FORECASTING MORTALITY IN HIGH AGES USING MIXING MORTALITY DATA METHODOLOGY

TOMÁŠ KAREL, JAN FOJTÍK, MARTIN MATĚJKA, PAVEL ZIMMERMANN, DAGMAR BLATNÁ

University of Economics, Prague, Faculty of Informatics and Statistics,
Department of Statistics and Probability,
W. Churchill Sq. 4, Prague, Czech Republic
e-mail: tomas.karel@vse.cz, xfojj00@vse.cz, martin.matejka@vse.cz, zimmerp@vse.cz, blatnad@vse.cz

Abstract

This article presents one of the possibilities for modelling high age mortality. Lack of data is one of the most important problems in high age mortality modelling. Using multiple data sources may increase precision of estimates. An approach to modelling high age death rates for the Czech Republic using a mixed mortality dataset methodology is presented in this article. The underlying dataset is based on publicly available data provided by the Human Mortality Database, extended by the extinct cohorts methods to improve reliability of very high mortality data. Czech mortality data are mixed with data from surrounding countries with weights taking into account a different exposure of each particular population.

Key words: *high-age mortality, mixed data model, multi-population mortality models, extinct cohorts.*

1. Introduction

Many approaches of mortality modelling have been studied and applied in many fields of science. Different approach is usually applied for adult ages and for high ages (e.g. over 80 years). For adult ages, common models are usually formulated using age specific parameters because, typically, the number of observations is sufficient for valid parameter estimates. For high ages, popular models are usually formulated as one dimensional parametric regression functions of age. Popular examples area for example the logistic model used by Kannisto et al. (1998) or the exponential model assumed by Gompertz (1825). Burcin et al. (2010) provides an extensive list of high age models. Subsequently the growth of mortality is extrapolated using the regression function to very higher ages, where only limited number or even no observation exist.

The major problem of high age mortality modelling is often associated with lack of relevant observations as well as high occurrence of errors in the data. This problem is observed especially in small populations. In this article, we apply the following approach to mitigate these problems. First, data for the population size exposed to risk of death collected for the highest ages were replaced by estimates calculated using the Extinct cohorts method. Second, a method taken over from Ahcan et al. (2014), based on exploiting the information collected for similar populations is applied. Every population has its specifics, but it seems reasonable to find suitable methods to incorporate information collected from areas or countries geographically or economically similar to the population of interest.

In this paper we apply a general linear model to fit the high age mortality in the Czech Republic. Estimates based on original data, data calculated with the Extinct cohorts method as well as on mixed data are compared based on life expectancy at a particular age and price of an in advance whole life annuity.

2. Dataset

The source of all data used in this paper is The Human Mortality Database (Wilmoth et al., 2012). The input data represents a particular cohort, the number of deaths and the population of five countries (i.e. Austria, Hungary, Germany, Czech Republic and Slovakia). These countries were chosen with respect to geographical and social similarities to Czech Republic. In order to estimate the high age model parameters, only the age range from 70 to 95 years was considered. We assume observations for age under 70 years irrelevant for high age modelling, and above the age of 95 years being too inaccurate to be used. To improve the quality of the exposed population data, observations in ages from 81 to 95 years were replaced by corresponding estimates calculated using method of Extinct cohorts. Cohorts between years 1886 and 1901 years, were considered as extinct.

Population exposed to death of a cohort c at age x is denoted $E_{x,c}$, the number of deaths is $D_{x,c}$, and specific death rates are $m_{x,c}$. Where necessary, the upper right script is used to distinguish particular populations. Population indexed with [0] is the population of interest.

3. Extinct Cohorts Methodology

The Extinct cohort method assumes that older population barely migrates or does not migrate at all. Therefore, the only changes in population status in high ages are caused by death. If we sum number of deaths in one cohort backward per ages and do not assume any migration effect, the result gives population states of this cohort (Wilmoth et al., 2012).

