ISSN 1440-771X

MONASH University

Department of Econometrics and Business Statistics

http://www.buseco.monash.edu.au/depts/ebs/pubs/wpapers/

Two-dimensional smoothing of mortality rates

Alexander Dokumentov, Rob J Hyndman

November 2013

Working Paper 26/13

Two-dimensional smoothing of mortality rates

Alexander Dokumentov
Department of Econometrics and Business Statistics,
Monash University, VIC 3800
Australia.
Email: alexander.dokumentov@monash.edu

Rob J Hyndman
Department of Econometrics and Business Statistics,
Monash University, VIC 3800
Australia.
Email: rob.hyndman@monash.edu

26 November 2013

JEL classification: C14, C33, J11

Two-dimensional smoothing of mortality rates

Abstract

We propose three new practical methods of smoothing mortality rates (the procedure known in demography as graduation) over two dimensions: age and time. The first method uses bivariate thin plate splines. The second uses a similar procedure but with lasso-type regularization. The third method also uses bivariate lasso-type regularization, but allows for both period and cohort effects. Thus the mortality rates are modelled as the sum of four components: a smooth bivariate function of age and time, smooth one-dimensional cohort effects, smooth one-dimensional period effects and random errors. Cross validation is used to compare these new methods of graduation with existing approaches.

Keywords: Mortality rates, nonparametric smoothing, graduation, cohort effects, period effects.

1 Introduction

Mortality rates are used to compute life tables, life expectancies, insurance premiums, and other items of interest to demographers and actuaries. However, mortality rates are noisy, and so it is useful to smooth them in order to obtain better estimates with smaller variance.

Let $M_{x,t}$ denote an observed mortality rate $M_{x,t}$ for a particular age x and for a particular year t, defined as $M_{x,t} = D_{x,t}/E_{x,t}$, where $D_{x,t}$ is the number of deaths during year t for people who died being x years old, and $E_{x,t}$ is the total number of years lived by people aged x during year t. In practice, $E_{x,t}$ is usually approximated by the mid-year population of people aged x in year t.

As we can see from the definition, mortality rates are two dimensional: one dimension is time and the other dimension is age. Our aim is to smooth the bivariate surface in both the age (x) and time (t) dimensions, and to allow occasional period effects (along x for a specific t) and cohort effects (along $x = t + k$ for a specific k).

Several nonparametric smoothing (or graduation) approaches have been proposed in the past (e.g., Schuette, 1978; Hyndman and Ullah, 2007), but none to our knowledge that exploit all the features of mortality rates. The combination of bivariate smoothing with cohort and period effects is not possible using any existing smoothing methods. In this paper, we propose several new bivariate smoothing methods for mortality data, the last of which also allows for both period and cohort effects. We compare our new methods, and some existing methods, using a cross-validation procedure.

To stabilize the variance of the noise, and to make the smoothness more uniform, it is necessary to take logarithms: $m_{x,t} = \log(M_{x,t})$. Moreover, features of the data for low mortality rates (for ages from 1 to 40) have clearer shape after taking logarithms. Taking logarithms also makes sense from the point of view that different factors affect mortality in a multiplicative manner, and after taking logarithms the effects are then additive. Figure 1 shows log mortality rates for females in France from 1950 to 1970. The data is taken from the demography package for R (Hyndman, 2012); it was originally sources from the Human Mortality Database (2008).

After taking logs the following features of the log mortality surface become evident:

- In the age dimension, the log mortality abruptly decreases for the early ages and reaches a minimum at age about 10 years. There is a “bump” around age 20, after which it increases almost linearly to the very old ages.

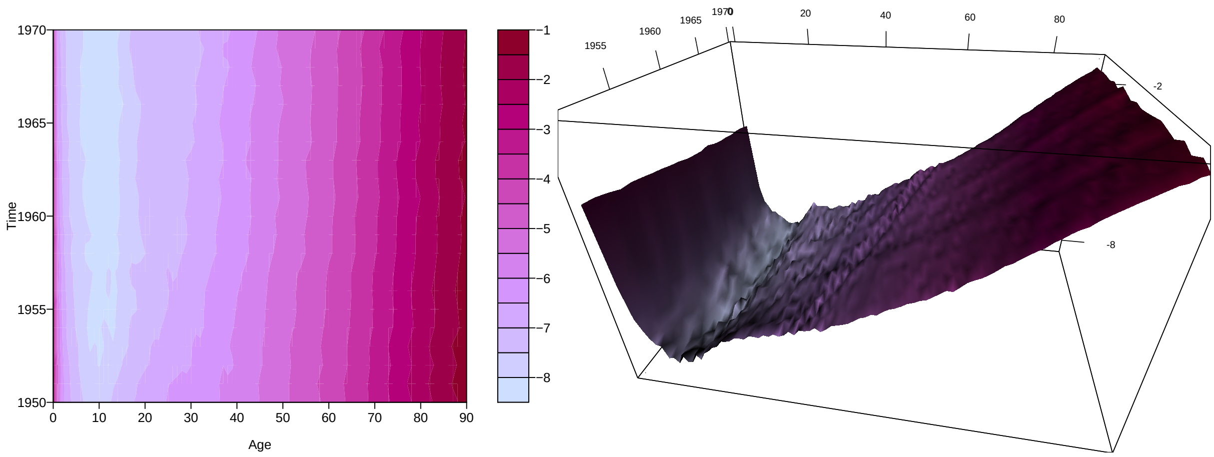

Figure 1: *Natural logarithm of French female mortality rates.*

- For almost every age, mortality has decreased with time over the period of these data. It decreases more steeply (on the log scale) for younger ages then for older ages.
- The highest mortality is reached for older ages.
- There are diagonal patterns due to cohort effects. While these are not easy to see in Figure 1, later graphs will highlight their existence. Such patterns may be due to some extreme events experienced by a cohort of people when they were born, or in their early years.
- There are horizontal patterns (for a fixed year) due to period effects. Such patterns are usually due to some extreme environmental event such as a war or pandemic, affecting all people (with different magnitude) during a particular year. These are more evident during the periods 1914–1918 and 1939–1945 than in the years shown in Figure 1.

Each of these features should be preserved when smoothing mortality data because they represent "real" effects.

A classic method for smoothing mortality rates is to use a parametric non-linear function of age dependent on only a few coefficients. Heligman and Pollard (1980) were among the first to propose a formula which covers all living ages. Having such a function, the smoothing can done by simply estimating the coefficients using, for example, least squares. This approach is one-dimensional (in x) but can be extended to two dimensions relatively easily. We do not follow this path because we do not want to be restricted to features easily described parametrically; we want the freedom to model any data features that appear. Forfar et al. (1988) discusses some alternative one-dimensional parametric approaches for smoothing mortality rates.

Hyndman and Ullah (2007) proposed a one-dimensional non-parametric approach for smoothing, based on penalized regression splines with a monotonicity constraint. We discuss this approach in more detail in Section 2.

In normal circumstances (when wars and pandemics are relatively uncommon), it is reasonable to assume that mortality rates are smooth in two dimensions: time and age, and so the main idea behind bivariate smoothing is to use both dimensions for estimating mortality at some two-dimensional point defined by time and age. The one-dimensional approach is (usually) to find a smooth function $f_t(x) = \mathrm{E}[m_{x,t}]$ for each t. In contrast, bivariate smoothing looks for a smooth bivariate function $f(x,t) = \mathrm{E}[m_{x,t}]$. By allowing the assumption of smoothness in both dimensions, better performance should be possible due to the additional information included in the estimation.

Currie et al. (2004) proposed using two-dimensional P-splines for smoothing and forecasting mortality rates. Camarda (2012) implemented this approach along with a one-dimensional version in the R package MortalitySmooth.

Camarda et al. (2010) proposed to use special bases for P-splines to fit logarithms of mortality rates. This helps to overcome problems related to the abrupt changes in data at early ages (which is the obstacle for the method described in Currie et al. (2004)). Unfortunately we do not have access to any implementation of this method.

