Max-Planck-Institut für demografische Forschung

Max Planck Institute for Demographic Research

Konrad-Zuse-Strasse 1 ▪ D-18057 Rostock ▪ Germany ▪ Tel +49 (0) 3 81 20 81 - 0 ▪ Fax +49 (0) 3 81 20 81 - 202 ▪ www.demogr.mpg.de

MPIDR Technical Report TR 2018-003 I January 2018

Analyzing the Young Adult Mortality Hump in R with MortHump

Adrien Remund
Carlo G. Camarda
Tim Riffe I riffe@demogr.mpg.de

For additional material see www.demogr.mpg.de/tr/

This technical report has been approved for release by: Vladimir Shkolnikov (shkolnikov@demogr.mpg.de), Editor of the MPIDR Technical Reports.

Technical reports of the Max Planck Institute for Demographic Research receive only limited review. Views or opinions expressed in technical reports are attributable to the authors and do not necessarily reflect those of the Institute.

Analyzing the Young Adult Mortality Hump in R with MortHump

Adrien Remund[1,2], Carlo G. Camarda[2], and Tim Riffe[3]

[1] *University of Geneva*
[2] *Institut national d'études démographiques*
[3] *Max Planck Institute for Demographic Research*

December 1, 2017

Abstract

`MortHump` is an R package designed to provide ready-to-use methods to analyse the young adult mortality hump. It contains functions to

- **format** all-cause and cause-of-death data from the Human Mortality Database (HMD) and the Human Cause-of-Death Database (HCD) respectively,
- **identify** and group causes of death that are likely to contribute to the young adult mortality hump,
- **estimate** parametric and non-parametric models that isolate the young adult mortality hump from the rest of the force of mortality, decomposing when needed by cause of death,
- **measure** the young adult mortality hump by computing summary statistics about its magnitude, location and spread, optionally by cause of death.

This technical paper is meant as a user guide for the `MortHump` package and provides examples on how to use its functions.

1 Introduction

Human mortality patterns usually include a brief period of excess mortality in young adult ages, often called the young adult mortality hump. Although the hump was first described long ago (Thiele, 1871), recognizability has not led to extensive theoretical or analytic attention. Consequently, empirical research on the hump has been scarce. Parametric models that do separate the hump have done so for the sake of a better fit to all-cause mortality, but these have not been used to study the hump specifically. Important questions therefore remain unanswered. It was for instance claimed that the hump is a universal feature of male populations (Heligman and Pollard, 1980; Goldstein, 2011), although this assertion is disputable (Remund, 2012). Moreover, despite the extensive use of the term "accident hump", its composition by cause of death remains poorly studied. These considerations make the development of dedicated tools all the more important for the study of the evolution and international comparison of the young adult mortality hump.

Human mortality is characterized by the level and shape of the age-specific death rates, which are defined as the ratio between the observed number of deaths by age, and the person-years lived in this age. The set of observed rates can be conceptualized as a realization of a latent *force of mortality.* The force of mortality can be divided into three main phases that characterize specific periods in the lifecourse (Figure 1). During childhood, the force of mortality decreases in a process known as ontogenescence (Levitis, 2011). During adulthood, the force of mortality increases exponentially due to senescence (Gompertz, 1825), until about age 90, when it appears to decelerate to a plateau (Vaupel, 1997; Horiuchi and Wilmoth, 1998). Between childhood and adulthood, the force of mortality often

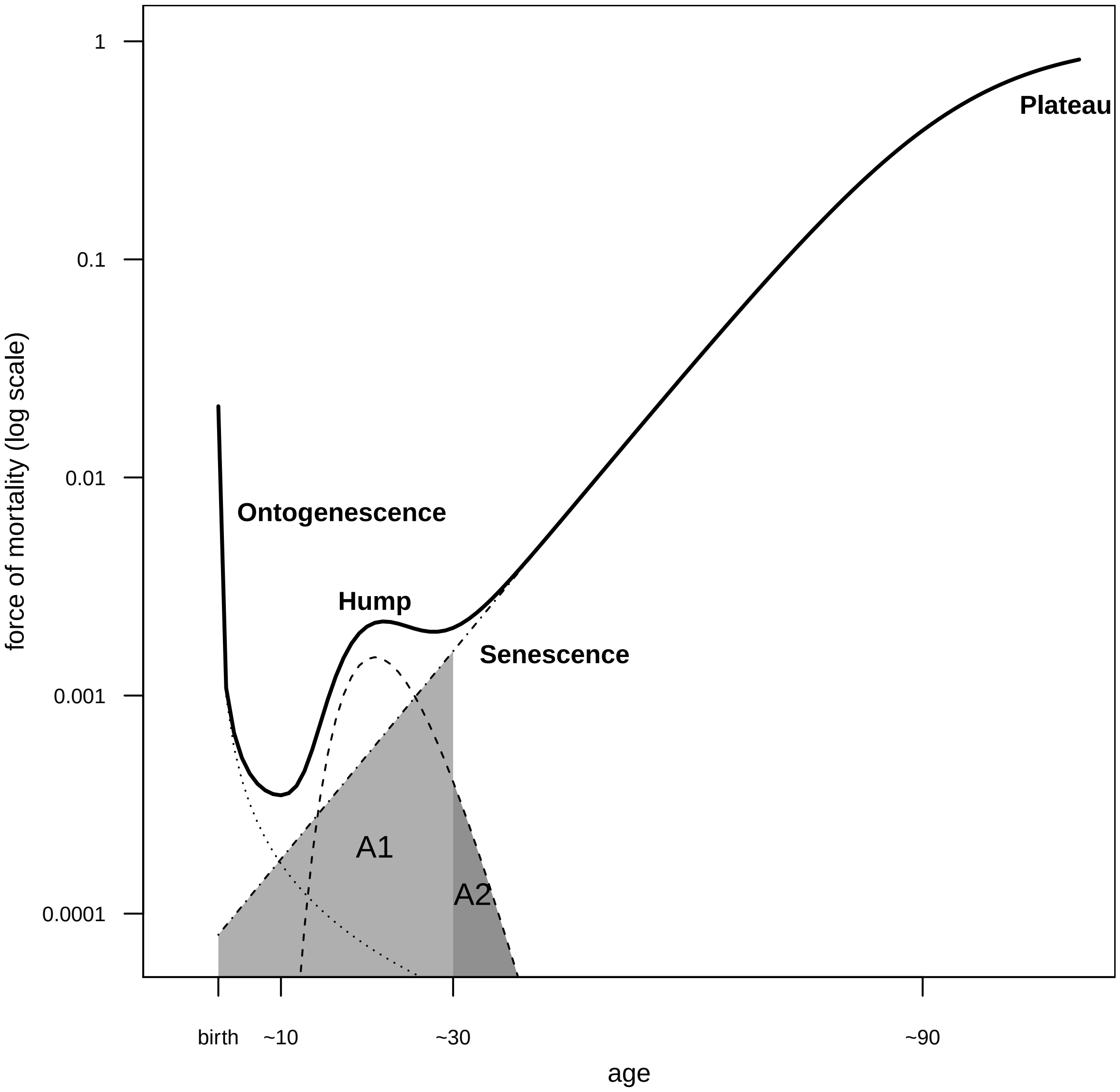

Figure 1: Schematic evolution of the risk of death over the life course. The force of mortality is usually composed of three phases: a decreasing trend during the first decade of life, a hump in the second and third decade, and an increasing trend thereafter, marked by a progressive deceleration in very old age.

includes what can be described as a hump. This feature is mostly visible between about 10 and 30 years of age, although it may extend further.

Studies addressing young adult mortality must take into account that at all ages deaths can be attributable to any of these three processes, although they each dominate a specific period of life. In particular, young adults are likely influenced by the same forces shaping senescent mortality in higher ages. If we accept this possibility, then the young-adult senescent pattern may be projected from the observations at older ages, leaving the hump as an identifiable excess. In this sense, it is possible to decompose the force of mortality using an additive model in which it is the sum of different components corresponding to the phases described. Figure 1 illustrates this additive construction and hints at the arbitrariness of setting strict age bounds for the hump. In this example the total force of mortality starts increasing again around age 30. Setting age 30 as the end of early adulthood would however result in attributing senescent deaths before age 30 (area $A1$) to the hump and ignoring deaths after age 30 that belong to the hump component (area $A2$).

The decomposition of the force of mortality into additive component has traditionally been done using parametric models because they are relatively easy to fit by least squares or maximum likelihood, and because their parameters can be relatively well interpreted. Examples of such parametric models include those proposed by Thiele (1871), Heligman and Pollard (1980), Mode and Busby (1982) or Kostaki (1992), to cite only some that include a young adult mortality hump.

Parametric models have the advantage of being easy to estimate but suffer from the important flaws of being relatively arbitrary and rigid in their formulation, which translated in the literature by a race to the best model that culminated in the 1980s and early 1990s and resulted in the presence of dozens of competing models whose respective merits are often hard to compare (Wunsch et al., 2002). A promising alternative consists in estimating the force of mortality with the help of non-parametric models, such as *splines*. This line of research has been explored in the last decade, for instance by Camarda (2008); Camarda et al. (2016).

In the `MortHump` package we implement several models, both parametric and non-parametric, that can be fitted to real data. Moreover, we offer a ready-to-use application of a new model that combines the best of parametric (i.e. formulation in terms of components) and non-parametric models (i.e. adaptiveness to any mortality context). This so-called *Sum of Smooth Exponentials* (SSE) model was defined by Camarda et al. (2016) and was extended to cause-of-death analysis by Remund et al. (2017). The `MortHump` package is not limited to the fitting procedure, but also includes methods to format data from popular demographic databases, identify and group causes of death that possibly contribute to the hump, as well as provide summary statistics about the magnitude, location and spread of the hump.

