STATISTICS IN MEDICINE
Statist. Med. 2008; **27**:3430–3441
Published online 8 January 2008 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3188

Modulation models for seasonal time series and incidence tables

Paul H. C. Eilers[1], Jutta Gampe[2], Brian D. Marx[3] and Roland Rau[4,*,†]

[1]*Faculty of Social and Behavioural Sciences, Utrecht University, Utrecht, The Netherlands*
[2]*Max Planck Institute for Demographic Research, Rostock, Germany*
[3]*Department of Experimental Statistics, Louisiana State University, Baton Rouge, LA, U.S.A.*
[4]*Terry Sanford Institute of Public Policy, Duke University, Durham, NC, U.S.A.*

SUMMARY

We model monthly disease counts on an age–time grid using the two-dimensional varying-coefficient Poisson regression. Since the marginal profile of counts shows a very strong and varying annual cycle, sine and cosine regressors model periodicity, but their coefficients are allowed to vary smoothly over the age and time plane. The coefficient surfaces are estimated using a relatively large tensor product B-spline basis. Smoothness is tuned using difference penalties on the rows and columns of the tensor product coefficients. Heavy over-dispersion occurs, making it impossible to use Akaike's information criterion or Bayesian information criterion based on a Poisson likelihood. It is handled by selective weighting of part of the data and by the use of extended quasi-likelihood. Very efficient computation is achieved with fast array algorithms. The model is applied to monthly deaths due to respiratory diseases, for U.S. females during 1959–1998 and for ages 51–100. Copyright © 2008 John Wiley & Sons, Ltd.

KEY WORDS: generalized linear array models; over-dispersion; P-splines; tensor product

1. INTRODUCTION

Many disease-related events reveal considerable seasonal variation, mostly striking harder in winter than in summer. The size of this seasonal effect may depend on the age and gender of the victims as well as on the type of disease. In a seasonal incidence table, one that presents event counts with a time precision of months or quarters and age in years, we will see clear ripples with a period of one year. An example is shown in Figure 1 for females of ages 51–100 in the United States who died of respiratory diseases during 1959–1998. The top panel shows the raw counts and can also

*Correspondence to: Roland Rau, Terry Sanford Institute of Public Policy, Population, Policy, and Aging Research Center, Duke University, Box 90309, 302 Towerview Road, Durham, NC 27708-0309, U.S.A.
†E-mail: roland.rau@duke.edu

Contract/grant sponsor: National Science Foundation
Contract/grant sponsor: AFOSR; contract/grant numbers: IIS-0326387, FA9550-05-1-0454

Received 9 January 2007
Accepted 13 November 2007

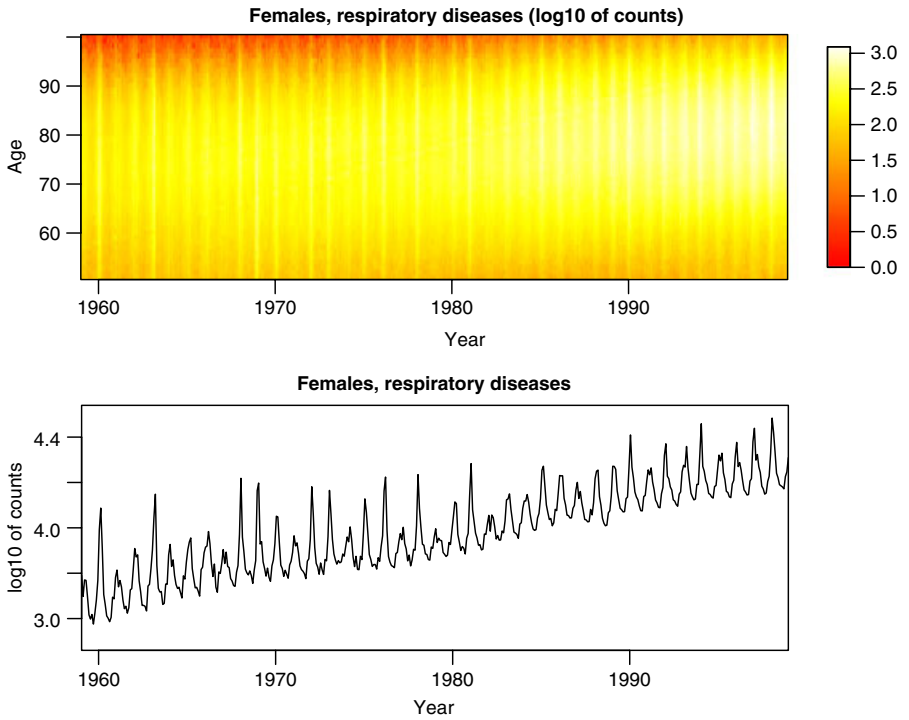

Figure 1. Image plot of death counts due to respiratory diseases, U.S. females, ages 51–100 during years 1959–1998 (top) and the marginal plot of monthly totals (bottom) reflecting strong and variable seasonality as well as a large trend.

be viewed as an image of a flattened large contingency table of monthly death counts with 24 000 cells (50 ages $\times$ 40 years $\times$ 12 months). For fixed year, we generally find that the death counts increase with age, until approximately 85, following the typical density of deaths from human populations. Deaths from respiratory diseases are influenced not only by lifestyle factors such as smoking but also by seasonally fluctuating forces such as the climate and air pollution. Figure 1 (bottom panel) shows the monthly totals, aggregated over all ages, to visualize marginal seasonal variation and in fact shows very strong and varying cyclical behavior due to seasonal effects.

It is of interest to quantify the character and the strength of such seasonal patterns. This includes the identification of the peak season and whether it is stable across ages and years, and also whether (relative) seasonal amplitudes change over time or differ among age-groups. Comparisons of different diseases or causes of death may also be of interest. In this paper, we present a very general model to give the answers. We generalize the work of Gampe and Rau [1], who presented modulation models for demographic time series.