To solve the problem of abrupt changes in the data, we prefer the L_1 norm in place of the L_2 norm. We use it for regularization as well as a measure of the closeness of the approximation to the data. The L_1 norm is often used because it is robust when the data contain outliers (see, for example, Schuette, 1978). While this is a useful feature, the main reason we adopt the L_1 norm here is different.

The L_2 norm penalizes smaller errors much less than bigger errors; this leads to a situation when relatively small errors are mixed with much greater "features", and the "features" are then penalized more than errors. Consequently features are distorted while attempting to reduce noise. In such cases, the L_1 norm behaves better than the L_2 norm because it is less harsh to outstanding "features".

In the next five sections, we present describe five smoothing algorithms, the last of which is our preferred procedure:

1. The Hyndman and Ullah (2007) algorithm (Section 2), is implemented in the demography R package (Hyndman, 2012), and smooths mortality rates only in the age dimension. This algorithm is presented for comparison only.

2. The Camarda (2012) algorithm (Section 3), is implemented in the MortalitySmooth package, and smooths mortality rates in both dimensions, although it is only designed to work for ages greater than 10. This algorithm is also presented for comparison only.

3. Section 4 describes a new algorithm that uses two-dimensional thin plate splines and therefore uses both dimensions — time and age — for smoothing.

4. Section 5 describes another new algorithm that uses Lasso-type regularization and also uses both dimensions for smoothing. This algorithm copies thin plate splines in many ways, but it uses the L_1 norm instead of the L_2 norm.

5. The last algorithm (Section 6) also uses Lasso-type regularization and both dimensions for smoothing, but incorporates cohort and period effects. This improves the performance and also provides greater insight to the structure of the mortality data.

The minimisation problem of the last two algorithms can be reduced to quantile regression minimisation problems (see Section 5), and therefore we use quantile regression software (Koenker, 2013) to implement these algorithms.

These five algorithms are compared in Section 7 using a cross-validation procedure. Finally, we provide some discussion and conclusions in Section 8.

2 Hyndman-Ullah (2007) method

Hyndman and Ullah (2007) proposed a method for smoothing mortality rates across ages in each year. The method is intentionally one-dimensional to allow for a forecasting procedure, applied after smoothing, that takes into account variation in the time dimension. An implementation of the method is provided in the demography package for R (Hyndman, 2012).

This smoothing method uses constrained weighted penalized spline regression applied independently for each year. Weighted penalized spline regression involves calculating a vector β which minimizes the expression

$$\|w(y - X\beta)\|^2 + \lambda^2 \beta^T D\beta,$$

where y is a vector of observations, X is a matrix representing linear spline bases, $D = \text{diag}(0,0,1,1,...,1)$ is a diagonal matrix, w is a vector of weights and λ is a parameter (see, for example, Ruppert et al., 2003).

In the case of smoothing mortality rates, observations in year t are given by $y_i = m_{x_i,t}$ for age group x_i years old. The weights w_i are taken as the inverse of the estimated variances of y_i. Assuming deaths follow a Poisson distribution, and using a Taylor series approach, Hyndman and Booth (2008) estimate the variance of y_i as $\sigma_i^2 \approx (E_{x_i,t} M_{x_i,t})^{-1}$, where $E_{x_i,t}$ is the mid-year population of people aged x_i years in year t.

Moreover such splines are constrained to ensure that the resulting function $f(x)$ is monotonically increasing for $x \geq c$ for some c (for example 50 years). Hyndman and Ullah (2007) use a modified version of the method described in Wood (1994) to implement this constraint.

The result of this approach is a surface which is smooth in the age dimension but still "wiggly" in the time dimension (Figure 2).

The residuals (Figure 3) show some serial correlation for early ages as well as diagonal patterns which are cohort effects (effects related to people born in the same year). For example Figure 4 reveals some serial correlation of the residuals for ages 1 and 2. However, it is clear that the residuals do not show any horizontal patterns due to period effects. This is expected, because separate smoothing has been done independently for each year.

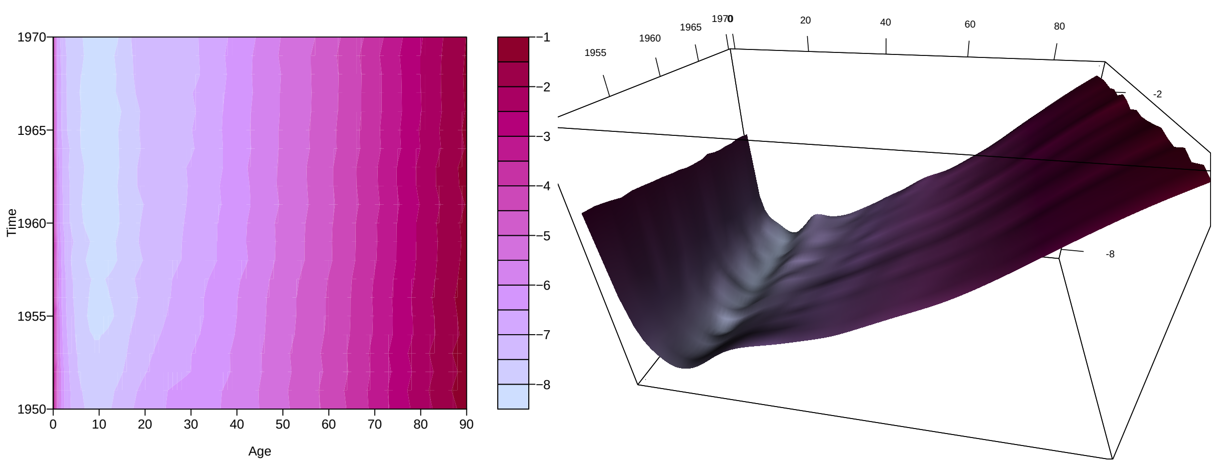

Figure 2: *French female mortality rates smoothed by Hyndman and Ullah (2007) method.*

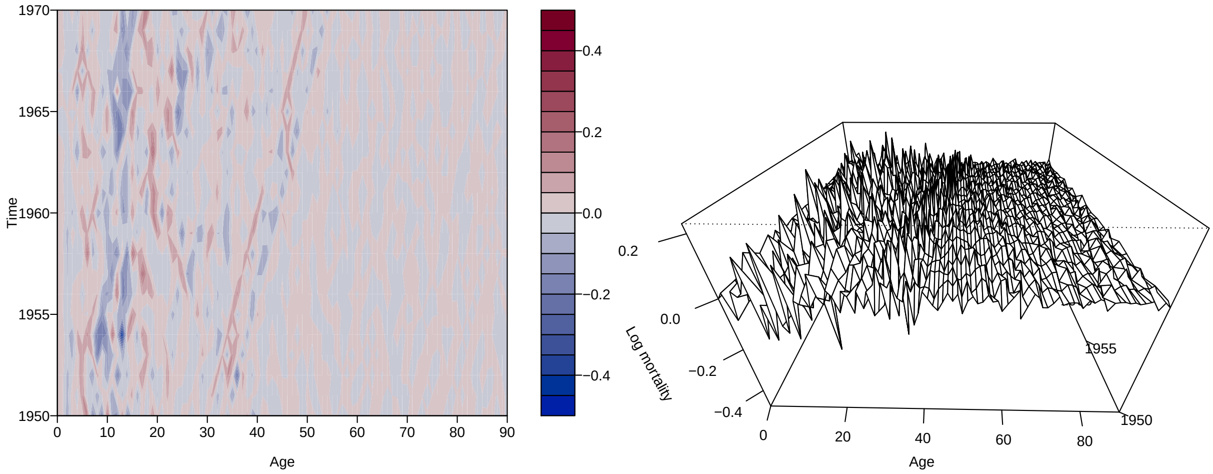

Figure 3: *French female mortality rates residuals after smoothing by Hyndman and Ullah (2007) method.*

