DETECTION OF OUTLIER AGE-SPECIFIC MORTALITY RATES BY PRINCIPAL COMPONENT METHOD IN R SOFTWARE: THE CASE OF VISEGRAD FOUR CLUSTER

Ondřej Šimpach

Abstract

The empirical studies have shown many conclusions about the similar development of many economic and social statistics of the Visegrad Four countries (V4). This dependence is partly due to a similar history of these countries and currently as well due to the joint cooperative policy. This is the reason why are these countries included into one cluster in many analyses and thus evaluated together as one group. In this paper we will focus on the development of age-specific mortality rates in this cluster of V4 countries. On the basis of the modern approach for detection the outliers in R software there will be shown the possibilities how to detect the outlying years using the Principal Components method with the programmed code. We will show and capture the years in which the development of these rates were statistically significantly different. For the practical application we used the data about the numbers of *x*-year-olds deaths by sex and the exposure to risk from the Human Mortality Database (for the mentioned countries, i.e. the Czech Republic, Slovakia, Hungary and Poland). Support package for the analysis in R software is the "demography" and "rainbow". The results will be compared with each other and there will be discussed the relations and mutual consequences.

Key words: Visegrad Four cluster, R software, Principal Component method, age-specific mortality rates, detection of outliers

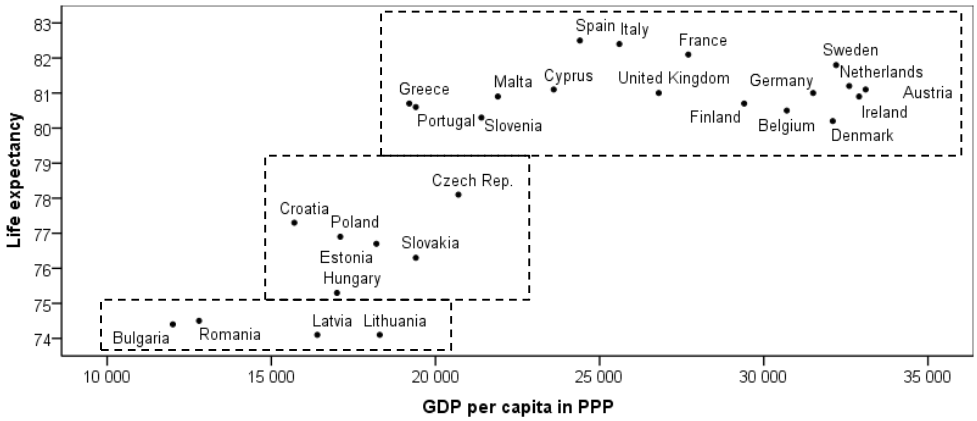
JEL Code: C38, C87, J11

Introduction

The quality of life of people in the European Union is evaluated by Eurostat on the basis of selected statistics from the field of national economic indicators, health, education (see Marek, 2013, Marek, Vrabec, 2013 or Šimpach, Langhamrová, 2014a) and also demographics (e.g. Fiala, Langhamrová, 2013 or Pivoňka, Löster, 2013a). There are groups of countries that develop in similar trend, because they are coupled, for example, by united foreign and economic policy (e.g. Pechrová, 2013 or Šimpach, Langhamrová, 2014b), by historical roots and

historical development or by social mentality of the people (e.g. Řezanková, Löster, 2013 or Bartošová, Želinský, 2013). Let's look at the clusters of countries of the European Union (except Luxembourg, which is outlying observation) in the scatter plot, depending on the selected evaluation criteria (Löster, Pavelka, 2013). On the Fig. 1 is shown the dependence of life expectancy at birth (without distinction of sex) vs. GDP per capita in purchasing power parity (PPP). The Visegrad Four countries (Czech Republic, Slovakia, Hungary and Poland) are in this cluster together (in addition with the countries of Croatia and Estonia).