In the next sections, we present how these tasks can be performed using our pre-defined functions. We first present the tools designed to study all-cause mortality and then move to cause-specific analyses. Two cases are used as main examples, namely Swiss males in 2010 for all-cause mortality, and American males in 2000 for cause-specific mortality. These data were obtained from the *Human Mortality Database* (HMD) and the *Human Cause-of-Death Database* (HCD, 2017), respectively.

2 All-cause mortality

In this section we present how to (1) extract and format data from the Human Mortality Database (HMD), (2) fit different models of mortality, including parametric and non-parametric options, (3) compute summary statistics about the hump from these models. We use data from the HMD and show all steps required to perform the analysis using the `MortHump` package. To access it, install it from CRAN and load it with the following commands.

```r
install.packages("MortHump")

library(MortHump)
```

2.1 Format data

Mortality data from the HMD can be accessed either online or locally with the function `HMD2MH()`, which can be used in the following manner.

```r
# For local access, replace "path" with the file path to the HMD data.
# For online access, replace "user" and "pass" with valid HMD username and password.
# To register for the HMD, go to www.mortality.org

# Period: Swiss males in 2010 (truncated at age 90)
 CHE2010m <- HMD2MH(country = "CHE", year = 2010, sex = "males",
                   dim = "period", min = 0, max = 90,
                   password = pass, username = user)
```

```r
# Cohort: Swiss females born in 1980 (extrapolation after last available year)
 CHE1980fc <- HMD2MH(country = "CHE", year = 1980, sex = "females",
                    dim = "cohort", xtra = TRUE, min = 0, max = 90,
                    path = path)
```

The function takes several arguments, some of them to select the desired combination of `country`, `years`, and `sex`, others to indicate where to find the data (`path` for local access, `username` and `password` for online access), and others to format the data (`min` and `max` to truncate by age). Additionally, it is possible to chose the dimension of the data (period or cohort), by using the `dim` argument. In the cohort case, the `xtp` option allows extrapolating the data for the non-extinct cohorts using a variant of the Lee-Carter model (Hyndman and Ullah, 2007, see documentation of the `xtp()` function).

The `HMD2MH()` function produces a data frame with the following variables: `x` = age, `d` = death counts, `n` = exposures, `m` = rates. User-supplied datasets can be used but they need to be supplied in the same format in order to be usable by the other functions of the `MortHump` package. The format of the object must be as follows.

```r
str(CHE2010m)

## 'data.frame': 91 obs. of  4 variables:
##  $ x: int  0 1 2 3 4 5 6 7 8 9 ...
##  $ d: num  151 7 4 5 5 5 2 1 4 1 ...
##  $ n: num  40069 40337 39880 39240 39276 ...
##  $ m: num  0.003768 0.000174 0.0001 0.000127 0.000127 ...
```

2.2 Estimate models

Studying the young adult mortality hump requires modelling the shape of the force of mortality in a flexible way. The general aim is to describe the age-specific death rates m_x as a function of age (Equation 1), where the parameters θ can represent explicit aspects of the force of mortality (such as the β of the Gompertz law that represents the rate of ageing due to senescence) or a more descriptive set of coefficients (such as the coefficients of a spline basis).

$$m_x = \mu(x) + \epsilon_x = \gamma_C(x, \theta_C) + \gamma_H(x, \theta_H) + \gamma_S(x, \theta_S) + \epsilon_x \tag{1}$$

Whatever the form of the model, it is designed to fit the observed mortality rates and decompose the underlying force of mortality into three additive components corresponding to specific periods of the life course: $\gamma_C(x, \theta_C)$ for the decrease in the risk of death during childhood, $\gamma_H(x, \theta_H)$ for the hump that characterizes young adult excess mortality, and $\gamma_S(x, \theta_S)$ for the senescence process characterizing adulthood. In this section we compare three parametric and one non-parametric models that can be all estimated with the `MortHump` package in a straightforward fashion.

2.2.1 Parametric models

Dozens of parametric models have been published (Wunsch et al., 2002), but only three of them are implemented here[1]. They are all based on the structure of the Heligman-Pollard model (Heligman and Pollard, 1980), and are conceived as nested models that offer a range of ways to address the specificity of young adult mortality.

The `"hps"` model, which is inspired by the Siler model (Siler, 1979), is the simplest because it only models ontogenescence and senescence, and does not take the hump into account. Its role is mainly to serve as a "null" model to be compared with more complex models that incorporate a hump. The `"hp"` model implements the Heligman-Pollard model. This assumes a symmetrical hump with a given

[1] For a more comprehensive implementation of other mortality laws, see the `MortalityLaws` R package (Pascariu, 2017).

height (D), spread (E) and location (F). The "hpk" model, which corresponds to the model suggested by Kostaki (1992), relaxes the assumption of symmetry by introducing a different spread before and after the peak of the hump. Algebraically, "hp" is equal to "hps" iff $E = 0$, and "hpk" is equal to "hp" iff $E1 = E2$.

"hp" Published by Heligman and Pollard (1980), this model contains eight parameters in three additive terms.

$$\mu(x) = A^{(x+B)^C} + D \cdot exp(-E \cdot (log(x) - log(F))^2) + \frac{G \cdot H^x}{1 + G \cdot H^x} \tag{2}$$

where (1) A, B and C describe the infant mortality component, (2) D, E and F describe the hump component, and (3) G and H describe the senescence component.

"hpk" Published by Kostaki (1992), this extension of "hp" contains nine parameters in three additive terms. This model is defined by two equations that describe the evolution of the force of mortality before and after the peak of the hump respectively.

$$\mu(x) = \begin{cases} A^{(x+B)^C} + D \cdot exp(-E_1 \cdot (log(x) - log(F))^2) + \frac{G \cdot H^x}{1+G \cdot H^x} & \forall x \leq F \\ A^{(x+B)^C} + D \cdot exp(-E_2 \cdot (log(x) - log(F))^2) + \frac{G \cdot H^x}{1+G \cdot H^x} & \forall x > F \end{cases} \tag{3}$$

In contrast to the "hp" model, it allows the hump to be asymmetrical by differentiating the spread of the hump before (E_1) and after (E_2) its peak. Another formulation of this model consists in substituting $E_2 = E_1 \cdot k$, leaving thus a single value for the E parameter, whose spread is scaled after the peak of the hump by a coefficient k. This second formulation is used in the outputs of the MortHump package.

"hps" This model is an adaptation of the "hp" model that does not account for the young adult hump. It resembles in this respect the model published by Siler (1979), but is nested in the "hp" model.

$$\mu(x) = A^{(x+B)^C} + D + \frac{G \cdot H^x}{1 + G \cdot H^x} \tag{4}$$

In contrast to the "hp" model, it only has one parameter for the hump component (D) that has the same function as the constant term of the Makeham model (1860), i.e. to absorbe the "white noise" of mortality unrelated with age.

For each of these models, the response variable is defined as m_x (i.e. the age-specific mortality rates), x stands for age and is the only covariate, while capital letters represent the parameters that need to be estimated. In their original paper, Heligman and Pollard (1980) use the odds of death as the response variable ($\frac{q_x}{1-q_x}$), a choice that they justify by the fact that this limits the possible values to the interval [0-1]. We chose instead to define all three models on m_x for ease of comparison and because its domain spans all real numbers.

Estimation is done by weighted least squares to avoid heteroskedasticity, using the inverse of rates ($\frac{1}{m_x}$) as the weights. Heligman and Pollard (1980) advise using quadratic weights (or relative least squares, which is the same[2]), while Brillinger (1986) advises to use the inverse of the death counts as weights[3]. A comparison of these options shows that $\frac{1}{m_x}$ respects the hypotheses of the Gauss-Markov theorem (Remund, 2015). However, the user can choose other weights if desired by specifying other values for the w argument.

[2] $S^2 = \sum_{x=0}^{\omega}(\frac{\widehat{m_x}}{m_x} - 1)^2 = \sum_{x=0}^{\omega}(\frac{1}{m_x} \cdot (\widehat{m_x} - m_x))^2 = \sum_{x=0}^{\omega}\frac{1}{m_x^2} \cdot (\widehat{m_x} - m_x)^2$

[3] By default, parametric models are estimated with the "port" algorithm from the nls() function. Is some rare cases, it may become stuck into local minimums. If this happens, try switching to the Levenberg-Marcquart algorithm using the "method" argument of the morthump() function.

