The Impact of Smoothing of Age-specific Death Rates by the Gompertz-Makeham Function on the Results of Stochastic Projections by Lee-Carter Model

Ondrej Simpach* and Petra Dotlacilova

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

Abstract. Variability and stability are two important issues in time series modelling and forecasting. The aim of this paper is to use stochastic modelling approach (Lee-Carter model) for the case of mortality of the Czech population. For the case of the Czech Republic we have empirical data from the Czech Statistical Office (CZSO) database for the period from 1920 to 2012. We compare two approaches for modelling between each other, one is based on the empirical time series of age-specific death rates and the other is based on smoothed time series by the Gompertz-Makeham function, which is currently the most frequently used function for smoothing of mortality curves at the highest ages. Based on the results of these two approaches we compare two mentioned issues of time series forecasting - variability and stability. Sometimes stable development of time series can be the correct issue which ensure significant and realistic prediction, sometimes not. In the case of mortality is necessary to consider both unexpected or stochastic changes and long-term stable deterministic trend.

Keywords: Gompertz-Makeham function, Lee-Carter model, time series of mortality, demographic projection, the Czech population

1 Introduction

Mortality is an important component of population's reproduction and its development always has been very interesting topic (not only for demographers). The trend of mortality is one of the most important indicators of standard of living. If mortality is going to be better, then people live longer. The reason for improvement in mortality and also for increasing in life expectancy could be better health care. The second one is greater interest in healthy life style. On

* Corresponding author, Department of Statistics and Probability, office phone for Prague 3 - Zizkov areal +420 224 09 5273 or +420 224 09 4315 for Prague 4 - Jizni Mesto areal, mobile +420 737 665 461.

the other hand the increase in values of life expectancy means population ageing (Lundstrom, Quist [19]). More and more people live to the highest ages, so it is very important to have the best imagination about mortality trend at these ages. In the previous years it was not so important, because only a few people live to the highest ages. It is also important to say that this data is unreliable (Coale, Kisker [8]). Therefore it is necessary to use some of the existing models used for smoothing of age-specific death rates at the highest ages.

The level of mortality also affects the length of life. When we analyse the development of mortality, it is important to know that the biggest changes of mortality come out at the highest ages (approximately 60 years and above), where mortality has got different character in comparison with lower ages. This is not only caused by small numbers of deaths ($D_{x,t}$), but also by small numbers of living at the highest ages ($E_{x,t}$). 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 it is good idea to make minor adjustments. This is mainly related to smoothing of mortality curve and possibility of its extrapolation until the highest ages. For smoothing we can use several existing models. The oldest one (but still very often used) is the Gompertz-Makeham function (Gompertz [13], Makeham [20]). It is suitable for the elimination of fluctuations in age-specific death rates and then for their subsequent extrapolation until the highest ages. The disadvantage is that it can not be used for projection of future mortality and therefore neither for the calculation of demographic projections (Arltova [1] or Simpach, Pechrova [25]).

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 (Simpach [23]). 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 [17]), explaining 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 [8], or comparing the multiple results of populations from study by Booth, Tickle, and Smith [4]).

In this paper we focus on the evolution of data about mortality in the Czech Republic, provided by the Czech Statistical Office (CZSO). The length of the time series is sufficient for statistically significant projections, but the empirical data contain high variability in the highest ages. We use two approaches for our analysis (see also Simpach, Dotlacilova, Langhamrova [26]). The first one uses the empirical data for the period from 1920 to 2012 and the second one uses smoothed values of age-specific death rates by the Gompertz-Makeham function. The first model is strong in the length and unexpected variability of the time series, the second one in stability and absence of unexpected stochastic changes. These models will be evaluated to each other and 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 about mortality from the Czech Statistical Office (CZSO): numbers of deaths at complete age x-year-old $D_{x,t}$ (for males and females separately), and the exposure to risk $E_{x,t}$, which is estimated as the mid-year population at age x (for males and females separately). We use the annual data for the reporting period from 1920 to 2012. The age-specific death rates (see e.g. Erbas et al. [9]) we calculate as

$$m_{x,t} = \frac{D_{x,t}}{E_{x,t}} \tag{1}$$

and these empirical rates (in logarithms $\ln(m_{x,t})$) we can see in 3D perspective charts (Charpentier, Dutang [7], or Hyndman [16]) in the Fig. 1 (top left for males, top right for females).

