Fitting Multi-Population Mortality Models to Socio-Economic Groups

Jie Wen, Andrew J.G. Cairns and Torsten Kleinow[*]

Department of Actuarial Mathematics and Statistics and the Maxwell Institute for
Mathematical Sciences, School of Mathematical and Computer Sciences, Heriot-Watt
University, EH14 4AS, Edinburgh, U.K.

July 19, 2019

Abstract

We compare results for twelve multi-population mortality models fitted to ten distinct socio-economic groups in England, subdivided using the Index of Multiple Deprivation. Using the Bayes Information Criterion to compare models, we find that a special case of the Common Age Effect (CAE) model fits best in a variety of situations, achieving the best balance between goodness of fit and parsimony.

We provide a detailed discussion of key models to highlight which features are important. Group-specific period effects are found to be more important than group-specific age effects, and non-parametric age effects deliver significantly better than parametric (e.g. linear) age effects. We also find that the addition of cohort effects is beneficial in some cases but not all. The preferred CAE model has the additional benefit of being coherent in the sense of Hyndman et al. (2013); some of the other models considered are not.

Keywords: Multi-population mortality models, Deprivation, Mortality inequality, Socio-economic models

1 Introduction

It is well-known that socio-economic inequalities in death rates and life expectancies exist in many countries. Those inequalities have been documented in the literature, see for example, Mackenbach et al. (1997) and Balia & Jones (2008). From an actuarial point of view it is important to take those differences into account for the pricing of annuities or life insurance products, and this is best achieved by considering stochastic mortality models that aim to capture the dynamics of death rates in different socio-economic groups while also allowing for common features that are shared by all.

In this paper we consider socio-economic groups defined with reference to the Index of Multiple Deprivation (IMD) in England. This index measures relative deprivation and allows us to identify ten groups. A brief empirical study of those ten groups can be found in Kleinow et al. (2019).

The main question we aim to answer in this paper is what model should be fitted to the group-specific data keeping in mind that all groups are sub-populations of the English national population, which suggests that they might share some common characteristics. Therefore, we are looking for a stochastic mortality model for multiple populations that allows us to capture common features as well as group specific factors.

Many stochastic multi-population models for mortality have been proposed in the literature. The most well-known model is that suggested by Li & Lee (2005). Their model is an extension of the Lee-Carter model, Lee & Carter (1992), that combines common age and time trends affecting all populations with population-specific components to allow for derivations from the common mortality table. An alternative is the common age effect model (CAE) suggested by Kleinow (2015). This model is also a modification of the Lee-Carter model. But, in contrast to the Li & Lee model, the CAE model treats all or some age effects as common while allowing for population specific period effects.

Some models that were originally developed for a single population can easily be adapted to multiple populations by allowing for some parameters to be common to all of them. One

---

[*]corresponding author, t.kleinow@hw.ac.uk

example is the model proposed by Plat (2009) which is one of the models that we adapt to fit multiple populations.

To be more specific, the contribution of this paper is a quantitative comparison of a number of stochastic multi-population models fitted to the mortality experiences in ten socio-economic groups in England. We introduce the different models and first compare them using the Bayesian Information Criterion. We then take a closer look at the best preforming models discussing the parameter estimates obtained for those models with a focus on identifying those parameters that capture the differences between groups and parameters that are common to all groups.

We also compare the fitted model-specific death rates with those actually observed in the ten groups to get a better idea of the advantages and shortcomings of the different models. While we will mostly focus on data for the female population, we find that models with cohort effects perform well for the male population and we therefore investigate those in more detail for both sexes.

While we concentrate in this paper on quantitative measures and goodness of fit for model selection, we should mention that other model properties, often called qualitative criteria, are also important to consider when models are chosen for a particular application. For example, models that result in perfect correlation between mortality improvements in different groups are not a good choice as they ignore group-specific trends. There are, of course, a number of other considerations, see for example, Villegas et al. (2017) and Cairns et al. (2018), and we will mention some of those when we discuss individual models. However, the focus of this paper is on quantitative criteria.

The remainder of the paper is organised as follows. We first describe the available data and discuss some empirical findings in section 2. In section 3 we then introduce the set of mortality models which we fit to the available data, and we then provide a first quantitative comparison of models in terms of their BIC in section 4. Parameter estimates for the some selected models are presented in section 5 and the fitted death rates implied by different models are discussed in section 6. In this section we find that some models might benefit from the inclusion of a cohort effect which we investigate in section 7. Finally, we summarise our main conclusions in section 8.

2 The Data

The data for our study consist of sex-specific deaths counts and exposures for ten socio-economic groups in England. The groups are determined by the Index of Multiple Deprivation (IMD) published by the UK Office for National Statistics. In the following we provide a brief description of the IMD and the related mortality data.

2.1 Index of Multiple Deprivation

The ONS measures different aspects of deprivation in small geographic areas in England, called Lower Layer Super Output Areas (LSOA). All LSOAs are of similar population size dividing the entire population into 32,844 small groups. The IMD is then calculated for each LSOA as a weighted average of seven indices measuring different aspects of deprivation: income (weight 22.5%), employment (22.5%), education (13.5%), crime (9.3%), health (13.5%), barriers to housing and services (9.3%), and living environment (9.3%). Each of those seven indices is based on a basket of indicators from the most recently available statistics. The IMD score for an individual LSOA represents a measure for the relative deprivation of the population living in that LSOA compared to other LSOAs. The LSOAs are then ranked according to their IMD score from the most deprived to the least deprived area in England and ten deciles are formed. Detailed information about the IMD can be found in Smith et al. (2015a) and Smith et al. (2015b) where the construction of the index is explained in detail.

2.2 Mortality Data and Notation

The ONS has published sex-specific population sizes, denoted by E_{xti}, and deaths counts, D_{xti}, for each IMD decile i, age x and year t for ages 0 to 89 (and 90+ as one group) and calendar years 2001 to 2017, see the ONS data portal on their website[1].

---

[1] The mortality data can be sourced at ONS website https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsand marriages/deaths/adhocs/009299numberofdeathsandpopulationsindeprivationdecileareasbysexandsingleyearofageenglandandwales registeredyears2001to2017, accessed on 24 June 2019.

Population sizes in 2017 are shown in Table 1. The total populations (over all ages) in the different deciles are of similar size, although they are smaller in the more deprived areas than in the less deprived. Although the numbers of LSOAs in each decile are the same, differences arise in Table 1 because (a) the individual LSOAs vary in size, (b) less deprived LSOAs tend to be larger and (c) more deprived LSOAs tend to have a greater proportion of the population aged less than 40.

Group	IMD males	IMD females
g1 (most deprived)	1,094,182	1,142,736
g2	1,138,784	1,213,750
g3	1,198,198	1,269,708
g4	1,274,284	1,348,478
g5	1,334,111	1,424,273
g6	1,392,196	1,490,977
g7	1,414,449	1,519,678
g8	1,420,182	1,530,570
g9	1,435,837	1,553,131
g10 (least deprived)	1,439,358	1,562,401

Table 1: Total population size $\sum_x E_{xti}$ in individual deciles i by sex for ages 40 to 89 in the year 2017.

2.3   Mortality in the Ten Groups

Before turning to mortality models we briefly discuss some empirical features of the mortality expierences in the ten socio-economic groups. In figure 1 we plot the crude death rates D_{xti}/E_{xti} by age for the year 2015, and the rates at age 65 by year. We find that both, males and females, have very similar mortality pattern at age 65 and in 2015. The death rates in the most deprived areas are very high compared to the least deprived. In fact, there seems to be an almost perfect ranking of death rates with respect to the deprivation deciles. The differences between groups seem to be more pronounced in the male populations, especially at younger ages. For both sexes those differences decrease with age. We also observe that mortality improvements have been smaller for the most deprived groups compared to the strong improvements experienced by the least deprived.
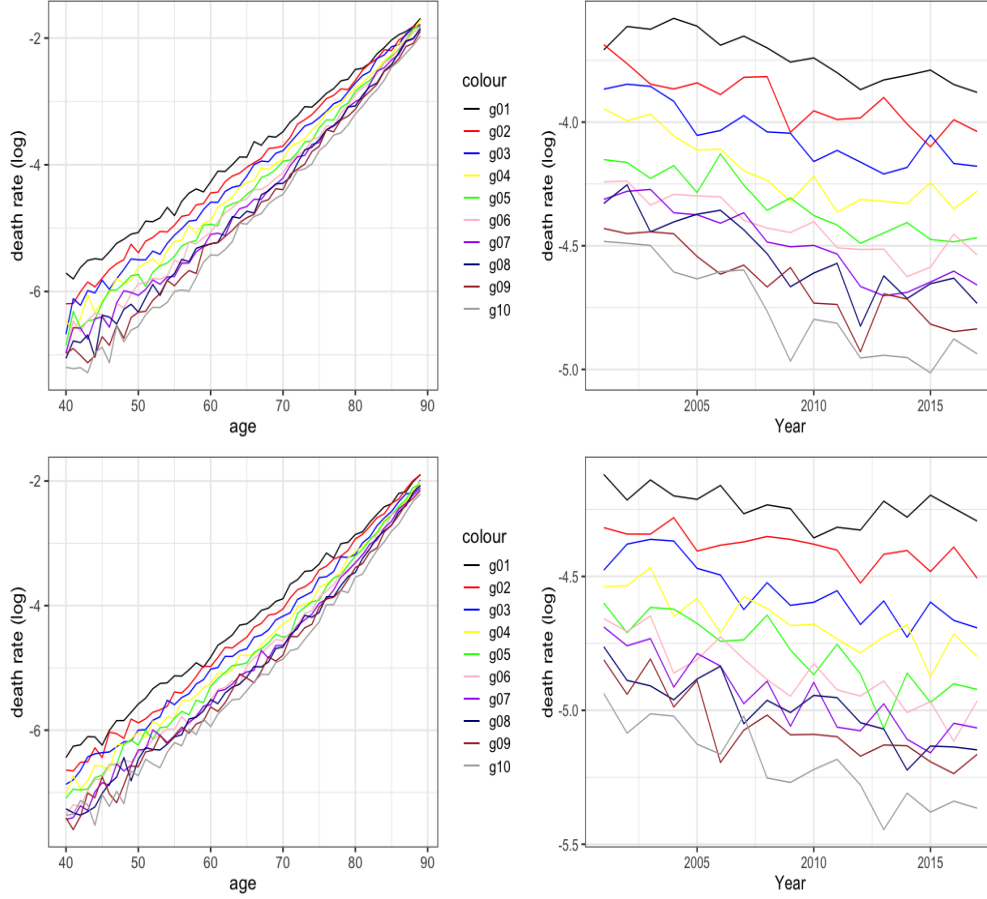
Figure 1: Crude death rates (log-scale) in 2017 (left) and at age 65 (right) for males (top row) and females.