Fitting any of these models is straightforward with the MortHump package, using the morthump() function. The user only needs to specify the data (formatted by the HMD2MH() function or similarly structured), and the type of model ("hp", "hpk" or "hps"). For instance, fitting the Heligman-Pollard model on Swiss males in 2010 is as follows.

```
# load data for Swiss males in 2010 (HMD)
data(CHE2010m)

# fit the Heligman-Pollard model
fit.hp.default <- morthump(data = CHE2010m, model = "hp")

# the fitted value of the F parameter (location of the peak of the hump)
# corresponds to its upper bound (25)
coef(fit.hp.default)

##               A               B               C               D
##  0.000225945531  0.069297497186  0.152846905932  0.000406578030
##               E               F               G               H
##  2.026465548113 25.000000000000  0.000004093011  1.128283819765
```

In this example, the F parameter that indicates the location of the peak of the hump is problematic because using the default starting, lower and upper values, this parameter is estimated at 25 years of age, which corresponds to its default upper bound. It is also possible to remove this upper bound by setting it to Infinity, using the following code.

```
# change upper bound for the F parameter from 25 to Infinity (!)
st <- list(
 start = list(A = 1e-3, B = 5e-3, C = 0.11, D = 15e-4, E = 8, F = 20, G = 3e-5, H = 1.105),
 lower = c(1e-4, 1e-6, 1e-4, 0, 1, 16, 1e-7, 0.5),
 upper = c(0.1, 0.5, 1, 0.01, 50, Inf, 0.01, 1.5))

# refit the model with new constraints
fit.hp.new <- morthump(data = CHE2010m, model = "hp", start = st)

# look at the new fitted parameters
coef(fit.hp.new)

##            A            B            C            D            E
## 1.781458e-04 1.000000e-06 3.789006e-02 1.000000e-02 1.687957e+00
##            F            G            H
## 1.294310e+02 5.138385e-07 1.155075e+00
```

This results in a new (absurd) fitted value of 129 years for the peak of the hump (F), and to new problematic values for parameters B and D, which is a clear sign that the hump component is misused. More specifically, the hump component is not really used to fit a young adult mortality hump, but instead corrects for a slight bend in the senescence component around age 60. Although the hump component is intended to capture the young adult mortality hump, which is usually the most prominent feature deviating from the senescence trend, in some (not so rare) cases this senescence trend does not exactly follow a strictly exponential trend. In this case the optimization algorithm "recycles" the hump to compensate for these departures from the exponential trend.

This problem is obvious on the left panel of Figure 2 that represents the age-specific death rates and the estimated force of mortality resulting from this unconstrained Heligman-Pollard model. One can see that, between about 40 and 70 years of age, the observed death rates fall slightly above the expected exponential trend, which is why the hump component is relocated around these ages[4]. On

[4] Do not forget that the estimation is done on the absolute rates and not on the logged values (as showed in the plot).

the right panel of Figure 2, the hump component is represented alone as a share of the observed age-specific death rates (which can be reduced to a density by rescaling the hump to sum to 1). Its median and mean are marked (see the next section about the measures of the hump). By definition, the mode of the hump corresponds to the F parameter in the Heligman-Pollard model (not shown here due to its extreme value of 129 years of age), while the median indicates the age at which half of the people have died from the hump (here 74 years of age), and the mean indicates the mean age at death of the people who died from the hump (here 71 years of age). In this case all of these measures of centrality confirm that the hump component was used for other purposes than describing the young adult mortality hump.

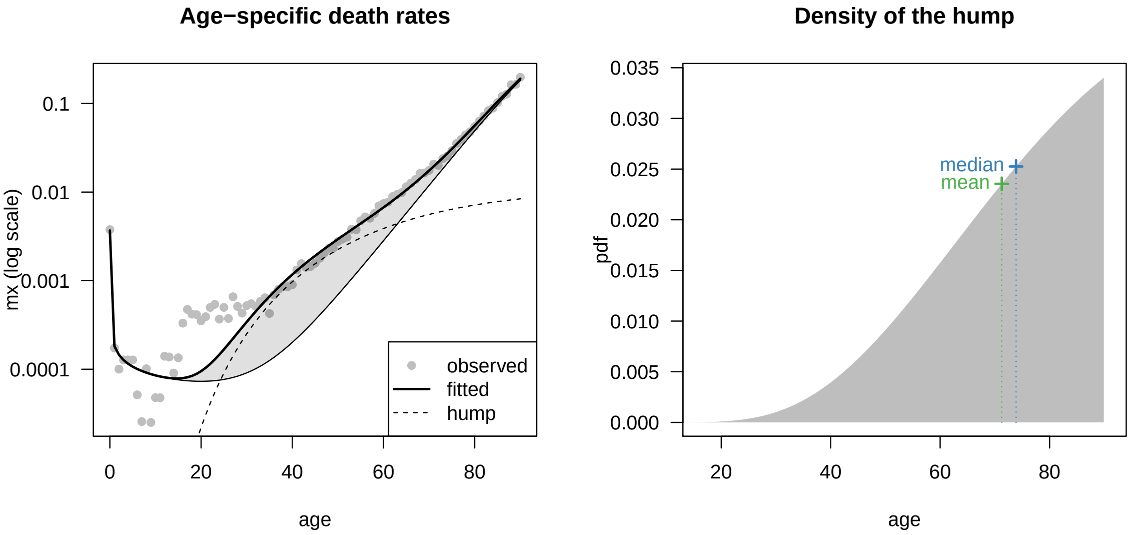

Figure 2: Fitted Heligman-Pollard model on Swiss males in 2010

A possible way around this problem is to fit the slightly more complex Kostaki model. This extension of the Heligman-Pollard formula relaxes the assumption of symmetry by allowing the spread of the hump to vary before and after its peak. This model can be easily fitted by specifying the "hpk" model in the morthump() function as follows.

```
# load data for Swiss males in 2010 (HMD)
data(CHE2010m)

# fit the Kostaki model
fit.hpk <- morthump(data = CHE2010m, model = "hpk")

# look at the fitted parameters
coef(fit.hpk)

##                A                B                C                D
##   0.000190851683  0.007068709528  0.086459317334  0.000495567376
##                E                F                G                H
##   8.191326548161 22.000000000000  0.000003114126  1.131899023381
##                k
##   0.000000000000
```

It may thus very well be that in absolute terms the deviation from the exponential trend is actually stronger around age 50 than around age 25, although the graph suggests otherwise.

There is now a ninth parameter k, which indicates by how much the spread after the peak of the hump differs from the spread before the hump. When $k = 1$, $E_1 = E_2$ and the Kostaki model reduces to the Heligman-Pollard model, and when $k = 0$, this means that the hump is in fact flat after the peak. According to the estimated values of the parameters, this is what happens, which allows for a better goodness of fit of the death rates between 20 and 40 years of age (Figure 3). Note however that: (1) the F parameter still sticks to the upper bound which in the Kostaki model is set at 22 by default, (2) the fit is poor between 40 and 60 years of age, and (3) the hump now takes a totally unreasonable shape. While the mode is located by definition at 22 years of age, the mean and the median now reach values above 50 years of age. This is because the hump does not decrease at all after reaching its peak.

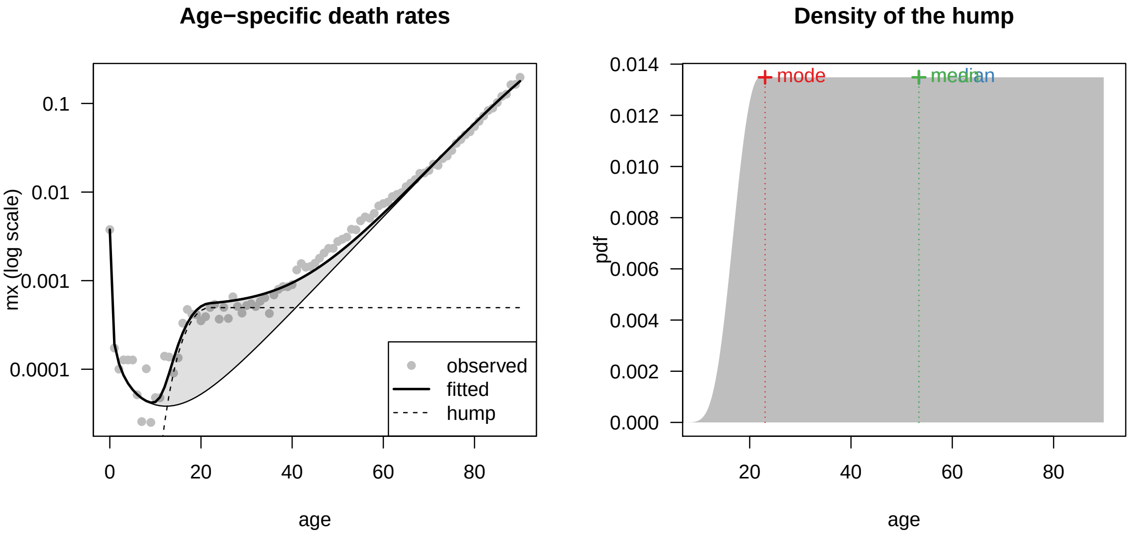

Figure 3: Fitted Kostaki model on Swiss males in 2010

We demonstrate particular example because the Heligman-Pollard model does not offer a good fit in all ages and tends to misuse the hump component to capture features of the force of mortality that have nothing to do with the young adult mortality hump. The Kostaki model may be better at modelling the overall force of mortality, but it tends to estimate unreasonable shapes for the hump. These drawbacks are less obvious when these models are estimated on data from the 1960s and 1970s, on which they were first developed, because at that time the hump had a more symmetrical shape that was easier to capture parametrically. To these limitations, one should also add that the parameters are highly correlated (Remund, 2015), which is why some studies have argued for the use of Bayesian estimation techniques of these parametric models (e.g. Dellaportas et al., 2001). Still, a potentially more promising option is to estimate the three components of the force of mortality with non-parametric techniques.

2.2.2 Non-parametric models

Age-specific mortality rates can alternatively be fitted with non-parametric models, such as P-splines (Camarda, 2008). These are are more flexible and are useful for smoothing, but they are less helpful in parsimoniously describing the characteristics of a mortality curve than parametric models. A possible compromise would keep the additive approach from parametric models that is needed to isolate the young adult mortality hump from the rest of the force of mortality, while defining each component non-parametrically. This is what does the *Sum of Smooth Exponentials* (*SSE*) model (Camarda et al., 2016).

