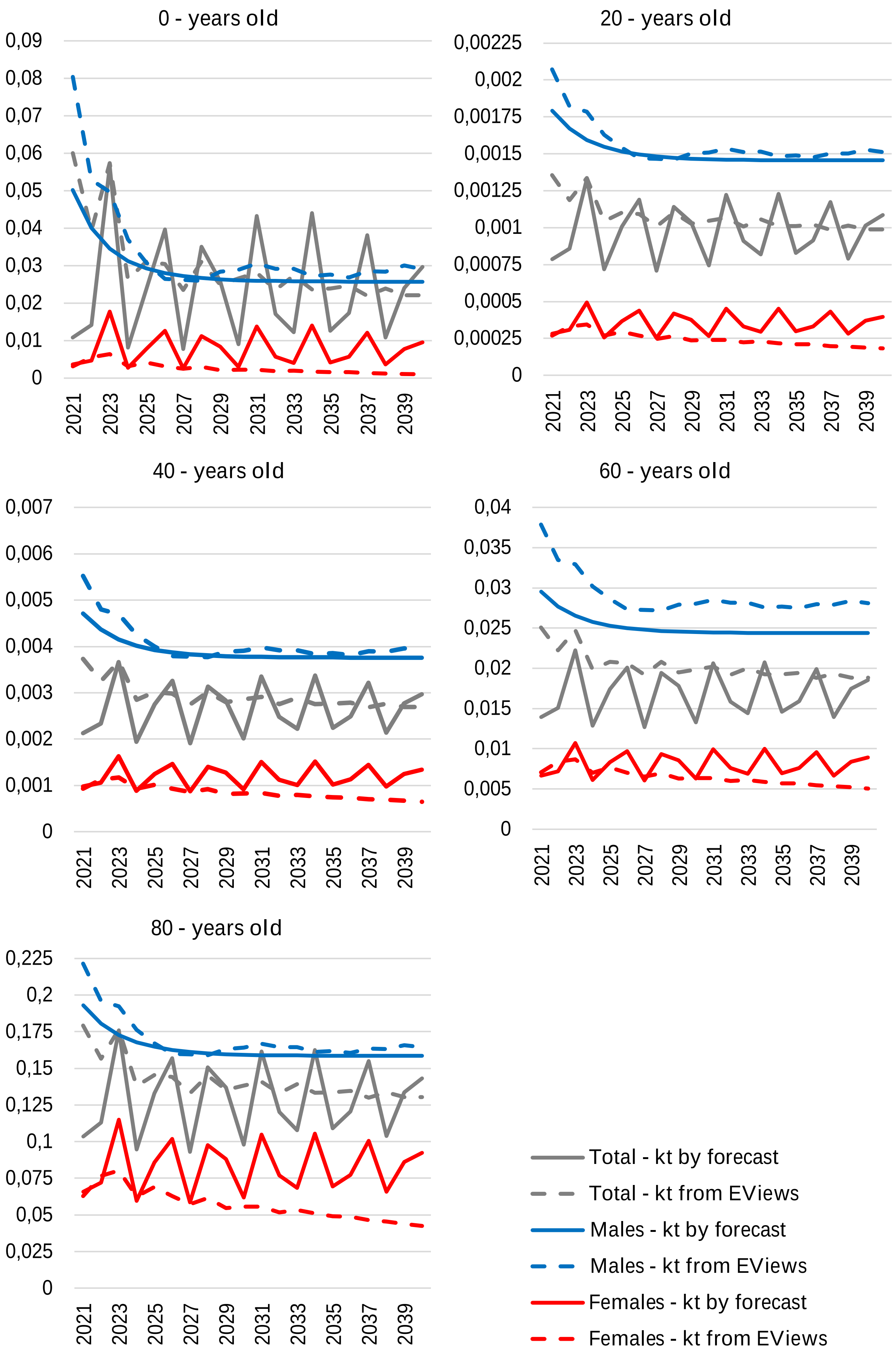

Figure 5. Forecasted mortality rates in selected ages for years 2021–2040 – comparison of two approaches

that also mortality rates in 2021 and 2022 would be affected, but the effect diminishes later. Šimpach and Šimpachová Pechrová (2021) found out that despite the Covid-19 pandemic, the mortality rates will still decrease in their forecast, and that adding data for year 2020 did not change much the decreasing trend that is very strong and rooted in the data since the 1990s when the increase of living standards, after the end of the communist era, caused that the mortality rates started to decrease more.

