Longevity forecasting by socio-economic groups using compositional data analysis

Kjærgaard, Søren; Ergemen, Yunus Emre; Bergeron-Boucher, Marie Pier; Oeppen, Jim; Kallestrup-Lamb, Malene

Published in: Journal of the Royal Statistical Society Series A: Statistics in Society

DOI: 10.1111/rss.a.12555

Publication date: 2020

Document version: Accepted manuscript

Citation for published version (APA):

Go to publication entry in University of Southern Denmark's Research Portal

Terms of use
This work is brought to you by the University of Southern Denmark. Unless otherwise specified it has been shared according to the terms for self-archiving. If no other license is stated, these terms apply:

- You may download this work for personal use only.
- You may not further distribute the material or use it for any profit-making activity or commercial gain.
- You may freely distribute the URL identifying this open access version.

If you believe that this document breaches copyright please contact us providing details and we will investigate your claim. Please direct all enquiries to puresupport@bib.sdu.dk
Longevity forecasting by socio–economic groups using compositional data analysis

Søren Kjærgaard
Interdisciplinary Centre on Population Dynamics, University of Southern Denmark

Yunus Emre Ergemen
CREATES and Department of Economics and Business Economics, Aarhus University

Marie-Pier Bergeron-Boucher
Interdisciplinary Centre on Population Dynamics, University of Southern Denmark

Jim Oeppen
Interdisciplinary Centre on Population Dynamics, University of Southern Denmark

Malene Kallestrup-Lamb
CREATES and Department of Economics and Business Economics, Aarhus University
Pension Research Centre (PeRCent), Copenhagen Business School

Abstract

Several OECD countries have recently implemented an automatic link between the statutory retirement age and life expectancy for the total population to ensure sustainability in their pension systems due to increasing life expectancy. As significant mortality differentials are observed across socio-economic groups, future changes in these differentials will determine whether some socio-economic groups drive increases in the retirement age leaving other groups with fewer pensionable years. We forecast life expectancy by socio-economic groups and compare the forecast performance of competing models using Danish mortality data and find that the most accurate model assumes a common mortality trend. Life expectancy forecasts are used to analyse the consequences of a pension system where the statutory retirement age is increased when total life expectancy is increasing.

Keywords: Compositional data, forecasting, longevity, pension, socio-economic groups.

JEL Codes: C22, C23, C53, I12.
1 Introduction

Recently, several OECD countries have established an automatic link between their pension systems and increases in life expectancy: examples are Finland, Denmark, Portugal, Italy, the Netherlands, the Slovak Republic, and Sweden (OECD, 2017). The link is either between pension payments and life expectancy, as in Sweden and Italy, or between the statutory retirement age and life expectancy, as in Denmark and the Netherlands (OECD, 2012). In the latter case, life expectancy for the total population is used to regulate the pension system, such that the statutory retirement age will be increased if life expectancy for the whole population increases. Thus, the pace by which life expectancies change in different socio-economic groups (SEG) will have implications for the number of pensionable years. This article analyses the consequences of such a pension system on socio-economic inequalities by forecasting life expectancy by SEG. The aim is to identify the most accurate model for forecasting mortality by SEG by comparing different models, basing the selection on their inclusion of dependence among the SEG. Forecasts are used to measure the implication of the current pension scheme in Denmark in terms of expected years with pension. Our empirical setting focuses on Denmark but as mortality differs significantly by SEG for almost all developed countries (Mackenbach et al., 2003) the outline of the results presented are relevant for all countries which link life expectancy changes with their pension systems.

A successful forecast of mortality differentials between socio-economic sub-populations relies on the model’s ability to capture different aspects of the differentials. Villegas and Haberman (2014) showed that changes in mortality differentials can be modelled successfully using multi-population mortality models and the analysis presented in this article uses this approach. Multi-population models aim at coherently modelling and forecasting mortality data from several populations or sub-populations.

Mortality forecasts are, currently, almost exclusively performed using models which decompose age-specific mortality rates into age, period, and sometimes also cohort effects, inspired by the Lee and Carter (1992) (LC) model, which is the most popular mortality forecasting model in countries with data of high quality (Booth, 2006; Cairns et al., 2009; Coelho and Nunes, 2011; Enchev et al., 2017). One major limitation with the LC type of models is that they generally underestimate improvements in mortality as a result of assuming constant age-specific and
relative improvements (Bergeron-Boucher et al., 2017; Booth and Tickle, 2008). As the LC model is fitted to historical data the model gives a high weight to improvements in mortality for relatively young ages when fitted to data from developed countries as mortality has declined most in these age groups. When this pattern is imposed in forecasts the model fails to capture improvements at higher ages.

Oeppen (2008) and later Bergeron-Boucher et al. (2017) suggested using life table deaths to forecast mortality based on compositional data analysis (CoDa) to alleviate this limitation in the LC models. CoDa mortality models shift deaths from younger ages towards older ages because the covariance structure in compositional data is utilised when analysing life table deaths with CoDa. For example, a decreasing number of deaths at young ages implies that more deaths occur at older ages. Thus, the use of life table deaths has an advantage, compared to models that use mortality rates as it allow the rate of mortality improvements to change. CoDa mortality models are especially useful when forecasting mortality for populations where the mortality patterns are changing e.g. if life expectancy has been stagnating and then begins to experience improving mortality again. This is the case for the Danish population because Denmark in the 1980’s and first part of the 1990’s experienced a stagnation in the improvement of mortality (Jarner et al., 2008). Other countries experienced similar stagnation periods, for example, the Netherlands and the U.S. from around from 1984 to 2000 (Meslé and Vallin, 2006).

The LC type of models, (for example the Lee-Li model), do not capture the shifting patterns which leads to less accurate life expectancy forecasts. In contrast CoDa models allow more interactive dynamics in the observed mortality trajectories, and more accurate forecasts are often found when these models are fitted to data (Bergeron-Boucher et al., 2017). Therefore, the CoDa models offer an attractive alternative to LC type models when forecasting mortality by SEG.

