ROYAL
STATISTICAL
SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

Appl. Statist. (2019)
68, *Part* 5, *pp.* 1351–1370

Forecasting causes of death by using compositional data analysis: the case of cancer deaths

Søren Kjærgaard,

University of Southern Denmark, Odense, Denmark

Yunus Emre Ergemen,

Aarhus University, Denmark

Malene Kallestrup-Lamb

Aarhus University and Copenhagen Business School, Denmark

and Jim Oeppen and Rune Lindahl-Jacobsen

University of Southern Denmark, Odense, Denmark

[Received August 2018. Final revision May 2019]

Summary. Cause-specific mortality forecasting is often based on predicting cause-specific death rates independently. Only a few methods have been suggested that incorporate dependence between causes. An attractive alternative is to model and forecast cause-specific death distributions, rather than mortality rates, as dependence between the causes can be incorporated directly. We follow this idea and propose two new models which extend the current research on mortality forecasting using death distributions. We find that adding age, time and cause-specific weights and decomposing both joint and individual variation between different causes of death increased the forecast accuracy of cancer deaths by using data for French and Dutch populations.

Keywords: Cancer forecast; Cause-specific mortality; Compositional data analysis; Forecasting methods; Population health

1. Introduction

Forecasting mortality by cause of death can provide valuable information for healthcare and social services planning in general. In particular, the future distribution of deaths by cause is of interest given the potential to target public health actions efficiently. The future cause-specific death distribution may be related to the relative risk of a disease and be used to predict future rates of incidence. In many countries, incidence forecasts are used for planning future hospital capacities, e.g. for cancer patients (Rapiti *et al.*, 2014). A major challenge when modelling cause-of-death data is the competing risks between causes. In a conventional all-causes-combined life table there is a one-to-one relationship between the mortality rates and the cumulative incidence of deaths. This relationship is lost when cause-specific deaths are analysed because survival may be influenced by all the causes, even if the cause-specific mortality rates are independent; see

Address for correspondence: Søren Kjærgaard, Department of Public Health, Centre on Population Dynamics, University of Southern Denmark, J. B. Winsløws Vej 9B, Odense Syddanmark 5000, Denmark.
E-mail: skjaergaard@sdu.dk

0035–9254/19/681351

Andersen *et al.* (2012). Analysts who are interested in the future pattern of deaths from a specific cause cannot ignore the effects of the other competing causes.

Among demographers and actuaries, cause-specific mortality is conventionally analysed by using a multiple-decrement life table where age- and cause-specific mortality rates are calculated by using age- and cause-specific deaths and the total number of individuals at risk in the population at a given age for a specific year. The underlying assumption in the multiple-decrement life table is that cause-specific death rates are independent so a change in the death rate for cause i does not imply a change in the death rate for cause j. However, the survival probabilities in the life table for each cause are not independent because they are calculated on the basis of all the cause-specific death rates so the proportion of people dying from a specific cause may be affected by the mortality from other causes; see Preston *et al.* (2001) for further details.

The multiple-decrement life table is a standard tool in demography and enables the calculation of cause-specific mortality and thus it is used to calculate the input for all the models that are analysed. The same approach is followed by all other cause-specific mortality forecasting models. Dependence between cause-specific mortality also relates to the time-dependent processes underlying changes in all-cause and cause-specific mortality. For example, behavioural changes in the population may affect multiple causes of death over a longer time period. Modelling of this dependence relates to the choice of model for modelling and forecasting. Generally, this dependence has been ignored as each cause is modelled independently of the others. Several studies use the Lee–Carter (LC) model (Lee and Carter, 1992) to forecast cause-of-death mortality or age–period–cohort models and apply the models independently to each cause, e.g. Wilmoth (1995), Peltonen and Asplund (1996), Knorr-Held and Rainer (2001) and Cesare and Murphy (2011). Regression models have also been used to forecast mortality by cause of death (Mathers and Loncar, 2006), but in their approach causes are treated individually. Girosi and King (2008) related causes of death in a Bayesian framework by treating the different causes as spatially different groups; however, the estimation was not carried out jointly for all the causes—data from different countries are included in a similar way.

Independent modelling and forecasting of cause-specific mortality is not only unattractive because it ignores dependence patterns among the causes, but also because forecasts often fail to be coherent in the sense that cause-specific deaths must sum to the total number of deaths, which could lead to implausible forecasts. Recently, some mortality forecasting models have been suggested which include dependence between different causes of death and thus incorporate competing risks among causes of deaths: Oeppen (2008), Arnold-Gaille and Sherris (2013), Foreman *et al.* (2017) and Hirz *et al.* (2017). In contrast with the other models proposed, Oeppen (2008) used life table deaths to forecast cause-specific mortality by using compositional data analysis (CODA). CODA is a well-established set of statistical methods for the analyses of compositional data which is defined as a data vector with positive elements summing to a constant value and thereby contains only relative information (Pawlowsky-Glahn and Buccianti, 2011). Life table deaths sum to the life table's initial birth cohort in each year across all causes and age. Thus, CODA enables a coherent and correct modelling of dependent causes of death by recognizing the sum constraint.

Oeppen (2008) used an LC type of model within CODA to forecast both all-cause mortality and cause-specific mortality. Wilmoth (1995) argued that an aggregation of individual causes always leads to more pessimistic forecasts than an all-causes-combined forecast but Oeppen (2008) showed that this is not so when modelling life table deaths because of the dynamics in CODA. CODA was also used by Bergeron-Boucher *et al.* (2017) for coherent forecasts of all-cause mortality in different countries. One important difference between modelling cause-specific deaths rates and deaths distributions is that deaths are directly dependent on each other on an aggregated level, as avoided deaths from one cause will result in an increase of deaths from some other cause. The same mechanism does not apply to death rates as death rates are defined as deaths divided by the number of people at risk. Avoided deaths from one cause thus affect both the numerator and the denominator, and the dependence between rates is therefore not as easy to predict (Preston *et al.*, 2001).

This study examines the limitations of the Oeppen (2008) model by suggesting two new models: the CODA model with cointegrating vector error correction model VECM-CoDA that allows for multiple time trends and the two-step CODA model 2S-CoDA which extends the Oeppen (2008) model by introducing time-, age- and cause-specific weights and by adding cause-specific information to the forecasts. 2S-CoDA is found to produce more accurate cancer forecasts for the populations that are analysed than the Oeppen (2008) model. VECM-CoDA and 2S-CoDA make use of compositional data analysis similarly to the model that was presented in Bergeron-Boucher *et al.* (2017) but differ fundamentally in the way that mortality is modelled. VECM-CoDA and 2S-CoDA model several causes of death from a multiple-decrement life table in a single population. In contrast Bergeron-Boucher *et al.* (2017) considered a single-decrement life table but several populations. Further analysis can be found in the on-line supplementary material, which examines the influence of the number of causes of deaths included on the accuracy of the cancer forecasts.

It is well known that the dependence between causes of death is not identifiable for individuals, (see A (1975)), as it is not possible to distinguish between independent competing risks and a large number of dependent competing risks producing the same cause-specific hazards. CODA mortality models, which make use of aggregated data, can respect covariances between ages and causes but only identify net transitions of shifted deaths between ages and/or causes. They do suggest that the latent mortality patterns in the absence of competing risks are identifiable.

In this study we focus on forecasting cancer mortality and evaluate the proposed models' ability to forecast cancer deaths. Cancer is selected as it has been a major cause of death over the last 50 years, is relatively well diagnosed as the principal cause of death and has experienced an increasing share of deaths in many industrialized countries. In 2014, 26.4% of all deaths in European Union countries (Eurostat, 2017) were caused by cancer and €83.2 billion were spent on healthcare related to cancer (Jönssona *et al.*, 2016). Despite the importance of cancer mortality, limited research has been carried out on improving current forecasting methods for cancer deaths and, hence, this paper provides valuable information for social planners and society.

The data that are analysed in the paper and the programs that were used to analyse them can be obtained from

```
https://rss.onlinelibrary.wiley.com/hub/journal/14679876/series-
c-datasets
```

2. Notation and methods

Throughout the paper mortality is measured by using life table information to account for the age composition of mortality. Standard life table calculations are used following Preston *et al.* (2001). Cause-specific mortality is often available in the form of actual death counts $D_{t,x,i}$ by year t, age x and cause of death i, where $t \in (1, 2, \ldots, T)$, $x \in (1, 2, \ldots, N)$ 5-year age groups and $i \in (1, 2, \ldots, K)$ causes. Knowing the death density for the all-cause mortality in the population, i.e. the life table deaths $d_{t,x}$, enables the calculation of cause-specific life table deaths $d_{t,x,i}$ following the equation

$$d_{t,x,i} = d_{t,x} \frac{D_{t,x,i}}{D_{t,x}}, \tag{1}$$

where $D_{t,x}$ denotes the total all-cause deaths count (Preston *et al.*, 2001). Further, by knowing the number of person-years lived at age x, $L_{t,x}$, it is possible to calculate the associated cause-specific death rates $m_{t,x,i}$ by using

$$m_{t,x,i} = \frac{d_{t,x,i}}{L_{t,x}}. \tag{2}$$

These two relationships are used to transform the observed data and to relate variables within the life table.