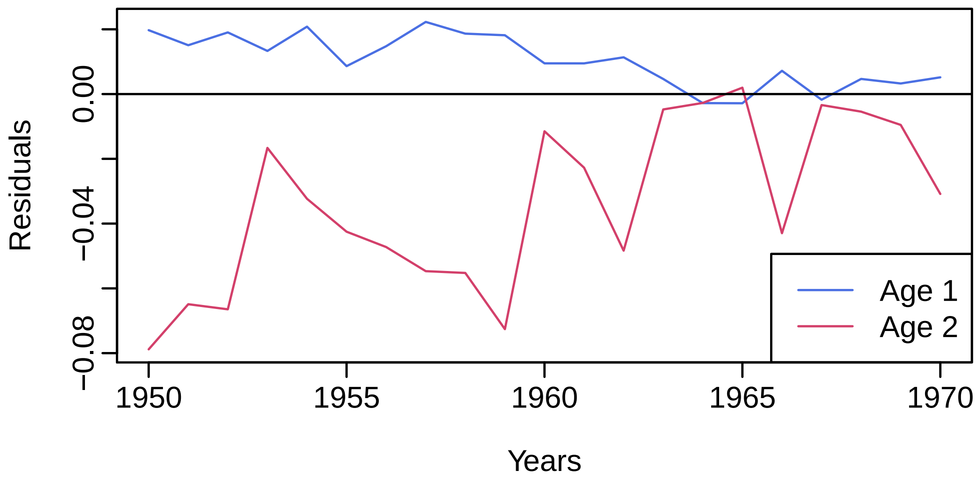

Figure 4: *French female mortality rates residuals for ages 1 and 2 after smoothing in age dimension.*

3 Camarda (2012) method

Camarda (2012) implements a two-dimensional method using P-splines for smoothing mortality rates in the MortalitySmooth package for R. For ages 0 to 10, the result of smoothing is notably biased (Figures 5 and 6). As we can see in Figure 6 the residuals are serially correlated for early ages. Also diagonal patterns due to cohort effects and horizontal patterns due to period effects are visible.

Camarda (2012) acknowledge that the method was not designed for smoothing of the youngest ages. Nevertheless, we still use it as the comparisons clearly show what problems we faced and have overcome.

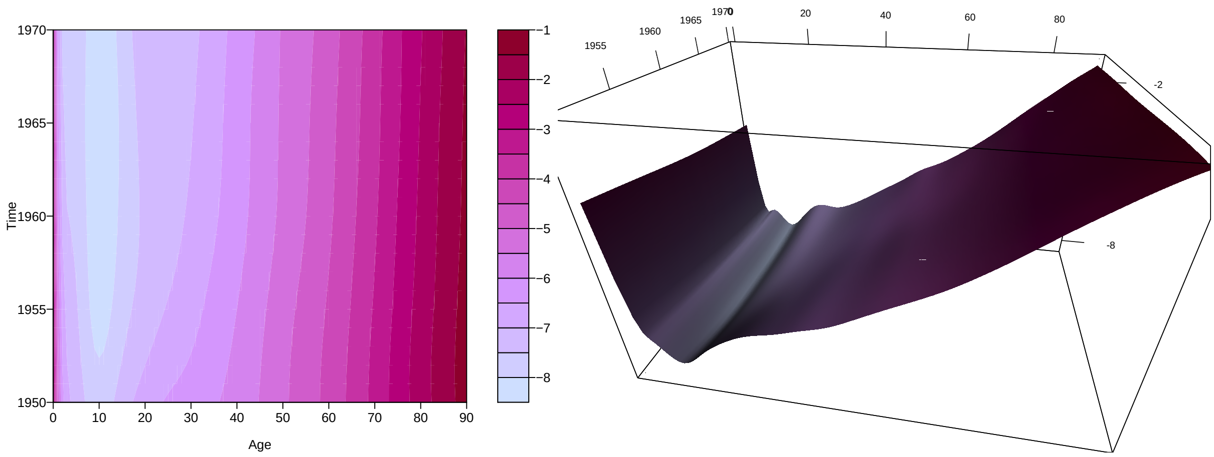

Figure 5: *French female mortality rates smoothed by Camarda (2012) method.*

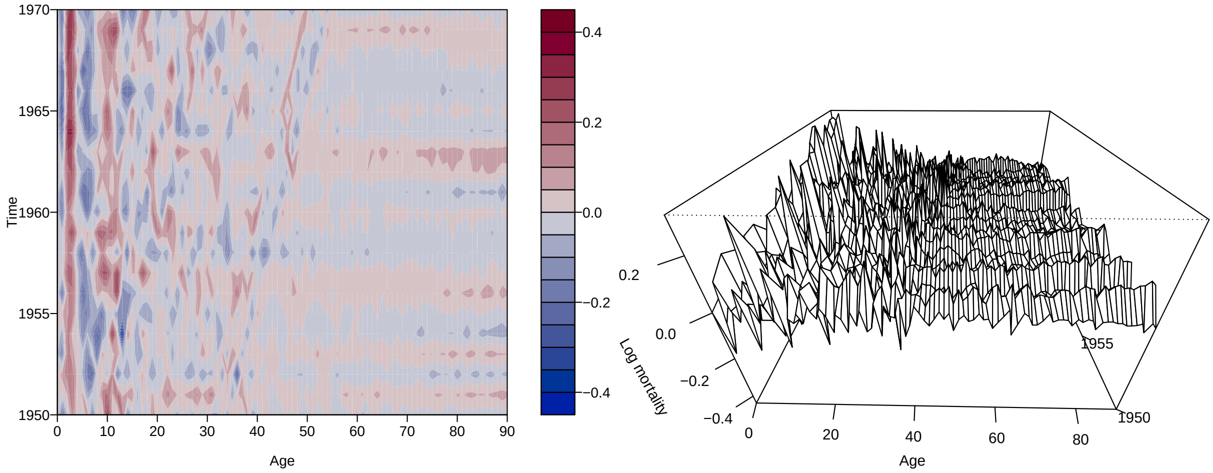

Figure 6: *French female mortality rates residuals after smoothing by Camarda (2012) method.*

4 Boosted thin plate splines

Thin plate splines work well in many cases (Wood, 2006). However, in case of mortality data, direct application of thin plate splines (or adaptive thin plate splines where flexibility varies) to the logs of mortality rates does not lead to precise and unbiased results, especially for early ages. This is the same problem that the Camarda (2012) method experiences for early ages.

We speculate that the reasons for the problems are twofold:

(a) The log mortality rate surface is very abrupt at early ages, and twisted along the time dimension due to the rapid decrease in mortality for younger ages; and

(b) Thin plate splines penalize big errors much heavier than small errors.

This leads to the situation when abrupt jumps in the mortality data generate errors in the proximity of the jumps, causing unsatisfactory performance of thin plate splines over the abrupt surface.

Consequently, improved performance is possibly by first "flattening" the surface and then applying adaptive thin plates splines to the flattened surface before reversing the flattening procedure.

The inputs for the flattening procedure are log mortality rates from year t_0 to year t_1. We denote these inputs as $m_{x,t}$, where x is the age group and t is the time. We split the data into two halves with the earliest years of observation in the first set of data, and the later years in the second set of data. We then estimate a very crude surface based on the median log mortality by age for these two sets of data. This crude surface is subtracted from the data, and the resulting differences are then smoothed using thin plate splines.

This has some similarities to the "twicing" procedure proposed in Tukey (1977). It can also be considered as rudimentary two-step "boosting" (see, for example, Bühlmann and Yu, 2003).

We now outline the procedure.

1. The mid-point year is at time $t_{0.5} = (t_0 + t_1)/2$. Data where $t \le t_{0.5}$ will belong to set one, and the remaining data will belong to set two. The age-specific medians for the two halves of the data are given by

$$m^*_{x,0} = \underset{t_0 \le t \le t_{0.5}}{\text{Median}}(m_{x,t}) \qquad \text{and} \qquad m^*_{x,1} = \underset{t_{0.5} < t \le t_1}{\text{Median}}(m_{x,t}).$$

2. Each of the two sets of medians — for older and recent years — is smoothed using a standard smoothing method. Let us denote the resulting curves by $m_0(x)$ and $m_1(x)$.