Several CoDa mortality models have been suggested (Oeppen, 2008; Bergeron-Boucher et al., 2017, 2018; Kjærgaard et al., 2019) and this paper discusses the suitability of using CoDa models to forecast mortality in SEG by comparing existing models as well as proposing a new model. The CoDa models were developed in a multi-population framework but their performance has not been tested in a socio-economic setting. The analysis could have included other multi-population models formulated on death rates such as the models presented in Villegas and Haberman (2014) or alternative models using the probability of death such as
Cairns et al. (2006) or Cairns et al. (2009). We do not consider these models because they have the tendency to produce lower life expectancy forecasts with constant age-specific improvements without redistribution of deaths as in the CoDa mortality models, see Bergeron-Boucher et al. (2019) for further details and discussions.

For the total population in a country, an independent modelling of the populations might be reasonable, but for sub-populations within the same country it is likely that factors such as health care, public policy, technology, etc. affect all sub-populations. Factors which affect all sub-populations could be incorrectly specified by the independent model and could lead to implausible forecasts with possibly unbounded divergence among the sub-groups (Villegas and Haberman, 2014). Hence, modelling the dependence between SEG is important but it is not straightforward to determine how and to what extent dependence should be included to obtain the most accurate forecasts. A key contribution of this article is in the selection of models which include models with different levels of dependence, imposed both in the time and age structures of mortality. The models used in the analysis are selected based on their inclusion of time and age dependent parameters so that different degrees of flexibility with respect to the time- and age-dimensions are allowed for. The inclusion of dependence spans from on one side an independent treatment (including 6 parameter vectors both for the time and the age dimension), that is the Independent-CoDa (Inde-CoDa) and Lee-Carter (LC) models. On the other side, we include models with one parameter vector for both the time and age dimension, that is the Relative-CoDa (Rela-CoDa), Three Dimensional CoDa (3D-CoDa), and Lee-Li (LL) models. The LL model is an extension of the LC model to multiple populations. The new model suggested in this article, Dynamic Factor CoDa (Dynam-CoDa), places itself in between these extremes by having one age dimension and multiple time dimensions. Further, by comparing CoDa mortality and models formulated on death rates (LC and LL) we check whether it is necessary to include the intergenerational dynamics that are imposed in the CoDa models.

We find that, out of six different models, CoDa mortality models that model mortality changes for each SEG proportional to a common trend provide the most accurate life expectancy forecasts for Danish males and females. Thus, models allowing for multiple trends and an individual treatment of the SEG are less suitable when forecasting, indicating a high degree of homogeneity among the socio-economic specific mortality trends. Further, models based on death rates (the LC and LL models) provided in general the lowest life expectancy forecasts and less accurate
forecasts compared to the CoDa models assuming a common trend. Hence, these models were not able to capture the mortality development observed in a developed country with shifting mortality patterns. By measuring life expectancy forecasts at the statutory retirement age from 2016 to 2030 we find large socio-economic differences in the number of pensionable years. These differences are expected to persist until 2030 meaning that the changing retirement ages in the Danish pension system are predicted to have the same relative effect by SEG.

2 Danish pension system and socio-economic groups

In 2007, Denmark implemented a pension scheme that gradually increases the pension age in line with the increase in life expectancy, targeting receipt of pension for 14.5 years. The scheme was implemented to finance an increasing number of retirees from large birth cohorts while also taking into account increasing life times in the entire population. The exact rules are complicated (Finansministeriet, 2017) but basically the pension age will be increased if life expectancy exceeds 14.5 years at the statutory retirement age. Hence, the pension age will increase if life expectancy increases regardless of the population subgroups experiencing mortality improvements. Widening life expectancy differentials across SEG would therefore not only imply larger inequality in life span but also a larger difference in the number of years people can expect to receive a labour market and public pensions. For example, if the highest SEG is experiencing a decline in mortality and the other groups experience no change, the pension age will increase leaving the lower SEG with fewer expected years with a pension. Analysing and forecasting life expectancy across SEG is therefore highly relevant when studying the distributional consequences of a pension system.

SEG are, in this study, based on an individual’s income and wealth and not educational status, as is more common, because we want to capture mortality trends that measure the underlying life span inequality in the population over time. This is not well captured by analysing mortality by education if the population experiences large changes in the national educational level as well as compositional changes over recent decades (Brønnum-Hansen and Baadsgaard, 2012). Mortality trends by education thus include a selection effect where, in particular, people with very low or no education consist of a small selected group, in addition to mortality differences caused by different health conditions (Colardyn and Baltzer, 2008). This selection effect is
referred to as the downward bias in mortality from education (Hendi, 2015). Using income and wealth for measuring SEG, groups of approximately equal size can be found because income and wealth are continuous variables, constituting an individual ranking basis. Time consistent mortality trends can thereby be calculated. Cairns et al. (2019) show that it is important to consider both income and wealth as both high income and high wealth are associated with low mortality. For example, an individual person can be well off with a low or medium level of income if he enjoys a sufficiently high accumulation of wealth.

2.1 Data for Danish socio-economic groups

The Danish population was divided into five gender specific SEG, of (almost) equally large size, based on an affluence index following the procedure suggested by Cairns et al. (2019), which is found to produce a consistent and relevant classification of SEG in relation to life expectancy in each sub-group. Cairns et al. (2019) define SEG by weighting individuals’ gross annual income with a factor of 15, compared to their net wealth, that is \( A = W + K \times Y \), where \( A \) is the affluence index, \( W \) is net wealth, \( K = 15 \) is the weighting factor, and \( Y \) is gross annual income. Cairns et al. (2019) find that migration between SEG should be allowed until age 67 after which the groups are fixed. Individual information about income, wealth and marital status was obtained from the Danish central registers. Data are only available from 1985 because reliable information about income at an individual level does not exist for the whole study population prior to 1985. Further details about the socio-economic measure and data can be found in Cairns et al. (2019).

