20th International Scientific Conference AMSE
Applications of Mathematics and Statistics in Economics 2017
Szklarska Poręba, Poland 30 August 2017 – 3 September 2017

NORMALITY AND PRINCIPAL COMPONENT APPROACH TOWARDS FERTILITY TIME SERIES IN POLAND AND CZECHIA

ONDŘEJ ŠIMPACH
University of Economics Prague, Faculty of Informatics and Statistics,
W. Churchill sq. 4, 130 67 Prague 3, Czech Republic
email: ondrej.simpach@vse.cz

Abstract

The aim of this paper is to analyse changes of trends in time series of age-specific fertility rates (functional data, where fertility rate is a function of women's age) in Poland and Czechia. Data from Eurostat database are available from 1990 to 2014 for both countries. During this period, the behaviour of the population has changed in terms of family planning. The birth of the first child is being continuously postponed to later ages. Also, the fertility rates are lower. However, the situation between Poland and Czechia differs. Therefore, we compare the development.

First, it is searched whether and when occurs normal shape of fertility function. Series of Jarque-Bera tests is applied on individual time series (functional data, where fertility rate is a function of women's age). The analysis revealed that occurs normal shape of fertility function in the age groups 15–49+ in both countries during the whole period. It has not been skewed to the right (to higher ages) yet. Second, the medians of fertility rates were analysed using principal components method implemented in "demography" and "rainbow" packages of the RStudio software. The changes in median fertility curves were examined based on two 10-years long time intervals: 1990–1999 and 2005–2014. We found that they were significantly shifted to the right.

These results are important for subsequent analyses because for working with demographic data about fertility it is important to consider the most recent data, which are not significantly skewed and influenced by a range of factors. Estimated parameters of shape of fertility function can be also used for predictions.

Key words: *normality, principal components method, fertility rates, Poland, Czechia*

JEL Codes: *C22, C32, J13*

DOI: 10.15611/amse.2017.20.35

1. Introduction

Fertility together with mortality and migration are important demographical process of the natural population change. Modelling and estimation of fertility is more complicated than of mortality, because fertility is influenced by several external factors that are hard to be assumed. The population development and improvement of the living standards in the country are closely related to the postponement of first childbirth to the later ages and with the decline of number of live births in total. This decrease of fertility is below the level of simple reproduction of the population (2.08 live born children per 1 female on average within the reproduction period) in Poland and Czechia. Therefore, the aim of the paper is to explore the development of fertility rates (functional data, where fertility rate is a function of women's age) in those countries and compare the situation.

A period when occurs the normal shape of fertility function in Poland and Czechia was chosen for examination. Therefore, first, a Jarque-Bera test was used to evaluation of the shape of individual time series of age-specific fertility rates from 1990 to 2014. In majority of the results the test does not reject null hypothesis claiming, that the fertility function has normal shape during the analysed period. Shape of fertility function is not it those cases skewed or biased. It is flat, with median almost equal to mean. This information is necessary for the next step – analysis of demographic data.

Second, using the principal component approach the changes in the development of median fertility curves during two 10-years long time intervals (1990–1999 and 2005–2014) were analysed. Both periods have (in majority of cases) normal shape of fertility function and thus on the bases of the median fertility curve the parameters of the shape of this curve were estimated. Those parameters could be used instead of a_x parameter of the Lee-Carter model (Lee, Carter, 1992, Lee, Tuljapurkar, 1994) for predicting levels of fertility in future population studies.

The database of Polish and Czech age-specific fertility rates recorded significant changes from 1990 to 2014, because the development of these rates was affected by a wide range of social changes after the end of the Communist regime in both countries.

