COMPARISON OF THE CALIBRATION OF MORTALITY MODELS ON THE CZECH DATA

Petr Sotona

Abstract

The paper compares well known stochastic mortality models applied on the Czech historical data since 1920. The Comparison is performed for males and females, ages 0 – 95 and for different lengths of observation period (since 1920, 1950 and 1990, respectively). We compare the following models: Lee-Carter model (LC model), Renshaw-Haberman model (RH model), Currie model (APC model), Cairns-Blake-Dowd models (CBD models) and Plat model. For the comparison purposes the new extended models are defined based on the combination of another models. These results are reviewed against the original models.

The first part presents the development of mortality in past and motivates the necessity to capture mortality development with mortality models. This includes short overview of considered models and calibration methods used.

Next, we present qualitative and quantitative measures that are used for the comparison of mortality models. We consider the following measures: quality of fit, reasonableness, parsimony, robustness and completeness of these models. Comparison of mortality models is illustrated using tables and charts presenting the results with the ranking of each model.

Finally, the paper summarizes the results and concludes the main results of the comparison on the Czech mortality data with the recommendation of the most appropriate model.

Key words: model calibration, stochastic mortality models, mortality

JEL Code: C60. J10

Introduction

The important part of the mortality modeling is to understand the observation data of mortality in past. In general we can observe that mortality rates are decreasing in time. This declining trend in mortality rates has many reasons. One of them is the increasing standard of life. The substantial is the level of medicine which improved over time, e.g. particular cures for

previously fatal diseases have been discovered (tuberculosis) or medical treatment is much more successful for some diseases (AIDS, cancer) or some causes of death (heart attack, accidents).

For reasonable modeling of mortality rates we need to understand the development of mortality rates in past and corresponding model has to be able to replicate this development so it can produce also reasonable projection of mortality rates. It is well known that the trend of mortality rates is not the same for all ages. Therefore we aim to compare mortality models on the same set of data to find the most appropriate model.

1 Facts about comparison

1.1 Development of mortality rates

The history of mortality tables started in the middle of the 17th century with the first life (mortality) table. In the middle of the 18th century the first national data began to be collected. Life expectancy at birth in more developed countries in Europe in that time was around 35-40 years. One hundred years later, in the middle of the 19th century, it was around 40-45 years. In the middle of the 20th century it was much higher, around 60-65 years. Currently in some countries the life expectancy at birth is above 80 years. From these numbers we immediately see the huge improvement of life expectancy in past, especially in the 20th century. It is obvious that the level of mortality rates changes over time and we can observe a trend in mortality rates (different for different age and sex) reflecting these developments.

Further description of the development of mortality rates during the 20th century in the Czech Republic is presented in (Sotona, 2014b).

1.2 Considered mortality models and calibration methods

In abstract we listed the mortality models we consider. Basic formulas are presented in Table 1, for further details including complete definition of models, or identifiability constraints we refer to (Lee, Carter, 1992, Renshaw, Habernam, 2006, Cairns et al., 2009, or Sotona, 2014a). We also propose three new models (M8, M9, M10) which are based on the combination of CBD models and general models containing age depending parameter α_x. Basic formulas are presented in Table 1 below and the further details will be presented in doctoral thesis. In this paper we present results only for males (results for females would provide consistent conclusions).