In brief, this model describes the force of mortality as the sum of three smooth functions of age that correspond to ontogenescence, the young adult hump, and senescence. These smooth functions are estimated non-parametrically using penalised splines, forcing the three components to be monotonically decreasing, log-concave, and increasing, respectively. The model assumes that age-specific death counts are draws from Poisson distributions, and exponentiation ensures non-negative values. The model is formulated as a Penalized Composite Link Model (Thompson and Baker, 1981) and fitted using Iterative Re-weighted Least Squares (Eilers, 2007).

In the `MortHump` package, estimating the *SSE* model is done similarely to the parametric models, using the `morthump()` function and specifying `"sse"` in the `model` argument as follows.

```
# load data for Swiss males in 2010 (HMD)
data(CHE2010m)

# fit the SSE model
fit.sse <- morthump(data = CHE2010m, model = "sse")
```

In the case of Swiss males in 2010, the *SSE* model proves more efficient than its parametric counterparts. Not only does the estimated force of mortality more closely match the shape of the age-specific death rates at all ages, but the hump is also more symmetrical and concave (Figure 4). Indeed, the lack of fit observed between age 40 and 70 with the parametric models completely disappears thanks to the more flexible estimation of the rate of ageing (i.e. the β parameter of the Gompertz model). Moreover, the three measures of location (mode, mean and median) are very close to each other around age 22, which confirms the visual impression of symmetry.

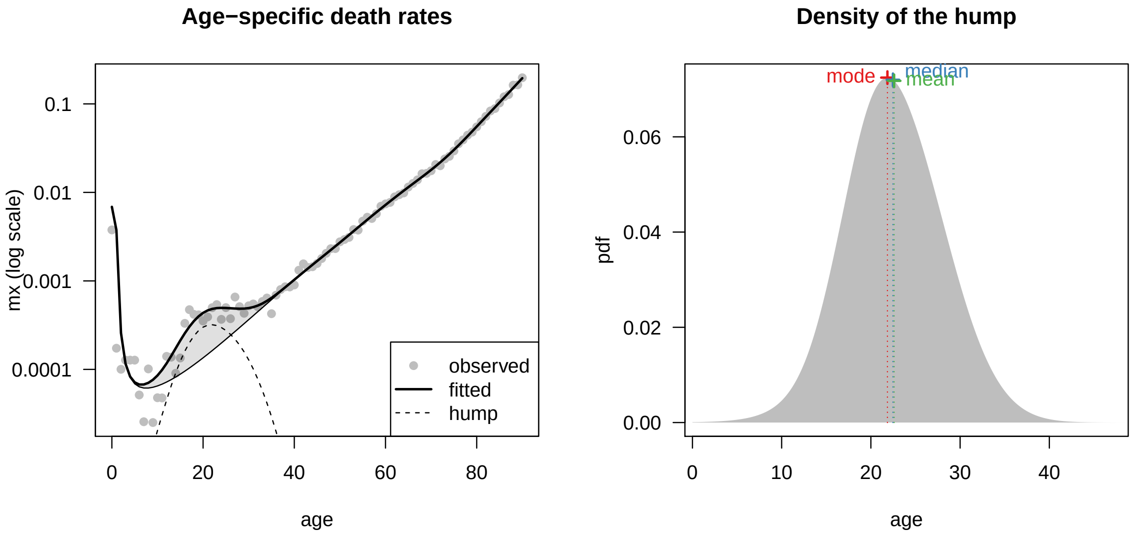

Figure 4: Fitted *SSE* model on Swiss males in 2010

This example shows the flexibility of the *SSE* model. It can adapt to virtually any shape of the force of human mortality, as long as the three components that it assumes are present. This nuance is important as there are examples of populations where the hump does not exist, or to such a small extent that it becomes indistinguishable from the stochastic noise due to the population size and the level of mortality. In this case, some fine tuning may help model convergence, but it does not mean that the result is meaningful. Several arguments might prove useful: a larger `maxit` increases the number of iterations and slows down the convergence, a larger `x1` helps capture a wider hump, while `lambda.hump` and `lambda.sen` control the rigidity of the hump and senescence components, respectively[5]. These arguments can be used for instance in the following way on the case of Swiss females in 1950.

[5] For more information see the documentation of the `morthump()` and `sse.fit()` functions

```
library(MortHump)

# load data for Swiss females in 1950 (HMD)
data(CHE1950f)

# fit sse model with default parameters
fit.sse.default <- morthump(data = CHE1950f, model = "sse")

# change starting values for the hump (x1) and senescence (x2) components,
# and amount of smoothing for the senescence component (lambda.sen)
fit.sse.custom <- morthump(data = CHE1950f, model = "sse", x1 = 25, x2 = 40,
    lambda.sen = 1)
```

The resulting fit of the *SSE* model on the case of Swiss females in 1950 shows how flexible this method is, but also that its results must be sometimes interpreted with caution (Figure 5). Just looking at the observed age-specific death rates, we can clearly see that there is almost no perceptible deviation around age 20. Consequently, if we estimate the *SSE* model using the default values for the parameters (Figure 5, left panel), the resulting hump component is unreasonable. It is possible to force the *SSE* model to focus on a more specific age range and reduce the smoothing parameter of the senescence component to tolerate more variation in the slope of the senescence component. The result is a very small hump estimate around age 20 (Figure 5, right panel), although its significance remains dubious. We see in the next section how to treat the question of statistical significance, both in parametric and non-parametric models.

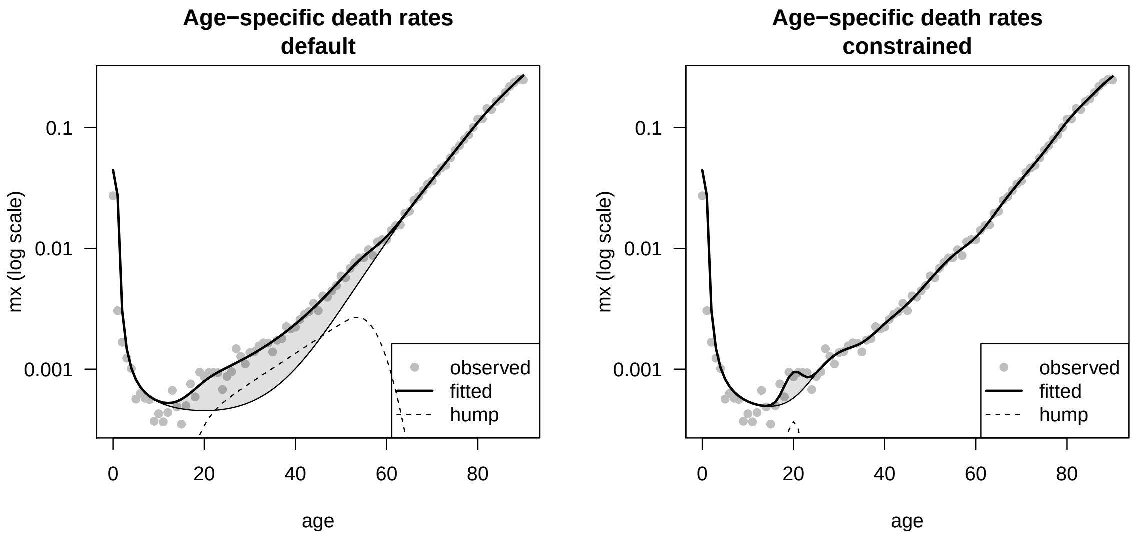

Figure 5: Fitted *SSE* model on Swiss females in 1950, using default (left) and custom (right) parameters

2.3 Measure the hump

We provide a small suite of summary methods for interpreting model output. Three dimensions are especially relevant in this context: the magnitude, the location and the spread of the hump. There are different ways to capture these concepts and we will now illustrate some of them using the same data of Swiss females in 1950. Let us thus estimate a parametric (Heligman-Pollard) and non-parametric (SSE) model on these data. Because this population has a very small hump, we need to use custom parameters for both models in order to avoid that the hump and the senescence components overlap.

In particular, for the Heligman Pollard model the mode of the hump must be kept low (F parameter). For the *SSE* model this involves chaning the age range used to find starting values for the hump (`"x1"`) and senescence component (`"x2"`), and allowing more flexibility to the senescence component in order to absorbe the variations observed at older ages (`"lambda.sen"`).

```
# load data for Swiss females in 1950 (HMD)
data(CHE1950f)

# fit the HP model with custom constraints
st <- list(
 start = list(A = 1e-3, B = 5e-3, C = 0.11, D = 15e-4, E = 40, F = 20, G = 3e-5, H = 1.1),
 lower = c(1e-4, 1e-6, 1e-4, 0, 30, 19, 1e-7, 0.5),
 upper = c(0.1, 0.5, 1, 0.01, 50, 21, 0.01, 1.5))

fit.hp.custom <- morthump(data = CHE1950f, model = "hp", start = st)

# fit the SSE model with custom starting values and smoothing parameter
fit.sse.custom <- morthump(data = CHE1950f, model = "sse", x1 = 25, x2 = 40, lambda.sen = 1)
```

Figure 6 represents the fit of the observed age-specific death rates for Swiss females in 1950 by each model. The *HP* model systematically underestimates mortality between 30 and 60 years of age, which is not the case of the *SSE* model. They however estimate similar humps, centered around age 20, although it is not obvious from the observed rates that the underlying force of mortality contains a hump.

