Stochastic Modelling of Age-specific Mortality Rates for Demographic Projections: Two Different Approaches

Ondřej Šimpach[1], Jitka Langhamrová[2]

Abstract. Nowadays it is not enough to construct the demographic projections based on the deterministic models. Stochastic models are more computationally intensive, but they are more expressive and may take account of many more factors, including e.g. random component. For the purposes of demographic projections it is necessary to know the expected future development of mortality, fertility and migration. In this paper we will show two different approaches to modelling age-specific mortality rates, and from the obtained models we prepare a projection of these rates to 2050. Data cover the period 1920–2012. The first type of model is an individual random walk with drift. Male and female population will be analysed for the range of 0–100+ years, and each completed age will be analysed as the individual time series with an annual frequency. The second model will be Lee-Carter, which is currently often used. It is based on Principal components method, which can capture and explain the main factors of mortality. Based on this models there will be constructed the forecasts of age-specific death rates for the period from 2013 to 2050. The obtained forecasts will be confronted and according to their provided results and their computational complexity there will be determined the recommendations about their suitability or their further improvement.

Keywords: population projection, age-specific mortality rates, random walk, Lee-Carter model.

JEL Classification: C32, C55, J11
AMS Classification: 60G25

1 Introduction

For a significant demographic projection it is necessary to know the possible future development of mortality. Mortality is an important component of population reproduction. Its level affects the length of life (Šimpach, Pechrová [19]). When we analyse the development of mortality, it is important to know that it is changeable during the human life (Gavrilov, Gavrilova [10]). The biggest changes come out at the highest ages (approximately 60 years and above), where the mortality has the different character in comparison with its character at lower ages (Keyfitz [14] or Thatcher, Kanistö, and Vaupel [20]). This is not only caused by small numbers of deaths, but also by small number of living at the highest ages. It is also necessary to realize that these data are affected by systematic and random errors. 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 and possibility of its extrapolation until the highest ages. For smoothing we can use several existing models. Among the most famous are included Coale-Kisker model (see e.g. Boleslawski, Tabeau [3] or Gavrilov, Gavrilova [10]), Thatcher model or Kanisto model (see e.g. Thatcher, Kanisto, and Vaupel [20]), or the oldest one (but still used) is the Gompertz-Makeham function (see comparison study by Gavrilov, Gavrilova [10] and the application in Šimpach [18]). The disadvantage of these models is that *they cannot be used for projection of future mortality* and hence for the calculation of demographic projections. Demographic projections of possible future evolution of population are essential information channel, which is used for providing of important information about the potential evolution of mortality rates, birth rates, immigration and emigration, or other demographic statistics. Each projection is based on the assumptions, which could but might not occurr (Gardner, McKenzie [9]). Sophisticated stochastic demographic projections are based on the main components (Bell, Monsell [2] and Lee, Carter [15]), explaining the trend, which is included in the development of time series of age-specific demographic rates. The length of the time series has a major influence on results (see e.g. Coale, Kisker [7], or comparing the multiple results of populations from study by Booth, Tickle, and Smith [4]). In this article we focus on the evolution of the mortality in the Czech Republic. We use two approaches of stochastic modelling. The first one is the individual models of random walk with drift (see e.g. Bell [1]). We analyse the male and female population sepa-

[1] University of Economics in Prague, Faculty of Informatics and Statistics, Dept. of Demography and Dept. of Statistics and Probability, W. Churchill sq. 4, 130 67 Prague 3, Czech Republic, ondrej.simpach@vse.cz
[2] University of Economics in Prague, Faculty of Informatics and Statistics, Department of Demography, W. Churchill sq. 4, 130 67 Prague 3, Czech Republic, langhamj@vse.cz

rately for the range of 0–100+ years and the period from 1920 to 2012. We considered each completed age as the individual time series with an annual frequency of 92 observations. Based on 101 individual models for males and 101 for females respectively, we calculate the predictions up to the year 2050 (by Box, Jenkins [5] methodology). These forecasts are mainly influenced by trend, which is present in the evolution of time series. In addition to this approach we use the second one. Lee-Carter's model (Lee, Carter [15] or Lee, Tuljapurkar [16]) based on Principal components method can capture and explain the main factors of mortality. We estimate the parameters of that model for population of Czech males and females. We forecast the death rates up to the year 2050 using the main components of mortality. In the discussion we compare the obtained forecasts and according to their results and their computational complexity we determine the recommendations about their suitability or their further improvement. We determine pros and cons of these approaches in the conclusion.

