WAVELET-BASED FEATURE EXTRACTION FOR MORTALITY PROJECTION

BY

DONATIEN HAINAUT AND MICHEL DENUIT

ABSTRACT

Wavelet theory is known to be a powerful tool for compressing and processing time series or images. It consists in projecting a signal on an orthonormal basis of functions that are chosen in order to provide a sparse representation of the data. The first part of this article focuses on smoothing mortality curves by wavelets shrinkage. A chi-square test and a penalized likelihood approach are applied to determine the optimal degree of smoothing. The second part of this article is devoted to mortality forecasting. Wavelet coefficients exhibit clear trends for the Belgian population from 1965 to 2015, they are easy to forecast resulting in predicted future mortality rates. The wavelet-based approach is then compared with some popular actuarial models of Lee–Carter type estimated fitted to Belgian, UK, and US populations. The wavelet model outperforms all of them.

KEYWORDS

Discrete wavelet transform, mortality smoothing, Poisson regression, regularization, Lasso

JEL codes: G22, C01

1. INTRODUCTION AND MOTIVATION

Mortality projections are known to be of crucial importance for life insurance companies and pension funds. This topic has thus attracted a lot of attention in the actuarial literature. A variety of mortality projection models emerged over the last 30 years, ranging from basic regression models in which age and time are viewed as continuous features, to sophisticated nonparametric models. We refer the interested readers to Pitacco *et al.* (2009) for a general overview of the topic and to dedicated chapters in Denuit *et al.* (2019a,b) for applications.

Astin Bulletin 50(3), 675-707. doi:10.1017/asb.2020.18 © 2020 by Astin Bulletin. All rights reserved.

To be successful, a mortality projection model must be flexible enough to capture the underlying longevity dynamics and produce time-dependent components exhibiting a clear trend. The latter aspect largely explained the success of the pioneering model proposed by Lee and Carter (1992) whose time index generally appears to be markedly linear. Despite being simple and transparent, this model is, however, not very flexible and may fail to capture some important aspects of mortality data under study. For this reason, Hyndman and Ullah (2007) proposed to extend the approach proposed by Lee and Carter (1992) by adopting a functional data paradigm combined with nonparametric smoothing (penalized regression splines). Univariate time series are then fitted to each component coefficient (or level parameter). However, some of these coefficients time series do not exhibit clear trends making them difficult to forecast. The new approach for age-specific mortality projection proposed in this paper suggests that wavelets analysis remedies to this problem since time-varying coefficients have clear trends. Extrapolation is therefore easy, and the resulting mortality forecasts appear to be accurate in terms of backtesting.

Let us now explain why wavelets may outperform alternative functional data approaches. These approaches have in common that mortality curves are decomposed into a basis of functions. Estimation is known to be particularly easy if the functions comprised in the selected basis are orthogonal. Examples of orthogonal bases are orthogonal polynomials (as in Renshaw *et al.*, 1996, who decomposed mortality curves into Legendre polynomials) and the Fourier basis. The disadvantage of the approach based on these families is that the basis functions are not compactly supported so that finding coefficients providing a reasonable fit in one region can cause the mortality curve to become implausible in remote regions. Splines are compactly supported, but they are not orthogonal. Wavelets have the advantage that they are compactly supported and can be defined so as to possess the orthogonality property. The mortality curve is projected into the space of wavelets and expressed as a sum of functions weighted by coefficients. Most of these wavelet coefficients are close to zero and considered as random to perturbations of the underlying mortality structure. Noise is then removed by thresholding the smallest of these coefficients. Morillas *et al.* (2016) were among the first authors to apply the wavelets technique to mortality modeling. They propose a year-by-year method for smoothing mortality rates based on a wavelets decomposition, combined with piecewise polynomial harmonic techniques. The quality of the smoothing is then assessed by mean relative errors and by Whittaker–Henderson smoothness indicators. More recently, Jurado and Sampere (2019) use wavelet techniques to smooth mortality curves together with bootstrapping to obtain confidence bands around best-estimate mortality. In the present paper, we extend these previous works to a dynamic setting and propose more formal statistical procedures based on penalized Poisson likelihood maximization.