It can be concluded that the forecasts of mortality rates based on the k_t parameter projected in software EViews is smoother and more realistic.

4. Discussion

Our model expects decline of the mortality rates. However, it shall not be that sharp as before, because the times of speed reduction are already over (From 1980 till approximately 2000 in the Czech Republic). According to Kannisto et al. (1994), developed countries have made progress in reducing death rates even at the highest ages and the pace of this progress has accelerated over the course of the twentieth century. "In most developed countries outside Eastern Europe, average death rates at ages 80–99 have declined at a rate of 1 to 2 percent per year for females and 0.5 to 1.5 percent per year for males since the 1960s." (Kannisto et al., 1994).

We can compare our forecasts with study of Šimpach (2013). He applied Lee-Carter method on Czech population in years 1920–2012 and 1948–2012. He found out that logarithms of age-specific mortality rates in the Czech Republic of men have a visible tendency to decrease faster and with a higher intensity in the future (in particular in the lowest age groups) than is the case for Czech women." (Šimpach, 2013). This would indicate the expected fact that male and female life expectancy could converge. Our forecasts do not show this trend in the near future.

Šimpach (2013) choose to use ARIMA (1,1,0)c model to forecast the mortality rates of males and females and ARIMA(0,1,0)c for total population. We used ARIMA (1,1,1) for males and females and ARIMA (0,1,0)c for total population.

We found out that ARIMA model chosen by `forecast` package in R is not optimal as the projection of k_t and forecast of $ln\ m_{x,t}$ are too volatile. For this reason, we can suggest smoothing of the time series before its inclusion into the Lee-Carter model. In general, the higher the age group, the higher the variability in the data. Especially at very high ages can be observed significant deviations in mortality. The existence of an outlier can cause problems in the forecast. That is why it is important to smooth mortality curve for obtaining better results. Dotlačilová et al. (2014) suggested to use polynomial functions for levelling and for extrapolation of mortality curves at the advanced ages. One of the mostly used method is e.g. Kannisto elaborated by Kannisto et al. (1994) and Thatcher et al. (1998). Kanisto method was suggested e. g. by Šimpach (2013) or Šimpach and Dotlačilová (2012). "A better result would probably be obtained if the time series of age-specific death rates were matched to some of the existing smoothing models, but this could cause the loss of some additional information needed for principal component analysis." Šimpach and Dotlačilová (2012) concluded that the model based on smoothed data fits better the reality, because it refers to the expected development of the ln $m_{x,t}$ – it declines through the all age groups.

5. Conclusions

The aim of the paper was to compare two possible methods of projections of the vector k_t (indexes of intensity of mortality) in the stochastic Lee-Carter model and to forecast the age and sex-specific mortality rates in the Czech Republic for next 20 years. Lee-Carter model is based on Principal Component Analysis and is estimated by Singular Value Decomposition method. It consists of two age-dependent parameters a_x and b_x and time-dependent vector k_t. Projection of the model's vector k_t was done by two approaches. First, the best ARIMA model was chosen by package `forecast` in software R. Second, average ARIMA model based on the smoothed Akaike information criterions was calculated in EViews software. Both k_t were used for forecasting of the mortality rates for period 2021–2040. Forecasting was done in software R package `demography`.

We found out that parameters a_x and b_x had expected development, so Lee-Carter model works well when it is applied on Czech data. However, the k_t was not calculated according to the expectations. Regarding the projection of index k_t, the best ARIMA model chosen by package `forecast` in R did not project k_t realistically. External calculation in EViews chose average ARIMA model that created better k_t projections which followed expected decreasing trend. This is reflected in the results of the forecasts of the mortality rates $m_{x,t}$. The forecasts of mortality rates $m_{x,t}$ based on the vector k_t projected in software EViews is smoother and more realistic. We suggest external projection of vector k_t or improvement of `forecast` package in software R. Or some changes in the data (smoothing) shall be done to calculate and project realistic vector k_t and hence mortality rates directly in R.

Therefore, the challenge for future research is to forecast the future mortality rates based on the smoothed data of mortality rates in the Czech Republic and to calculate and project more realistic development of vector k_t.

