International Journal of Epidemiology, 2019, 571–582
doi: 10.1093/ije/dyy183
Advance Access Publication Date: 26 September 2018
Original article

Methods

How to estimate mortality trends from grouped vital statistics

Silvia Rizzi,[1]* Ulrich Halekoh,[1] Mikael Thinggaard,[1] Gerda Engholm,[2] Niels Christensen,[2] Tom Børge Johannesen[3] and Rune Lindahl-Jacobsen[1]

[1]Institute of Public Health, Unit of Epidemiology, Biostatistics and Biodemography, University of Southern Denmark, Odense, Denmark, [2]Danish Cancer Society, Copenhagen, Denmark and [3]Norwegian Cancer Registry, Oslo, Norway

*Corresponding author. Institute of Public Health, Unit of Epidemiology, Biostatistics and Biodemography, University of Southern Denmark, J.B. Winsløws Vej 9, 5000 Odense, Denmark. E-mail: srizzi@health.sdu.dk

Editorial decision 3 August 2018; Accepted 15 August 2018

Abstract

Background: Mortality data at the population level are often aggregated in age classes, for example 5-year age groups with an open-ended interval for the elderly aged 85+. Capturing detailed age-specific mortality patterns and mortality time trends from such coarsely grouped data can be problematic at older ages, especially where open-ended intervals are used.

Methods: We illustrate the penalized composite link model (PCLM) for ungrouping to model cancer mortality surfaces. Smooth age-specific distributions from data grouped in age classes of adjacent calendar years were estimated by constructing a two-dimensional regression, based on B-splines, and maximizing a penalized likelihood. We show the applicability of the proposed model, analysing age-at-death distributions from cancers of all sites in Denmark from 1980 to 2014. Data were retrieved from the Danish Cancer Society and the Human Mortality Database.

Results: The main trends captured by PCLM are: (i) a decrease in cancer mortality rates after the 1990s for ages 50–75; (ii) a decrease in cancer mortality in later cohorts for young ages, especially, and very advanced ages. Comparing the raw data by single year of age, with the PCLM-ungrouped distributions, we clearly illustrate that the model fits the data with a high level of accuracy.

Conclusions: The PCLM produces detailed smooth mortality surfaces from death counts observed in coarse age groups with modest assumptions, that is Poisson distributed counts and smoothness of the estimated distribution. Hence, the method has great potential for use within epidemiological research when information is to be gained from aggregated data, because it avoids strict assumptions about the actual distributional shape.

Key words: Vital statistics, ungrouping, two dimensions, smoothing, penalized composite link model, oldest old

Key Messages

- Vital statistics, such as disease incidence and cause-of-death data, are often provided in coarse age groups, as is the case for the elderly aged 85+. Such aggregated data hide the age-specific trajectories at higher ages and consequently hinder accurate data analysis in an ageing population. Moreover, ungrouping is needed so that data that follow different groupings can be comparable.
- The two-dimensional penalized composite link model for ungrouping has a double goal: to estimate detailed age-specific trajectories (of mortality or incidence of various diseases) from grouped data and to model a smooth trend over calendar time to control for small fluctuations or noise and misreporting in the data collection.
- Application of this approach in epidemiology and public health research can improve the detail of vital statistics used for further analysis.

Introduction

Vital statistics data often lack fine resolution when publicly released. Death counts are usually provided in aggregated form. Databases of the World Health Organization (WHO)[1] and NORDCAN, a Nordic tool for cancer information, planning, quality control and research,[2] are two examples in which deaths are grouped in age classes, for example age groups of 5 years with a coarse open-ended interval that gathers together the elderly aged 85 and over. The reasons for such aggregation might be privacy protection and the wish for a compact and easy illustration. With increasing longevity in Western countries, wide grouping schemes at older ages are particularly problematic for analysis, given that an increasing proportion of the population dies in the given open-ended age group, for example people aged 85 and over. Records from the Human Mortality Database (HMD) show, for example, that deaths in Denmark of people aged 85 and over rose from 29% of total deaths in the year 2000 up to almost 35% in the year 2014.[3] The analysis of mortality and leading causes of death at older ages is therefore problematic, not only because of the known variation in mortality reporting by physicians,[4] but also because of coarse age groupings. Additionally, grouping schemes might differ across time and countries, which makes comparability challenging.[1]

The most widely used methods to estimate univariate distributions, for example age-specific deaths for a single calendar year, from grouped data are flexible parametric models (with, however, a high number of parameters that lead to overfitting[5–8]) and non-parametric methods such as kernel density estimators and spline interpolations.[5]

When both detailed age-specific mortality patterns and the underlying mortality time trends are of interest, it is necessary to model mortality surfaces, for example a bivariate distribution. We propose using the penalized composite link model (PCLM)[9] to estimate detailed age-at-death distributions of adjacent calendar years from coarsely grouped death counts. The PCLM is a flexible tool that requires modest assumptions. The coarsely grouped counts observed by the researcher can be regarded as indirect observations of a latent sequence of expected counts. These expected counts represent the distribution on a fine resolution that we seek to estimate. This distribution is assumed to be smooth, that is expected counts on the detailed grid are similar to each other and can be estimated from the composite observed data by maximizing a penalized Poisson likelihood. We use a two-dimensional regression analysis, following the P-spline method by Currie *et al.* 2004,[10] for ungrouping the age-specific distributions from the coarsely grouped data and smoothing across adjacent calendar years. Thus, the suggested methodology combines two approaches: the PCLM for ungrouping in one dimension,[11] and two-dimensional smoothing with P-splines.[10,12–14]

The PCLM is a powerful tool to model aggregated epidemiological and demographic data. In a one-dimensional setting it has been found to model age-specific grouped data,[11] outperforming kernel density estimator and spline interpolation methods in the presence of open-ended intervals especially,[15] and data suffering from digit preferences.[16] In two- or three-dimensional settings it has been applied to aggregated spatial counts[17] and to fertility rates grouped by age, time and birth order.[18] The methodology was also explored in a Bayesian framework.[19] The use of this model is as yet low. Here, we focus on smooth estimations of mortality surfaces from data grouped in age classes, with particular attention to open-ended intervals at the right-hand tail of the age-at-death distributions.