The present paper extends previous research on smoothing and forecasting of mortality rates in several directions. First, we use a chi-square statistical test

https://doi.org/10.1017/asb.2020.18 Published online by Cambridge University Press

to determine the optimal threshold under which wavelet coefficients are canceled. Next, we propose an alternative approach based on a penalized Poisson log-likelihood. Specifically, Lasso regularization techniques are applied to select the optimal wavelets. Thirdly, we perform a wavelets analysis of the Belgian mortality observed over the period 1965–2015. This numerical illustration reveals that the relevant information contained in the observed death rates is carried by a few wavelet coefficients common to all mortality curves. Since these coefficients exhibit clear trends, we propose and test a multivariate regression model. A parallel may be drawn with the work Hyndman and Ullah (2007) who proposed to extrapolate coefficients of a functional principal component analysis. However, compared to their approach, wavelet coefficients generally exhibit stable trends that are easy to extrapolate. Finally, we benchmark the predictive power of the wavelet model to the Lee and Carter (1992), Renshaw and Haberman (2003), Renshaw and Haberman (2006), and Cairns *et al.* (2006, 2009) models based on a backtesting analysis. This benchmarking is performed for Belgian, US, and UK populations.

The remainder of the text is organized as follows. We start in Section 2 with a brief introduction to wavelets, gathering detailed results needed for numerical implementation in the appendix to this paper. Particular attention is paid to discrete wavelets transform (DWT) since this algorithm is used in numerical illustrations. Next, we present the chi-square test that can be used to adjust the level of smoothing of log-mortality rates. We also propose an alternative smoothing method based on a least absolute shrinkage and selection operator (Lasso). This is an L_1-penalization of the log-likelihood used in high-dimensional regressions. In the second part of this paper (Section 3), we study the dynamics of wavelets coefficients for the Belgian population from 1965 to 2015. The trends in these coefficients suggest that a simple regression model can be used to forecast future mortality. We conclude with a numerical comparison with some popular actuarial models of Lee–Carter (LC) type based on backtesting and validate our conclusions with the US and UK datasets. Section 4 briefly concludes this paper.

2. Wavelets decomposition of the mortality curve

2.1. Wavelets for nonparametric regression

We provide here a gentle introduction to wavelets for nonparametric regression. A comprehensive presentation addressing all the issues needed for application to mortality is provided in the appendix to this paper.

Wavelets are functions that integrate to zero, “waving” above and below the x-axis, hence their name. Like sines and cosines in Fourier analysis, wavelets are used as basis functions in representing other functions. According to Hastie *et al.* (2016, Chapter 5), wavelets produce a dictionary $\mathcal{D}$ consisting of a very large number of basis functions that can be used to approximate any

https://doi.org/10.1017/asb.2020.18 Published online by Cambridge University Press

well-behaved unknown function of interest (here, the force of mortality on the log scale). Selection and regularization methods can then be used to restrict the entire dictionary to an optimal subset. Wavelets can thus be seen as new features entering the score in Poisson regression. It is interesting to note that, contrarily to Generalized additive model involving splines or local Generalized linear model which typically assume that the force of mortality is a smooth function of age, wavelets are able to represent both smooth and/or locally bumpy functions in an efficient way. Wavelets can thus capture the transitory effect of epidemics for instance.

Wavelets are defined by parent functions: a "father" wavelet ϕ and a "mother" wavelet ψ both assumed to be compactly supported. Once the mother wavelet ψ has been selected, the wavelet basis is obtained by dilating and translating ψ to form the dictionary $\mathcal{D}$, that is, $\psi(\frac{x-b}{a})$ for $a > 0$ and $b \in \mathbb{R}$. It is convenient to take special values for a and b in defining the wavelet basis, to ensure sparsity: $a = 2^j$ and $b = k2^{-j}$, where k and j are integers. The father wavelet ϕ plays the role of scaling function. The functions

$$\phi_{j,k}(x) = 2^{j/2}\phi\left(2^j x - k\right) \text{ and } \psi_{j,k}(x) = 2^{j/2}\psi\left(2^j x - k\right) \text{ for } k, j \in \mathbb{Z}$$

form the available dictionary $\mathcal{D}$ to model the mortality curve.

The simplest type of wavelet is certainly the Haar basis. The mother wavelet ψ for the Haar family is the so-called Haar function defined on the interval $[0, 1)$, being equal to 1 on $[0, 0.5]$ and to -1 on $(0.5, 1)$. The corresponding scaling function ϕ is equal to 1 on the interval $[0, 1)$ and to 0 otherwise. Dilations and translations $\psi_{j,k}$ of the Haar function ψ form an orthogonal basis in the space of all square-integrable functions. This means that any such function can be represented as a linear combination (possibly infinite) of these basis functions.

The Haar wavelets are simple to understand, but not smooth enough for representing smooth mortality curves. Continuous basis functions, such as Daubechies wavelets, are better choices in that respect. The mother wavelet is not explicitly defined but is implicitly computed from the method for making the wavelet decomposition.

