Modelling and decomposing vital rates a non-parametric approach

Carlo G. Camarda* Paul H.C. Eilers† Jutta Gampe‡

Extended abstract submitted to the
XXVII IUSSP International Population Conference, Busan, Republic of Korea
October 2012

1 Introduction

Demographic events, such as marriage, migration or death, have characteristic age specific patterns of occurrence. Finding so called model schedules to summarize the age pattern has a long tradition, however, parametric models are predominantly used. Among others see Heligman and Pollard (1980); Hoem et al. (1981); Rogers and Castro (1981); Rogers et al. (2005); Rogers and Little (1994); Romaniuk (1973); Siler (1983).

One feature of many demographic rates is that their overall shape is rather complex, but the pattern can be attributed to different distinct components. As an example, migration rates are high at very young ages (infants migrating with their parents), in the young adult ages (labour motivated migration), and again after retirement. Models that can separate these components are particularly welcome.

While some of the components can be described well by a parametric model, such as adult mortality by the Gompertz hazard, many others cannot. An additional complication arises if data are provided only in age groups, which is still the case in many official statistics, and is standard if one goes back in time or analyzes age specific disease incidence.

In the following we propose a general model that allows to specify a (demographic) rate across a wide range of ages as the sum of several components, which are modelled on the log scale and are assumed to be smooth, but do not have to follow a particular parametric form. If several nonparametric components are additively combined, shape constraints are necessary to identify the components correctly. The data can be given in grouped form, and the age groups can be of variable lengths. Furthermore the model can cope with two-dimensional settings in which age-patterns change over time. We will illustrate the performance of the proposed method by a series of common demographic datasets.

2 Sums of Smooth Exponentials

In this section we will briefly illustrate the proposed model in a uni-dimensional setting and emphasizing its demographic aspects. For a fuller statistical treatment we refer to Camarda et al. (2010, 2012).

Let $\boldsymbol{e}$ be the m-dimensional vector of exposures at the m ages considered, and $\boldsymbol{y}$ be the corresponding vector of numbers of counts, e.g. deaths, migrants, births. The actually observed counts are assumed to be realizations from a Poisson distribution, $\boldsymbol{y} \sim \mathcal{P}(\boldsymbol{\mu})$. The expected values $\boldsymbol{\mu}$ are the product of exposures $\boldsymbol{e}$ and the actual vital rates at the respective ages, which we denote by $\boldsymbol{\theta}$.

The rates are assumed to be the sum of K components $\boldsymbol{\gamma}^k$, $k = 1, \ldots, K$, each of length m. For easy of

*Corresponding author: Institut National d'Études Démographiques. 133, Bd Davout, 75980 Paris Cédex 20, France; Tel. +33(0)1 5606 2155, email: carlo-giovanni.camarda@ined.fr
†Department of Biostatistics, Erasmus Medical Centre, Rotterdam, The Netherlands; email: p.eilers@erasmusmc.nl
‡Max Planck Institute for Demographic Research, Rostock, Germany; email: gampe@demogr.mpg.de

presentation we set $K = 3$. The vital rates are then equal to

$$\boldsymbol{\theta} = \begin{bmatrix} \gamma_1^1 + \gamma_1^2 + \gamma_1^3 \\ \vdots \\ \gamma_m^1 + \gamma_m^2 + \gamma_m^3 \end{bmatrix} = \underbrace{\begin{bmatrix} \gamma_1^1 & \gamma_1^2 & \gamma_1^3 \\ \vdots & \vdots & \vdots \\ \gamma_m^1 & \gamma_m^2 & \gamma_m^3 \end{bmatrix}}_{\boldsymbol{\Gamma}} \cdot \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \boldsymbol{\Gamma} \cdot \mathbf{1} \,. \tag{1}$$

If $\boldsymbol{\gamma} = \mathtt{vec}(\boldsymbol{\Gamma})$ is the vector of length mK, then we can write the expected values $\boldsymbol{\mu}$ as follows

$$\boldsymbol{\mu} = \boldsymbol{C}_0 \boldsymbol{\gamma} = \underbrace{\left[\begin{array}{ccc|ccc|ccc} e_1 & 0 & \cdots & e_1 & 0 & \cdots & e_1 & 0 & \cdots \\ 0 & \ddots & \cdots & 0 & \ddots & \cdots & 0 & \ddots & \cdots \\ \cdots & \cdots & e_m & \cdots & \cdots & e_m & \cdots & \cdots & e_m \end{array}\right]}_{\boldsymbol{C}_0 = \mathbf{1}_{1,3} \otimes \mathtt{diag}(\boldsymbol{e})} \cdot \begin{bmatrix} \boldsymbol{\gamma}^1 \\ \boldsymbol{\gamma}^2 \\ \boldsymbol{\gamma}^3 \end{bmatrix} \tag{2}$$

In other words, the composition matrix $\boldsymbol{C}_0$ additively combines the $\boldsymbol{\gamma}^k$ and simultaneously matches the exposures.