Consider a particular cohort c divided in two parts $L_{x,c}$ and $U_{x,c}$, representing triangles of Lexis diagram. The lower triangle $L_{x,c}$ represents number of deaths at age x and time $c + x$ and the upper triangle $U_{x,c}$ stands for number of deaths at age x and time $c + x + 1$. The sum of $L_{x,c}$ and $U_{x,c}$ gives total number of deaths at age x and in cohort c:

$$D_{x,c} = L_{x,c} + U_{x,c}. \tag{1}$$

The sum of the triangle part then leads to estimation of population states $P(x,k)$:

$$P_{x,c} = \sum_{i=0}^{\omega} \left(U_{x+i,c} + L_{x+1+i,c} \right). \tag{2}$$

For the estimated age category from 81 to 95 we put $E_{x,c} = P_{x,c}$.

4. Credibility Mixing of Data

In order to incorporate the information contained in the data from surrounding countries the method for credibility mixing, originally developed in (Ahcan et al., 2014), was applied. First step of this method is to construct a linear combination of the observed age specific death rates of the surrounding countries

$$m_{x,c}^{[AVE]} = \sum_{k=1}^{n} w_k m_{x,c}^{[k]}, \tag{3}$$

where w_k are weights determined by the following optimization problem

$$\arg\left(\min_{w_k} \sum_x \sum_c \left(m_{x,c}^{[0]} - w_k m_{x,c}^{[k]} \right)^2 \right), \tag{4}$$

where

$$\begin{aligned} & w_k \geq 0, \quad k = 1,\ldots,n, \\ & \sum_{k=1}^{n} w_k = 1. \end{aligned} \tag{5}$$

The second step is to calculate specific death rate $m_{x,c}^{[R0]}$ using the following credible estimation

$$m_{x,c}^{[RO]} = m_{x,c}^{[0]} z_{z,c} + m_{x,c}^{[AVE]} \left(1 - z_{x,c}\right), \tag{6}$$

where $z_{x,c}$ is the credible coefficient calculated using the following credibility formula

$$z_{x,c} = \frac{E_{x,c}^{[0]}}{E_{x,c}^{[0]} + \sum_{k=1}^{n} w_k E_{x,c}^{[k]}}, \tag{7}$$

and $E_{x,c}^{[k]}$ is the population size for the *k*-th surrounding country and $E_{x,c}^{[0]}$ is the population size for the Czech Republic.

5. Extrapolation of Specific Death Rates

The specific death rates $m_{x,c}^{[0]}$ of the Czech Republic and the credible specific death rates $m_{x,c}^{[RO]}$ are subsequently modelled for extrapolation to higher ages. It is assumed that specific death rate for individual cohort *c* is described by the logistic function

$$\log\left(\frac{m_{x,c}}{1 - m_{x,c}} \right) = a_c + b_c x, \tag{8}$$

which means that

$$m_{x,c} = \frac{1}{1 + e^{-(a_c + b_c x)}}, \tag{9}$$

where a_c and b_c are parameters to be estimated.

Parameters of the logistic model a_c and b_c are then estimated using maximum likelihood method analogously to Brouhns et al. (2002). It is assumed that the number of deaths $D_{x,c}$ follows Poisson distribution with expected value $m_{x,c,}$ $E_{x,c,}$. Parameters a_c and b_c are then estimated maximizing the following logarithmic transformation of the likelihood

$$l = \sum_{x=70}^{90} \left(D_{x,c} \ln\left(m_{x,c} E_{x,c} \right) - E_{x,t} m_{x,c} - \ln\left(D_{x,c}! \right) \right). \tag{10}$$

6. Numerical Results

Table 1 illustrates the results of data mixing for one particular cohort 1886 in the Czech Republic. E_{EC} denotes number of population exposed to risk of death calculated using the method of extinct cohorts. $E^{[0]}$ denotes number of population exposed to risk of death, for which ages up to 90 years are taken from the Human Mortality Database (HMD) and above 90 years are equal to E_{EC}. $D^{[0]}$ denotes number of deaths taken from the HMD. $D^{[RO]}$ is the number of deaths obtained by multiplying the mixed death rates $m_{x,c}^{[R0]}$ and $E^{[0]}$.