2.1. Data

To evaluate the cause-specific mortality models, data from France and the Netherlands are used to fit the models. Data for France were downloaded from Institut National d'Etudes Démographiques (2016) divided into the following causes: infectious diseases, cancer, cardiovascular diseases, diseases of the respiratory system, diseases of the digestive system, other diseases, and injury and poisoning. The death classification is taken as given from the source of data. Data were divided into 5-year age groups, censored at age above 100 years, and into single years of time from 1925 to 2008. We restricted the analysis to the years 1955 and onwards to avoid fluctuating patterns from the Second World War. As few deaths occur at younger ages we restricted the data to age groups older than 25 years, which simplified the estimation of the CODA models. These cannot be estimated for age groups with zero deaths; hence 0s are replaced by half a death—a similar imputation method was used by Bergeron-Boucher *et al.* (2015). The imputation has no appreciable effect on the results in this study.

Data for the Netherlands were downloaded from Statistics Netherlands (2018) and divided into the following causes: infectious diseases, cancer, diseases of the endocrine system, mental diseases, diseases of the nervous system, circulatory diseases, diseases of the respiratory system, diseases of the digestive system, other diseases and external causes. Some causes were aggregated from the original source to match the French data. A close match between the two data sets is not possible because data are not in general available at a sufficiently detailed international classification of diseases level. Hence, the results for the two countries are not directly comparable but are used to illustrate that the models can be applied to different subdivisions of the causes of death. Similarly to the French data, analyses were restricted to the age groups 25–95 years and older, where data were censored, for the years 1958–2014. 1958 is selected to avoid very frequent shifts in the international classification of diseases classification system of causes of death with changes in 1941, 1950 and 1955 (Koren *et al.*, 2012).

Sex-specific all-cause life tables for France and the Netherlands for the relevant years were downloaded from the human mortality database (Human Mortality Database, 2018).

2.2. Compositional data analysis

CODA refers to a broad set of statistical methods used to analyse compositional data. This section briefly describes CODA that is essential to the models suggested and further details can be found in Aitchison (1982, 1986), Pawlowsky-Glahn and Buccianti (2011) and Bergeron-Boucher *et al.* (2017). As life table deaths are constructed by multiplying age-specific probabilities of dying into an arbitrary initial birth cohort which is also known as the radix, they contain only relative information and sum to the initial cohort in each year (Bergeron-Boucher *et al.*, 2017). Changes in the number of life table deaths for a specific age group must therefore be offset by changes in the other age groups, which is the fundamental feature in compositional data. Traditional decomposition techniques provide inconsistent results when applied to compositional data as they do not recognize the implicit constraints of summing to a constant (Aitchison, 1982, 1986): mathematically, compositional data lie in the bounded space of the simplex and traditional decomposition techniques are defined for data in the real space. Aitchison (1986) showed that by making log-ratio transformations it is possible to express compositional data in the real space where the data can be analysed with conventional models and then transformed back into the simplex.

The analysis that is presented in the present paper makes use of the centred log-ratio (CLR) transformation to express the data in the real space. The CLR transformation takes the logarithm of each observation divided by the geometric mean, i.e.

$$\operatorname{clr}(d_{t,x}) = \log\left\{\frac{d_{t,x}}{g(t,x)}\right\}, \tag{3}$$

where $g(t,x)$ is the geometric mean by age in the year, i.e. $g(t,x) = (d_{t,x_1} \cdot d_{t,x_2} \cdots d_{t,x_n})^{1/n}$. The CLR transformation maintains the initial constraint in the data as its elements sum to 0 by construction but resulting values are real. The inverse CLR, clr^{-1}, is defined as

$$\operatorname{clr}^{-1}(d_{t,x}) = C\{\exp(d_{t,x})\}, \tag{4}$$

where C is the closure operator that divides each entry by the sum of all entries and multiplies by the initial number of life table deaths so the result meets the restriction of all life table deaths summing to a certain constant. For a vector $Y = (y_1, \ldots, y_n)$ the closure of Y is $C(Y) = (y_1/\Sigma y_i, y_2/\Sigma y_i, \ldots, y_1/\Sigma y_n)K$, where K is a constant (Bergeron-Boucher *et al.*, 2017).

Aitchison (1986) also defined addition and subtraction operations obeying conventional rules of arithmetic, called perturbations, maintaining the result of the operation in the simplex (Pawlowsky-Glahn and Buccianti, 2011). The following operations will be used in the analysis and are essential when modelling death distributions. Assume that $X = (x_1, \ldots, x_n)$ and $Y = (y_1, \ldots, y_n)$; then

$$X \ominus Y = C\left(\frac{x_1}{y_1}, \ldots, \frac{x_n}{y_n}\right)$$

and

$$X \oplus Y = C(x_1 y_1, \ldots, x_n y_n).$$

$X \ominus Y$ is called a perturbation and measures the distance between X and Y in compositional data similarly to subtraction in some data on the real axis. $X \oplus Y$ is the opposite operation and can be compared with addition on the real axis.

2.3. *Forecasting models*

The model that was suggested by Oeppen (2008), denoted the CODA model with common time trend CT-CoDA, is compared with two new CODA models which accommodate limitations in CT-CoDA, i.e. the two-step CODA model 2S-CoDA and the CODA with cointegrating vector error correcting model VECM-CoDA. Finally, all three CODA models are compared with the LC model as a standard benchmark from the literature.

2.3.1. Compositional data analysis model with common time trend

CT-CoDA was suggested by Oeppen (2008) for forecasting cause-specific mortality in Japan. It uses life table deaths from a multiple-decrement life table to forecast mortality and recognizes the compositional nature of life table deaths over cause and age where

$$\sum_i \sum_x d_{x,i} = 1, \tag{5}$$

when the radix of the life table is equal to 1. The model makes use of equation (5) by stacking cause-specific death matrices horizontally, forming a $T \times NK$ matrix, and calculates the CLR of each row to map the life table deaths onto the real space. Next, data are centred by subtracting the geometric mean for each year and a singular value decomposition (SVD) is used to estimate the model parameters. The model can be written as

$$\mathrm{clr}(d_{t,x,i} \ominus \alpha_{x,i}) = \beta^1_{x,i} k^1_t + \beta^2_{x,i} k^2_t + \ldots + \beta^p_{x,i} k^p_t + \epsilon_{t,x,i}, \tag{6}$$

where $\alpha_{x,i}$ is the geometric mean, k^p_t measure changes over time, $\beta^p_{x,i}$ age-specific changes, p is the number of extracted components from the SVD and $\epsilon_{t,x,i}$ is the time-, age- and group-specific independent and identically distributed (IID) error term. Note that the SVD is constructed so that the first component explains most of the total variation in the data, the second component the second most, etc.

The k^p_t parameter vectors are the only time-dependent parameters and are constructed to be orthogonal to each other. Thus, forecasts can be calculated by forecasting k^p_t by using autoregressive integrated moving average models (Box and Jenkins, 1970). Forecasts of k^p_t are used in the model and the result is transformed back by using the inverse CLR procedure and the geometric mean $\alpha_{x,i}$ is added, i.e.

$$\hat{d}_{t,x,i} = \mathrm{clr}^{-1}(\beta^1_{x,i} \hat{k}^1_t + \beta^2_{x,i} \hat{k}^2_t + \ldots + \beta^p_{x,i} \hat{k}^p_t) \oplus \alpha_{x,i}, \tag{7}$$

where the circumflex indicates forecast values.

The basic dynamics in the model can most easily be understood by considering the first component alone so that the total variation is decomposed by using only the first rank component. Higher order components can be interpreted in a similar way. With one component CT-CoDA reduces to

$$\mathrm{clr}(d_{t,x,i} \ominus \alpha_{x,i}) = \beta_{x,i} k_t + \epsilon_{t,x,i}. \tag{8}$$

As the summation constraint, equation (5), is maintained by the CLR transformation, deaths are redistributed in the model by $\beta_{x,i}$ when mortality is changing over time. This means that, if some deaths do not occur at a specific age and cause, they will shift to a different age and/or cause group. These deaths are not redistributed randomly but towards the ages and causes where they are most likely to occur according to the parameter estimates in the model. Here, $\beta_{x,i}$ estimates that are below 0 indicate that deaths are reallocated from these ages and causes to ages and causes with a positive $\beta_{x,i}$ when k_t is increasing (Oeppen, 2008; Bergeron-Boucher *et al.*, 2017). Thus, the net effects of competing risks between causes of death are modelled explicitly in CT-CoDA. The redistributing effects are one of the main advantages of using the death distribution instead of deaths rates for modelling cause-specific mortality. Redistribution is the result of CODA's ability to handle a covariance structure where changes over time in one of the elements in the distribution of deaths must be offset by changes in other elements (Bergeron-Boucher *et al.*, 2017).