If the observed counts $\boldsymbol{y}$ are given only for age groups but the exposures are available by single years of age, which is not uncommon, then the composition matrix can be adapted to represent also the additional grouping. For n age classes, which can be of different widths, the new composition matrix $\boldsymbol{C}$ has the following form:

$$\boldsymbol{C} = \boldsymbol{E} \cdot \boldsymbol{C}_0 \,. \tag{3}$$

The elements η_{ji} of $\boldsymbol{E}$ are equal to 1, if age i is contained in age group j, and zero otherwise.

If the exposures are also grouped, then the methodology presented in Lambert and Eilers (2009) is used in a first step to ungroup the exposure numbers. The single age exposures obtained by this approach are then used in equation (3).

In both single-ages and grouped-ages settings, we manage to express the expected values $\boldsymbol{\mu}$ as a linear combination of a composition matrix ($\boldsymbol{C}_0$ and $\boldsymbol{C}$) and the vector $\boldsymbol{\gamma}$. This allows us to embed our model in a composite link model framework (Thompson and Baker, 1981).

Each component can be described by parametric or non-parametric structures, as needed. In this way, the composed mean $\boldsymbol{\mu}$ can be viewed as sum of K exponential components, which generally are smooth. Hence we call this a Sum of Smooth Exponentials (SSE) model.

The K coefficient vectors $\boldsymbol{\beta}^k$ associated to each component can be estimated by a penalized iteratively re-weighted least squares (IRWLS) algorithm (Eilers, 2007) and the matrix $\boldsymbol{P}$ combines the penalty matrices $\boldsymbol{P}^k$ for the K components. The specific form of $\boldsymbol{P}^k$ depends on the assumptions we make on the different components $\boldsymbol{\gamma}^k$. In practice, such assumptions often come naturally. For example, infant mortality is supposed to quickly drop from high values to zero, while the "accident hump" at young adult ages, as the name suggests, is supposed to have a unimodal, log-concave shape. Old-age mortality is a strictly increasing function of age. The component specific penalty can be devised to combine different assumptions as well as parametric models, e.g. a Gompertz curve.

3 Applications

We now illustrate the performance of this model by four examples. Mortality and fertility data are taken from Human Mortality Database (2012) and Human Fertility Database (2012), respectively. Data for the migration example come from the Eurostat web-site.

3.1 Mortality Data in 1D by single year of age

Figure 1 shows the results obtained from the SSE model on mortality of males in Italy in 2008, ages 0-105. Death counts and population are given by single year of age. We use a combination of a parametric (hyperbolic) function for child mortality, a smooth and monotonically increasing ageing related mortality and smooth and log-concave component related to the accident-hump.

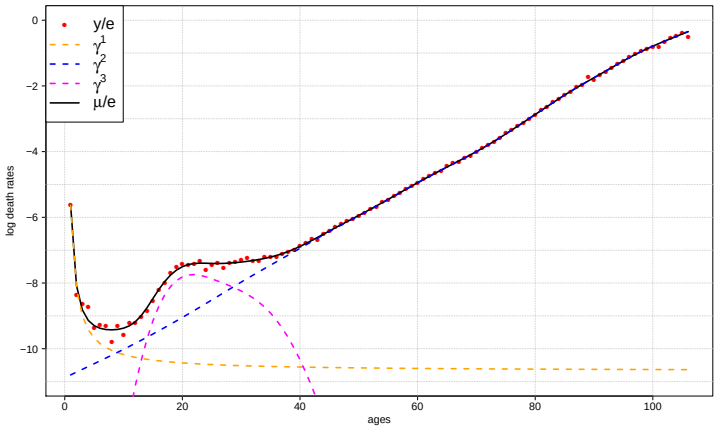

Figure 1: Death rates, log-scale. Italy, males, 2008, ages 0-105. Three smooth components are additively combined to give the overall trajectory.

3.2 Mortality Data in 2D by single year of age

The second dataset is a two-dimensional generalization of the first example. We analyze how age pattern changes over time for the men in England and Wales between 1970 and 1990, ages 0-105. Figure 2 presents the actual and fitted values over ages for selected years. For the infant mortality component we employed a series of hyperbolic functions. A two-dimensional penalized surface is used for describing the ageing mortality pattern and monotonicity is enforced over ages. The development of the accident hump over the period is modelled by a smooth surface which is log-concave over ages.