To compare mortality in different populations over a wider age range we standardise death rates to compensate for differences in the demographic structures of the populations. To this end we calculate an age standardised mortality rate (ASMR) which is a weighted average of the crude death rates over a defined age range. The ASMR for group i in calendar year t for specific ages $\mathcal{X}$ is defined as:

$$ASMR(ti) = \sum_{x \in \mathcal{X}} \frac{D_{xti}}{E_{xti}} w_x \text{ with weights } w_x = \frac{E_x^s}{\sum_{x \in \mathcal{X}} E_x^s}$$

where the weights are determined by the age-specific exposures, E_x^s, in some standard population. For our empirical results, we use the European Standard Population (ESP) in 2013 and the age range $\mathcal{X} = \{40, \ldots, 89\}$ (Revision of the European Standard Population, 2013 Edition)[2].

In Figure 2 we plot the ASMRs for the available calendar years for both sexes.

---

[2]Report is available at: https://ec.europa.eu/eurostat/documents/3859598/5926869/KS-RA-13-028-EN.PDF/e713fa79-1add-44e8-b23d-5e8fa09b3f8f, downloaded 24 June 2019
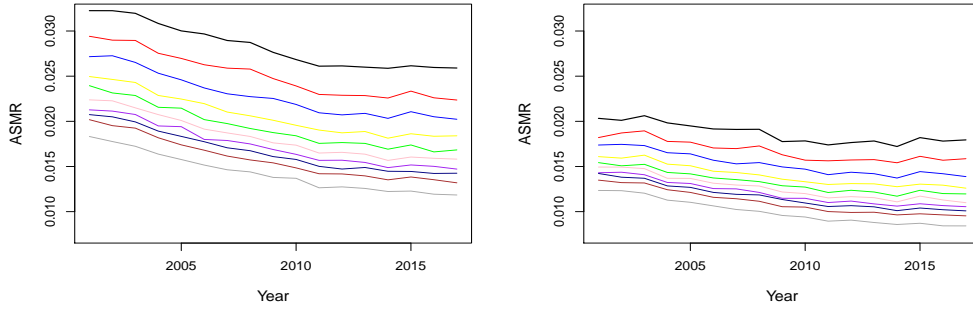
Figure 2: Group-specific ASMRs based on ages 40 to 89 for the ten IMD deciles, males (left) and females (right). The weighting is based on the European Standard Population recalibrated in 2013.

We find in Figure 2 that the ASMRs are clearly ranked with respect to the level of deprivation for both sexes. The rates decrease steadily over time for all groups until 2011, after which there appears a slow down of mortality improvements, a feature which was not so obvious in the crude death rates in Figure 1.

2.4 The IMD as an indicator of Mortality

Since the focus of this paper is on a quantitative comparison of mortality models for different socio-economic groups we should make some comments about the suitability of the IMD as a means to identify different groups, in particular, with a view towards mortality modelling. After all, the index was not specifically created for identifying groups which experience different mortality patterns.

In fact there are a few issues with the IMD in that respect. First of all, one of the components in the IMD is related to health. In the research report by Smith et al. (2015a) it is explained that the health domain score is constructed by several indicators including comparative illness measure, morbidity rate and illness and disability ratio among others. These indicators are very relevant to the death rates in a population, and their inclusion in the IMD is likely to have an impact on our results. However, we would argue that the impact is relatively small. Most importantly, there is a high correlation between individual components of the IMD. Specifically the correlation between health and income deprivation is 0.83 and between health and employment deprivation 0.88. So income and employment deprivation would act as effective proxies if health deprivation was excluded from the IMD.

The IMD deciles are derived from the ranks of the LSOAs rather than the index values. The ranks clearly do not capture the actual differences between deprivation in different LSOAs. Instead, they only show that one LSOA is more or less deprived than another but not by how much.

While the ten deciles are constructed such that each decile contains an equal number of LSOAs, the population sizes are different as shown in Table 1. This will have an effect on parameter uncertainty but we would expect that effect to be rather small, since the population sizes are similar.

The IMD measures deprivation rather than affluence. While the two concepts are related, it is important to keep this in mind when considering our empirical results. A good example for illustrating this issue is income deprivation. The income deprivation score is constructed from data on low-income state benefits, meaning that two LSOAs with a similar number of people receiving similar amounts of income benefits will have a similar rank with respect to the income deprivation score. On the other hand, the average income in those two LSOAs might be very different as this is determined by the differences in income of those people who receive high incomes and therefore, no or little income benefits.

All of the issues mentioned here suggest that a more detailed analysis of individual aspects of deprivation might be more suitable for the identification of socio-economic groups. However, the strong differences between deciles with respect to mortality suggest that the IMD is well suited to improve mortality models.

m1	$\log m_{xti} = \alpha_{xi} + \beta^1_{xi}\kappa^1_{ti} + \beta^2_{xi}\kappa^2_{ti}$	(Renshaw & Haberman (2003))
m2	$\log m_{xti} = \alpha_{xi} + \beta^1_{xi}\kappa^1_{ti} + \beta^2_x\kappa^2_{ti}$	(m1 with common β^2_x)
m3	$\log m_{xti} = \alpha_{xi} + \beta^1_x\kappa^1_t + \beta^2_{xi}\kappa^2_{ti}$	(Li & Lee (2005))
m4	$\log m_{xti} = \alpha_{xi} + \beta^1_{xi}\kappa^1_{ti}$	(Lee & Carter (1992))
m5	$\log m_{xti} = \alpha_{xi} + \beta^1_x\kappa^1_{ti} + \beta^2_x\kappa^2_{ti}$	(Kleinow (2015))
m6	$\log m_{xti} = \alpha_x + \beta^1_x\kappa^1_{ti} + \beta^2_x\kappa^2_{ti}$	(m5 with common α_x)
m7	$\log m_{xti} = \alpha_{xi} + \kappa^1_{ti} + (x - \bar{x})\kappa^2_{ti}$	(Plat (2009))
m8	$\log m_{xti} = \alpha_x + \kappa^1_{ti} + (x - \bar{x})\kappa^2_{ti}$	(m7 with common α_x)
m9	$\log m_{xti} = \alpha_{xi} + \kappa^1_t + (x - \bar{x})\kappa^2_{ti}$	(m7 with common κ^1_t)
m10	$\log m_{xti} = \alpha_{xi} + \kappa^1_{ti} + (x - \bar{x})\kappa^2_t$	(m7 with common κ^2_t)
m11	$\log m_{xti} = \alpha_{xi} + \kappa^1_t + (x - \bar{x})\kappa^2_t$	(m7 with common κ^1_t and κ^2_t)
m12	$\log m_{xti} = \kappa^1_{ti} + (x - \bar{x})\kappa^2_{ti}$	(Cairns et al. (2006))

Table 2: List of models considered in this study.

3 Multi-Population Mortality Models

The focus of this study is on comparing multi-population stochastic mortality models with respect to their ability to fit the mortality experiences in the ten IMD deciles simultaneously. In this section we look at twelve models with different parametric structures and analyse the results. All models are fitted to data for females and males. They all have an 'Age-Period-Cohort' (APC) structure.

3.1 Basic Modelling Assumptions and Estimation

We assume that the number of deaths D_{xti} at age x, in year t and for IMD decile i has a Poisson distribution,

$$D_{xti} \sim \text{Pois}\left(m_{xti}E_{xti}\right),$$

with intensity parameter $m_{xti}E_{xti}$ where m_{xti} is the death rate, and E_{xti} denotes the exposure at risk, that is, the mid-year population estimate as introduced earlier.

In the following, the death rate m_{xti} will be assumed to follow a parametric model and we use maximum likelihood estimation to estimate its unknown parameters. It is well-known that identifiability problems exist in all of the specific models for m_{xti} which we consider in our study. We, therefore, impose constraints on the parameters to obtain unique parameter estimates. As this is a common topic in the literature on stochastic mortality models we do not provide details on the constraints. Instead, we refer the reader to the relevant literature, in particular, the references given in Table 2 where the specific models are introduced.

3.2 Model specification

All considered models are for the log death rates, $\log m_{xti}$. They are listed in Table 2 where they are roughly ordered according to their complexity with model $m1$ being the model with the most parameters. All other models can be derived from $m1$ by imposing restrictions on some of the parameters.

The unknown parameters α, β and κ capture age and period effects. The parameter α acts as the 'baseline' age pattern of mortality while κ describes the period effect. Finally, β rescales this period effect to obtain different mortality improvements at different ages.

3.3 Relationships between models

As mentioned above, model m1 is the most complex in the sense that it has the largest number of parameters. All other models are nested in m1, and some are also nested in others. Figure 3 shows the model hierarchy with arrows pointing from the nested model to the more complex model.
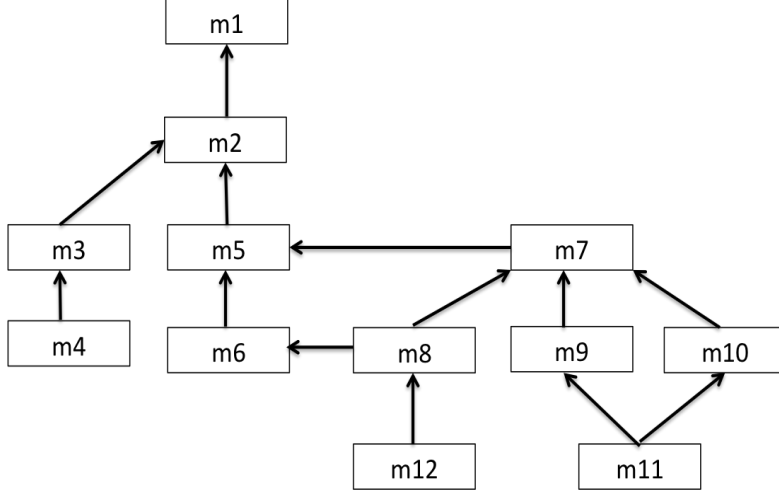
Figure 3: Tree plot specifying nested models (arrows are pointing to the more detailed model)

Figure 3 shows the distinction between two families of models at high level: Lee-Carter-type models with nonparametric effects β on the left hand side, and Cairns-Blake-Dowd-type models on the right hand side where the age effects β are either constant or linear in age x but the parameter α is chosen to be nonparametric as suggested by Plat (2009).