Fig. 1: Quality of life evaluated by the selected criteria according to Eurostat methodology



Source: Eurostat, author's illustration in IBM SPSS Statistics

It is expected the growing future trend in the development of life expectancy of these countries (Miskolczi et al., 2011 or Pavelka et al., 2014). It should catching up the developed (especially the northern and western European countries), as well as the development of GDP per capita (Pivoňka, Löster, 2013b). The mortality rate is closely related with the evolution of life expectancy (Dotlačilová, Šimpach, 2013 or Šimpach, Dotlačilová, Langhamrová, 2013). If we estimate the future development of life expectancy it is necessary to analyse the development of age-specific mortality rates and predict their future trend. Therefore, we focus our attention on the cluster from Fig. 1, in which are included the countries of the Visegrad Four (V4). In this paper we will find the years in which are the age-specific mortality rates significantly distant from others using the RStudio (R Development Core Team, 2008) and Principal Component method, and data matrix from the Human Mortality Database (number of deaths *x*-year-old and the exposure to risk for the population of males, females and total). These identified years may affect the predicted trend of these rates and the resulting population projections could be biased. (More information about outliers in multidimensional statistical methods provided e.g. Řezanková et al., 2011, Löster, Langhamrová, 2012 or Marek, Vrabec,

2014). It would be useful to balance the found years by particular levelling function or even not consider in the analysis (Erbas et al., 2012), (if their remoteness is really significant).

1 Materials and Methods

For the analysis there will be used the data from the Human Mortality Database. For case of the Czech Republic, Slovakia and Hungary there are available the observations from 1950 to 2009 in a one-year age groups, for Poland we have unfortunately shorter database (from 1958 to 2009). Using the numbers of *x*-year-old deaths in the year *t* and the country *c*, exposure to risk at age *x*, time *t* and the country *c*, there will be established the *BASE* (Hyndman, 2012) in software RStudio (R Development Core Team, 2008). The *BASE* will be filled by logarithms of age-specific mortality rates for each of the analysed countries as

$$BASE \leftarrow \ln\left(m_{x,t}^{c}\right) = \ln\left(\frac{D_{x,t}^{c}}{E_{x,t}^{c}}\right). \tag{1}$$

This *BASE* of logarithms is established due to the reason that this structure together with a package "demography" and "rainbow" (Hyndman, Shang, 2009 and Hyndman, 2012) is used for stochastic modelling of mortality by Lee-Carter model (Lee, Carter, 1992)

$$\ln\left(m_{x,t}^{c}\right) = a_{x}^{c} + b_{x}^{c}k_{t}^{c} + \varepsilon_{x,t}^{c}, \tag{2}$$

where a_x^c are the age-specific profiles independent of time, b_x^c are the additional age-specific components determine how much each age group changes when k_t^c changes and finally k_t^c are the time-varying parameters - the mortality indices. Using the principal components method, which can be acquired using the packages "demography" and "rainbow", we are able to identify the outlying years of the analysed countries and populations. These will be graphically displayed based on the results of the PC scores (Hyndman, Shang, 2009).

2 Results and Discussion

Shang, Hyndman (2010) introduced the functional and bivariate bagplots for clear visualization of outliers in functional data. We use their approach to identify the outliers in the development of age-specific mortality rates of the analysed populations. The functional and bivariate bagplots always contain two regions. One region is dark grey, the other one light grey. Dark grey region contains 50 % of all observations while also there is a black median curve. In the surroundings of this median curve there are located 2 dashed lines, which are 95% confidence intervals. As noted Shang, Hyndman (2010): "functional curves that are outside the fence region are considered outliers". These curves are in functional bagplot shown in colour. The bivariate