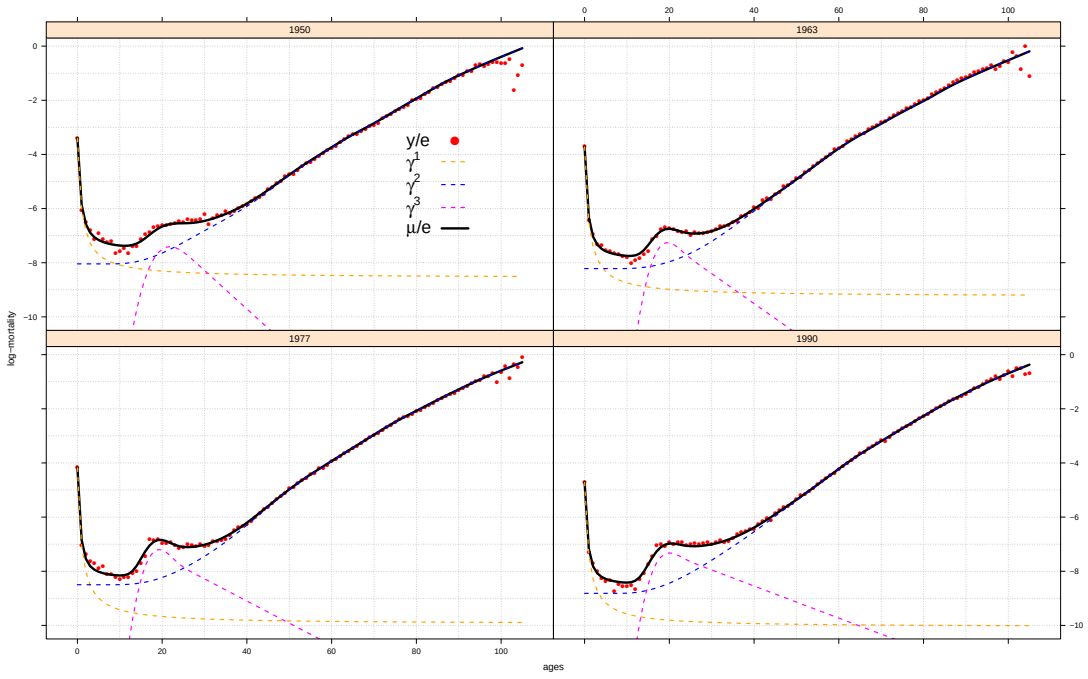

Figure 2: Death rates, log-scale. England and Wales, males, 1950-1990, ages 0-105. Three smooth components are additively combined and smoothly changing over time.

3.3 Migration Data in 1D with age-groups

In this example we look at the age specific migration rates from Germany to Spain for 2004-2008, see Figure 3. The migration flows are given in five year age intervals, the population counts are available by single year of age. The overall rate was decomposed into four (non-)parametric components. The first one was assumed to be monotone decreasing and relates to migration of children, which is mirrored in the relatively high rates

of their parents. Two other components were penalized to be log-concave and pertain to job and retirement related migration. A fourth component describes an age independent migration propensity.

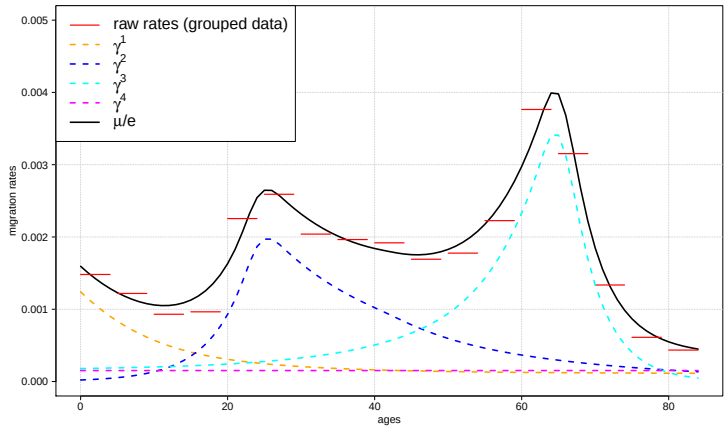

Figure 3: Migration rates from Germany to Spain for the years 2004-2008 and ages under 85. The step function shows the raw rates for 5 year age groups. Four (smooth) components are additively combined to give the overall trajectory.

3.4 Fertility Data in 2D with changing age-groups

The fertility example is employed for assessing the performance of the model in a two-dimensional setting where the length of the age groups varies over time and ages. Specifically we fit the SSE model on Bulgarian data from 1947 to 2009, ages 12-54. Births are aggregated in age classes and female population by single year of age. No clear component are present in the data, therefore we employ our model to smooth over age and time, accounting for the grouping structure. Figure 4 presents the outcomes.