4 Quantitative Comparison of Models

To compare the different models we first consider two quantitative criteria: the Bayesian Information Criterion and the explanation ratio.

4.1 Model ranking with respect to the BIC

For an initial ranking we fit the twelve models to female and male data separately and calculate the BIC values. The BIC is defined as

$$BIC = k \log(n) - 2 \log(\hat{L})$$

where k represents the degrees of freedom, which is the number of parameters reduced by the number of constraints required for the model. The sample size n is calculated as the product of the number of ages, years and groups. Finally, $\log(\hat{L})$ denotes the log likelihood function. Due to the way the BIC is defined here, lower BIC values indicate a better fit of the model.

| Model | k | females | | males | |
		$\log(\hat{L})$	BIC	$\log(\hat{L})$	BIC
m1	1800	-34398.54	85083.16	-35634.78	87555.64
m2	1359	-34652.44	81600.86	-35900.31	84096.61
m3	1215	-34848.27	80689.64	-36065.10	83123.31
m4	1150	-35083.25	80571.50	-36293.44	82991.87
m5	918	-35058.84	78423.59	-36242.45	80790.81
m6	486	-35336.06	75069.36	-36702.57	77802.39
m7	820	-35653.80	78726.80	-37422.19	82263.59
m8	388	-37375.07	78260.70	-38213.32	79937.20
m9	676	-36104.58	78325.48	-37821.39	81759.10
m10	676	-35746.95	77610.23	-37491.71	81099.75
m11	532	-36760.83	78335.10	-38171.44	81156.32
m12	340	-46822.86	96721.99	-41385.10	85846.45

Table 3: BIC and log likelihood values for all models fitted to IMD deciles for ages 40 to 89. The degrees of freedom are denoted by k.

In general, Table 3 shows that models with more parameters tend to have higher log-likelihood values, but they are then penalised for over-parametrization in the BIC calculation. For females and males, Model m6 (the CAE model with common α_x) turns out to be the preferred model in the sense that it has the lowest BIC value, and therefore, seems to strike a good balance between goodness of fit and number of parameters. Model m8, which imposes a specific form on the parameters β^1 and β^2, is the second-best model in our collection. The feature that both models share, is that the baseline age factor α is common to all groups.

Considering the results in Table 3 for some of the other models, we notice that switching from group-specific α_{xi} in model m5 to common α_x (m6) does not deteriorate the log-likelihood too much, and as there are less parameters after the change, the BIC is improved significantly from m5 to m6. A similar effect is observed when comparing models m7 and m8. This indicates that the different socio-economic groups have a very similar basic age structure.

Comparing the BIC values in Table 3 for models m6 and m7 shows that the assumptions $\beta^1 = 1$ and $\beta^2 = x - \bar{x}$ are not justified for the data set that we consider here. The number of parameters in model m7 is actually greater than the number of parameters for model m6. Nevertheless, even the log likelihood value of m6 is better than that of m7.

Models m7 and m8 explicitly assume that annual changes in log death rates are linear in age. It is well-known that this is assumption is justified for relatively old ages, but not for younger ages. To investigate the impact of the chosen age range on our results in Table 3 we repeat our analysis for the age range 65-89, see Table 4. In that table we find that model m6 still provides a better fit than model m8 (although the gap is smaller), which is a further indication that the assumption of linear improvements (in age) log death rates does not seem to be justified for our data set.

| | | females | | males | |
Model	k	$\log(\hat{L})$	BIC	$\log(\hat{L})$	BIC
m1	1050	-19284.74	47341.88	-19570.98	47914.38
m2	834	-19424.59	45816.98	-19700.23	46368.27
m3	690	-19596.88	44958.48	-19853.85	45472.42
m4	650	-19773.19	44976.91	-20051.25	45533.04
m5	618	-19663.73	44490.65	-19898.83	44960.84
m6	411	-19817.36	43068.49	-20179.78	43793.34
m7	570	-19964.27	44690.70	-20407.43	45577.02
m8	363	-20194.42	43421.58	-20546.68	44126.10
m9	426	-20764.87	45088.84	-21090.39	45739.86
m10	426	-20051.93	43662.95	-20482.63	44524.35
m11	282	-21015.19	44386.39	-21180.54	44717.09
m12	340	-20858.08	44556.74	-20704.75	44250.09

Table 4: BIC and log likelihood values for all models fitted to IMD deciles for ages 65 to 89. The degrees of freedom are denoted by k.

4.2 Explanation ratio

We are also interested in how much of the information contained in empirical deaths rates, D_{xti}/E_{xti}, is explained by our models. To this end we consider the explanation ratios for our ten socio-economic groups following the definition by Li and Lee (2005):

$$R_i = 1 - \frac{\sum_{xt} \left(\log \frac{D_{xti}}{E_{xti}} - \log \hat{m}_{xti} \right)^2}{\sum_{xt} \left(\log \frac{D_{xti}}{E_{xti}} - \alpha_{xi}^c \right)^2}$$

where $\log \hat{m}_{xti}$ denotes the fitted log death rate for a specific model in table 2 but with the unknown parameters replaced by their maximum likelihood estimates. The baseline age factor α_{xi}^c in the denominator is defined as the average over time of the log death rates at certain ages and groups:

$$\alpha_{xi}^c = \frac{1}{n_Y} \sum_t \log \frac{D_{xti}}{E_{xti}}$$

where n_Y is the total number of years; for our data we have $n_Y = 17$. The explanation ratio R_i describes the percentage of observed mortality variation in group i that can be explained by a model.

In Table 5 we show the obtained values for R_i for some of our models. In general, the explanation ratios are maybe smaller than we would expect from a large population like England & Wales. However, we should keep in mind that an individual IMD devile has a much smaller population size, and the uncertainty about parameter estimates and fitted death rates is correspondingly high, see for example Enchev et al. (2015).

Group	R_{m1}	R_{m3}	R_{m5}	R_{m6}	R_{m7}	R_{m8}
1	0.81	0.77	0.77	0.75	0.70	0.56
2	0.78	0.74	0.74	0.71	0.70	0.58
3	0.78	0.75	0.74	0.72	0.68	0.64
4	0.79	0.77	0.75	0.73	0.69	0.68
5	0.76	0.74	0.72	0.69	0.65	0.61
6	0.82	0.79	0.76	0.75	0.71	0.65
7	0.74	0.74	0.69	0.67	0.63	0.55
8	0.75	0.72	0.69	0.68	0.63	0.58
9	0.77	0.71	0.68	0.66	0.61	0.52
10	0.78	0.75	0.73	0.69	0.71	0.63

Table 5: Explanation ratios R_i for all groups and some models.

We also clearly observe in Table 5 that models with more parameters tend to have higher explanation ratios. In particular, we find that the two models with the best BIC values and relatively few parameters, m6 and m8, have rather low explanation ratios with model m6 dominating model m8.

5 Estimated Parameters

We will now turn to comparing our models with respect to qualitative aspects. In particular, we are interested in how the estimated parameter values compare to those estimated from our baseline model m1. Therefore, we start with a short discussion of m1 and then provide estimates and discussions for some of the other models where we concentrate on those that provide a good fit to our data in the sense of a low BIC value, see Table 3.

5.1 Our Baseline Model - m1

As mentioned above, our model m1 was proposed by Renshaw and Haberman (2003) as an extension to the Lee-Carter model for the mortality experience in a single population. Therefore, all parameters in this model are group specific, see Table 2. Using this model would implicitly assume that mortality patterns are very different between groups in both, ages and years. We consider the model here as a baseline, and we will compare the estimated age and period effects in other models with those obtained for model m1.

As there are some identifiability issues with this model, we need to impose constraints to obtain a unique solution when maximising the likelihood function. We have chosen to apply the following set of constraints for each group i:

$$\sum_x (\beta_{xi}^1)^2 = 1, \quad \sum_x (\beta_{xi}^2)^2 = 1, \quad \kappa_{0i}^1 = 0, \quad \sum_t \kappa_{ti}^2 = 0.$$

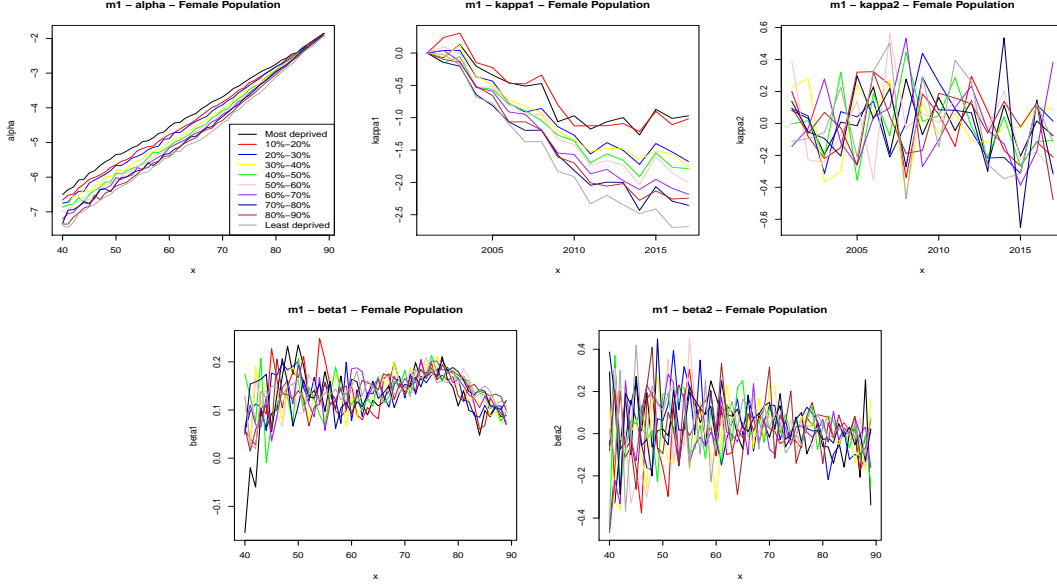Figure 4 shows the estimated parameters.

