Stochastic Modeling of Mortality in Two Population: application to limited data situation

Bamidele, M.A, Adejumo O.A, Ruth Ella B
Department of Statistics, University of Ilorin, PMB1515, Ilorin, Nigeria
Department of Economics, University of Ilorin, Ilorin, Nigeria
Email: bamidelemoyo@gmail.com; oadejumo@gmail.com

Abstract

The appropriateness of mortality models varies from country to country depending on the prevailing parameter of the country. The deficiencies are sometimes introduced via model misspecification by estimating parameters that are not peculiar to a particular country, limited data for countries lacking adequate historical mortality rates for model building and through unforeseen variation like wars and epidemics. The aim of this paper is to extend stochastic mortality model to capture mortality situation in two populations with application to limited data situation. Data were generated from binomial distribution using Monte-carlo simulations; the mortality rate for age 0 to 100years and 3-year points to capture the limited data condition and the model were applied to Nigeria mortality dataset. Bayesian Information Criterion (BIC), Deviance and Log-likelihood were used to assess the performance of the models. The result from the parameter estimates shows that AP model capture the age effect better compared to all other models. Though ASPC has a better fitting result compared to AP, but AP estimates the age effect better. The period effects were not adequately captured by all the models due to fewer number years' data that is available. The confidence intervals of the forecast from the ARIMA models were very wide indicating that the future pattern of mortality is not stable but over time might come to equilibrium while the cohort effect estimate from the Nigeria was not significant. The findings of from this study have shown the appropriate of AP and ASPC model in modeling mortality in two-population situation.

Keyword: Modeling, Stochastic, Mortality

1. Introduction

The study of mortality relates to the survival and death of individuals within a particular population. The future development of human mortality, together with its wider implications, has attracted increasing interest in recent decades. Modeling mortality data has been the major focus of many authors in actuarial science. The major concern of most authors within the context of mortality is to establish a model that will adequately fits the historical data. In recent years a limited number of studies have explored the use of multi-population models for mortality modeling (see for example [Cairns et al., 2011b], [Cairns et al., 2011c] and [Lee and Li, 2005]). These references jointly model mortality rates from multiple populations, accounting for interactions between both populations. This is in line with 'biological reasonableness' that multiple populations do not necessarily behave independently of each other, as is the case in single-population mortality models. The aim of this paper therefore is to extend the univariate model proposed by Bamidele et al (2014) to bivariate case.

The rest of the paper is structured as follows; section 2 reviews existing literature, section 3 explores multi-population models, while the fitting methodology, stimulation study, result and discussion were presented in section 4,5,6 and 7 respectively.

2. Literature Review

As noted by most authors, the following literature guides on stochastic models that are often used for modeling mortality force $\mu_{x,t}$ or mortality risks $q_{x,t}$; Lee and Carter (1992), Renshaw and Haberman (2006), Cairns et al (2006a), Currie et al (2004) and Currie (2006).

2.1 *Stochastic Mortality Models*

Stochastic mortality models are often used to models mortality force $\mu_{x,t}$ or mortality risk $q_{x,t}$.

The mortality rate or mortality force is often defines as;

$$\mu_{x,t} = \frac{D_{x,t}}{E_{x,t}} \qquad (1)$$

Where $D_{x,t}$ is the number of death at age x during calendar year t and $E_{x,t}$ is the mid-year population of age x during calendar year t.

The initial mortality rate $q_{x,t}$ is the probability that a person aged x dies within the next year. The different mortality measures are linked by the following approximation:

$$q_{x,t} = 1 - \exp(-\mu_{x,t}) \qquad (2)$$

Based on fact that some of the existing models considered in literatures suffers from one problem or another, which well means there is no universal model to model mortality force or risk as the case may be. Bamidele et al (2014) proposed the following models considering its simplicity, easy estimation and relevant parameter in the target population.

Age-Period model (AP)

$$\log(\mu_{x,t}) = \alpha_x + \kappa_t(x - \bar{x}) \qquad (3)$$

Where α_x and κ_t is defined as the age and period effect. The basis for the above model is that in the absence of adequate information apriori the estimation, the component of the data structure should be the first thing to estimate. The Additional variable $(x - \bar{x})$ where $\bar{x}$ is the mean age, tends to adjust for the trivial correlation between period and age. Also, as noted by Currie (2013) that cohort effects is not independent of age and year of death effects, which thus implies that the additional information from cohort is just a linear combination of age and year which might introduce collinearity or confounding to the estimation thus making estimation difficult.