2.4. Two-step compositional data analysis

2S-CoDA introduces two new elements compared with CT-CoDA to improve the fit and forecast of the model:

(a) age, cause and time weights are imposed and
(b) cause-specific information is added to a common trend forecast determined by CT-CoDA.

These two elements also diversify 2S-CoDA from the model that was presented in Bergeron-Boucher *et al.* (2017) together with the modelling of multidecrement life tables instead of single-decrement life tables as in Bergeron-Boucher *et al.* (2017). CT-CoDA centres the observed life table deaths by perturbating its geometric mean. The importance of each cause of death, i.e. the number of deaths, is therefore not relevant when estimating $\beta_{x,i}$ and k_t. A less common cause like infectious diseases is thus weighted equally with a large cause like cancer when determining the common trend k_t^1. The size of the cause of death is potentially important when forecasting because of competing risks and thus 2S-CoDA introduces a weighting scheme where age- and cause-specific weights are imposed according to the average number of deaths for each age and cause. Further, equal weighting of the causes means that results can depend on how causes are aggregated. For example, if cancer is split into whether or not the deaths are related to smoking, the SVD would effectively give twice the weight to overall cancer compared with the situation where cancer is treated as a single cause. Hence, 2S-CoDA neutralizes some of the undesired consequences of aggregation.

Further, $\beta_{x,i}$ is determined in CT-CoDA by an equal weighting of each year and is assumed to be constant over time. As there has been a considerable change in the relative sizes of the various causes in recent years this assumption is likely to fail. Further, improvements in mortality have shifted from the young ages towards improvements at increasingly higher ages (Rau *et al.*, 2008; Bergeron-Boucher *et al.*, 2015). Estimation of $\beta_{x,i}$ where recent years are given more weight may thus produce more accurate forecasts. Hyndman *et al.* (2013) suggested a declining weighting scheme where the highest weight is placed on the most recent year and thereafter reduced exponentially. A similar approach is introduced in 2S-CoDA, meaning that both the year and the age dimensions are weighted. A comparison between a CT-CoDA forecast and a 2S-CoDA forecast will determine how useful the weighting scheme described is for forecasting.

2S-CoDA follows the same first steps as CT-CoDA but using only a first-rank approximation as higher rank approximations are modelled by cause-specific terms, i.e.

$$\mathrm{clr}(d_{t,x,i} \ominus \alpha_{x,i}) = \beta_{x,i} k_t + \epsilon_{t,x,i}. \tag{9}$$

Generalized SVD is used to estimate the model parameters, imposing weights on the age, cause and time dimensions. Generalized SVD is a generalization of the SVD which imposes constraints on a rectangular matrix so that a weighting of rows and columns is possible (Loan, 1976; Abdi, 2011). 2S-CoDA uses age- and cause-specific weights according to the average relative size of each age and cause combination, i.e.

$$w_{x,i}^{\mathrm{age}} = \frac{\bar{d}_{x,i}}{\sum_{x=1}^{\omega} \sum_{i=1}^{K} \bar{d}_{x,i}}, \tag{10}$$

where $\bar{d}_{x,i}$ is the temporal mean of the life table deaths.

We adopt the approach that was described in Hyndman *et al.* (2013) and impose the following weights on the time dimension when estimating $\beta_{x,i}$ and k_t:

$$w_t^{\mathrm{time}} = \rho(1-\rho)^{T-t}, \tag{11}$$

where $t \in (1, 2, \ldots, T-1)$ and ρ determines the percentage weight on the recent year, so $\rho = 0.05$ implies that the last year is weighted with 5%.

The common trend assumption in CT-CoDA implies that only the variation that is shared between the causes is used in the fit and forecast of the model. In 2S-CoDA this component is approximated by $\beta_{x,i} k_t$ and is denoted the joint component. Cause-specific variation is thereby contained in the residuals and is potentially important when forecasting. To improve the fit and forecast we suggest decomposing the cause-specific residuals that are obtained from estimating equation (9) by using SVD and the individual information when forecasting. 2S-CoDA follows the underlying idea in Lock *et al.* (2013) by estimating a joint and individual component but differs in the estimation procedure. Lock *et al.* (2013) used an iterative process for the estimation of both joint and individual components whereas 2S-CoDA uses a simpler approach without an iterative process. Further, 2S-CoDA uses weights based on the number of age-specific deaths whereas Lock *et al.* (2013) suggested applying weights based on standardized variation of the data.

The model can be written as

$$\mathrm{clr}(d_{t,x,i} \ominus \alpha_{x,i}) = \beta^{\mathrm{J}}_{x,i} k^{\mathrm{J}}_t + \beta^{\mathrm{I}}_{x,i} k^{\mathrm{I}}_{t,i} + \epsilon_{t,x,i}, \tag{12}$$

where J and I denote the joint and individual cause components of the model respectively and $\epsilon_{t,x,i}$ is an IID error term.

2.5. Compositional data analysis model with cointegrating vector error correction model

VECM-CoDA introduces one new element in comparison with CT-CoDA by estimating more than one time trend for each rank approximation. CT-CoDA implicitly assumes that different causes share one time trend forming stationary relationships. Equivalently, this can be expressed as cointegrating relationships for all the causes. This assumption might be acceptable for some causes but does not necessarily hold for all causes. Furthermore, the assumption may be appropriate for some populations but not for others. The forecast implication of the common trend assumption will be analysed by comparing forecasts from CT-CoDA with forecasts from VECM-CoDA where multiple stochastic time trends are allowed, in which case the modelling of mortality trends follows the underlying idea in Arnold-Gaille and Sherris (2013).

VECM-CoDA transforms horizontally stacked life table death matrices by using the CLR operator and centres the data by perturbation with the geometrical mean, similarly to CT-CoDA. But instead of applying SVD on the stacked $T \times NK$ matrix, VECM-CoDA decomposes each centred cause-specific matrix by using SVD. Thus, K $k_{t,i}$ time trends and $\beta_{x,i}$ age-specific responses are estimated, i.e.

$$\mathrm{clr}(d_{t,x,i} \ominus \alpha_{x,i}) = \beta^1_{x,i} k^1_{t,i} + \ldots + \beta^p_{x,i} k^p_{t,i} + \epsilon_{t,x,i}, \tag{13}$$

where $k^p_{i,t}$ is cause specific and $\epsilon_{t,x,i}$ an IID error term. The different $k_{t,i}$ time trends might be dependent, i.e. some causes might be sharing trends. The Johansen trace test, which is described in the on-line supplementary material section A, provides a statistical test of stationary and non-stationary relationships between time series, i.e. whether some $k^1_{i,t}$s follow the same time trends. The number of stationary relationships among $k^1_{i,t}$ is found (denoted r) with the Johansen test and thereby also the number of trends (denoted $n = K - r$) that drive the $k^1_{i,t}$-system. Having determined stationary relationships in $k^1_{i,t}$, a cointegrated vector error correction model is used when forecasting $k^1_{i,t}$. Dependence between the causes is thereby taken into account when forecasting cause of death. The cointegrated vector error correction model can be written as

$$\boldsymbol{\Delta}\mathbf{k}_t = \boldsymbol{\Pi}\mathbf{k}_{t-1} + \sum_{j=1} \boldsymbol{\Gamma}_j \boldsymbol{\Delta}\mathbf{x}_{t-j} + \mathbf{B} + \epsilon_t, \tag{14}$$

where $\mathbf{k}_t = (k^1_{1,t}, \ldots, k^1_{K,t})'$, $\mathbf{\Pi}$ is a matrix with rank r determining the stationary relationships among $k^1_{i,t}$, $\mathbf{\Gamma}_j$ a matrix measuring the auto-regressive part of the system and $\mathbf{B}$ a $k \times 1$ vector specifying deterministics in the model; further details are in the on-line appendix A. We specify a deterministic trend in the non-stationary part of the model, meaning that the variables are trending: this is similar to a drift term in the auto-regressive integrated moving avearage specification that was used for the other models. Equation (14) is estimated and used to calculate forecasts of $k_{i,t}$ which are multiplied by $\beta_{x,i}$ and used to calculate forecasts of causes of death.