Figure 4: Estimated parameters for model m1 (female population).

We observe in Figure 4 that the estimated values of α_{xi} as the overall age pattern confirm our earlier observations that clear differences exist between socio-economic groups and that those differences are decreasing with age. This seems to indicate that it would be reasonable to choose α as a group-specific parameter, that is, a different basic age pattern for all IMD deciles. However, we should keep in mind that α is not identifiable in this model and that our results in Table 3 clearly point towards the opposite conclusion. When considering models m6 and m8 we will see how other parameters pick up the age-related differences between groups when α is common to all groups.

The parameter κ_{ti}^1 is the leading period effect. It clearly shows a downward trend indicating longevity improvements during the observation period 2001-2017 for all IMD deciles. However, we notice the kink appearing in 2011 at which the downward slope of κ_{ti}^1 becomes less steep. This corresponds to the well-documented slow-down of longevity improvements since 2011, which we here observe for all groups. Figure 4 also reveals that the mortality improvement rates are very different between groups with the least deprived experiencing the strongest improvements. We should mention here that the slope of κ^1 is only identifiable up to a group-specific constant. However, we have chosen our constraint on β^1 such that the parameters β^1 are on a similar level for all groups. This means that the different slopes of κ^1 can only be explained by mortality improvement rates that are different for different IMD deciles. This would make it unlikely that κ^1 can be chosen to be common to all groups, which is consistent with our results in Table 3 where we show that the BIC of model m3 is less good than many of the others.

We also observe that there is much variability in β_{xi}^1 for ages up to about sixty, but that there seems to be a pattern for older ages. In particular, we find that β^1 increases from age sixty to a maximum value at around age 75, and then decreases. This indicates that at age 75 we observe the greatest mortality improvements. The age at which we observe maximum improvements seems to be similar in all groups. Also, more generally, it seems that the parameters β^1 and β^2 are similar across groups with non-systematic differences between them. This would suggest that those parameters can indeed by modelled as common without loosing too much quality of fit.

Finally, both κ_{ti}^2 and β_{xi}^2 are rather noisy and without any regular pattern, as they absorb second order effects which are not covered by the other parameters.

5.2   The Best Fitting Model m6 - Common Age Effects

As mentioned earlier, the models that fit our data best are the two models that have no group-specific age effects: m6 and m8 with m6 providing the better fit. The model m6 is a modification of the CAE model m5, proposed by Kleinow (2015). To investigate the effect of choosing common age effects rather than groups specific effects we compare the estimates of the age effects, α, β^1 and β^2 of the two models (m5 and m6) with each other and with

those obtained for our baseline model m1.

The original CAE model m5 suggests that age-related mortality improvements over time are the same across all groups while the basic age structure captured by α_{xi} are group specific. In contrast, in model m6 even that basic age structure is not specific to the socio-economic group. This might be surprising given the big differences between the α_{xi} in Figure 4. To investigate this further we show the estimated values of the parameters α in Figure 5.

Figure 5: MLE estimates of the parameters in models m5 and m6. The dashed line in the plot of α is for m6. The dotted lines in the plots of β^1 and β^2 represent the estimated parameters in model m1 for comparison.

We find in Figure 5 that the slope of the estimated α_x in model m6 (the dashed line) is roughly equal to the average slope of the group specific age effects α_{xi} for model m5, which look very similar to those estimated in m1, see Figure 4.

Comparing the time effects we find that the differences between groups are clearly picked up by the period effects κ^1 in model m6. The differences between those period effects together with the common age effect β_1 in m6 would only result in a parallel shift of death rates from one group to another. The observed convergence of rates for high ages (Figures 1 and 4) is captured by the parameters β^2 and κ^2 in m6. While those two parameters only capture second order effects in models m1 and m5, we find that the common basic age structure α in model m6 gives β^2 and κ^2 a stronger meaning. This change of interpretation of β^1 and β^2 explains why they are rather different across the three models.

In Figure 5 we have also shown graphs of the estimates for β_1 and β_2 in model m1 to compare them with our estimates in the CAE models m5 and m6. We clearly see that the estimates for those parameters in m5 follow the same general pattern as the estimates in m1. Of course, β_1 and β_2 are only uniquely identifiable in m1 and m5 up to a constant factor, but we have chosen the identifiability constraints for the two models such that they are on the same scale.

5.3 Models m7 and m8 - Constant and Linear Age Effects

One way of further reducing the number of parameters in the CAE models m5 and m6 is imposing a parametric structure on the common age effects β^1 and β^2. This is our motivation for considering models m7 and m8. The model m7 was suggested by Plat (2009) as a model for an individual population, so with a group-specific parameter α_{xi}. The model can be considered as an extension to the CBD model (Cairns et al. (2006)) with an extra 'baseline' α_{xi}, or a simplification of the CAE model m5 with $\beta^1_{xi} = 1$ and $\beta^2_{xi} = x - \bar{x}$ for all groups i.

Comparing the BIC values for m6 and m8 (or m5 and m7) in Tables 3 and 4 we find that the goodness of fit is reduced by introducing the constant and linear structure for the age effects. On the other hand, when comparing the quality of fit of m7 and m8, we find again that choosing the basic age structure α to be common to all IMD deciles is improving the BIC.

Figure 6: MLE estimates of the parameters in models m7 and m8. The dashed line in the plot for α is for m8.

Figure 6 shows the obtained parameter estimates. Our conclusions are similar to those drawn when we compared models m5 and m6: it is clearly so that the differences in death rates between groups are captured by α, but if that parameter is chosen to be common to all groups, the parameters κ^1 and κ^2 take over as the factors that distinguish the death rates in different groups from each other.

Interestingly, we observe an almost perfect ranking of κ^1 in model m8 (and m6) with the lowest level of mortality in the least deprived group of the population. For κ^2 in m8 this ranking is the other way around indicating that the slope of the Gombertz-line is steepest for the least deprived groups. This observation is consistent with our findings in Figure 1 that mortality differences between socio-economic groups are greatest at young ages with the least deprived having the lowest rates, but that those differences are very small at old ages, meaning that the age-related increase in mortality is strongest for the least deprived.

12

5.4 Model m3 - Common Time Effect

While we have found in tables 3 and 4 that common age effects seem to improve the BIC of the considered models, this is not found for a common period effect κ^1 as in model m3. This model was first proposed by Li and Lee (2005) as a model which captures the common trend in a number of populations and combines that common trend with population-specific factors β^2 and κ^2.
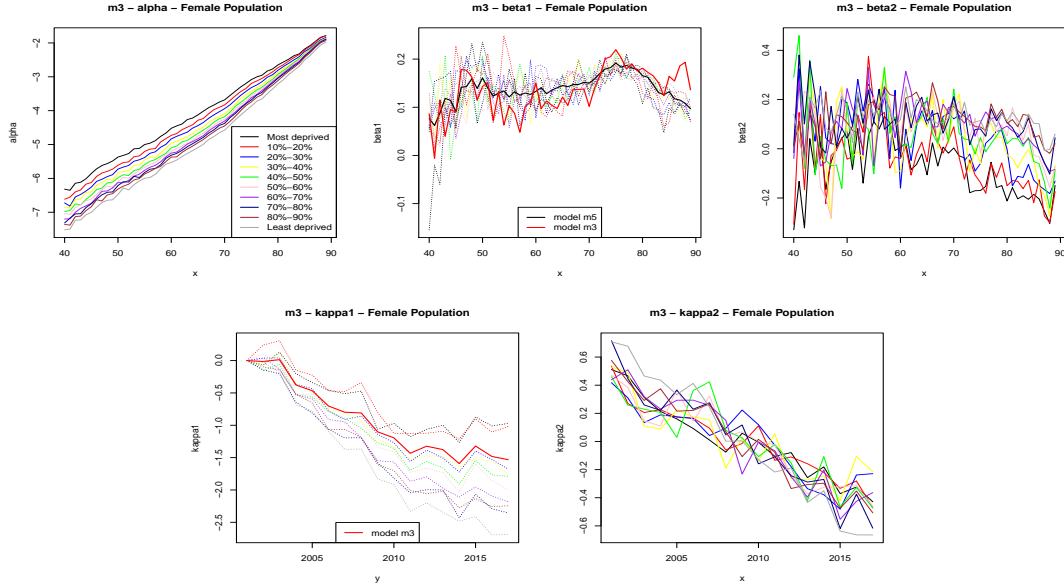

Figure 7: MLE estimates of the parameters in model m3. The dotted lines in the plots of β^1 and κ^1 represent the estimated parameters in model m1 for comparison. The plot for β^1 also includes the estimates for model m5.

We show the estimated parameter values in Figure 7 and find that the common period effect picks up general development of death rates over time across all groups. However, there are of course differences between groups that we discussed earlier. In particular, there are differences in the mortality improvement rates in the different groups and a common time trend κ^1 together with a common parameter β^1 is clearly not able to capture those differences. It seems that the additional group-specific age and period effects β^2 and κ^2 are rather similar to each other, and are not able to capture those differences and other second order effects. Those observations together with the obtained BIC values for all our models lead us to the conclusions that common period effects are not present in our data, and that the parameters κ^1 and κ^2 are best chosen to be group-specific.

6 Goodness of Fit

The BIC values presented in Section 4 indicate the relative goodness of fit of individual models when compared to others. In this section we further investigate the fit of some of our models to the observed data by considering graphical diagnostics, namely residual plots and plots that compare observed and fitted death rates at specific ages or years.

6.1 Standardised residuals

We start our analysis with Pearson's residuals, Z_{txi}, defined as the standardised difference between crude observations:

$$Z_{xti} = \frac{D_{xti} - E_{xti}\hat{m}_{xti}}{\sqrt{E_{xti}\hat{m}_{xti}}}$$

where $\hat{m}_{xti}$ is the model-specific fitted death rate at age x in year t.

A good model should result in standardised residuals which show no trends or clusters along any of the dimensions. Studying the distribution of the residuals Z over the underlying data range can therefore give indications for systematic effects in the data that a model fails to capture.