3. These two curves are used to create an approximate, but very smooth surface for years from t_0 to year t_1, by connecting points corresponding to same age in the smoothed curves by straight lines:

$$s(x,t) = \frac{(3t_1 + t_0 - 4t)m_0(x) + (4t - 3t_0 - t_1)m_1(x)}{2(t_1 - t_0)}.$$

4. The smooth surface $s(x,t)$ is subtracted from the original data to give $r_{x,t} = m_{x,t} - s(x,t)$. These flattened values lie on a surface that is not abrupt at early ages and is not twisted.

5. The flattened values $r_{x,t}$ are smoothed using adaptive thin plate splines to give $r(x,t)$.

6. The last step is to "un-flatten" the result obtained in the previous step: $f(x,t) = r(x,t) + s(x,t)$.

The resulting smooth surface is shown in Figure 7. The residuals, $m_{x,t} - f(x,t)$, are shown in Figure 8. They now look more uncorrelated than in Figures 3 and 6. Remaining diagonal patterns due to cohort effects, and remaining horizontal patterns due to period effects, are still clearly visible.

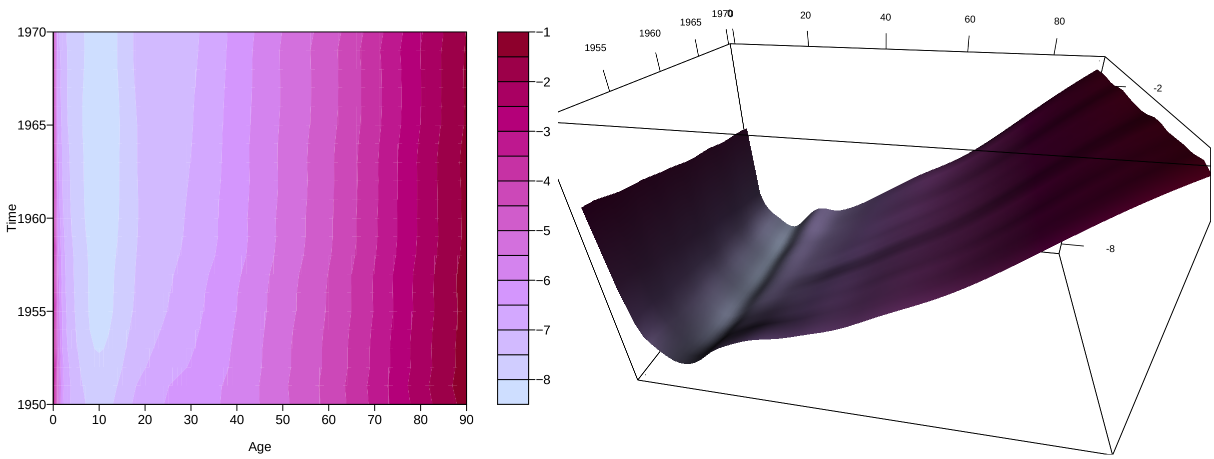

Figure 7: *French female mortality rates smoothed with thin plate splines.*

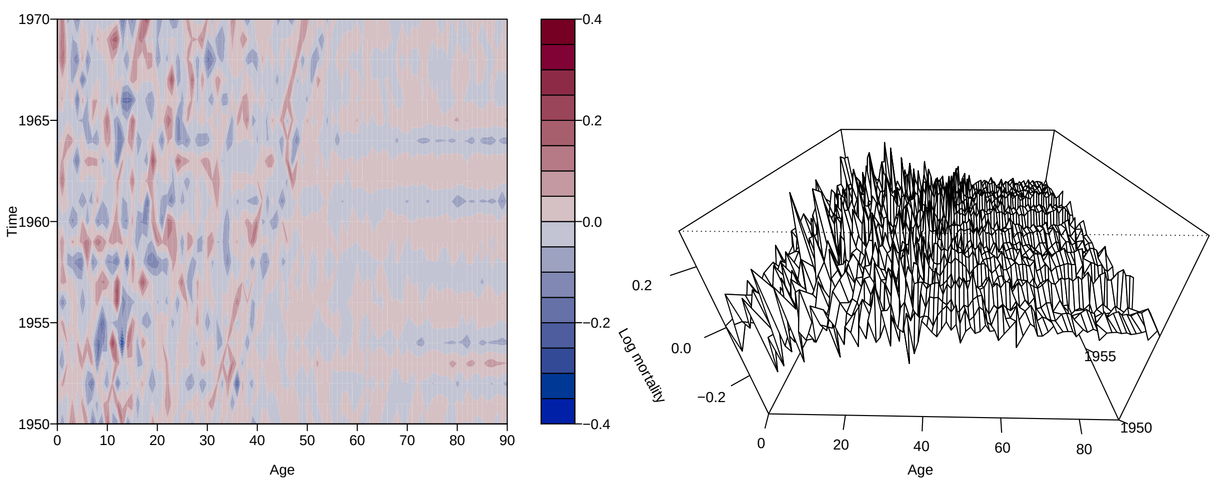

Figure 8: *French female mortality rates residuals after smoothing with thin plate splines.*

5 Quantile lasso smoothing

A two dimensional thin plate spline is defined as the function $f(x,t)$ which minimises

$$J\left(\{y_i\}_{i=1}^n, f\right) = \sum_{i=1}^n (y_i - f(x_i,t_i))^2 + \lambda \int \left[\frac{\partial^2 f}{\partial x^2} + 2\frac{\partial^2 f}{\partial x \partial t} + \frac{\partial^2 f}{\partial t^2}\right] \mathrm{d}x\,\mathrm{d}t$$

for some smoothing parameter $\lambda > 0$, knots $\{(x_i,t_i)\}_{i=1}^n$ and values $\{y_i\}_{i=1}^n$ (see for example Wood, 2006).

If the knots form a fine regular grid, then the integral can be approximated by a sum and so $J(\{y_i\}_{i=1}^n, f)$ can be approximated as

$$J\left(\{y_i\}_{i=1}^n, f\right) \approx \sum_{i=1}^n (y_i - f(x_i,t_i))^2 + \frac{\lambda}{n} \sum_{i=1}^n \left[\frac{\partial^2 f}{\partial x^2}(x_i,t_i) + 2\frac{\partial^2 f}{\partial x \partial y}(x_i,t_i) + \frac{\partial^2 f}{\partial y^2}(x_i,t_i)\right].$$

Also if the knots form a fine regular grid, then the second partial derivatives at knots can be approximated as linear combinations of function values at nearby knots.

Denoting $\{y_i\}_{i=1}^n$ as vector y and $\{f(x_i,y_i)\}_{i=1}^n$ as vector z, then $J(\{y_i\}_{i=1}^n, f)$ can be approximated as

$$J(y,z) \approx \|y - z\|_{L_2}^2 + \frac{\lambda}{n}\left(\|D_{xx}z\|_{L_2}^2 + 2\|D_{xt}z\|_{L_2}^2 + \|D_{tt}z\|_{L_2}^2\right)$$

where D_{xx}, D_{xt} and D_{tt} are linear operators (matrices) which calculate approximations of vectors $\left\{\frac{\partial^2 f}{\partial x^2}(x_i,t_i)\right\}_{i=1}^n$, $\left\{\frac{\partial^2 f}{\partial x \partial t}(x_i,t_i)\right\}_{i=1}^n$ and $\left\{\frac{\partial^2 f}{\partial t^2}(x_i,t_i)\right\}_{i=1}^n$.

Using the above expression, we can approximate a thin plate spline computed at its knots as

$$S(y) = \arg\min_z \left(\|y - z\|_{L_2}^2 + \frac{\lambda}{n}\left(\|D_{xx}z\|_{L_2}^2 + 2\|D_{xt}z\|_{L_2}^2 + \|D_{tt}z\|_{L_2}^2\right)\right).$$

In the case of smoothing mortality rates, y becomes the data vector containing log mortality rates (two-dimensional data packed as vector). The order of packing affects only the representation of matrices D_{xx}, D_{xt} and D_{tt}.