2. Materials and Methods

The data about age-specific fertility rates $f_{x,t}$ (functional data, where fertility rate is a function of women's age) for Poland and Czechia were calculated by Eurostat on the basis of known numbers of live-born persons to x-year-old mothers in the calendar year t (can be labelled as $B_{x,t}$) and the numbers of mid-year females' x-year-old in the calendar year t (can be labelled as $E_{x,t}$), which is an exposure time. Using RStudio software (RStudio Team, 2015, R Core Team, 2017) the *BASE* was established according to Hyndman (2012) approach as

$$BASE \leftarrow f_{x,t} = \frac{B_{x,t}}{E_{x,t}}. \tag{1}$$

These *BASEs* of rates are established because this structure together with "demography" and "rainbow" package (Hyndman et al., 2017, and Shang, Hyndman, 2016) is used for stochastic fertility modelling by Lee-Carter model (Lee, Carter, 1992, and Shang, Hyndman, 2010)

$$f_{x,t} = a_x + b_x \cdot k_t + \varepsilon_{x,t}. \tag{2}$$

where a_x are the age-specific fertility profiles independent of time, b_x are the additional age-specific components determine how much the fertility in each age group changes when indices k_t change and k_t are the time-varying parameters – the fertility indices. $\varepsilon_{x,t}$ is an error term with the classical characteristics of white noise process, where expected value $E(\varepsilon_{x,t}) = 0$, dispersion $D^2(\varepsilon_{x,t}) = \sigma_\varepsilon^2$, covariance $cov(\varepsilon_{x,t}\,;\,\varepsilon_{x,t}') = 0$, $\varepsilon_{x,t} \approx \mathrm{N}(0,\sigma_\varepsilon^2)$ distribution, $x = 15, 16, ..., 49+$ (50 years and older females are added to the 49-year-old) and $t = 1, 2, ..., T$.

According to the approach elaborated by Jarque, Bera (1987) it is possible to examine, whether $f_{x,t}$ has normal shape. Null hypothesis H_0: data is normally distributed, is tested against the alternative H_1: non H_0. The test criterion is as follows

$$JB(TC) = \frac{N}{6}\left(S^2 + \frac{(K-3)^2}{4}\right), \tag{3}$$

(where S is skewness, K is kurtosis and N is number of observation = population).

By principal components (PC) method (using "demography" (Hyndman et al., 2017) and "rainbow" package (Shang, Hyndman, 2016) implemented in in RStudio software (RStudio

Team, 2015, R Core Team, 2017) it is possible to identify the outlying years of the analysed demographic data. This is graphically displayed based on the results of the principal component scores. Hyndman, Shang (2010) denoted $i = 1, \ldots, N$ as index of the rows and $j = 1, \ldots, M$ as index of the columns. Hence, according to their approach, the combination of variables (columns) can be linearized as

$$Z_{i,1} = c_{i,1}Y_{i,1} + c_{i,2}Y_{i,2} + \quad + c_{i,M}Y_{i,M} . \tag{4}$$

According to Hyndman, Shang (2010) the "formula basically says to multiply row elements with a certain value c (loadings) and sum them by columns. Resulting values (Y values times the loading) are scores. A principal component (PC) is a linear combination $Z_1 = (Z_{1,1}, ..., Z_{N,1})$ (values by columns which are called scores). Basically, the PC should present the most important features of variables (columns)". Two highest eigenvalues and corresponding eigenvectors of covariance (or correlation) matrix are used to get optimal loadings in the equation (4).

Shang, Hyndman (2010) introduced the functional and bivariate bagplots for clear visualization of the outliers in functional data. Their approach to identify the outliers in the development of age-specific fertility rates was used in this paper. The functional and bivariate bagplots always contain two regions: dark grey and light grey. The first one contains 50% of all observations. There are also three curves – black curve that shows median and 2 dashed lines that represents 95% confidence intervals. As noted by Shang, Hyndman (2010): "functional curves that are outside the border region are considered to be outliers". These curves are in functional bagplot shown in colour. The bivariate bagplot does not show the median curve, but Tukey depth median (see Tukey, 1975, methodology for finding the median is presented by Chan, 2004). Coloured points with year label outside the threshold region are outliers.

3. Results

First, a normal shape of the time series was tested. Second, a principal component analysis was applied.