The socio-economic data are available from age 50 and grouped at 100+. Mortality is measured by the death rates \((m_x)\), life table deaths distributions \((d_x)\), or life expectancy \((e_x)\), all calculated using standard life table techniques following Preston et al. (2001). All variables are measured by single age \( x = (50, \ldots, X) \), single year \( t = 1985, \ldots, T \), and \( g = (1, \ldots, G) \) is used to denote a socio-economic group constituting a sub-population, where group 1 is the economically best well off group. To avoid a problem of artificial compression in the life table forecasts we estimate single year age-specific deaths and exposures after the age 100, when fitting the CoDa models (Bergeron-Boucher et al., 2017). We use a penalized composite link model for this as suggested by Rizzi et al. (2015) and calculate single yearly age mortality until age 105.
Other studies have, like Cairns et al. (2019), used income to define SEG, for example: Tarkiainen et al. (2012), who use taxable income and find widening life expectancy differentials for the lowest socio-economic group compared with the others in the Finnish population. Mortality trends by SEG, similar to those presented in this study, are found for Danish males and females by Brønnum-Hansen and Baadsgaard (2012). A steep increase for the lowest socio-economic group during the first part of the data period is found by Brønnum-Hansen and Baadsgaard (2012), suggesting that the large improvements for the lowest group are due to changing conditions in the labour market and the flexibility model of the Danish labour market. The socio-economic classification used by Brønnum-Hansen and Baadsgaard (2012) differs from the one suggested by Cairns et al. (2019) by using disposable income, missing the inclusion of any wealth measure, and by the number of groups.

3 Methods

The LC model is included in the analysis as a benchmark and compared to the proposed CoDa models. For comparison, all models use similar notation for the time and age index.

3.1 The Lee-Carter model (LC)

The LC model is a single population model and, treats to sub-populations independently. The model decomposes age-specific mortality rates $m_{t,x,g}$ by Singular Value Decomposition (SVD), using only the first rank, after having subtracted the average level of mortality. That is,

$$
\log(m_{t,x,g}) = \alpha_{x,g} + \beta_{x,g} k_{t,g} + \epsilon_{t,x,g},
$$

where $\alpha_{x,g} = \frac{1}{T} \sum_t \log(m_{t,x,g})$ is the temporal arithmetic average measuring the general mortality age pattern, $k_{t,g}$ an index over time of the general level of mortality, $\beta_{x,g}$ age-specific responses to the index, and $\epsilon_{t,x,g}$ the iid. error term. Mortality forecasts are calculated with the LC model by extrapolating $k_{t,g}$ using an ARIMA model to produce mortality rate forecasts.
3.2 The Li-Lee model (LL)

The CoDa models are also compared with the multi-population extension of the LC model, the Li-Lee (LL) model. The LL model (Li and Lee, 2005) estimates a common factor term for the average population by applying an LC model to the average death rates. Next, the common factor is subtracted from each socio-economic sub-population and SVD is used to decompose deviations by sub-population specific age and time parameters. The LL model can be written as,

$$\log(m_{t,x,g}) = \alpha_{x,g} + B_x K_t + \beta_{x,g} k_{t,g} + \epsilon_{t,x,g},$$

(2)

where $K_t$ is an index of the general level of mortality for the average population, $B_x$ is the age specific response to changes in the index, and $\epsilon_{t,x,g}$ the iid. error term. The group-specific parameters are interpreted similar to the parameters in the LC model but relative to changes in the common factor term. Following Li and Lee (2005), we assume that the $k_{t,g}$ follow an AR(1) model which is used to calculate life expectancy forecasts, and changes in sub-population specific mortality thus converge to the national level described by the common factor.

3.3 Compositional Data Methods

We present and analyse four different CoDa models and forecast mortality for each sub-population. Three of the models have already been used to forecast mortality in different settings whereas the Dynamic Factor CoDa model is proposed for the first time. The CoDa models differ from the traditional LC modelling by using life table death distributions instead of death rates and by applying compositional data analysis techniques to introduce redistribution of deaths.

Compositional data are defined as a composition with only positive entries summing to a fixed constant and life table deaths are densities summing to 1 in each year, if rescaled to the life table radix, and therefore only contain relative information. The radix of a life table is the assumed number of new born individuals in the life table, also called the root of the life table. The number can be set arbitrarily as it only has a relative meaning (Preston et al., 2001). Aitchison (1982) showed that traditional decomposition methods, for example SVD, do not apply to compositional data as data coordinates cannot vary freely but are constrained to vary.
between 0 and a constant. Instead, it is necessary to transform the data so it can vary freely and back-transform after the decomposition has been carried out (Pawlowsky-Glahn and Buccianti, 2011).

The analysis presented in this paper uses the centered log-ratio transformation (clr) transformation which is the log ratio of the composition of life table deaths divided by its geometric mean, \( g_t = d_{t,1} \cdot d_{t,2} \cdots d_{t,X} \). That is,

\[
clr(d_{t,x}) = \ln \left( \frac{d_{t,x}}{g_t} \right).
\]

All the models except the Inde-CoDa and the LC models account for dependence among sub-populations but differ in their assumptions on the nature of dependence. The Dynam-CoDa model restricts the age dimension but allows different mortality time trends and by including the Dynam-CoDa model in the analysis we explore whether flexibility in the time dimension is important when forecasting mortality by SEG. The Rela-CoDa model constraints both age and time dimensions to a national trend for all sub-populations by modelling deviations from a national common trend. The 3D-CoDa also restricts the sub-groups to follow the same common age and time factors but instead of modelling the residuals, as with the Rela-CoDa model, a third dimension, related to the population-specific pace of mortality, is introduced.

Summing up, we analyse whether dependence among SEG is most useful for forecasting when it is introduced in the time dimension (Dynam-CoDa), by a common national trend (Rela-CoDa), or by the structure in the age and time dimensions (3D-CoDa). It is important to analyse different levels of flexibility in order to determine the most suitable forecasting model. A very complex model might provide a very accurate fit of the observed mortality but could lack ability to forecast mortality due to bias-variance trade off (Hastie et al., 2008). By analysing several models that provide different forecast methods we test for alternative ways that mortality by SEG can be related.