We have found that models m6 and m8 provide the best fit in terms of the BIC compared to the other models in Table 2. Therefore, we focus on their standardised residuals, and we start by comparing the mean squared error of the two models for each individual socio-economic group.

Group	m6	m8
g1	0.674	1.209
g2	0.804	1.145
g3	0.723	0.954
g4	0.760	0.974
g5	0.824	1.096
g6	0.689	1.037
g7	0.8323	1.184
g8	0.767	1.079
g9	0.792	1.177
g10	0.743	1.007

Table 6: Mean-squared error (MSE) calculated from standardised residuals of models m6 and m8 at individual group level.

Table 6 shows that m6 has much lower MSEs than m8, which would expect to find since m6 has more parameters than m8. Despite model m8 having greater MSEs they are still close to 1 for all groups. Extending our analysis to all twelve models in Table 2, we find that model m1, the most complex model with the greatest number of parameters, has the lowest MSE, which is much lower than 1 indicating that the model is over-fitting the data.

Another important aspect of standardised residuals is their distribution over the data range and this is best assessed by graphical analysis, for example, heatmaps of the residuals over the underlying age and year range.

In Figure 8 we show the heatmaps for the common age effect model m6 for groups g1, g5 and g10 with years on the horizontal axis and ages on the vertical axis. The heatmaps for all three groups show no obvious pattern or bias indicating that the obtained residuals are randomly distributed and that model m6 captures all of the structure in our data. In particular, there is no structure along individual axis (corresponding to ages and years), and there is no obvious diagonal (corresponding to a cohort effect). This indicates that m6 has captured age, year and cohort effects from the IMD males data well.
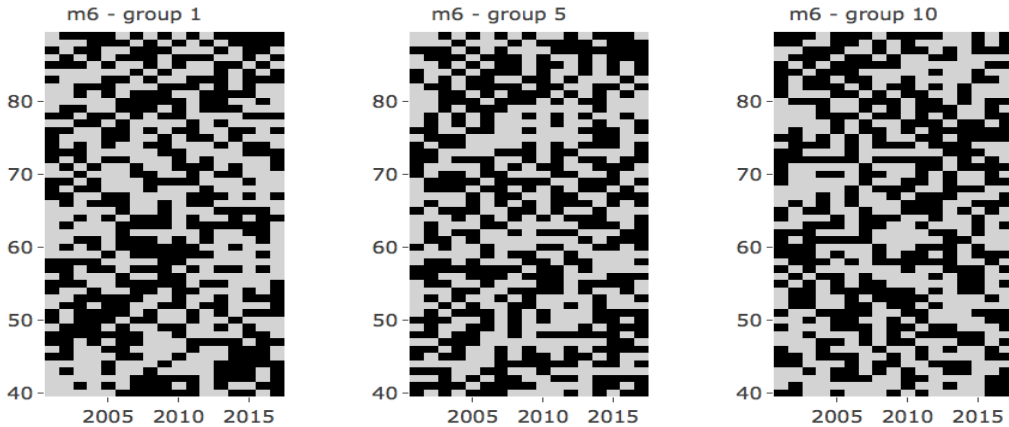

Figure 8: Heatmaps for standardised residuals from model m6 (fitted to female population) over the underlying data range for group 1 (left), 5 (middle) and 10 (right). Black cell indicate positive residuals Z_{xti} and grey cell indicate negative values.

While we do not report heatmaps for other groups, we have assessed those and found similar results. None of the heatmaps shows any non-random cluster of positive or negative residuals.
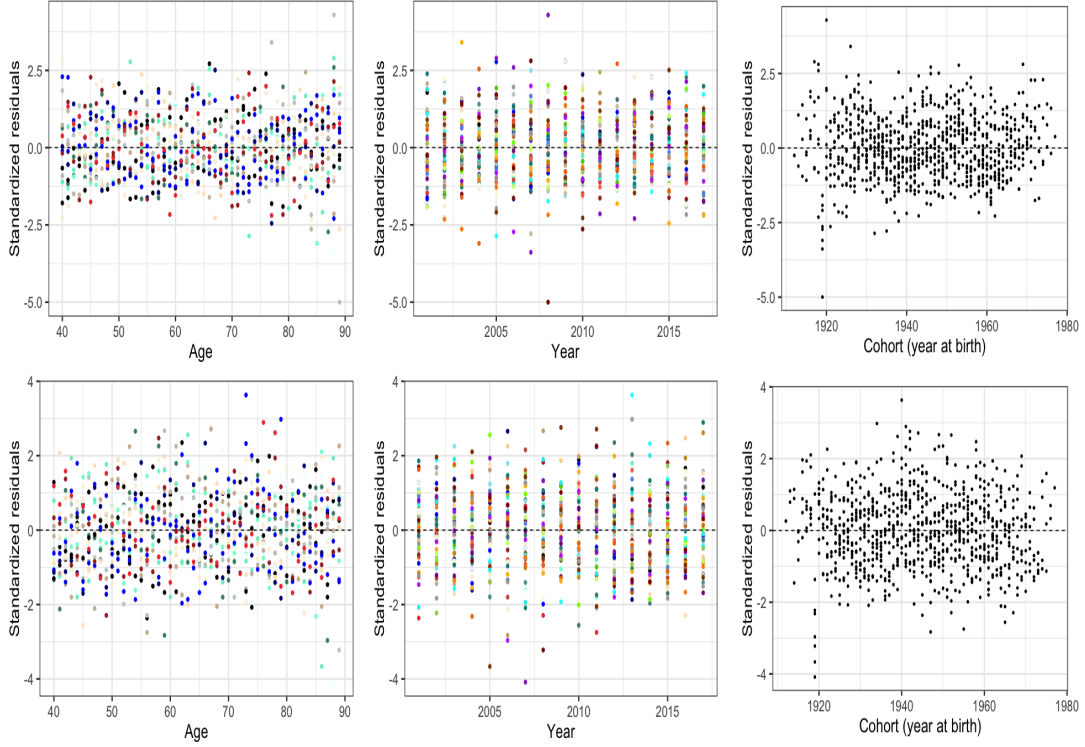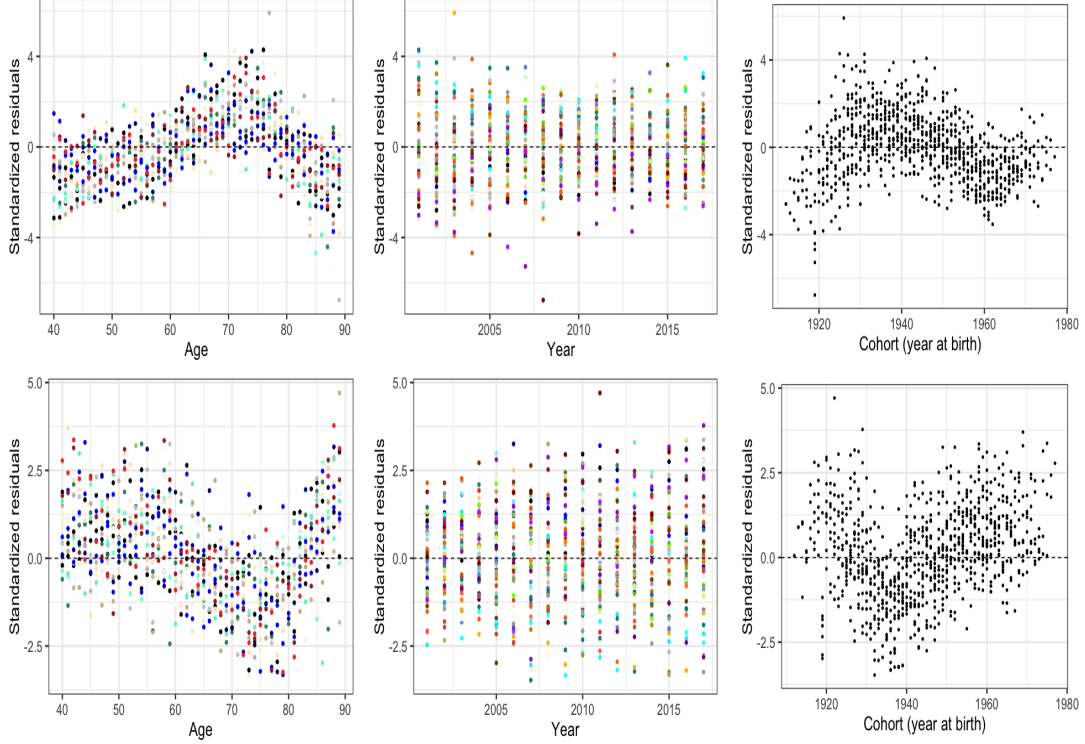
Figure 9: Standardised residuals of m6 (fitted to female population) over age (left), year (middle) and cohort (right) for subgroup 1 (top) and 10 (bottom). The colours in age and year plot represent each underlying year/age.

To investigate residuals further we also consider plots for the residuals as functions of age, year and cohort separately in figure 9. We find that most residuals are between -2 and 2 as we would expect. While there are some larger residuals we find no significant structure or clusters. In particular, there seem to be no data ranges for which residuals have a particularly large or small variance a feature we would be unable to detect from heatmaps. There is, of course, one noticeable exemption: the residuals for cohorts born around 1918 show significantly larger residuals than other cohorts. Again, this shows that model m6 provides a very good fit to the observed mortality data.

Turning to model m8, we plot the heatmaps for the three groups g1, g5 and g10 in Figure 10.

Figure 10: Heatmaps of standardised residuals of model m8 (fitted to female population) for subgroups 1, 5 and 10.

We find in Figure 10 that the residuals from model m8 are significantly different from

those obtained from m6. There are clearly some clusters of positive and negative residuals, in particular, along ages and cohorts. The difference between models m6 and m8 is that the age pattern of mortality is assumed to be linear in model m8. The heatmaps in Figure 10 indicate that this is a rather strong assumption as it seems that there is some remaining structure with respect to age in the residuals for the three groups. The heatmaps also suggest that might be appropriate to include a cohort effect in model m8.

Figure 11: Standardised residuals of m8 (fitted to female population) over age (left), year (middle) and cohort (right) for subgroup 1 (top) and 10 (bottom). The colours in age and year plot represent each underlying year and age.