2.2. DWT of mortality curves

Let $\mu(t, x)$ be the force of mortality (or hazard rate) for an individual aged x at time t. The survival probability to time $s \geq t$ is then given by

$${}_sp_x(t) := \exp\left(-\int_t^s \mu(v, x + v - t)\, dv\right).$$

Henceforth, we assume that the force of mortality is constant on each square of the Lexis diagram, that is, for every integer x and t,

$$\mu(t + \tau, x + \xi) = \mu(t, x) \text{ for all } 0 \leq \tau < 1 \text{ and } 0 \leq \xi < 1.$$

We assume that we have at our disposal a set of observations where time ranges from year t_m to t_M and age from x_m to x_M. The number of observations for each year is denoted as $n = x_M - x_m + 1$. Available demographic data consist of the number of deaths observed at age x last birthday during year t, $n_{t,x}$, and the corresponding exposure to risk, $E_{t,x}$. Here, $E_{t,x}$ (sometimes called central exposure to risk at age x) is the total time lived by people aged x last birthday in calendar year t. An unbiased estimator $\widehat{\mu}(t,x)$ of the force of mortality is then given by

$$\widehat{\mu}(t,x) = \frac{n_{t,x}}{E_{t,x}}.$$

In Denuit and Legrand (2018), it is formally shown that it is not restrictive to conduct inference under the Poisson assumption for death counts, that is, by assuming that the observed number of deaths is the realization of a random variable $N_{t,x}$ that has a Poisson distribution with parameters $E_{t,x}\,\mu(t,x)$ as proposed by Brouhns *et al.* (2002). The corresponding expected number of deaths is thus $E_{t,x}\,\mu(t,x)$. The use of wavelets in Poisson regression is discussed, for example, in Besbeas *et al.* (2004).

The DWT decomposes the force of mortality on the log scale viewed as a function of age x, for fixed time t, according to formula (A.9) in appendix. It requires a number of data points equal to a power of 2. Since n is generally not a power of 2 in applications, we have to interpolate the log force of mortality. Morillas *et al.* (2016) propose a piecewise polynomial harmonic technique. We use instead a linear regression which appears to be sufficient for our purposes. We set $\Delta_x = \frac{x_M - x_m}{2^J - 1}$ and calculate $\widehat{\mu}(t,x)$ for noninteger ages $x \in \{x_k \mid x_k = x_m + k\Delta_x,\ k = 0, \ldots, 2^J - 1\}$ by linear interpolation, in order to produce a data set with $n = 2^J$ observations. DWT allows us to decompose the log-mortality rates $\widehat{\mu}(t,x)$ into a sum of wavelets:

$$\ln \widehat{\mu}(t,x) = c_0(t)\phi(x) + \sum_{j=0}^{J-1} \sum_{k=0}^{2^j-1} d_{j,k}(t)\psi_{j,k}(x). \tag{2.1}$$

The vector of parameters $\boldsymbol{d}(t) = \left(d_{j,k}(t)\right)_{j,k}$ of dimension 2^J is therefore itself the realization of a multivariate random variable, denoted by $\boldsymbol{D}(t)$. Indeed, if we denote by

$$\boldsymbol{y}(t) = (\ln \widehat{\mu}(t,x))_{x=x_m,\ldots,x_M}$$

the vector of the log force of mortality, wavelets coefficients are linear combinations of realized log-mortality rates:

$$\boldsymbol{d}(t) = \boldsymbol{T}\boldsymbol{y}(t), \tag{2.2}$$

where $\boldsymbol{T}$ is an orthogonal matrix, that is, $\boldsymbol{T}\boldsymbol{T}^\top = \boldsymbol{I}$. The vector $\boldsymbol{d}(t)$ is sparse: the information carried by $\boldsymbol{y}(t)$ is redistributed among a smaller number of

https://doi.org/10.1017/asb.2020.18 Published online by Cambridge University Press

coefficients, significantly different from zero. Wavelet fitting can be performed with the help of the wavethresh package of R. In the next section, we test the relevance of including all wavelets in the sum (2.1).

The curve of log forces of mortality, stored in a vector $\boldsymbol{y}(t)$, is converted into a sparse vector $\boldsymbol{d}(t)$ of the same dimension. Knowing $\boldsymbol{d}(t)$ or $\boldsymbol{y}(t)$ is equivalent because we can reconstruct $\boldsymbol{y}(t)$ from $\boldsymbol{d}(t)$ with equation (2.2). This sparse vector can be layered into subvectors $\boldsymbol{d}_j(t)$ for $j = 0$ to J which contains enough information for approaching the original signal by a smooth curve with an increasing accuracy when $j \to J$. Details and an illustration with the Haar wavelet are provided in Appendix A.2.