All models capture level differentials by calculating sub-population specific means \( a_{x,g} \) but mortality improvement differentials are captured differently in the models; in Dynam-CoDa, \( g \) mortality time indexes measuring time changes \( k_{t,g} \); in Rela-CoDa by both age specific dynamics \( b_{x,g} \) and \( k_{t,g} \); and in 3D-CoDa by an additional third dimensional parameter vector.
related to population-specific pace of mortality. Life table deaths could also be modelled using other statistical methods than CoDa as long as the constraint in the data is fulfilled so the death sums to the total. For example, in a study by Basellini and Camarda (2019) life table deaths are modelled using Segmented Transformation Age-at-death Distributions model. The models used and developed in this article have the strength of identifying age and time dimensions of the data which makes it possible to answer questions about trend differences among SEG. The CoDA methodology makes it possible to apply these age and time factorized models to life table deaths.

The CoDa models cannot be estimated if the population experiences zero death counts as the log of the life table is calculated (Bergeron-Boucher et al., 2017). However, this was not a problem for the data used in this article. Section H of the supplementary material presents common imputation procedures that can be applied in the case of zero death counts.

Next we introduce the CoDa models used in the study. Figure 1 summaries the models by illustrating the parameter structure.

### 3.4 Independent (Inde-CoDa)

The simplest way to analyse socio-economic sub-populations is to treat each sub-population independently, similar to the LC model but in a CoDa setting. Using CoDa, that means applying the model suggested by Oeppen (2008) to each sub-population. The Inde-CoDa model centres life table deaths for each sub-population \( (d_{t,x,g}) \) by differencing out the age-specific geometric mean. The model uses the operator \( \ominus \) which is the subtraction operator in CoDa. For \( d_x \) and \( \alpha_x \) in a specific year and sub-group, \( d_x \ominus \alpha_x = C[\frac{d_1}{\alpha_1}, \frac{d_2}{\alpha_2}, ..., \frac{d_x}{\alpha_x}] \). The closing operator \( C(\cdot) \) is defined as, \( C[d_x] = [\frac{d_1}{\sum d_1}, \frac{d_2}{\sum d_2}, ..., \frac{d_x}{\sum d_x}] \cdot \bar{K} \), where \( \bar{K} \) is a constant equal to the sum (Pawlowsky-Glahn and Buccianti, 2011). The centred \( d_{t,x,g} \) are approximated by SVD. That is,

\[
clr(d_{t,x,g} \ominus \alpha_{x,g}) = b_{x,g}^1 k_{t,g}^1 + \ldots + b_{x,g}^p k_{t,g}^p + \epsilon_{t,x,g},
\]

where \( k_{t,g}^p \) is an index representing the overall mortality development over time for rank-\( p \) approximation, \( b_{x,g}^p \) the age-specific response changes in \( k_{t,g}^p \), and \( \epsilon_{t,x,g} \) the iid. error term. The

This article is protected by copyright. All rights reserved
changes in mortality are thus decomposed into an age and time dimension. \( b_{x,g}^{p} \) displays how deaths are redistributed in the forecasts, fulfilling the restriction that the total number of deaths needs to be maintained. For positive \( k_{t,g}^{p} \) values, which is the case for all forecast years, deaths are redistributed from ages with negative \( b_{x,g}^{p} \) values towards those with positive \( b_{x,g}^{p} \) values. Mortality forecasts are calculated by extrapolating \( k_{t,g}^{p} \) using ARIMA models. Estimates and forecasts of \( d_{t,x,g} \) are transformed back using the inverse of the \( clr \) transformation and estimates of \( \alpha_{x,g} \) are added. The inverse procedure of \( clr \) ensures that the initial life table constraint is fulfilled so deaths sum to the radix in each year.

### 3.5 Dynamic Factor Coda model (Dynam-CoDa)

One way to incorporate dependence among sub-populations is to estimate \( k_{t} \) time trends for each sub-group and forecast these in a system estimating dependence among time trends but use the same \( \beta_{x} \) all sub-groups. To do so, we propose to stack the life table deaths for each sub-population vertically and perform an analysis similar to the Inde-CoDa. A matrix of size \((T \cdot G) \times X\) of life table deaths is first centred and transformed and from an SVD, \( g \) group-specific time parameter vectors and one age parameter vector are calculated. That is,

\[
clr(d_{t,x,g} \ominus \alpha_{x,g}) = b_{x}^{1}k_{t,g}^{1} + ... + b_{x}^{p}k_{t,g}^{p} + \epsilon_{t,x,g}
\]  

The Dynam-CoDa model is thus assuming \( g \) time indexes describing the time dimension and one age parameter vector common to all sub-populations measuring the age dimension, for each rank approximation. \( \epsilon_{t,x,g} \) is the iid. error term.

If the sub-population mortality trends move together, it is possible that they share common time trends which can be modelled together. We use a multi-level dynamic factor model here to model \( k_{t,g} \) jointly. The multi-level dynamic factor model determines a factor common to all sub-groups and factors which are only shared by one or more of the sub-groups. More specifically the \( k_{t,g}^{1} \) are factorized using:

\[
k_{t,g}^{1} = \gamma_{g}^{1}P_{t} + \lambda_{g}^{1}R_{t} + \epsilon_{t,g}, \quad g = 1, \ldots, G,
\]

where \( P_{t} \) is the vector of factors that pervade all groups, \( R_{t} \) is the vector of factors that pervade
only a subset of groups, and \( \gamma_g \) and \( \lambda_g \) are the corresponding loadings. Using the selection method suggested by Hallin and Liska (2007), one \( P_t \) factor is estimated for Danish males and females, two \( R_t \) factors for Danish females, and one \( R_t \) factor for Danish males. The Dynam-CoDa is thereby incorporating dependence between sub-groups by estimating a common factor for all groups but also factors (\( R_t \)) which are shared only by some of the sub-groups. Details about the multi-level dynamic factor model are presented in the supplementary material section A.