Considering scatterplots again, we find in figure 11 that the residuals for model m8 show indeed some structure, in particular, along the age and cohort dimensions. This confirms our conclusions drawn from the heatmaps.

6.2 Fitted mortalities diagnostic

A more straight forward graphical diagnostic is to directly compare the shape of the crude death rates with the fitted rates obtained from the models in Table 2 using the estimated parameters. The crude deaths rates are given in Figure 1. We observed in Section 2.3 that there is a clear ranking of socio-economic groups with no strong difference in the variability of rates in different groups.

To compare the observed crude deaths rates to the fitted rates obtained for our models we now reproduce the plots from Figure 1 but for the fitted rates from common age effect model m6, see Figure 12.
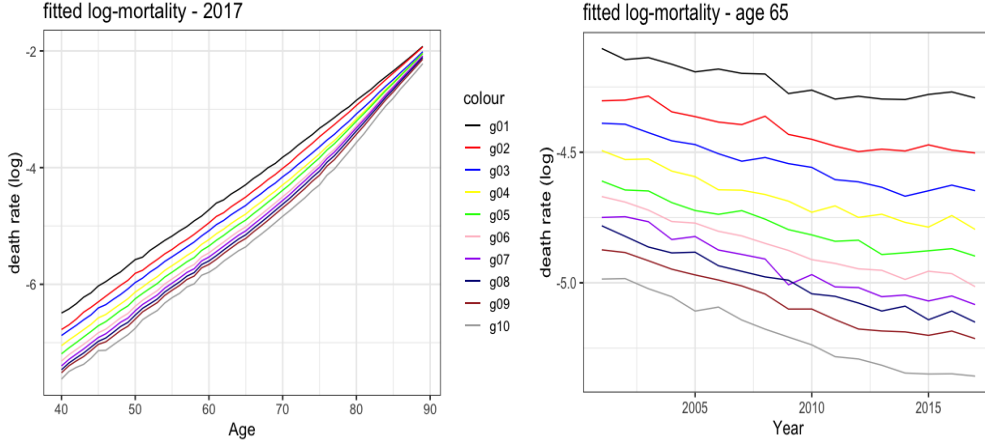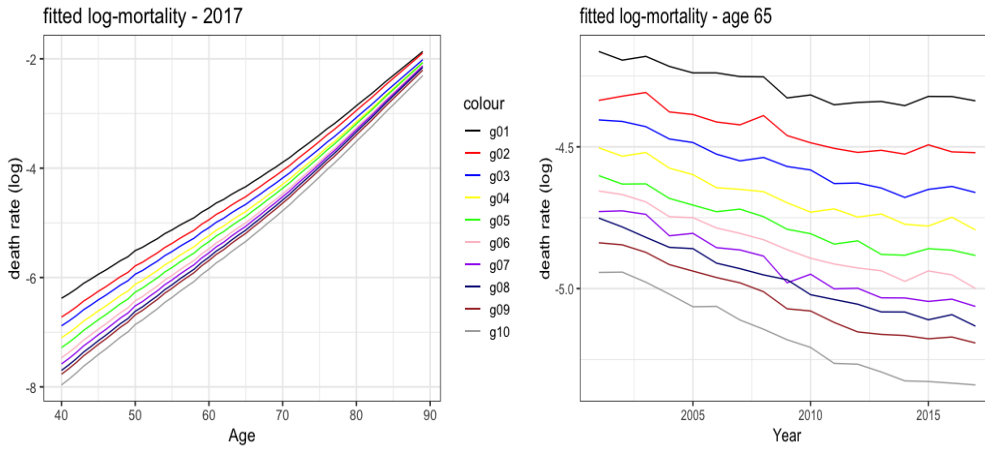
Figure 12: Fitted log death rates (female population) using model m6 for the year 2017 (left) and at age 65 (right).

We observe in Figure 12 that the obtained rates from model m6 are smoother than the observed rates in Figure 1 as we would expect. The model is able to capture some of the features of observed rates, in particular, the differences between groups are decreasing with age (estimated β_x^1 is decreasing in age).

To compare the fitted rates form model m8 with the observed rates we again reproduce the plots in Figure 1 with the fitted rates, see Figure 13.

Figure 13: Fitted log death rates (female population) using model m8 for the year 2017 (left) and at age 65 (right).

We observe in Figure 13 that the parametric forms of β^1 and β^2 in model m8 lead to even smoother functions of age and calendar year than we observed for model m6.

It seems that both models, m6 and m8 are able to capture some of the important features of group-specific death rates. While we have seen that m6 is able to mimic some of the age specific features in our data, we find that the much simpler model m8 is producing very similar fitted rates. So, if for applications, simplicity of the model is more important than the quality of fit, model m8 might be the preferred model.

7 Cohort Effects

We found in Figure 10 that there is some evidence for a cohort effect in the residuals of model m8. This suggests that extending model m8 with a cohort effect will improve the quality of fit. Therefore, we introduce the extended model, called m8c:

$$\log m_{xti} = \alpha_x + \kappa_{ti}^1 + (x - \bar{x})\kappa_{ti}^2 + \gamma_{ci} \qquad \text{for} \qquad c = t - x$$

where γ_{ci} denotes a group-specific effect for the cohort born in year $c = t - x$.

We continue to use maximum likelihood estimation to obtain estimated parameter values. There are, of course, identifiability issues as for other models and we apply appropriate constraints to obtain a unique parameter estimate.
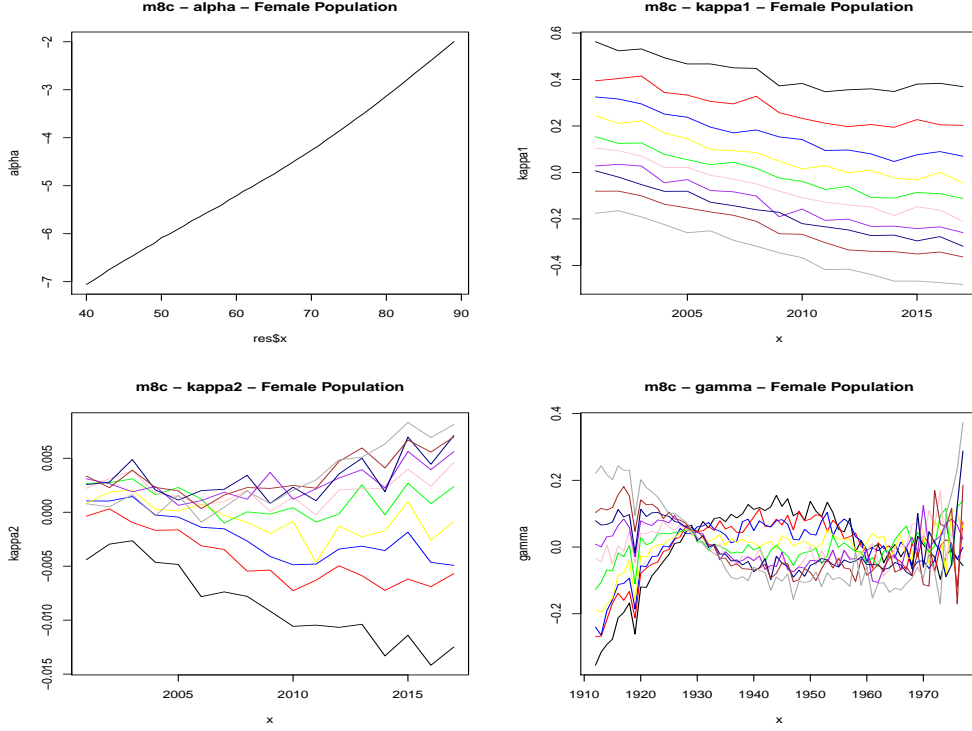

Figure 14: Parameter estimates for model m8c with group-specific cohort effects fitted to the female population.

Figure 14 shows the estimated parameter vectors. Comparing these estimates with those obtained for model m8 in Figure 6, we find that the estimates for α and κ^1 are almost identical for the two models, m8 and m8c. The largest differences are observed in the second order period effect κ_2. This indicates that the cohort effects in model m8c are capturing features in the residuals which the leading parameters α and κ_1 did not pick up. Therefore, it seems that the added cohort effects are providing an improvement to model m8.

The estimated cohort effects look very volatile which suggests a high degree of uncertainty. We also notice that differences between socio-economic groups are narrow for cohorts born at about 1930, but those differences are rather large for other cohorts. Given the high variability of the cohort effects it seems that those large differences between groups seem to be unrealistic.

Ultimately, stochastic mortality models are applied to project death rates. Projections for individual cohort effects need to be constructed carefully to avoid divergence of death rates for different socio-economic groups from each other (see, for example, Hyndman et al., 2013). The task of projecting rates would be simplified if there was just one common cohort effect rather than one effect for each group.

Keeping those comments and the uncertainty about γ_{ci} in mind, we investigate the possibility of a common cohort effect, γ_c, by fitting an extended model m8 with a cohort effect that is common to all socio-economic groups:

$$\log m_{xti} = \alpha_x + \kappa_{ti}^1 + (x - \bar{x})\kappa_{ti}^2 + \gamma_c \qquad \text{for} \qquad c = t - x$$

where γ_c does not depend on the socio-economic group i. The parameter estimates for this model are shown in Figure 15.
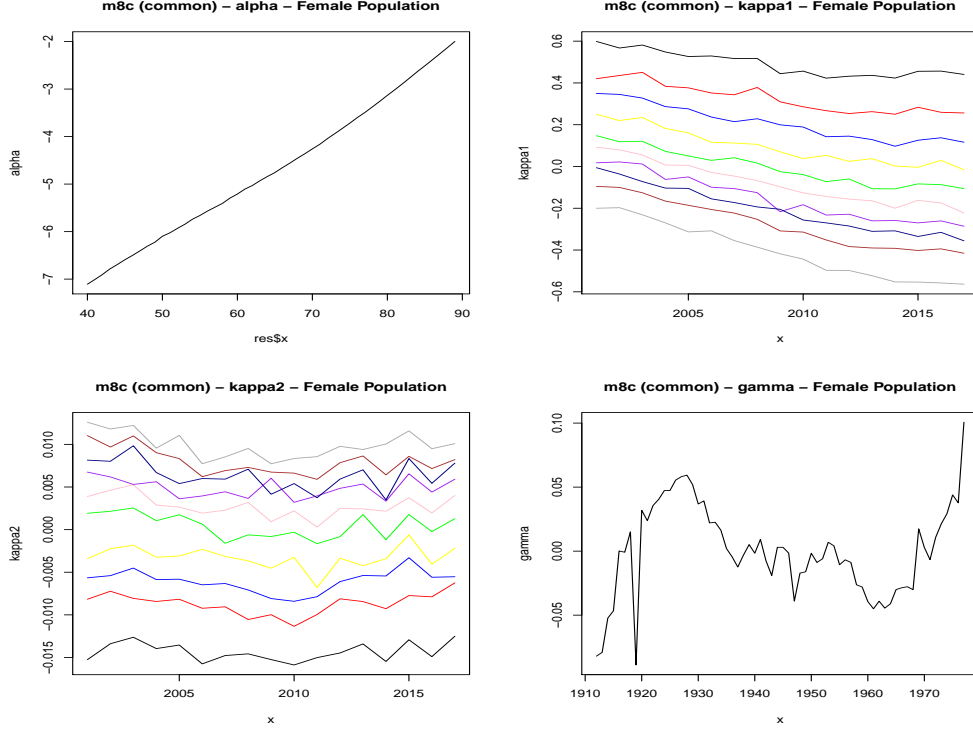
Figure 15: Parameter estimates for model m8c with common cohort effects fitted to the female population.

