Effect of the Length and Stability of the Time Series on the Results of Stochastic Mortality Projection: An application of the Lee-Carter model

Ondrej Simpach*, Petra Dotlacilova, and Jitka Langhamrova

University of Economics in Prague, Faculty of Informatics and Statistics,
W. Churchill sq. 4, 130 67 Prague, Czech Republic
{ondrej.simpach,petra.dotlacilova,langhamj}@vse.cz
http://kdem.vse.cz/

Abstract. For a significant demographic projection is necessary to know the possible future development of mortality. The aim of this paper is to use stochastic modelling approach (Lee-Carter model) in the case of mortality of the Czech population. For the case of the Czech Republic we have available data from the database of the Czech Statistical Office for the period 1920 to 2012, but during the 2^{nd} World war there was an unstable development of mortality in most countries in the world. We will compare the two approaches for modelling between each other, one based on the complete length of the time series and the other based on shortened series from 1948 (restructuring of Czechoslovakia after the war and political regime change). Based on the results of these two approaches we compare the two key issues of time series forecasting, namely, the length and stability. Not always can be the stable development of time series the correct issue which ensure the significant and realistic prediction. In the case of mortality is also necessary to consider the unexpected changes.

Keywords: mortality, demographic projection, Lee-Carter model, length and stability, Czech population

1 Introduction

Mortality is an important component of population reproduction. Its level affects the length of life. When we analyse the development of mortality, it is important to know that the biggest changes of it come out at the highest ages (approximately 60 years and above), where the mortality has got the different character in comparison with its character at lower ages. This is not only caused by small numbers of deaths, but also by small numbers of living at the highest ages. It is also necessary to realize that these data are affected by systematic and random errors. So if we want to capture the most accurately mortality of oldest people

* Corresponding author, Department of Demography, office phone +420 224 09 5273 for ZI areal or +420 224 09 4315 for JM areal, mobile +420 737 665 461.

it is good idea to make minor adjustments. This is mainly related to smoothing of mortality and possibility of its extrapolation until the highest ages. For smoothing we can use several existing models. Nowadays, the estimates of the unknown parameters of these models is possible to obtain using the professional software (e.g. *DeRaS*, see Burcin, Hulikova Tesarkova, and Komanek [7]). Among the most famous are included Coale-Kisker model (see e.g. Boleslawski, Tabeau [4] or Gavrilov, Gavrilova [14])

$$m_x = e^{ax^2+bx+c}, \tag{1}$$

where m_x are age-specific death rates and x is age. The calculated model corresponds with an exponential quadratic function, where a, b and c are parameters. Next one the Thatcher model (Thatcher, Kanisto, and Vaupel [29])

$$\mu_x = \frac{z}{1+z} + \gamma, \tag{2}$$

where $z = \alpha e^{\beta x}$, α, β, γ are parameters of model, x is age (and μ_x is the intensity of mortality at exact age x). The other one is Kannisto model (Thatcher, Kanisto, and Vaupel [29]

$$\mu_x = \frac{e^{[\theta_0+\theta_1(x-80)]}}{1+e^{[\theta_0+\theta_1(x-80)]}}, \tag{3}$$

where θ_0 and θ_1 are parameters of model. This model is special case of the logistic function, where the LOGIT transformation of mortality rates is expressed like linear function of age. The oldest model (but still used) is the Gompertz-Makeham function (Gompertz [15], Makeham [23])

$$\mu_x = a + bc^x, \tag{4}$$

where a, b and c are estimated parameters. (more e.g. in Simpach [28]) All these models are suitable for the elimination of fluctuations in age-specific death rates and their subsequent extrapolation. The disadvantage of these models is that they can not be used for projection of future mortality and therefore neither for the calculation of demographic projections. Demographic projections of possible future evolution of population are essential information channel, which is used for providing of key information about the potential evolution of mortality, birth rates, immigration and emigration, or other demographic statistics. Each projection is based on the assumptions, which could but might not be occurred. Stochastic demographic projections are based on the main components (Lee, Carter [20]), explaining the trend, which is included in the development of time series of age-specific demographic rates. A major influence on results has the length of the time series (see e.g. Coale, Kisker [9], or comparing the multiple results of populations from study by Booth, Tickle, and Smith [5]). In this paper we focus on the evolution of the data of mortality in the Czech Republic. The length of the time series is sufficient for statistically significant projections, because the statistics are detailed recorded by statisticians in the annual period since 1920. The stability is, unfortunately, problematical in most populations in the world. When we consider the data from 1920, it will probably come out the influence of the global economic crisis since 1929, the 2^{nd} World war in 1939-1945 and the consequences of centrally planned economy in some countries, which worked from fifties under the communist regime. On the basis of the two Lee-Carter's models there will be compared two key characteristics of good projections - the length and the stability of the time series. The first model will use the data for the period 1920 to 2012 and the second one will be based on the database for the period 1948 to 2012. The first model is strong in the length of the time series, the second one in their stability. These models will be evaluated to each other and the final projections of age-specific death rates for the Czech population until 2050 will be compared with each other.