We will first introduce the PCLM for ungrouping coarse age-at-death distributions of adjacent calendar years and, secondly, analyse age-at-death distributions from cancers of all sites in Denmark from 1980 to 2014, using the proposed methodology. We conclude with a discussion and give a demo R-code of the model in the Appendix and in the Supplementary material, with the aim of bridging the statistical methodology with a view to its application in the public health field.

Methods

The statistical method for ungrouping coarse age-at-death distributions of adjacent calendar years

Suppose that one could observe death counts by single year of age for several adjacent calendar years. Let us denote these death counts on a fine grid by the matrix $Z = (z_{jn})$ of dimension $J \times N$, where $j = 1, \ldots, J$ corresponds to the ages in a single year step and $n = 1, \ldots, N$ to adjacent calendar years. Death counts are assumed to follow a Poisson process, with expected value $E(z_{jn}) = \gamma_{jn}$. In practice, we are only able to observe these death counts grouped in coarser age classes. We denote the observed death counts aggregated in coarse age groups for adjacent calendar years by the matrix $Y = (y_{in})$ of dimension $I \times N$, where $i = 1, \ldots, I$ corresponds to the age classes and $n = 1, \ldots, N$ to adjacent calendar years. Death counts y_{in} follow a Poisson distribution, with expected value $E(y_{in}) = \mu_{in}$, where μ_{in} can be seen as expected values composed of the latent unobserved expectations γ_{jn}. Hence, γ_{jn} is the expected mortality surface on a fine grid, which we aim to estimate from the composite means μ_{in}. This is done by maximizing a penalized Poisson likelihood via maximum likelihood estimation.

For the purpose of the regression, we arrange the data by column vector, that is $y = (y_{11}, y_{21}, \ldots, y_{IN})'$, $\mu = (\mu_{11}, \mu_{21}, \ldots, \mu_{IN})'$ and $\gamma = (\gamma_{11}, \gamma_{21}, \ldots, \gamma_{JN})'$. The PCLM for ungrouping coarse age-at-death distributions of adjacent calendar years is then given by:

$$\mu = C\ \gamma = C\ \exp(B\theta), \tag{2.1}$$

where γ is the sequence of detailed expected death counts that we aim to estimate, C is a composition matrix, B a B-spline basis with θ being the corresponding coefficient. The matrix C is a 0/1 block matrix that 'composes' μ from γ and describes how the latent distribution was mixed before generating the data. The C matrix is derived from the Kronecker product of the two marginal composition matrices $C = C_t \otimes C_a$, where C_t is an identity matrix of dimension $N \times N$ and C_a of dimension $I \times J$. Elements of C_a are zero, except for those $c_{ij} = 1$ that indicate the elements of the expected counts γ that are summed to get the aggregated expected counts μ for each calendar year n. In order to model rates instead of counts, each column of C is multiplied by the corresponding exposure to risk, that is person-years at risk of dying, by single year of age (if the exposures are also grouped, one can produce detailed estimates again with the PCLM).

In the PCLM, γ is overparametrized with respect to μ. This is solved by two regularizations following the P-splines methodology by Eilers and Marx (1996).[20] First, the set of γ is restricted by using a B-spline basis of lower dimension than $J \times N$; second, the linear space of the coefficients θ is restricted by introducing a penalty matrix P.

As the first step, we use as a two-dimensional regression matrix a B-spline basis B, which reduces the number of parameters θ to be estimated, since the number of elements in γ is very large. Following[10] $B = B_t \otimes B_a$, where B_t is the marginal B-spline basis in the calendar year direction of dimension $N \times K_t$ and B_a is the marginal B-spline basis in the age direction of dimension $J \times K_a$, K_t and K_a are equal to the number of knots of the corresponding marginal B-spline basis, plus the degree of the spline. Throughout the paper cubic splines are used.

As the second step, we assume that the latent distribution γ is smooth, that is adjacent elements of γ are similar, and implement this assumption by a roughness penalty on the parameters θ, with θ being the vector of coefficients associated with the regression matrix B of length $K_t K_a$. Then, the penalty matrix follows as $P(\theta) = \lambda_a \left(I_{K_t} \otimes D_a' D_a\right) + \lambda_t \left(D_t' D_t \otimes I_{K_a}\right)$, with λ_a and λ_t the smoothing parameters in the age and calendar year directions, respectively, that control the amount of smoothness, and D_a and D_t the matrices that compute the d^{th} differences of the regression coefficients θ. In the paper, the roughness of the coefficients θ is measured by second-order differences following Currie *et al.* (2004).[10]

Estimation procedure

The PCLM for ungrouping coarse age-at-death distributions of adjacent calendar years can be estimated by maximum likelihood estimation, by maximizing the following penalized Poisson log likelihood:

$$l^* = l - P = \sum_{i=1}^{I} \sum_{n=1}^{N} \left(y_{in} \ln \mu_{in} - \mu_{in}\right) - P. \tag{2.2}$$

