User versus Automatic Selection of Models in Actuarial Demographics: The Impact on the Expected Development of the Probability of Death in the Czech Republic

Ondřej Šimpach[1], Marie Pechrová[2]

Abstract. For the needs of actuarial demographics, it is not sufficient to construct only deterministic models, but it is suitable to combine the results also using stochastic ones. Finding optimal form of the model is a complicated process, because it is necessary to test not only statistical significance of the parameters, but also the significance of the whole model and to respect the results of the diagnostics tests. While applying of this modelling on data of age-and-sex specific mortality rates, there is a problem in the number of those models. There are about 2 × 101 or more models in advanced populations, as each gender is necessary to model separately and the age range is usually surveyed by statistical offices in the intervals 0–100+ completed years of life.

The aim of the paper is to assemble 2 × 101 optimal ARIMA models on the data of logarithms of age-and-sex specific mortality rates in the Czech Republic. Those models are diagnosed and consequently used for predictions. Other 2 × 101 ARIMA models are created on the same data, but using automatic process in software RStudio. It is applied a compromise form of the model on all time series. Both approaches - user and automatic are compared based on the results of the projection, that is recalculated on the probability of death of *x*-year old person. It was proved that automatic process is significantly faster and that the results of the projections are not distorted.

Keywords: ARIMA, mortality, forecast, probability of death, RStudio.

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

1 Introduction

"Mortality rates of human populations in developed countries are declining with time" (Finkelstein, [6]). With increasing life expectancy improving long-term care and sustaining the pension system are becoming an important issue. Besides, establishing methodologically sound longitudinal data sets is necessary in order to examine the phenomena (Andel [1]). For the purposes of actuarial demography, it is not sufficient to construct only deterministic models, but it is necessary to combine them and to compare the results also which stochastic ones. "The actuarial and demographic literature has introduced a myriad of (deterministic and stochastic) models to forecast mortality rates of single populations" (Antonion, Bardoutsos and Ouburg [2]). Work of Gompertz [10] played an important role in shaping the emerging statistical science. Gompertz model provided a powerful stimulus to examine the patterns of death ("law of mortality") across the life course not only in humans but also in a wide range of other organisms (Kirkwood [13]). Since that many of models have been developed. Lee and Carter [14] published a new statistical method for forecasting mortality in 1992. Since that it has been applied on many real data of the populations. For example, Li and Lee [15] applied the Lee-Carter model to a group of populations, allowing each its own age pattern and level of mortality but imposing shared rates of change by age.

However, there are more models used. For example, Antonion, Bardoutsos and Ouburg [2] presented in their paper a Bayesian analysis of two related multi-population mortality models of log-bilinear type, designed for two or more populations. Gogola [8] used stochastic mortality – Cairns, Blake and Dowd model that is well suited at very high ages to calculate mortality rates of age categories from 85 to 115 for selected countries. Godunov [9] compared several models applied on particular populations and found that „the most appropriate models of smoothing mortality curve are Kannistö-Thatcher (UK) Martinell (Sweden) and Kannistö (Canada). On the other side, the least suitable models are Coale-Kisker (Singapore), Gompertz-Makeham and modified Gompertz-Makeham (Czech Republic, Slovakia, Germany)". Mortality rates in the Czech Republic was calculated for example by Jindrová and Slavíček [12]. They applied Lee-Carter model on Czech population in the period of 1950–2009 and

[1] University of Economics Prague, Department of Statistics and Probability, W. Churchill sq. 4, 130 67 Prague 3, Czech Republic, ondrej.simpach@vse.cz.
[2] Institute of Agriculture Economics and Information, Mánesova 1453/75, 120 00 Prague 2, Czech Republic, pechrova.marie@uzei.cz.

predicted the development of specific mortality rates and consequently life expectancy for period 2010–2029. In the broader context, Fiala and Langhamrová [5] analysed of the development of the sex-and-age structure of the Czech population of productive age based on the latest population projection of the Czech Statistical Office. Probabilistic projections of age-specific mortality and fertility rates were done for example by Ševčíková et al. [16] in order to apply probabilistic population projections on United Nations (UN) countries.