2 Materials and Methods

For purposes of mortality analysis in the Czech Republic we use the data from the Czech Statistical Office (CZSO) about the numbers of deaths x-year-old in total $D_{x,t}$ (without distinction of sex), and the exposure to risk $E_{x,t}$, which is estimated as the midyear population x-year-old in total (without distinction of sex). We use the annual data for the period 1920 to 2012. The age-specific death rates in population (e.g. Charpentier, Dutang [8] or Erbas et al. [11]) we calculate as

$$m_{x,t} = \frac{D_{x,t}}{E_{x,t}} \tag{5}$$

and these empirical rates we can see in the black and white "rainbow" chart (Hyndman [18]) in Fig. 1 (left). The lighter curves represent the past, the darker curves represent the present. The border of 1.0 means that from 1,000 living x-year-old persons die exactly 1,000 in year t. Therefore it is a maximal logical value, the higher values are due to the instability of the data at the highest ages. (Coale, Kisker [9]) In 1929 was the World economic crisis, which expanded later into the Czechoslovakia. Miserable conditions increased the numbers of deaths and the subsequent years showed the instability in the development of death rates. Between years 1939-1945 there took place the 2^{nd} World war, which caused again an entirely different development of mortality in the most populations throughout the world. In the post-war Czechoslovakia occurred the restructuring for the period 1945 to 1948, while between 17^{th} and 25^{th} February 1948 there became the Communist coup, the onset of nationalization and the beginning of hard totalitarian regime for the next 40 years. Dark period of economic, social oppression and imprisonment of people. The time series of the numbers of deaths, however, began to be stable and did not happen any significant changes. After the fall of the totalitarian regime during the Velvet Revolution from 17^{th} November to 29^{th} December 1989 became a period of the permanent increasing in the standard of living of the population and the death rates were improved more and more. The Czech Republic currently catching up with their statistics about the mortality development the advanced Western European countries. Prolongations

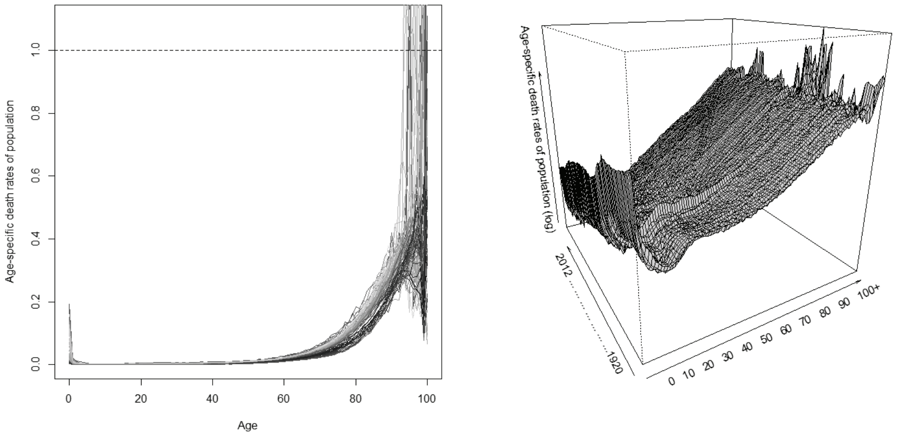

Fig. 1. Empirical data of age-specific death rates $m_{x,t}$ of the Czech population in total for the period 1920 to 2012 in "rainbow" chart (*left*) and the logarithms of these rates $\ln(m_{x,t})$ in perspective 3D chart (*right*).