2 Materials and Methods

For mortality analysis in the Czech Republic we use the data from the Czech Statistical Office (CZSO) about the number of deaths x-year-old $D_{x,t}^{M/F}$ (males and females separately), and the exposure to risk $E_{x,t}^{M/F}$, which is estimated as the midyear population x-year-old (males (M) and females (F) separately). We use the annual data from 1920 to 2012. The logarithms of age-specific death rates in population (e.g. Lee, Carter [15], Charpentier, Dutang [6] or Erbas et al. [8]) we calculate as

$$\ln\left(m_{x,t}^{M/F}\right) = \ln\left(\frac{D_{x,t}^{M/F}}{E_{x,t}^{M/F}}\right), \tag{1}$$

where $x = 0, 1, \ldots, \omega$, and $t = 1, 2, \ldots, T$. These empirical data can be seen in 3D perspective charts, (for R code see Charpentier, Dutang [6]) in Fig. 1, (where males are on the left side and females on the right side).

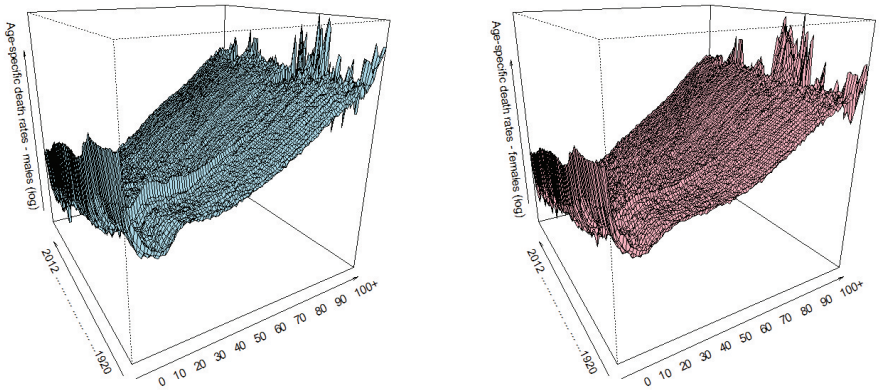

Figure 1 Age-specific death rates (in logarithms) of Czech males (left) and females (right) in 1920–2012. Source: CZSO, own construction and illustration in RStudio (R Development Core Team [17])

From the charts it is evident, that there were very unstable time series of death rates until 1948 (especially in the case of male population). It is well known that the instability of the time series reduces their predictive capability (Bell [1] or Gardner, McKenzie [9]). 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 [4]). In our paper we do not smooth the empirical data according to Gavrilov, Gavrilova [10] and rather use the random walk models with drift directly on raw data. (Lee-Carter model will be used on raw data as well). We denote Random walk model for male (M), female (F) population respectively as

$$\ln\left(m_{x,t}^{M/F}\right) = c_x^{M/F} + \ln\left(m_{x,t-1}^{M/F}\right) + \varepsilon_{x,t}^{M/F}, \tag{2}$$

where $c_x^{M/F}$ is constant and $\varepsilon_{x,t}^{M/F}$ is the error term with characteristics of white noise. This formula can be modified according to Box, Jenkins [5]. We get $ARIMA(0,1,0)$ model with constant for male, female population respectively as

$$\ln\left(m_{x,t}^{M/F}\right) = \ln\left(m_{x,0}^{M/F}\right) + c_x^{M/F} \cdot t_x^{M/F} + \sum_{t=1}^{T} \varepsilon_t^{M/F} \,, \tag{3}$$

where $c_x^{M/F}$ $t_x^{M/F}$ is the deterministic trend. This trend is linear decrease of death rates in time. Our second used approach, the logarithms of age-specific death rates, can be decomposed (see Lee, Carter [15] or Lee, Tuljapurkar [16]) as

$$\ln\left(m_{x,t}^{M/F}\right) = a_x^{M/F} + b_x^{M/F} \cdot k_t^{M/F} + \varepsilon_{x,t}^{M/F} \,, \tag{4}$$

where $x = 0, 1, \ldots, \omega$, $t = 1, 2, \ldots, T.$, $a_x^{M/F}$ are the age-specific profiles independent of time, $b_x^{M/F}$ are the additional age-specific components determine how much each age group changes when k_t changes and finally $k_t^{M/F}$ are the time-varying parameters - the mortality indices. The estimation is based on Singular Value Decomposition (SVD) of matrix of age-specific death rates, presented e.g. by Bell, Monsell [2] and Lee, Carter [15]. For predicting the future age-specific death rates it is necessary to forecast the values of parameter $k_t^{M/F}$ only. This forecast is mostly calculated by *ARIMA*(*p,d,q*) models with or without constant (Box, Jenkins [5]). 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 [16]).