Finding optimal form of the model age-and-sex specific mortality rates $m_{x,t}$ (respectively probability of death ($q_{x,t}$) after recalculation according to life table algorithm) is a complicated process, because it is necessary to test not only statistical significance of the parameters, but also the significance of the whole model and to respect the results of the diagnostics tests. Finding a suitable model for $m_{x,t}$ raises a problem that there are about 2 × 101 or more models in advance populations, because each gender and age[3] is necessary to model. Therefore, the aim of the paper is to compare user and automatic ARIMA approaches towards the model selection based on the results of the mortality projection that is recalculated on the probability of death of x-year old person. Paper shows that automatic process is significantly faster and that the results of the projections are not significantly distorted.

2 Methods and Data

Mortality is one of the demographic indicators showing the percentage of deaths for a certain period from a group of people. The age-specific mortality rates are one of the basic quantities used in modelling mortality. They are calculated according to the formula (1)

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

where $m_{x,t}$ is age-specific mortality rate in age x and time t, $D_{x,t}$ is the number of deaths at completed age x in time t, and $E_{x,t}$ is mid-year state of x-year old population in time t, i.e. exposure to risk. Based on the observation of large population groups using demographic methods it is possible to estimate the probability of death for males and females of different ages and other important characteristics. From mortality tables at each year are calculated probabilities of death ($q_{x,t}$) at age x and time t as (2)

$$q_{x,t} = 1 - e^{-m_{x,t}}, \tag{2}$$

where e represents Euler's constant. Probabilities of survival ($p_{x,t}$) at age x and time t are therefore calculated as $p_{x,t} = 1 - q_{x,t}$. (Next steps can be seen, for example, in paper by Šimpach and Langhamrová [17]).

Data about of age-and-sex specific mortality rates for the period 1920–2015 in the Czech Republic were obtained from Czech Statistical Office (CZSO). They were consequently transformed to logarithms, tested by Dickey-Fuller test (ADF test) whether they were stationary and Box-Jenkins (Box and Jenkins [4]) methodology was applied on them. There are 3 types of ADF test with constant and trend, with constant only, and without constant and trend (i.e. without deterministic elements). The first case is calculated according the equation (3)

$$\Delta Y_t = \beta_1 + \beta_2 t + \beta_3 Y_{t-1} + \sum_{i=1}^{m} \alpha_i Y_{t-i} + \varepsilon_t, \tag{3}$$

where Δ is the first difference of the examined variable, t is the trend variable in this case, ε_t is pure white noise error term, m is the maximum length of the lagged dependent variable, and α, β are parameters (β_1 represents the constant). Box and Jenkins [4] introduced the models that are working with autoregressive (AR(p)) and moving average (MA(q)) processes. When the time series is not stationary, its difference of d^{th} order must be done. Diagnostic of the model type is done by Autocorrelation function (ACF) and Partial Autocorrelation function (PACF) that were plotted in order to determine the order p of AR process and order q of MA process. General ARIMA(p,d,q) model is formulated as (4).

$$Y_t = \beta + \sum_{i=1}^{p} \alpha_i Y_{t-i} + \sum_{j=1}^{q} \delta_j \varepsilon_{t-j} \tag{4}$$

There were assembled 2 × 101 optimal ARIMA models on the data of logarithms of age-and-sex specific mortality rates in the Czech Republic. From above stated information can be seen that wide diagnostic of the

[3] Age range is usually surveyed by statistical offices in the intervals 0–100+ completed years of life.

models is needed. Appling all tests to 202 time series manually can be time consuming. Therefore, other 2 × 101 ARIMA models were applied on the same data, but using automatic process in software RStudio. Both approaches user and automatic are compared based on the results of the projection, that is recalculated on the probability of death of x-year old person.