2.6. Lee–Carter model

Forecast results from the CODA models will be compared with the LC model, which is one of the most used mortality forecasting models. The LC model was suggested by Lee and Carter (1992) for forecasting all-cause mortality but has also been used to forecast cause-specific mortality (Peltonen and Asplund, 1996). The LC model centres age-specific deaths rates by subtracting the age-specific arithmetic time mean and decomposes the result by using SVD, i.e.

$$m_{t,x,i} = \alpha_{x,i} + k_{t,i}\beta_{x,i} + \epsilon_{t,x,i}, \tag{15}$$

where $\alpha_{x,i}$ is the arithmetic mean, $k_{t,i}$ is a cause-specific index describing the general pattern of mortality, $\beta_{x,i}$ is the age- and cause-specific response to changes in $k_{t,i}$ and $\epsilon_{t,x,i}$ is an IID error term. The LC model forecasts mortality rates and life table deaths are calculated from these when comparing with the CODA models. Standard life table calculations are used following Preston *et al.* (2001). We do not include a jump-off correction in the LC model as argued by Lee and Miller (2001) as it was shown to introduce a forecast bias (Booth *et al.*, 2006). A jump-off correction might be useful for short-term mortality forecasts but as we focus on long forecasts of 20 years we want to avoid introducing a forecast bias.

2.7. Comparing model forecasting performance

The models' abilities to forecast future cancer deaths are evaluated by fitting them to a subset of the data and forecasting the remaining years. Observed and forecast cancer life table deaths are compared by averaging the root-mean-square error (RMSE) over age, i.e.

$$\mathrm{aRMSE}^{d_{x,t}} = \frac{1}{N} \sum_{x=1}^{X} \sqrt{\left\{ \frac{\sum_{h=1}^{M} (\hat{d}_{h,x,\mathrm{cancer}} - d_{h,x,\mathrm{cancer}})^2}{M} \right\}} \tag{16}$$

where a denotes an average over age, h the forecast year and M the total number of forecast years. (This statistic is used to give an idea about forecast comparisons of different models. Asymptotically, the test statistic follows a normal distribution for large $d_{h,x,\mathrm{cancer}}$, following the arguments in Czado *et al.* (2009).)

aRMSE is calculated for different forecast periods, rolling the origin of the forecasts from 10 to 20 years. For example, when using the French data, a 20-years forecast is calculated by fitting the models to the years 1955–1988. Next, a 19-years forecast is calculated by rolling the origin of the forecast 1 year so data from 1955 to 1989 are included in the fitting period. This procedure is continued until a 10-years forecast is calculated.

The average forecast error over the different forecasts is calculated and used for comparing the models. Thus, the effect of particular years used for forecasting is reduced as different forecast periods are considered. By using the RMSE, ages with a large number of life table deaths are implicitly weighted more than ages with few life table deaths. We do this as the main objective is to predict the total number of cancer deaths most accurately.

As the LC model is formulated in death rates we also evaluate the models with the RMSE measured on death rates, i.e.

$$\mathrm{aRMSE}^{m_{x,t}} = \frac{1}{N}\sum_{x=1}^{X}\sqrt{\left\{\frac{\sum_{h=1}^{M}(\hat{m}_{h,x,\mathrm{cancer}} - m_{h,x,\mathrm{cancer}})^2}{M}\right\}}. \tag{17}$$

2.8. Other specifications in the models

The time weighting parameter, in 2S-CoDA, is set to $\rho=0.1$ throughout the paper as determined by cross-validation of a 15-years out-of-sample window across the four populations. A rank 3 approximation is used for CT-CoDA and a rank 2 approximation for VECM-CoDA as adding higher ranks changes the fit and forecast only negligibly. The specific auto-regressive integrated moving average models are selected by using the augmented Dickey–Fuller test (Dickey and Fuller, 1979) to determine the order of integration and the numbers of auto-regressive and moving average terms are selected on the basis of the Akaike information criterion (Akaike, 1974). For all rank 1 k_t-terms, a deterministic drift term is included to account for upward and downward slopes.

3. Results

Fits of the four models are illustrated for French females and Dutch males and results of the forecast performance of the models are shown for all populations. Remaining results for French males and Dutch females are shown in the on-line supplementary material Figs D3 and D11.

3.1. Distributions of cancer deaths

The cancer deaths distribution has both shifted and been compressed over the data period meaning that cancer deaths are occurring at increasingly higher ages (Fig. 1) for French females (results for Dutch males are in the on-line supplementary material Fig. D4). The geometric mean is the geometric average over time and close to the distribution in 1980. The log-transformed and centred life table deaths are positive in all years and ages, which means that deaths are relatively transferred towards cancer throughout the data period (Fig. 1). The centred distribution in 2005 is increasing at older ages, meaning that deaths for these ages especially are increasing. The CODA models fit the changes in the centred deaths distributions for all causes when modelling cause-specific mortality.

3.2. Parameter estimates

To illustrate the four models we show the parameters estimated by using data for French females. For k_t^1 a 20-years forecast is shown together with the model estimates. Parameter estimates for the first rank are shown in this section and higher orders are presented in the on-line supplementary material Figs D5 and D6. The negative estimates of the LC model are plotted to be comparable with the CODA models.

The time pattern of mortality that is captured by k_t^1 in Fig. 2 shows an increasing pattern for the three CODA models; measuring that across all causes of deaths there has been a decline in mortality. Comparing k_t^1 from CT-CoDA and 2S-CoDA shows that weighting the causes