Age Specific-Period Cohort Model (APSC)

$$\log(\mu_{x,t}) = \alpha_x + \kappa^1{}_t + \kappa^2{}_t(x - \bar{x}) + \gamma_{t-x} \qquad (4)$$

All the models considered in this paper except CBD suffers from identifiability problem, thus estimation of the models considered depends on the following constraints;

$$\sum_{t=1}^{n_y} \kappa_t = 0 \qquad (5)$$

$$\sum_{c=1}^{n_c} \gamma_c = 0 \qquad (6)$$

$$\sum_{c=1}^{n_c} c\gamma_c = 0 \qquad (7)$$

where $c = t - x$ runs from 1 (youngest cohort) to n_c (oldest cohort).

3. Multi-population Mortality Modeling

A multi-population mortality model considers more than one population in a joint mortality model. Such models can be used for example to model both male and female Nigeria mortality data into one model as well as interactions between both populations. For example we might expect some form of correlation between mortality improvements of both populations. Other applications of such models are joint models for population and portfolio data, or joint models for smokers/non-smokers, or different socio-economic classes in portfolio data.

3.1 Extension of univariate stochastic mortality models to bivariate

Few researchers have contributed to the subject of two-population stochastic mortality modeling. Li and Lee (2005) developed two extensions of the Lee-Carter model for modeling multiple populations. The two extensions were subsequently used by Li and Hardy (2011) to quantify the population basis risk in an index-based q-forward longevity hedge. Cairns et al. (2011) discussed the core hypotheses and desirable criteria for use in constructing two-population mortality models. Dowd et al. (2011) proposed a two-population gravity model, in which autoregressive processes are used to reduce any increasing spread between each pair of stochastic factors in the model back to a constant level and Zhou et al. (2014) investigated the use of different multivariate stochastic processes for projecting the stochastic factors in a multi-population mortality model. Zhou et al. (2013) introduced a two-population Lee-Carter model with transitory jump effects. The mentioned models are meticulously constructed so that they will not lead to diverging long-term forecasts, which do not seem to be biologically reasonable.

Most of the existing two-population mortality models are built on either the Lee-Carter (LC) or the Age-Period-Cohort (APC) structure. Nonetheless, these two model structures, which were developed many years ago, do not always provide the best fitting and forecasting results. In an analysis based on data from the US and EW populations, Cairns et al. (2009) found that Model M7 (the Cairns-Blake-Dowd model with cohort and quadratic effects) performs better than the LC and APC structures in terms of the Bayesian Information Criterion and the robustness of parameter estimates. Similar studies conducted by Cairns et al. (2011) and Dowd et al. (2010a,b) also point to the conclusion that the LC and APC structures are not always the most preferred. Cairns et al 2011 proposed the bivariate model of Age-period-Cohort Model (APC) for modeling

two closely related populations. The approach was applied to the England and Wales (EW) Male and UK male for ages 65-99. The model is thus;

$$\text{Log}\,\mu_1(x,t) = \alpha_x^{11} + \kappa_t^{21} + \gamma_{t-x}^{31}$$

$$\text{Log}\,\mu_2(x,t) = \alpha_x^{12} + \kappa_t^{22} + \gamma_{t-x}^{32}$$

And generally the above model can be extended to j closely related population define as;

$$\text{Log}\,\mu_j(x,t) = \alpha_x^{1j} + \kappa_t^{2j} + \gamma_{t-x}^{3j}$$

The table below summarize the extension of the models considered in this paper to the bivariate condition where $j = 2$

Table 1: The proposed models with respect to their structure and number of constraints that will make the model identifiable.