Tab. 1: Overview of mortality models and key characteristics

Mark	Type	Formula	Method[1]	Assumption[2]
M1a	LC model	$\ln m(t,x) = \alpha_x^{(1)} + \beta_x^{(1)}\kappa_t^{(1)}$	MLE	Poisson distribution
M1b			WLS	Normal distribution
M1c			SVD	Normal distribution
M2	RH model	$\ln m(t,x) = \alpha_x^{(1)} + \beta_x^{(1)}\kappa_t^{(1)} + \beta_x^{(2)}\chi_{t-x}^{(2)}$	MLE	Poisson distribution
M3	APC model	$\ln m(t,x) = \alpha_x^{(1)} + \kappa_t^{(1)} + \chi_{t-x}^{(2)}$	MLE	Poisson distribution
M4a	CBD model	$\ln \frac{q(t,x)}{p(t,x)} = \kappa_t^{(1)} + \kappa_t^{(2)}x$	MLE	Poisson distribution
M4b			WLS	Normal distribution
M5a		$\ln \frac{q(t,x)}{p(t,x)} = \kappa_t^{(1)} + \kappa_t^{(2)}(\bar{x} - x)$	MLE	Poisson distribution
M5b			WLS	Normal distribution
M6a		$\ln \frac{q(t,x)}{p(t,x)} = \kappa_t^{(1)} + \kappa_t^{(2)}(\bar{x} - x) + \chi_{t-x}^{(3)}$	MLE	Poisson distribution
M6b			WLS	Normal distribution
M7	Plat model	$\ln m(t,x) = \alpha_x^{(1)} + \kappa_t^{(1)} + \kappa_t^{(2)}(\bar{x} - x) + \kappa_t^{(3)}(\bar{x} - x)^+ + \chi_{t-x}^{(4)}$	MLE	Poisson distribution
M8a	Proposed model	$\ln \frac{q(t,x)}{p(t,x)} = \alpha_x^{(1)} + \kappa_t^{(1)} + \kappa_t^{(2)}x$	MLE	Poisson distribution
M8b			WLS	Normal distribution
M9		$\ln \frac{q(t,x)}{p(t,x)} = \alpha_x^{(1)} + \kappa_t^{(1)} + \kappa_t^{(2)}(\bar{x} - x)$	MLE	Poisson distribution
M9b			WLS	Normal distribution
M10a		$\ln \frac{q(t,x)}{p(t,x)} = \alpha_x^{(1)} + \kappa_t^{(1)} + \kappa_t^{(2)}(\bar{x} - x) + \chi_{t-x}^{(4)}$	MLE	Poisson distribution
M10b			WLS	Normal distribution

Source: Author's calculation

Assumptions and methods presented in table are described further in this paragraph. In general there is not one universal approach for the calibration of models because each mortality model has its specific assumptions, different number of parameters, or different characteristics. We will focus on the calibration of LC and CBD models and compare some of known approaches. The original LC model was calibrated using the least squares method (LSM) with minimization of the objective function as defined for example in (Pitacco, 2009). This can be done by singular value decomposition (SVD) approach or by Newton-Raphson method which is based on recursive formulas. The result is equivalent to the maximum likelihood estimation (MLE) under the condition that error term in LC model has Normal distribution (Pitacco, 2009).

[1] Methods are further explained in the paragraph below table
[2] Assumptions used are further described in the paragraph below table

We further consider extension introduced by (Wilmoth, 1993) called weighted least-squares method (WLS). Last, we consider MLE but we assume that the number of deaths D(t:x) at age x in year t has a Poisson random variation with parameter E(t; x)m(t; x). In this situation the values of parameters are estimated by maximizing the log-likelihood based on the Poisson distributional assumption.

Practical calculation is processed using Newton-Raphson method. For CBD models we consider WLS and MLE to compare calibration approaches.

1.3 Measures used for comparison of models

There are many ways in which we can compare mortality models and we ourselves must decide which properties or criterion are the most important for our purposes. Together with different ways of comparison there are also different sets of desirables characteristics that should good mortality model satisfy, see for example (Cairns et al., 2009). Here we summarize desirability criteria as proposed in (Hunt, Blake, 2014) that should good mortality model fulfill:

- Quality of fit - Provide an adequate fit to the data, with sufficient terms to capture all the significant structure in the data.
- Reasonableness - Be demographically significant in the sense that each age function can
    - be identifiable with specific biological and socio-economic processes occurring at the ages of interest, and
    - be biologically reasonable.
- Parsimony - Be parsimonious, with the smallest number of terms needed to capture this structure, and with each term using as few parameters as possible.
- Robustness - Be robust, in that parameter uncertainty should be low and small changes in the data should not result in significant changes in the estimates of the parameters and in our interpretation of them.
- Completeness - Span the full age range, with sufficient terms to model the complex shape of and dynamics observed in mortality rates at younger ages.
- Cohort effect - Include cohort effects if justified by the data and allow for these to be clearly distinguished from age/period effects to allow plausible projections of the model.