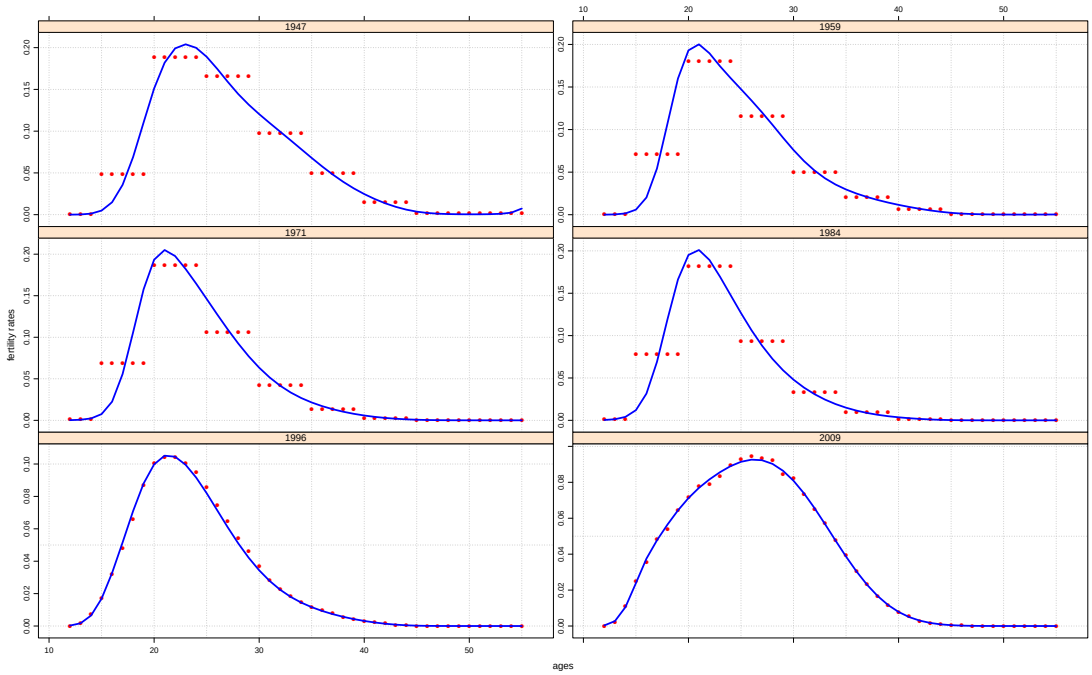

Figure 4: Fertility rates. Bulgaria, 1947-2009, ages 12-54. The dots show the raw rates for varying age groups.

References

Camarda, C. G., P. H. C. Eilers, and J. Gampe (2010). Sums of Smooth Exponentials. In A. Bowman (Ed.), *Proceedings of the 25th International Workshop of Statistical Modelling*, pp. 113–118.

Camarda, C. G., P. H. C. Eilers, and J. Gampe (2012). Additive Decomposition of Vital Rates from Grouped Data. In A. Komárek and S. Nagy (Eds.), *Proceedings of the 27th International Workshop of Statistical Modelling*, pp. 57–62.

Eilers, P. H. C. (2007). Ill-posed Problems with Counts, the Composite Link Model and Penalized Likelihood. *Statistical Modelling 7*, 239–254.

Heligman, L. and J. H. Pollard (1980). The Age Pattern of Mortality. *Journal of the Institute of Actuaries 107*, 49–80.

Hoem, J. M., D. Madsen, J. L. Nelsen, E. M. Ohlsen, H. O. Hansen, and B. Rennermalm (1981). Experiments in modelling recent danish fertility curves. *Demography 18*, 231–244.

Human Fertility Database (2012). *Max Planck Institute for Demographic Research (Germany) and Vienna Institute of Demography (Austria)*. Available at `www.humanfertility.org`. (Data downloaded on August 2012).

Human Mortality Database (2012). *University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany)*. Available at `www.mortality.org`. (Data downloaded on August 2012).

Lambert, P. and P. Eilers (2009). Bayesian density estimation from grouped continuous data. *Computational Statistics & Data Analysis 53*, 1388–1399.

Rogers, A. and L. J. Castro (1981). Model migration schedules. Technical Report 81-30, International Institute for Applied Systems Analysis, Laxenburg, Austria.

Rogers, A., L. J. Castro, and M. Lea (2005). Model Migration Schedules: Three Alternative Linear Parameter Estimation Methods. *Mathematical Population Studies 12*, 17–38.

Rogers, A. and J. Little (1994). Parameterizing Age Patterns of Demographic Rates with the Multiexponential Model Schedule. *Mathematical Population Studies 4*, 175–194.

Romaniuk, A. (1973). A three parameter model for birth projections. *Population Studies 28*, 467–478.

Siler, W. (1983). Parameters of Mortality in Human Populations with Widely Varying Life Spans. *Statistics in Medicine 2*, 373–380.

Thompson, R. and R. J. Baker (1981). Composite Link Functions in Generalized Linear Models. *Applied Statistics 30*, 125–131.