S/N	Model	Structure	Constraints
1	**AP Model**	$\log \mu_1(x,t) = \alpha_x^{11} + \kappa_t^{21}(x-\bar{x})$ $\log \mu_2(x,t) = \alpha_x^{12} + \kappa_t^{22}(x-\bar{x})$	**1*2**
2	**ASPC Model**	$\log \mu_1(x,t) = \alpha_x^{11} + \kappa_t^{21} + \kappa_t^{31}(x-\bar{x}) + \gamma_{t-x}^{41}$ $\log \mu_2(x,t) = \alpha_x^{12} + \kappa_t^{22} + \kappa_t^{32}(x-\bar{x}) + \gamma_{t-x}^{42}$	**3*2**

4.Fitting Age-Period-Cohort Models using Poisson GLMs

Fitting mortality models using Poisson GLM follow the same procedure as in the usual Poisson GLMs explained by McCullagh and Nelder (1989), the additional task involved is the addition of the constrains to make the models identifiable. The main task is to specify the design matrix $\boldsymbol{X}$ which is the matrix of the predictors. Here $\boldsymbol{X}$ can be a combination of age effect, period effect and cohort effect as the case may be in any of the model.

Let $\boldsymbol{d} = vec(d_{x,t})$ and $\boldsymbol{e} = vec(e_{x,t})$ be the vectors of observed deaths and central exposures; here, the vec operator stacks the columns of a matrix in column order on top of each other. It is pertinent to note that with this definition the age suffix varies faster than the year suffix in $\boldsymbol{d}$ and $\boldsymbol{e}$. With the above definition, it's easier to describe the Poisson GLM fitting

methodology for the APC model. The fitting procedure implemented in R as explained by Currie (2013) was used in estimating the models. Thus suppose $\boldsymbol{\theta}$ is the vector of parameters, the maximum likelihood estimator for parameter $\boldsymbol{\theta}$ is;

$$\bar{\theta} = \left(X^{'}X + H^{'}H\right)^{-} (X^{'}X + H'H_{R})\bar{\theta_{R}}$$

R are the constraint matrix for the model and the one used by R. Using the above formulation we $\boldsymbol{ML}$ since the result from R statistical package software assumes MLE framework. Thus

$$\bar{\theta}_{ML} = \left(X^{'}X + H^{'}H\right)^{-} (X^{'}X + H'H_{R})\bar{\theta_{R}}$$

Using mortality models for forecasting and simulation purposes based on the specific dynamics for the stochastic processes $\kappa_1(t)$, $\kappa_1(t)$ and $\gamma(t-x)$. This was modeled using Box Jenkins methodology (Box & Jenkins, 1976)

4.1. Criteria for Assessing the Goodness of Fit of the Models

The assessments of the models considered in this work were based on the following criteria.

i. Deviance
ii. Log-Likelihood
    a. Bayesian Information criterion (BIC)

ii. Deviance

The measure of discrepancy in a GLM to assess the goodness of fit of the model to the data is called the deviance. Deviance is defined as − 2 times the difference in log-likelihood between the current model and a saturated model (i.e. a model that fits the data perfectly). Because the latter does not depend on the parameters of the model, minimizing the deviance is the same as maximizing the likelihood. Deviance is estimated in different ways for different families within GLM.

The analysis is complicated slightly by the fact that some of the models we consider directly model the death rate $\mu_{x,t}$ while others model the mortality rate $q_{x,t}$. In order to ensure that our comparison of the different models is carried out in a consistent way, our analyses of the models for q(t, x) involve an additional step. First, for a given set of parameters we calculate the q(t, x). We then transform these into death rates using the identity $\mu_{x,t} = -\log[1 - q_{x,t}]$. We can then

calculate the likelihood for all models consistently based on the $\mu_{x,t}$ values. The equation below summarizes the deviance for the Poisson family of GLMs used in this thesis.

$$Dev_{Poisson} = 2\sum_{xt}[d_{xt}\log(d_{xt}/\hat{d}_{xt}) - (d_{xt} - \hat{d}_{xt})]$$

Where $\hat{d}_{xt}$ is the fitted death for Poisson models

Log-likelihood

Log-likelihood is another formal evaluation of the goodness of fit. The log-likelihood of the Poisson is given as;

$$Log\left(L(\theta)\right) = \sum_{xt}[D_{xt}\log[e_{xt}\hat{\theta}] - e_{xt}\hat{\theta} - \log(D_{xt}!)]$$