2 Results of comparison

2.1 Calibration methods

We tested three calibration methods for LC model (M1a, M1b, M1c) and two methods for CBD models (M4a, M4b, M5a, M5b, M6a, M6b) and proposed models (M8a, M8b, M9a, M9b, M10a, M10b).

We compared the results using several measures discussed later in in this paper. Nevertheless, result were not strongly supporting only one of methods and we observed that for various observation periods, sex or mortality models the best results are achieved with various calibration methods. Therefore we do not conclude with the best calibration method and in our further mortality models comparison we present the best achieved results for each model.

2.2 Quality of fit

Due to limited space we do not present graphs showing the quality of fit of particular models to observed mortality rates. Some graphs can be found in (Sotona, 2014b) and the rest will be published in the doctoral thesis. Nevertheless here are the main observations. The most of models fit the observed death rates quite accurately, especially in recent years. However longer fitting period makes the accuracy of fitting more difficult especially to cover volatility movements in higher ages. Next, we observe that CBD models did not fit observed rates at young ages and especially at age 0 and underestimated death rates for higher ages.

First quantitative measure is the attained maximum likelihood of particular models (which have been calibrated using MLE method). Values of loglikelihood functions (without constant parts) for males and three observation periods and the order according to this criterion are shown in the Table 2 below.

Tab. 2: Maximum Likelihood and corresponding ranking of mortality models for males and all observation periods

Model	1990 -2011		1950 -2011		1920 -2011	
	Likelihood	Order	Likelihood	Order	Likelihood	Order
M1a	-5 262 333	4	-15 219 292	4	-24 359 980	4
M2	-5 261 843	1	-15 209 622	1	-24 339 715	2
M3	-5 262 930	5	-15 224 184	7	-24 448 552	7
M4a	-5 291 265	9	-15 590 028	10	-26 078 004	10
M5a	-5 291 265	8	-15 590 028	9	-26 078 004	9
M6a	-5 292 453	10	-15 571 053	8	-25 706 506	8
M7	-5 262 217	2	-15 210 146	2	-24 336 088	1
M8a	-5 262 985	7	-15 220 385	6	-24 364 608	6
M9a	-5 262 985	6	-15 220 385	5	-24 364 608	5
M10a	-5 262 283	3	-15 214 335	3	-24 343 157	3

Source: Author's calculation

According to this criterion the best model is M2 and M7, respectively. CBD models (M4a, M5a, M6a) are the worst ones. One of proposed models, model M10a is on the third place.

Another measure of the quality of fit is MAPE defined for example in (Pitacco, 2009). Table 3 below shows MAPE and results of this comparison for each model

Tab. 3: MAPE and corresponding ranking of mortality models for males and all observation periods

Model	1990 -2011		1950 -2011		1920 -2011	
	MAPE	Order	MAPE	Order	MAPE	Order
M1	9,1%	2	12,2%	3	14,4%	5
M2	8,6%	1	8,8%	1	11,0%	2
M3	10,7%	7	15,4%	7	29,2%	7
M4	19,9%	10	29,5%	10	34,8%	10
M5	19,9%	9	29,5%	9	34,8%	9
M6	17,8%	8	27,0%	8	31,2%	8
M7	9,5%	3	9,3%	2	10,3%	1
M8	10,1%	6	13,4%	6	16,2%	6
M9	9,9%	5	12,4%	4	14,3%	4
M10	9,6%	4	12,4%	4	12,3%	3

Source: Author's calculation

Similarly to the results of maximum likelihood we see that the models M1, M2 and M7 are the best models. New proposed models M8, M9 and M10 are following them. In general models with more parameters provide usually better fit than the models with less parameters. CBD models provide worse results compared to other models due to their construction which is based on the characteristics of only higher ages (let's say above 60) for which these models were created.

2.3 Reasonableness

Reasonableness of the model can be assessed through the level of understanding the values of model parameters and explanation of these values with reasonable biological and / or socio-economic facts.

First observation is that all models show reasonable declining trend in time dependent parameter κ_t which corresponds to a decrease in mortality rates over last twenty, sixty and ninety years, respectively. Second thing is that parameters α_x have also very similar development for all models except for M7 model where we observe different shape. This can

be caused by higher number of parameters and mutual compensation of effects between parameters (result of overparametrisation).