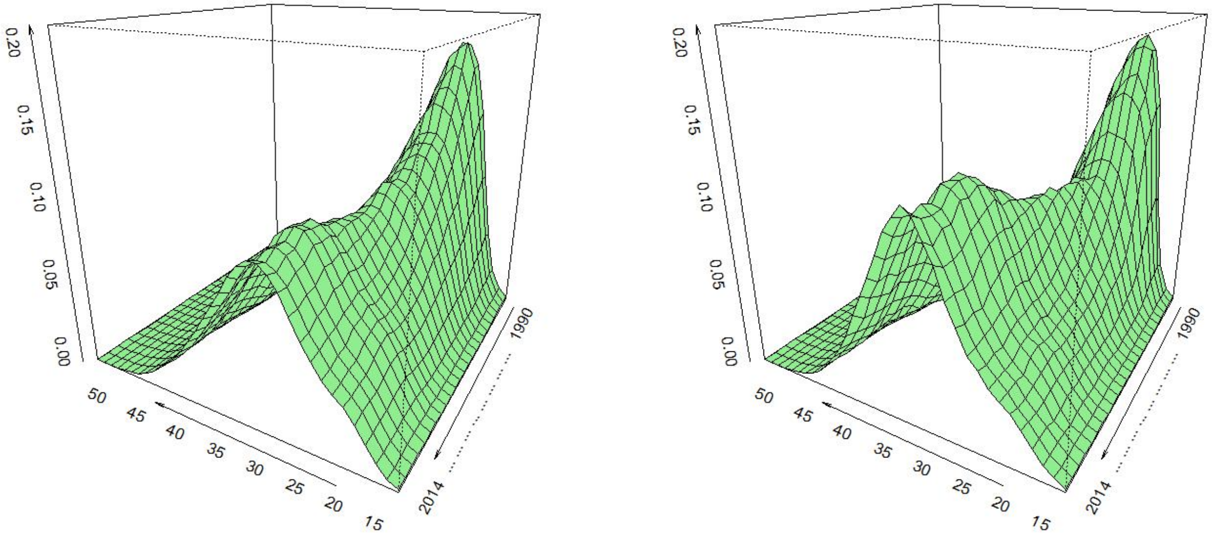
3.1 Examination of normal shape of fertility function

Empirical age-specific fertility rates $f_{x,t}$ of Polish (left chart) and Czech (right chart) females in 1990–2014 can be seen in Fig. 1a. It is evident that modes of those distributions are significantly changed to the advanced ages after nineties. Interpretation of values on the y axis is the number of live births per 1 female in the reproductive age of x years and time t.

Transformation of fertility during the nineties can be clearly seen from bottom chart (Fig. 1b). Distribution started to decline. The modus began to shift to higher ages later.

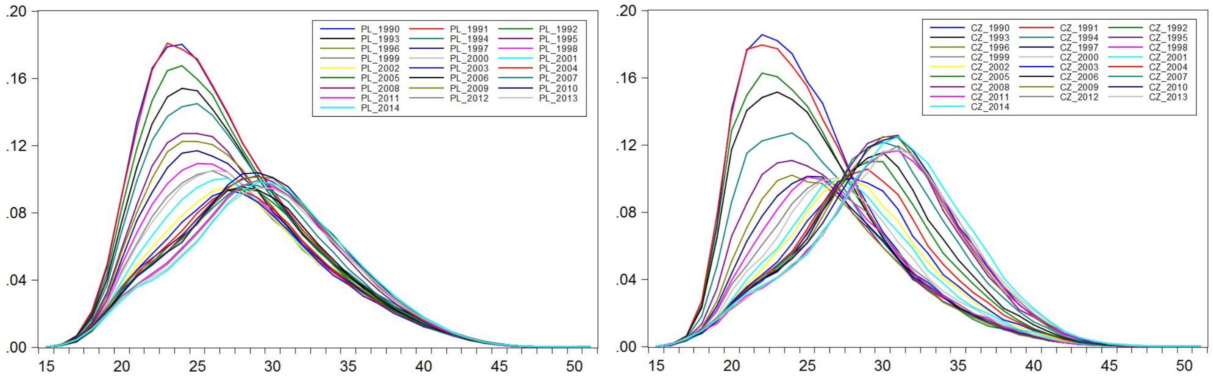
Every single calendar year t was examined as an individual dataset. H_0 of normality was tested. Calculated Jarque-Bera TC's according to formula (3) are shown in Fig. 2. When the value is (approximately) less than 6.25, the hypothesis assuming normal shape of fertility function is not rejected at the 5% significance level. This situation occurred in the years 1990, 1991 and 1992 in Czechia. Distribution of fertility rates in those years was positively skewed. Beside the above-mentioned period, the shape of Polish and Czech fertility function was approximately normal. Hence, it is possible to estimate the parameters of those distributions. Known parameters can help with the projection of age-specific fertility rates to the future because this shape can be used as an estimate of the average profile of fertility by age independent of time (parameter a_x of the Lee-Carter model (formula 2).