Where $\hat{\theta}$ is the estimate of the parameters of the model, a significant increase in the log-likelihood suggests that a particular model provides a better fits to the historical data. However, under the principle of parsimony increase in log-likelihood does not necessarily justify the adequacy of a model since increase in model parameters will lead to increase in log-likelihood. Therefore, to justify for the extra parameters introduced to a particular model, we use the following criterion:

Bayesian Information Criterion (BIC) (Schwarz, 1978)

$$BIC = Log\left(L(\theta)\right) - \frac{1}{2}Klog(n)$$

Where K is the number of parameters in the model. The intuition for BIC is to adjust for the number of parameters and sample size.

5. Simulation Study

The scheme used under univariate case as described by Bamidele et al (2014) was extended to capture bivariate situation. Assuming a starting population size N which is regarded as initial exposure to risk size, let q_{xt}^{i} denote the mortality rate for a fixed number of years n_y and varying ages $i = 1, 2, \ldots, n_a$. Also, it's very easy to define $1 - q_{xt}^{i}$ as the probability of

surviving from age x to $x+i$. Assuming baseline survival probability of 0.5, $q_{x,t}^{\,i}$ is generated as follows;

1. $1-q_{x,t}^{\,i}$ =c(seq(0.4,0.5,length=5), seq(0.5,0.01,length=n-5))
2. $y_{1i} \sim bernoulli(1-q_{1,x,t}^{\,i})$
3. $y_{2i} \sim bernoulli(1-q_{2,x,t}^{\,i})$
4. $\overline{q_{1,x,t}}$=1-$\bar{y}_1$
5. $\overline{q_{2,x,t}}$=1-$\bar{y}_2$

The sample matrices generated were replicated 1000 times to ensure stability. The model was therefore applied to Nigeria mortality data extracted from WHO Global health Observatory Data Repository http://apps.who.int/gho/data/view.main.60630?lang=en

5. Results

In this section, we present the result from the monte-carlo simulation and the real life data set; Nigeria mortality data. The dataset was extracted from R package demography within R the statistical software environment (www.cran.org). The data consists of numbers of deaths $D_{x,t}$ and the corresponding exposures $E_{x,t}$. The data simulation and analysis was implemented using the same software.

5.1 Simulation Case

Table 1: Model assessment criteria results using the univariate and bivariate method under limited data condition: $n_u = 101$ and $n_v = 3$

Gender	Model	UNIVARIATE			BIVARIATE			ENP
		DEVIANCE	LogLik	BIC	DEVIANCE	LogLik	BIC	
Male	AP	5575.617	-4549.45	-4882.61	5575.617	-4549.45	-4882.61	104
	ASPC	5644.006	-4583.64	-5256.37	5644.005	-4583.64	-5256.37	210
Female	AP	5228.494	-4375.66	-4708.82	5228.494	-4375.66	-4708.82	104
	ASPC	5255.223	-4389.02	-5061.75	5255.223	-4389.02	-5061.75	210

*** **ENP: Effective number of parameters**

Nigeria Mortality Data

Table 3: Model assessment criteria results using the univariate and bivariate method under limited data condition: $n_a = 22$ and $n_y = 3$

Gender	Model	UNIVARIATE			BIVARIATE			ENP
		DEVIANCE	LogLik	BIC	DEVIANCE	LogLik	BIC	
Male	AP	1209.804	-884.112	-945.147	1209.804	-905.444	-966.479	25
	ASPC	202.7759	-380.598	-507.551	202.7809	-380.6	-500.229	210
Female	AP	1159.911	-860.833	-921.868	1159.911	-878.749	-939.784	104
	ASPC	280.8332	-421.294	-548.246	280.8379	-421.296	-540.924	210