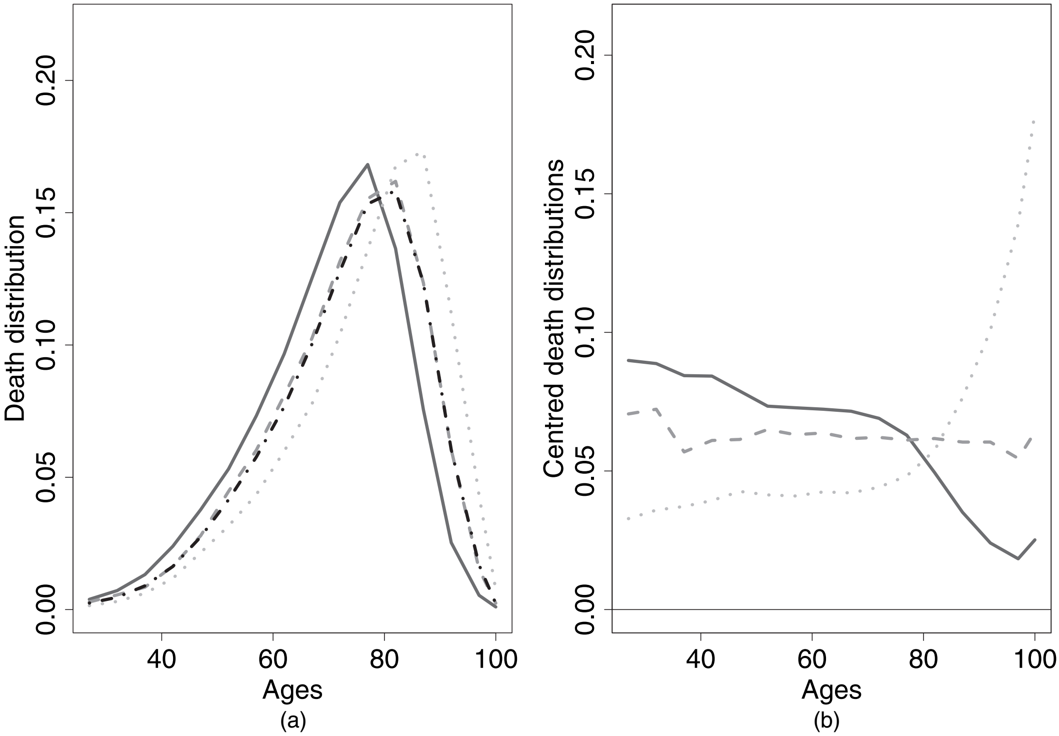
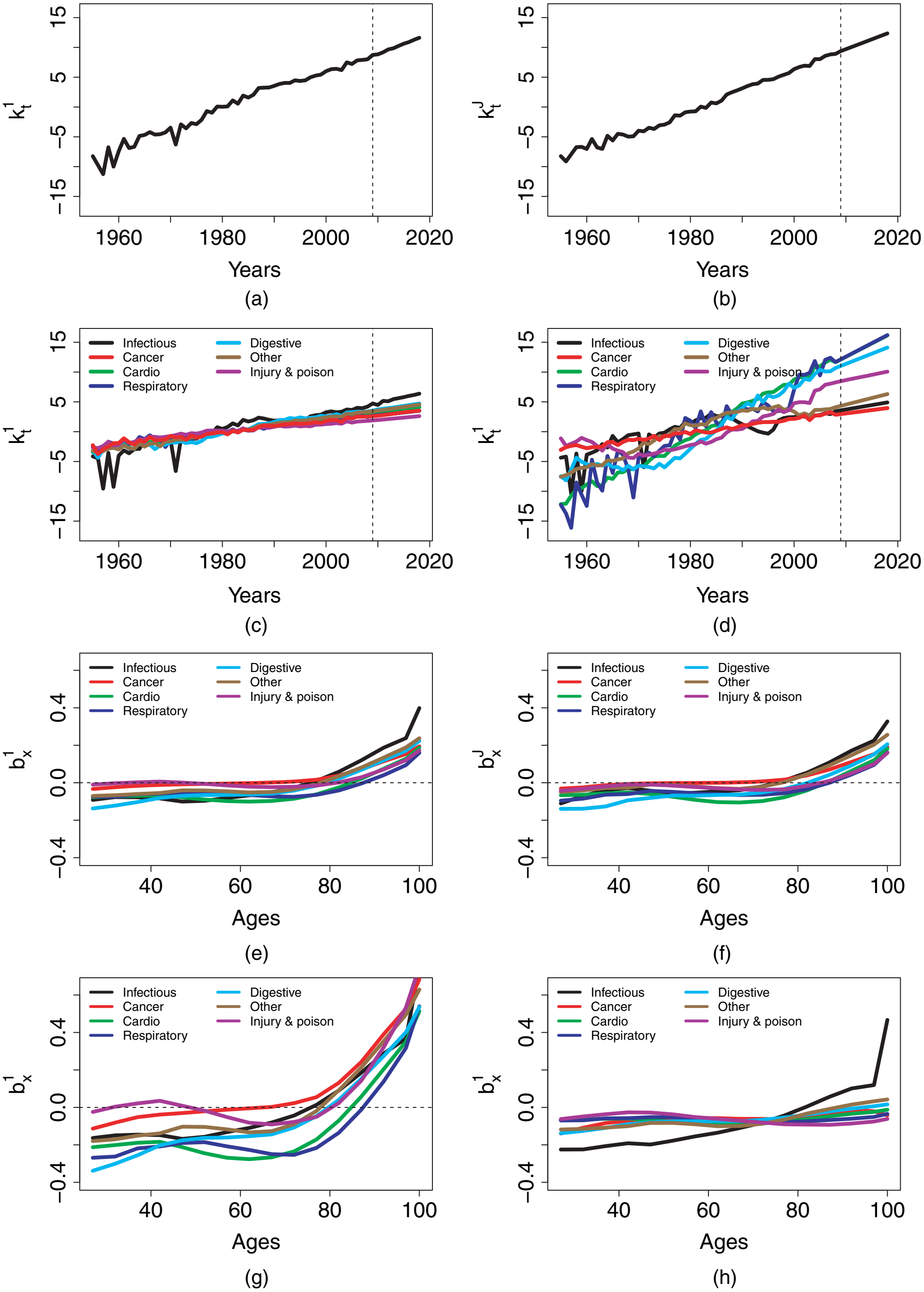


Fig. 1. (a) Distribution and (b) centred distribution of cancer life table deaths for French females in selected years: ———, 1955; – – – –, 1980; · · · · , 2005; · — · —, geometric mean

by size reduces the fluctuations in k_t^1, making it more linear and potentially easier to forecast. From VECM-CoDA it is clear that infectious disease contributes significantly to this fluctuating pattern in k_t^1 for CT-CoDA, as it is the only cause with three major fluctuations from 1955 to approximately 1972. The $k_{t,i}^1$-estimates in VECM-CoDA also indicate that causes do not follow the same time trend through the fitting period. In particular infectious diseases and injury and poisoning follow separate patterns: this is tested with the Johansen trace test (Johansen, 1991) and results are shown in the on-line supplementary material. The LC model's parameters are not directly comparable with the CODA models' as it uses death rates as input. An increase in the negative of k_t in the LC model implies that mortality is declining and thus the LC model also estimates a decline in mortality for all the causes. The LC model and VECM-CoDA show diverging time trends but with different patterns.

The estimated $\beta_{x,i}^1$-parameters show similar patterns across the three CODA models. Negative $\beta_{x,i}^1$ indicate redistribution of deaths away from these ages and causes when k_t is positive. Thus, Fig. 2 shows that deaths are transferred from younger ages towards older ages, and away from cardiovascular and respiratory diseases when time progresses, as the lowest estimates are found for these causes. The LC parameter estimates are again not directly comparable with the CODA models' as the LC model uses death rates and causes which are not related in the model. A positive $\beta_{x,i}$ implies that mortality declines for this age group, but an automatic redistribution between causes cannot be claimed because the deaths rates are not closed by a closure operator as in the CODA models. The negative $\beta_{x,i}$-estimates for the LC model for infectious diseases are increasing by age with large positive values for high ages, meaning that the LC model



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 2. k_t^1 (measuring the general development in mortality) and β_x^1 (measuring the redistribution of deaths) estimates for French females from CT-CoDA, 2S-CoDA and VECM-CoDA and the LC model for the years 1955–2009 (the negative estimates are plotted for the LC model to be comparable with the CODA models; for 2S-CoDA the joint components are plotted, i.e. k_t^J and β_t^J): (a), (e) CT-CoDA; (b), (f) 2S-CoDA; (c), (g) VECM-CoDA; (d), (h) LC model finds an increasing number of deaths for these ages groups when considering infectious diseases only.

The time, age and cause weights that were used for 2S-CoDA are shown in Fig. D1 in the on-line supplementary material. As most French females die from cardiovascular disease and cancer at an age of around 80 years, these causes and ages are given the highest weights.

3.3. Model fit in and out of sample

3.3.1. *In-sample fit and forecast of withheld years*

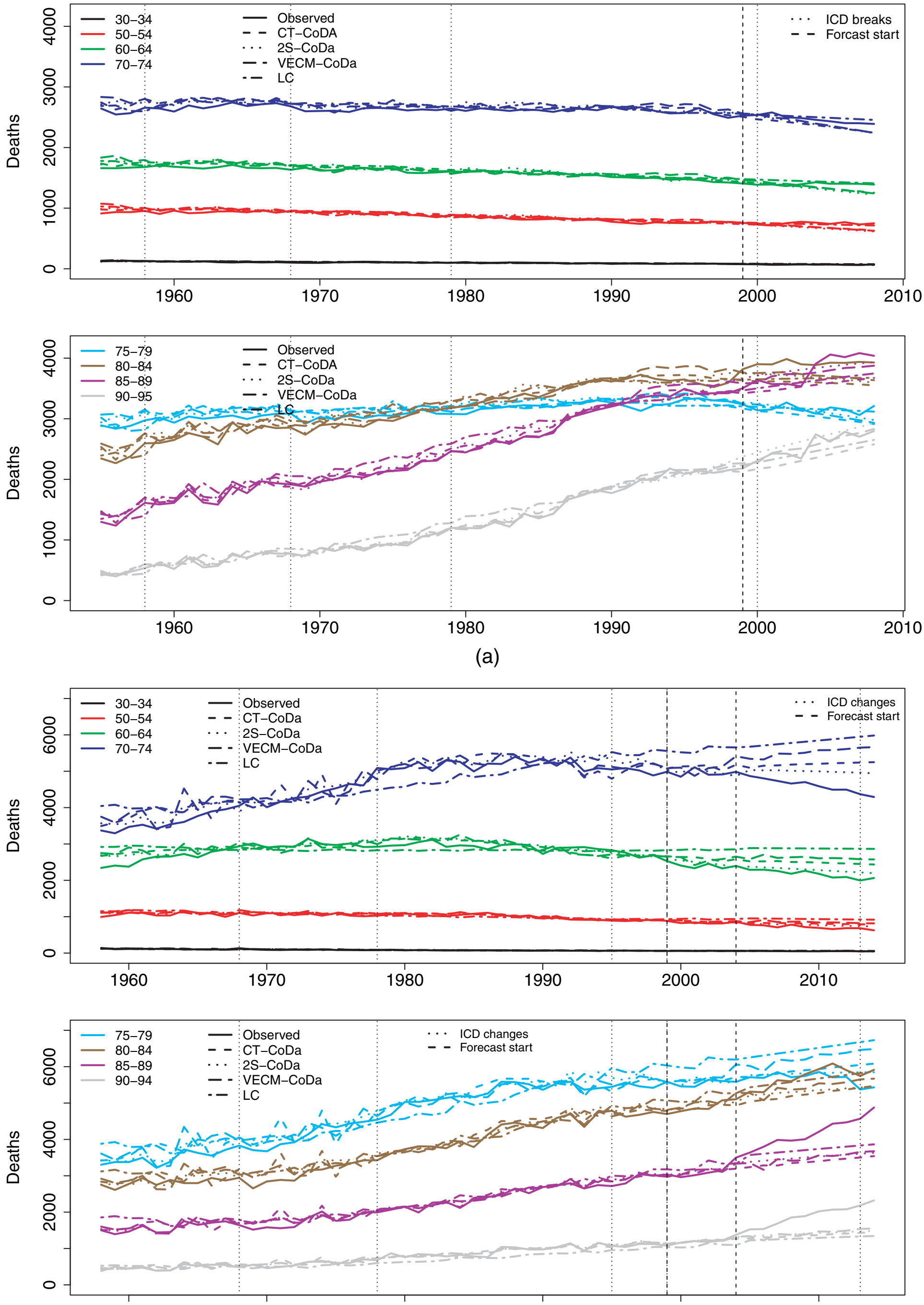
The main objective of this paper is to forecast the number of cancer deaths; hence observed, fitted and 10 out-of-sample forecasts for cancer are shown for different age groups illustrated for French females and Dutch males. Similar plots are shown for French males and Dutch females in the on-line supplementary material Fig. D3.