of human life and also the population structure gets its typical regressive form. From the empirical data is evident, that until 1948 were the time series of death rates very unstable. (We can see this process detailed in the Fig. 1 (right), where the age-specific mortality rates are displayed in the logarithms in 3D perspective chart, for R code see Charpentier, Dutang [8].) It is known that the instability of the time series reduces their predictive capability. (Bell [2] or Gardner, McKenzie [13]) The history although has the lowest weight in the prediction model, but for the modelling of mortality, which is a long term process that has for each population its long-term trend, the history even with a little weight could be quite important (Booth, Tickle, and Smith [5]). Therefore we consider two models. One unabridged, based on the data matrix for the period 1920 to 2012 and the second one shortened, where the database will be only for the period 1948 to 2012. The logarithms of age-specific death rates can be decomposed (Lee, Carter [20] or Lee, Tuljapurkar [21]) as

$$\ln m_{x,t} = a_x + b_x k_t + \varepsilon_{x,t} \tag{6}$$

where $x = 0, 1, ..., \omega-1$, $t = 1, 2, ..., T$, a_x are the age-specific profiles independent of time, b_x are the additional age-specific components determine how much each age group changes when k_t changes, k_t are the time-varying parameters - the mortality indices and $\varepsilon_{x,t}$ is the error term. The age-specific death rates $m_{t,x}$ at age x and time t create $\omega - 1 \times T$ dimensional matrix

$$\mathbf{M} = \mathbf{A} + \mathbf{B}\mathbf{K}^\top + \mathbf{E}, \tag{7}$$

and the identification of Lee-Carter model is ensured by

$$\sum_{x=0}^{\omega-1} b_x = 1 \text{ and } \sum_{t=1}^{T} k_t = 0. \tag{8}$$

The estimation of b_x and k_t is based on Singular Value Decomposition (SVD) of matrix of age-specific death rates, presented by Bell, Monsell [3], Lee, Carter [20] or Lundstrom, Quist [22] and finally

$$a_x = \frac{\sum_{t=1}^{T} \ln m_{x,t}}{T} \tag{9}$$

is the simple arithmetic average of the logarithms of age-specific death rates. In the past there was often used the approach of linear extrapolation of the logarithms of age-specific death rates over time. The sufficient information was that the series of $\ln(m_{x,t})$ are approximately linear for each age x and decreasing over time. There was possible to conclude, that if we find a suitable intercept (b_x^0) and slope (b_x^1) of the linear regression, we can easily make a linear extrapolation to the future as

$$\ln m_{x,t} = b_x^0 + b_x^1 t + \varepsilon_{x,t} \tag{10}$$

where t is the time. In the Fig. 2 (left) there are shown the logarithms of age-specific death rates in black and white "rainbow" chart, while on the right side we can see their decreasing trends over time. There are the ages in which the development of time series is approximately linear and actually decreasing. But especially at the advanced ages, unfortunately, we can see the great variability, which cannot be explained by linear models only. This approach is simple, however, because the data in the conditions of most populations in the world are quite variable and the character of the time series for each logarithms of age-specific death rates are unstable, (Murphy [25]) the residues from these regressions would be strongly autocorrelated and the variance would not be explained.

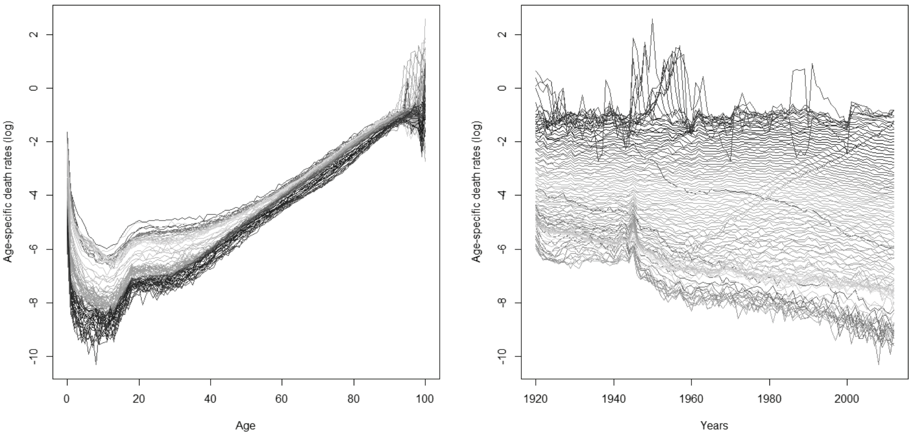

Fig. 2. Empirical data of logarithms of age-specific death rates $\ln(m_{x,t})$ of the Czech population in total for the period 1920 to 2012 in "rainbow" chart (*left*) and the development of these rates over time (*right*).