Figure 1a: 3D empirical age-specific fertility rates of Polish (left) and Czech (right) females.



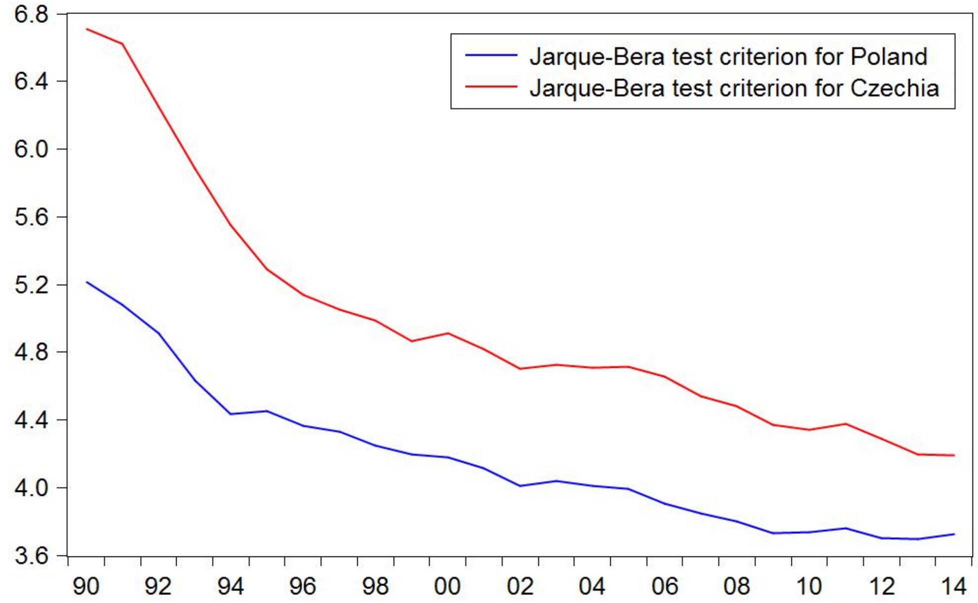
Source: data Eurostat (2015), author's construction and illustration.

Figure 1b: Age-specific fertility rates of Polish (left) and Czech (right) females over time



Source: data Eurostat (2015), author's construction and illustration.

Figure 2: Jarque-Bera test criterion for Poland (blue) and Czechia (red).



Source: author's construction and illustration.

Table 1: Estimates of mean (μ) and variance (σ^2) of the normal shape of fertility function (at 5% level of significance) of Polish (PL) and Czech (CZ) females in the years 2005–2014.