First we describe our data in Section 2. We then start modeling in Section 3, by revisiting the one-dimensional modulation model for a time series of counts, and the use of P-splines in this context. When tuning the smoothness parameters we encounter heavy over-dispersion; we handle it with selective weighting and the use of quasi-likelihood instead of a Poisson model. In Section 4, we extend the model to incidence data, allowing the amplitude of the seasonal components to vary

DOI: 10.1002/sim

both with time and age. There we also sketch how recently developed array methods [2] allow very efficient computation. We illustrate the use of the model with data from the U.S. in Section 5, finishing with a Discussion.

2. DESCRIPTION OF THE DATA

We are using the 'Mortality Detail Files' for the United States for the years 1959–1998. They have been obtained from the Population, Policy, and Aging Research Center (PPARC) at Duke University for the years 1959–1967 and downloaded from the Inter-university Consortium for Political and Social Research (ICPSR) at the University of Michigan `http://www.icpsr.umich.edu/` for the remainder of years (ICPSR Study 7632). The original collection of the data has been conducted by the National Center for Health Statistics.

The data record every single individual death by month and year of death, sex, age at death, cause of death and numerous other variables. Cause of death has been coded using the various revisions of the 'International Classification of Diseases' (ICD). In this paper we restrict ourselves to the analysis of 'diseases of the respiratory system', which constitute about 7 per cent of all causes of death for the entire observation period. Besides pneumonia, the major cause of death in this group, diseases such as asthma, influenza, and bronchitis belong to this category as well. Previous studies have shown the strong seasonal pattern of respiratory diseases, which supports the choice of this category of disease to serve as an illustrative example [3–5]. Table I gives an overview of the years and ICD codes used to extract deaths from respiratory diseases correctly.

3. THE MODULATION MODEL IN ONE DIMENSION

In Figure 1, we see strong (and varying) cyclical behavior, and we choose to model such seasonality using sine and cosine functions. In one dimension, a modulation model for the expected values μ_t for the observed series of counts y_t has the following structure:

$$\log(\mu_t) = \log(e_t) + v_t + f_t \cos(\omega t) + g_t \sin(\omega t) \tag{1}$$

where $t = 1, \ldots, T$. Here $\omega = 2\pi/p$, where p is the period (e.g. $p = 12$ when time is given in months and $p = 4$ when it is given in quarters). The smooth trend is represented by v, while f and g are smooth series that describe the local amplitudes of the cosine and sine waves. The model allows for exposures e, when the objective is to model mortality rates, instead of raw counts (in which case $e_t \equiv 1$). Gampe and Rau [1] and Rau [6] compared this model with a number of established models for seasonal data and found it to be superior, especially when seasonal patterns change over time.

Table I. ICD coding of respiratory diseases.

Revision	Years	ICD codes used
7	1959–1967	241, 470–527
8	1968–1978	460–519
9	1979–1998	460–519

DOI: 10.1002/sim

3.1. Splines and penalties

Technically, (1) is a combination of a varying-coefficient model (VCM), for f and g, and a generalized additive model (GAM), for v. VCMs allow the regression coefficients to vary smoothly with another variable and were first proposed by Hastie and Tibshirani [7]. GAMs were also proposed by Hastie and Tibshirani [8, 9]. VCMs and GAMs can be estimated by 'backfitting'; an iterative procedure that improves v, f and g in turn. Eilers and Marx [10] proposed a non-iterative procedure, based on P-splines [11], which they used to model seasonal counts. The P-spline approach is not only more elegant than backfitting but also computationally more efficient and offers complete control over convergence. It also allows straightforward computation of error bands and Akaike's information criterion (AIC) to optimize the amount of smoothing.

One-dimensional P-spline VCMs and GAMs use B-spline bases to model series such as v, f and g. The bases are 'rich', in the sense that they allow far more flexibility than needed. To increase smoothness, different penalties on the B-spline coefficients are introduced. Only the weights of the penalties are used to tune the model to the data, avoiding discussions about the number and placement of the knots of the equally spaced B-splines.

Now we will discuss the details of implementing the modulation model with P-splines. We start from the basis matrix $B=[b_{tj}]=[B_j(t)]$, where t is the time index of the observations, $t=1,\ldots,T$, and j, $j=1,\ldots,J$, indexes the B-splines. Figure 2 shows a basis with 13 cubic B-splines (some of them only partially visible). A smooth trend function is approximated and constructed by scaling and adding the B-splines: $v_t=\sum_j \alpha_j B_j(t)$ or $v=B\alpha$ in matrix–vector notation.

The modulation series f and g are constructed in the same way: $f=B\beta$ and $g=B\gamma$. The modulated series can be constructed directly if we scale the rows of the B-spline basis using

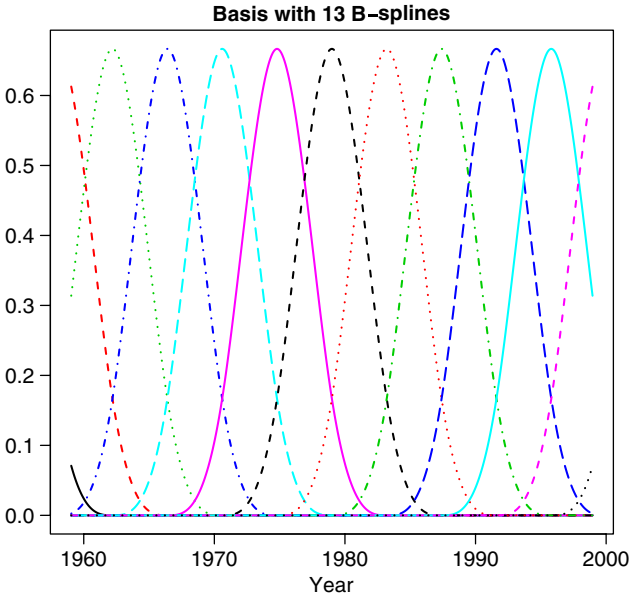

Figure 2. Illustration of a basis consisting of 13 cubic B-splines. The domain is divided into 10 regions, between knots. In each region, four B-spline segments are non-zero.