One of the most famous model used for smoothing of age-specific death rates is the Gompertz-Makeham function, which can be written as

$$\mu_x = a + bc^{x+0.5}, \tag{2}$$

where a, b and c are estimated parameters (more e.g. in Gavrilov, Gavrilova [12] or Simpach [24]). We estimated these parameters in *DeRaS* software (see Burcin, Hulikova Tesarkova, Komanek [6]) and calculated values of smoothed age-specific death rates for males (bottom left) and females (bottom right) are shown in the Fig. 1 as well.

It is known that the instability of the time series reduces their predictive capability (Bell [2] or Gardner, McKenzie [11]). The history although has the lowest weight in the prediction model. But for the modelling of mortality, which is a long term process that it has for each population its long-term trend, the history even with a little weight could be quite important (Booth, Tickle, and Smith [4]). It will be interesting our idea to consider these two data matrices, (one empirical and the other one smoothed), for calculation of mortality forecast up to the year 2050. The logarithms of age-specific death rates can be decomposed (Lee, Carter [17] or Lee, Tuljapurkar [18]) as

$$\ln m_{x,t} = a_x + b_x k_t + \varepsilon_{x,t} \tag{3}$$

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 which determine how much each age group changes when k_t changes, k_t are the time-varying parameters - mortality indices and $\varepsilon_{x,t}$ is the error term. The age-specific death rates $m_{x,t}$ at age x and year t create $(\omega - 1) \times T$ dimensional matrix

$$\mathbf{M} = \mathbf{A} + \mathbf{B}\mathbf{K}^{\top} + \mathbf{E}, \tag{4}$$

and the identification of Lee-Carter model is ensured by

$$\sum_{x=0}^{\omega-1} b_x = 1 \text{ and } \sum_{t=1}^{T} k_t = 0. \tag{5}$$

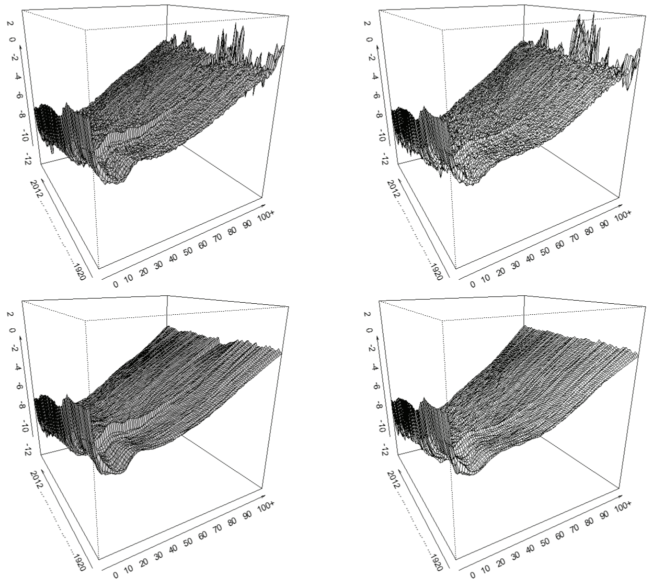

Fig. 1. Empirical data of age-specific death rates in logarithms $\ln(m_{x,t})$ of the Czech males *(top left)* and females *(top right)* for the period from 1920 to 2012. *Bottom left* are these rates for males (respectively *bottom right* for females) smoothed by the Gompertz-Makeham function. *Source: Data CZSO, authors' calculations.*

The estimation of parameters b_x and k_t is based on the Singular Value Decomposition (SVD) of matrix of age-specific death rates, presented by Bell, Monsell [3], Lee, Carter [17] or Lundstrom, Quist [19] and finally

$$a_x = \frac{\sum_{t=1}^{T} \ln m_{x,t}}{T} \tag{6}$$

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 in each age x and also 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

$$\ln m_{x,t} = b_x^0 + b_x^1 t + \varepsilon_{x,t} \tag{7}$$

where t is time. In the Fig. 2 (top left) shown the logarithms of age-specific death rates for males, (respectively top right for females), in the black and white "rainbow" chart over time, while on the bottom charts are represented male's and female's logarithms of age-specific death-rates smoothed by the Gompertz-Makeham function. There are the ages at which the development of time series