Following Schuette (1978), we now replace the L_2 norm with the L_1 norm to give smoothing with the quantile lasso. In addition, we use three different λ coefficients before every derivative to separately adjust the influence of each derivative on the smoothing. Therefore in this method we define smoothing as $Q(y) = \arg\min_z(K(y,z))$ where

$$K(y,z) = \|y - Mz\|_{L_1} + \lambda_{xx}\|D_{xx}z\|_{L_1} + \lambda_{xt}\|D_{xt}z\|_{L_1} + \lambda_{tt}\|D_{tt}z\|_{L_1}$$

and y, D_{xx}, D_{xt}, D_{tt} are as described above. Since we use the same number of knots and data points, each positioned at the same places, our matrix M becomes an identity matrix.

For this approach, we do not use the "flattening" procedure described in Section 4. While it would be possible to use it, the effect is negligible, and so for simplicity we have not included it.

Minimization of $K(y,z)$ appears to be difficult, but due to a well known procedure (described for example in Wood, 2006) the problem can be reduced to a quantile regression problem, which then can be solved with existing software (Koenker, 2013). In this study we adopt the following reduction procedure:

- Matrices M, $\lambda_{xx}D_{xx}$, $\lambda_{xt}D_{xt}$, and $\lambda_{tt}D_{tt}$ are stacked on top of each other to give $R = [M', \lambda_{xx}D'_{xx}, \lambda_{xt}D'_{xt}, \lambda_{tt}D'_{tt}]'$.
- Vector y is extended by zeros until its length is equal to the number of rows in R: $y_{ext} = [y', 0']'$.
- $K(y,z) = \|y-Mz\|_{L_1} + \lambda_{xx}\|D_{xx}z\|_{L_1} + \lambda_{xt}\|D_{xt}z\|_{L_1} + \lambda_{tt}\|D_{tt}z\|_{L_1}$ is replaced with the equivalent expression $K(y,z) = \|y_{ext} - Rz\|_{L_1}$.

Then finding $Q(y) = \arg\min_z(K(y,z))$ is a quantile regression problem.

The smoothing method described above is defined for some fixed parameters λ_{xx}, λ_{xt} and λ_{tt}, which need to be optimised to get maximum performance. As a measure of performance we use the predictive ability of the procedure, estimated using the mean absolute error based on five-fold cross validation. The function measuring performance depends on parameters λ_{xx}, λ_{xt} and λ_{tt}. This function may have many local minima which makes the process of finding optimal parameters difficult. We optimize parameters λ_{xx}, λ_{xt} and λ_{tt} using the optimization procedure "malschains" (Bergmeir et al., 2012) which tends to avoid local minima and therefore has a greater chance to find a global optimal solution than standard gradient descent algorithms.

Every subset of data has about 20% of missing values and they each have the same (but shifted) pattern (Figure 9). The points are missed in a regular pattern to ensure the distance between them is as large as possible. Assuming that distant points affect smoothed value less than close points, the result is a fair compromise between closeness to leave-one-out cross validation and good computational time. Clearly such pseudo leave-one-out cross validation requires about five times more resources (processor time) comparing to a single smoothing task over the whole data set.

The result of smoothing is shown in Figure 10. It is less “smooth” than the results of the previous two methods. Nevertheless we will see in Section 7 that this smoothing method reflects “features” of the data more precisely than two previous smoothing methods.

The residuals are shown in Figure 11. Visually it is difficult to find any serial correlation in the errors. However, cohort and period effects are clearly visible.

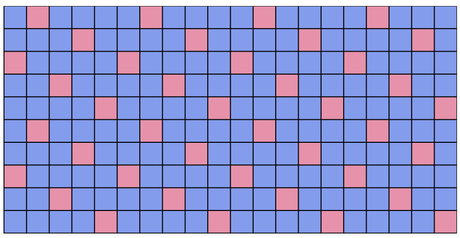

Figure 9: *Missed data pattern for pseudo leave-one-out cross validation.*

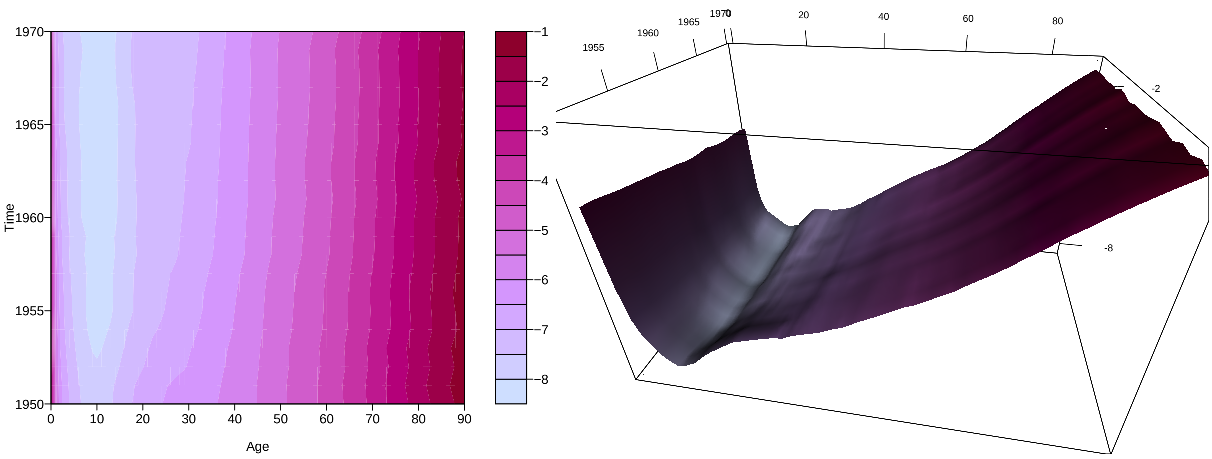

Figure 10: *French female mortality rates smoothed with the quantile lasso.*

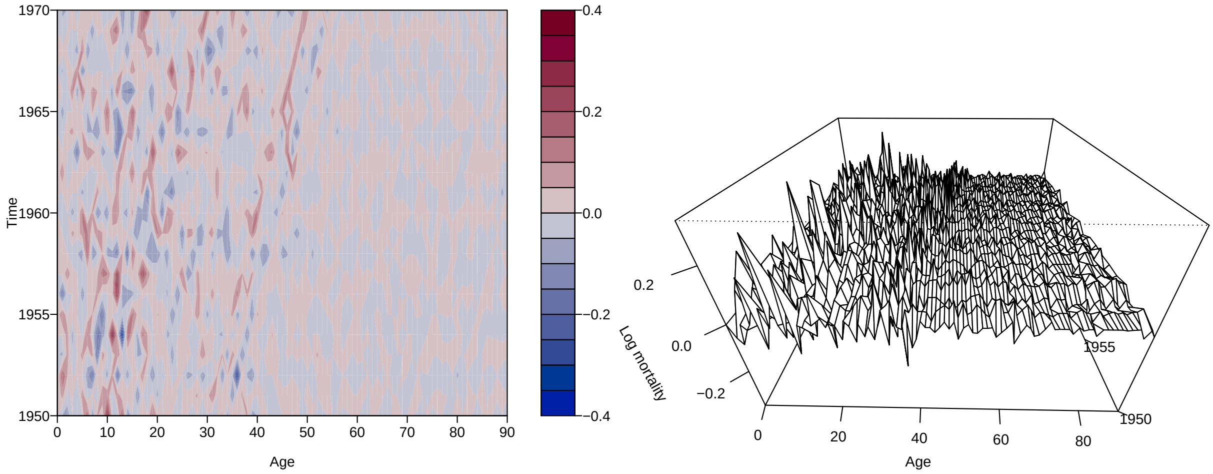

Figure 11: *French female mortality rates residuals after smoothing with the quantile lasso.*

6 Quantile lasso smoothing with cohort and period effects

The cohort and period effects seen in Figure 11 suggest that the smoothing model can be improved by incorporating these features explicitly. This leads us to our new and final smoothing method in which we first apply the quantile lasso algorithm of the previous section, and then identify and incorporate significant period and cohort effects. We call this the SMILE method: Smooth Mortality Involving Lasso and period and cohort Effects.