DOI: 10.1002/sim

element-wise multiplication:

$$h_t = f_t \cos(\omega t) = \cos(\omega t) \sum_j b_{tj} \beta_j \tag{2}$$

This can be expressed in matrix–vector notation if we introduce the matrix $C = \text{diag}\{\cos(\omega t)\}$: $h = CB\beta$. In a similar way, we construct the matrix SB with $S = \text{diag}\{\sin(\omega t)\}$). Expressing the smooth 'intercept' as $v = B\alpha$, the model can now be expressed as

$$\log(\mu) = \log e + B\alpha + CB\beta + SB\gamma = \eta \tag{3}$$

a generalized additive model with linear predictor η.

We minimize the penalized Poisson deviance

$$d^*(y;\mu) = 2 \sum_{t=1}^{T} y_t \log(y_t/\mu_t) + \lambda_1 \|D\alpha\|^2 + \lambda_2 \|D\beta\|^2 + \lambda_2 \|D\gamma\|^2 \tag{4}$$

Here D is a $(J-2) \times J$ matrix that forms second-order differences of a J-vector. If the coefficients α show relatively strong variation, their second differences $D\alpha$ will be relatively large and $\|D\alpha\|^2$ will be a large number. In (4) it functions as a penalty that directly imposes smoothness on α. The positive penalty hyper-parameter λ_1 balances smoothness against fit to the data. The penalties on β and γ operate in the same way. We allow different penalty weights for trend (λ_1) and modulation functions (λ_2).

Dropping the offset $\log e$ from now on, to simplify notation, the model in (3) can be expressed as

$$\eta = [B|CB|SB][\alpha'|\beta'|\gamma']' = \mathbf{B}\boldsymbol{\theta} \tag{5}$$

The penalties can be represented as $\boldsymbol{\theta}'\mathbf{P}\boldsymbol{\theta}$ with the block-diagonal matrix $\mathbf{P} = \Lambda \otimes D'D$ and $\Lambda = \text{diag}(\lambda_1, \lambda_2, \lambda_2)$. We use the boldface notation only to emphasize which matrices ($\mathbf{B}$ and $\mathbf{P}$) and vector ($\boldsymbol{\theta}$) refer to the compound basis for trend and amplitudes of cosine and sine.

To minimize the penalized deviance, the familiar iterated GLM scoring algorithm is being used.

$$(\mathbf{B}'\tilde{M}\mathbf{B} + \mathbf{P})\boldsymbol{\theta} = \mathbf{B}'(y - \tilde{\mu}) + \mathbf{B}'\tilde{M}\mathbf{B}\tilde{\boldsymbol{\theta}} \tag{6}$$

where $M = \text{diag}(\mu)$ and a tilde, as in $\tilde{\boldsymbol{\theta}}$ indicates an approximation to the solution. As can be seen, the penalties introduce only slight changes. Details are presented by Eilers and Marx [10].

3.2. Tuning the smoothness parameters

We have two tuning parameters, λ_1 and λ_2, for the smoothness of the model components. Although playing with them and looking at the resulting model fits are instructive, it is desirable to have a more or less objective procedure to set their values. Initially we measured model performance by AIC, which is a compromise between goodness of fit and model complexity and is defined as

$$\text{AIC} = d(y;\mu) + 2\text{ED} = d(y;\mu) + 2\,\text{trace}(H) \tag{7}$$

where ED is the effective model dimension and trace(H) approximates it. See Hastie and Tibshirani [9] for a discussion of effective dimensions and approximations. Alternative definitions involve H^2, computation of which is very unattractive if H is large. For one-dimensional data

DOI: 10.1002/sim

this is not a real problem, but we will encounter a matrix with 24 000 rows and columns in the following section. Upon convergence, the matrix $\mathbf{H}$ is the effective 'hat' matrix:

$$\mathbf{H}=\hat{M}^{1/2}\mathbf{B}(\mathbf{B}'\hat{M}\mathbf{B}+\mathbf{P})^{-1}\mathbf{B}'\hat{M}^{1/2} \tag{8}$$

It is a large matrix, with T rows and T columns. We only need the trace; hence, it is more efficient to use cyclical permutation:

$$\mathrm{ED}=\mathrm{trace}(H)=\mathrm{trace}[\mathbf{B}'\hat{M}\mathbf{B}(\mathbf{B}'\hat{M}\mathbf{B}+\mathbf{P})^{-1}] \tag{9}$$

where only a matrix of size $J\times J$ is computed.

The original idea was to search (using a linear grid on a log scale) for a reasonable set of λ's by monitoring AIC and choosing λ's with minimum AIC. Unfortunately, our data show severe over-dispersion. This is not a problem for estimating the trend and modulation surfaces for given values of the penalty parameters. However, it makes it impossible to optimize λ's using AIC based on a Poisson likelihood. First we give a quick impression of the size of the over-dispersion. Then we describe an effective combination of selective weights and quasi-likelihood.

For Poisson-distributed data, the deviance would approximately equal the effective degrees of freedom, that is, the number of observations (T) minus the effective model dimension (ED). However, we find values of $d(y;\mu)/(T-\mathrm{ED})$ of the order of 100 and higher. Separate calculations for the 12 months lead to values in the range from 10 to more than 100, the largest numbers occurring in winter (December–March). We call these the winter spikes. Flu epidemics and other cold-related infections occur mainly in these months and might be a major, but not the only, component. The severity of the spikes changes strongly from year to year; hence, we have here an obvious source of over-dispersion.