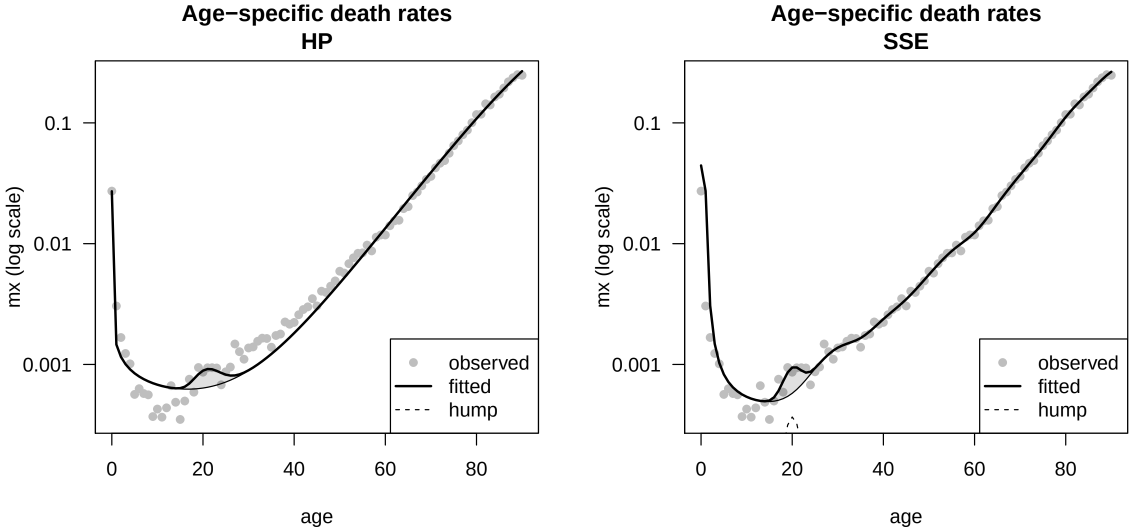

Figure 6: Fitted *HP* (left) and *SSE* (right) models on Swiss females in 1950, using custom parameters

2.3.1 Magnitude

Magnitude refers to the overall size of the hump component and not only its height. We suggest three measures of magnitude which have straightforward interpretations. They can all be computed using the `summary` method of `morthump` objects.

LEL

The life expectancy at birth lost due to the hump (LEL) is the difference between the life expectancy computed on the fitted force of mortality ($\widehat{e}(0)$), and life expectancy computed on the hump-free force of mortality ($\widehat{e^{-H}}(0)$). Algebraically,

$$LEL = \widehat{e^{-H}}(0) - \widehat{e}(0) = \int_0^\omega \mathrm{e}^{-\int_0^a \gamma_C(y)+\gamma_S(y)dy} da - \int_0^\omega \mathrm{e}^{-\int_0^a \gamma_C(y)+\gamma_H(y)+\gamma_S(y)dy} da \tag{5}$$

where γ_C, γ_H and γ_S, defined in Equation 1, are the childhood, hump and senescence components, respectively. These components can be estimated with a parametric or non-paremetric model.

This difference can be interpreted as the mean years of life lost in the population due to the presence of the hump, or alternatively as the potential gain in life expectancy that could be reached in the absence of the hump. In the case of Swiss females in 1950, LEL amounts only to 0.09 years of life according to both parametric and non-parametric models.

```
# life expectancy lost to the hump
summary(fit.hp.custom)$loss

## [1] 0.09

summary(fit.sse.custom)$loss

## [1] 0.09
```

In other words, were the hump not present, Swiss females would have lived about one extra month on average. This should be compared with the uncertainty of the observed period life expectancy at birth (about 70 years in 1950), which depends on the amount of stochasticity due to the size of the population exposure and the level of mortality. In this case, the 95% confidence interval estimated either analytically (Chiang, 1978) or numerically (Andreev and Shkolnikov, 2010) reaches a width of 0.24 year. The LEL of 0.09 due to the hump falls thus within the confidence interval of the observed life expectancy and does not stand out from stochastic noise. One can conclude from this that the estimated hump does not significantly lower life expectancy.

```
# confidence intervals
ci.hp <- confint(fit.hp.custom, method = "chiang")

## Loss of life expectancy due to the hump (years): 0.09
## Half-confidence interval of fitted life expectancy at birth: 0.102
## The hump is not statistically significant
## Significance level: 0.95

ci.sse <- confint(fit.sse.custom, method = "chiang")

## Loss of life expectancy due to the hump (years): 0.09
## Half-confidence interval of fitted life expectancy at birth: 0.121
## The hump is not statistically significant
## Significance level: 0.95
```

In the case of a parametric model, the statistical significance of the hump can be further tested by comparing the fit of a model that contains a hump component, with a model that does not. For instance, using the same population, it is possible to fit a *HP* model and compare its goodness-of-fit with that of a *HPS* model. This can be done with a F-test of nested models, since *HPS* is a special case of *HP*. In this case, this test indicates a p-value of 1, which means that there is about 100% chance to be wrong in claiming that the *HP* model (with a hump) outperforms the *HPS* model (without a hump). This confirms the conclusion reached with the confidence intervals of the fitted life expectancy.

```
# p-value
summary(fit.hp.custom)$pval

## [1] 1
```

YLL

Another measure of magnitude is an adaptation from the *years of life lost* (YLL), which is a well-known measure of premature mortality, defined by the World Health Organization as the product of the age-specific distribution of death with age-specific life expectancy (WHO, 2013). In this case, we only consider the deaths that are generated by the hump to compute the years of life lost due to the hump (YLL_H).

$$YLL_H = \sum_x \gamma_H(x) \cdot n_x \cdot e(x) \tag{6}$$

YLL_H is defined as the product of the hump component of the force of mortality (γ_H) and the age-specific exposure (n_x), which are then multiplied again with the remaining life expectancy ($e(x)$) and summed over age in order to obtain the total number of years that could have been lived by those who died because of the hump. In the case of Swiss females in 1950, this measure amounts to 3309 years of life according to the *SSE* model (3483 according to the *HP* model).

```
summary(fit.hp.custom)$yll

##      [,1]
## [1,] 3483

summary(fit.sse.custom)$yll

##      [,1]
## [1,] 3309
```

Deaths

The third measure of magnitude is the absolute number of deaths that would have been averted in the absence of the hump. It is obtained by multiplying the hump component of the force of mortality (γ_H) by the exposures (n_x):

$$d_H = \sum_x \gamma_H(x) \cdot n_x \tag{7}$$

This measure is probably the simplest to grasp and is thus easy to communicate to a non-scientific audience, but is not advisable if the goal is to compare populations. Indeed, the number of people who die because of the hump not only depends on the shape of the force of mortality, but also on the population at risk. In two populations with the same hump, the one with the larger population in the age range that contains the hump will experience more deaths. Note that this limitation also applies to the years of life lost to the hump. In the case of Swiss females in 1950, 62 young women would have been "saved" if the hump had been suppressed according to the *SSE* model (66 according to the *HP* model).

```
summary(fit.hp.custom)$d

## [1] 66

summary(fit.sse.custom)$d

## [1] 62
```

2.3.2 Location

Measures of location are inspired by classical measures of centrality that can be applied to any distribution. We treat the hump components as a density for such measures by rescaling the hump to sum to 1.

Mode

The mode of the hump is the most straightforward measure of location. This corresponds to the age at which the hump component reaches its peak.

$$Mode_H = \underset{x}{\text{argmax}} \quad \gamma_H(x) \tag{8}$$

In the case of parametric models (*HP* and *HPK*), $Mode_H = F$ by definition. With the *SSE* model, this quantity can be easily computed by predicting values for non-integer ages. In the case of Swiss females in 1950, the modal age at death was 21 according to the *HP* model, against 20.05 according to the *SSE* model. In this case parametric and non-parametric agree thus more or less on the age at which the hump reaches its maximum strength.

```
summary(fit.hp.custom)$mode

## [1] 20.99999

summary(fit.sse.custom)$mode

## [1] 20.04776
```

Mean

Another measure of location is the mean age at death for those who died because of the hump. It is the weighted mean of the age, weighted by the hump component.

$$Mean_H = \frac{\int \gamma_H(x) \cdot x}{\int \gamma_H(x)} \tag{9}$$

This mean computes to 20.05 and 21.5 years of age for the *SSE* and *HP* models respectively. The difference between the two models is slightly larger on this measure because, unlike the mode, it is sensitive to the shape of the tails of the hump. An asymetrical hump that extends further into older ages may have a higher mean age at death but not necessarily a higher mode.

```
summary(fit.hp.custom)$mean

## [1] 21.53162

summary(fit.sse.custom)$mean

## [1] 20.04734
```

Median

A last measure of location is the median age at death from the hump. As its name suggests, it is the age at which half of the people who died from the hump have done so.

$$Median_H = x \mid \frac{\int_0^x \gamma_H(a)da}{\int_0^\omega \gamma_H(a)da} \geq 0.5 \quad \text{and} \quad \frac{\int_x^\omega \gamma_H(a)da}{\int_0^\omega \gamma_H(a)da} \geq 0.5 \tag{10}$$

This measure computes to 20.06 and 21.35 years of age for the *SSE* and *HP* models respectively.

```
summary(fit.hp.custom)$median

## [1] 21.35294

summary(fit.sse.custom)$median

## [1] 20.06337
```

The combination of these three measures of location (or centrality) is useful in determining the overall shape of the hump. Indeed, as this example demonstrates (albeit to a relatively small extent), the mean is more sensitive to extreme values than the median and the mode. If these three measures are close this is thus a sign that the hump is symmetrical, while if the mode and median are located before the mean this is a sign of a longer right tail (positive skew), and inversely if they are located after the mean this suggests a longer left tail (negative skew). These three measures of location should thus be ideally interpreted together.