2.3. A chi-square test for wavelets shrinkage of mortality curves

Following the idea of Donoho and Johnstone (1994, 1995), large values of wavelet coefficients most likely correspond to the true signal, whereas small coefficients are related to noises. Hence, an efficient estimate $\widehat{\boldsymbol{d}}(t)$ of $\boldsymbol{d}(t)$ only keeps coefficients that are sufficiently large. Donoho and Johnstone (1994) propose two types of thresholding: the so-called hard and soft ones. In case of hard thresholding, we cancel all wavelet coefficients smaller in absolute value than a threshold, noted d^*. Precisely,

$$\widehat{d}_{j,k}(t) = \begin{cases} 0 \text{ if } |d_{j,k}(t)| < d^* \\ d_{j,k}(t) \text{ otherwise,} \end{cases}$$

for $j \in \{0, ..., J-1\}$, $k \in \{0, ..., 2^j - 1\}$. We denote by $\widehat{\mu}^S(t, x)$ the shrinked wavelet representation of $\widehat{\mu}(t, x)$:

$$\ln \widehat{\mu}^S(t, x) = c_0(t)\phi(x) + \sum_{j=0}^{J-1} \sum_{k=0}^{2^j-1} \widehat{d}_{j,k}(t)\psi_{j,k}(x). \tag{2.3}$$

The algorithm reconstructing the $\ln \widehat{\mu}^S(t, x)$ is called the Mallat's pyramid. It is summarized in Appendix A.2, and we refer the reader to Mallat (1989a,b) for further explanations. The Mallat's pyramid provides the value of $\widehat{\mu}^S(t, x)$ at noninteger ages, and log forces of mortality at integer ages are next retrieved by linear interpolation. In order to determine an acceptable threshold, we use a chi-square test. As mentioned earlier, it is not restrictive for estimation purposes to assume that the number of deaths $N_{t,x}$ obeys the Poisson distribution with parameters $E_{t,x}\mu(t, x)$. Given that the expectation and variance of $N_{t,x}$ are equal to $\mu(t, x)E_{t,x}$, the first two moments of $\widehat{\mu}(t, x)$ are given by

$$\mathbb{E}\big[\widehat{\mu}(t, x)\big] = \mu(t, x) \text{ and } \mathbb{V}\big[\widehat{\mu}(t, x)\big] = \frac{\mu(t, x)}{E_{t,x}}.$$

If the size of the population is large enough, the expected number of deaths is also large and the Poisson distribution for death counts can be approximated

https://doi.org/10.1017/asb.2020.18 Published online by Cambridge University Press

by the normal one. This assumption of normality holds if the Cochran (1952) criterion is satisfied. Let us denote by $P_{t,x}$ the size of the population of age x on year t and $\widehat{q}_{t,x} = \frac{n_{t,x}}{P_{t,x}}$ an estimate of the death probability. The assumption of normality is accepted if $n_{t,x}\widehat{q}_{t,x} \geq 5$ and $n_{t,x}(1-\widehat{q}_{t,x}) \geq 5$. In practice, these conditions are fulfilled for a wide range of ages. This is why we can consider the following Gaussian approximation for the distribution of $\widehat{\mu}(t,x)$:

$$\widehat{\mu}(t,x) \sim \text{Normal}\left(\mu(t,x),\ \frac{\mu(t,x)}{E_{t,x}}\right). \tag{2.4}$$

To determine if a threshold is admissible, we first build the mortality curve $\widehat{\mu}^S(t,x)$ from shrinked coefficients with the Mallat's pyramid. Next, we test for

$$\begin{cases} H_0 : & \mu(t,x) = \widehat{\mu}^S(t,x), \\ H_1 : & \mu(t,x) \neq \widehat{\mu}^S(t,x), \end{cases}$$

with the help of statistics

$$S_t = \sum_{x=x_{min}}^{x_{max}} E_{t,x} \frac{\left(\widehat{\mu}^S(t,x) - \widehat{\mu}(t,x)\right)^2}{\widehat{\mu}^S(t,x)}.$$

This statistics is approximately chi-square distributed with $n-p-1$, where n and p are, respectively, the number of observations and the number of non-null wavelet coefficients. If S_t is too large, that is, exceeds the corresponding chi-square quantile, we reject the null assumption H_0. The age range $[x_m,\ x_M]$ is restricted to $[x_{min},\ x_{max}] \subset [x_m,\ x_M]$, so that the Gaussian approximation for $N_{t,x}$ is sufficiently accurate.