Maximizing (2.2) leads to a system of equation that can be solved by an appropriately modified version of the iteratively reweighted least squares (IRWLS) algorithm. The system of equations becomes:

$$\left(\breve{B}' \tilde{W} \breve{B} + P\right)\theta = \breve{B}' \tilde{W} \left[\tilde{W}^{-1} (y - \tilde{\mu}) + \breve{B}\tilde{\theta}\right], \tag{2.3}$$

where $\breve{B} = \tilde{W}^{-1} C \tilde{\Gamma} B$ with $\tilde{\Gamma} = diag(\tilde{\gamma})$ and $\tilde{W} = diag(\tilde{\mu})$ which are diagonal matrices of weights, with tilde indicating the approximation to the solution of a specific iteration, and P the penalty matrix. For a given value of the smoothing parameters, the system can be solved. To obtain the optimal values of smoothing, we computed the Akaike

Information Criterion (AIC) for a two-dimensional grid of λ-values (on a log scale) and chose the optimal smoothing corresponding to its minimal value. Standard errors for the estimated $\hat{\gamma}$ can be obtained from a sandwich-estimator for $var(\hat{\theta})$, given by $var(B\hat{\theta}) = \left(\left(\breve{B}'W\breve{B} + P\right)^{-1}\left(\breve{B}'W\breve{B}\right)\left(B'W\breve{X} + P\right)^{-1}\right)$ or via a Bayesian approach, such as $var(B\hat{\theta}) = \left(B'W\breve{X} + P\right)^{-1}$. Standard errors are obtained by taking the square root of the diagonal elements as $s.e. = \sqrt{diag\left(B\,var(B\hat{\theta})\,B'\right)}$, and confidence intervals for the estimated $\hat{\gamma}$ are consequently given by $e^{(B\hat{\theta} \pm 2\,s.e)}$. Standard errors do not reflect the uncertainty introduced by the choice of the smoothing parameters λ_a and λ_t that are treated as fixed. The estimating procedure was implemented in R version 3.2.2,[21] and the code is provided in the Appendix to this paper.

Data

Age-specific cancer deaths in adjacent calendar years

Age-specific cancer mortality of all sites, including non-melanoma skin cancer, men and women together, in Denmark from 1980 to 2014, is analysed with the proposed method. Raw death counts were retrieved from the Danish Cancer Society from the database for NORDCAN.[2] NORDCAN is a database that contains incidence, mortality, prevalence and survival statistics for 50 major cancers in Nordic countries. Danish data for cancer mortality in the NORDCAN database stem from the Danish Registry of Causes of Death.[22] An internet application provides access to summary data with graphic and tabulation facilities. From the Danish Cancer Society, overall cancer death counts were obtained on a finer resolution, that is from age 0 up to the last age of recorded events in the time window 1980–2014 for each consecutive calendar year and for both sexes combined. Deaths were classified according to the International Classification of Diseases, Tenth Revision (ICD-10, codes CXX.X + D09.0–1+ D30.1–9 + D35.2–4 + D41.1–9 + D32-33 + D42-43 + D44.3–5 + D46-47). The quality of this classification relies mainly upon the reporting of physicians; the analyses of overall cancer mortality reduces the impact of possible misclassifications or discontinuity in registrations of specific cancer sites.[23,24] The total number of cancer deaths among the Danish population in the 35 years considered was 529 511. The year 1995 had the highest number of cancer deaths, with 15 914 cases, whereas the age of 77 was the most affected age across the entire time window, with 17 675 deaths.

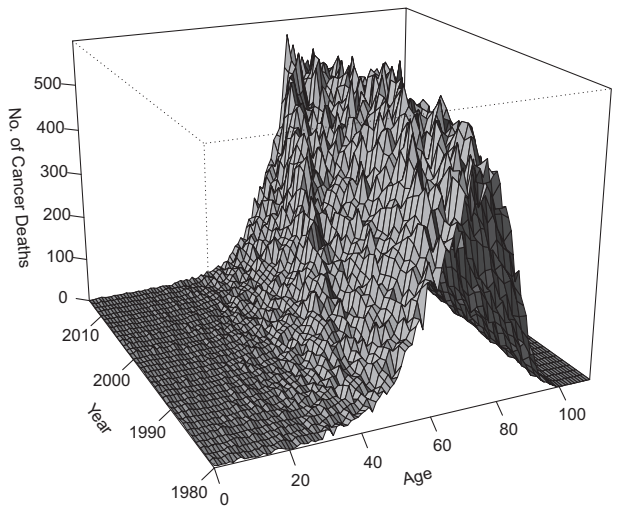

Figure 1. Age-at-death distributions from cancers of all sites, including non-melanoma skin cancer, in Denmark from 1980 to 2014, men and women together from raw data. Source: Danish Cancer Society.

The age-at-death distributions for adjacent years from the raw data are illustrated in Figure 1.

Person-years at risk of dying, by single year of age for each calendar year, were taken from the Human Mortality Database,[3] which is a free database containing data on death from all causes, exposure to risk, and population size, by age and calendar year for 38 countries. Danish data in the Human Mortality Database stem from Statistics Denmark. We used exposures to risk provided for Denmark for each calendar year at 1-year age intervals from 0 to 109 years, and then a final group of 110+.