2.3.3 Spread

Measures of spread are inspired of the classical measures of dispersion that can be applied to any distribution. By treating the hump component as a density, these measures provide alternative perspective on the age range affected by the hump. The spread (or dispersion) of the hump is also an explicit parameter in the *HP* and *HPK* models. Indeed, in both cases, the E parameter controls the concentration of the hump around the peak. Additionally, in the *HPK* model this value can be different before and after the peak depending on the value of the k parameter. The E parameter is thus inversely correlated with the spread of the hump, but its units are not easily interpretable.

Standard deviation

Unlike the E parameter, the standard deviation of the age at death from the hump comes in years units. This quantity can be easily computed from the density of the hump with standard formula.

$$sd_H = \sqrt{\frac{\int \gamma_H(x) \cdot (x - Mean_x)^2}{\int \gamma_H(x)}} \tag{11}$$

In the case of Swiss females in 1950, the standard deviation of the age at death from the hump amounts to 1.9 years according to the *SSE* model, against 2.8 years according to the *HP* model.

```
summary(fit.hp.custom)$sd

## [1] 2.791342

summary(fit.sse.custom)$sd

## [1] 1.888632
```

Quantile

Other measures of spread can be computed using the ready-to-use functions generated by the `summary` method. In particular, the quantile function ($qtl(x)$) can be used to compute any quantile of the hump. For instance, one can compute the interquartile range (IQR), which for Swiss females in 1950 ranges from 19.6 to 23.3 years according to the *HP* model, and from 18.7 to 21.3 years according to the *SSE* model.

```
qtl <- summary(fit.hp.custom)$qtl
qtl(c(0.25, 0.75))  # IQR

## [1] 19.57224 23.29560

qtl <- summary(fit.sse.custom)$qtl
qtl(c(0.25, 0.75))  # IQR

## [1] 18.70538 21.32667
```

3 Cause-specific mortality

In recent decades, the hump has been almost exclusively associated in the literature with accidents, as reflected by the widespread use of the term "accident hump", likely coined by Heligman and Pollard

(1980). However this hypothesis has never been tested properly due to methodological limitations.

A solution to this problem was proposed by Remund et al. (2017). It is based on the same premises as the all-cause "sse" model, but generalizes it to decompose the hump into contributions from each cause of death. Common age-cause decompositions of mortality differences such as those proposed by Arriaga (1984), Pollard (1982), Andreev (1982) and Pressat (1985) do not isolate the hump. The MortHump package implements this model and allows (1) formatting the data, (2) generating a cause-of-death typology, (3) estimating the model and (4) measuring the cause- and age-specific contributions to the young adult mortality hump.

3.1 Format data

The MortHump package includes a data grabber for the Human Cause-of-Death Database (HCD, 2017), which offers cause-of-death series reconstructed across ICD transitions (Meslé and Vallin, 1996). It currently only works locally and thus requires previously downloading the data from the website. All the data must be located in country-specific folders named after the country short names, ideally directly taken from the zipped files available on the HCD website. The function HCD2MH() works along the same principles as the HMD2MH() function, but has slightly different options.

```
# For local access, replace "path" with the file path to the HCD data.
# To register for the HCD and download the data, go to www.causesofdeath.org
#
# US males in 2000 (intermediate list with 101 causes, abridged data)
 USA2000m <- HCD2MH(country = "USA", year = 2000, sex = "males",
                    unabr = FALSE, list = "interm", path = path)

 str(USA2000m, max.level = 1)

## List of 7
##  $ mxc  :'data.frame': 22 obs. of  101 variables:
##  $ dxc  :'data.frame': 22 obs. of  101 variables:
##  $ nx   : num [1:22] 2005790 7825242 10461439 10556762 10438509 ...
##  $ x    : num [1:22] 0.5 3 7.5 12.5 17.5 22.5 27.5 32.5 37.5 42.5 ...
##  $ age  : chr [1:22] "0" "1-4" "5-9" "10-14" ...
##  $ inter: num [1:22, 1:2] 0 1 5 10 15 20 25 30 35 40 ...
##  $ lab  :'data.frame': 101 obs. of  3 variables:

# US males in 2000 (short list with 16 causes, unabridged data)
 USA2000m <- HCD2MH(country = "USA", year = 2000, sex = "males",
                    unabr = TRUE, list = "short", path = path)

 str(USA2000m, max.level = 1)

## List of 7
##  $ mxc  :'data.frame': 110 obs. of  16 variables:
##  $ dxc  :'data.frame': 110 obs. of  16 variables:
##  $ nx   : num [1:110] 2005790 1965719 1940427 1943665 1975431 ...
##  $ x    : num [1:110] 0.5 1.5 2.5 3.5 4.5 5.5 6.5 7.5 8.5 9.5 ...
##  $ age  : int [1:110] 0 1 2 3 4 5 6 7 8 9 ...
##  $ inter: num [1:109, 1:2] 0.5 1.5 2.5 3.5 4.5 5.5 6.5 7.5 8.5 9.5 ...
##  $ lab  :'data.frame': 16 obs. of  3 variables:
```

HCD data are available in three different lists of causes of death: short, intermediate and full. The short list contains 16 causes of death that define broad families of causes (e.g. *neoplasms*, *heart diseases*, or *external causes*). The intermediate list contains 104 causes of death that provide a finer display of the etiological processes (e.g. *malignant neoplasm of stomach*, *pulmonary heart diseases*, or *suicide and self-inflicted injury*). The full list depends on the original data but is as close as possible to the 4-digit codes of the 10th revision of the ICD. Depending on the desired level of analysis, the type of list can be easily selected in the `HCD2MH()` function using the `list` argument, which takes three possible values: `"short"`, `"interm"` and `"full"`. In practice, for the study of the young adult mortality hump, it is probably advisable to work with the intermediate list.

Another option that is specific to the `HCD2MH()` function concerns the nature of the age groups. By default, all HCD data come into five-year age groups, except the first year of life. The open interval varies from 85+ to 100+ depending on the original data. These abridged data can be used directly in the other functions of the `MortHump` package, but they also can be *unabridged* thanks to the `unabr` option. When this argument is used (`unabr = TRUE`), a single-age dataset is produced by applying a monotonic spline to the cumulative distribution of the cause- and age-specific death counts. This method is inspired by the method protocol of the Human Mortality Database and is fully described in the documentation of the `unabridge()` function, together with explanations on how to produce diagnostic plots in order to check the coherence of the resulting unabridged data.

User-supplied datasets can also be loaded instead of the HCD data. In this case, the dataset must be structured in the same way as the output of the `HCD2MH()` function. Particularly, it must be a list containing data frames with the cause- and age-specific rates and death counts, as well as vectors for the age-specific exposures, mid-points, and labels, intervals for each age groups and a data frame containing the label for each cause of death (short and long).

3.2 Group causes of death

The first step in the decomposition of cause- and age-specific contributions to the young adult mortality hump is to identify the causes that are likely to contribute to the hump (Remund et al., 2017). This task is unavoidable because there is no way to fully automatize this selection. This selection is not an apriori estimation of the respective weight of each cause, and different typologies can be tested in sensitivity analyses. As a general rule of thumb, it is advisable to select between 2 and 6 causes of death (or groups of causes) that are susceptible to contribute to the hump. This task can be done manually, by creating a list containing the desired typology. The user needs to define this by creating a list of which each element is a vector containing the column index of the desired causes in the `mxc` and `dxc` data frames which are stored within the data object generated by the `HCD2MH()` function. For instance, using the dataset of US males in 2000, one can define five user-defined groups of causes in the following way. Note that in this typology, causes 93 (Accidental poisoning by alcohol) and 94 (Accidental poisoning by alcohol) are aggregated to form a single cause of death labeled as poisoning (poi). Likewise, causes 95 (Other accidental threats to breathing), 98 (Event of undetermined intent) and 100 (Other accidents and late effects of accidents (remainder)) are grouped into a common cause of death labeled as other accidents (oac).

```
# load data for US males in 2000 (HCD)
data(USA2000m)

# manual typology
typ <- list()
typ$tac <- 89
typ$sui <- 96
typ$hom <- 97
typ$poi <- c(93,94)
typ$oac <- c(95,98,100)

# display long labels
lapply(typ,function(x){USA2000m$lab$label[x]})

## $tac
## [1] "Transport accidents"
```

```
##
## $sui
## [1] "Suicide and self inflicted injury"
##
## $hom
## [1] "Assault"
##
## $poi
## [1] "Accidental poisoning by alcohol"
## [2] "Accidental poisoning by other substance"
##
## $oac
## [1] "Other accidental threats to breathing"
## [2] "Event of undetermined intent"
## [3] "Other accidents and late effects of accidents (remainder)"
```