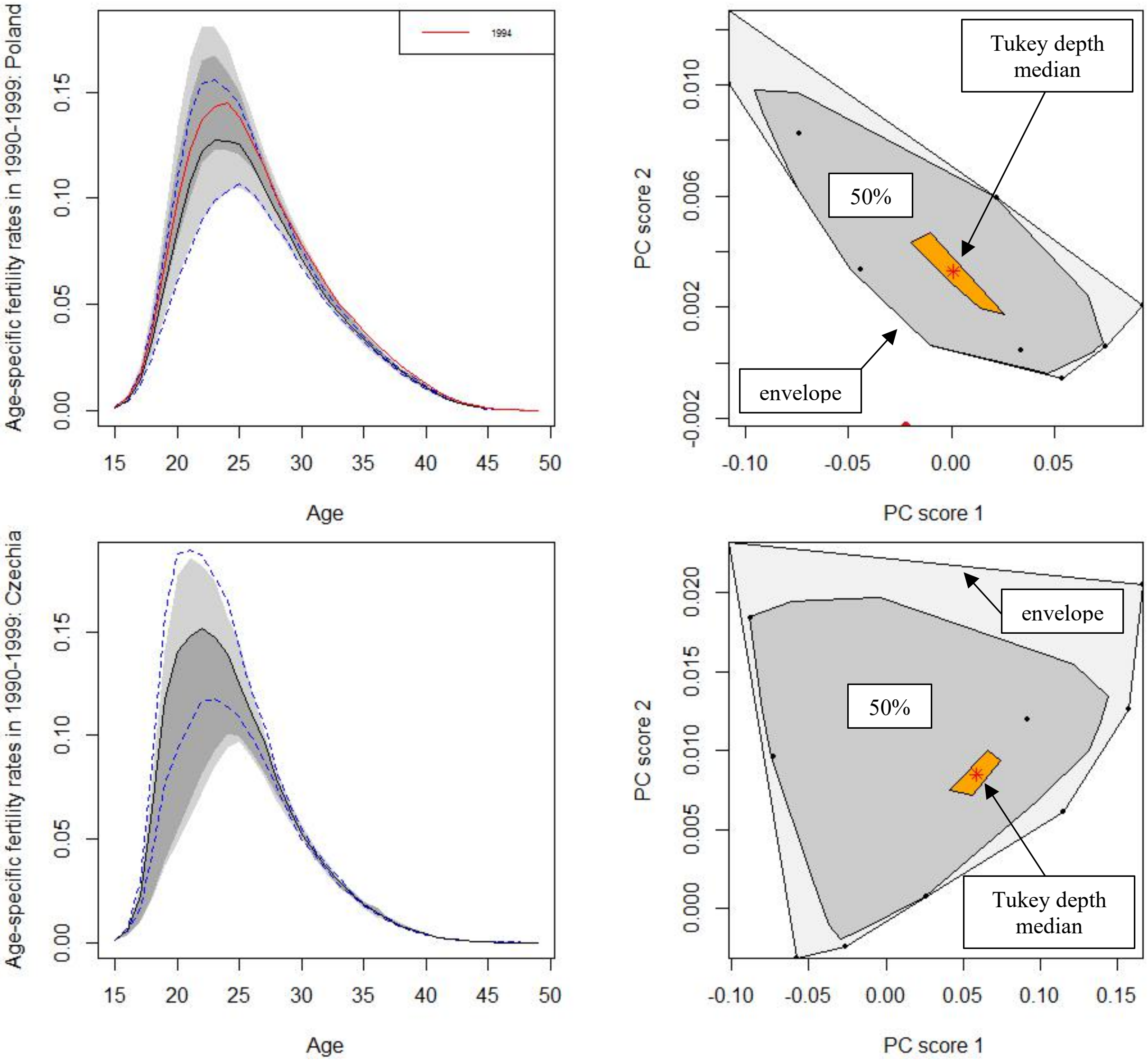
	PL 2005	**PL 2006**	**PL 2007**	**PL 2008**	**PL 2009**	**PL 2010**	**PL 2011**	**PL 2012**	**PL 2013**	**PL 2014**
μ	0.0336	0.0342	0.0353	0.0376	0.0378	0.0382	0.0360	0.0360	0.0348	0.0357
σ^2	0.0011	0.0011	0.0012	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0012
	CZ 2005	**CZ 2006**	**CZ 2007**	**CZ 2008**	**CZ 2009**	**CZ 2010**	**CZ 2011**	**CZ 2012**	**CZ 2013**	**CZ 2014**
μ	0.0348	0.0361	0.0391	0.0409	0.0408	0.0409	0.0385	0.0392	0.0393	0.0413
σ^2	0.0014	0.0015	0.0017	0.0018	0.0018	0.0018	0.0016	0.0016	0.0016	0.0017

Source: author's construction and illustration.

3.2 Shifting of median curves

Second, the analysed period was divided on two 10-years long series (1990–1999 and 2005–2014) based on knowledge of the characteristics of the shape of fertility function.

Figure 3: Principal components method: Age-specific fertility rates in 1990–1999 in Poland and Czechia – functional and bivariate bagplots.