Application and results

The high-quality data described in the section above[22,25] allowed us to investigate the performance of the proposed method, by comparing the raw detailed death rates with the PCLM ungrouped estimates. To construct the mortality surface on a fine grid (Figure 2, right-hand panel), death counts for each age and calendar year, obtained from the Danish Cancer Society, were divided by the corresponding exposures to risk retrieved from the Human Mortality Database.[3]

To study the performance of the proposed model, we grouped the raw death counts from cancer into age classes of different lengths, that is 5-year age classes with an open-ended interval starting at age 85, following the age grouping of most of the freely available databases.[1,2] After grouping the raw cancer deaths, we applied the PCLM to model the mortality surface at a fine resolution, that is from age zero up to age 110+, for the time window 1980–2014, in 1-year age groups. Age 110+ was set as the maximum age, assuming that no cancer deaths are

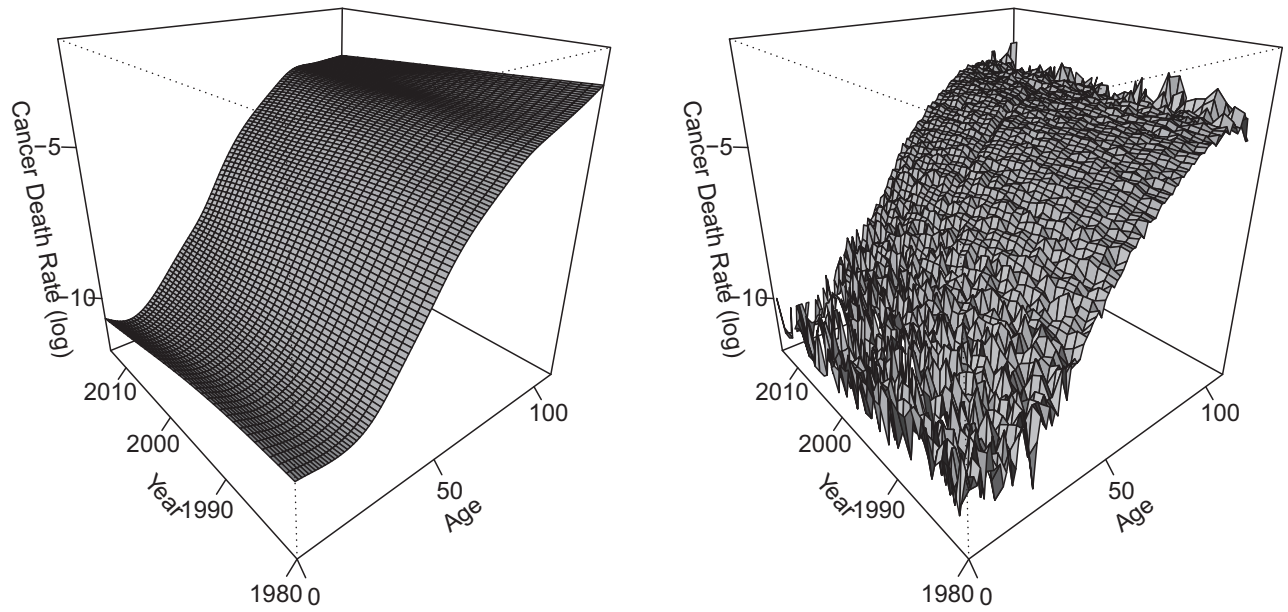

Figure 2. Age-specific death rates from cancers of all sites, including non-melanoma skin cancer, in Denmark from 1980 to 2014. Left-hand panel: smooth cancer mortality surface estimated by the PCLM for ungrouping age-at-death distributions of adjacent calendar years from coarsely aggregated data. Right-hand panel: detailed cancer mortality surface from raw data. Sources: Danish Cancer Society and Human Mortality Database.

expected after that age. In the present application, the age groups are equal to $I = 18$, whereas $J = 111$ are the detailed ages and $N = 35$ the calendar years analysed. Z of dimension $J \times N$ denotes the raw cancer deaths with expected value γ, and Y of dimension $I \times N$ denotes the aggregated cancer death counts with expected value μ. We aimed at estimating γ from μ, with $\mu = C\gamma = C\exp(B\theta)$. The two-dimensional regression basis B has a dimension of 3885×112, as it results from the Kronecker product of the two marginal B-spline bases: B_t of dimension 35×7, where 35 is the number of calendar years analysed, and B_a of dimension 111×16, where 111 is the number of detailed ages. The number of columns of the marginal B-spline bases is the sum of the number of knots chosen and the degree of the spline. Following the rule of thumb described in Currie *et al.* (2004),[10] one knot every eight elements of γ equally spaced and cubic splines are used, giving seven columns, K_t, (floor(35/8)+3) for B_t and 16 columns, K_a, (floor(111/8)+3) for B_a. The resulting model has 112 parameters $\theta = (\vartheta_1, \ldots, \vartheta_{112})$ in order to estimate the mortality surface, that is $7 \times 16 = 112$. The composition matrix C of dimension 630×3885 is the Kronecker product of the two marginal matrices C_t and C_a, of dimensions 35×35 and 18×111, respectively. C_t is an identity matrix with as many rows and columns as the N calendar years analysed, whereas C_a has as many rows as the I age classes and as many columns as the J ages on the fine resolution. When modelling mortality rates, the composition matrix C is further multiplied by the respective exposures as offset so that the estimated γ is a mortality surface.

Table 1. Root mean squared error (RMSE) calculated in the year and age dimensions. RMSE measures the differences between estimates and raw data. Values close to 0 indicate a good fit. The RMSE of the PCLM is compared with the RMSE derived by only smoothing the raw detailed data. Results do not differ much, indicating that the ungrouping procedure fits the data well

	RMSE PCLM	RMSE smoothing
Year	0.0151	0.0088
Age	0.0034	0.0013

From the Human Mortality Database, detailed exposure figures, that is for single year of age from zero up to 110+ for all calendar years analysed, were retrieved. In order to find the solution of (2.3), the optimal values of the smoothing parameters are selected via AIC: $\hat{\lambda}_a = 0.316$ in the age direction and $\hat{\lambda}_t = 3.160$ in the calendar year direction. The so obtained smooth cancer mortality surface, γ of dimension $J \times N$, is illustrated in Figure 2, left-hand panel.