Table 1: The results of data mixing for the particular cohort 1886

Year	Age	$D^{[0]}$	$D^{[RO]}$	$E_{EC.}$	$E^{[0]}$	$\ddot{a}_x^{[0]}$	$\ddot{a}_x^{[R0]}$	$e_x^{[0]}$	$e_x^{[R0]}$
1956	70	1,354	1,216.3	22,327	23,483	1.73	1.85	9.25	9.58
1957	71	1,405	1,280.2	20,895	22,146	1.72	1.84	8.77	9.06
1958	72	1,353	1,278.9	19,597	20,732	1.72	1.83	8.31	8.57
1959	73	1,333	1,314.8	18,200	18,956	1.71	1.82	7.84	8.08
1960	74	1,359	1,251.8	16,868	17,532	1.70	1.81	7.37	7.63
1961	75	1,449	1,361.0	15,555	16,002	1.70	1.80	6.93	7.16
1962	76	1,355	1,259.9	14,043	14,582	1.69	1.79	6.54	6.75
1963	77	1,325	1,277.2	12,733	13,217	1.68	1.78	6.13	6.31
1964	78	1,226	1,229.5	11,424	11,885	1.67	1.77	5.72	5.90
1965	79	1,326	1,256.9	10,134	10,584	1.66	1.76	5.29	5.49
1966	80	1,263	1,200.3	8,819	9,303	1.65	1.74	4.92	5.12
1967	81	1,256	1,187.9	7,628	7,628	1.63	1.73	4.57	4.75
1968	82	1,106	1,047.2	6,422	6,422	1.62	1.71	4.30	4.47
1969	83	1,023	996.4	5,362	5,362	1.60	1.69	4.01	4.17
1970	84	930	848.7	4,384	4,384	1.58	1.66	3.75	3.92
1971	85	789	746.8	3,503	3,503	1.55	1.63	3.51	3.65
1972	86	655	640.2	2,788	2,788	1.52	1.60	3.27	3.40
1973	87	619	549.2	2,157	2,157	1.49	1.55	3.01	3.15
1974	88	456	442.0	1,595	1,595	1.44	1.50	2.84	2.92
1975	89	389	365.6	1,190	1,190	1.37	1.43	2.62	2.70
1976	90	284	267.3	859	859	1.29	1.34	2.44	2.49
1977	91	212	204.9	611	611	1.18	1.22	2.19	2.21
1978	92	165	168.5	422	422	1.02	1.06	1.90	1.89
1979	93	121	122.6	294	294	0.80	0.83	1.57	1.58
1980	94	87	79.8	176	176	0.48	0.50	1.11	1.14
1981	95	55	51.6	105	105	0.00	0.00	0.50	0.50

Source: The authors and the HMD.

Life expectancy, price of the whole life annuity and particular extrapolations to higher ages were compared for both the original data as well as for the data with the adjustments described above to evaluate the impact of the adjustments on several quantities of interest.

Comparison of the life expectancy e_x at the age of 70 is for both datasets and all cohorts displayed in Figure 1. It can be seen that on average, the impact of the data adjustments did not vary significantly over the cohorts. On average, the difference was 0.268 which is around 3.2 %.