3 Results

From the empirical data of age-and-sex specific mortality rates in logarithms (see Fig. 1, top charts) is evident, that there was unstable development of time series of mortality rates until 1950 (especially in the case of male population). The age-and-sex specific mortality rates are lower in the case of female population, because, in general, the intensity of mortality ($\mu_{x,t}$ according to Gompertz law) of the female population is lower in most age groups. Significant difference between male and female mortality is in age group 18–32 years and at the oldest age groups (85+). Higher mortality level of young males is caused by suicides, poisoning, dangerous behaviour, gambling, etc. (this is unfortunately a long-term trend in most populations, not only in the Czech Republic). Recalculated values of age-and-sex specific probabilities of death according to the life table algorithm are shown in Fig. 1 (bottom charts). It is well known that the instability of the time series reduces their predictive capability (Bell [3] or Gardner, McKenzie [7]). The history, although, has the lowest weight in the prediction model. For 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. We do not smooth the empirical data in this paper according to Gogola [8] or Godunov [9] and rather use the ARIMA models with or without constant on raw data.

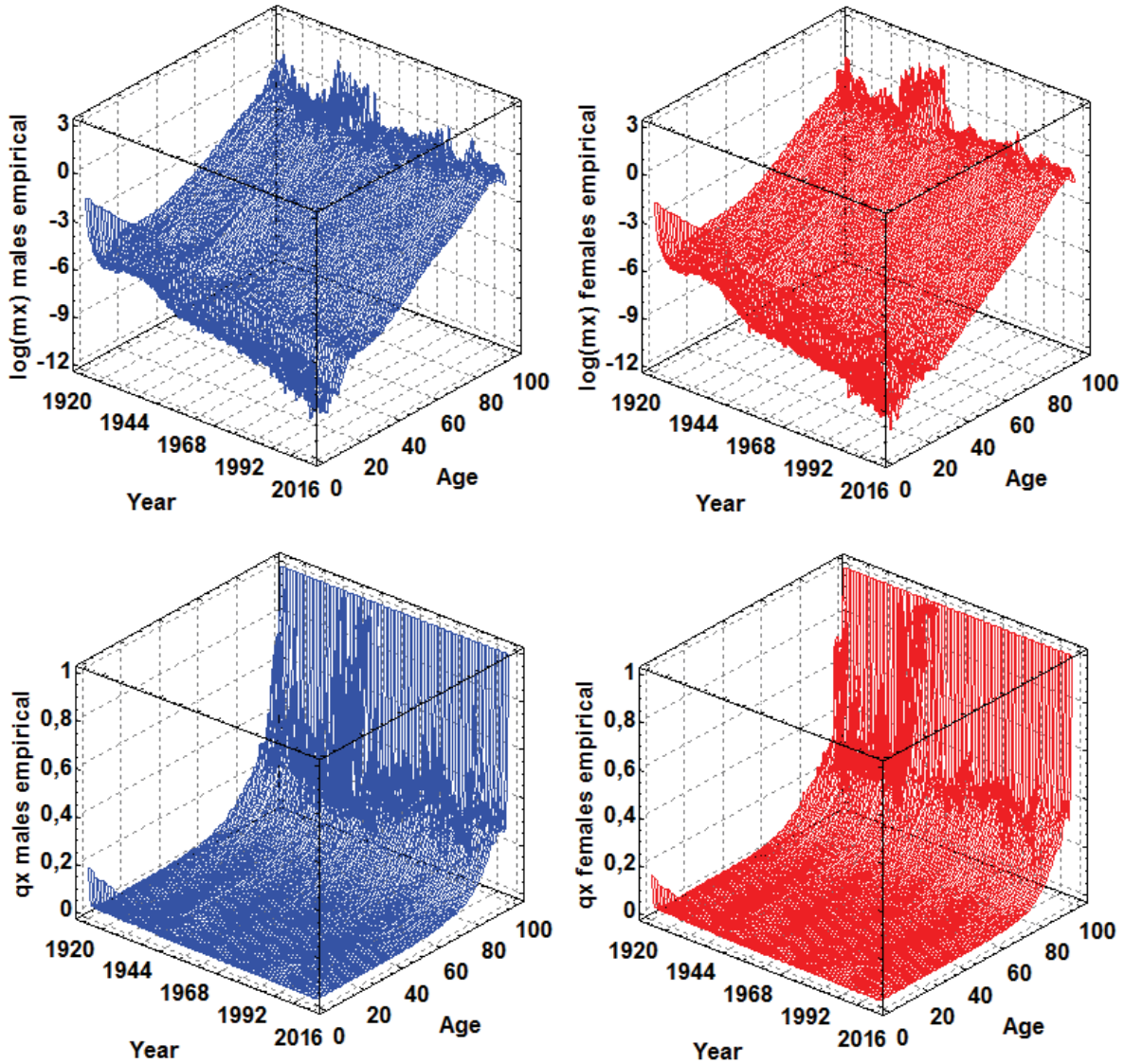

Figure 1 Logs of age-and-sex specific mortality rates for males and females in the Czech Republic (top charts) and calculated probabilities of death (bottom charts). Data source: CZSO, authors' illustration.