Comparing the raw data by single year of age with the PCLM-ungrouped distributions, we find that the model fit is good. This is captured visually in Figure 2 and it is further quantified by the root mean squared error (RMSE) reported in Table 1.

When modelling a smooth mortality surface, the estimated trends in the age and year directions are of particular interest. To better show such patterns modelled by the PCLM, we report estimates versus raw data in a

cross-sectional cut in the age and year direction in Figures 3 and 4, respectively. The level of uncertainty of the estimates is captured by the width of the confidence intervals (Figure 3; and Figure A1 in the Appendix): uncertainty increases at high ages, particularly after 100 years of age, where death counts and exposures are of very small numbers.

The main trends captured by PCLM are: (i) a decrease in cancer mortality rates after the 1990s for ages 50–75 (Figure 3; and for more details, see Figure A2 in the Appendix); (ii) a decrease in cancer mortality for later cohorts, particularly for young ages and very advanced ages (Figure 4); (iii) a peak of cancer mortality around age 90, followed by a levelling off (Figure 4; and Figure A1). These results are in line with the existing literature on cancer mortality patterns in Denmark[26,27] and similar to those observed in international studies.[28–31] Modelling cancer mortality surfaces with the PCLM allows us to gain information about the mortality patterns, of the oldest old especially, from data that are otherwise aggregated at the higher ages. The levelling off of cancer mortality around age 90 could not be captured with data coarsely grouped in the traditional 85+ open-ended age class. Smoothing in the calendar year direction allows us at the same time to capture the trend across time and to correct for fluctuations or noise in data collection.

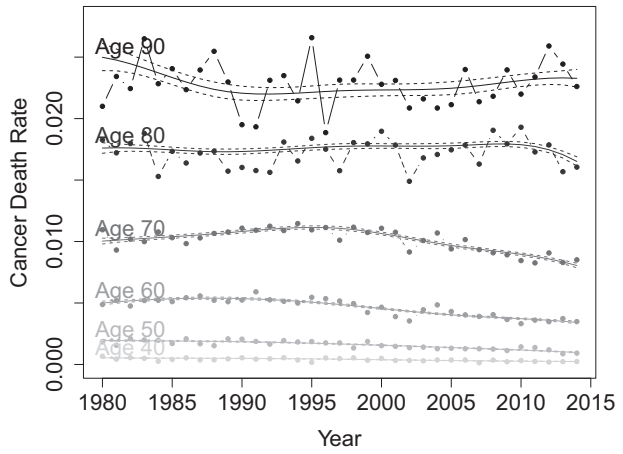

Figure 3. Death rates from all cancers, including melanoma skin cancer, in Denmark from 1980 to 2014 for ages 40, 50, 60, 70, 80 and 90 years. Smooth lines are PCLM smooth estimates with confidence intervals (dashed lines), and dots are raw death rates. Sources: Danish Cancer Society and Human Mortality Database.

Discussion

We have presented the PCLM to model smooth mortality surfaces from age grouped data. In our application, we were able to reconstruct from aggregated data a cancer mortality surface on a fine age grid equivalent to the raw ungrouped data. At the same time we model a smooth time trend across calendar years.

In the PCLM setting, the user has the freedom of selecting the number of knots in the marginal B-spline bases, the degree of the spline and the order of the penalty plus the smoothing parameters. To choose the optimal number of knots in the B-spline basis, no information criterion is adopted: the idea is to choose a rich enough basis and let the smoothing parameters control the optimal level of smoothness. Following Currie *et al.* (2004),[10] we suggest using one knot about every eight elements of the latent and detailed distribution γ, cubic splines and quadratic penalties. This choice is not crucial, since a different degree of the spline or order of the penalty leads to a similar fit. The optimal smoothing parameters $\hat{\lambda}_a$ and $\hat{\lambda}_t$ are obtained using a grid search of possible values of λ_a and λ_t and selecting those that minimize AIC. In our application, the range of smoothing parameter values is set as $(10^{-1}, 10^{-4})$. We found that allowing too small values of (λ_a, λ_t) leads to singularity of the system of equations: This occurs because the parameter θ is free to vary where it is also not informative.

The main assumptions of the PCLM for ungrouping in two dimensions are smoothness of the estimated surface and

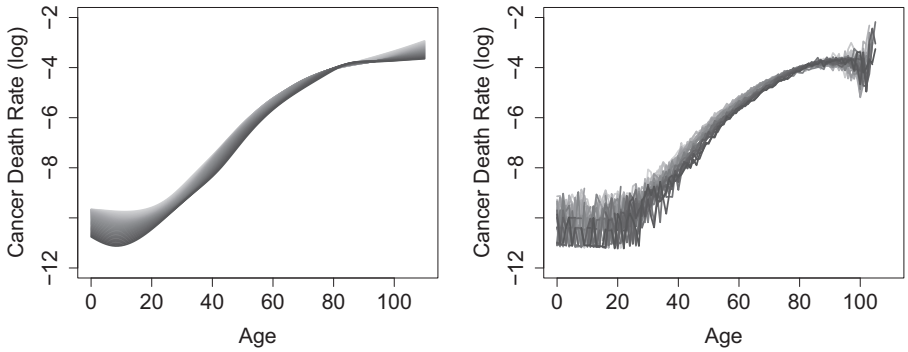

Figure 4. Age-specific death rates from all cancers, including melanoma skin cancer, in Denmark from 1980 to 2014. Different calendar years are displayed with different shades of grey: lighter grey corresponds to earlier years and progressively darker grey to more recent years. Left-hand panel: PCLM smooth estimates. Right-hand panel: raw death rates. Sources: Danish Cancer Society and Human Mortality Database.