3 Results and Discussion

Despite the fact that the random walk model with drift is relatively simple, it provides significant prediction, which involve the informations from the previous trend only. Neither additional expectations nor technological progress is implemented in results by random walk model (Coale, Kisker [7] or Hyndman et al. [11]). We estimated 101 male and 101 female individual models of random walk with drift (for particular series of age-specific death rates). Based on them we calculated the forecast of these rates up to the year 2050 in STATGRAPHICS Centurion XVI. From the output shown in the Fig. 2 it is clear that we can accept the general assumptions about the future declines of age-specific death rates, because the predicted time series for each year of life have a tendency to visibly decline in the future. In the case of the female population we can see only one irregularity at the age of 99 years. Due to the high variability of raw data (see Fig. 1) it is the predicted trend increasing at this age, but it is not so significant error.

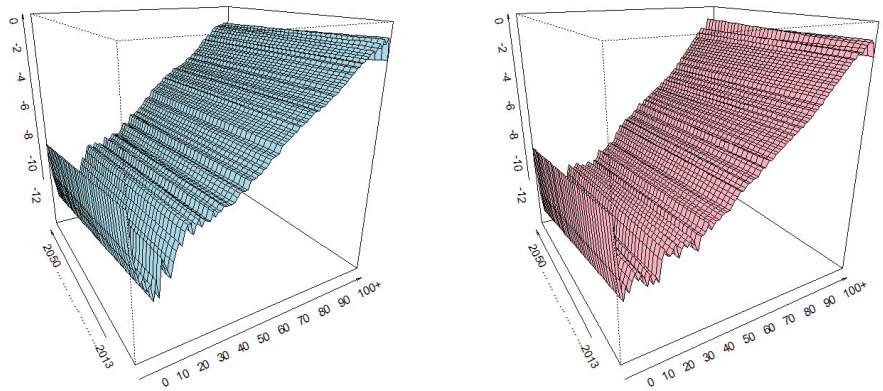

Figure 2 Forecast of age-specific death rates (in logarithms) for Czech males (left) and females (right) in period 2013–2050 by individual Random walk models with drift. Source: own calculation in STATGRAPHICS Centurion XVI, own construction and illustration in RStudio.

In the next step, using the SVD method implemented in the package "demography" (see Hyndman, [13]), which is developed for RStudio (R Development Core Team, [17]), we estimate the parameters a_x (age-specific profiles independent in time) and b_x (additional age-specific components determine how much each age group changes when k_t changes) for male and female Lee-Carter model. We can see these parameters in the Fig. 3, from which it is also clear the comparison between the different evolutions of parameters by sex. The age-specific profiles independent of time (a_x) are lower in the case of female model, because in general are the mortality rates of the female population lower in most age groups. The most significant difference between male and female mortality is in age group 18–32 years and at the oldest age groups. Higher mortality level of young males is caused by suicides, poisoning, dangerous behaviour, gambling, etc., (this is unfortunately a long-term trend).