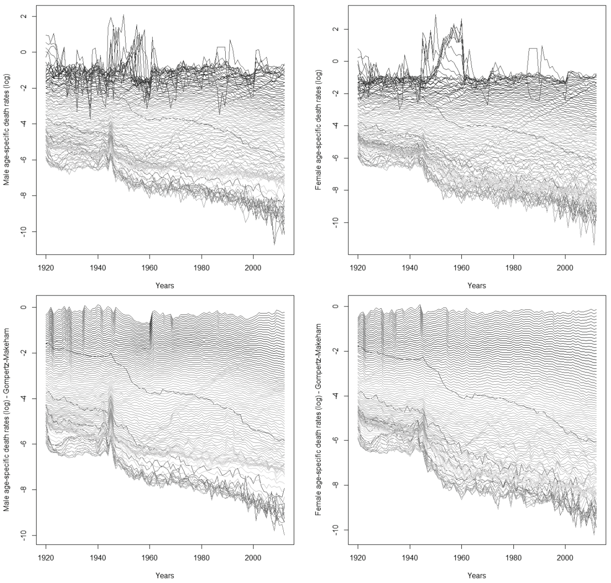

Fig. 2. Empirical data of logarithms of age-specific death rates $\ln(m_{x,t})$ of the Czech males *(top left)* and females *(top right)* over time and the development of these rates smoothed by the Gompertz-Makeham function for males *(bottom left)*, females *(bottom right). Source: Data CZSO, author's calculations, (based on Hyndman [16]).*

is approximately linear and actually decreasing. But especially at the advanced ages in the case of empirical data (bottom charts), we can see greater variability, which cannot be explained by linear models only.

Because we want to explain the major components of mortality of the Czech population, we use stochastic Lee-Carter model (3), where for the purposes of prediction of the future age-specific death rates it is necessary to forecast only the values of parameter k_t. This forecast is mostly calculated by $ARIMA(p, d, q)$ models (Box, Jenkins [5] or Melard, Pasteels [21]). 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 [18]).

3 Results and Discussion

Using the SVD method implemented in the package "demography" (Hyndman, [16]), which is developed for RStudio (R Development Core Team, [22]), we estimated the parameters $\hat{a}_x$ (age-specific profiles independent of time) and $\hat{b}_x$ (additional age-specific components determine how much each age group changes when k_t changes) for both Lee-Carter's model. We can see them in the Fig. 3, from which it is also clear the comparison between the different evolutions of these parameters, depending on the input variability.

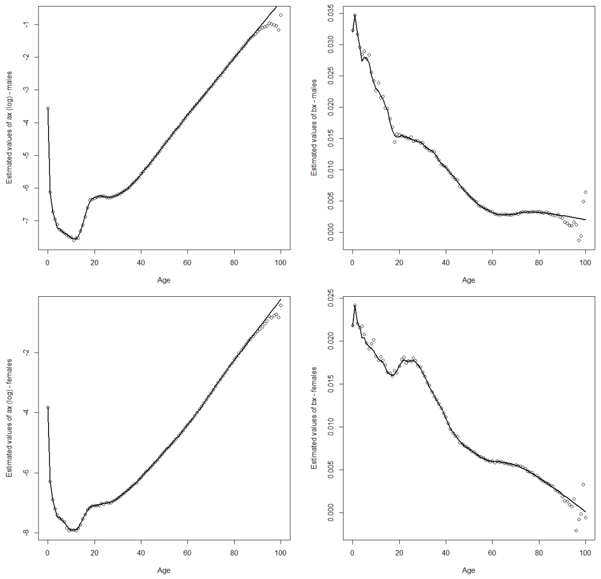

Fig. 3. Comparison of two Lee-Carter's models - The estimates of parameter $\hat{a}_x$ *(left)* and parameter $\hat{b}_x$ *(right)*. Black wheels represent model based on the empirical data matrix, black line represents model based on smoothed data matrix. *Top* charts are for males, *bottom* for females. *Source: authors' calculations, (based on Hyndman [16]).*