Poisson distributed counts. When consecutive observations are expected to be characterized by drastic differences in magnitude, the smoothness assumption is violated. In age-at-death analysis appreciable differences are found, for example between mortality levels at age 0 years and following ages. Explicitly including a point mass for infant mortality can solve this limitation. Another assumption that does not always hold for epidemiological data is the Poisson distribution of counts, that is $Var(y) = E(y)$. In some circumstances data are overdispersed, that is $Var(y) > E(y)$, and therefore the variability of the estimates might be underestimated.

In our application, we chose the same age grouping for all calendar years analysed. However, in practice, the age grouping scheme can differ across calendar years. This is particularly true for historical data or data of the past few decades, aggregated, for example, in 10-year age classes with open-ended intervals starting at age 75 or 80.[1] By changing the composition matrix C accordingly, the PCLM can handle different age groupings. In population-based data, deaths are usually reported by each single calendar year.[1–3] However, in some cases, time might also be grouped, for example in classes of 5 calendar years. Modifying the marginal composition matrix C_t accordingly, the PCLM can be extended to model smooth mortality surfaces from both grouped ages and calendar years.

The proposed methodology can be applied to deaths from various causes—since no assumption about the shape of the sequence γ is made. Because of its flexibility, it can also serve other epidemiological applications with bivariate data that can be assumed to be Poisson distributed, for exampple grouped data on individuals' body mass index (BMI) and height.

Conclusion

The PCLM has proven useful in this study in gaining detailed information when vital statistics are aggregated in coarse age groups. The PCLM applied to age-specific mortality data allows us to capture the underlying trends in two dimensions, that is age and calendar time, when limited information is to hand.

Supplementary Data

Supplementary data are available at *IJE* online.

Conflict of interest: None declared.

References

1. World Health Organization. *Global Health Observatory Data Repository*. http://www.who.int/ghodata/ (1 August 2017, date last accessed).
2. Engholm G, Ferlay J, Christensen N *et al*. *NORDCAN: Cancer Incidence, Mortality, Prevalence and Survival in the Nordic Countries, Version 7.3 (08.07.2016)*. Association of the Nordic Cancer Registries. Danish Cancer Society. http://www.ancr.nu (1 August 2017, date last accessed).
3. Human Mortality Database. *University of California, Berkeley (USA) and Max Planck Institute for Demographic Research (Germany)*. 2015. http://www.mortality.org (10 October 2017, date last accessed).
4. Lozano R, Naghavi M, Foreman K *et al*. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet* 2012;**380**:2095–128.
5. Kostaki A, Panousis V. Expanding an abridged life table. *DemRes* 2001;**5**:1–22.
6. Kostaki A. The Heligman-Pollard formula as a tool for expanding an abridged life table. *J Off Stat* 1991;**7**:311–23.
7. Hsieh JJ. Construction of expanded continuous life tables—a generalization of abridged and complete life tables. *Math Biosci* 1991;**103**:287–302.
8. Kostaki A, Lanke J. Degrouping mortality data for the elderly. *Math Popul Stud* 2000;**7**:331–41.
9. Eilers PHC. Ill-posed problems with counts, the composite link model and penalized likelihood. *Stat Model* 2007;**7**:239–54.
10. Currie ID, Durban M, Eilers PHC. Smoothing and forecasting mortality rates. *Stat Model* 2004;**4**:279–98.
11. Rizzi S, Gampe J, Eilers PHC. Efficient estimation of smooth distributions from coarsely grouped data. *Am J Epidemiol* 2015; **182**:138–47.
12. Eilers PHC, Marx BD. Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. *Chemometr Intell Lab Syst* 2003;**66**:159–74.
13. Eilers PHC, Currie ID, Durban M. Fast and compact smoothing on large multidimensional grids. *Comput Stat Data Anal* 2006; **50**:61–76.
14. Camarda CG. MortalitySmooth: An R Package for Smoothing Poisson Counts with P-Splines. *J Stat Softw* 2012;**50**:1–24.
15. Rizzi S, Thinggaard M, Engholm G *et al*. Comparison of nonparametric methods for ungrouping coarsely aggregated data. *BMC Med Res Methodol* 2016;**16**:59.
16. Camarda C, Eilers PHC, Gampe J. Modelling general patterns of digit preference. *Stat Model* 2008;**8**:385–401.
17. Ayma D, Durban M, Lee D-J, Eilers PHC. Penalized composite link models for aggregated spatial count data: a mixed model approach. *Spat Stat* 2016;**17**:179–98.
18. Ayma D, Camarda C. Modelling fertility rates by age, time, and birth order from coarsely grouped data: a penalized composite link model approach. *Extended Abstract for European Population Conference, 31 August -3 September 2016*. Mainz, Germany, 2016.
19. Lambert P. Smooth semiparametric and nonparametric Bayesian estimation of bivariate densities from bivariate histogram data. *Comput Stat Data Anal* 2011;**55**:429–45.
20. Eilers PHC, Marx BD. Flexible smoothing with B-splines and penalties. *Stat Sci* 1996;**11**:89–121.
21. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