Therefore, we use the stochastic Lee-Carter model (6), where for the purposes of prediction the future age-specific death rates is necessary to forecast only the values of parameter k_t. This forecast is mostly calculated by $ARIMA(p, d, q)$ models (Box, Jenkins [6] or Melard, Pasteels [24]). The values of the parameters a_x and b_x are independent of time and the prediction using the Lee-Carter model is therefore purely extrapolative (Lee, Tuljapurkar [21]).

3 Results and Discussion

Using the SVD method implemented in the package "demography" (Hyndman, [18]), which is developed for RStudio (R Development Core Team, [27]), there were estimated the parameters $\hat{a}_x$ (age-specific profiles independent in time) and $\hat{b}_x$ (additional age-specific components determine how much each age group changes when k_t changes) for unabridged and shortened Lee-Carter model. We can see them in the Fig. 3, from which it is also clear the comparison between the different evolutions of these parameters. The age-specific profiles independent of time ($\hat{a}_x$) are lower in the shortened model, because in the considered period there were already the death rates in the population lower. (Post-war period, stability). Given that the length of the analysed time series is shorter, the variability of the estimated additional age-specific components ($\hat{b}_x$) is higher, especially at the advanced ages. This variability is later reflected in the balanced and predicted $\ln(m_{x,t})$, estimated by Lee-Carter model. The mortality indices $\hat{k}_t$

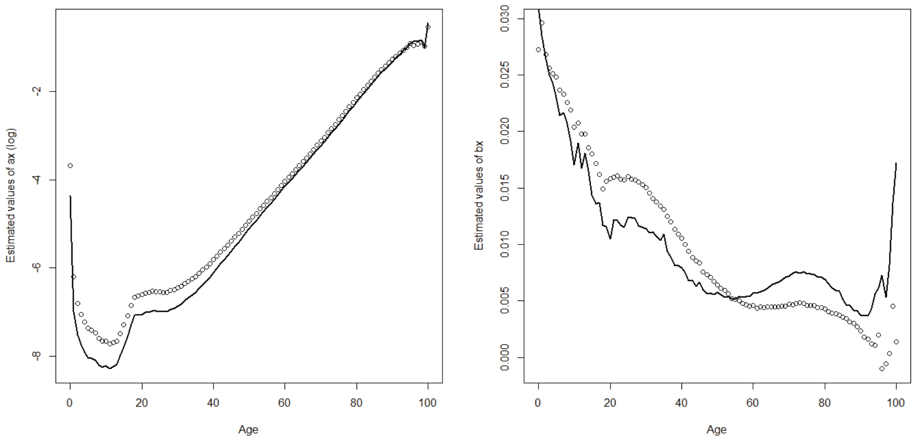

Fig. 3. Comparison of two Lee-Carter's models - The estimates of age-specific profiles independent in time (parameter $\hat{a}_x$, *left*) and additional age-specific components determine how much each age group changes when k_t changes (parameter $\hat{b}_x$, *right*). Black wheels represent the model based on data for the period 1912 to 2012, black line the model based on data for the period 1948 to 2012.

(the time-varying parameters) were estimated for the period 1920 to 2012 and 1948 to 2012. We can see these estimates in the Fig. 4. To these estimates there were calculated the predictions up to the year 2050 based on the methodological approach of *ARIMA*, (Box, Jenkins, [6]) and ran by "forecast" package in R. (Hyndman et al. [16], Hyndman, Shang [17]) Parameters of *ARIMA* models are written in Tab. 1. From these predictions with 95% confidence intervals, (which can be seen in Fig. 4 too) is clear, that the unabridged model provides slightly lower values of these estimates. (More information about the prediction models provides also e.g. Zhang et al. [30].) Confidence intervals are slightly wider at the shortened model. Note that the theoretical basis for these predictions in sophisticated systems and automatic prediction software prepared as well e.g. Bell [2], Bell, Monsell [3], Keyfitz [19] or Ord, Lowe [26]. Now we evaluate two Lee-Carter models on the basis of approach, which is presented by Charpentier, Dutang [8]. Using RStudio we will display the Pearson's residuals first for the unabridged and then for the shortened model. Each model will be evaluated on

Table 1. Estimated parameters of two *ARIMA* models

		AR	(s.e.)	Drift	(s.e.)
Unabridged model	ARIMA(1,1,0) with drift	-0.1545	0.1031	-2.2367	0.5023
Shortened model	ARIMA(0,1,0) with drift			-1.8108	0.5157