Overview of ARIMA models used on 2 × 101 time series is for males shown in the Table. 1, for females in Table 2. These models were evaluated by residual diagnostic tests (autocorrelation, heteroskedasticity and normality; see e.g. paper by Jarque and Bera [11]) and they are correct in all cases at 5% statistical significance level. It is clear from this overview that the most common form of models is ARIMA(0,1,1) with a constant. (This situation occurred in 77 cases (out of 101) in male population and in 61 cases (out of 101) in female population). This form was used as a compromise and a script was programmed to fit the 2 × 101 time series once again in the RStudio software. It was done this time without diagnostic tests, which could not be at 5% significance level statistically significant in some cases. Therefore, some parameters might have be deflected.

Age	Model	Age	Model	Age	Model	Age	Model	Age	Model	Age	Model	Age	Model	Age	Model
0	(0,2,1) c	13	(0,1,1) c	26	(0,1,1) c	39	(0,1,1) c	52	(0,1,1) c	65	(0,1,1) c	78	(0,1,1) c	91	(0,1,2) c
1	(0,1,1) c	14	(0,1,1) c	27	(0,1,1) c	40	(2,2,1)	53	(0,1,1) c	66	(0,1,1) c	79	(1,1,1) c	92	(2,1,0) c
2	(2,2,1)	15	(0,1,1) c	28	(0,1,1) c	41	(0,1,1) c	54	(0,1,1) c	67	(0,1,1) c	80	(0,1,1) c	93	(0,1,1) c
3	(0,1,1) c	16	(0,1,1) c	29	(0,1,1) c	42	(0,1,1) c	55	(0,1,1) c	68	(0,1,1) c	81	(0,1,1) c	94	(0,1,2) c
4	(0,1,1) c	17	(0,1,1) c	30	(0,1,1) c	43	(0,1,1) c	56	(2,1,0) c	69	(0,1,1) c	82	(0,1,1) c	95	(1,1,1) c
5	(0,1,1) c	18	(0,1,1) c	31	(2,2,1)	44	(0,1,1) c	57	(0,1,1) c	70	(0,1,1) c	83	(1,1,1) c	96	(0,1,2) c
6	(0,1,1) c	19	(0,1,1) c	32	(0,1,1) c	45	(0,1,1) c	58	(0,1,1) c	71	(0,1,1) c	84	(0,1,1) c	97	(0,1,1) c
7	(0,1,1) c	20	(0,1,2) c	33	(0,1,1) c	46	(0,1,1) c	59	(0,1,1) c	72	(0,1,1) c	85	(0,1,1) c	98	(0,1,1) c
8	(0,1,1) c	21	(0,1,1) c	34	(0,1,1) c	47	(2,1,0) c	60	(0,1,1) c	73	(1,1,0) c	86	(0,1,1) c	99	(0,1,2) c
9	(0,1,1) c	22	(0,1,1) c	35	(0,1,1) c	48	(1,1,0) c	61	(0,1,1) c	74	(1,1,0) c	87	(0,1,1) c	100+	(0,1,1) c
10	(0,1,1) c	23	(0,1,1) c	36	(0,1,1) c	49	(0,1,1) c	62	(0,1,1) c	75	(2,1,0) c	88	(0,1,1) c		
11	(0,1,1) c	24	(0,1,1) c	37	(0,1,1) c	50	(2,1,0) c	63	(0,1,1) c	76	(0,1,1) c	89	(0,1,2) c		
12	(0,1,1) c	25	(2,1,0) c	38	(0,1,1) c	51	(1,1,0) c	64	(0,1,1) c	77	(0,1,1) c	90	(2,1,0) c		

Table 1 ARIMA (*p,d,q*) models with or without constant for male's logarithms of age-specific mortality rates at the exact age 0–100+ in the Czech Republic. Source: authors' illustration.