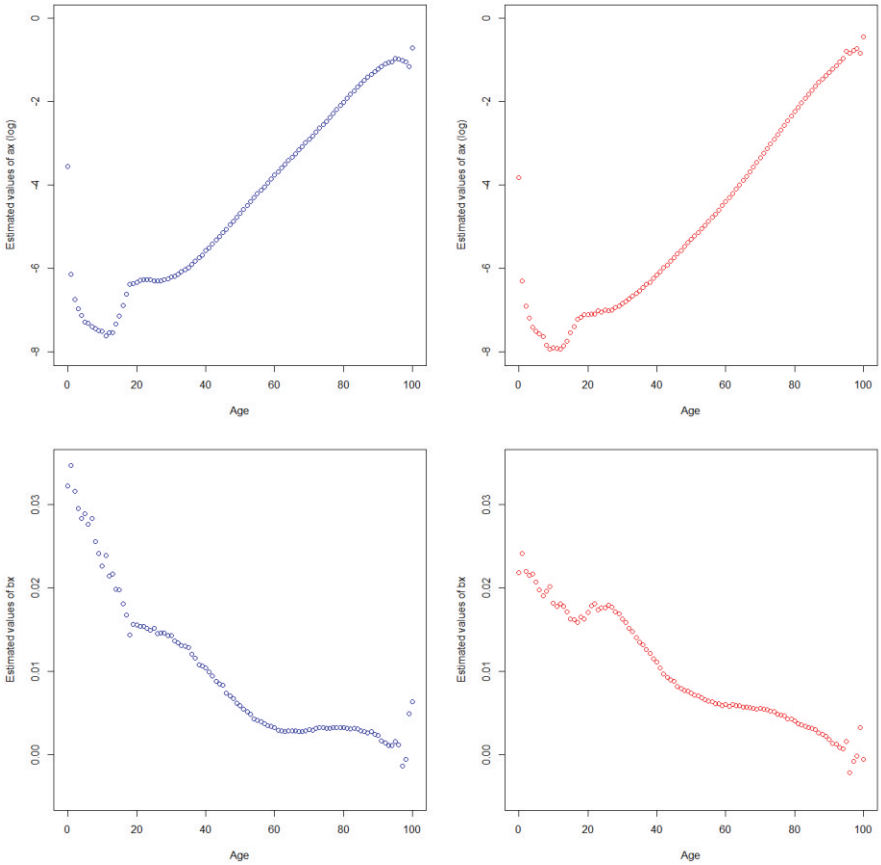

Figure 3 The estimates of age-specific profiles independent in time (parameter $\hat{a}_x$, top left for males and right for females) and the additional age-specific components determine how much each age group changes when k_t changes (parameter $\hat{b}_x$, bottom right for males and left for females). Source: own construction and illustration.

We also estimate the mortality indices k_t (the time-varying parameters) for the period 1920–2012 and these estimates are shown in the Fig. 4. To these estimates we calculate the predictions up to the year 2050 based on the *ARIMA* methodological approach, (Box, Jenkins [5]) and ran by "forecast" package in R (Hyndman et al. [11] and Hyndman, Shang [12]). Parameters of *ARIMA* models with drift are written in Tab. 1. From these predictions with 95% confidence intervals, (which can be seen in Fig. 4 too) it is clear that the model for female population provides slightly lower values of these estimates. Confidence intervals are wider in the model for male population (this is due to absence of AR parameter). Note that the theoretical basis for these predictions in sophisticated systems and automatic prediction software are prepared as well by Bell [1], Bell, Monsell [2] or Keyfitz [14].

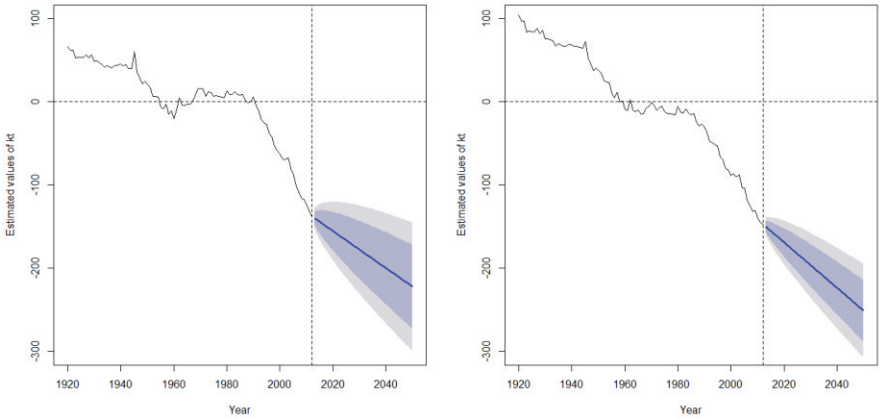

Figure 4 The estimates of the time-varying parameters $\hat{k}_t$ - the mortality indices with attached forecasts of these indices from 2013 to 2050 by ARIMA models. On the left side is the model for males, on the left side for females respectively. Source: own construction and illustration.