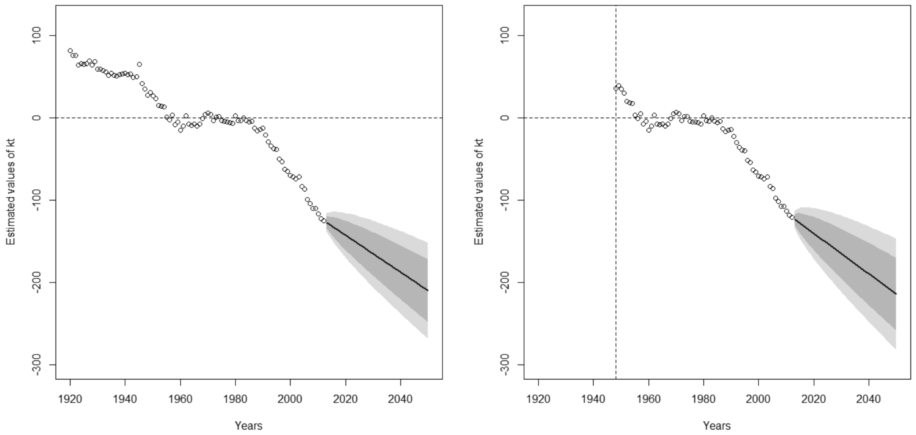

Fig. 4. Comparison of two Lee-Carter's models - The estimates of the time-varying parameters $\hat{k}_t$ - the mortality indices. On the *left* side is the model based on data for the period 1912 to 2012, on the *right* side is the model based on data for the period 1948 to 2012.

the basis of the residues by age x and of the residues at time t. The main requirement is that the unexplained system should not be seen in the estimated model. Most residues are concentrated around 0, the more variability is explained by the estimated model. The Pearson's residues for unabridged model for the period 1920 to 2012 are shown in the Fig. 5 (*top*), where residues by age x are on the left side and the residues at time t are on the right side. Given that this model also includes the World war period, it is understandable that the residues will be much more variable than in the case of shortened model. The residues of the shortened model for the period 1948 to 2012 are shown in the Fig. 5 (*bottom*). Stable period in both models is approximately from 1960 to the present. The most stable age groups are around middle and advanced ages. Based on the estimated parameters $\hat{a}_x$, $\hat{b}_x$ and $\hat{k}_t$ of two Lee-Carter's models we now fit and then