Notice that Donoho and Johnstone (1995) apply instead soft shrinkage to wavelets coefficients. In this approach, coefficients are trimmed as follows:

$$\widehat{d}_{j,k}(t) = \begin{cases} 0 \text{ if } |d_{j,k}(t)| < d^* \\ \\ sgn(d_{j,k}(t))\left(\left|d_{j,k}(t)\right| - d^*\right) \text{ otherwise,} \end{cases}$$

where d^* is the threshold level. This rule is closely related to optimal coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$ of an L_1-penalized linear regression:

$$\boldsymbol{\beta} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 + \lambda||\boldsymbol{\beta}||_1,$$

where X is a $n \times p$ matrix of p covariates for n observations, $\boldsymbol{y}$ is the n vector of measurements, and $\lambda \in \mathbb{R}^+$ is a penalty. If the matrix X is orthogonal (i.e., $X^\top X = I_p$, where I_p is the identity matrix), this optimization problem admits a closed-form solution

$$\widehat{\boldsymbol{\beta}} = \text{sgn}\left(\widehat{\boldsymbol{\beta}}_{LS}\right) \max\left(0,\ \left|\widehat{\boldsymbol{\beta}}_{LS}\right| - \lambda\right),$$

https://doi.org/10.1017/asb.2020.18 Published online by Cambridge University Press

that is, precisely the trimming formula used in soft shrinkage. Nevertheless, we do not follow this direction because empirical experiments reveal that the χ^2 statistics rejects curves smoothed with the soft-shrinkage approach.

2.4. Application to Belgian mortality

Let us apply the method described in the preceding sections to mortality rates observed for the Belgian population (both genders combined) during 2015 and for ages ranging from 0 to 109 years. The data set comes from the Human Mortality Database (HMD, www.mortality.org). In practice, raw mortality rates at older ages are noised due to the lack of observations. There exist multiple approaches for managing this issue, for example, as the one proposed by Gbari *et al.* (2017). We circumvent this drawback by using HMD tables that contain smoothed mortality rates for ages above 90. The HMD protocol of Wilmoth *et al.* (2019) mentions indeed that "above age 80, population estimates are derived by the method of extinct generations for all cohorts that are extinct and by the survivor ratio method for nonextinct cohorts who are older than age 90 at the end of the observation period."

We work with Daubechies wavelets of order 4 (the latter order leading to the most sparse models). The left and right plots of Figure 1, respectively, show the evolution of the number of non-null $\widehat{d}_{j,k}$ and chi-square statistics in function of the threshold d^*. Table 1 reports the values of the test statistics S_t for increasing thresholds. This statistics is computed with $x_{min} = 0$ and $x_{max} = 100$. The null assumption is not rejected by decreasing the number of wavelets coefficients to 22. This reveals that we can explain the term structure of mortality with only 22 wavelets instead of the 128 initial ones. The last column reports the power ratio $\|\widehat{\boldsymbol{d}}(t)\|^2 / \|\boldsymbol{y}(t)\|^2$ that quantifies the information captured by the shrinked model. The values of AIC and BIC obtained under the normal approximation for mortality rates are reported in Table 1. Given a set of candidate models, the preferred model is the one with the lowest AIC or BIC. The AIC and BIC reward goodness of fit assessed by the likelihood function but also penalize models with a large number of parameters. According to the AIC and BIC, the best smoothed curves are, respectively, the ones built with 44 wavelets and 22 wavelets. The BIC usually favors sparse models compared to the AIC. Figure 2 compares the smoothed and original curves of log-mortality rates for the year 2015. The smoothed curve still presents some oscillations at younger ages. Increasing the threshold partly removes these oscillations, but the chi-square test statistics S_t lead to a rejection of these curves.

Figure 3 presents the 22 wavelets selected among the 128 ones for smoothing log forces of mortality. It is not possible to assimilate these wavelets to particular age effects. By essence, the wavelet transform converts a signal in a sum of wavelet functions chosen for their mathematical properties (orthogonality and scalable), but these functions are not easily interpretable.

https://doi.org/10.1017/asb.2020.18 Published online by Cambridge University Press

TABLE 1

BELGIAN POPULATION, YEAR 2015. GOODNESS-OF-FIT STATISTICS FOR DIFFERENT WAVELETS THRESHOLDS.

Threshold	$p =$ # of	S_t	χ^2	χ^2			Power
d^*	$\widehat{d}_{j,k} \neq 0$		2.5%	97.5%	AIC	BIC	ratio
0.01	93	10.08	1.69	16.01	−1328.73	−1077.58	1
0.03	60	24.74	24.43	59.34	−1380.04	−1218.01	1
0.05	49	38.99	33.16	72.62	−1387.8	−1255.47	1
0.07	44	41.95	37.21	78.57	−1394.84	−1276.02	1
0.09	37	59.34	42.95	86.83	−1391.44	−1291.52	0.9999
0.11	35	60.24	44.6	89.18	−1394.52	−1300	0.9999
0.13	32	74.4	47.09	92.69	−1385.86	−1299.45	0.9999
0.15	30	78.19	48.76	95.02	−1386.07	−1305.05	0.9998
0.17	29	79.81	49.59	96.19	−1386.45	−1308.14	0.9998
0.19	26	87.56	52.1	99.68	−1384.76	−1314.54	0.9997
0.21	24	96.3	53.78	102	−1380.05	−1315.23	0.9996
0.23	22	98.1	55.47	104.32	−1382.24	−1322.83	0.9995
0.25	22	98.1	55.47	104.32	−1382.24	−1322.83	0.9995