bagplot does not show the median curve, but Tukey depth median (see e.g. Tukey, 1975). Coloured points with year label outside the fence region are outliers. For the first time we look at the population of the Czech Republic (see Fig. 2 in Appendix). The male population has outlying years 1950, 1951 and 1958. It is a post-war period and the beginning of the communist regime in the former Czechoslovakia, but this remoteness is rather determined by the volatility of mortality in the highest age groups. In the case of female population is the situation similar. For the outlying years we consider 1950– 1952 and additionally 2000. Period of 50s were more volatile, but the outlying year 2000 is due to other reasons. In the highest age groups there was a significant decline. This is probably caused by random error in data. If we consider the population without distinction of sex, all identified outliers are from the 50s (years 1950, 1951 and 1954–1956). Higher values of age-specific mortality rates are particularly evident in the age groups of 5–45 years and in the highest age groups. There are much more visible the detected outliers in the Slovakia (see Fig. 3 in Appendix). In the case of the male population are the outlier years 1950 and 1951, while the significantly higher values are the rates in the age group of 5–35 years. Females have much more outliers, these are the years 1950–1952 and also 1964, 1966 and 1969. Outlying years of the first group are characterized by higher values of age-specific mortality rates in the age groups 5–65 years. The second group has a high variability of the data in the highest age groups. In the case of Slovak population in total there are the outlying years only from the beginning of 50s (i.e. 1950–1953). This is caused by higher mortality rate of the total population in the age group of 5–45 years. The Hungarian population is similar to the development in the Czech Republic and Slovakia. However, there will be probably the one interesting difference. Looking at the Fig. 4 in Appendix, we can see that at the beginning of 50s there were the outliers for male, female and the population without distinction of sex. Males have the outlying years 1950–1952 and 1956, females 1950–1952 and the total population also 1950–1952. But in the case of male population and the total population there was identified as outlier also the year 2009. In this year there were the age-specific mortality rates significantly lower than in other years. This was influenced especially by the male population in the age group of 0–40 years. Last Polish population was analysed only from 1958 because for the earlier years there are not available data. If the data were available, it probably would been identified as outliers the years from the beginning of 50s. In the case of male population (see Fig. 5 in Appendix) are outlier years 1959 and 1991. From this development is clear that there is no significant remoteness. In 1991 there were slightly higher the age-specific mortality rates in the age groups of 40–70 years. For female population there was practically no outlier observation observed. The functional bagplot identified no

observation, the bivariate one found the year 1952. We see that this value is only slightly behind the border of remoteness. In the case of population without distinction of sex there are recorded the outliers from the end of 50s, namely 1958 and 1959.

Conclusion

Based on the methodology above, we identified the outliers' age-specific mortality rates of the population of males, females and total in the Czech Republic, Slovakia, Hungary and Poland (V4 countries cluster). The results indicate that if we carried out the projection of these rates, it would be suitable to fit these rates in which were identified the remoteness in the highest age groups. To fit these rates can be used some of the known levelling function (e.g. Gompertz–Makeham, Kanistö, Thatcher or Coale–Kisker). With fitted values it increases the accuracy of the calculated mortality projection.

Acknowledgment

This paper was supported by the Internal Grant Agency of University of Economics Prague under the No. IGS MF/F4/6/2013 called "Evaluation of the results of cluster analysis in economic issues".

References

BARTOŠOVÁ, J., ŽELINSKÝ, T. (2013). Extent of Poverty in the Czech and Slovak Republics Fifteen Years after Split. *Post-Communist Economies*, vol. 25, no. 1, pp. 119-131.

DOTLAČILOVÁ, P., ŠIMPACH, O. (2013). Vybrané logistické modely používané pro vyrovnávání a extrapolaci křivky úmrtnosti a vliv použitých modelů na hodnotu střední délky života. In: *Hradecké ekonomické dny 2013/I*. Hradec Králové: Gaudeamus, pp. 110–114.

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.

FIALA, T., LANGHAMROVÁ, J. (2013). Vývoj ekonomického a sociálního zatížení a stárnutí populace [Development of Economic and Social Dependency and Population Ageing]. *Politická ekonomie*, vol. 61, no. 3, pp. 338-355.

HYNDMAN, R. J., SHANG, H. L. (2009). Forecasting functional time series. *Journal of the Korean Statistical Society*, 38(3), pp. 199-221 (With discussion)

HYNDMAN, R. J. (2012). *demography: Forecasting mortality, fertility, migration and population data*. R package v. 1.16. URL: <http://robjhyndman.com/software/demography/>

LEE, R. D., CARTER, L. R. (1992). Modeling and forecasting U.S. mortality. *Journal of the American Statistical Association*, vol. 87, pp. 659-675.