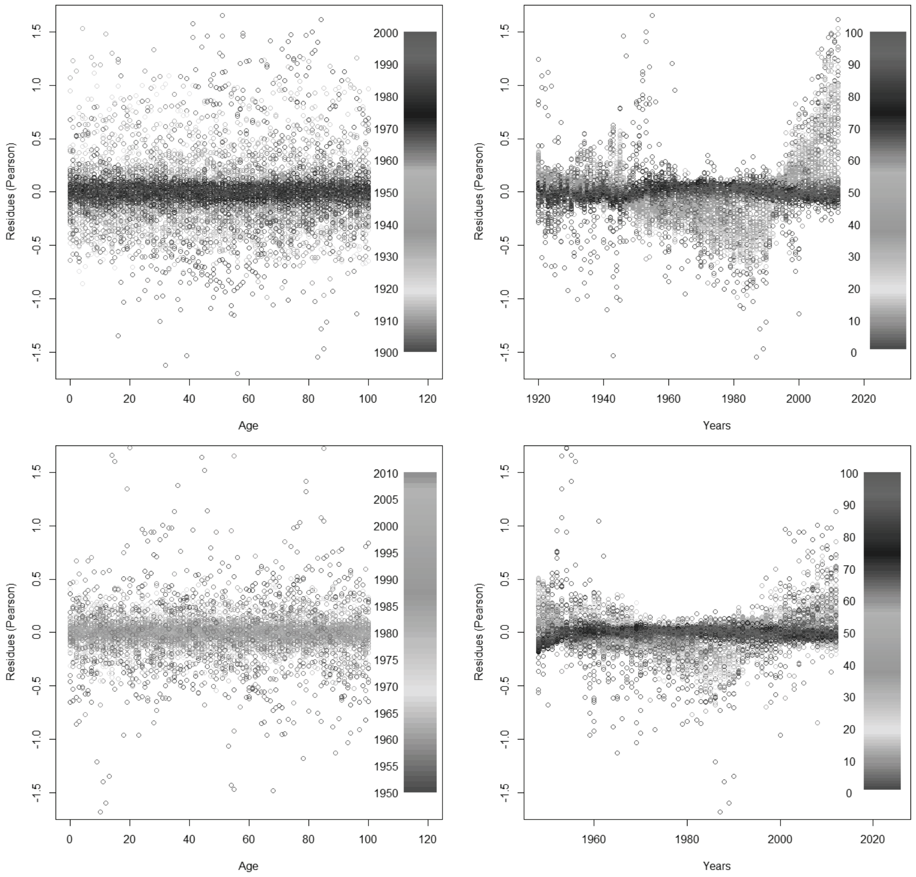

Fig. 5. Diagnostic control of the Lee-Carter model - Pearson's residues (data for the period 1920 to 2012 *top charts* and for the period 1948 to 2012 *bottom charts*).

estimate the future values of $\ln(m_{x,t})$ as

$$\ln m_{x,t} = \hat{a}_x + \hat{b}_x \hat{k}_t, \tag{11}$$

where for the unabridged model we use $x = 0, 1, ..., 100$ and $t = 1920, ..., 2050$ and for the shortened model we use x the same and $t = 1948, ..., 2050$. Fitted values with the attached estimates of $\ln(m_{x,t})$ based on the unabridged model are displayed in 3D perspective chart in the Fig. 6 (*top*). Below are displayed the fitted and then attached the estimated values based on the shortened model.

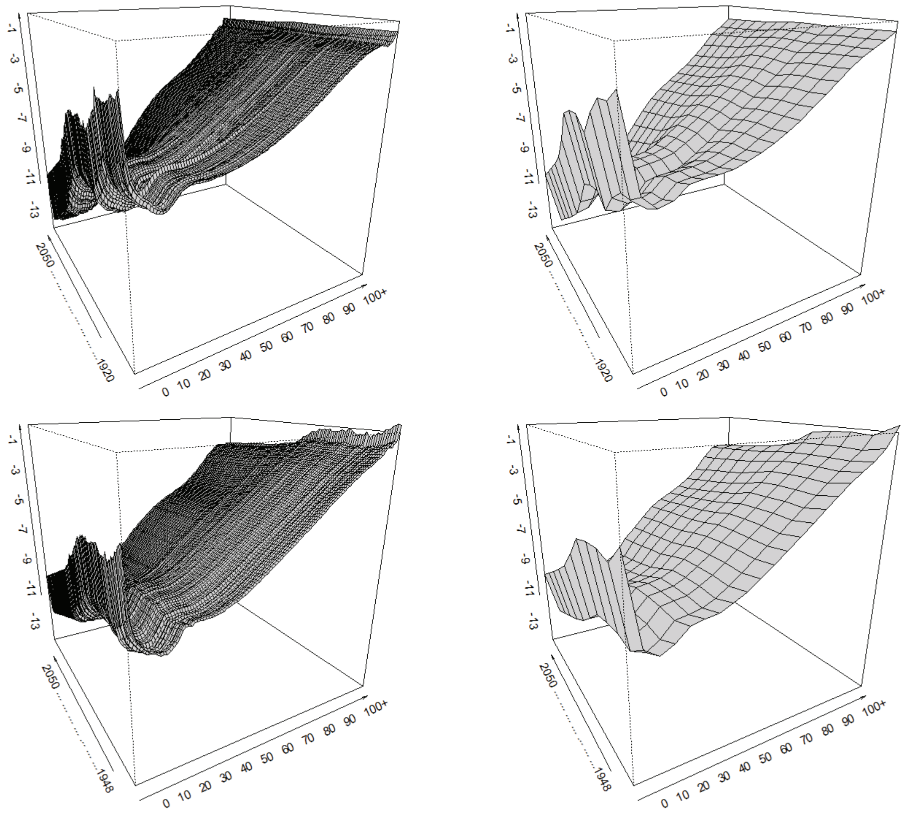

Fig. 6. Fitted values of logarithms of age-specific death rates $\ln(m_{x,t})$ of the Czech population in total for the period 1920 to 2012 with attached forecasts of these rates for the period 2013 to 2050 (*top left* in 1-year and *top right* in 5-years age interval based on full data matrix for the period 1920 to 2012) and the fitted values of these rates for the period 1948 to 2012 with attached forecasts for the period 2013 to 2050 (*bottom left* in 1-year and *bottom right* in 5-years age interval based on shortened data matrix for the period 1948 to 2012).