Fig. 3 shows that all models produce a reasonable fit and forecast for most age groups. For French females, the older age groups 85–89 and 95–99 years old are more difficult to fit and forecast compared with the younger age groups. As 2S-CoDA is weighted on the time dimension it produces a worse fit compared with CT-CoDA at the beginning of the fitting period but a better fit at the end. VECM-CoDA has the worst forecast of the four models whereas the LC model produces a good fit and forecast at the younger ages but largely deviates at the older age groups 80–84 and 85–89 years old. This could be a consequence of the restricted age dynamics that are allowed in the LC model because of the constant $\beta_{x,i}$ assumption. For completeness, Fig. D10 in the on-line supplementary material shows the average RMSE over age for each year across the whole fitting period for French females and Dutch males. It is important to note that all the models applied are trend models, which cannot predict breaks and deviations from a trend. For example, with French females at age 85–89 years we see that the observed number of life table deaths is increasing very rapidly around 2003 and none of the models can predict this break. This is a general limitation with extrapolative trend models but a successful mortality forecasting model which can predict breaks does not exist in the literature.

The fit of 2S-CoDA is illustrated in Fig. 4 where the variation that is explained by the joint and individual components of the models is shown. The explained variation is calculated by using the Frobenius norm (Golub and van Loan, 1996).

The joint component is the first-rank approximation from the generalized SVD, which for French females captures a high proportion of the variation for the diseases cancer, cardiovascular, respiratory, digestive and other diseases, and injury and poisoning. These diseases are therefore also well described by CT-CoDA. In contrast much less variation is captured by the joint component for infectious diseases and 2S-CoDA improves the fit by adding an individual component. For Dutch males the joint component only exceeds 60% of the variation for cancer, circulatory diseases and other diseases, whereas a low proportion of the variation is explained for the rest of the causes. Thus 2S-CoDA has the potential to perform much better than CT-CoDA as the individual component explains a large part of the variation for Dutch males. Note that CT-CoDA captures more variation than the joint component as higher order rank approximations are added but, in contrast with 2S-CoDA, these higher order terms are still joint for all causes.

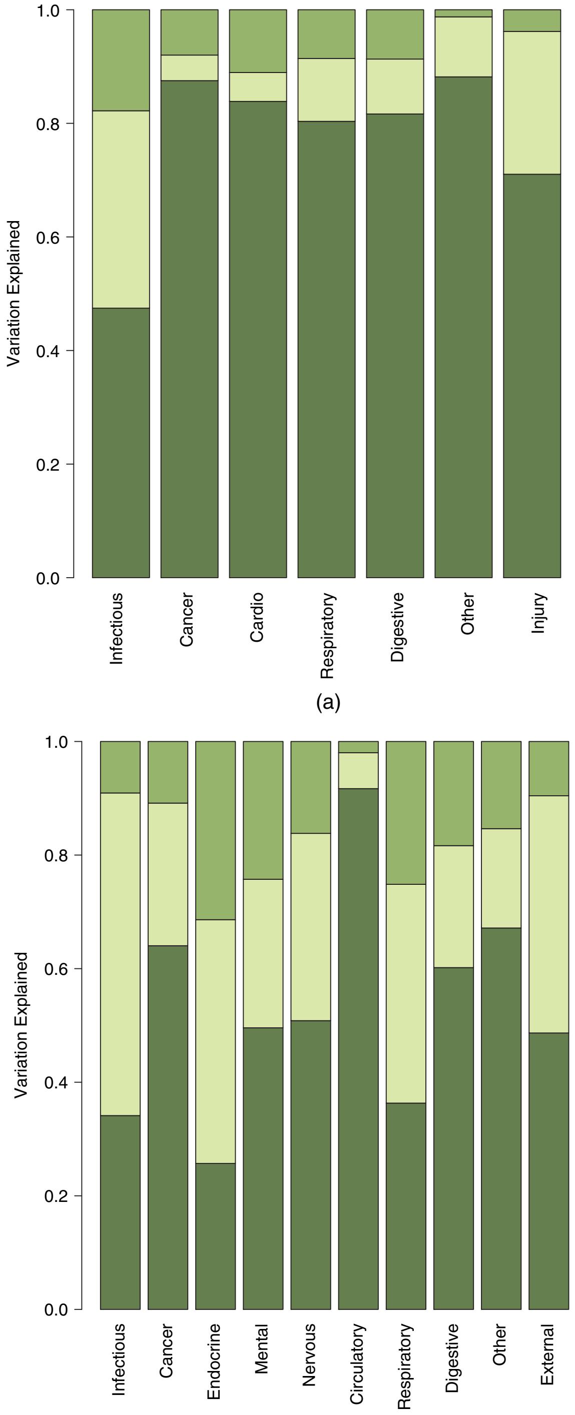
Residuals for the four models are plotted in the on-line supplementary material Figs D8 and D9 for French females and Dutch males. No evidence of cohort effects is found for residuals for French females whereas some cohort effects are seen for Dutch males. Including cohort effects could therefore potentially improve the forecasts for Dutch males, but it is beyond the scope of this paper to explore the inclusion of cohort effects.



(a)

(b)

Fig. 3. 10-year out-of-sample forecasts of cancer life table deaths across age groups for (a) French females and (b) Dutch males using CT-CoDA, 2S-CoDA, VECM-CoDA and the LC model: ⋮, changes in the classification of causes of deaths for each country



(a)

(b)

Fig. 4. Explained variation by the joint and individual component in 2S-CoDA for (a) French females and (b) Dutch males: ■, joint; □, individual; ▨, residual

3.3.2. *Out-of-sample comparison*

Cancer forecasts based on the three CODA models and the LC model are compared by using the scheme that was described in Section 2.7. Table 1 shows the average RMSE comparing observed and forecast life table deaths and for death rates by using data from France and the Netherlands.

Considering life table deaths, 2S-CoDA produces the lowest RMSE for both French populations and for Dutch males and the VECM model produces the lowest RMSE for Dutch females. The results are similar when measured in death rates for the male populations but the LC model produces the lowest RMSE for French females and 2S-CoDA for Dutch females when death rates are used to measure mortality.

As 2S-CoDA is performing better than CT-CoDA for three of the four populations it indicates that 2S-CoDA can improve the forecast over CT-CoDA when forecasting cancer. Accommodating the limitations about equal weighting and modelling of cause-specific behaviour in CT-CoDA thus leads to a better fit and a better forecast performance for most of the populations that we consider. VECM-CoDA in contrast did not improve the forecast of CT-CoDA, indicating that allowing different $k^1_{t,i}$-trends did not improve the forecast of cancer, or that VECM-CoDA could not forecast these well for the populations selected. A significance test of the out-of-sample errors was carried out and results are shown in the on-line supplementary material.