Again, we observe only minor changes to the estimated age and period effects indicating that the residuals from model m8 are rather stable and the cohort effects are capturing structure in the residuals of this model. The observed common cohort effect is in the same range as the individual effects.

7.1 Goodness of Fit

To obtain further insights into the importance of cohort effects for model m8 we fit the model without a cohort effect, with individual effects and with common cohort effect to our data and compare the BICs. In Table 7 we report the obtained BIC values for males and females. For easy comparison we also include the BIC values for the best fitting model m6.

Model	k	females		males	
		$\log(\hat{L})$	BIC	$\log(\hat{L})$	BIC
m8	388	-37375.07	78260.70	-38213.32	79937.20
m8c	1045	-34612.99	78680.96	-35604.34	80663.65
m8c (common)	451	-36340.44	76761.45	-36562.17	77204.91
m6	486	-35336.06	75069.36	-36702.57	77802.39

Table 7: BIC and log likelihood values for model m8 without cohort effect, with individual cohort effects and with common cohort effect. All models are fitted to IMD deciles for ages 40 to 89. The degrees of freedom are denoted by k.

We find in Table 7 that models with a common cohort effect are better suited to our data than models with no cohort effects or individual cohort effects. However, when we compare model m8 with a common cohort effect to our best fitting model m6 we noticed that the results are not so clear. While model m8 with a common γ outperforms model m6 for the male population we find that the opposite is true for the female population. This suggests that the inclusion of non-parametric age effects in our model is more important for the female population than it is for the male population.

The quality of fit can be investigated further with the graphical diagnostics already applied in Section 6. Heatmaps of standardised residuals are shown in Figure 16.
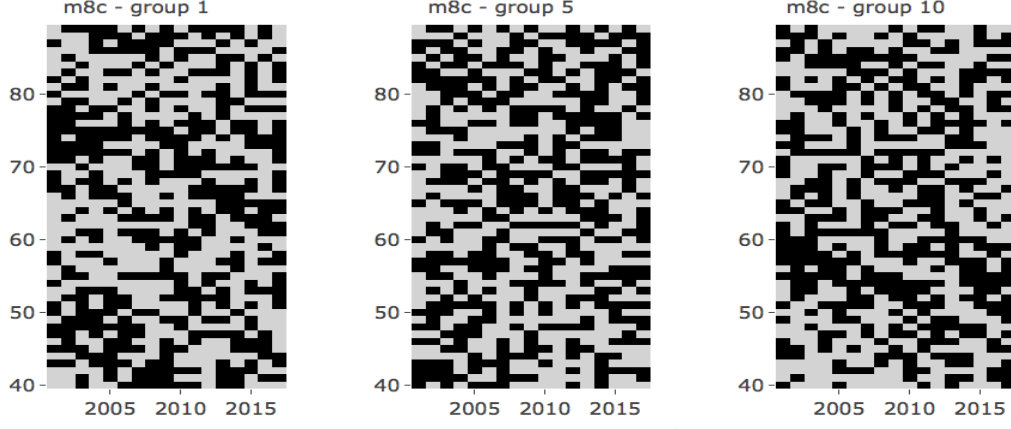
19

Figure 16: Heatmaps for standardised residuals from m8c (fitted to female population) with group-specific cohort effects γ_{ci} for group 1 (left), group 5 (middle) and group 10 (right). Black cell: positive figures; Grey cell: negative figures.

Comparing the heatmaps in Figure 16 with those in Figure 10 we find that the inclusion of group-specific cohort effects removed the remaining structure observed in the residuals of model m8.
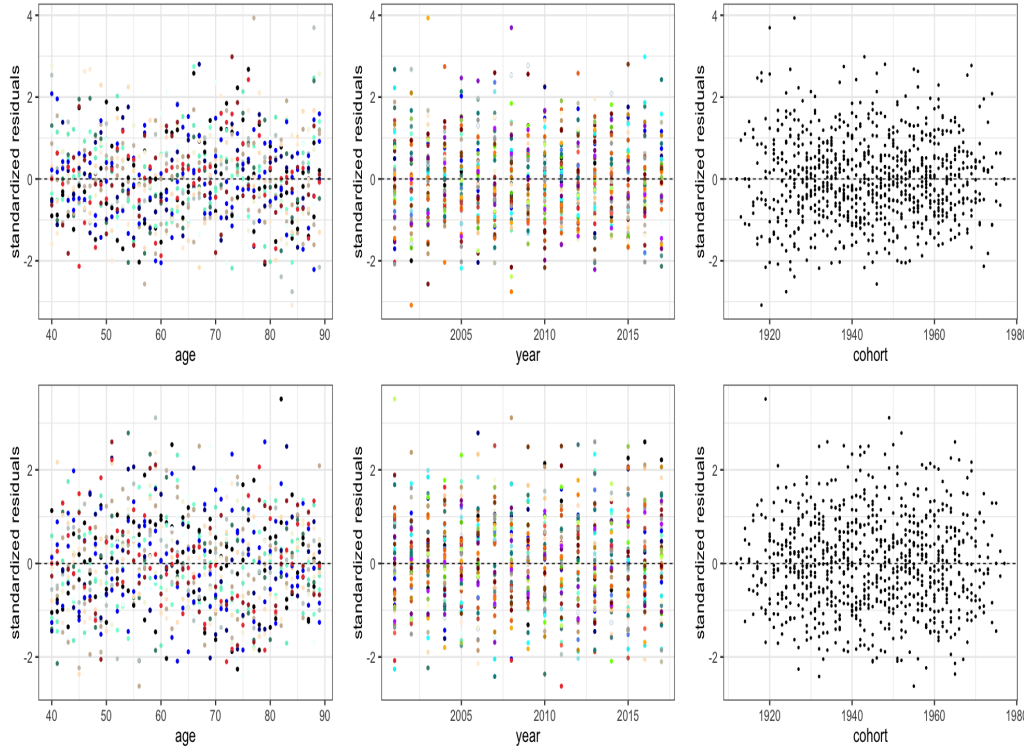

Figure 17: Distribution of standardised residuals by m8c (fitted to female population) over age (left), year (middle) and cohort (right) for group 1 (top) and 10 (bottom) with γ_{ci} applied. Colours in age and year plot represent different years/ages.

Turning to the model with common cohort effect γ_c we again plot the heatmaps of the obtained residuals, see Figure 18. In this figure we clearly observe that there is some (quadratic) structure along the age dimension present in the residuals for groups 1 and 10 that the common cohort effect is unable to capture. This is again an indication that model m6 with the non-parametric age effects is a better choice then model m8 with cohort effects.
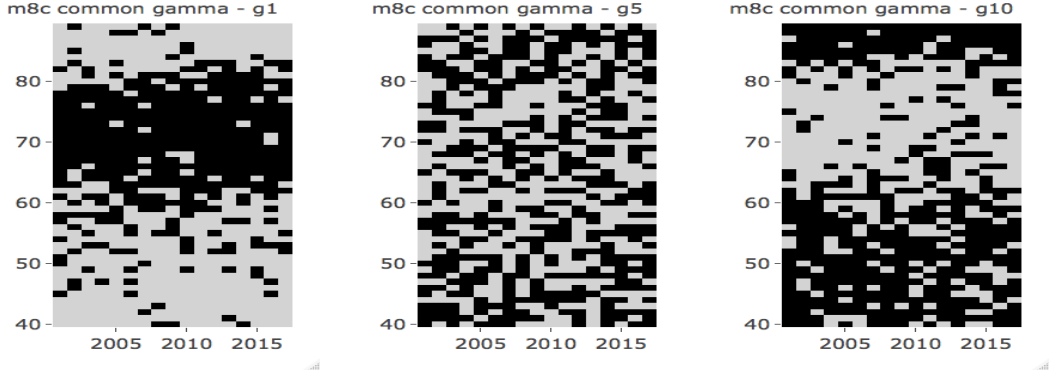
Figure 18: Heatmaps of standardised residuals fitted from m8c with common γ_c (fitted to female population) in group 1 (left), 5 (middle) and 10 (right). Black cell: positive figures; Grey cell: negative figures.
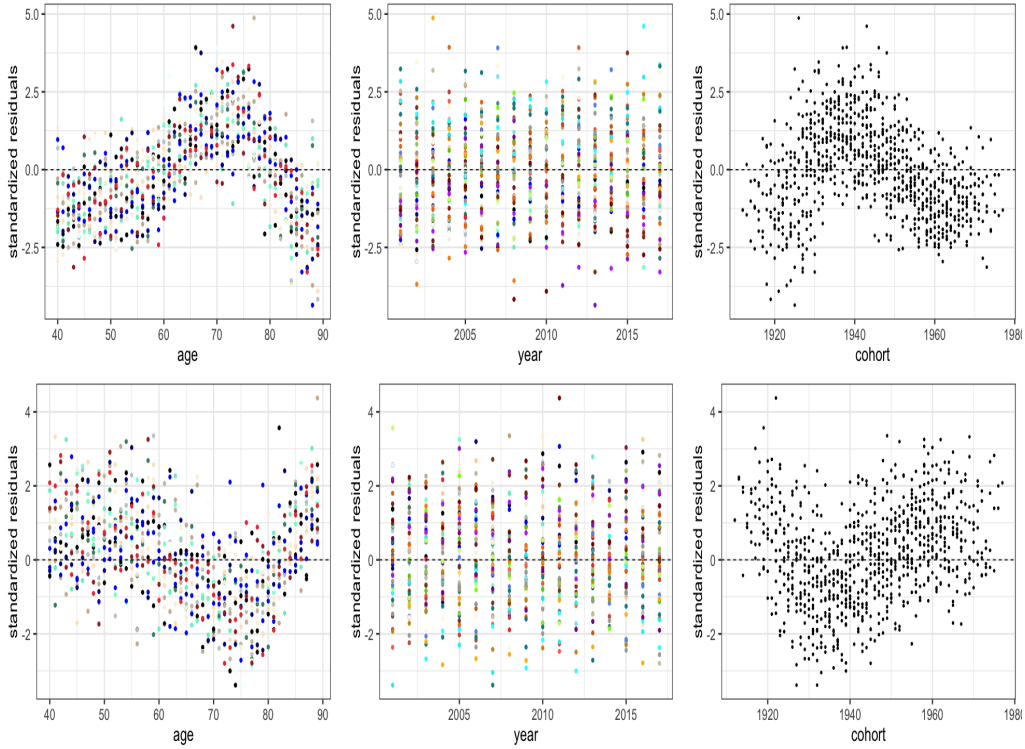