The mortality indices $\hat{k}_t$ (the time-varying parameters) were estimated for both models (empirical and smoothed) and it was found that the results are almost identical. This is due to the fact that these indices are also almost independent on the input variability of age-specific death rates. We can see these estimates in the Fig. 4. For these estimates we calculated the predictions up to the year 2050 based on the methodological approach of *ARIMA*, (Box, Jenkins, [5]) and ran by "forecast" package in R (Hyndman et al. [14], Hyndman, Shang [15]). Results are four *ARIMA*(1,1,0) models with drifts (see Table 1). Parameters AR(1) signed by † are equal to zero at the 5% significance level. From these predictions with 95% confidence intervals, (which can be seen in the Fig. 4 too) it is clear, that models for females provides slightly lower values of these estimates.

Now we evaluate both Lee-Carter's models on the basis of approach, which is presented by Charpentier, Dutang [7]. Using the RStudio we display Pearson's residuals firstly for the empirical males' model, secondly for the smoothed males' model, thirdly for the empirical females' model and lastly for the females' smoothed model. Each model will be evaluated on the basis of the residues by age x (left side of the page) and of the residues at time t (right side of the page). Most residues are concentrated around 0, higher variability is explained by the estimated model. The Pearson's residues for empirical and for smoothed models (for males and for females) are shown in the Fig. 5, the top four charts are for male's population, bottom four charts for female's population. (See labels at y axes).

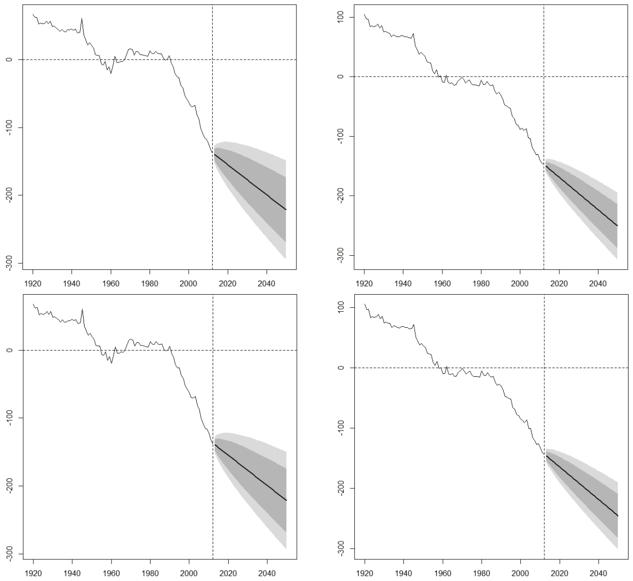

Fig. 4. Comparison of both Lee-Carter's models - The estimates of the time-varying parameters $\hat{k}_t$ - mortality indices. *Top* charts represent model based on the empirical data matrix, *bottom* charts represent model which is based on smoothed data matrix. Males are on the *left side*, females on the *right side*. *Source: authors' calculations, (based on Hyndman [16]).*

Table 1. Estimated parameters of four *ARIMA* models.

		AR	(s.e.)	Drift	(s.e.)
Empirical model - males	ARIMA(1,1,0) with drift	-0.0546†	0.1043	-2.2153	0.6313
Smoothed model - males	ARIMA(1,1,0) with drift	-0.0835†	0.1042	-2.2210	0.6184
Empirical model - females	ARIMA(1,1,0) with drift	-0.2182	0.1020	-2.7256	0.4844
Smoothed model - females	ARIMA(1,1,0) with drift	-0.2321	0.1019	-2.6888	0.4776

Based on the estimated parameters $\hat{a}_x$, $\hat{b}_x$ and $\hat{k}_t$ of both Lee-Carter's models for males and for females we now fit and then we estimate the future values of $\ln(m_{x,t})$ as

$$\ln m_{x,t} = \hat{a}_x + \hat{b}_x \hat{k}_t, \tag{8}$$

where $x = 0, 1, ..., 100$ and $t = 1920, ..., 2050$.