Alternatively, if one wants to adopt a more inductive approach, data mining methods are available in the `codgroup()` function. They consist essentially in computing the first difference of the cause- and age-specific death rates, and applying Principal Component Analysis (PCA) and Hierarchical Cluster Analysis (HCA) to study the differences in age-shape between causes. Keeping the US males in 2000 as a working example, one can see that the `codgroup()` function, applied to the age range 10 to 29 and using a six-group cluster solution (`k = 6`, including the group of all other causes that do not contribute to the hump) suggests the following typology.

```
# load data for US males in 2000 (HCD)
data(USA2000m)

# Find grouping automatically
groups <- codgroup(USA2000m, k = 6, x.range = 10:29)

# store the automatic typology
typ <- groups$typ

# display long labels
lapply(typ,function(x){USA2000m$lab$label[x]})

## $B
## [1] "HIV disease"
##
## $C
## [1] "Transport accidents"
##
## $D
## [1] "Accidental poisoning by other substance"
##
## $E
## [1] "Suicide and self inflicted injury"
##
## $F
## [1] "Assault"

# rename groups
names(typ) <- c("hiv","tac","poi","sui","hom")
```

```
# display short labels
typ

## $hiv
## INF_hiv
##       6
##
## $tac
## EXT_transp
##         89
##
## $poi
## EXT_poison.other
##               94
##
## $sui
## EXT_suicide
##          96
##
## $hom
## EXT_assault
##          97
```

Several diagnostic plots are available in order to visualize distance between each cause and cluster of causes (Figure 7). This typology isolates all the causes that stand out from the general trend in the force of mortality between 10 and 29 years of age: HIV, traffic accidents, poisonings other than alcohol (i.e. drug overdoses), suicides, and homicides. The projection of these causes on the first two dimensions of a Principal Component Analysis (Figure 7, left) shows that these five causes have very different shapes than the other ones during early adulthood. The first two dimensions of the PCA capture 96% of the between-cause variation. The comparison of the first difference of the force of mortality by cluster of causes (Figure 7, right) shows that the proposed clusters distinguish causes that experience a rapid increase before 20 years of age followed by a decrease (transports accidents, and to a lesser extent suicides and homicides), from causes that are progressively accelerating (non-alcoholic poisonings and HIV) compared to the general trend (cluster 1). In other words, the former probably have narrower contributions to the hump than the latter, but all of these causes deviate from the general trend.

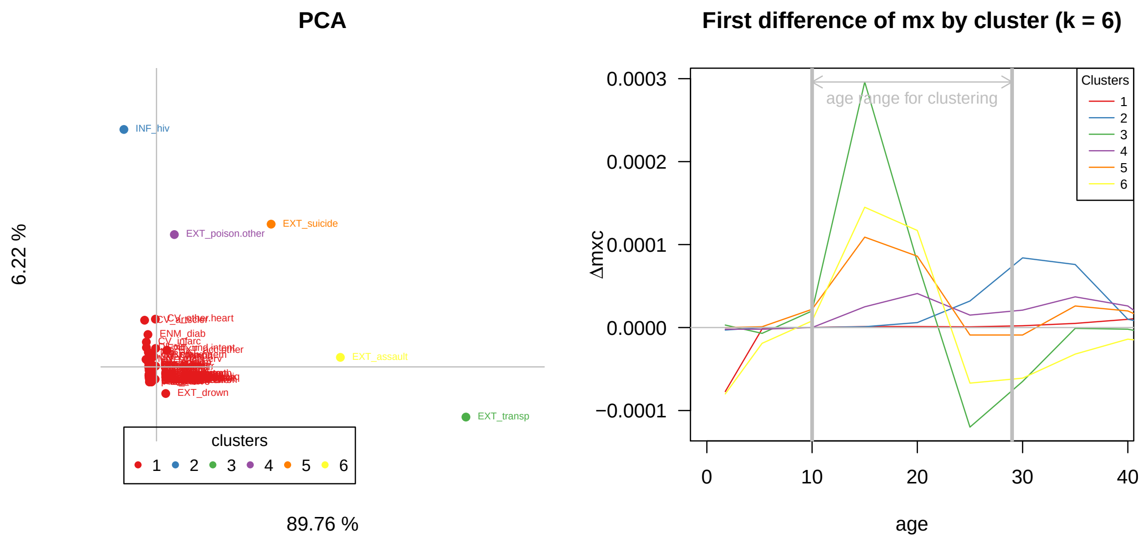

Figure 7: Constructing a cause-of-death typology with automatized tools

Other tools can be used to help chosing the typology, using measures of cluster quality to pick the optimal size of the typology (see documentation of `codgroup()` and `plot.codgroup()`). The general aim is to obtain (1) a remaining group that does not display any hump, and (2) as many groups of causes of death as necessary to obtain a satisfactory level of detail in the analysis.

3.3 Estimate models

The modelling of the cause-of-death decomposition of the hump is described in detail in Remund et al. (2017). It consists first in fitting an *SSE* model to the overall force of mortality from age 10 (or the lowest oberved death rate), only keeping the hump and senescence components in order to simplify the computation. Then the model is refit with cause deleted and replaced in sucession. The resulting perturbations in the hump component after deletion of a cause is interpreted as the contribution of this cause to the hump. The estimation is done simultaneously on all the causes as the sum of all cause-specific contributions to the hump needs to equal the overall hump estimated in the first step. In order to ensure this, a constrained optimization algorithm is used.

The model returns a set of values for the contribution of each cause (κ) to each component (γ_H for the young adult mortalit hump and γ_S for senescence).

$$m_x = \mu(x) + \epsilon_x = \gamma_H(x,\theta_H) + \gamma_S(x,\theta_S) + \epsilon_x = \sum_{\kappa} \gamma_H^{\kappa}(x,\theta_H^{\kappa}) + \sum_{\kappa} \gamma_S^{\kappa}(x,\theta_S^{\kappa}) + \epsilon_x, \quad \forall x \geq 10 \quad (12)$$

In the `MortHump` package, this model can be estimated with the `codhump()` function, which takes as necessary arguments `data`, a list typically resulting from a call to `HCD2MH()`, and `typ`, a list that describes the structure of the cause-of-death typology (see previous section). All the other arguments have default values that help the algorithm to converge, but can be modified if necessary. For instance, the

```
# load data for US males in 2000 (HCD)
data(USA2000m)

# automatically generate and rename a typology
groups <- codgroup(USA2000m, k = 6, x.range = 10:29)
```

```
typ <- groups$typ[c(2:5, 1)]
names(typ) <- c("tac", "poi", "sui", "hom", "hiv")

# fit the model
fit.full <- codhump(data = USA2000m, typ = typ)

# remove HIV from the contributing causes
typ <- typ[-5]
fit.nohiv <- codhump(data = USA2000m, typ = typ)
```

Applying this approach on US males in 2000, and using the default typology suggested by the inductive approach, we obtain a decomposition that shows no significant contribution from HIV (Figure 8, left)[6]. This can be explained by the fact that the force of mortality for this cause of death only deviates from the overall force of mortality after the mid-twenties (Figure 7), whereas the overall hump peaks around 20 years of age. Consequently, although HIV deviates from the general trend, it does so in a way that does not contribute to the hump. Incidentally, the quality of the fit is also affected as we can see that the overall hump (black line) does not overlap perfectly with the stacked cause-specific contributions. One way to deal with this issue is to remove HIV from the contributing causes. Doing this allows the model to converge more easily and the sum of all cause-specific contributions to overlap with the overall hump (Figure 8, right).

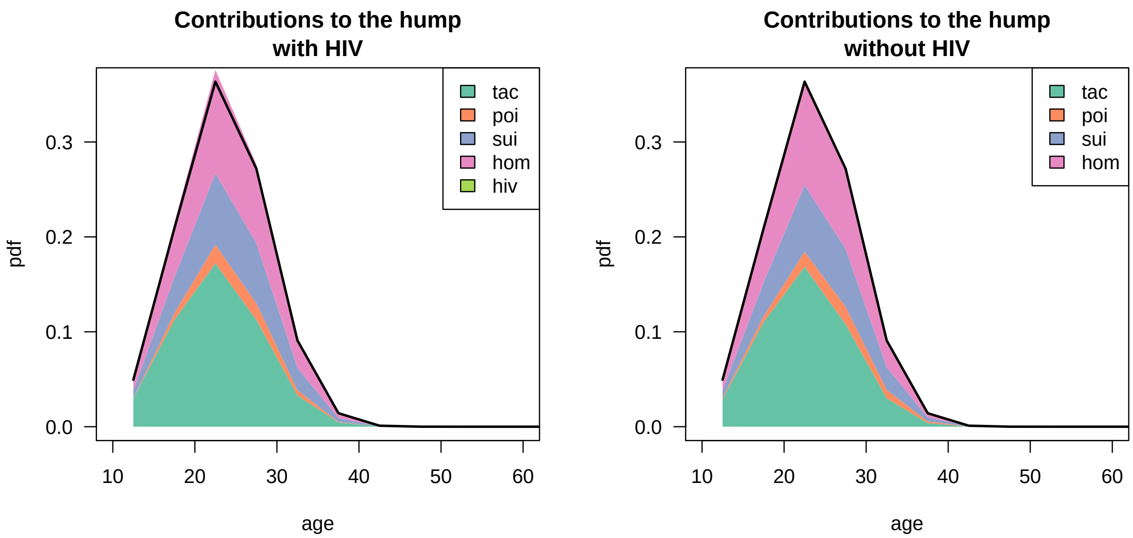

Figure 8: Estimating a cause-of-death decomposition of the hump on US males in 2000, with (left) and without (right) HIV

3.4 Measure the hump

The measures developed for the all-cause hump can be generalized to cause-of-death contributions. Here again, we distinguish measures of intensity (magnitude), location (centrality) and spread (dispersion). They are defined by applying the same formulas that for all-cause mortality, this time on the cause-specific contribution to the hump (γ_H^κ). We refer thus to the previous chapter for the formulas.