Figure 1: The life expectancy at age 70 for both Czech and mixed data

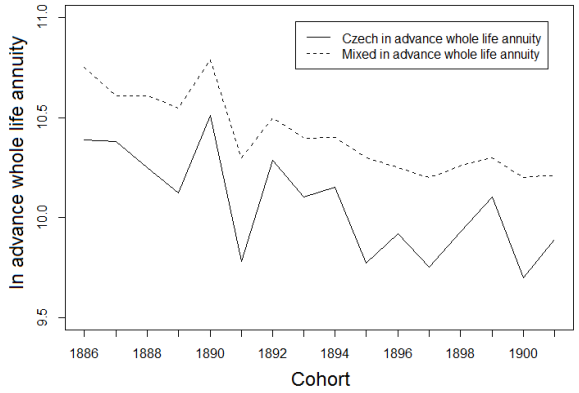
Source: the authors.

Subsequently the in advance whole life annuity for both Czech and mixed data is calculated using the following formula

$$\ddot{a}_x = \sum_{k=1}^{\omega} {}_k p_x (1+i)^{-k}, \tag{11}$$

where ${}_kp_x$ is the probability that a person at age x survives additional k years. The age x=70 years and the interest rate i=*1 %* is assumed.

Figure 2: In advance whole life annuity for both Czech and mixed data



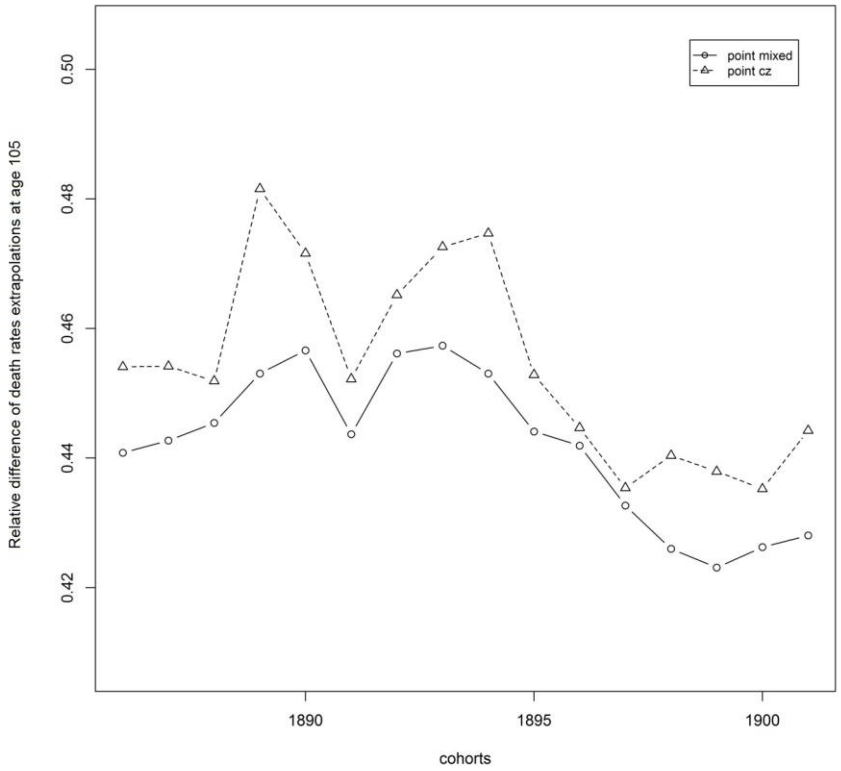
Source: the authors.

It can be seen that, again on average, the impact of the data adjustments didn't vary significantly over the cohorts. On average, the difference was 0.362, which is around 3.6 %.

Finally, extrapolations to age x=105 years assuming the logistic curve were calculated for both data sets. Comparison of the extrapolations for both datasets and all cohorts is displayed in Figure 3.