To identify the period and cohort effects, we compute the residuals from the quantile lasso smoothing algorithm described in the previous section. Then we carry out the following tests.

1. Perform two-sided t-tests of the residuals over all diagonals to find diagonals with residual mean values significantly different from zero.
2. Perform one-sided sample correlation tests for residual diagonals to find diagonals with positively correlated errors (every diagonal is tested for serial correlation with lag one).
3. Perform two-sided t-tests of the residuals over all years to find years with residual mean values significantly different from zero.
4. Perform one-sided sample correlation tests to find years with positively correlated errors (residuals for every particular year are tested for serial correlation with lag one along age dimension).

The new smoothing model involves summing four components: smooth mortality rates, cohort effects being non zero only along the diagonals identified in tests 1 and 2 above, period effects being non zero only along years identified in tests 3 and 4 above, and the noise. These four components are estimated using the following model:

$$Q(y) = \underset{z_{sm}, z_{coh}, z_{long}}{\arg\min} \, (K(y, z_{sm}, z_{coh}, z_{long})),$$

where

$$\begin{aligned} K(y, z_{sm}, z_{coh}, z_{long}) = {} & \|y - (z_{sm} + z_{coh} + z_{long})\|_{L_1} + \lambda_{xx}\|D_{xx} z_{sm}\|_{L_1} + \lambda_{xt}\|D_{xt} z_{sm}\|_{L_1} + \lambda_{tt}\|D_{tt} z_{sm}\|_{L_1} \\ & + \lambda_{coh}\|D_{coh} z_{coh}\|_{L_1} + \theta_{coh}\|z_{coh}\|_{L_1} + \lambda_{long}\|D_{long} z_{long}\|_{L_1} + \theta_{long}\|z_{long}\|_{L_1}; \end{aligned}$$

- y, D_{xx}, D_{xt} and D_{tt} are as described above;
- z_{sm}, z_{coh} and z_{long} are estimated components representing respectively smooth mortality surface, cohort effects restricted to some diagonals and period effects restricted to some years;
- D_{coh} is a linear differentiation operator representing a discrete version of the second directional derivative in the direction of vector $(1, 1)$;
- $D_{long} = D_{tt}$ is a linear differentiation operator representing a discrete version of the second derivative along the years axis;
- $\lambda_{xx}, \lambda_{xt}$, and λ_{tt} are parameters responsible for the smoothness of the mortality surface;
- λ_{coh} and λ_{long} are parameters responsible for the smoothness of the cohort effects and the period effects respectively;
- θ_{coh} and θ_{long} are parameters responsible for shrinking (respectively) the cohort effects and the period effects towards zero.

It may appear that components z_{sm}, z_{coh} and z_{long} duplicate each other. However, this is not the case because λ_{coh} and λ_{long} are restricted to values much greater than values of parameters λ_{xx}, λ_{xt} and λ_{tt}. Such restrictions make it difficult for z_{sm}, z_{coh} and z_{long} to compete for the same features in the data. High values of λ_{coh} and λ_{long} make z_{coh} and z_{long} tend to reflect trends along the diagonals and years. This occurs because observations which affect z_{coh} and z_{long} at some point are "one dimensional" (one observation narrow and long), although observations which affect z_{sm} at the same point are "two dimensional" (they make a shape which is rather small and approximately round).

It is also worth mentioning that the above tests for cohort and period effects are done only for the purpose of reducing computational complexity of the minimization problem. The multiple testing that is carried out means that the selected cohort and period effects are not necessarily statistically significant overall. Some or all of these cohort and period effects will be dropped in the subsequent minimization.

As in the previous section, to minimize $K(y, z_{sm}, z_{coh}, z_{long})$, we use the corresponding quantile regression problem in which:

- vectors z_{sm}, z_{coh} and z_{long} are stacked on top of each other as a single vector, $z_{ext} = [z'_{sm},\ z'_{coh},\ z'_{long}]'$;
- matrices I, $\lambda_{xx}D_{xx}$, $\lambda_{xt}D_{xt}$, $\lambda_{tt}D_{tt}$, $\lambda_{coh}D_{coh}$, $\lambda_{long}D_{long}$, $\theta_{coh}I$, and $\theta_{long}I$ are combined in one matrix,

$$R = \begin{bmatrix} I & I & I \\ \lambda_{xx}D_{xx} & 0 & 0 \\ \lambda_{xt}D_{xt} & 0 & 0 \\ \lambda_{tt}D_{tt} & 0 & 0 \\ 0 & \lambda_{coh}D_{coh} & 0 \\ 0 & \theta_{coh}I & 0 \\ 0 & 0 & \lambda_{long}D_{long} \\ 0 & 0 & \theta_{long}I \end{bmatrix};$$

- vector y is extended by zeros to have its length equal to the number of rows in R: $y_{ext} = [y',\ 0']'$;

- and

$$K(y, z_{sm}, z_{coh}, z_{long}) = \|y-(z_{sm}+z_{coh}+z_{long})\|_{L_1}+\lambda_{xx}\|D_{xx}z_{sm}\|_{L_1}+\lambda_{xt}\|D_{xt}z_{sm}\|_{L_1}+\lambda_{tt}\|D_{tt}z_{sm}\|_{L_1}+ \lambda_{coh}\|D_{coh}z_{coh}\|_{L_1} + \theta_{coh}\|z_{coh}\|_{L_1} + \lambda_{long}\|D_{long}z_{long}\|_{L_1} + \theta_{long}\|z_{long}\|_{L_1}$$

  is replaced with the equivalent expression $K(y, z_{ext}) = \|y_{ext} - Rz_{ext}\|_{L_1}$.

Then $Q(y) = \underset{z_{ext}}{\arg\min}(K(y, z_{ext}))$ is a quantile regression problem.

The resulting smoothed surface z_{sm} is shown in Figure 12. The cohort effects have been completely removed, but there are the shadows of some period effects remaining.

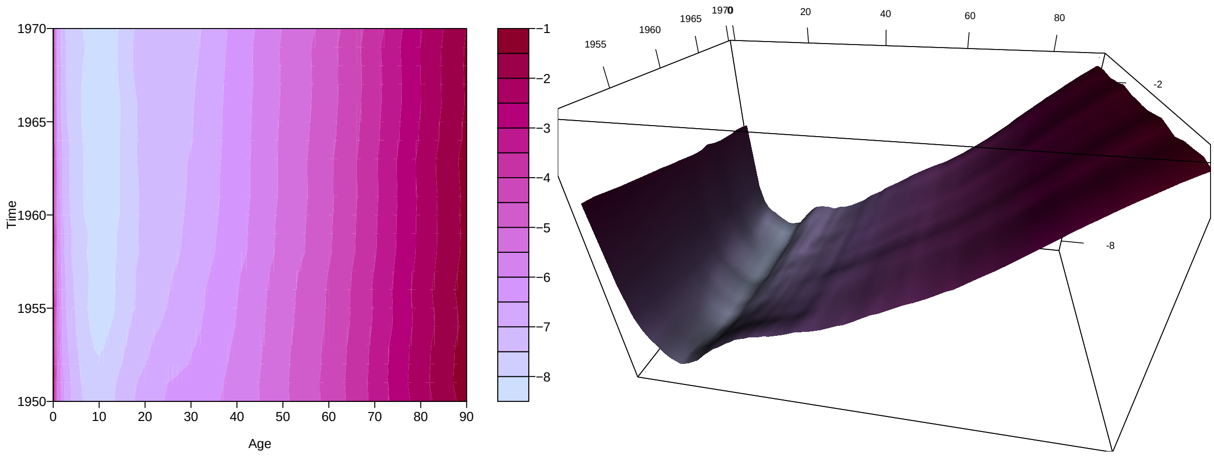

Figure 12: *French female mortality rates smoothed with the quantile lasso and taking into account cohort and period effects.*