Further, we take into consideration the assumption about Poisson distribution of number of deaths which we have used in MLE method. If the assumption about number of deaths being independent Poisson random variables is true then standardized residuals $z_{t,x}$ will be approximately independent random variables with standard Normal distribution. According to Kolmogorov-Smirnov test on confidence level 95% we reject for all listed models the null hypothesis that these residuals $z_{t,x}$ have standard Normal distribution. Therefore we can conclude that the assumption about Poisson distribution of number of deaths is not supported. However according to (Cairns et al., 2009) this problem is an issue in many countries and this overdispersion does not have significant impact on estimates of the future dynamics of mortality rates. On the other side Poisson assumption could result in underestimation of the future variability of the actual death rates regarding the true underlying rates.

Next, we check the independence and identical distribution (i.i.d.) assumption of standardized residuals $z_{t,x}$ using surface graphs of these residuals. Under this assumption we should not observe any clustering or systematic patterns. Some of these graphs can be found in (Sotona, 2014b). When considering the shorter history 1990 - 2011 these patterns seem quite random except for the M1 model where we observe diagonal clustering which is caused by the lack of cohort parameters in this model. Looking at calibration period 1950 - 2011 we observe more significant clustering in M1 and M3 models. For models M2 and M7 the clustering is not that visible however there is clear worsening for these models as well. For longest period 1920 - 2011 the clustering is obvious for all models but M1c and M3 are clustered significantly. Regarding model M1 we suppose it is caused mainly by the fact of missing cohort parameters and in terms of M3 model we observe diagonal clustering even if this model contains cohort parameters. This could be caused by the low robustness of this model arising from independence of age, period and cohort parameters.

2.4 Parsimony

Taking into account the parsimony of mortality models, first, we compare the models according to the number of parameters. The results are summarized in the following Table 4.

Tab. 4: Number of parameters and corresponding ranking of mortality models for all observation periods

Model	1990 -2011	1950 -2011	1920 -2011

	Parameters	Order	Parameters	Order	Parameters	Order
M1	214	8	254	6	284	5
M2	368	10	448	10	501	10
M3	176	6	256	7	309	6
M4	44	1	124	1	184	1
M5	44	1	124	1	184	1
M6	102	3	222	5	312	7
M7	220	9	380	9	493	9
M8	140	4	220	3	280	3
M9	140	4	220	3	280	3
M10	198	7	318	8	401	8

Source: Author's calculation

We note here that in case of cohort parameters we allow for this parameter only when at least 10 observations are available and at least age 50 is included in that cohort. This supports the idea that cohort effect should be more visible in higher ages than in childhood and low ages. Number of parameters is the lowest in CBD models M4 and M5. However, low number of parameters and structure of CBD models (the purpose of these models is to model mortality rates for higher ages and not for all ages) caused significantly worse quality of fit to observed data. Proposed models M8, M9 are also very good. M2 and M7 are the worst ones

Another comparison is based on BIC (Cairns et al., 2009) which combines the quality of fit expressed by maximum likelihood and the number of parameters in the model. Following Table 5 shows BIC for all models (which were fitted by MLE method).

Tab. 5: BIC and corresponding ranking of mortality models for all observation periods

Model	1990 -2011		1950 -2011		1920 -2011	
	BIC	Order	BIC	Order	BIC	Order
M1	-5 263 152	3	-15 220 395	4	-24 361 270	4
M2	-5 263 251	6	-15 211 569	1	-24 341 992	2
M3	-5 263 604	7	-15 225 297	7	-24 449 956	7
M4	-5 291 433	9	-15 590 567	10	-26 078 840	10
M5	-5 291 433	8	-15 590 567	9	-26 078 840	9
M6	-5 292 844	10	-15 572 017	8	-25 707 892	8
M7	-5 263 059	2	-15 211 797	2	-24 338 328	1
M8	-5 263 153	5	-15 221 341	6	-24 365 880	6
M9	-5 263 153	4	-15 221 341	5	-24 365 880	5
M10	-5 263 041	1	-15 215 717	3	-24 344 543	3

Source: Author's calculation

The best model is different for each observation period and it is caused by combination of various structure of models and increasing number of time dependent parameters with longer

observation period. Nevertheless the best models are M2, M7 and M10. CBD models are on the other side of scale even considering low number of parameters.