Catching individual winter spikes is not the purpose of analysis, rather we are interested whether long-term trends in seasonal fluctuations of death counts exist. Therefore, we choose to ignore them, by giving zero weights to the months December–March. The remaining eight months carry enough information to reliably estimate the model. As a bonus we hope to estimate the size of the epidemics (as the difference between observations and model-interpolated values) more reliably, because they are not allowed to increase the trend or the amplitude of the seasonal components. We will return to this issue in the Discussion.

This strategy reduces over-dispersion but does not eliminate it. Therefore, instead of the Poisson likelihood, we switch to the extended (log-)quasi-likelihood [12]:

$$Q(y;\mu)=d(y;\theta)/\phi+T\log\phi \tag{10}$$

where ϕ is the over-dispersion parameter, which is estimated as $\hat{\phi}=d(y;\mu)/(T-\mathrm{ED})$. With this modification, we compute a 'quasi-AIC' (*QIC*), as

$$\mathrm{QIC}=Q(y;\mu)+2\mathrm{ED}=T+\mathrm{ED}+T\log\hat{\phi} \tag{11}$$

Plate 1 shows the fit of the model to monthly totals of deaths counts of females (ages 51–100) caused by respiratory diseases. The upper panel shows the data: the trend function v and the fitted values $\hat{\mu}$. The lower panel shows the ratios $y/\hat{\mu}$, the modulated component $f_t\cos(\omega t)+g_t\sin(\omega t)$ and the combined amplitude $\rho=\sqrt{f^2+g^2}$.

This is a presentation on a logarithmic scale, which is natural, because we are modeling the logarithms of the expected values. On a linear scale, $e^{2\rho}$ might be interesting: it is approximately

DOI: 10.1002/sim

the ratio between the nearest peak and the nearest valley of the seasonal component. As we can see in Plate 1 (accounting for the 10-based logarithms), ρ varies roughly between 0.1 and 0.25. Accordingly the peak-to-trough ratio varies between approximately 1.25 and 1.65.

The model uses 23 cubic P-splines (20 interior knots with cubic B-splines), with $\lambda_1 = 10^{2.8}$ for trend and $\lambda_2 = 10^{1.6}$ for the seasonal component. These values were found after a search on a grid with steps of 0.2 on the log (base 10) scale.

4. MODULATION MODELS FOR INCIDENCE TABLES

We now generalize the modulation model to a two-dimensional table of counts $Y = [y_{ta}]$, indexed by both time, $t = 1, \ldots, T$, and age $a = 1, \ldots, A$. Again we consider the Poisson regression using a log link function

$$\log(\mu_{ta}) = \log(e_{ta}) + v_{ta} + f_{ta}\cos(\omega t) + g_{ta}\sin(\omega t) = \eta_{ta} \tag{12}$$

with $\mu_{ta} = E(y_{ta})$. In case that age- and time-specific information on exposures e_{ta} is available, it is included via an offset-term $\log(e_{ta})$. Otherwise this term is dropped from (12). The matrices $V = [v_{ta}]$, $F = [f_{ta}]$ and $G = [g_{ta}]$ represent smooth surfaces for trend and modulation along age and time. We model them by relatively many tensor products of B-splines and tune smoothness with difference penalties.

Figure 3 shows a portion (rank 9) of a full cubic tensor product B-spline basis. This figure is presented only to give a qualitative idea of the basis. In reality the density of the tensor product splines is two times larger (in both directions) than shown. Each tensor product has its own coefficient, which scales its height.

Tensor product B-spline bases can produce very general surfaces, however, one difficulty in using such a basis is the choice of the number and the placement of the tensor products in the indexing plane. To avoid a difficult numerical optimization, the P-spline approach takes two steps toward smoothness: (1) A relatively rich (gridded) tensor product basis is used (usually such

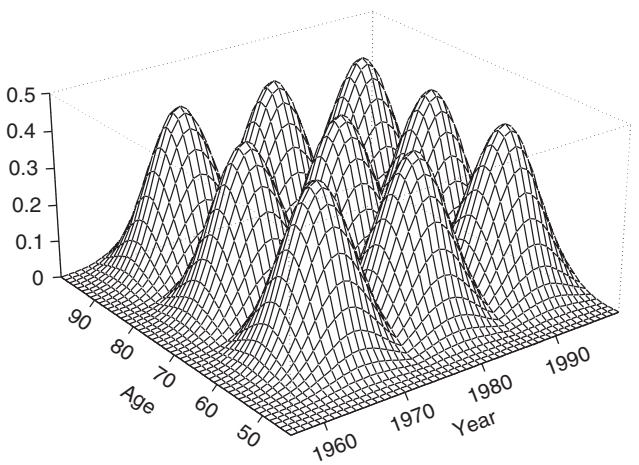

Figure 3. Portion of a tensor product B-spline basis.