3.3.3. *20-years forecast*

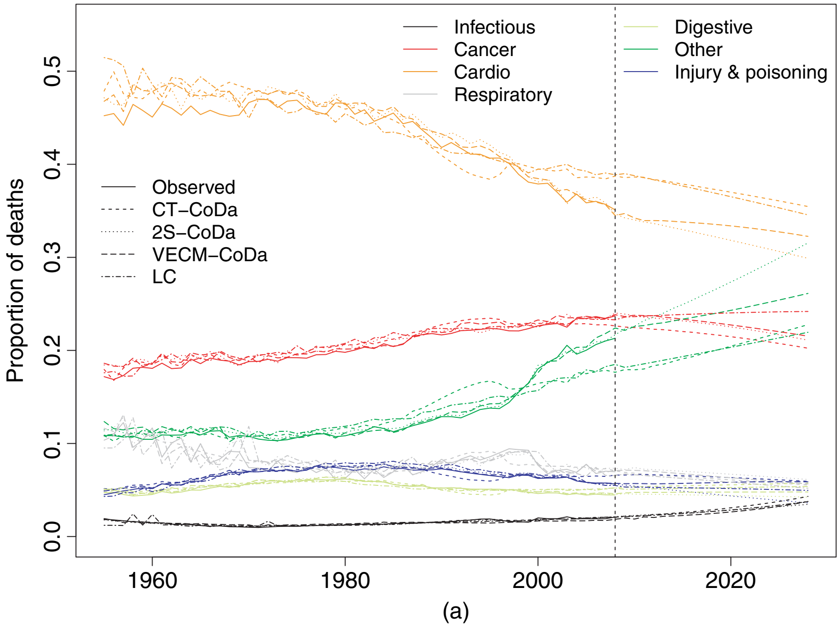
Even though the main objective of this paper is to forecast the number of cancer deaths, it is important to check how the models forecast the other causes because of dependence between the causes. Hence, the proportions of life table deaths are calculated across age for the causes and 20-year forecasts calculated for each model. The compositional analysis in the three CODA models ensures that the proportions sum to 1 in each year, whereas this is not ensured in the LC model which fits each cause independently.

Fig. 5 shows the fitted and a 20-years forecast of proportions of life table deaths across age for French females and Dutch males, and similar plots are shown for French males and Dutch females in the on-line supplementary material Fig. D11. The improved fit for 2S-CoDA, compared with CT-CoDA, not only applies to cancer but also to the other causes: for example for French females 2S-CoDA provides a better fit for other diseases. All the models provide a similar fit and forecast for the causes respiratory diseases, digestive diseases, infectious diseases

Table 1. 20-year out-of-sample rolling window RMSEs for observed *versus* forecast cancer life table deaths, for French and Dutch populations†

Model	*Results for French females*	*Results for French males*	*Results for Dutch females*	*Results for Dutch males*
RMSE measured in life table deaths (equation (16))				
CT-CoDA	105.4	292.3	140.03	316.9
2S-CoDA	*90.9*	*217.2*	168.55	*259.1*
VECM-CoDA	114.6	369.4	*108.13*	391.6
LC	99.6	263.9	153.56	484.3
RMSE measured in death rates (equation (17))				
CT-CoDA	0.00076	0.00320	0.00631	0.01366
2S-CoDA	0.00070	*0.00239*	*0.00586*	*0.01349*
VECM-CoDA	0.00085	0.00570	0.00634	0.01490
LC	*0.00056*	0.00324	0.00654	0.01571

†The lowest RMSE forecast error is indicated by italics.



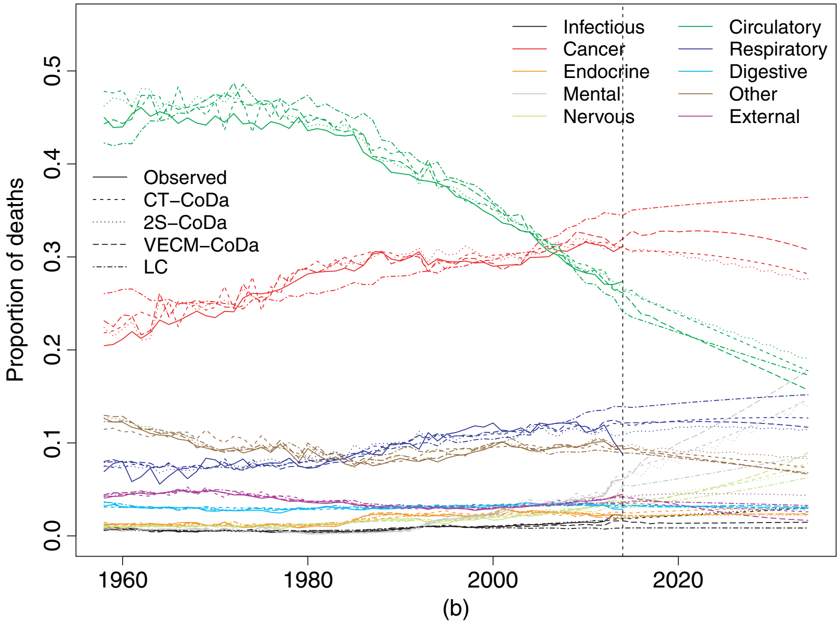


Fig. 5. Proportion of life table deaths: fitted and 20-year forecasts for (a) French females and (b) Dutch males (⁞, start of the forecast)

and injury and poisoning but differ for cancer, and other and cardiovascular disease, which are the largest causes of deaths. (Tables E1 and E2 in the on-line supplementary material present the RMSEs of the proportion of life table deaths for cause for French females and Dutch males respectively.) Hence, the better forecast performance of 2S-CoDA in terms of cancer is not only due to cancer alone but also the other causes to which it is linked through the sum constraint (see equation (5)). Similar results are found for Dutch males where the fit and forecast differ between the models for circulatory diseases, cancer, respiratory diseases and mental diseases, whereas the fit and forecasts for the rest of the causes are similar for all models. Note that mental diseases in the French data are aggregated in other diseases, which is likely to be the cause of the predicted increase in deaths from other diseases.

4. Discussion

The analysis that is presented in this paper extends and improves cause-specific death forecasts using CODA by developing a CODA-based forecasting model with the benefit of explicit modelling of dependence between cause-specific mortality, introduction of age-, period- and cause-specific weights, and modelling of both joint and cause-specific variation. We find that the new 2S-CoDA model provides more accurate forecasts of the number of cancer deaths. CODA enables coherent modelling of cause-specific mortality where dependences between causes are explicitly modelled, so a relative improvement in survival for one cause leads to a decline in the relative survival for the remaining causes. Thus, CODA models provide a more satisfactory modelling of cause-specific mortality than applying the LC model to each cause separately as dependence between causes is ignored in the LC model. This paper focuses on forecasting cancer by improving CT-CoDA, suggested by Oeppen (2008), and finds that through adding age-, cause- and time-specific weights and by introducing a cause-specific decomposition the forecast accuracy of CT-CoDA can be improved. The best performing model (2S-CoDA) has a better forecast performance in three of the four populations that were used in the analysis compared with CT-CoDA and the LC model. 2S-CoDA is also less sensitive to aggregation among causes of deaths as each cause is weighted by its size. VECM-CoDA, which allows for multiple trends, does not provide better forecasts compared with CT-CoDA despite a rejection of CT-CoDA's assumption about all causes being described by one common trend. Assuming one time trend for all causes thus constitutes a reasonable assumption when forecasting as a simplification of a more complicated set of different time trends. Hence, a simpler model such as CT-CoDA or 2S-CoDA performs better when compared with the more complex VECM-CoDA. That simpler models perform better than complex models is not new in forecasting. For example Green and Armstrong (2015) found, in a meta-analysis study, that complex models failed to improve the forecast of simple models in 81 out of 97 comparisons.

The suggested CODA models provide valuable information that can be used for the planning of healthcare and targeting of public health actions. In healthcare planning, the CODA models provide information about the future number of cancer deaths relative to the total population. Thus, potentially, implied incidence rates could be calculated for cancer forecasts with information about the relative risk of cancer. The proportion of cancer deaths is expected to decline slightly for French females and Dutch females and males in 2030 when considering forecasts from 2S-CoDA, which was the most accurate of the models that were considered. Only for French males is the proportion of cancer deaths expected to increase in 2030. Hence, cancer mortality is expected to decline faster relative to the other causes for three of the four populations. As the CODA models are forecasting cause-specific mortality coherently for all causes they also predict which causes will be the main causes of death in the future. For example, for Dutch males the relative numbers of circulatory diseases and cancer deaths are predicted to decrease whereas an increase for mental diseases is forecast. To improve the general survival for Dutch males it is thus necessary to consider treatments or life extending procedures for mental diseases. Our results demonstrate that causes which, today, are considered as the natural public health target because of their size might not be large in the future. As medical research takes years before better treatments are ready, research resources should be allocated so that they contribute to future survival. The CODA models suggested have the potential to inform such public health strategies.

The CODA models are, like the LC model, trend models which extrapolate time trends that are identified in the data. Thus, the models are sensitive to breaks in the data and cannot necessarily predict new trends after a break. 2S-CoDA is less sensitive to new trends as recent observations are weighted more than past observations and hence new trends will be fitted better in 2S-CoDA. Users of the models should be careful when using them on data with many trend breaks.