Forecasts of \( P_t \) and \( R_t \) are calculated using ARIMA procedures. Finally, \( k_{t,g}^{1} \) forecasts are perturbed on \( \beta_x \) and forecasts of the life table deaths are calculated by back-transforming and centre the life table deaths forecast.

3.6 Relative model (Rela-CoDa)

The third model forecasts the mortality for each sub-population in relation to the national mortality and is a variation of the model suggested by Bergeron-Boucher et al. (2017).

Rela-CoDa is estimated in two steps. In step one, a simple CoDa mortality model, using rank-1 SVD, is fitted to the national life table deaths and national forecasts of age specific responses and a mortality index are produced. The subscript \( g \) is left out in the first step and a superscript \( N \) added when denoting the national mortality, \( d_{t,x}^N \).

\[
clr(d_{t,x}^N \ominus \alpha_x) = B_x K_t + \epsilon_{t,x},
\]

(7)

In a second step each sub-population is considered after subtracting the geometric means of national deaths from the sub-population specific geometrical means using CoDa perturbation. Both the sub-population specific and national life table deaths are first rescaled so each row sums to one. A rank-1 SVD is used to calculate sub-population specific estimates of \( b_x, k_t \), and a \( \epsilon_{t,x,g} \) iid. error term. That is,

\[
clr(d_{t,x,g} \ominus \alpha_{x,g} \ominus d_{t,x}^N) = b_{x,g} k_{t,g} + \epsilon_{t,x,g}.
\]

(8)
The Rela-CoDa model is thus a variation of the Li and Lee (2005) model within a CoDA framework. Forecasts of $K_t$ and $k_{t,g}$ are calculated with an ARIMA model and an ARMA model, respectively, and the specific AR and MA terms selected using the Akaike Information Criterion (AIC). Rela-CoDa thus assumes stationary $k_{t,g}$ parameters such that the change in mortality will converge for the sub-populations towards the national level, similar to what is normally assumed in the LL model. Forecasts of the life table deaths in each sub-population are calculated by back-transforming and centring the life table deaths forecast according to the initial data constraint. The difference between this model and that of Bergeron-Boucher et al. (2017) is that they use $B_xK_t$ estimates from an average population instead of the national deaths. Arguably in our approach, more of the national variation is accounted for due to the use of national deaths.

3.7 Three-dimensional model (3D-CoDa)

The fourth CoDa model we consider introduces a third dimension to capture sub-population differentials. The model is suggested by Bergeron-Boucher et al. (2018) and applied to Canadian regions. Bergeron-Boucher et al. (2018) suggest, similar to the other CoDa models, to first centre and transform the life table deaths but instead of an SVD approximation, the 3D-CoDa model uses a three-way principal component analysis (Tucker, 1966). The model can be written as,

$$clr(d_{t,x,g} \ominus \alpha_{x,g}) = \sum_{q=1}^{Q} \sum_{p=1}^{P} \sum_{r=1}^{R} w_{qpr}(k_{t,q} \beta_{x,p} \gamma_{g,r}) + \epsilon_{t,x,g},$$

(9)

where the $w_{qpr}$ are elements in a weighting array containing weights between the loading matrices $\beta$, $k$, and $\gamma$.

Thus, the model assumes that all sub-populations, for each rank, share the same mortality index $k_{t,q}$ and the same age responses $b_{x,p}$ but that each sub-population experiences changes in mortality at a different pace measured by the parameter vector $\gamma_{g,r}$. $\epsilon_{t,x,g}$ is an iid. error term. A high degree of similarity in the time and age structures of the mortality development in each sub-population is therefore assumed. In line with the analysis by Bergeron-Boucher et al. (2018) we only consider equal elements of $Q$, $P$, and $R$ for the 3D-Coda model but unequal
elements are analysed with the more general Population Value Decomposition (PVD) which is described in the supplementary material Section C together with comparative results. The PVD model did not provide more accurate forecasts than the 3D-CoDa model.  

Figure 1: Graphical representation of the CoDa models

Sub-populations independently (Inde-CoDa)

Dynamic Factor CoDa (Dynam-CoDa)

Relative model (Rela-CoDa)
4 Results

4.1 Estimation findings

We show parameter estimates for the rank-1 SVD in this section, fitting the models to the whole data period for Danish males (Parameter estimates for Danish females are shown in the supplementary material in Figures A3 to A7.). The Inde-CoDa and 3D-CoDa ($P=R=Q=2$), models are estimated using two ranks and parameter estimates for the second rank are shown in the supplementary material Figures A8 and A9. A rank-1 approximation was found to be sufficient for the Dynam-CoDa model as higher rank approximations did not improve the forecast accuracy for this model. Mortality forecasts are calculated by extrapolating $k_{t,g}$ using ARIMA models. A random walk with drift is generally used as a suitable model similar to other extrapolative mortality models as for example the LC model (Booth and Tickle, 2008).
Figure 2 shows estimates for $\alpha_x$ for the CoDa models and for the LC/LL models. $\alpha_x$ for the CoDa models is bell shaped as it describes the general death distribution whereas $\alpha_x$ is increasing for LC/LL as these models use log death rates for modelling, where the rate of mortality is increasing with age. Differences in $\alpha_x$ by socio-economic status follow the ordering of the SEG with a higher mortality at lower ages for G1 compared to G5. Hence, the part of the mortality differentials between the SEG that is related to differences in the level of mortality follows the ordering of the SEG. Further, because $\alpha_x$ is assumed to be stable over time the ordering in the level differentials component will persist in the forecasts (Villegas and Haberman, 2014).

Figure 3 shows $k^1_t$ estimates and forecasts for all the models which capture the overall mortality development over time in an index. The time indexes are increasing for the CoDa models and for LC and the LL models meaning that mortality has been declining over time from 1985 to 2016 for all models. Parameters for the LC and the LL models are plotted on a different scale for visibility reasons. The scale differences are a consequence of the LC and LL models being based on death rates, as with the $\beta_x$ estimates.