LÖSTER, T., LANGHAMROVÁ, J. (2012). Disparities between regions of the Czech Republic for non-business aspects of labour market. In: *The 6th International Days of Statistics and Economics*, Slaný: Melandrium, pp. 689-702.

LÖSTER, T., PAVELKA, T. (2013). Evaluating of the Results of Clustering in Practical Economic Tasks. In: *The 7th International Days of Statistics and Economics*. Slaný: Melandrium, pp. 804-818.

MAREK, L. (2013). Some Aspects of Average Wage Evolution in the Czech Republic. In: *The 7th International Days of Statistics and Economics*. Slaný: Melandrium, pp. 947–958.

MAREK, L., VRABEC, M. (2013). Probability Models for Wage Distributions. In: *Mathematical Methods in Economics*. Jihlava: College of Polytechnics Jihlava, pp. 575–581.

MAREK, L., VRABEC, M. (2014). Mixture of Johnson Distribution. In: *Quantitative Methods in Economics (multiple Criteria Decision Making XVII)*. Bratislava: EKONÓM, pp. 162–170.

MISKOLCZI, M., LANGHAMROVÁ, J., FIALA, T. (2011). Unemployment and GDP. In: *The 5th International Days of Statistics and Economics at VŠE, Prague*. Prague: VŠE, pp. 407-415.

PAVELKA, T., LÖSTER, T., LANGHAMROVÁ, J., MAKOVSKÝ, P. (2014). *Vybrané aspekty flexibility trhu práce České republiky [Selected aspects of labor market flexibility Czech Republic]*. Slaný: Libuše Macáková, Melandrium, 147 p.

PECHROVÁ, M. (2013). The Influence of European Union's Subsidies on the Development of Rural Villages in the Czech Republic. In: *The 7th International Days of Statistics and Economics at VŠE, Prague*. Prague: Libuše Macáková, MELANDRIUM, pp. 1100-1109.

PIVOŇKA, T., LÖSTER, T. (2013a). Cluster Analysis as a Tool of Evaluating Clusters of the EU Countries before and during Global Financial Crisis from the Perspective of the Labor Market. *Intellectual Economics*, vol. 7, no. 4, pp. 411-425.

PIVOŇKA, T., LÖSTER, T. (2013b). Clustering of the countries before and during crisis. In: *Löster, T., Pavelka, T. (ed.). The 7th International Days of Statistics and Economics. Conference Proceedings*. Slaný: Melandrium, pp. 1110-1121.

R DEVELOPMENT CORE TEAM. (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: <http://www.R-project.org/>

ŘEZANKOVÁ, H., LÖSTER, T., HÚSEK, D. (2011). Evaluation of Categorical Data Clustering. In: *Advances in Intelligent Web Mastering – 3. Fribourg*, 26.01.2011 – 28.01.2011. Berlin: Springer Verlag, pp. 173-182.

ŘEZANKOVÁ, H., LÖSTER, T. (2013). Shluková analýza domácností charakterizovaných kategoriálními ukazateli [Cluster Analysis of Households Characterized by Categorical Indicators]. *E+M. Ekonomie a Management*, vol. XVI, no. 3, pp. 139-147.

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, August 3, 2010, pp. 1-9.

ŠIMPACH, O., DOTLAČILOVÁ, P., LANGHAMROVÁ, J. (2013). Logistic and ARIMA models in the Estimation of Life Expectancy in the Czech Republic. In: *Mathematical Methods in Economics*. Jihlava: College of Polytechnics Jihlava, pp. 915–920.

ŠIMPACH, O., LANGHAMROVÁ, J. (2014a). Changes in Demographic Structures of ED5-6 Graduates with An Impact on Their Economic (In)Activity. In: *Efficiency and Responsibility in Education (ERiE)*. Praha: Czech University of Life Sciences Prague, pp. 736–743.

ŠIMPACH, O., LANGHAMROVÁ, J. (2014b). 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 2014*. Ostrava: VŠB TU, pp. 675–683.