None of the models considered include cohort effects to account for specific survival in some cohorts. It is very likely that cohort components could improve the fit and forecast in some populations, as has been found by Renshaw and Haberman (2006) for all-cause mortality. However, it is not straightforward to implement cohort effects because of the unique relationship between age, time and cohort which makes it problematic to identify each component (Holford, 1983). A natural extension of cause-specific mortality forecasting using CODA models is to include cohort components. Further, models forecasting all-cause mortality which relate a forecast for a single country to international trends have been suggested to provide more stable and accurate forecasts (Li and Lee, 2005; Cairns *et al.*, 2009; Hyndman *et al.*, 2013). It is possible that changes in the relative importance of causes are shared by multiple countries since medical interventions and health trends can be shared: for example a decline in circulatory diseases is found in both France and the Netherlands. Another possible extension of the cause-specific forecasting models is thus to include shared trends among countries.

Acknowledgements

The authors are grateful to the Joint Editor and two referees whose helpful suggestions and constructive comments have led to an improved version of the paper. The authors also thank James W. Vaupel and the participants in the Longevity 14 conference for their helpful comments and discussions. The work by Søren Kjærgaard and Jim Oeppen was completed with the support of the Axa Research Fund. The work of Malene Kallestrup-Lamb has been supported by PeRCent, which receives funding from the Danish pension funds and Copenhagen Business School.

References

A, T. (1975) A nonidentifiability aspect of the problem of competing risks. *Proc. Natn. Acad. Sci. USA*, **72**, 20–22.

Abdi, H. (2011) *Encyclopedia of Measurement and Statistics*. New York: Sage.

Aitchison, J. (1982) The statistical analysis of compositional data (with discussion). *J. R. Statist. Soc.* B, **44**, 139–177.

Aitchison, J. (1986) *The Statistical Analysis of Compositional Data*. London: Chapman and Hall.

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Trans. Autom. Control*, **19**, 716–723.

Andersen, P. K., Geskus, R. B., de Witte, T. and Putter, H. (2012) Competing risks in epidemiology: possibilities and pitfalls. *Int. J. Epidem.*, **41**, 861–870.

Arnold-Gaille, S. and Sherris, M. (2013) Forecasting mortality trends allowing for cause-of-death mortality dependence. *Nth Am. Act. J.*, **17**, 273–282.

Bergeron-Boucher, M.-P., Canudas-Romo, V., Oeppen, J. and Vaupel, J. W. (2017) Coherent forecasts of mortality with compositional data analysis. *Demog. Res.*, **37**, 527–566.

Bergeron-Boucher, M.-P., Ebeling, M. and Canudas-Romo, V. (2015) Decomposing changes in life expectancy: compression versus shifting mortality. *Demog. Res.*, **33**, 391–424.

Booth, H., Hyndman, R. J., Tickle, L. and de Jong, P. (2006) Lee-Carter mortality forecasting, a multi-country comparison of variants and extensions. *Demog. Res.*, **15**, 289–310.

Box, G. and Jenkins, G. (1970) *Time Series Analysis Forecasting and Control*. San Francisco: Holden-Day.

Cairns, A. J. G., Blake, D., Down, K., Coughland, G. D., Epstein, D., Ong, A. and Balevich, I. (2009) A quantitative comparison of stochastic mortality models using data from England and Wales and The United States. *Nth Am. Act. J.*, **13**, 1–35.

Cesare, M. D. and Murphy, M. (2011) Forecasting mortality, different approaches for different cause of deaths?: The cases of lung cancer; influenza, pneumonia, and bronchitis; and motor vehicle accidents. *Br. Act. J.*, **15**, 185–211.

Czado, C., Gneiting, T. and Held, L. (2009) Predictive model assessment for count data. *Biometrics*, **65**, 1254–1261.

Dickey, D. A. and Fuller, W. A. (1979) Distribution of the estimators for autoregressive time series with a unit root. *J. Am. Statist. Ass.*, **74**, 427–431.

Eurostat (2017) European cancer information system, cancer statistics. *Technical Report*. Eurostat, Luxembourg.

Foreman, K. J., Li, G., Best, N. and Ezzati, M. (2017) Small area forecasts of cause-specific mortality: application of a Bayesian hierarchical model to US vital registration data. *Appl. Statist.*, **66**, 121–139.

Girosi, F. and King, G. (2008) *Demographic Forecasting*. Princeton: Princeton University Press.

Golub, G. H. and van Loan, C. F. (1996) *Matrix Computations*, 3rd edn. Baltimore: Johns Hopkins University Press.

Green, K. and Armstrong, J. (2015) Simple versus complex forecasting: the evidence. *J. Bus. Res.*, **68**, 1678–1685.

Hirz, J., Schmock, U. and Shevchenko, P. (2017) Actuarial applications and estimation of extended CreditRisk$^+$. *Risks*, **5**, no. 2, 1–23.

Holford, T. R. (1983) The estimation of age, period and cohort effects for vital rates. *Biometrics*, **39**, 311–324.

Human Mortality Database (2018) Human mortality database. *Technical Report*. University of California, Berkeley, and Max Planck Institute for Demographic Research, Rostock. (Available from `www.mortality.org`.)

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

Institut National d'Etudes Démographiques (2016) Database. *Technical Report*. Institut National d'Etudes Démographiques, Paris. (Available from `https://www.ined.fr/en/`.)

Johansen, S. (1991) Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. *Econometrica*, **59**, 1551–1580.

Jönsson, B., Hofmarcher, T., Lindgren, P. and Wilking, N. (2016) The cost and burden of cancer in the European Union 1995–2014. *Eur. J. Cancer*, **66**, 162–170.

Knorr-Held, L. and Rainer, E. (2001) Projections of lung cancer mortality in West Germany: a case study in Bayesian prediction. *Biostatistics*, **2**, 109–129.

Koren, W., Harteloh, P., Kardaun, J. and van der Stegen, R. (2012) Reconstruction possibilities of long-term time series of causes of death. *Technical Report*. Statistics Netherlands, The Hague.

Lee, R. D. and Carter, L. R. (1992) Modeling and forecasting U. S. mortality. *J. Am. Statist. Ass.*, **87**, 659–671.

Lee, R. D. and Miller, T. (2001) Evaluating the performance of the Lee-Carter method for forecasting mortality. *Demography*, **38**, 537–549.

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

van Loan, C. F. (1976) Generalizing the singular value decomposition. *SIAM J. Numer. Anal.*, **13**, 76–83.

Lock, E. F., Hoadley, K. A., Marron, J. S. and Nobel, A. B. (2013) Joint and individual variation explained (jive) for integrated analysis of multiple data types. *Ann. Appl. Statist.*, **7**, 523–542.

Mathers, C. D. and Loncar, D. (2006) Projections of global mortality and burden of disease from 2002 to 2030. *PLOS Med.*, **3**, 1–20.

Oeppen, J. (2008) Coherent forecasting of multiple-decrement life tables: a test using Japanese cause of death data. *European Population Conf.*

Pawlowsky-Glahn, V. and Buccianti, A. (2011) *Compositional Data Analysis: Theory and Applications*. Chichester: Wiley.

Peltonen, M. and Asplund, K. (1996) Age-period-cohort effects on stroke mortality in Sweden 1969-1993 and forecasts up to the year 2003. *Stroke*, **27**, 1981–1985.

Preston, S., Heuveline, P. and Guillot, M. (2001) *Demography, Measuring and Modeling Population Processes*. Oxford: Blackwell Publishers.

Rapiti, E., Guarnori, S., Pastoors, B., Miralbell, R. and Usel, M. (2014) Planning for the future: cancer incidence projections in Switzerland up to 2019. *BMC Publ. Hlth*, **14**, 95–102.

Rau, R., Soroko, E., Jasilionis, D. and Vaupel, J. W. (2008) Continued reductions in mortality at advanced ages. *Popln Devlpmnt Rev.*, **34**, 747–768.

Renshaw, A. and Haberman, S. (2006) A cohort-based extension to the Lee–Carter model for mortality reduction factors. *Insur. Math. Econ.*, **38**, 556–570.

Statistics Netherlands (2018) Database. Statistics Netherlands, The Hague. (Available from `https://opendata.cbs.nl/statline//CBS/en/dataset`.)

Wilmoth, J. R. (1995) Are mortality projections always more pessimistic when disaggregated by cause of death? *Math. Popln Stud.*, **5**, 293–319.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article.

'Supplementary material to: Forecasting causes of death using compositional data analysis: the case of cancer deaths'.