In order to visible mutual comparison of these values, all of the charts have the same scale. On the right side there are always displayed $\ln(m_{x,t})$ in 5-year intervals to make more clear the long-term trend. It is evident that the unabridged model provides more optimistic values of $\ln(m_{x,t})$ than the shortened model. In the case of the unabridged model there is predicted the more declining trend. We believe that the unabridged model provides the prediction very close to reality. The Czech Statistical Office (CZSO) periodically calculates the deterministic forecasts in 3 possible scenarios. Our prediction is very close to the middle scenario and its particular advantage is that it also takes into account the random component. The mortality is explained and predicted by its main components which is a much more sophisticated approach than expect a simple linear decline. Another study in the Czech Republic (Fiala, Langhamrova JI., and Prusa [12]) also predict the death rates, but their research is based on a deterministic approach and expert estimates only. We would like to enforce the suitability of the stochastic mortality approach in the Czech Republic.

4 Conclusion

The aim of this paper was to show the two key (and sometimes also contradictory) characteristics of the time series analysis (the length and the stability). These properties have been demonstrated in the case of the mortality development of the Czech population without distinction of sex. We examined whether provides the better predictions of future $\ln(m_{x,t})$ the unabridged or shortened Lee-Carter's model. The advantage of the shortened model should consist in the fact, that the time series are stable over time. However, the shorter development in the past showed that the predicted decline in age-specific death rates would be lower than in the case of unabridged model, which analysed more variable time series. Residues of both models seem to be favourable. On the basis of this test is no doubt about one of the models. Authors Arltova, Langhamrova JI., and Langhamrova JA. [1] analysed the development of life expectancy in the Czech Republic and using the Lee-Carter's model they calculated the predictions. The database was for the period 1920 to 2010 and the calculated predictions to a certain extent corresponded to medium variant of the estimated future development of life expectancy in the Czech Republic, which periodically calculates the Czech Statistical Office (CZSO). This may involve the one important conclusion. In the case of modelling and estimating the development of mortality may be in some populations a key issue for analysis the length of the analysed time series. Human life and its length develops for long time and therefore is preferable to analyse the population with the longest possible last development. This finding to a certain extent corresponds with the conclusions of Booth, Tickle, and Smith [5]. From our comparison we can claim that the unabridged model is optimal, because it refers to the expected future development of Czech population. In our future research we will focus on the backwards projection of mortality in the Czech Republic according to authors Dowd et al. [10]. We try to compute the mortality projection on the basis of empirical data from the pre-war period and we will compare the differences between the reality and the model. It will be interesting to examine how much the war period affect future predictions for the short and long term period.

Acknowledgments. This paper was supported by the Internal Grant Agency of University of Economics in Prague under the No. VSE IGS F4/24/2013 and also by the Czech Science Foundation project No. P402/12/G097 DYME - *Dynamic Models in Economics*.

References

1. Arltova, M., Langhamrova, JI., Langhamrova, JA.: Development of life expectancy in the Czech Republic in years 1920-2010 with an outlook to 2050. *Prague Economic Papers*, 22(1): pp. 125-143 (2013)
2. Bell, W.R.: Comparing and assessing time series methods for forecasting age-specific fertility and mortality rates. *Journal of Official Statistics* 13(3): pp. 279-303 (1997)
3. Bell, W.R., Monsell, B.: Using principal components in time series modelling and forecasting of age-specific mortality rates. *In: Proceedings of the American Statistical Association, Social Statistics Section*, pp. 154-159 (1991)
4. Boleslawski, L., Tabeau, E.: Comparing Theoretical Age Patterns of Mortality Beyond the Age of 80. *In: Tabeau, E., van den Berg Jeths, A., and Heathcote, Ch. (eds.): Forecasting Mortality in Developed Countries: Insights from a Statistical, Demographic and Epidemiological Perspective*, pp. 127–155 (2001)
5. Booth, H., Tickle, L., Smith, L.: Evaluation of the variants of the Lee-Carter method of forecasting mortality: a multi-country comparison. *New Zealand Population Review* 31(1): pp. 13-34 (2005)
6. Box, G.E.P., Jenkins, G.: *Time series analysis: Forecasting and control.* San Francisco, Holden-Day, 537 p. (1970)
7. Burcin, B., Hulikova Tesarkova, K., Komanek, D.: *DeRaS: software tool for modelling mortality intensities and life table construction.* Charles University in Prague, `http://deras.natur.cuni.cz` (2012)
8. Charpentier, A., Dutang, Ch.: *L'Actuariat avec R.* [working paper]. Decembre 2012. Paternite-Partage a lindentique 3.0 France de Creative Commons, 215 p. (2012)
9. Coale, Ansley J., Kisker, Ellen E.: Mortality Crossovers: Reality or Bad Data? *Population Studies.* Vol. 40, pp. 389–401 (1986)
10. Dowd, K., Cairns, Andrew J.G., Blake, D., Coughlan, G.D., Epstein, D., Khalaf-Allah, M.: Backtesting Stochastic Mortality Models: An Ex-Post Evaluation of Multi-Period-Ahead Density Forecasts. *CRIS Discussion Paper Series 2008.IV*, 42 p., Nottingham University (2008)
11. Erbas, B., Ullah, S., Hyndman, R.J., Scollo, M., Abramson, M.: Forecasts of COPD mortality in Australia: 2006-2025. *BMC Medical Research Methodology*, vol. 2012, pp. 12–17 (2012)
12. Fiala, T., Langhamrova, JI., Prusa, L.: Projection of the Human Capital of the Czech Republic and its Regions to 2050. *Demografie*, vol. 53, no. 4, pp. 304-320 (2011)
13. Gardner Jr. E.S., McKenzie, E.: Forecasting Trends in Time Series. *Management Science*, 31(10): pp. 1237–1246 (1985)