Age	Model	Age	Model	Age	Model	Age	Model	Age	Model	Age	Model	Age	Model	Age	Model
0	(0,2,1) c	13	(0,1,1) c	26	(0,1,1) c	39	(0,1,1) c	52	(0,1,1) c	65	(1,1,0) c	78	(2,1,2) c	91	(1,1,1) c
1	(0,1,1) c	14	(1,1,1) c	27	(0,1,1) c	40	(2,1,0) c	53	(0,1,1) c	66	(1,1,0) c	79	(0,1,1) c	92	(2,1,0) c
2	(1,1,1) c	15	(0,1,1) c	28	(0,1,1) c	41	(1,1,0) c	54	(0,1,1) c	67	(0,1,1) c	80	(0,1,1) c	93	(0,1,1) c
3	(2,1,0) c	16	(2,1,0) c	29	(0,1,1) c	42	(0,1,1) c	55	(0,1,1) c	68	(0,1,1) c	81	(0,1,1) c	94	(2,1,1) c
4	(2,1,0) c	17	(0,1,1) c	30	(0,1,1) c	43	(0,1,1) c	56	(0,1,1) c	69	(0,1,1) c	82	(0,1,1) c	95	(1,0,0) c
5	(0,1,1) c	18	(0,1,2) c	31	(0,1,1) c	44	(0,1,1) c	57	(0,1,1) c	70	(0,1,1) c	83	(0,1,1) c	96	(1,0,0) c
6	(1,1,1) c	19	(0,1,1) c	32	(0,1,1) c	45	(0,1,1) c	58	(2,1,0) c	71	(1,1,0) c	84	(0,1,1) c	97	(0,0,2) c
7	(0,1,1) c	20	(0,1,1) c	33	(0,1,1) c	46	(0,1,1) c	59	(1,1,1) c	72	(2,1,0) c	85	(0,1,1) c	98	(0,0,2) c
8	(2,1,0) c	21	(0,1,1) c	34	(0,1,1) c	47	(0,1,1) c	60	(0,1,1) c	73	(1,1,0) c	86	(0,1,1) c	99	(1,1,0) c
9	(2,1,0) c	22	(2,1,0) c	35	(2,1,0) c	48	(0,1,1) c	61	(0,1,2) c	74	(2,1,0) c	87	(0,1,1) c	100	(0,1,1) c
10	(0,1,1) c	23	(0,1,1) c	36	(0,1,1) c	49	(1,1,1) c	62	(1,1,0) c	75	(2,1,0) c	88	(0,1,1) c		
11	(0,1,1) c	24	(1,1,0) c	37	(0,1,1) c	50	(0,1,1) c	63	(1,1,0) c	76	(0,1,1) c	89	(1,1,1) c		
12	(1,1,1) c	25	(0,1,1) c	38	(0,1,1) c	51	(0,1,1) c	64	(0,1,1) c	77	(0,1,1) c	90	(2,1,0) c		

Table 2 ARIMA (*p,d,q*) models with or without constant for female's logarithms of age-specific mortality rates at the exact age 0–100+ in the Czech Republic. Source: authors' illustration.

Optimized (user) models and compromised ARIMA (0,1,1) c (automatic) models were subsequently used to predict the indicator up to the year 2050. These mortality rates were recalculated using classical life table algorithm into the probability of death. Results are shown in the Figure 2, where at the top are the results of prediction by user models, in the middle part by automatic (ARIMA (0,1,1) c) models. For mutual comparison, the differences were calculated as

$$diff\left(q_{x,t}\right) = q_{x,t}^{user} - q_{x,t}^{automatic} \tag{5}$$

and these results are subsequently shown in the Figure 2 on the bottom. As can be seen from the results, there are almost no differences in prediction of the probability of death in the age groups of 0–95 years. Only in the highest age groups over 95 years there are deviations, which are probably caused due to variability in empirical data. If we would program a script that will smooth the input data using one of the existing levelling models, the ARIMA model (0,1,1) c would probably work better and the detected deviations would not occur. Because the variability in the highest age groups in input data is lower in the case of male population, the resulting deviations are therefore significantly smaller.