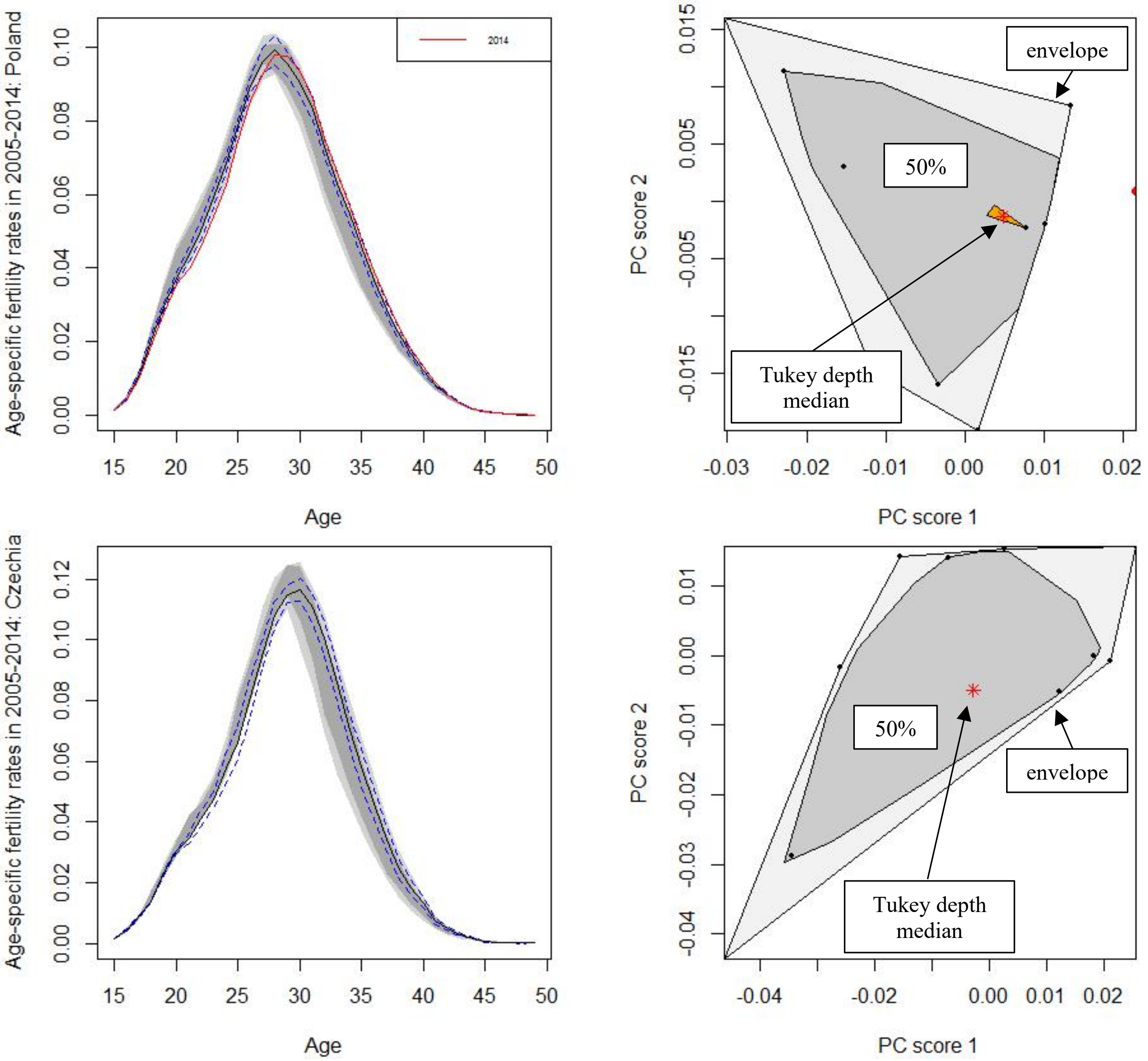


Source: data Eurostat (2015), author's construction and illustration according to Shang, Hyndman (2010, 2016) approach.

A mutual comparison of the estimates of mean (μ) and variance (σ2) of the normal shape of fertility function (at 5% level of significance) of Polish and Czech females in the years 2005–2014 is provided in Tab. 1. Years 2005–2014 can be considered as relatively stable period.