The Inde-CoDa model estimates for $k_t$ show differences across SEG with the steepest increase...
for the lowest SEG meaning that this group experiences the largest improvements in mortality. The second lowest socio-economic group in the Inde-CoDa model is predicted to have the lowest future improvements in mortality. Similar $k_1^1$ patterns are estimated with the Dynam-CoDa model, but different forecasts are produced when dependence among sub-groups is taken into account by the multi level dynamic factor model. The multi level dynamic factor model estimates a global common factor, identifying dependency among the SEG in the mortality index $k_1^1$. The five SEG are predicted to have a similar increase in $k_1^1$ meaning that they are expected to exhibit similar mortality improvements. The common factor is therefore the dominating part of the Dynam-CoDa forecasts implying only small trend differences for the SEG. The CoDa models with common factor terms, 3D-CoDa and Rela-CoDa, both identify
an upward sloping time trend in both estimates and forecasts.

Figure 4: First rank beta estimates and forecasts using different models for Danish males

Note: The LC and LL parameter estimates are plotted on a different scale compared to the CoDa models for visibility reasons.

$\beta_x$ estimates for the models are shown in Figure 4. All the CoDa-models follow a similar pattern which can be described by looking at the $\beta_x$ estimates from the Inde-CoDa model. Over age, the generally increasing pattern in $\beta_x$ means that when $k_{t,g}$ is increasing and becomes positive, deaths are shifted from ages with negative $\beta_x$ towards older ages where $\beta_x$ is positive. For the Inde-CoDa model $\beta_x$ is, for G1 and G2, decreasing at ages 50 to 70 years and increasing over age until around age 102 followed by a decrease. For G3 to G5 no decrease is found for ages 50 to 70. G1 and G2 have the highest $\beta_x$ estimates for the ages 50 to 60 but lowest from age 60 to 95 compared to other groups. The relatively high $\beta_x$ values at ages 50 to 60 and subsequent low values at ages 60 to 96 for the G1 and G2 groups imply that fewer deaths are transferred to
higher ages for the same shift in $k_{t,g}$ values: these groups will experience a lower improvement in life expectancy for the same increase in $k_{t,g}$ compared to the other groups.

$\beta_x$ parameter estimates for models assuming a common $\beta_x$ vector generally follow patterns observed for the groups G3 to G5. Assuming a common $\beta_x$ parameter thus provides a better fit for these groups than for G1 and G2. Similar patterns are found for the LC/LL models with relative high values for G1 and G2 at ages 50 to 60. Note that these models do not directly imply transfer of deaths as with the CoDa-models.

Figure 5: Group specific kappa and beta estimates for the LL and Rela-CoDa models for Danish males

Figure 5 shows the group specific parameter estimates for the Rela-Coda and LL models, that is the group specific adjustments to the common trend identified in the Rela-CoDa and LL models. Because the group specific $k_{t,g}$ terms are assumed to be stationary, and thus mean reverting, all SEG are assumed to follow the same long-run mortality trend dominated by the common $K_t$ term, as seen in panel a). No particular trending pattern is found for $k_{t,g}$ parameters in the Rela-CoDa model meaning that the stationarity and thus mean reverting
pattern fits the Rela-CoDa model well. In panel c, the $k_{t,g}$ parameters in the LL model display more upward and downward trending patterns making it harder to forecast and less suitable to assume a mean reverting pattern. This makes the Rela-CoDa formulation more attractive for forecasting as the patterns are easier to predict. The different patterns in the Rela-CoDa and the LL models are a consequence of Rela-CoDa subtracting $d_{t,g}^N$ instead of parameter estimates $K_tB_x$ as in the LL model. The age- and group-specific age responses ($\beta_{x,g}$) are similar for all age groups and describe how each age group responds to changes in $k_{t,g}$. Note that $\beta_{x,g}$, in panel d) Figure 5, for Rela-CoDa shows large differences by age and less by SEG and, thus, plotted close together.

Figure 6: Gamma estimates for Danish males using the 3D-CoDa model

The $\gamma_g$ estimates for the 3D-CoDa model show how the common factor terms are scaled for each socio-economic group (Figure 6). Higher $\gamma_g$ estimates are found for the lower SEG meaning that, in the view of the 3D-Coda model, mortality is decreasing more slowly for those groups compared to higher SEG. The only exception is G1 where the fastest decline is found.

Figures A12 to A16 in the supplementary material show standardized residuals for all the models studied across all SEG. None of the residuals shows any particular pattern and thus we do not include specific cohort terms in the models.²

4.2 Out-of-sample comparison and selection of forecasting model

To determine which mortality model is most suitable for forecasting mortality we compare the out-of-sample forecast performance of the different models. Data are available from 1985 to 2016. To have a sufficient number of years for fitting the models, we consider forecasts with
a length of 5 to 11 years calculated by rolling the onset of the forecasts (Shang, 2015). That is, for the first forecast we use a fitting period from 1985 to 2005 and the period from 2005 to 2016 for validation. For the second forecast, one year is added to the fitting period by reducing the validation period by one year. From this a minimum of 2/3 of the data period is used to fit the models and 1/3 for validation. Forecast errors are measured by the root-mean-square error (RMSE) comparing observed and forecast life expectancy at age 50,

$$\text{RMSE} = \sqrt{\frac{1}{H} \sum_{h=1}^{H} (e_{h,50} - \tilde{e}_{h,50})^2}$$

where $h \in \{1, 2, ..., H\}$ is the number of forecast years and $e_{50}$ the observed life expectancy at age 50 and $\tilde{e}_{50}$ the corresponding forecast. The average RMSE is calculated by averaging over the different forecast horizons and used to compare the models. We use life expectancy at age so for comparison as it is calculated using mortality information for all the considered age groups. A good in-sample fit can be achieved by introducing a large number of parameters but this will not guarantee a good forecast. Because forecasting is the objective of this paper, only the out-of-sample performance of the models is considered when selecting the most suitable model. RMSE's are shown in Table 1 and 2 for Danish males and females, respectively. In-sample fit measures can be found in the supplementary material Section F.