Acknowledgments: 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.

Conflict of interest: none.

References

Booth, H., Maindonald, J., & Smith, L. (2002). Applying Lee-Carter under conditions of variable mortality decline. *Population Studies*, *56*(3), 325–336. https://doi.org/10.1080/00324720215935

Booth, H., Tickle, L., & Smith, L. (2005). Evaluation of the variants of the Lee-Carter method of forecasting mortality: a multi-country comparison. *New Zealand Population Review*, *31*(1), 13–34.

Box, G., & Jenkins, G. (1970). *Time Series Analysis: Forecasting and Control*. Holden-Day.

Danesi, I. L., Haberman, S., & Millossovich, P. (2015). Forecasting mortality in subpopulations using Lee–Carter type models: A comparison. *Insurance: Mathematics and Economics*, *62*, 151–161. https://doi.org/10.1016/j.insmatheco.2015.03.010

De Jong, P., Tickle, L., & Xu, J. (2020). A more meaningful parameterization of the Lee–Carter model. *Insurance: Mathematics and Economics*, *94*, 1–8. https://doi.org/10.1016/j.insmatheco.2020.05.009

Dotlačilová, P., Šimpach, O., & Langhamrová, J. (2014). The Use of Polynomial Functions for Modelling of Mortality at the Advanced Ages. In *32nd International Conference on Mathematical Methods in Economics (MME 2014)*, (pp. 174–179).

Eurostat (2022). *Database*. https://ec.europa.eu/eurostat/data/database

Fiala, T., Langhamrová, J., & Průša, L. (2011). Projection of the Human Capital of the Czech Republic and its Regions to 2050. *Demografie*, *53*(4), 304–319.

Kanisto, V., Lauritsen, J., Thatcher, A. R., & Vaupel, J. W. (1994). Reductions in Mortality at Advanced Ages: Several Decades of Evidence from 27 Countries. *Population and Development Review*, *20*(4), 793–810. https://doi.org/10.2307/2137662

Lee, R., & Carter, L. (1992). Modeling and Forecasting U. S. Mortality. *Journal of the American Statistical Association*, *87*(419), 659–671. https://doi.org/10.2307/2290201

Lee, R., & 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., & Gerland, P. (2013). Extending the Lee-Carter Method to Model the Rotation of Age Patterns of Mortality Decline for Long-Term Projections. *Demography*, *50*(6), 2037–2051. https://doi.org/10.1007/s13524-013-0232-2

Russolillo, M. (2017). Assessing Actuarial Projections Accuracy: Traditional vs. Experimental Strategy. *Open Journal of Statistics*, 7, 608–620. https://doi.org/10.4236/ojs.2017.74042

Šimpach, O. (2013). *Statistické metody v demografickém prognózování*. Disertační práce. Praha: Vysoká škola ekonomická v Praze. https://theses.cz/id/d8gh16/

Šimpach, O., Dotlačilová, P., & Langhamrová, J. (2014). Effect of the Length and Stability of the Time Series on the Results of Stochastic Mortality Projection: An application of the Lee-Carter model. In *International work-conference on time series (ITISE 2014)*, (pp. 1375–1386).

Šimpach, O., & Dotlačilová, P. (2016). Age-Specific Death Rates Smoothed by the Gompertz–Makeham Function and Their Application in Projections by Lee–Carter Model. In I. Rojas, & H. Pomares (Eds.), *Time Series Analysis and Forecasting. Contributions to Statistics*. Springer, Cham.

Šimpach, O., & Langhamrová, J. (2014). Stochastic Modelling of Age-specific Mortality Rates for Demographic Projections: Two Different Approaches. In *32nd International Conference on Mathematical Methods in Economics (MME 2014)*, (pp. 890–895).

Šimpach, O., & Šimpachová Pechrová, M. (2021). Implications of the SARS-Cov-2 Pandemic for Mortality Forecasting: Case Study for the Czech Republic and Spain. *Engineering Proceedings*, *5*(1), 58. https://doi.org/10.3390/engproc2021005058

Thatcher, A. R., Kannisto, V., & Vaupel, J. W. (1998). *The Force of Mortality at Ages 80 to 120*. Odense University Press.

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. https://doi.org/10.1007/978-3-642-05253-8_6