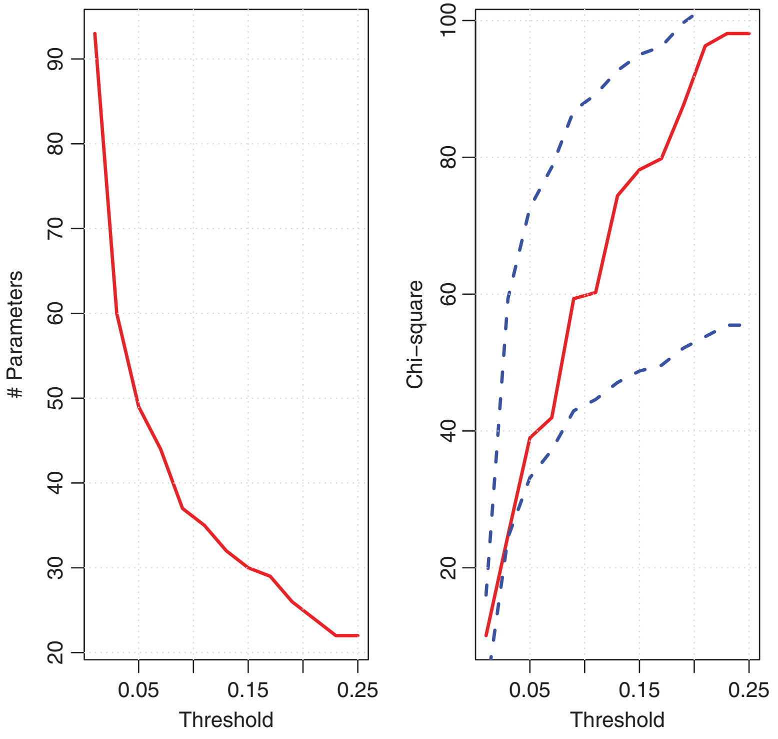

FIGURE 1: Belgian population, year 2015. Left and right plots, respectively, show the evolution of the number of non-null $\widehat{d}_{j,k}$ and chi-square statistics in the function of the threshold d^*. In the right graph, blue lines correspond to the 2.5% and 97.5% chi-square quantiles with $n - p - 1$ degrees of freedom.

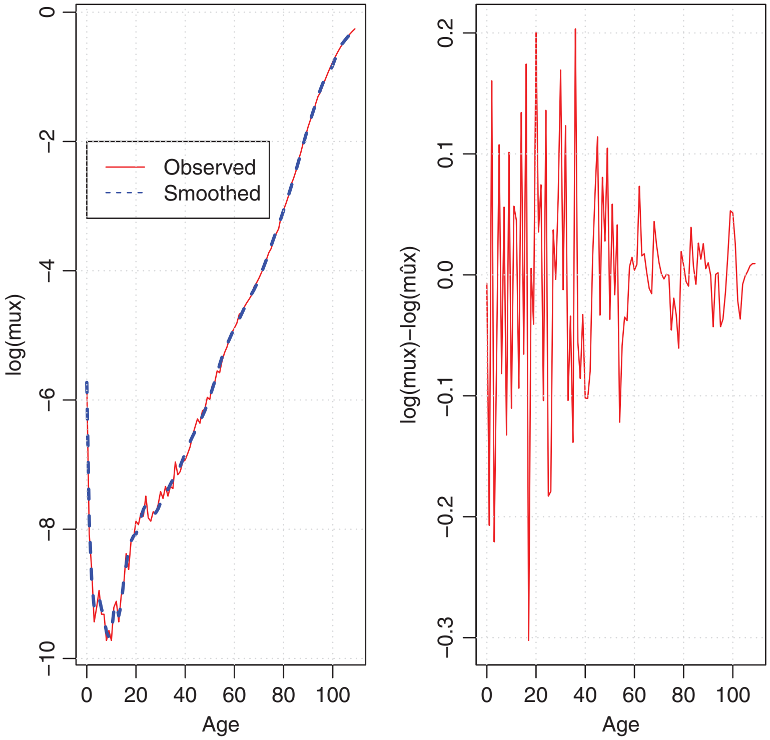

FIGURE 2: Belgian population, year 2015. The left plot shows the smoothed and original curves of log forces of mortality. The right graph reports the spreads between the observed and smoothed log-mortality rates.