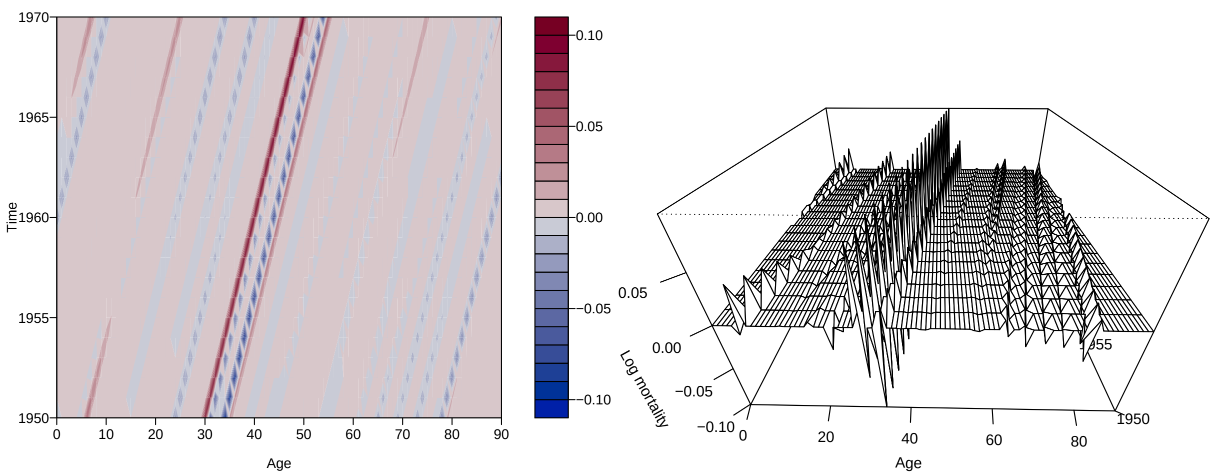

Figure 13: *Cohort effects of French female mortality rates.*

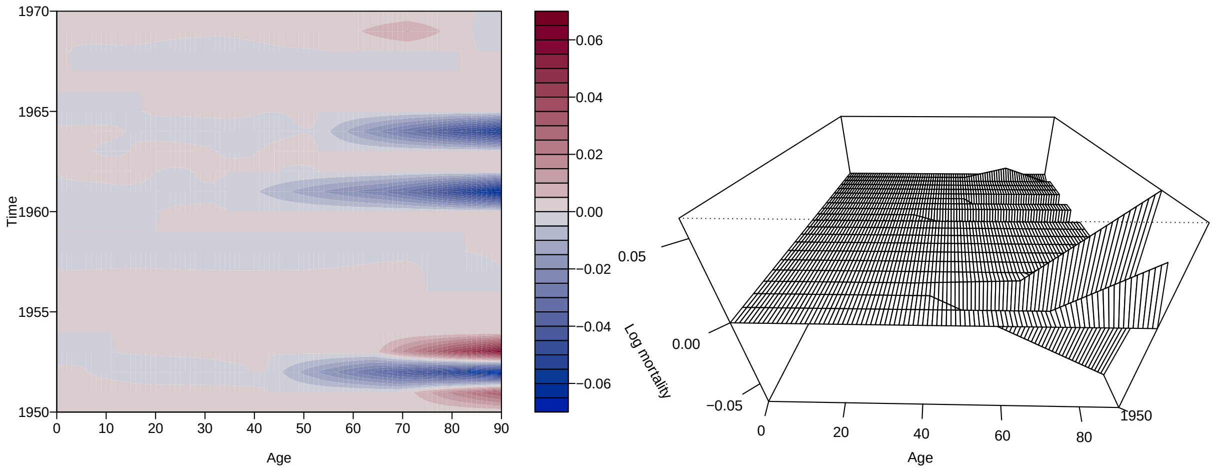

Figure 14: *Period effects of French female mortality rates.*

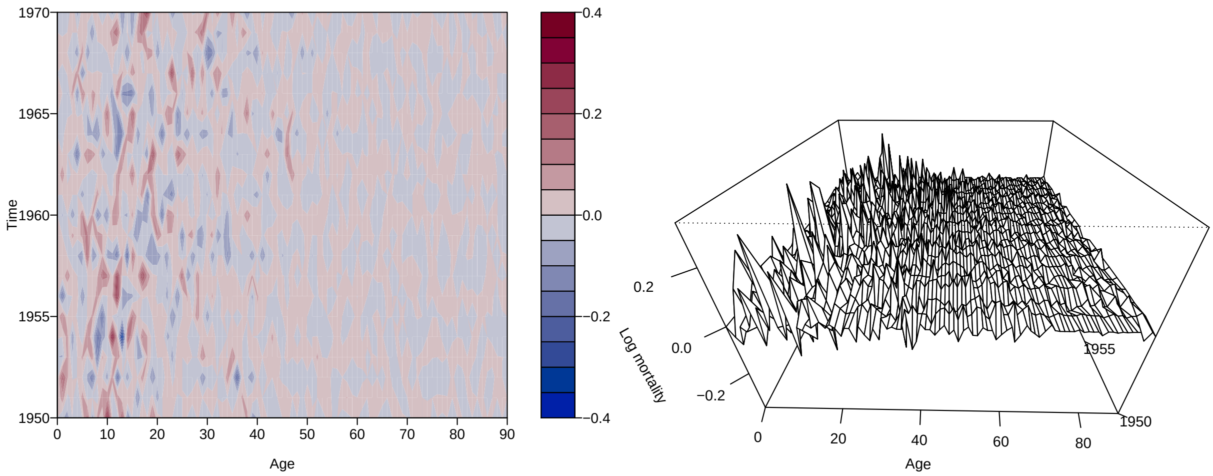

Figure 15: *French female mortality rates residuals after smoothing with the quantile lasso and taking into account cohort and period effects.*

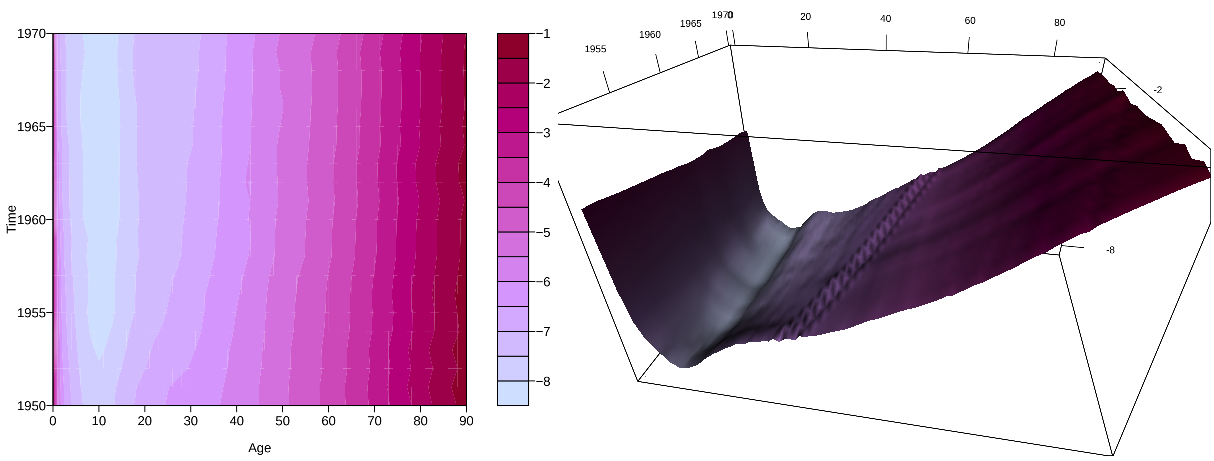

Figure 16: *French female mortality rates smoothed with the quantile lasso and cohort/period effects applied.*

The estimated cohort effects z_{coh} are shown in Figure 13. The strongest effects shown are starting at ages and years: (0, 1960), (6, 1950), (30, 1950), (31, 1950), (34, 1950), (35, 1950), (78, 1950).