DOI: 10.1002/sim

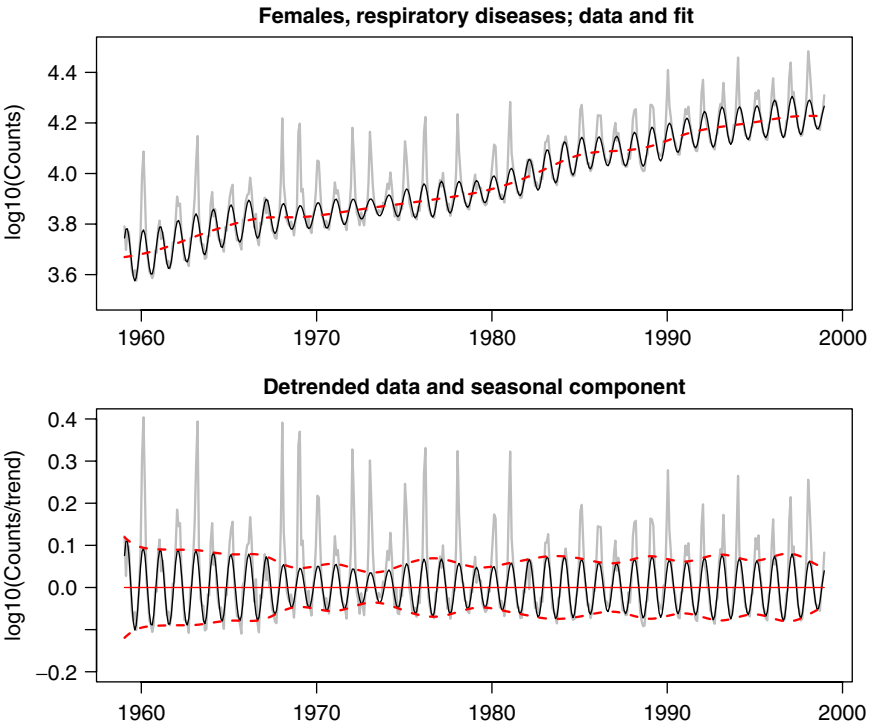

Plate 1. Illustration of the one-dimensional modulation model data as applied to female (ages 51–100) respiratory deaths. Upper panel: observed counts (thick gray line), model estimates (thin black line) and trend (red broken line). Lower panel: log of counts divided by the trend (thick gray line), seasonal component (thin black line) and modulation amplitude (red broken line).

DOI: 10.1002/sim

that $K \times L$ is manageable, say smaller than 1000) to purposely over-fit the estimated coefficient surfaces. (2) Penalties are imposed on the rows and columns of the coefficient matrix, such that the influence of each penalty is regulated by its own positive tuning parameter, λ. Hence smoothing is anisotropic.

Let $B = [b_{tj}]$ be the $T \times J$ B-spline basis on the time domain, which we used in the one-dimensional model. Let $\breve{B} = [\breve{b}_{ak}]$ be an $A \times K$ basis, but defined on the age domain. Let $\mathscr{A}$, $\mathscr{B}$ and $\mathscr{C}$ be the $J \times K$ matrices of tensor product coefficients for V, F and G. Then we can express (in a model without exposures)

$$V = B\mathscr{A}\breve{B}', \quad F = B\mathscr{B}\breve{B}', \quad G = B\mathscr{C}B' \tag{13}$$

with $V = [v_{ta}]$, $F = [f_{ta}]$, $G = [g_{ta}]$, and the model as

$$\log(M) = B\mathscr{A}\breve{B}' + CB\mathscr{B}\breve{B}' + SB\mathscr{C}B' \tag{14}$$

where $M = [\mu_{ta}]$, $C = \text{diag}\{\cos(\omega t)\}$ and $S = \text{diag}\{\sin(\omega t)\}$ as for the one-dimensional model.

Penalties are applied to the rows *and* the columns of $\mathscr{A}$, $\mathscr{B}$ and $\mathscr{C}$. Let D be the $(J-2) \times J$ matrix that forms second differences of J-vectors (such as D in the one-dimensional model), and $\breve{D}$ the $(K-2) \times K$ matrix that forms second differences of K-vectors. The penalty is

$$\text{pen} = \lambda_1 \|D\mathscr{A}\|_F + \lambda_2 \|D\mathscr{B}\|_F + \lambda_2 \|D\mathscr{C}\|_F + \breve{\lambda}_1 \|\mathscr{A}\breve{D}'\|_F + \breve{\lambda}_2 \|\mathscr{B}\breve{D}'\|_F + \breve{\lambda}_2 \|\mathscr{C}\breve{D}'\|_F \tag{15}$$

where $\|X\|_F$ indicates the Frobenius norm of X, i.e. the sum of the squares of all its elements. The penalized deviance is

$$d^*(Y; M) = 2 \sum_t \sum_a y_{ta} \log(y_{ta}/\mu_{ta}) + \text{pen} \tag{16}$$

The model has the structure of a generalized linear array model (GLAM) [2], an extension of generalized linear models to tensor product structures for data in multi-dimensional arrays. GLAMs offer very efficient computation [2, 13]. In our application, a speed improvement by at least a factor 20 was obtained, compared with the more traditional approach, expanding the bases with Kronecker products. We do not repeat the details of GLAM, but to illustrate the basic principle we sketch the core, using a simplified example, a model with only a smooth trend (and no seasonal component), that is, $\log M = V = B\mathscr{A}\breve{B}'$ and penalties on $\mathscr{A}$ as given in (15).

An equivalent vectorized (i.e. column-stacked) form is

$$\log(\mu) = \text{vec}(M) = (\breve{B} \otimes B)\alpha = (\breve{B} \otimes B)\text{vec}(\mathscr{A}) \tag{17}$$

a GLM with basis $\breve{B} \otimes B$. The penalties on columns and rows of the $K \times L$ coefficient matrix $\mathscr{A}$ are translated into equivalent penalties on α:

$$\alpha' P \alpha = \lambda_1 \|D\mathscr{A}\|_F + \breve{\lambda}_1 \|\mathscr{A}\breve{D}'\|_F = \alpha'[\lambda_1(\breve{I} \otimes D'D) + \breve{\lambda}_1(\breve{D}'\breve{D} \otimes I)]\alpha \tag{18}$$

where I is a $K \times K$ identity matrix and $\breve{I}$ an $L \times L$ identity matrix. The scoring algorithm solves

$$[(\breve{B} \otimes B)'\tilde{W}(\breve{B} \otimes B) + P]\alpha = (\breve{B} \otimes B)'(y - \tilde{\mu}) + (\breve{B} \otimes B)'\tilde{W}(\breve{B} \otimes B)\tilde{\alpha} \tag{19}$$

repeatedly, where $W = \text{diag}(\mu)$. The essence of a GLAM is that it computes $(\breve{B} \otimes B)'\tilde{W}(\breve{B} \otimes B)$ in a very efficient way, avoiding the Kronecker products. However, it solves the same equations with penalties on the vector α.