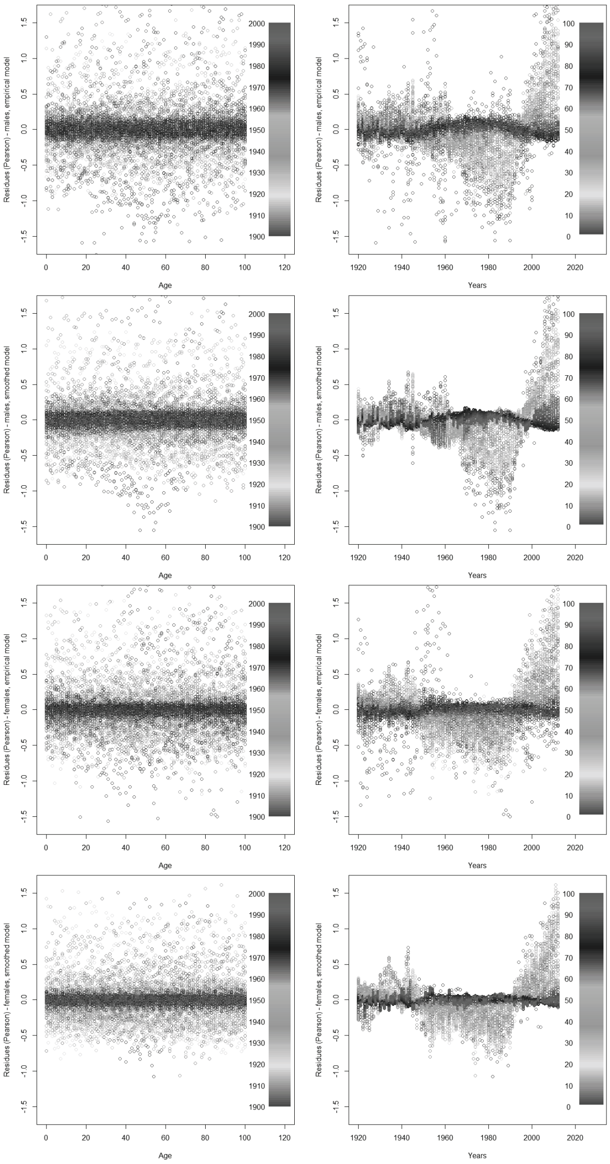

Fig. 5. Diagnostic control of four Lee-Carter's models - Pearson's residues. Empirical values for males, smoothed values for males, empirical values for females and smoothed values for females, see labels at y axes. *Source: author's calculations.*

Obtained values with the attached estimates of $\ln(m_{x,t})$ based on the empirical model (empirical data matrix) are displayed in 3D perspective chart in the Fig. 6 (*top*). *Below* are displayed obtained and then attached estimated values based on the smoothed model (data matrix smoothed by the Gompertz-Makeham function). In order to visible mutual comparison of these values, all charts have the same scale. On the *left side* there are always displayed male's $\ln(m_{x,t})$ and on the *right side* female's rates. It is evident that the empirical model provides more variable values of $\ln(m_{x,t})$ than the smoothed model, especially at the highest ages (60 years and older).

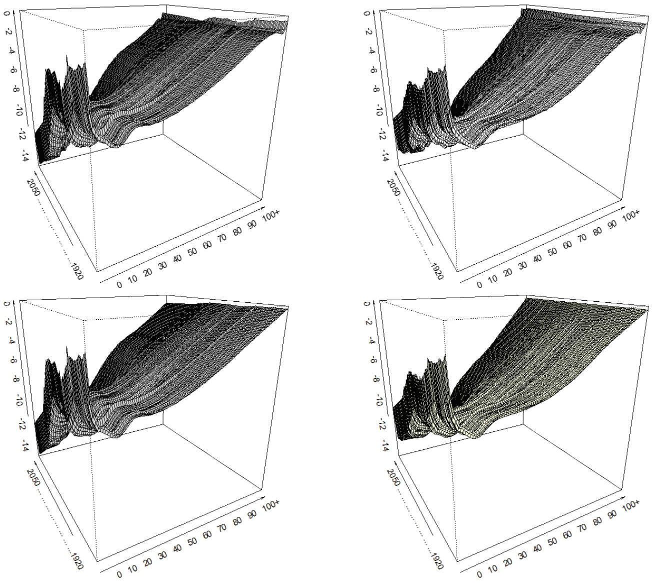

Fig. 6. Obtained values of logarithms of age-specific death rates $\ln(m_{x,t})$ for the Czech males *(left charts)* and for females *(right charts)* for the period from 1920 to 2012 with attached forecasts of these rates for the period from 2013 to 2050. *Top charts* were constructed by model based on empirical data, *bottom charts* by model based on smoothed data by the Gompertz-Makeham function. *Source: authors' calculations.*