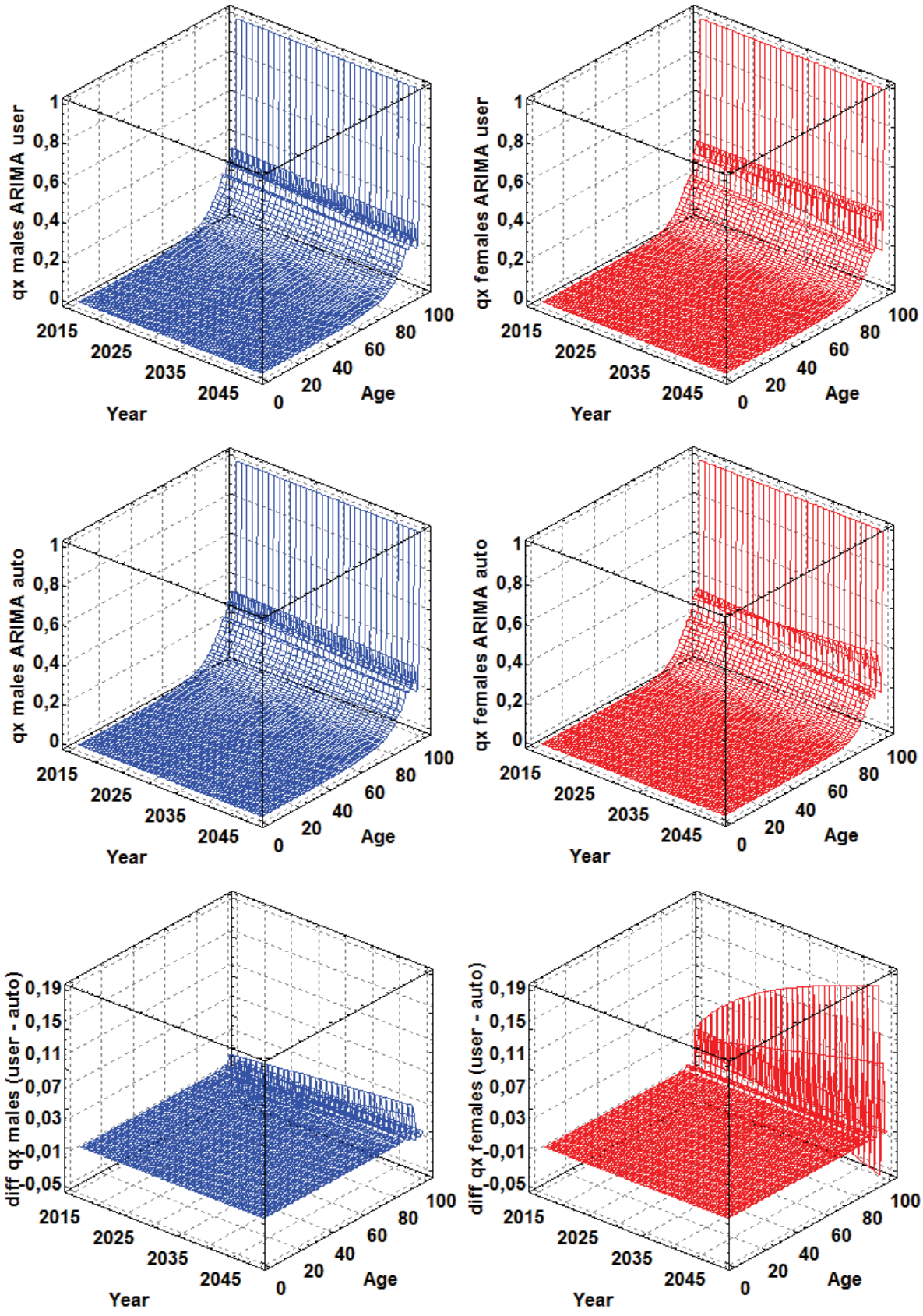

Figure 2 Forecasted age-and-sex-specific probabilities of deaths for male and female population in the Czech Republic up to the year 2050 using user and automatic approach. Source: authors' illustration.