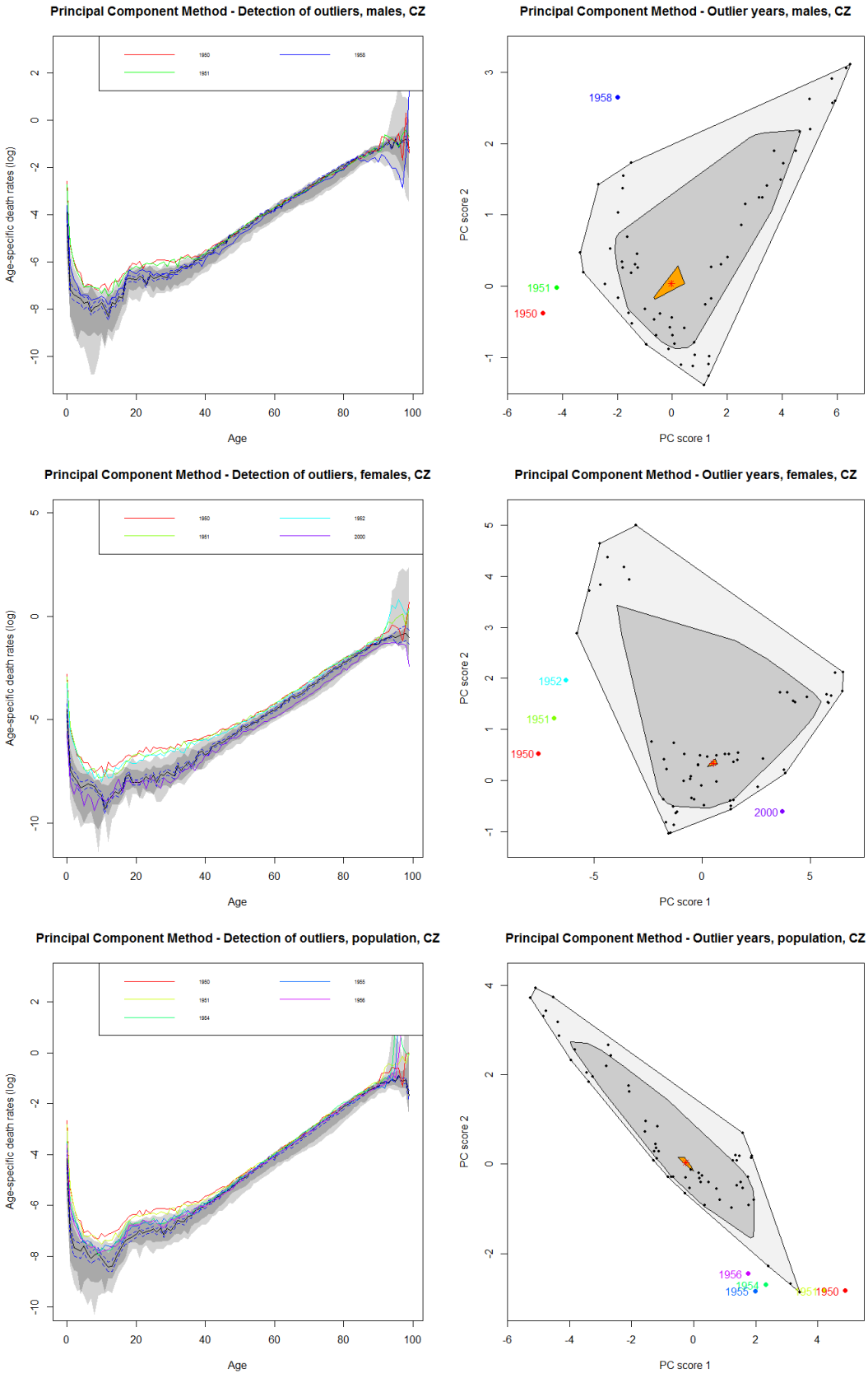
TUKEY, J. W. (1975). Mathematics and the picturing of data. In: *R. D. James, ed., "Proceedings of the International Congress of Mathematicians"*, vol. 2, Canadian mathematical congress, Aug. 21-29, 1974, Vancouver, pp. 523-531.