We speculate that the most visible cohort effects, starting at year 1950 for people aged 30–35, are due to the Spanish flu pandemic and World War I affecting the cohort of people born in 1915–1920. It might also be due to incorrect number of births/deaths registered (records could be lost or births could be registered in adjacent year) during those years which uniformly affected mortality figures for these cohorts.

(Barry, 2004) reports that in thirteen studies of hospitalised pregnant women during the Spanish flue pandemic the death rate ranged from 23% to 71%. 26% of pregnant women who survived childbirth lost their child. Therefore it is quite possible that such severe death rates could affect both health of the cohort born during those years as well as accuracy of the number of registered people in that cohort.

The estimated period effects z_{long} are shown in Figure 14. The most visible period effects can be observed for years 1951, 1952, 1953, 1961 and 1964. They become stronger for older years. Possibly the 1951 and 1953 period effects are due to extreme climate events, or unusual flu epidemics, which tend to affect older people more severely. Years 1952, 1961 and 1964 show some *reduction* in mortality for older ages, the cause of which is unknown.

The residuals from the model are shown in Figure 15. The period and cohort effects are no longer visible.

Figure 16 depicts the complete surface with the cohort and period effects added to the smooth surface. It is the "signal" which we have separated from the "noise" (represented by the residuals).

7 Comparison

We use a cross validation procedure for comparing the different smoothing methods we have discussed. We randomly choose a 15% subset of all points in the original data, which is then split into five other random "small" subsets. Therefore each of these "small" subsets has only 3% of the original data. The resulting cross validation error is the average of errors over those five "small" subsets.

We use four subsets of the French female mortality data for our comparisons.

1. Data for years 1950–1970 and ages 10–60 represent a relatively smooth surface. This comparison is useful to ensure that the most "responsive" algorithms using the L_1 norm do not perform any worse than the more "stable" algorithms based on the L_2 norm. This data set is also important because it is the only comparison satisfying the requirements for Method 2 (Camarda, 2012) which is designed to work for ages starting from 10 years and where there are no outliers.

2. Data for years 1950–1970 and ages 0–60 represent a period when no outliers happened — there were no wars or large pandemics. The younger ages from 0–9 have abrupt changes in mortality rates which are more challenging to smooth.

3. Data for years 1935–1955 and ages 10–60 represent a period including "outliers" due to World War II. It is important to mention that such "outlier" should be considered as an outlier only along time dimension and it is a smooth data curve along ages dimension. Therefore in our case, when only uncorrelated errors are considered as noise, such one-dimensional outlier should be preserved during smoothing as signal. These data are important for testing the smoothing abilities of the algorithms in the presence of one-dimensional outliers.

4. Data for years 1935–1955 and ages 0–60 represents the most complex dataset containing the one-dimensional outliers (WWII) and also a period of abrupt mortality changes from ages 0–9.

All calculations are done using R (R Development Core Team, 2013).

Data		MSE x 100					MAE x 100				
Years	Ages	H-U	Cam	BTPS	QL	SMILE	H-U	Cam	BTPS	QL	SMILE
1950-1970	10-60	0.82	0.46	0.51	0.41	0.37	5.45	4.93	5.16	4.69	4.48
1950-1970	0-60	0.79	0.82	0.41	0.42	0.40	5.43	6.55	4.91	4.77	4.51
1935-1955	10-60	0.44	3.68	0.73	0.25	0.21	3.75	13.21	6.12	3.73	3.39
1935-1955	0-60	1.84	3.78	0.79	0.43	0.37	6.03	12.90	6.79	4.76	4.34

Table 1: *Cross validation performance of different smoothing methods against French female mortality data (H-U = Hyndman-Ulah method, Cam = Camarda method, BTPS = Boosted Thin Plate Splines method, QL = Quantile Lasso method, SMILE = Smooth Mortality Involving Lasso and period and cohort Effects method). The best figures in the table are highlighted in green, the next best in yellow/green, and the worst in pink.*

Table 1 demonstrates that the SMILE method shows better or similar performance (in terms of MSE and MAE) compare to the other methods in all tests. Amongst the other methods, the Quantile lasso performs better than Hyndman-Ulah, Camarda and Boosted thin plate splines methods in most cases. Boosted thin plate splines work well except for the cases when there are outliers, for which the Hyndman-Ullah method does better. Overall, the SMILE method is the best performing method amongst those tested.

8 Conclusion

In this paper we have considered three new methods to smooth mortality data in two dimensions. We have also introduced a comparison technique that is computationally feasible. Using this technique we compared our new methods between each other and also with existing one- and two-dimensional methods. We found that our proposed SMILE method gave the best results. This method also provides us with some insights into the mortality data including the existence of cohort and period effects that might otherwise be overlooked.

We conclude that use of two-dimensional data and thin plate splines for graduation can lead to improvements compared to a one dimensional approach (except in terms of MAE on abrupt data). On the other hand, using the L_1 norm instead of L_2 can lead to further performance improvements for abrupt data. Moreover, the methods which use the L_1 norm do not require overcomplicated preprocessing of data, which was necessary for methods based on L_2. Further performance improvements can be achieved by building into the model abilities to project the cohort and period effects.

Additional improvements in these proposed methods are possible. Our boosted thin plate spline method used adaptive splines, while only partially adaptive splines were used for the quantile lasso methods. A fully adaptive approach applied to the quantile lasso methods requires further investigation and may provide further improvements.

The λ coefficients used in the quantile lasso methods were estimated using lengthy numerical methods. A simpler procedure, similar to what is used by Camarda (2012), would improve their practical usefulness. We leave this to a later paper.

References

Barry, J. M. (2004). *The Great Influenza: The Epic Story of the 1918 Pandemic*. Viking.

Bergmeir, C., Molina, D., and Benítez, J. M. (2012). *Rmalschains: Continuous Optimization using Memetic Algorithms with Local Search Chains (MA-LS-Chains) in R*. http://cran.r-project.

org/package=Rmalschains.

Bühlmann, P. and Yu, B. (2003). Boosting with the L2 loss: regression and classification. *Journal of the American Statistical Association*, 98(462):324–339.

Camarda, C. G. (2012). MortalitySmooth: An R package for smoothing Poisson Counts with P-Splines. *Journal of Statistical Software*, 50(1):1–24. http://cran.r-project.org/package=MortalitySmooth.

Camarda, C. G., Eilers, P. H. C., and Gampe, J. (2010). Additive decomposition of vital rates from grouped data. In *25th International Workshop on Statistical Modelling*, pages 113–118.

Currie, I. D., Durban, M., and Eilers, P. H. (2004). Smoothing and forecasting mortality rates. *Statistical Modelling*, 4(4):279–298.

Forfar, D. O., McCutcheon, J. J., and Wilkie, A. D. (1988). On graduation by mathematical formula. *Journal of the Institute of Actuaries*, 115(1):1–149.

Heligman, L. and Pollard, J. H. (1980). The age pattern of mortality. *Journal of the Institute of Actuaries (1886-1994)*, 107(1):pp. 49–80.

Human Mortality Database (2008). University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Data downloaded on 20 Feb 2008. http://www.mortality.org.

Hyndman, R. J. (2012). *demography: Forecasting mortality, fertility, migration and population data*. R package version 1.16. With contributions from Heather Booth and Leonie Tickle and John Maindonald. http://cran.r-project.org/package=demography.

Hyndman, R. J. and Booth, H. (2008). Stochastic population forecasts using functional data models for mortality, fertility and migration. *International Journal of Forecasting*, 24(3):323–342.

Hyndman, R. J. and Ullah, S. M. (2007). Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics & Data Analysis*, 51(10):4942–4956.

Koenker, R. (2013). *quantreg: Quantile Regression*. http://cran.r-project.org/package=quantreg.

R Development Core Team (2013). *R: A language and environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.r-project.org.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press, New York.

Schuette, D. R. (1978). A linear programming approach to graduation. *Transactions of society of actuaries*, 30.

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley.

Wood, S. N. (1994). Monotonic smoothing splines fitted by cross validation. *SIAM Journal on Scientific Computing*, 15(5):1126–1133.

Wood, S. N. (2006). *Generalized Additive Models: an introduction with R*. CRC Press.