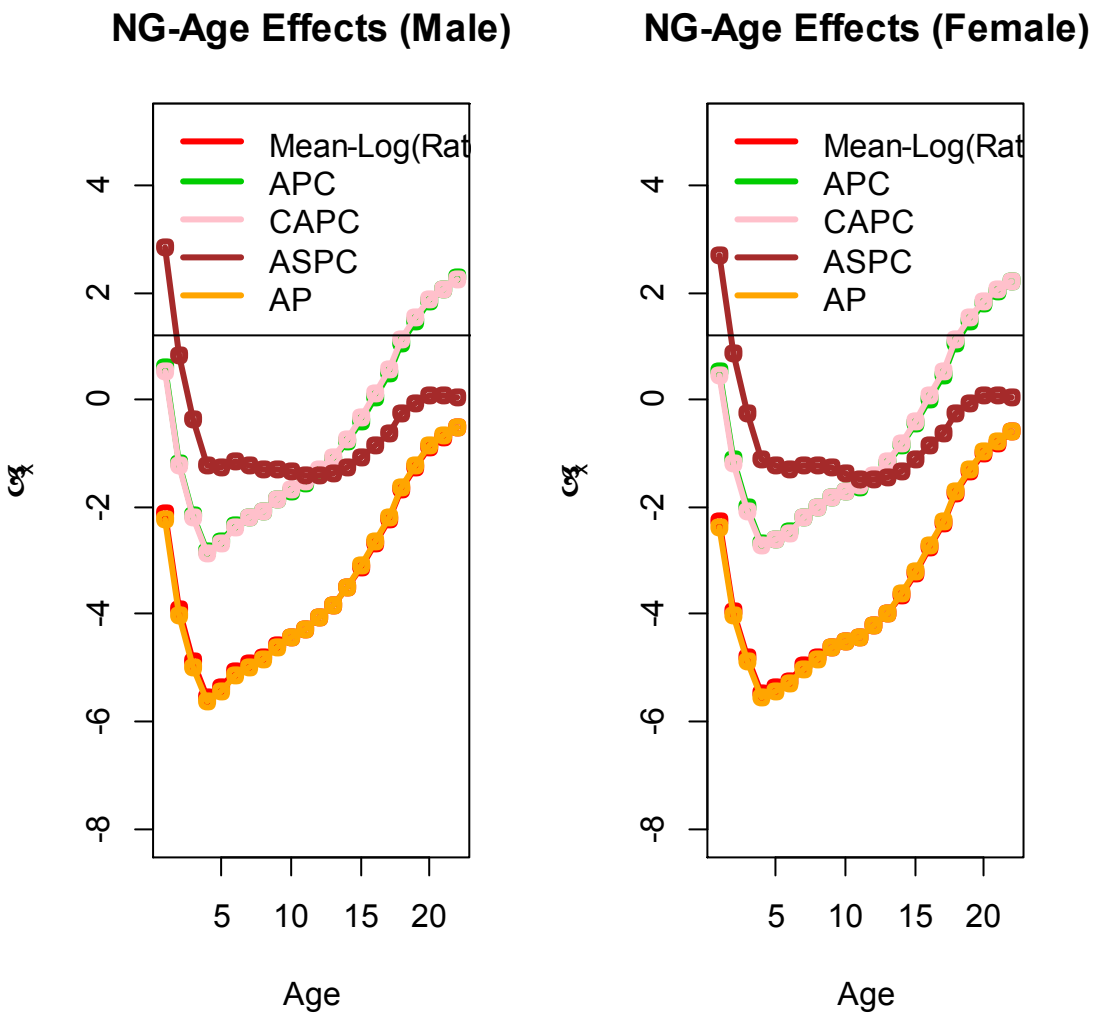
*** **ENP: Effective number of parameters**

7. Discussion of Results

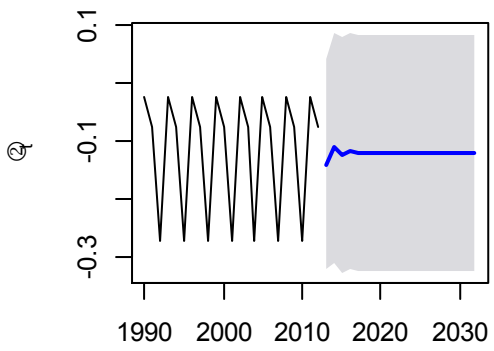
This paper has extended univariate model to capture for two-population stochastic models by the existing single-population mortality models to their two-population versions. The result of the bivariate and univariate for the simulated data revealed that the performance of the two procedures are identical with slight improvement in the performance for the bivariate case. Similar results were also observed using real life data sets. The parameter estimates for the dataset shows that AP model capture the age effect better compared to all other models. Though

ASPC has a better fitting result compared to AP, but AP estimates the age effect better. The period effects were not adequately captured by all the models due to fewer number years' data that is available. The confidence intervals of the forecast from the ARIMA models were very wide indicating that the future pattern of mortality is not stable but over time might come to equilibrium. The cohort effect estimate from the dataset was not significant. For further study, another area to explore is to extend the model to incorporate mortality jumps, which capture catastrophic mortality events such as a widespread pandemic and to historical data condition.