Contact

Ing. Ondřej Šimpach
University of Economics in Prague, Faculty of Informatics and Statistics
W. Churchill sq. 4, 130 67 Prague 3, Czech Republic
ondrej.simpach@vse.cz

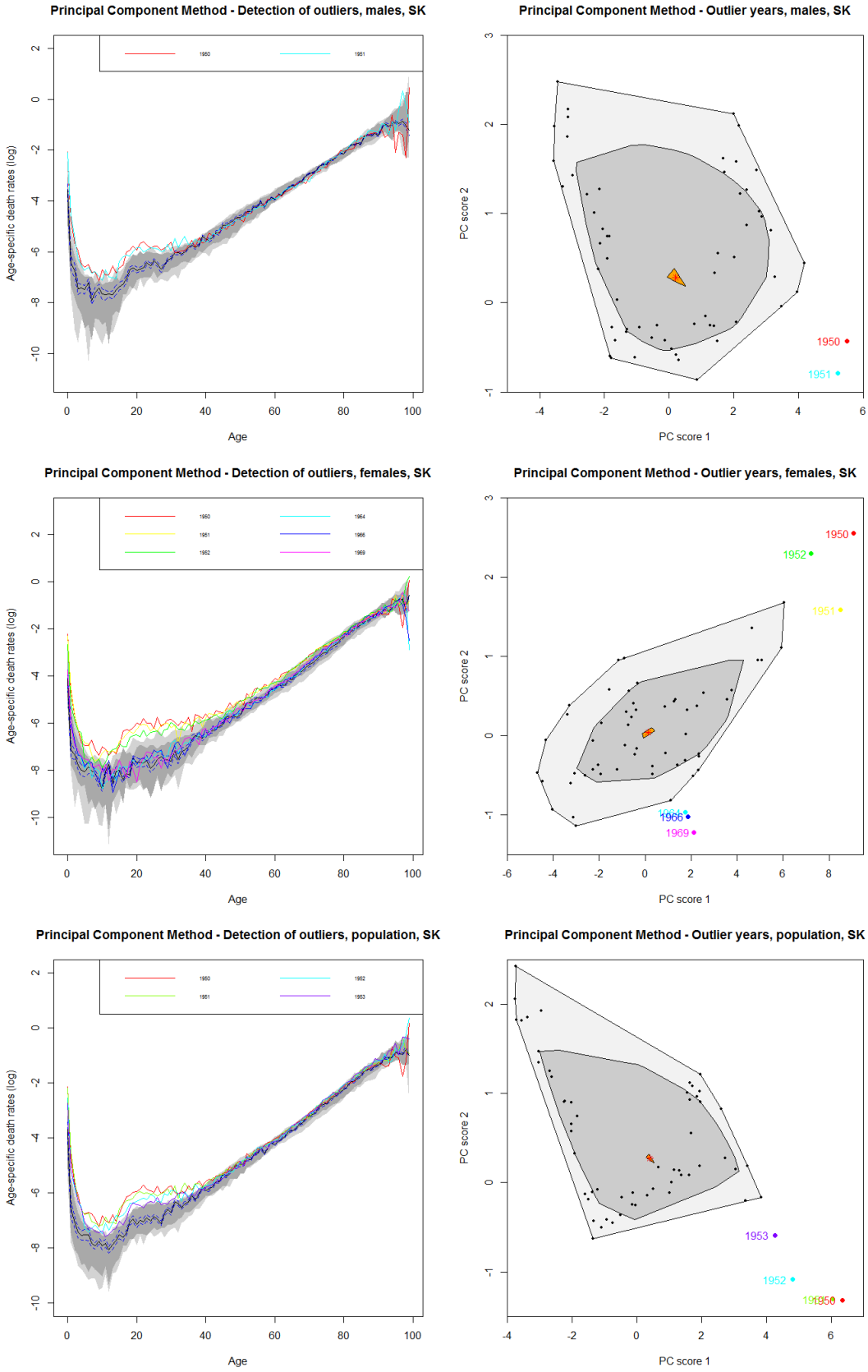
Appendix

Fig. 2: Age-specific death rates (log) in the Czech Republic (males, females and total)