<table>
<thead>
<tr>
<th></th>
<th>SEG1</th>
<th>SEG2</th>
<th>SEG3</th>
<th>SEG4</th>
<th>SEG5</th>
<th>Avr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inde-CoDa</td>
<td>0.4511</td>
<td>0.8159</td>
<td>0.8970</td>
<td>0.8657</td>
<td>0.5136</td>
<td>0.7087</td>
</tr>
<tr>
<td>Dynam-CoDa</td>
<td>0.7190</td>
<td>0.6016</td>
<td>0.7286</td>
<td>0.8333</td>
<td>0.6419</td>
<td>0.7049</td>
</tr>
<tr>
<td>Rela-CoDa</td>
<td>0.7088</td>
<td><strong>0.1980</strong></td>
<td><strong>0.3615</strong></td>
<td>0.5626</td>
<td>0.5884</td>
<td><strong>0.4838</strong></td>
</tr>
<tr>
<td>3D-CoDa</td>
<td>0.7200</td>
<td>0.5040</td>
<td>0.4641</td>
<td><strong>0.5117</strong></td>
<td><strong>0.4110</strong></td>
<td>0.5221</td>
</tr>
<tr>
<td>LC</td>
<td>0.6637</td>
<td>0.5284</td>
<td>0.5487</td>
<td>0.6442</td>
<td>0.4878</td>
<td>0.5746</td>
</tr>
<tr>
<td>LL</td>
<td>0.7459</td>
<td>0.3212</td>
<td>0.5617</td>
<td>0.8514</td>
<td>0.9750</td>
<td>0.6910</td>
</tr>
</tbody>
</table>

**Note:** Lowest RMSE forecast error is indicated with bold font

For the Danish males, the Rela-CoDa, 3D-Coda, and Inde-CoDa models provide the lowest forecast errors for the different SEG. For the Danish females, the Rela-CoDa and 3D-CoDa models are the most accurate. The Rela-CoDa model provides the lowest forecast error on average for all the groups for Danish males and thus we conclude that this model is the most accurate model for forecasting mortality by SEG. Similarly, the 3D-CoDa provides the lowest forecast error for females, but the accuracy is just slightly higher for the Rela-CoDa model. A
Table 2: RMSE of $e_{50}$, average over 7 forecast horizons for Danish females

<table>
<thead>
<tr>
<th></th>
<th>SEG1</th>
<th>SEG2</th>
<th>SEG3</th>
<th>SEG4</th>
<th>SEG5</th>
<th>ASvr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indt-CoDA</td>
<td>0.3243</td>
<td>0.8037</td>
<td>0.8056</td>
<td>0.8551</td>
<td>1.2521</td>
<td>0.8081</td>
</tr>
<tr>
<td>Dynam-CoDa</td>
<td>0.8396</td>
<td>0.8327</td>
<td>0.7752</td>
<td>0.9767</td>
<td>0.9456</td>
<td>0.8740</td>
</tr>
<tr>
<td>Rela-CoDa</td>
<td>0.6475</td>
<td><strong>0.6366</strong></td>
<td><strong>0.5778</strong></td>
<td><strong>0.7384</strong></td>
<td>0.8240</td>
<td>0.6848</td>
</tr>
<tr>
<td>3D-CoDa</td>
<td><strong>0.2797</strong></td>
<td>0.7305</td>
<td>0.7581</td>
<td>0.8308</td>
<td><strong>0.7846</strong></td>
<td><strong>0.6767</strong></td>
</tr>
<tr>
<td>LC</td>
<td>0.3460</td>
<td>0.8033</td>
<td>0.7827</td>
<td>0.8298</td>
<td>0.9123</td>
<td>0.7348</td>
</tr>
<tr>
<td>LL</td>
<td>0.7968</td>
<td>0.6709</td>
<td>0.6927</td>
<td>0.8805</td>
<td>0.9745</td>
<td>0.8030</td>
</tr>
</tbody>
</table>

Note: Lowest RMSE forecast error is indicated with bold font.

good fit for the Rela-CoDa and 3D-CoDa models indicates a high degree of homogeneity in the Danish mortality trends across SEG and a model based on a national trend is thus useful when forecasting. Further, it also shows the value of modelling the dependence between the subgroup specific mortality trends. As the Dynam-CoDa model did not provide better forecasts, in general, it does not seem relevant to allow for different patterns in the time trends. It is sufficient to have the same time and age pattern for each group and adjust by modelling the residuals from a national pattern as in the Rela-CoDa model or by allowing for a different pace as in the 3D-CoDa model. A model based on death rates (LC or LL) did not provide the lowest forecast error for any of the groups. Hence, better forecasts are obtained for Danish mortality data using life table deaths when forecasting life expectancy.  

5 Implication for the pension age and its developments

Having identified the CoDa models with a common trend as the most suitable models for forecasting life expectancy for Danish SEG, these forecasts and remaining life expectancy at the pension age are reported for both Danish males and females in Figure 7c and 7d until 2030. Life expectancy forecasts are calculated using the Rela-CoDa model for each sexes as this model provided the lowest RMSE forecast error for males and just slightly higher forecast errors for females compared to the 3D-CoDa model. This ensures coherence in the analysis of the consequences for the pension system as the same model is used for both sexes. Figure 7a shows the statutory retirement age in Denmark until 2030 and finally, Figure 7b shows life expectancy forecasts for Danish males at age 50 for completeness.

At age 50, life expectancy for the lowest socio-economic group is converging towards the other
groups at a decreasing pace from 1985 to 2026. Mortality differentials for the other groups stay roughly the same during the data period. Similar trends are observed for Danish females and shown in the supplementary material.