The middle 5-years period (2000–2004) was omitted in both countries because in this period a significant shift of modes to the higher ages occurred and low level of total fertility prevailed. The time series were processed by principal components method to identify outliers according to Hyndman, Shang, (2009) and Hyndman et al. (2017) approach and to find the median curves. Results can be seen in the Fig. 3 for the case of 1990–1999, where Polish population is shown on the top and Czech population on the bottom.

Figure 4: Principal components method: Age-specific fertility rates in 2005–2014 in Poland and Czechia – functional and bivariate bagplots.



Source: data Eurostat (2015), author's construction and illustration according to Shang, Hyndman (2010, 2016) approach.

The outliers of age-specific fertility rates (year 1994) were identified only in Poland. This remoteness is caused by slight deviation in the higher age groups (see top charts in Fig. 3). Shape of fertility function was skewed to the left during 90s in both populations. The mode was 23 years in the case of Polish and 22 years in the case of Czech females. According to estimated median curves it is possible to conclude that Polish level of total fertility rate declined more significantly than in Czech. However, it is obvious that the Czech data are more variable (changes are more significant). It is evident in both – in higher variation range of the confidence interval (left chart on the bottom of Fig. 3) and in larger areas of displayed regions of PC scores (right chart on the bottom of Fig. 3).

The most recent data (already summarized in the Tab. 1) are displayed at Fig. 4 in the functional and bivariate bagplots. The algorithm assessed year 2014 as an outlier in the case of Poland, but it is a minor deviation only caused by a slightly lower fertility level of females aged 21–26 years. Distribution of fertility in both countries is very similar, normal and has minimal variability. This development can be extrapolated to the future, because it represents a stable and low level of fertility typical for Western European countries.

4. Concluding remarks

The last 10 years of age-specific fertility rates in Poland (5) and Czechia (6) can be used to create an estimation of parameter a_x of the Lee-Carter model with parameters

$$\approx \mathrm{N}\left(\mu = 0.0391; \sigma^2 = 0.0016\right) \text{ and} \tag{5}$$

$$\approx \mathrm{N}\left(\mu = 0.0359; \sigma^2 = 0.0012\right). \tag{6}$$

In other words, those parameters can be used instead of classical average age-specific fertility profile independent of time

$$a_x = \frac{\sum_{t=1}^{T} f_{x,t}}{T}, \tag{7}$$

which is not robust, because it is easily influenced by outliers (see e.g. Zhang et al. 2011). Using explicitly specified parameters μ and σ^2 based on expert judgment eliminates, the possibility of bias of this profile is minimized. Hence, the predictions should have a more realistic development. Moreover, the predictions can be divided into several time intervals with possibility to select different parameters μ and σ^2 for each interval.

The prediction is illustrated on an example. To sketch the future development of the mean and variance of age-specific fertility rates we use the Eurostat database and the results of the study EUROPOP 2015 (Eurostat, 2015). The results for Poland and the Czechia for the period 2015–2060 (in 5-year time points only) are displayed in Tab. 2.

Table 2: Forecasted means (μ) and variances (σ^2) of shape of fertility functions for Polish (PL) and Czech (CZ) females up to the year 2060 in 5-year time periods.

	PL 2015	**PL 2020**	**PL 2025**	**PL 2030**	**PL 2035**	**PL 2040**	**PL 2045**	**PL 2050**	**PL 2055**	**PL 2060**
μ	0.0356	0.0392	0.0409	0.0421	0.0429	0.0436	0.0441	0.0446	0.0450	0.0454
σ^2	0.0012	0.0015	0.0016	0.0017	0.0017	0.0017	0.0017	0.0017	0.0017	0.0018
	CZ 2015	**CZ 2020**	**CZ 2025**	**CZ 2030**	**CZ 2035**	**CZ 2040**	**CZ 2045**	**CZ 2050**	**CZ 2055**	**CZ 2060**
μ	0.0425	0.0453	0.0463	0.0469	0.0474	0.0477	0.0480	0.0482	0.0484	0.0487
σ^2	0.0020	0.0024	0.0025	0.0026	0.0026	0.0026	0.0025	0.0024	0.0024	0.0023

Source: author's construction and illustration based on data by Eurostat (2015).