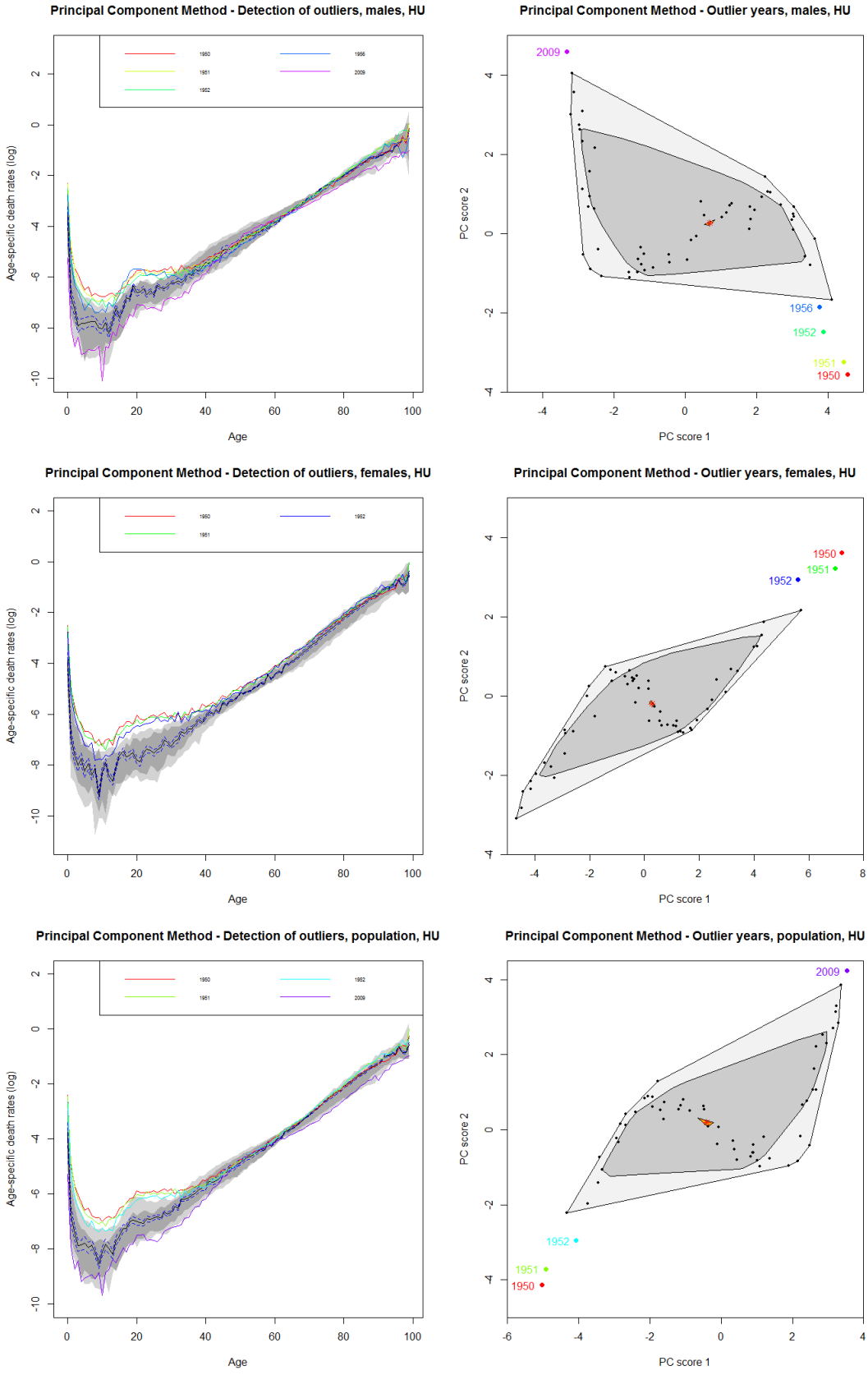
Source: Human Mortality Database, author's calculation and illustration in RStudio by "rainbow" package

Fig. 3: Age-specific death rates (log) in Slovakia (males, females and total)