14. Gavrilov, L. A., Gavrilova, N. S.: Mortality measurement at advanced ages: a study of social security administration death master file. *North American actuarial journal* 15(3): pp. 432–447 (2011)
15. Gompertz, B.: On the Nature of the Function Expressive of the Law of Human Mortality, and on a New Mode of Determining the Value of Life Contingencies. *Philosophical Transactions of the Royal Society of London* 115(1825): pp. 513-585
16. Hyndman, R.J., Koehler, A.B., Snyder, R.D., Grose, S. A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting*, 18(3): pp. 439-454 (2002)
17. Hyndman, R.J., Shang, Han Lin: Forecasting functional time series. *Journal of the Korean Statistical Society*, 38(3): pp. 199–221 (With discussion) (2009)
18. Hyndman, R.J.: *demography: Forecasting mortality, fertility, migration and population data*. R package v. 1.16, `http://robjhyndman.com/software/demography/` (2012)
19. Keyfitz, N. Experiments in the projection of mortality. *Canadian Studies in Population*, 18(2): pp. 1-17 (1991)
20. Lee, R.D., Carter, L.R.: Modeling and forecasting U.S. mortality. *Journal of the American Statistical Association*, vol. 87, pp. 659-675 (1992)
21. Lee R.D., Tuljapurkar, S.: Stochastic population forecasts for the United States: beyond high, medium, and low. *Journal of the American Statistical Association*, vol. 89, pp. 1175-1189 (1994)
22. Lundstrom, H., Qvist, J.: Mortality Forecasting and Trend Shifts: An Application of the Lee-Carter Model to Swedish Mortality Data. *International Statistical Review / Revue Internationale de Statistique*, 72(1): pp. 37–50 (2004)
23. Makeham, W. M.: On the Law of Mortality and the Construction of Annuity Tables. *The Assurance Magazine, and Journal of the Institute of Actuaries* 8(1860): pp. 301-310
24. Melard, G., Pasteels, J.M.: Automatic ARIMA Modeling Including Intervention, Using Time Series Expert Software. *International Journal of Forecasting*, vol. 16, pp. 497–508 (2000)
25. Murphy, M. J.: The prospect of mortality: England and Wales and the United States of America, 1962-1989. *British Actuarial Journal*, 1(2): pp. 331-350 (1995)
26. Ord, K., Lowe, S.: Automatic Forecasting. *The American Statistician*, 50(1): pp. 88–94 (1996)
27. R Development Core Team: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. `http://www.R-project.org/` (2008)
28. Simpach, O.: Faster convergence for estimates of parameters of Gompertz-Makeham function using available methods in solver MS Excel 2010. *In: Proceedings of 30th International Conference on Mathematical Methods in Economics, Part II.*, pp. 870–874 (2012)
29. Thatcher, R. A., Kanisto, V., and Vaupel, J. W.: *The Force of Mortality at Ages 80 to 120*. Odense University Press (1998)
30. Zhang, J., Xanthopoulos, P., Tomaino, V., and Pardalos Panos, M.: Minimum Prediction Error Models and Causal Relations between Multiple Time Series. *In: Wiley Encyclopedia of Operations Research and Management Science*, vol. 5, pp. 3271–3285, John Wiley & Sons Inc. (2011)