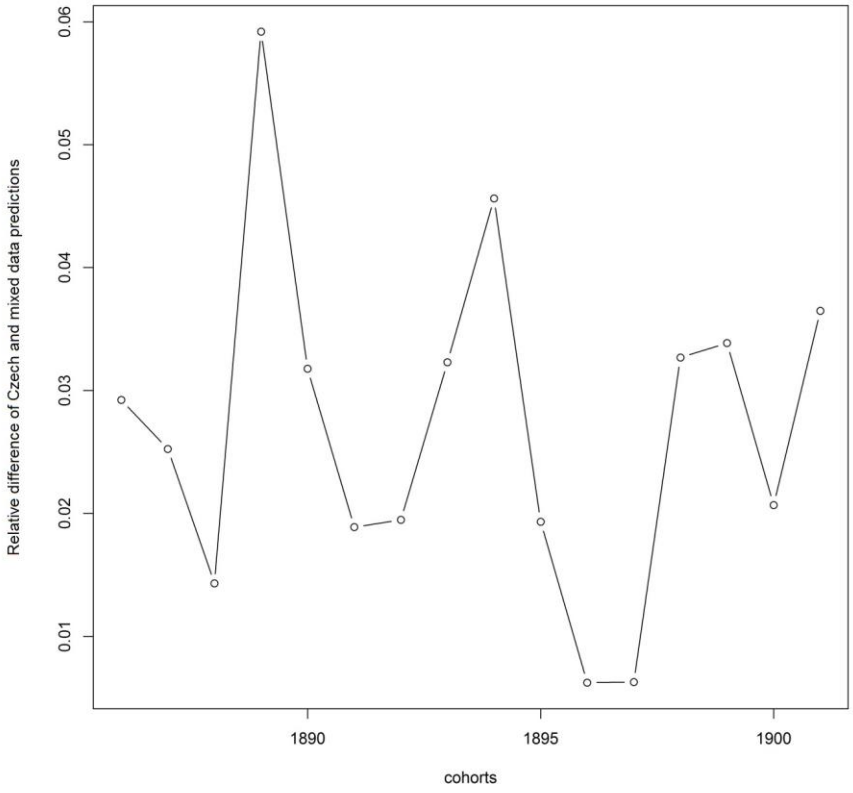
Predictions based on mixed mortality data have significantly smaller variance over the cohorts compared to those obtained by using the unadjusted Czech data. Figure 4 provides development of relative difference of Czech predictions compared to the predictions of model based on mixed data. It can be seen that the impact of the data adjustments on the extrapolations is negligible. On average the difference was 0.123, which is 2.7 %.

Figure 3: Comparison between extrapolations of death rates at age 105 – all examined cohorts



Source: the authors.

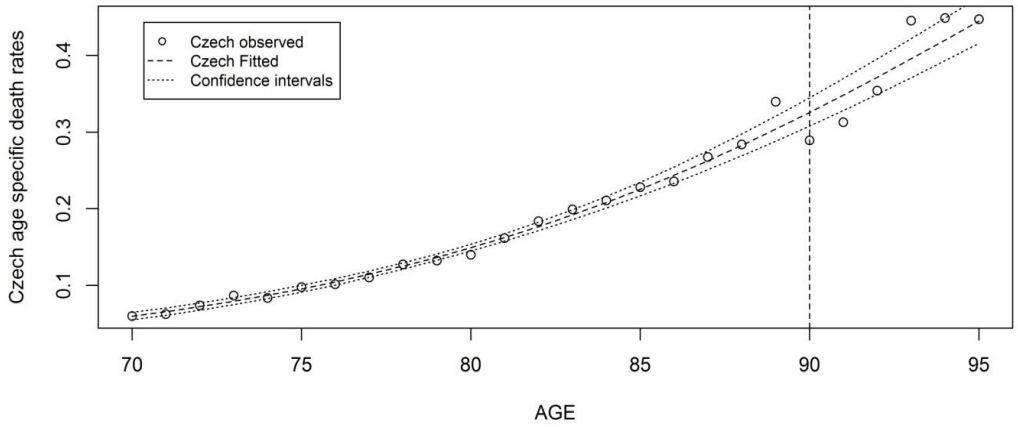
Figure 4: Difference of Czech predictions and mixed data predictions



Source: the authors.