2.5 Robustness

Here we compare the model robustness by the stability of model parameters estimated for three time intervals. From results we see that in case of models M1, M2, M3, M8, M9 and M10 values of fitted parameters are quite robust and do not differ significantly for various calibration periods. For CBD models M4 and M5 we observe that values of time dependent parameters coincide for the same calendar years which is a consequence of the fact that calibration of parameters is done for each calendar year separately. Values of M6 model very slightly differ from previous models. The reason is the application of additional constraints on model M6 (models M4, M5 do not have any additional constraint). For model M7 we see significant changes for parameters' pattern which indicates less robustness of these parameters. These changes are probably caused by high number of parameters in this model and mutual dependencies of these parameters.

2.6 Completeness

All models except for CBD models are designed to be used for whole range of ages. CBD models were developed to model mortality rates of annuitants. The fact is supported by poor results. Therefore CBD models (M4, M5, M6) are not complete models.

Further, based on the definition of each model we can easily say which model takes into account cohort effect and which does not. In our comparison the models allowing for cohort effect are M2, M3, M6, M7 and M10. Diagonal dependencies mentioned during testing of standardized residuals for LC model indicate the lack of allowance for cohort effect.

Conclusion

First conclusion is that differences between the models are more visible for longer calibration periods where particular disadvantages clearly influence the results. The most of models fit the observed death rates quite accurately.

It is easy to see that CBD models (M4, M5, M6) did not fit observed rates at young ages and underestimated death rates for higher ages (as a consequence of the incompleteness of CBD models). Therefore we do not recommend these models when whole age range should be modelled. Overall, we conclude that the best model is M2.However on some data it may be

difficult to calibrate it. Model M7 seems to be too complex for Czech data and shows signs of overparametrization. On the other side model M3 is not robust enough and results were not as good in majority of measures. Model M1 is good alternative when we believe that cohort effect is not significant. Calibration is easy and in insurance industry it is broadly used model.

New proposed models leverage on the characteristics of CBD models and at the same time can capture whole age range. In majority of tests the results were very close to top ranking. We believe that models M8 and M9 can be good alternative to model M1 and model M10 can replace M2 in case we want to cover cohort effect.

References

Cairns, A. J., Blake, D., Dowd, K., Coughlan, G. D., Epstein, D., Ong, A., & Balevich, I. (2009). A Quantitative Comparison of Stochastic Mortality Models Using Data From England and Wales and the United States. North American Actuarial Journal, 13(1), 1-35. doi:10.1080/10920277.2009.10597538.

Hunt, A., & Blake, D. (2014). A General Procedure for Constructing Mortality Models. North American Actuarial Journal, 18(1), 116-138. doi:10.1080/10920277.2013.852963.

Lee, R., & Carter, L. (1992). Modeling and Forecasting U.S. Mortality: Comment. Journal of the American Statistical Association, 87(419), 659-671.

Pitacco, E. et al (2009). Modelling longevity dynamics for pensions and annuity business. Oxford: Oxford University Press.

Renshaw, A., & Haberman, S. (2006). A cohort-based extension to the Lee–Carter model for mortality reduction factors. Insurance: Mathematics and Economics, 38(3), 556-570. doi:10.1016/j.insmatheco.2005.12.001

Sotona, P. (2014a). Modely úmrtnosti a jejich srovnání na datech České republiky. Pojistné rozpravy, 32(1), 111-133.

Sotona, P. (2014b). Vývoj úmrtnosti v České republice. Nastane rovnoprávnost pohlaví? Pojistný obzor, 2, 26-28.

Wilmoth. J. R. (1993). Computational Methods for Fitting and Extrapolating the Lee-Carter Model of Mortality change, Technical Report. Berkeley. Department of Demography. University of California.

Contact

Petr Sotona

The Faculty of Informatics and Statistics, University of Economics in Prague

W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic

Petr.Sotona@seznam.cz