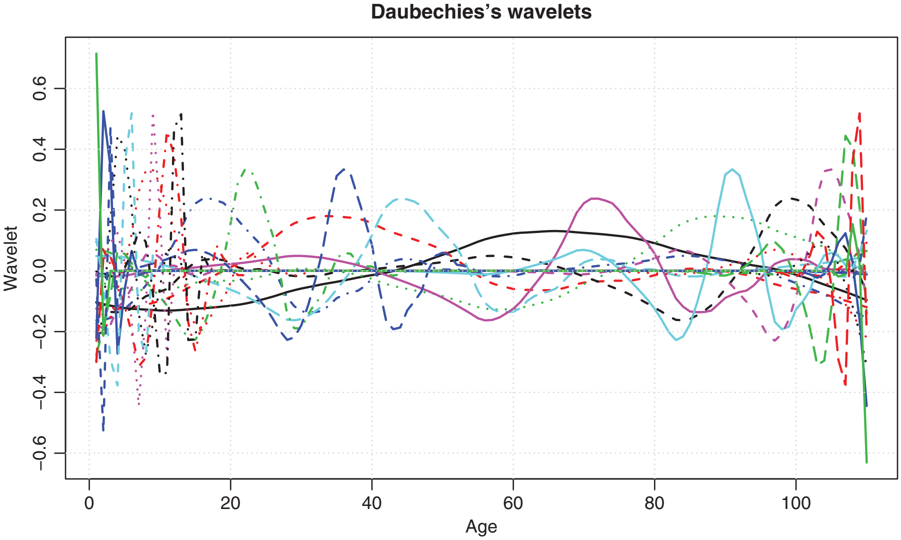

FIGURE 3: Plot of the 22 wavelets for the construction of the smoothed curve of mortality rates (year 2015).

https://doi.org/10.1017/asb.2020.18 Published online by Cambridge University Press

2.5. A Lasso approach for wavelets shrinkage

The chi-square test used in the preceding section can be seen as an exploratory technique since it relies on a normal approximation. An alternative way for tuning the degree of smoothing consists in selecting the shrinked model to minimize a penalized Poisson log-likelihood. The Lasso (least absolute shrinkage and selection operator) is an L_1-penalization selecting relevant variables in order to enhance the accuracy of prediction and interpretability of results. Lasso was popularized by Tibshirani (1996) for least squares regressions and can easily be adapted to wavelets shrinkage.

Our approach is based on a Poisson regression model, such as introduced in Brouhns *et al.* (2002). Under the assumption that the number of deaths is Poisson distributed, the likelihood of observations during year t is

$$\mathbb{P}\left[N_{t,x}=n_{t,x}\mid\mu(t,x)\right]=\frac{\left(\mu(t,x)E_{t,x}\right)^{n_{t,x}}}{n_{t,x}!}\exp\left(-\mu(t,x)E_{t,x}\right).$$

The log-likelihood for a hard skrinked model $\mu^S(t,x)$ is denoted by $\ln\mathcal{L}_{Pois}\left(d^*,\boldsymbol{\mu}^S\right)$ and is equal to the sum:

$$\begin{aligned}\ln\mathcal{L}_{Pois}(d^*,\boldsymbol{\mu}^S)&=\sum_{x=x_m}^{x_M}\ln\mathbb{P}\left[N_{t,x}=n_{t,x}\mid\mu^S(t,x)\right] \qquad (2.5)\\&=\sum_{x=x_m}^{x_M}\left(n_{t,x}\ln\left(\mu^S(t,x)E_{t,x}\right)-\mu^S(t,x)E_{t,x}-\ln\left(n_{t,x}!\right)\right).\end{aligned}$$

If the number of observations n is at least equal to the number of nonredundant parameters p, we can get a perfect fit by setting $\mu^S(t,x)=\widehat{\mu}(t,x)$. The corresponding model is the saturated one. This model is trivial and of no practical interest but since it perfectly fits data, its log-likelihood is the best attainable one for this distribution. The log-likelihood of the saturated model is

$$\ln\mathcal{L}_{Pois}(d^*,\widehat{\boldsymbol{\mu}})=\sum_{x=x_m}^{x_M}\left(n_{t,x}\ln\left(\widehat{\mu}(t,x)E_{t,x}\right)-\widehat{\mu}(t,x)E_{t,x}-\ln\left(n_{t,x}!\right)\right).$$