Many types of software (e.g. RStudio, Statgraphics Centurion, SAS, etc.) can generate functions based on specified parameters. If these parameters are modelled based on expert scenarios, they can be used instead of the universal parameter a_x of the Lee-Carter model which is commonly used constant in the extrapolation over time. Šimpach, Langhamrová (2014a, 2014b) were able to predict mortality of the population based on the estimated parameter a_x, but in the case of mortality is this parameter stable over time. For the case fertility, the variable distribution of the a_x parameter seems to be an innovative solution. This paper showed that estimated parameters of fertility functions can be used for the purpose of demographic forecasting.

Acknowledgements

The paper was financed from Internal Grant Agency of the University of Economics Prague No. 35/2017 "Demographic models in R software".

References

[1] Chan, T. M. 2004. An optimal randomized algorithm for maximum Tukey depth. In: SODA '04 Proceedings of the fifteenth annual ACM-SIAM symposium on Discrete algorithms, New Orleans, Louisiana, Jan. 11-14, 2004, pp. 430–436.

[2] Eurostat 2015. EUROPOP 2015: Population projections 2015 at national level [on-line]. URL: http://ec.europa.eu/eurostat/data/database

[3] Hyndman R. J., Shang, H. L. 2009. Forecasting functional time series (with discussion). Journal of the Korean Statistical Society, vol. 38, no. 3, pp. 199–221.

[4] Hyndman R. J., Shang H. L. 2010. Rainbow plots, bagplots, and boxplots for functional data. Journal of Computational and Graphical Statistics, vol. 19, no. 1, pp. 29–45.

[5] Hyndman, R. J., Booth, H., Tickle, L., and Maindonald, J. 2017. demography: Forecasting Mortality, Fertility, Migration and Population Data. R package version 1.20. URL: https://CRAN.R-project.org/package=demography/.

[6] Jarque, C. M., Bera, A. K. 1987. A test for normality of observations and regression residuals. International Statistical Review, vol. 55, no. 2, pp. 163–172.

[7] Lee, R. D., Carter, L. R. 1992. Modeling and forecasting U.S. mortality. Journal of the American Statistical Association, vol. 87, pp. 659–675.

[8] 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.

[9] R Core Team. 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

[10] RStudio Team. 2015. RStudio: Integrated Development for R. RStudio, Inc., Boston, MA. URL: http://www.rstudio.com/.

[11] Shang, H. L., Hyndman, R. J. 2010. Exploratory graphics for functional data. Working paper of the Department of Econometrics and Business Statistics, Monash University, Clayton, Australia, Aug. 3, 2010, pp. 1–9.

[12] Shang, H. L., Hyndman, R. J. 2016. rainbow: Rainbow Plots, Bagplots and Boxplots for Functional Data. R package version 3.4. URL: https://CRAN.R-project.org/package=rainbow/.

[13] Šimpach, O., Langhamrová, J. 2014a. Development of Socio-Economic Indicators and Mortality Rates during Ten Years of the CR Membership in the EU. In: Proceedings of the 2nd International Conference on European Integration (ICEI 2014). Ostrava: VŠB TU, pp. 675–683.

[14] Šimpach, O., Langhamrová, J. 2014b. Stochastic Modelling of Age-specific Mortality Rates for Demographic Projections: Two Different Approaches. In: Mathematical Methods in Economics 2014. Olomouc: Palacký University in Olomouc, pp. 890–895.

[15] Tukey, J. W. 1975. Mathematics and the picturing of data. In: R. D. James (ed.), Proceedings of the International Congress of Mathematicians, vol. 2, Canadian mathematical congress, Aug. 21-29, 1974, Vancouver, pp. 523–531.

[16] Zhang, J., Xanthopoulos, P., Tomaino, V., and Pardalos Panos, M. 2011. Minimum Prediction Error Models and Causal Relations between Multiple Time Series. Wiley Encyclopedia of Operations Research and Management Science, vol. 5, pp. 3271–3285, John Wiley & Sons Inc.