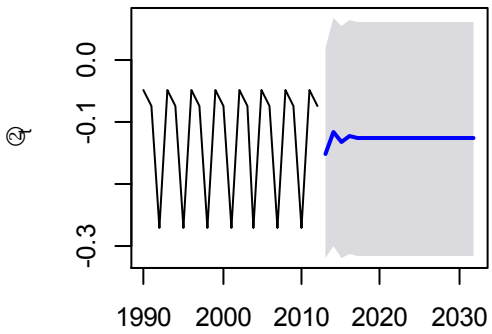
Parameter Estimates



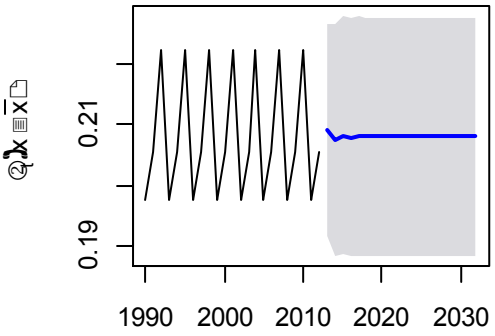
NG-ASPC Period Effect (Male)



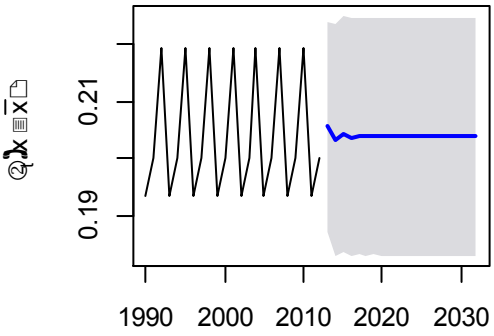
NG-ASPC Period Effect(Female)



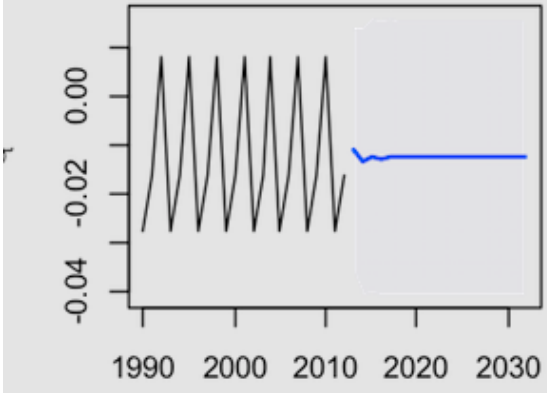
NG-ASPC C-Period Effect (Male)



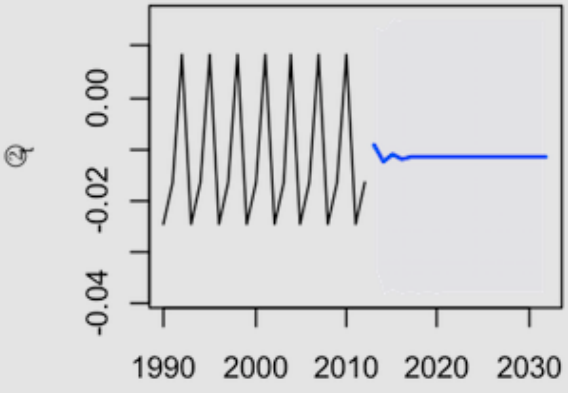
NG-ASPC C-Period Effect (Female



NG-AP Period Effect (Male)



NG-AP Period Effect (Female)



REFERENCES

Bamidele M.A, Adejumo A.O & Olaniran O (2014): *A New Approach to Stochastic Modeling and Fitting Mortality Risks,* Journal of Mathematical Analysis and Application, 1(4), Page 59-62

Box, G.E.P. & Jenkins, G.M. (1976) *Time series Analysis Forecasting and Control,* Holden-Day Press)

Currie, I.D., Durban, M. and Eilers, P.H.C. (2004) *Smoothing and forecasting mortality rates*. Statistical Modelling, 4, 279-298.