Figure 19: Distribution of standardised residuals by m8c with common cohort effect (fitted to female population) over age (left), year (middle) and cohort (right) for group 1 (top) and group 10 (bottom) with common γ_c applied. Colours in age and year plot represent different years/ages.

Summarising the results in this section for the female population, we conclude that non-parametric age effects are a better choice than cohort effects for ensuring a good fit. However, for the male population we have seen in Table 7 that model m8c with a common γ has a lower BIC value than model m6. To compare the two models further (when fitted to the male population), we plot the heat maps for the residuals of models m6 and m8c in Figure 20.
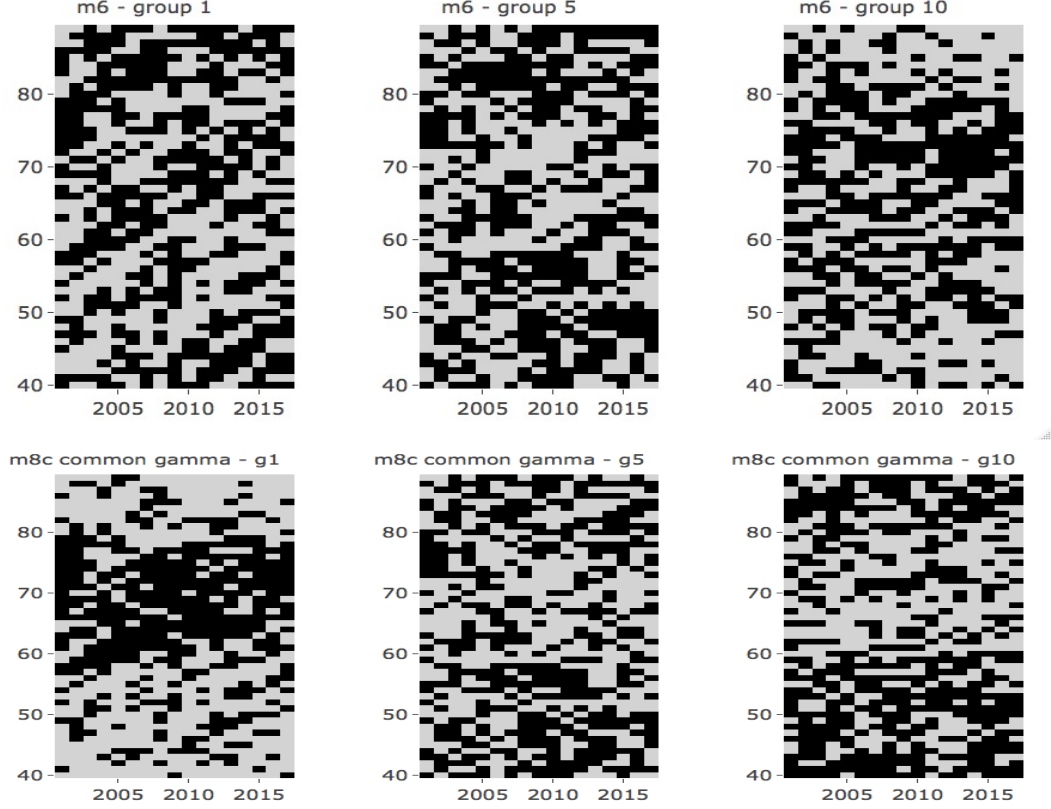
Figure 20: Heatmaps for standardised residuals from model m6 (top row) and model m8c (bottom row). Both models are fitted to the male population over the underlying data range for groups 1 (left), 5 (middle) and 10 (right). Black cell indicate positive residuals Z_{xti} and grey cell indicate negative values.

Comparing Figures 18 and 20 we notice that the residuals of model m8c (fitted to the male population, figure 20) still show some structure in the age dimension for group 1. However, this structure is less obvious than what we observed for the female population (Figure 18), and it has disappeared completely for males in groups 5 and 10. Comparing the model m8c residuals to the residuals from model m6 we now find rather similar pictures for the two models. Those observations are in line with the better BIC value of m8c compared to m6 as model m8c is able to capture most of the structure in the male mortality data with much fewer parameters than model m6.

8   Conclusions

In this paper we compared the goodness of fit of several multi-population mortality models to data for ten similarly-sized socio-economic groups. Our analysis suggests that models which allow for group-specific time trends outperform models with a common first order period effect, in particular the Li&Lee model, suggesting that mortality improvements in the ten IMD deciles are different. This reflects the very different improvement rates observed since the year 2000. For many applications, mortality projections will be required and choosing a model with individual period effects as suggested here, will create the new challenge of finding a parsimonious multivariate time series model for the period effects in the ten groups that produces meaningful projected rates.

Our second most important conclusion from this empirical study is that models with common non-parametric age effects β, like model m6, seem to provide a better fit than models with log death rates that increase linearly in age. Although the inclusion of a cohort effect can be beneficial (e.g. males m8c versus m6), its inclusion tends to be less significant if the model incorporates non-parametric age effects: that is, in models such as m8c the cohort effect is perhaps capturing what should really be modelled as age-period effects. Lastly, we found that individual age effects do not improve the fit of a model.

In summary, we conclude that the effect of age on mortality is similar in all socio-economic groups while period effects are different.

It is also useful to remark that the preferred CAE model, m6, also satisfies the principle of coherence (Hyndman et al., 2013). Other models do not satisfy this principle, and so it is reassuring that the model that (mostly) has the highest BIC is one of those that is coherent.

This analysis is a first step towards projecting death rates in the ten IMD deciles in England, but the empirical evidence found in our study suggests that mortality models for a country can be refined significantly by considering mortality patterns in different socio-economic groups.

Acknowledgments

References

Balia, S. & Jones, A. M. (2008), 'Mortality, lifestyle and socio-economic status', *Journal of Health Economics* **27**(1), 1 – 26.

Cairns, A. J. G., Blake, D. & Dowd, K. (2006), 'A two-factor model for stochastic mortality with parameter uncertainty: Theory and calibration', *Journal of Risk and Insurance* **73**(4), 687–718.

Cairns, A., Kleinow, T. & Wen, J. (2018), Trends in Canadian Mortality By Pension Level: Evidence From the CPP and QPP. Working Paper, Actuarial Research Centre, Heriot-Watt University.

Enchev, V., Kleinow, T. & Cairns, A. J. G. (2015), 'Multi-population mortality models: fitting, forecasting and comparisons', *Scandinavian Actuarial Journal* **online**, 1–24.

Hyndman, R. J., Booth, H. & Yasmeen, F. (2013), 'Coherent mortality forecasting: The product-ratio method with functional time series models', *Demography* **50**(1), 261–283.

Kleinow, T. (2015), 'A common age effect model for the mortality of multiple populations', *Insurance: Mathematics and Economics* **63**, 147 – 152. Special Issue: Longevity Nine - the Ninth International Longevity Risk and Capital Markets Solutions Conference.

Kleinow, T., Cairns, A. & Wen, J. (2019), 'Deprivation and life expectancy in the UK', *The Actuary, April 2019* .

Lee, R. D. & Carter, L. R. (1992), 'Modeling and Forecasting U.S. Mortality', *Journal of the American Statistical Association* **87**(419), 659–675.

Li, N. & Lee, R. (2005), 'Coherent mortality forecasts for a group of populations: An extension of the Lee-Carter method', *Demography* **42**(3), 575–594.

Mackenbach, J. P., Kunst, A. E., Cavelaars, A. E., Groenhof, F. & Geurts, J. J. (1997), 'Socioeconomic inequalities in morbidity and mortality in western Europe', *The Lancet* **349**(9066), 1655 – 1659.

Plat, R. (2009), 'On stochastic mortality modeling', *Insurance: Mathematics and Economics* **45**(3), 393 – 404.

Renshaw, A. & Haberman, S. (2003), 'Lee-Carter mortality forecasting with age-specific enhancement', *Insurance: Mathematics and Economics* **33**(2), 255 – 272. Papers presented at the 6th IME Conference, Lisbon, 15-17 July 2002.

Smith, T., Noble, M., Noble, S., Wright, G., McLennan, D. & Plunkett, E. (2015*a*), The English Indices of Deprivation 2015. Research Report, Department for Communities and Local Government.

Smith, T., Noble, M., Noble, S., Wright, G., McLennan, D. & Plunkett, E. (2015*b*), The English Indices of Deprivation 2015. Technical Report, Department for Communities and Local Government.

Villegas, A. M., Haberman, S., Kaishev, V. K. & Millossovich, P. (2017), 'A comparative study of two-population models for the assessment of basis risk in longevity hedges', *ASTIN Bulletin* **47**(3).