DOI: 10.1002/sim

The seasonal components can be handled in a similar way and added to the model, stacking $\beta = \mathrm{vec}(\mathscr{B})$ and $\gamma = \mathrm{vec}(\mathscr{C})$ under α and adding bases $\breve{B} \otimes (CB)$ and $\breve{B} \otimes (SB)$. This requires careful organization. For practical computation of the penalty matrices, it is in fact easier to rotate the matrices (rows for age and columns).

In a birds-eye view, one can simply approach a GLAM as a black box, allowing specification of a two-dimensional model similar in (14) with the corresponding penalties similar in (15), and returning a fitted model. That is also the way our software (written in R [14] and available on request) presents itself to the user.

Large values of λ enforce smoothness, whereas small values encourage roughness in either the row or column orientation. Similar in the one-dimensional case, we search for optimal values of λ's. Over-dispersion strikes here too, and we take the same countermeasures: excluding December–March from the fit and monitoring QIC instead of AIC.

We implement the search in a greedy way: each of the four λ's is changed in turn, by one step up and one step down on the grid. The step that gives the largest improvement, in *QIC* is kept. If there is no improvement, the current value is kept.

5. APPLICATION

For the female death counts due to respiratory diseases, all three surfaces (intercept, sine- and cosine-varying coefficients) were constructed on the time–age grid using a basis with $(10+3) \times (10+3)$ tensor products. λ's could take on values on a grid with width 0.2 (on the 10 base log scale) and were changed in turn to test for possible downhill moves. Result: $\lambda_1 = 10$, $\lambda_2 = 1$; $\breve{\lambda}_1 = 10^{0.6}$ and $\breve{\lambda}_2 = 10^4$.

The effective dimension of the model is 167.8, much smaller than the number of model parameters ($3 \times 13^2 = 507$), showing the effect of the penalties. Plate 2 shows the varying intercept term in the upper left panel. In the two following panels, the cosine- and sine-varying coefficients are transformed into amplitude and phase surfaces. The phase surface has been further mapped from the interval $(0,\ 2\pi)$ into (0, 365) and thus has a direct interpretation with the days in a year: January 1 (1)–December 31 (365). The final image shows the Pearson residuals. Their range illustrates the amount of over-dispersion. Large positive residuals, corresponding to flu epidemics, show up as bright vertical stripes. Interestingly, cohort effects are clearly visible as diagonal stripes running from the lower left to the upper right of the image. The marginal averages of amplitude and phase for `age` and `time` are depicted in Figure 4.

We can see in the upper left panel of Plate 2 that the number of deaths increases with age, peaking at about age 85, resembling closely the typical distribution of deaths over age. Over time (i.e. on the horizontal axis), two interesting features can be recognized, which can be traced back to two different developments: first, the number of deaths is increasing at ages 65 and higher. Most likely, this is the consequence of a compositional change: the number of women aged 65 and older increased from 8.8 million in 1959 to more than 20 million in 1998 (own calculations, based on—partly adjusted—official data as published in the Human Mortality Database, `http://www.mortality.org`). At the same time, we can see that the number of deaths from respiratory diseases below age 65 remained relatively constant or even decreased—despite a population growth of more 63 per cent of that age group ($44 \leqslant \mathrm{age} \leqslant 65$) between 1959 and 1998. This can serve as an indicator for the progress made against mortality in general by shifting deaths

DOI: 10.1002/sim

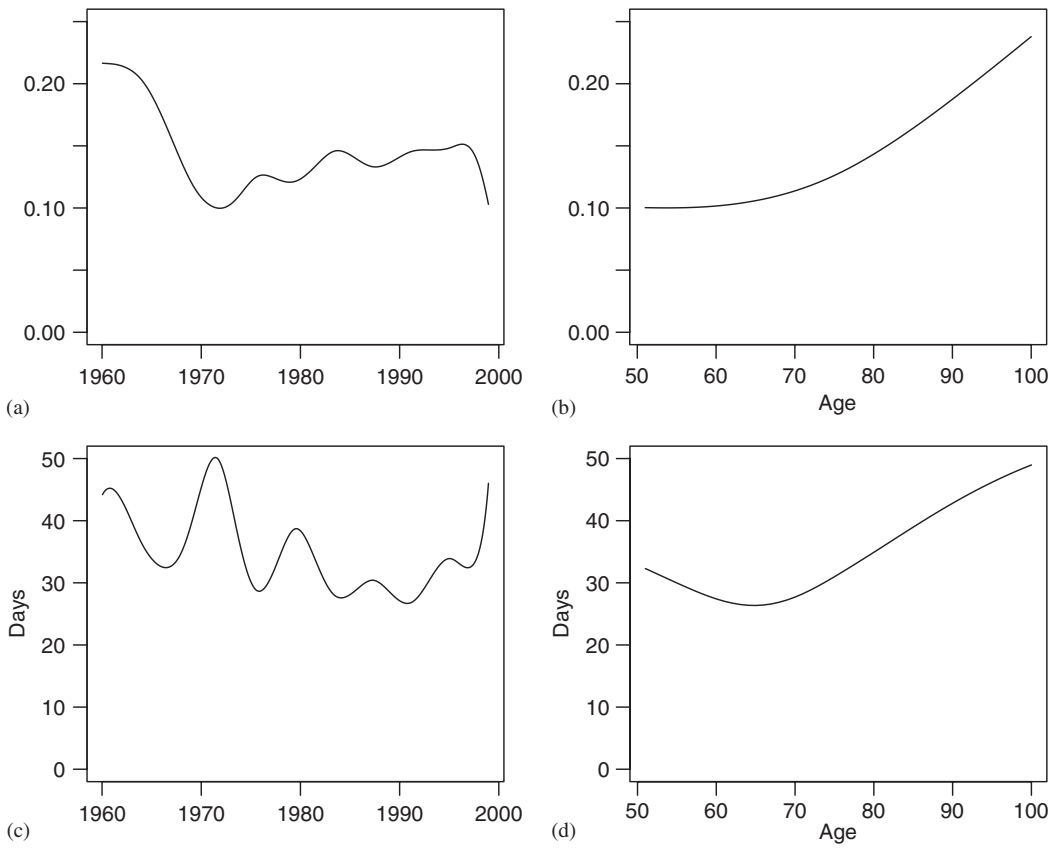

Figure 4. Marginal plots of the amplitude and phase of the seasonal effects: (a) marginal amplitude profile over time; (b) marginal amplitude profile over age; (c) marginal phase profile over time; and (d) marginal phase profile over age.

to a later age and possibly, more specifically, successful efforts to prevent deaths from respiratory mortality.