22. Helweg-Larsen K. The Danish register of causes of death. *Scand J Public Health* 2011;**39**(**Suppl 7**):26–29.
23. Storm HH, Michelsen EV, Clemmensen IH, Pihl J. The Danish cancer registry—history, content, quality and use. *Dan Med Bull* 1997;**44**:535–39.
24. Gjerstorff ML. The Danish cancer registry. *Scand J Public Health* 2011;**39**(**Suppl 7**):42–45.
25. Pukkala E, Engholm G, Schmidt L *et al*. Nordic Cancer Registries—an overview of their procedures and data comparability. *Acta Oncol* 2018;**57**:440–55.
26. Ewertz M, Christensen K, Engholm G *et al*. Trends in cancer in the elderly population in Denmark, 1980-2012. *Acta Oncol* 2016;**55**:1–6.
27. Pedersen JK, Engholm G, Skytthe A, Christensen K. Cancer and aging: epidemiology and methodological challenges. *Acta Oncol* 2016;**55**:7–12.
28. La Vecchia C, Bosetti C, Lucchini F *et al*. Cancer mortality in Europe, 2000-2004, and an overview of trends since 1975. *Ann Oncol* 2010;**21**:1323–60.
29. Liu L, Liu K. Age-specific cancer mortality trends in 16 countries. *Int J Public Health* 2016;**61**:751–63.
30. Smith DW. Cancer mortality at very old ages. *Cancer* 1996;**77**:1367–72.
31. Harding C, Pompei F, Wilson R. Peak and decline in cancer incidence, mortality and prevalence at old ages. *Cancer* 2012;**118**:1371–86.

Appendix

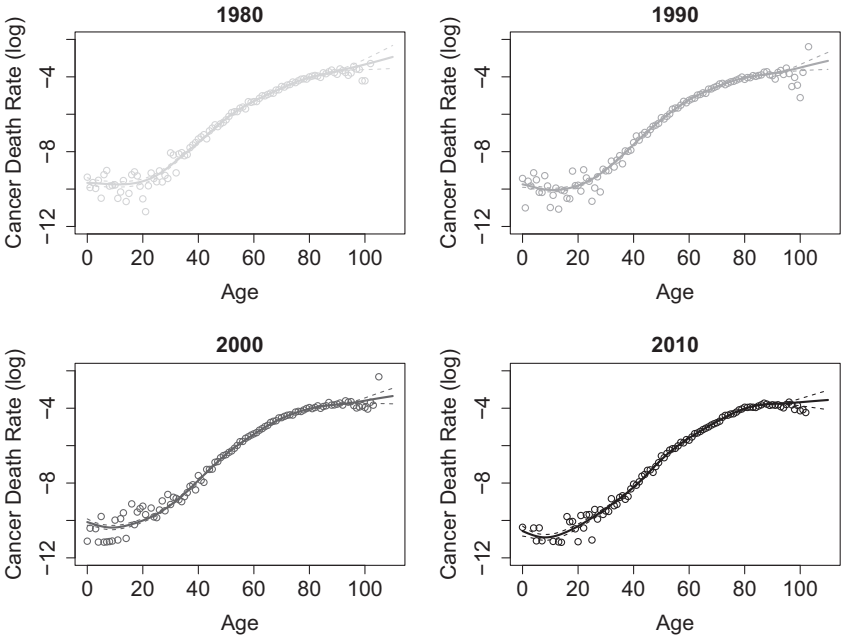

Figure A1. Age-specific death rates from all cancers including melanoma skin cancer in Denmark in 1980, 1990, 2000 and 2010. Smooth lines are PCLM smooth estimates with confidence intervals (dashed lines) from death counts grouped in age classes of 5 years, with open-ended age group 85+. Dots are raw death rates. Sources: Danish Cancer Society and Human Mortality Database.

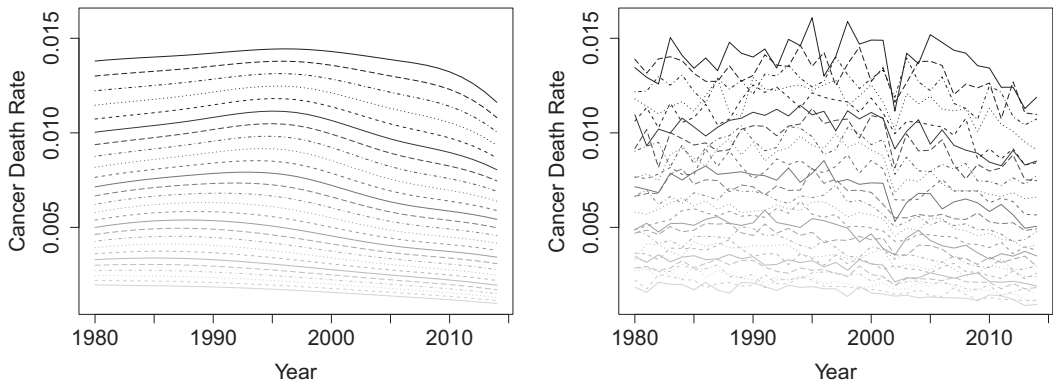

Figure A2. Death rates from all cancers, including melanoma skin cancer, in Denmark from 1980 to 2014 for ages 50–75 years. The mortality trajectories over calendar years are displayed with different shades of grey, according to the ages: lighter grey corresponds to earlier ages, starting at age 50, and progressively darker grey corresponds to older ages, up to age 75. Left-hand panel: PCLM smooth estimates. Right-hand panel: raw death rates. Sources: Danish Cancer Society and Human Mortality Database.