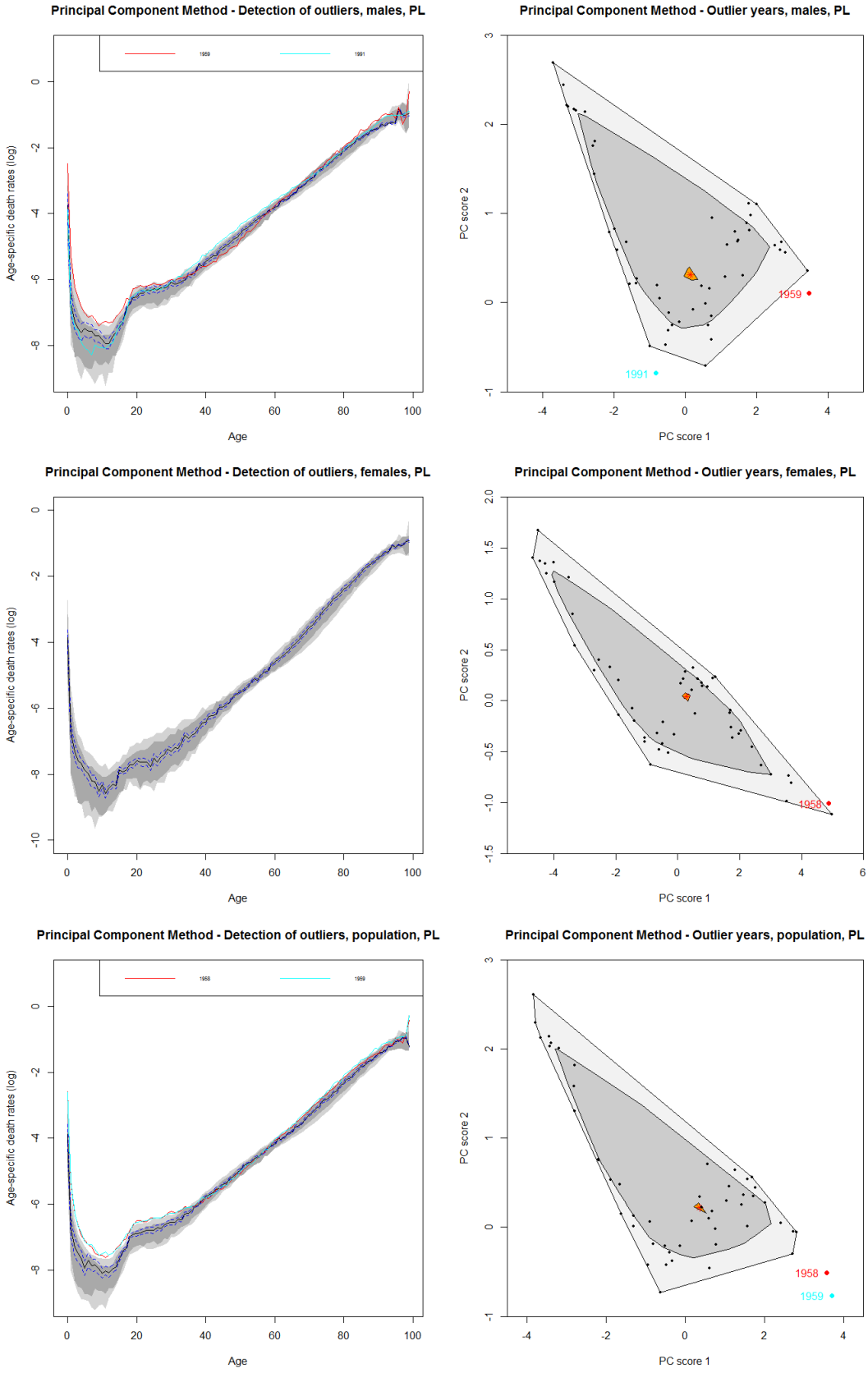
Source: Human Mortality Database, author's calculation and illustration in Rstudio by "rainbow" package

Fig. 4: Age-specific death rates (log) in Hungary (males, females and total)



Source: Human Mortality Database, author's calculation and illustration in RStudio by "rainbow" package

Fig. 5: Age-specific death rates (log) in Poland (males, females and total)



Source: Human Mortality Database, author's calculation and illustration in RStudio by "rainbow" package