The upper right panel depicts the changes in the seasonal amplitude over age and time. We are not sure whether the sudden increase in seasonality in the early 1970s can be traced back to data problems (the data for 1972 contained not the whole population but a 50 per cent random sample) or the impact of smaller influenza epidemics in the winters 1971/1972 and 1972/1973 [15]. In general, we can see that the annual fluctuations decreased until the mid-1970s. Since then, no more improvement has been made and one can even detect signs for increasing seasonality at ages below 80. Above age 80, seasonality continuously decreased since the 1970s. Similar patterns have been found previously in the one-dimensional case for corresponding age-groups [6, 16]. Seretakis *et al*. [17] discovered a comparable development for coronary heart disease. According to that article, the decrease and the subsequent increase can be traced back, first, to the spread of central heating, which helped reducing excess winter mortality and dampened the seasonal amplitude. The increased usage of air conditioning since the mid-1970s may have helped to reduce

DOI: 10.1002/sim

absolute mortality during summer, resulting in larger annual fluctuations in mortality—despite a lower overall level in mortality.

The lower left panel indicates the day of the year when mortality peaked. This timing varied throughout our 40 years of observation period. Our results indicate that the peak occurred between the end of January (30 days) and sometime before the middle of March. We interpret these findings to be less an outcome of technological advance as we did for the amplitude but rather the varying onset of cold spells triggering respiratory infections.

6. DISCUSSION

Our modulation model for incidence tables allows detailed quantitative description of many aspects of seasonality. The intercept surface captures the overall trend in event counts, as a function of year and age. After conversion to polar coordinates, the modulation surfaces tell us how the relative strength of seasonality changes with year and age, as well as changes in the phase, i.e. the time of the year, in which peaks occur. This can be a powerful tool for detailed epidemiological or demographic analysis. In this paper we studied mortality data, but of course the model can be applied to other types of events, such as disease incidence.

Because the estimated surfaces are smooth by design, they allow the computation of derivatives, with respect to age or time, for specialized research in which rates of change are of interest.

The combination of tensor products and discrete penalties is very effective for multi-dimensional smoothing. One can imagine extensions adding gender, with shared seasonal components, but differing trends for men and women. One can also envision extensions to more dimensions. The U.S. death counts can be subdivided by state. This gives the opportunity to introduce east–west or north–south spatial location for building a three-dimensional incidence table. A really challenging extension is to build four-dimensional models, in which surfaces live in the longitude–latitude–time–age space.

On the other hand, it will be worthwhile to study simplifications of the two-dimensional model. If a surface shows little interaction between time and age, it will be easier to understand and more parsimonious to be described by means of two additive functions, one for age and the other for time. This would result in models that contain combinations of one-dimensional and tensor-product P-spline components.

The Poisson model implies that the variance is equal to the expected value. One almost never observes this in real-life data, where over-dispersion is very common. Our data are no exception, showing ratios of variance to mean of the order of 10 to over 100. This is a serious obstacle to automatic tuning of penalties. In general, over-dispersion will then lead to surfaces that are more flexible than one would expect from prior knowledge.

Our solution to over-dispersion has two components: elimination of the most extreme months (December–March) and the use of quasi-likelihood. Eliminations of specific months is done by zero/one weights. The remaining months carry enough information to reliably estimate the model, although essentially the ‘summer valleys’ now determine the amplitude of the seasonal component and the assumption of a sinusoidal pattern is critical. We expect that our estimates of the winter spikes will be improved this way, because they are not influencing seasonality estimates. A possible refinement would be to model spikes as ‘shocks’: a smooth age distributions of variable shape and size for every individual month [18]. This is not of primary importance in the context of this paper; hence, we do not pursue this topic further here.