In Figure 7c, Danish males in the lowest socio-economic group would have a remaining life expectancy at pension age of around 16 years in 2016 which is 4.5 years less than the highest group. The other groups fall between with remaining life expectancy from around 17 to 19 years in 2016. Danish males are, for all groups, forecast to have slightly falling life expectancy at pension age meaning that pension age is expected to increase faster than life expectancy until 2030. For Danish females in Figure 7d, G1 and G2 have the same remaining life expectancy at pension age at around 19 years in 2016 and the similarity remains in the forecast. The other SEG follow with the highest life expectancy for G5 at around 23 years. Note that females in G1 and G2 have almost the same remaining life expectancy and thus their life expectancies are plotted close together. All groups, for both males and females, will have a life expectancy more than 14.5 years, which is the desired long term goal in the current pension scheme, despite the social inequality of pensionable years. Thus, all SEG can expect to receive a public pension in more years than the politically desired number of years if they retire at the statutory retirement age.
Future mortality differentials by SEG are highly relevant for determining the consequences of the current pension system in Denmark. Life expectancy forecasts show that relatively large differentials are expected 14 years ahead between SEG. As a strong common mortality trend was found across SEG, increases in the pension age are predicted to have a similar effect across SEG. No particular group is expected to drive future changes in the total life expectancy. Another implication of the strong common trend is that mortality differentials between the groups persist. Thus, the Danish pension reform in 2007 does not introduce further inequalities than already observed in the predicted number of life expectancy years after the statutory retirement age.
6 Concluding remarks

This study provides life expectancy forecasts for the Danish population by SEG. SEG were measured with an affluence index constructed by weighting income and wealth. Two models using death rates and five models using life table deaths were compared and, based on the models' out-of-sample performance, the Rela-CoDa and 3D-CoDa models were found to most accurately forecast life expectancy for males and females, respectively. Both models use a common trend for all sub-groups to forecast mortality. The six models differed by their method for of including dependence among the SEG: the Inde-CoDa and LC models treat the SEG independently, the LL and Rela-CoDa in relation to a common mortality level, the Dynam-CoDa model by relating multiple mortality trends, and the 3D-CoDa by scaling common age and time structures. That the Rela-CoDa and 3D-CoDa models provided the most accurate forecast indicates the existence of a high degree of homogeneity in the mortality trends between the Danish SEG, so common $k_t$ and $b_x$ parameters could be assumed for all SEG when forecasting.

Despite the similarity in the trend by which mortality changes, large mortality differentials were observed throughout the data period. Models formulated on death rates did not provide the most accurate forecast indicating that these models did not capture changes in life expectancy in the validation period.

Populations in many OECD countries are expected to age in the years ahead, increasing the cost of pensions and health care (European Commission, 2018). As a consequence several countries in the OECD have linked changes in their pension system to changes in life expectancy for the whole population (OECD, 2017). In this, the relatively large mortality differentials between SEG also constitute a distributional issue because the lower SEG have less private pension and lower capital income from saving making them more dependent on the public pension (Pensionskommissionen, 2015). The lower groups are thus affected more when the statutory retirement age is increased because they do not have sufficient wealth to retire before the statutory retirement age. The consequences of future mortality differentials are therefore larger today than before.

The mortality differentials are also important for the public support of the pension reforms as mortality improvements in the future could be distributed unequally and thus implicitly new reforms will have unanticipated consequences. However, the results for Denmark indicate that
the inequalities are not increased further for Denmark, but this might not be the case in other countries. Further research should include data from other countries but maintain the focus on forecasting mortality by SEG at retirement ages.

Forecasts of mortality in general and of mortality differentials are not only relevant from an individual and public point of view but also for private pension companies (Richards, 2008) that could experience a mismatch in the composition of SEG between their insured population and the national population. Determination of possible mismatch is also highly relevant for the risk management of pensions as it can ensure a better hedge of the risk that a pensions fund’s customers live longer than anticipated - longevity risk (Cairns et al., 2019). Forecasts of the mortality differential could be used to inform pension companies about their longevity risk and improve the hedging of the longevity risk.

Notes

1 None of the models include any covariates such as GDP, smoking prevalences, or other relevant factors. The main problem of including covariates is that many of these covariates are often harder to forecast than the mortality patterns themselves. Thus, forecasts including covariates are often found to be less reliable (Cairns et al., 2011). A few studies have included covariates in the LC model, for example French and O’Hare (2014), but it is beyond the scope of this paper to include covariates in the CoDa mortality model framework. These extensions are left for further research.

2 Cohort terms can be included in the models but cannot be identified uniquely because of the exact relation between age, calender year and cohort birth year (cohort = calender year - age).

3 We test the significance of forecast performance differences of the seven models using a Clark-West test (Clark and West, 2006) in the medium horizon. Details about the Clark-West test are presented in the supplementary material Section D. We test the forecast differences for each socio-economic group towards the best performing model for each sex and find that a large majority of the forecasts are significantly different. Results of the Clark-West test are presented in the supplementary material Table A3.

References


This article is protected by copyright. All rights reserved


This article is protected by copyright. All rights reserved


This article is protected by copyright. All rights reserved


Sub-populations independently (Inde-CoDa)

\[ \text{clr}(d_{txi} \Theta \alpha_{xi}) \rightarrow k_t \rightarrow \beta_x \]
This article is protected by copyright. All rights reserved.
Three dimensional model (3D-CoDa)

This article is protected by copyright. All rights reserved
a) Group specific $\kappa_t$ – Rela–CoDa

Years

$\kappa$

-0.6

-0.4

-0.2

0

b) Group specific $\beta_x$ – Rela–CoDa

Ages

$\beta$

-0.6

-0.4

-0.2

0

c) Group specific $\kappa_t$ – LL

Years

$\kappa$


-0.6

-0.4

-0.2

0

d) Group specific $\beta_x$ – LL

Ages

$\beta_x$

(50 60 70 80 90 100)

0.04

0.02

0.00

0.00

This article is protected by copyright. All rights reserved
a) Statutory retirement age by year

b) Life expectancy for Danish males using the Rela–CoDa model

c) Remaining life expectancy at the statutory retirement age, males

d) Remaining life expectancy at the statutory retirement age, females

This article is protected by copyright. All rights reserved