4 Conclusion

The aim of this paper was to perform two approaches to modelling age-and-sex specific mortality rates for male and female population in the Czech Republic. We estimated 2 × 101 optimal ARIMA models and these subjected to the diagnosis of residues (autocorrelation, heteroskedasticity and normality). Then we forecasted mortality rates up to the year 2050. In the second step, we found that the most frequently occurring form of the model is ARIMA (0,1,1) with constant. We programmed script that uses this model structure and applies it to all 2 × 101 time series again using RStudio software, this time without the evaluation by the diagnostic tests. Consequently, we calculated the forecasts up to the year 2050 as well. It was found that in the case of male population are the results almost comparable. Differences depending on used model are greater in the case of female population. It is mainly caused by great variability in the raw data at the highest ages. This problem could be resolved by smoothing of mortality data by some of the existing models (see e.g. study by Gogola [8] or Godunov [9]), but that would bring a lot of extra steps that would completely eliminate the effect of saving work and effort.

Acknowledgements

The paper was supported from the Internal Grant Agency of University of Economics Prague no. 35/2017 “Demographic models in R Software”.

References

[1] Andel, R. (2014). Aging in the Czech Republic. *Gerontologist*, vol. 54, no. 6, pp. 893-900.
[2] Antonion, K., Bardoutsos, A., and Ouburg, W. (2015). Bayesian Poisson log-bilinear models for mortality projections with multiple populations. *European Actuarial Journal*, vol. 5, no. 2, pp 245-281.
[3] Bell, W. R. (1997). Comparing and assessing time series methods for forecasting age-specific fertility and mortality rates. *Journal of Official Statistics*, vol. 13, no. 3, pp. 279-303.
[4] Box, G. E. P., Jenkins, G. (1970). *Time series analysis: Forecasting and Control*. San Francisco: Holden-Day.
[5] Fiala, T., Langhamrová, J. (2014). Increase of Labor Force of Older Age – Challenge for the Czech Republic in Next Decades. *Procedia Economics and Finance*, vol. 12, pp. 144-153.
[6] Finkelstein, M. S. (2005). Lifesaving explains mortality decline with time. *Mathematical Biosciences*, vol. 196, no. 2, pp. 187-197.
[7] Gardner Jr. E. S., McKenzie, E. (1985). Forecasting Trends in Time Series. *Management Science*, vol. 31, no. 10, pp. 1237-1246.
[8] Gogola, J. (2015). Modelling mortality rate of the very old. *In: 9th Professor Aleksander Zelias International Conference on Modelling and Forecasting of Socio-Economic Phenomena*, Cracow: Cracow University of Economics, pp. 58-67.
[9] Godunov, I. (2015). *Comparison of official life tables construction in selected countries [Bachelor's thesis]*. Charles University in Prague, Faculty of Science, Department of Demography and Geodemography.
[10] Gompertz, B. (1825) 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.
[11] Jarque, C. M., Bera, A. K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, vol. 6, no. 3, pp. 255-259.
[12] Jindrová, P., Slavíček, O. (2014). Life Expectancy Development and Prediction for Selected European Countries *In: 6th International Scientific Conference Managing and Modelling of Financial Risks*, Ostrava: VŠB-TU Ostrava, pp. 303-312.
[13] Kirkwood, T. B. L. (2015). Deciphering death: a commentary on Gompertz (1825) '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, B-Biological Sciences*, vol. 370, no. 1666, UNSP 20140379.
[14] Lee, R. D. Carter, L. (1992). Modeling and Forecasting the Time Series of U.S. Mortality. *Journal of the American Statistical Association*, vol. 87, pp. 659-671.
[15] Li, N., Lee, R. D. (2005). Coherent mortality forecasts for a group of populations: An extension of the Lee-Carter method. *Demography*, vol. 42, no. 3, pp. 575-594.
[16] Ševčíková, H., Li, N., Kantorová, V., Gerland, P., and Raftery, A. E. (2015). *Age-Specific Mortality and Fertility Rates for Probabilistic Population Projections*. Center for Statistics and the Social Sciences, University of Washington, Working Paper no. 150, March 17, 2015, pp. 1-31.
[17] 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)*, Olomouc: Palacký University in Olomouc, pp. 890-895.