DOI: 10.1002/sim

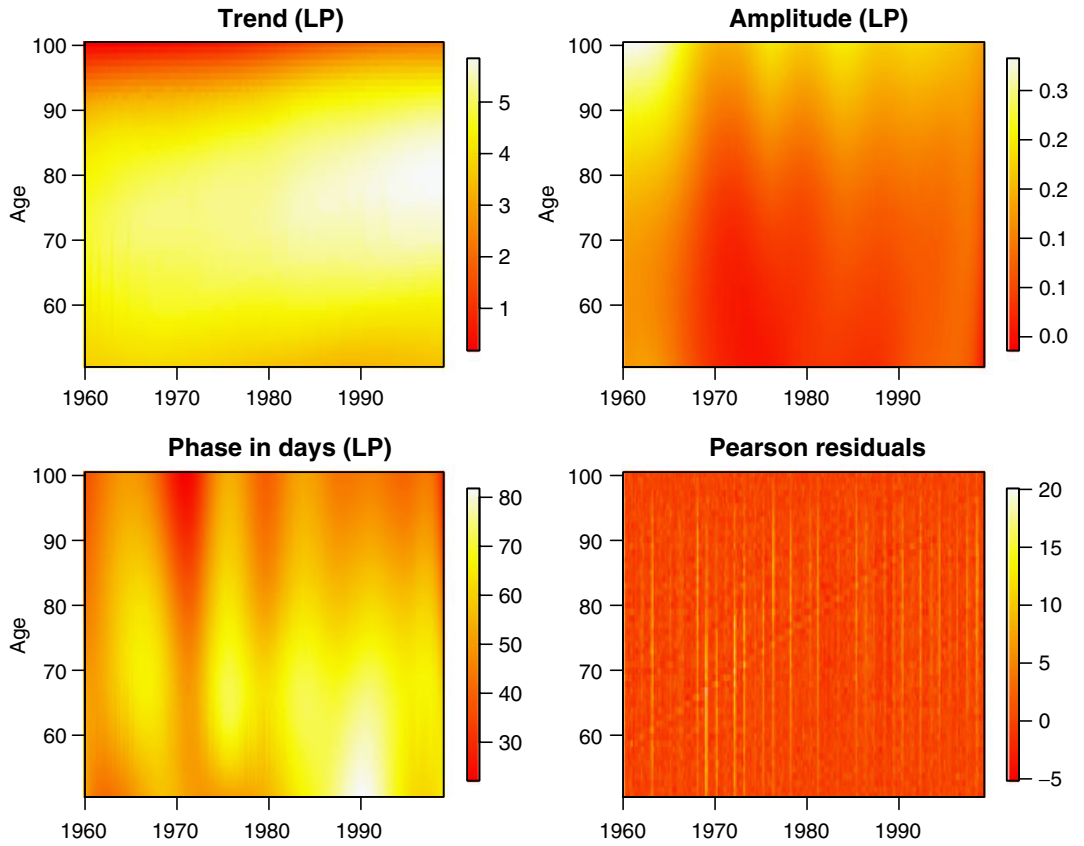

Plate 2. Image plot of the varying intercept term or overall trend (upper, left); amplitude and phase (month) resulting from the two-dimensional seasonal effects, (upper, right) and (lower, left), respectively; and the Pearson residuals (lower, right).

DOI: 10.1002/sim

Our use of quasi-likelihood assumes that the variance is proportional to the mean, with a ratio that is estimated from the data. Other options are available. The first is to use the negative binomial distribution [19]. An interesting second option is to introduce 'individual deviance effects', one for each cell of the incidence table [20]. This leads to more parameters than there are observations, but a ridge penalty (which is optimized) constrains the size of the effects allowing identification.

ACKNOWLEDGEMENTS

The research of Brian Marx was partially supported by a National Science Foundation grant and an AFOSR grant IIS-0326387, FA9550-05-1-0454.

REFERENCES

1. Gampe J, Rau R. Seasonal variation in death counts: P-spline smoothing in the presence of overdispersion. *Proceedings of the 19th International Workshop on Statistical Modelling*, Biggeri A, Dreassi E, Lagazio C, Marchi M (eds). Florence University Press: Florence, 2004; 405–409.
2. Currie ID, Durban M, Eilers PHC. Generalized linear array models with applications to multidimensional smoothing. *Journal of the Royal Statistical Society*, *Series B* (*Methodological*) 2006; **68**(2):259–280.
3. Rosenwaike I. Seasonal variation of deaths in the United States, 1951–1960. *Journal of the American Statistical Association* 1966; **61**(315):706–719.
4. Eurowinter Group. Cold exposure and winter mortality from ischaemic heart disease, cerebrovascular disease, respiratory disease, and all causes in warm and cold regions of Europe. *Lancet* 1997; **349**(9062):1341–1346.
5. van Rossum CTM, Shipley MJ, Hemingway H, Grobbee DE, Mackenbach JP, Marmot MJ. Seasonal variation in cause-specific mortality: are there high-risk groups? 25-year follow-up of civil servants from the first Whitehall study. *International Journal of Epidemiology* 2001; **30**(5):1109–1116.
6. Rau R. *Seasonality in Human Mortality. A Demographic Approach*. Springer: Heidelberg, 2007.
7. Hastie T, Tibshirani R. Varying-coefficient models. *Journal of the Royal Statistical Society*, *Series B* (*Methodological*) 1993; **55**(4):757–796.
8. Hastie T, Tibshirani R. Generalized additive models: some applications. *Journal of the American Statistical Association* 1987; **82**(398):371–386.
9. Hastie T, Tibshirani R. *Generalized Additive Models*. Chapman & Hall: London, 1990.
10. Eilers PHC, Marx BD. Generalized linear additive smooth structures. *Journal of Computational and Graphical Statistics* 2002; **11**(4):758–783.
11. Eilers PHC, Marx BD. Flexible smoothing with *B*-splines and penalties (with comments and rejoinder). *Statistical Science* 1996; **11**(2):89–121.
12. McCullagh P, Nelder JA. *Generalized Linear Models* (2nd edn). Chapman & Hall: London, 1989.
13. Eilers PHC, Currie ID, Durban M. Fast and compact smoothing on multi-dimensional grids. *Computational Statistics and Data Analysis* 2006; **50**(1):61–76.
14. R Development Core Team. *R*: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Austria, Vienna, 2007, ISBN 3-900051-07-0.
15. Barker WH. Excess pneumonia and influenza associated hospitalization during influenza epidemics in the United States, 1970–78. *American Journal of Public Health* 1986; **76**(7):761–765.
16. Feinstein CA. Seasonality of deaths in the U.S. by age and cause. *Demographic Research* 2002; **6**(17):469–486.
17. Seretakis D, Lagiou P, Lipworth L, Signorello LB, Rothman KJ, Trichopoulos D. Changing seasonality of mortality from coronary heart disease. *Journal of the American Medical Association* 1997; **278**(12):1012–1014.
18. Kirkby JG, Currie ID. Smooth models with period shocks, 2007; submitted.
19. Thurston SW, Wand MP, Weincke JK. Negative binomial additive models. *Biometrics* 2000; **56**(1):139–144.
20. Perperoglou A, Eilers PHC. Penalized random individual deviance effects, 2006; submitted.

DOI: 10.1002/sim