Demo R code to estimate mortality surfaces with the penalized composite link model for ungrouping

```
# PCLM core function
pclm2D <- function(y, C, B, P, show = 0){

    # Fit a 2D PCLM (estimate A in E(y) = C %*% exp(B %*% A))
      # y = the matrix of grouped counts of dimension I × N in vector format
      # C = the composition matrix, i.e. C = C_t ⊗ C_a
      # B = the two-dimensional B-spline basis, i.e. B = B_t ⊗ B_a
      # P = penalty matrix
      # A = vector of parameters, i.e. θ
      # gam = mortality surface γ of dimension J × N to be estimated
      # mu = composite distribution μ, expected value of y, of dimension I × N

      # Preparations
      nx <- dim(B)[2] # length of the parameters vector θ, i.e. K_t × K_a
      ly <- length(y)
      it <- 0
      Astart <- log(sum(y) /ly); # initial coefficient value for the algorithm
      A <- rep(Astart, nx); # θ vector of parameters to model mortality surface

      # Iterations for the IRWLS algorithm
      for (it in 1:50) {

          A0 <- A
          eta <- B %*% A
          gam <- exp(eta)
          mu <- c(C %*% gam)
          w <- mu

          Q <- (C * ((1 / mu) %*% t(gam)) ) %*% B
          z <- y - mu + w * Q %*% A
          QWQ = t(Q) %*% (w * Q)

          A = solve(QWQ + P, t(Q) %*% z) # solve system of equations
          da <- max(abs(A - A0))
          if (show) cat(it, " ", da, "\n")
          if (da < 1e-6) break # stop when two successive values of θ < 10^-6
      }
      cat(it, " ", da, "\n")
```

```
    fit = list(coeff = A, gamma = gam, mu = mu)
    H = solve(QWQ + P, QWQ)
    H0 <- solve(QWQ + P) # variance-covariance matrix Bayesian approach
    H1 <- H0 %*% QWQ %*% H0 # variance-covaraince matrix sandwitch estimator

    fit$trace <- sum(diag(H))
    ok <- y > 0
    fit$dev <- 2 * sum(y[ok] * log(y[ok] / mu[ok]))
    fit$psi2 <- fit$dev / (length(y)-fit$trace)
    fit$aic <- fit$dev + 2 * fit$trace
    fit$bic <- fit$dev + log(length(y)) * fit$trace
    fit$H0 <- H0
    fit$H1 <- H1
    fit$eta <- eta

    return(fit)

}


# Simultaneous estimation of smooth mortality from grouped death counts for adjacent
calendar years: An example using Human Mortality Database (HMD) data.

library(MortalitySmooth)
library(rgl)
library(svcm)

# Read HMD data for Sweden 1980-2014

library(xml2)
library(HMDHFDplus)
# Create an account under www.mortality.org to get access to the data
username <- ""
password <- ""
# Deaths
Deaths <- readHMDweb("SWE", "Deaths_1x1", username, password)
Deaths_subset <- subset(Deaths, Year >= 1980 & Year <=2014)$Total
d_counts_m <- matrix(Deaths_subset, nrow = 111)
# Exposures
Exposures <- readHMDweb("SWE", "Exposures_1x1", username, password)
Exposures_subset <- subset(Exposures, Year >= 1980 & Year <=2014)$Total

# Number of years
ny <- ncol(d_counts_m)

# Group counts
d_counts <- as.data.frame(d_counts_m)
d_counts$Groups_Counts <- c(rep(1:17, each=5), rep(18,26))
y <- aggregate (d_counts[, 1:35], by=list(d_counts$Groups_Counts), FUN="sum")
y <- y[, -1]

# Deaths from cancer for both sexes combined
# and put in vec format
y <- as.vector(unlist(y))

# Number of bins
nb <- rep(18,ny)
```

```
# Age grid for the underlying distribution for each calendar year
m = 111
x = 1:m

# Define the grouping
# e.g. 5 years age groups with 85+
ilo = seq(1, 86, by = 5)
ihi = ilo + 4
n = length(ihi)
ihi[n] = m
# intervals lengths
leng <- ihi-ilo+1

# Construct C matrix
# CA matrix in the age direction
CA = matrix(0, n, m)
for (i in 1:n) CA[i, ilo[i] : ihi[i]] = 1
# CY matrix in the year direction
CY = matrix(0, ny, ny)
diag(CY) = 1
# C as kronecker product
Ci = kronecker(CY,CA)
# exposures as offset
E <- Exposures_subset
C = Ci %*% diag(E)

# Construct B-spline basis
# for age
basisA = 1:m
xl <- min(basisA)
xr <- max(basisA)
xmax <- xr + 0.01 * (xr - xl)
xmin <- xl - 0.01 * (xr - xl)
BA <- MortSmooth_bbase(basisA, xmin, xmax, ndx=floor(m/15), deg=3)
# for year
basisY = 1:ncol(CY)
yl <- min(basisY)
yr <- max(basisY)
ymax <- yr + 0.01 * (yr - yl)
ymin <- yl - 0.01 * (yr - yl)
BY <- MortSmooth_bbase(basisY, ymin, ymax, ndx=floor(ncol(CY)/15), deg=3)
# B as kronecker product
B = kronecker (BY, BA)

# Second order penalties
DA = diff(diag(ncol(BA)), diff =2)
PA = kronecker(diag(ncol(BY)), t(DA) %*% DA)
DY = diff(diag(ncol(BY)), diff =2)
PY = kronecker(t(DY) %*% DY,diag(ncol(BA)))

lambdaA.hat = 0.1
lambdaY.hat = 0.1
P = (lambdaA.hat*PA) + (lambdaY.hat*PY)

# Model estimation
mod = pclm2D(y, C, B, P, show = 0)
cat(mod$aic, mod$bic, mod$trace, '\n')
```

```
Gamma <- matrix(mod$gamma, ncol=ncol(CY))

# 3D plot
x <- 0:110
y <- seq(1980,2014)
z <- Gamma

persp(x, y, log(z),
      theta = -40, phi = 25, r = sqrt(3), d = 1,
      xlab = "Age", ylab = "Year", zlab=" Mortality Death Rate (log)",
      col="lightgrey",
      main = "Mortality Surface: Sweden 1980 - 2014
      2D ungrouping with PCLM",
      ticktype="detailed", nticks=5,
      shade=T)
title(sub="Source: Human Mortality Database",adj=0, line=1)
```