[6] Other diagnostic plots are available, notably to make sure that the convergence happened as expected. One possible caveat is that the algorithm gets stuck around a local optimum and starts drifting away from a reasonable solution. Safeguards were designed to stop the algorithm in this case, but they may prevent it from reaching a satisfactory solution. This kind of issue can be identified with the help of the other diagnostic plots (see the documentation for the plot method of `codhump()`).

We only present a selection of these measures, keeping only, for *magnitude*, years of life expectancy lost to the hump (Equation 5), for *location*, mode (Equation 8) and mean age at death from the hump (Equation 9), and for *spread*, standard deviation of age at death from the hump (Equation 11). They can all be computed using the `summary` method of the `codhump` objects in the following way.

In the case of US males in 2000, the loss in life expectancy due to the hump amounts to 0.62 years, of which about 45% (0.28 years) is due to traffic accidents, 5% (0.03 years) to poisonings, 20% (0.12 years) to suicides and 30% (0.18 years) to homicides. The sum of these three contributions equals the life expectancy lost to the overall hump.

```
# life expectancy lost to the hump, total and by cause of death
summary(fit.nohiv)$loss

##  all  tac  poi  sui  hom
## 0.62 0.28 0.03 0.12 0.18
```

These values do not match the corresponding share in the absolute death couts for these causes between ages 10 and 35. Traffic accidents is the only cause for which the two measures are more or less equal (about 25%), but all the other causes have much larger contributions to the hump than their share in the absolute number of deaths. This is because the model isolates the contributions to the hump, which is stronger for these causes, and thus increases their weight compared to the observed deaths. Other causes account for about 40% of the original death counts in this age range, but the shape of their force of mortality is close enough to the overall trend that they do not generate any deviation.

Regarding measures of location, in the case of US males in 2000, the mode of cause-specific contributions to the hump coincide with the overall hump at 22.5 years of age. The mean also gives similar results across causes (22.6 to 24.3), which suggests that, at least in this case, the underlying forces that make young adult more vulnerable to specific causes of death during their transition to adulthood follow similar timings. This assertion should however be tested on single-age data instead of abridged data.

```
# mode of the hump, total and by cause of death
summary(fit.nohiv)$mode

##  all  tac  poi  sui  hom
## 22.5 22.5 27.5 22.5 22.5

# mean age at death from the hump, total and by cause of death
summary(fit.nohiv)$mean

##   all   tac   poi   sui   hom
## 23.47 22.64 25.60 24.26 23.79
```

Concluding with measures of spread, in the case of US males in 2000 the standard deviation of age at death from the hump also indicates similar values for all causes (5.23 to 5.52). Let us note here that depending on the age-specific shape of the cause-of-death contributions, the overall spread may be smaller, equal or larger than any of the cause-specific contributions. For instance, each cause-specific contributions can be narrow but centered on different ages, which would generate a wide overall hump. Inversely, several wide cause-specific contributions centered on the same age may generate a narrower overall hump.

```
# standard deviation of the age at death from the hump, total and by cause of death
summary(fit.nohiv)$sd

##                tac      poi      sui      hom
## 5.364672 5.231438 5.517979 5.467718 5.236948
```

4 Conclusion

The young adult mortality hump is probably the least studied of the three main components of the force of mortality. The terms used to speak of this phenomenon are even themselves debatable, as many publications use the expression "accident hump", even though there is evidence that accidents are not the only cause of death contributing to the hump, and maybe not even the largest one. More importantly perhaps, there is no commonly accepted measure of the hump that is really based on the definition of a deviation in the force of mortality. Consequently, theories about the source of this phenomenon remain fuzzy and have not been thoroughly tested.

The `MortHump` package is conceived as a methodological toolbox to fill this gap, by offering a user-friendly, adaptable, and open-source solution for research on the hump. It includes functions to format data, estimate models, generate diagnostic plots and compute summary measures in a simple straightforward fashion. Our hope is that it will help open a new line of research on the comparative and historical study of the young adult mortality hump, ultimately allowingcompeting and complementary theories on the forces that drive this phenomenon to be tested.

Acknowledgements

This work as been made possible by the generous funding of the MODICOD (Modicod: "Le projet AXA Mortality Divergence and Causes of Death") and DIMOCHA (Dimocha: "Project ANR-12-FRAL-0003-01 DIMOCHA") projects, as well as an *Early Postdoc.Mobility* scholarship from the Swiss National Science Foundation and the support of the NCCR LIVES - overcoming vulnerability: life course perspectives.

References

E.M. Andreev. Metod komponent v analize prodoljitelnosty zjizni. [the method of components in the analysis of length of life]. *Vestnik Statistiki*, 9:42–47, 1982.

Evgeny M Andreev and Vladimir M Shkolnikov. Spreadsheet for calculation of confidence limits for any life table or healthy-life table quantity. *Rostock: Max Planck Institute for Demographic Research (MPIDR Technical Report*, 5, 2010.

Eduardo E. Arriaga. Measuring and explaining the change in life expectancies. *Demography*, 21(1): 83–96, 1984.

David R. Brillinger. A biometrics invited paper with discussion: The natural variability of vital rates and associated statistics. *Biometrics*, 42(4):693–734, 1986.

Carlo G. Camarda. *Smoothing methods for the analysis of mortality development.* PhD thesis, Universidad Carlos III de Madrid. Departamento de Estadística, 2008. URL `http://e-archivo.uc3m.es/handle/10016/5133`.

Carlo G Camarda, Paul HC Eilers, and Jutta Gampe. Sums of smooth exponentials to decompose complex series of counts. *Statistical Modelling*, 16(4):279–296, 2016.

Chin Long Chiang. *Life table and mortality analysis.* Geneva : World Health Organization, division of health statistics, 1978.

Petros Dellaportas, Adrian F. M. Smith, and Photis Stavropoulos. Bayesian analysis of mortality data. *Journal of the Royal Statistical Society: Series A*, 164:275–291, 2001.

Paul HC Eilers. Ill-posed problems with counts, the composite link model and penalized likelihood. *Statistical Modelling*, 7(3):239–254, 2007.

Joshua Goldstein. A secular trend toward earlier male sexual maturity: Evidence from shifting ages of male young adult mortality. *PLoS ONE*, 6(8), 2011.

Benjamin Gompertz. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. *Philosophical Transactions of the Royal Society of London*, 115:513–583, 1825.

HCD. Human cause-of-death database. french institute for demographic studies (france) and max planck institute for demographic research (germany). available at www.causeofdeath.org., 2017.

L. Heligman and J. H. Pollard. The age pattern of mortality. *Journal of the Institute of Actuaries*, 107, 1980.

HMD. Human mortality database. university of california, berkeley (usa), and max planck institute for demographic research (germany). available at www.mortality.org or www.humanmortality.de.

Shiro Horiuchi and John R Wilmoth. Deceleration in the age pattern of mortality at older ages. *Demography*, 35(4):391–412, 1998.

Rob J Hyndman and Md Shahid Ullah. Robust forecasting of mortality and fertility rates: a functional data approach. *Computational Statistics & Data Analysis*, 51(10):4942–4956, 2007.

Anastasia Kostaki. A nine-parameter version of the heligman-pollard formula. *Mathematical Population Studies*, 3(4):277–288, 1992.

Daniel A. Levitis. Before senescence: the evolutionary demography of ontogenesis. *Proceedings of the Royal Society B: Biological Sciences*, 278(1707):801–809, 2011.

France Meslé and Jacques Vallin. Reconstructing long-term series of causes of death: the case of france. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 29(2):72–87, 1996.

Charles Mode and Robert Busby. An eight-parameter model of human mortality - the single decrement case. *Bulletin of Mathematical Biology*, 44(5):647–659, 1982.

Marius D. Pascariu. *MortalityLaws: Parametric Mortality Models, Life Tables and HMD*, 2017. URL `https://CRAN.R-project.org/package=MortalityLaws`. R package version 1.0.5.

J. H. Pollard. The expectation of life and its relationship to mortality. *Journal of the Institute of Actuaries*, 109(2):225–240, 1982.

Roland Pressat. Contribution des écarts de mortalité par âge à la différence des vies moyennes. *Population*, 40(4/5):766–770, 1985.

Adrien Remund. Is young adults' excess mortality a universal phenomenon? Chaire Quetelet, Université catholique de Louvain-la-Neuve, 2012. URL `http://130.104.5.100/cps/ucl/doc/demo/documents/RemundCQ.pdf`.

Adrien Remund. *Jeunesses vulnérables? Mesures, composantes et causes de la surmortalité des jeunes adultes.* PhD thesis, Université de Genève, 2015.

Adrien Remund, Carlo G Camarda, and Timothy Riffe. A cause-of-death decomposition of the young adult mortality hump. *MPIDR Working paper*, 2017.

William Siler. A competing-risk model for animal mortality. *Ecology*, 60(4):750–757, 1979.

Thorvald Nicolai Thiele. On a mathematical formula to express the rate of mortality throughout the whole of life, tested by a series of observations made use of by the danish life insurance company of 1871. *Journal of the Institute of Actuaries and Assurance Magazine*, 16(5):313–329, 1871.

R Thompson and RJ Baker. Composite link functions in generalized linear models. *Applied Statistics*, pages 125–131, 1981.

James W Vaupel. Trajectories of mortality at advanced ages. *Between Zeus and the salmon: The biodemography of longevity*, pages 17–37, 1997.

WHO. Metrics: Disability-adjusted life year (daly), 2013.

Guillaume Wunsch, Michel Mouchart, and Josianne Duchêne. *The life table : modelling survival and death.* Kluwer Academic, Dordrecht ; London, 2002.