	AR(1)	s.e.	Drift	s.e.	AIC	BIC
Males ARIMA(0,1,0) with drift	x	x	-2.2181	0.6663	598.85	603.90
Females ARIMA(1,1,0) with drift	-0.2182	0.1020	-2.7256	0.4844	578.40	585.97

Table 1 Estimated parameters of ARIMA models for male and female $\hat{k}_t$. Source: own calculation.

Predicted values of the logarithms of age-specific death rates (from Fig. 2) by random walk models with drift can be combined with the empirical values of these rates (from Fig. 1). The result for male and female population is shown in the Fig. 5 (top). In order to compare these values all of the charts have the same scale. 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 estimate the future values of $\ln(m_{x,t})$ for males and females as

$$\ln\left(m_{x,t}^{M/F}\right) = \hat{a}_x^{M/F} + \hat{b}_x^{M/F} \cdot \hat{k}_t^{M/F} . \tag{5}$$

The result is shown in the Fig. 5 (bottom). Also in the case of Lee-Carter models it is clear that at the highest age groups break the assumption that the trend of the age-specific death rates decreases in the future.

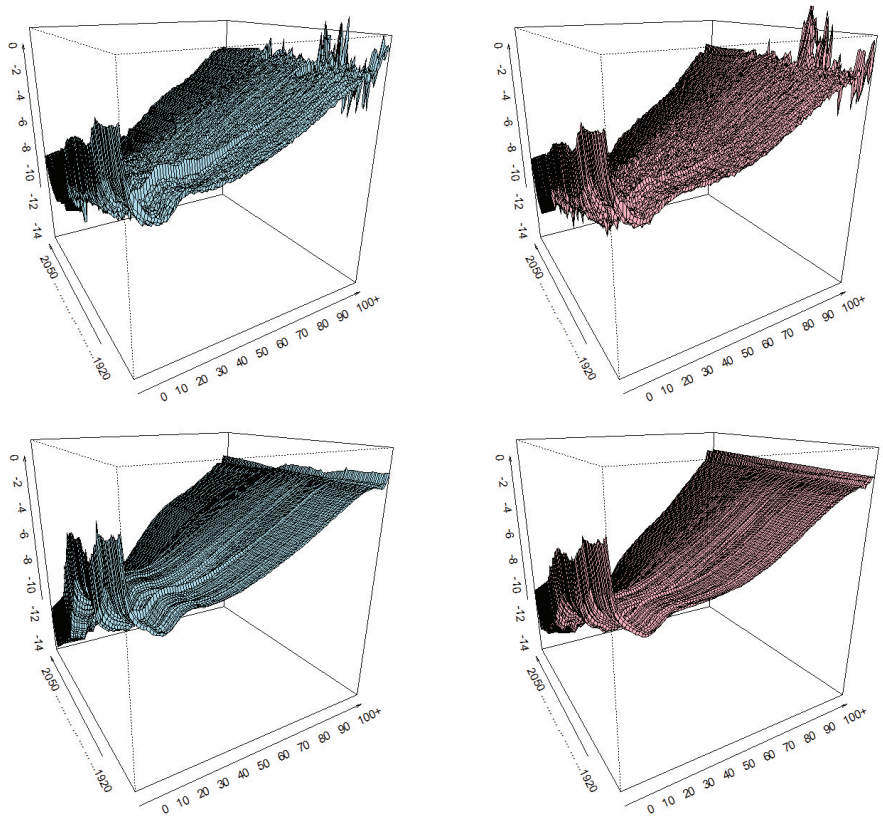

Figure 5 Age-specific death rates (empirical values in logarithms) of Czech males (top left) and females (top right) in 1920–2012 with attached forecasts of these rates from 2013 to 2050 by individual Random walk models with drift and age-specific death rates (fitted values in logarithms by Lee-Carter models) of Czech males (bottom left) and females (bottom right) in 1920–2012 with attached forecasts of these rates from 2013 to 2050 by Lee-Carter models. Source: own construction and illustration.