We believe that model which is based on smoothed data by the Gompertz-Makeham function provides the prediction close to reality. Mortality is explained and predicted by its main components which is much more sophisticated approach than expect a simple linear decline (see model (7) and Fig. 2, where the risk is that the trend will not be sufficiently explained by model and in residues remain the unexplained system). Another study in the Czech Republic (Fiala, Langhamrova, Prusa [10]) also predicts death rates, but their research is based on deterministic approach and expert estimates only. We would like to enforce the suitability of the stochastic mortality approach in the Czech Republic.

4 Conclusion

In our paper we examined whether the Lee-Carter's model provides better predictions of future $\ln(m_{x,t})$, which is based on the empirical data matrix or on smoothed data matrix obtained by the Gompertz-Makeham function (which is currently the most famous one for modelling and extrapolating of mortality curves). The advantage of empirical model was, that we analysed data without any modifications. Residues of both models seem to be favourable. On the basis of this test is no doubt about one of the used models. But if we look at our results in the Fig. 6, we can see, that the age-specific death rates decline through the all age groups in the smoothed model only, which is related to the law of mortality (Gompertz [13]). This may involve the one important conclusion. From our comparison we can claim that the model based on smoothed data fits better the reality, because it refers to the expected future development of the Czech population. In our future research we would like to calculate the impact of our conclusions, mainly on the evolution of life expectancy and together the future number of inhabitants in the appropriate age groups according to this scenario.

Acknowledgments. This paper was supported by the Czech Science Foundation project No. P402/12/G097 DYME - *Dynamic Models in Economics*.

References

1. Arltova, M.: *Stochasticke metody modelovni a predpovidani demografickych procesu [habilitation thesis].* Prague: University of Economics Prague, 131 p. (2011)
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. 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)
5. Box, G.E.P., Jenkins, G.: *Time series analysis: Forecasting and control.* San Francisco, Holden-Day, 537 p. (1970)
6. 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)
7. 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)
8. Coale, Ansley J., Kisker, Ellen E.: Mortality Crossovers: Reality or Bad Data? *Population Studies.* Vol. 40, pp. 389–401 (1986)
9. 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)
10. Fiala, T., Langhamrova, J., 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)

11. Gardner Jr. E.S., McKenzie, E.: Forecasting Trends in Time Series. *Management Science*, 31(10): pp. 1237–1246 (1985)
12. 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)
13. 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*, vol. 115, pp. 513-585 (1825)
14. 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)
15. Hyndman, R.J., Shang, Han Lin: Forecasting functional time series. *Journal of the Korean Statistical Society*, 38(3): pp. 199–221 (With discussion) (2009)
16. Hyndman, R.J.: *demography: Forecasting mortality, fertility, migration and population data*. R package v. 1.16, `http://robjhyndman.com/software/demography/` (2012)
17. Lee, R.D., Carter, L.R.: Modeling and forecasting U.S. mortality. *Journal of the American Statistical Association*, vol. 87, pp. 659-675 (1992)
18. 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)
19. 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)
20. 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 (1860)
21. 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)
22. 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)
23. Simpach, O.: Detection of Outlier Age-specific Mortality Rates by Principal Component Method in R Software: The Case of Visegrad Four Cluster. *In: International Days of Statistics and Economics.* Slany: Melandrium, pp. 1505-1515 (2014)
24. 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)
25. Simpach, O., Pechrova, M.: The Impact of Population Development on the Sustainability of the Rural Regions. *In: Agrarian perspectives XXIII. The Community-led Rural Development.* Prague: Czech University of Life Sciences Prague, pp. 129-136. (2014)
26. Simpach, O., Dotlacilova, P., Langhamrova, J.: 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: Proceedings ITISE 2014.* Granada: University of Granada, pp. 1375-1386. (2014)