Currie, I.D., Durban, M. and Eilers, P.H.C. (2006) *Generalized linear array models with applications to multidimensional smoothing*. Journal of the Royal Statistical Society, Series B, 68, 259-280.

Cairns, A.J.G., Blake, D., and Dowd, K. (2006a) "*Pricing death: Frameworks for the valuation and securitization of mortality risk"*, ASTIN Bulletin, 36: 79-120.

Cairns, D. Blake, K. Dowd, G.D. Coughlan, D. Epstein, and M. Khalaf-Allah (2011a). *Mortality Density Forecasts: An Analysis of Six Stochastic Mortality Models*. Insurance: Mathematics and Economics, 48:355–367, 2011a.

Cairns, D. Blake, K. Dowd, G.D. Coughlan, and M. Khalaf-Allah (2011b). *A gravity model of mortality rates for two related populations*. North American Actuarial Journal, 15:334–356

Currie, I. D. (2013). *Fitting models of mortality with generalized linear and non-linear models*. Maxwell Institute for Mathematical Sciences, Heriot-Watt University, Edinburgh, EH14 4AS, UK.

Dowd, K., Cairns, A.J.G., Blake, D., Coughlan, G.D., Epstein, D., and Khalaf-Allah, M. (2010a). *Back-testing Stochastic Mortality Models: An Ex-Post Evaluation of Multi-Period-Ahead Density Fore- casts*. North American Actuarial Journal, 14, 281-298.

Dowd, K., Cairns, A.J.G., Blake, D., Coughlan, G.D., Epstein, D., and Khalaf-Allah, M. (2010b). *Evaluating the Goodness of Fit of Stochastic Mortality Models*. Insurance: Mathematics and Economics, 47, 255-265.

Dowd, K., Cairns, A.J.G., Blake, D., Coughlan, G.D., Epsten, D., and Khalaf-Allah, M. (2011). *A Gravity Model of Mortality Rates for Two Related Populations*. North American Actuarial Journal, 15, 334-356.

Lee, R.D., and Carter, L.R. (1992) "*Modeling and forecasting U.S. mortality*", Journal of the American Statistical Association, 87: 659-675.

Li, J.S.H., and Hardy, M.R. (2011). *Measuring Basis Risk in Longevity Hedges*. North American Actuarial Journal, 15, 177-200.

Li, N., and Lee, R. (2005). *Coherent Mortality Forecasts for a Group of Population: An Extension of the Lee-Carter Method*. Demography, 42, 575-594.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*. London: Chapman and Hall.

R Core Team (2014) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. (http://www.R-project.org).

Renshaw, A.E., and Haberman, S. (2006) "*A cohort-based extension to the Lee-Carter model for mortality reduction factors",* Insurance: Mathematics and Economics, 38: 556-570.

Renshaw, A.E., and Haberman, S., (2003a), "*On the forecasting of mortality reduction factors",* Insurance: Mathematics and Economics, 32, 379–401.

Renshaw, A.E., and Haberman, S., (2003b), "*Lee-Carter mortality forecasting: a parallel generalized linear modelling approach for England and Wales mortality projections*", Journal of Royal Statistical Society Series C (Applied Statistics), 52(1), 119–137.

Searle, S.R. (1982). *Matrix Algebra Useful for Statistics*. New York: Wiley.

Zhou, R., Li, J.S.-H., and Tan, K.S. (2013). *Pricing mortality risk: A two-population model with transi- tory jump effects*. Journal of Risk and Insurance, 80, 733-774.

Zhou, R., Wang, Y., Kaufhold, K., Li, J.S.-H., and Tan, K.S. (2014) *Modeling Period Effects in Multi- Population Mortality Models: Applications to Solvency II*. North American Actuarial Journal, in press.