4 Conclusion

The aim of this paper was to show two different approaches to modelling age-specific death rates for male and female population. We estimated the models of random walk with drift for males and females. Then we forecasted and also estimated parameters of Lee-Carter model for both populations. Consequently we calculated the forecasts as well. In the case of the female population the results obtained by random walk models with drift and by the Lee-Carter model are almost comparable. Lee-Carter model just provides lower estimates of mortality rates at the lowest ages (which may be explained by the lower expected rate of future infant mortality). This result is fully consistent with study by Booth, Tickle, and Smith [4]. Differences depending on used model are

greater in the case of male population. Because we expect that the differences in male excess mortality will decrease in the future, it will be probably better to use the Lee-Carter model, which predicts a significant decline of logarithms of male age-specific death rates. Random walk models with drift are simpler, but in the case of male population we have to take into account also the additional information together with the past trend in time series. This additional information is significantly decreasing of the parameter b_x - the additional age-specific components determine how much each age group changes when k_t changes (see study by Lee, Tuljapurkar [16] and Hyndman et al. [11]). Results, which we obtained by Lee-Carter model, reflect better the expectations of some institutions. We can conclude that our predictions to a certain extent corresponds to the medium variant of the estimated future development of mortality in the Czech Republic, which periodically calculates the CZSO. In the conditions of the Czech Republic only the random walk models for the female population could be used.

Acknowledgements

This paper was supported by the project of Internal Grant Agency of VŠE Praha under the No. IGS F4/24/2013.

References

[1] Bell, W.R. (1997). Comparing and assessing time series methods for forecasting age-specific fertility and mortality rates. *Journal of Official Statistics*, 13(3): pp. 279-303.

[2] Bell, W.R., Monsell, B. (1991). 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.

[3] Boleslawski, L., Tabeau, E. (2001). 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.

[4] 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): pp. 13-34.

[5] Box, G.E.P., Jenkins, G. (1970). *Time series analysis: Forecasting and control*. San Francisco, Holden-Day, 537 p.

[6] Charpentier, A., Dutang, Ch. (2012). *L'Actuariat avec R*. [working paper]. Decembre 2012. Paternite-Partage a lindentique 3.0 France de Creative Commons, 215 p.

[7] Coale, Ansley J., Kisker, Ellen E. (1986). Mortality Crossovers: Reality or Bad Data? *Population Studies*. Vol. 40, pp. 389-401.

[8] Erbas, B., Ullah, S., Hyndman, R.J., Scollo, M., Abramson, M. (2012). Forecasts of COPD mortality in Australia: 2006-2025. *BMC Medical Research Methodology*, Vol. 2012, pp. 12-17.

[9] Gardner Jr. E.S., McKenzie, E. (1985). Forecasting Trends in Time Series. *Management Science*, 31(10): pp. 1237-1246.

[10] Gavrilov, L. A., Gavrilova, N. S. (2011). Mortality measurement at advanced ages: a study of social security administration death master file. *North American actuarial journal*, 15(3): pp. 432-447.

[11] Hyndman, R.J., Koehler, A.B., Snyder, R.D., Grose, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting*, 18(3): pp. 439-454.

[12] Hyndman, R.J., Shang, Han Lin (2009). Forecasting functional time series. *Journal of the Korean Statistical Society*, 38(3): pp. 199-221 (With discussion).

[13] Hyndman, R.J. (2012). *demography: Forecasting mortality, fertility, migration and population data*. R package v. 1.16, http://robjhyndman.com/software/demography/

[14] Keyfitz, N. (1991). Experiments in the projection of mortality. *Canadian Studies in Population*, 18(2): pp. 1-17.

[15] Lee, R.D., Carter, L.R. (1992). Modeling and forecasting U.S. mortality. *Journal of the American Statistical Association*, vol. 87, pp. 659-675.

[16] Lee R.D., Tuljapurkar, S. (1994). Stochastic population forecasts for the United States: beyond high, medium, and low. *Journal of the American Statistical Association*, Vol. 89, pp. 1175-1189.

[17] R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

[18] Šimpach, O. (2012). 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*., Opava: Universita Opava, pp. 870-874.

[19] Šimpach, O., Pechrová, M. (2013). Assessing the impact of standard of living on the life expectancy at birth using Vector Autoregressive Model. In: *Proceedings of 31st International Conference on Mathematical Methods in Economics, PTS I and II*, Jihlava: College of Polytechnics Jihlava, pp. 921-926.

[20] Thatcher, R. A., Kanisto, V., and Vaupel, J. W. (1998). *The Force of Mortality at Ages 80 to 120*. Odense University Press.