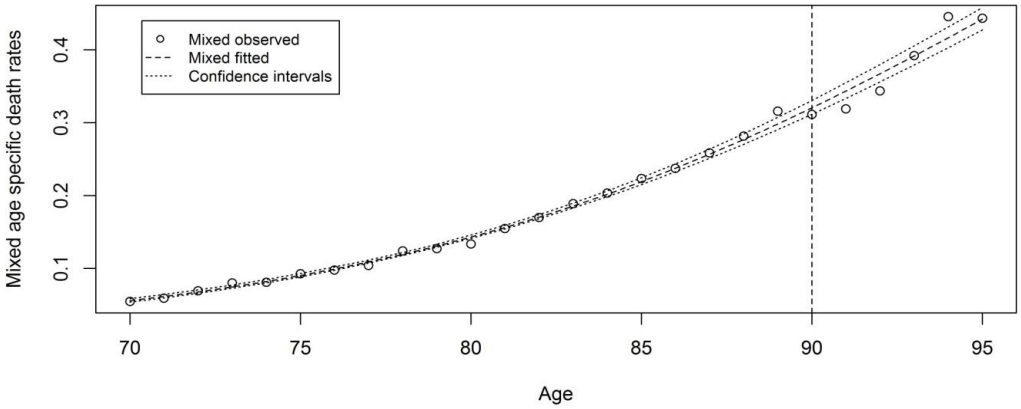
Extrapolated death rates for high ages (Czech dataset) in one particular cohort c=1886 are displayed in Figure 5 and 6. Extrapolations did not change significantly. Significant difference can, however, be observed in the precision of the estimates. The confidence bounds of the extrapolation based on the credibility mixed data are much narrower.

Figure 5: Expected value of death rates for high ages (Czech dataset) in cohort 1886



Source: the authors.

Figure 6: Expected value of death rates for high ages (Czech dataset) in cohort 1886



Source: the authors.

7. Conclusion

The impact of the data adjustments on expected lifetime, in whole life annuity and logistic curve extrapolations was evaluated. Based on the presented results, we can conclude that the data adjustments had minor impact on the life expectancy at the age x=70 years as well as for in advance whole life annuity and also for the extrapolations to the age x=105 years. The average relative impact was for all the quantities calculated approximately about 3 %. The impact of the data adjustments did not vary too much for different cohorts for the life expectancy. Despite the minor impact on the extrapolations to the age x= 105 years there was a significant impact on the confidence bounds of the predictions which were significantly narrower for the extrapolations based on the credibility mixed data.

References

[1] AHCAN, A. et al. 2014. Forecasting mortality for small populations by mixing mortality data. In Insurance : Mathematics and Economics, 2014, vol. 54, pp. 12-27.

[2] BROUHNS, N., DENUIT, M., VERMUNT, J. K. 2002. A Poisson log-bilinear regression approach to the construction of projected lifetables. In Insurance : Mathematics and Economics, 2002, vol. 31, pp. 373–393.

[3] BURCIN, B., TESÁRKOVÁ, K., ŠÍDLO, L. 2010. Nejpoužívanější metody vyrovnávání a extrapolace křivky úmrtnosti a jejich aplikace na českou populaci. In Demografie, 2010, vol. 52, pp. 77-89.

[4] 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. In Philosophical Transactions of the Royal Society of London, 1825, vol. 115, pp. 513–583.

[5] LEE, R. D., CARTER, L. R. 1992. Modeling and forecasting U.S. mortality. In Journal of the American Statistical Association, 1992, vol. 87, pp. 659–671.

[6] R CORE TEAM 2014. R: A language and environment for statistical. Vienna : R Foundation for Statistical Computing, 2014.

[7] THATCHER, A. R., KANNISTO, V., VAUPEL, J. W. 1998. The force of mortality at ages 80 to 120. Odense : Odense University Press, 1998.

[8] WILMOTH, J. R., SHKOLNIKOV, V., BARBIERI, M. 2012. Human mortality database, http://www.mortality.org.