The scaled deviance D^* is defined as the logarithm of the likelihood ratio test of the model under consideration against the saturated model:

$$\begin{aligned}D^*(\widehat{\boldsymbol{\mu}},\boldsymbol{\mu}^S)&=2\left(\ln\mathcal{L}_{Pois}(d^*,\widehat{\boldsymbol{\mu}})-\ln\mathcal{L}_{Pois}(d^*,\boldsymbol{\mu}^S)\right)\\&=2\sum_{x=x_m}^{x_M}E_{t,x}\left(\mu(t,x)\ln\left(\frac{\mu(t,x)}{\mu^S(t,x)}\right)+\mu^S(t,x)-\mu(t,x)\right).\end{aligned}$$

We choose the optimal threshold d^* among the vector of absolute values of wavelet coefficients:

$$d^* \in |\boldsymbol{d}| = \left\{ |d_{j,k}(t)| \qquad j \in \{0, ..., J-1\}, k \in \{0, ..., 2^j - 1\} \right\},$$

and the vector of shrinked wavelet coefficients $\widehat{\boldsymbol{d}}$ is such that $\widehat{d}_{j,k}(t) = 0$ if $\left|\widehat{d}_{j,k}(t)\right| < d^*$ and $\widehat{d}_{j,k}(t) = d_{j,k}(t)$ otherwise. The optimal Lasso threshold d^* minimizes the following penalized deviance:

$$d^* = \arg \min_{d^* \in |\boldsymbol{d}|} D^*(\widehat{\boldsymbol{\mu}}, \boldsymbol{\mu}^S) + \lambda \sum_{j=1}^{J-1} \sum_{k=1}^{2^J-1} \left|\widehat{d}_{j,k}(t)\right|, \tag{2.6}$$

where $\lambda \in \mathbb{R}^+$ is the Lasso parameter determining the level of shrinkage. The function in Equation (2.6) corresponds to the Lagrangian of the optimization problem:

$$d^* = \arg \min_{d^* \in |\boldsymbol{d}|} D^*(\widehat{\boldsymbol{\mu}}, \boldsymbol{\mu}^S) \text{ subject to } ||\widehat{\boldsymbol{d}}||_1 \leq \gamma, \tag{2.7}$$

for some upper bound $\gamma \in \mathbb{R}^+$ on the L_1 norm of $\widehat{\boldsymbol{d}}$. We have applied the Lasso approach to mortality rates of the Belgian population (both genders) observed in 2015 and for ages ranging from 0 to 109 years. We choose a Lasso weight from $\lambda = 7$ to 80. For each penalty level, we find the optimal threshold d^*. Next, we select the model with the lowest AIC or BIC and check the goodness of fit with the chi-square test of Section 2.3.

The left plot of Figure 4 shows the series of thresholds, $|\boldsymbol{d}|$, sorted by ascending order. The right graph shows the evolution of penalized deviances for $\lambda = 10$. Table 2 reports the statistics of smoothing for different Lasso penalties. When $\lambda = 10$, the lowest penalized deviance is achieved with 22 non-null wavelet coefficients, and the chi-square test does not reject the smoothed curve. The relation between the penalty weight and the number of wavelet coefficients is a staircase function of λ. Figure 5 compares the smoothed curves of log-mortality rates for different levels of penalty. Visually, smoothed curves built with 22 ($\lambda = 25$) or 95 ($\lambda = 8$) wavelets do not present significant differences. The Lasso penalty and the BIC leads to the selection of same wavelets. For this reason, we use BIC to measure goodness of fit in the remainder of this work.

3. Mortality projection

3.1. Shrinkage of the mortality surface

Let us now consider forces of mortality dynamically over time. As demonstrated in Section 2, the wavelet shrinkage is an efficient procedure for reducing the information contained in one curve of log forces of mortality. For a given calendar year, this information is carried by a few wavelet coefficients compared to the initial size of the curve. The natural question that arises is whether the set of relevant wavelets varies with time. If it remains stable over time and

https://doi.org/10.1017/asb.2020.18 Published online by Cambridge University Press

TABLE 2

GOODNESS-OF-FIT STATISTICS, YEAR 2015 FOR DIFFERENT LASSO PENALTIES.

Lasso penalty, λ	p: # of $\widehat{d}_{j,k} \neq 0$	d^*	log-likelihood	D	AIC	BIC	S_t	Rejection of H0 ?
7	101	0.01	−398.85	8.75	999.7	1272.45	8.68	No
8	95	0.01	−399.21	9.47	988.42	1244.96	9.4	No
9	95	0.01	−399.21	9.47	988.42	1244.96	9.4	No
10	22	0.42	−444.48	100.56	932.95	992.36	99.95	No
25	22	0.42	−444.48	100.56	932.95	992.36	99.95	No
50	22	0.42	−444.48	100.56	932.95	992.36	99.95	No
80	8	4.47	−2629.32	4472.37	5274.63	5296.23	3672.6	Yes

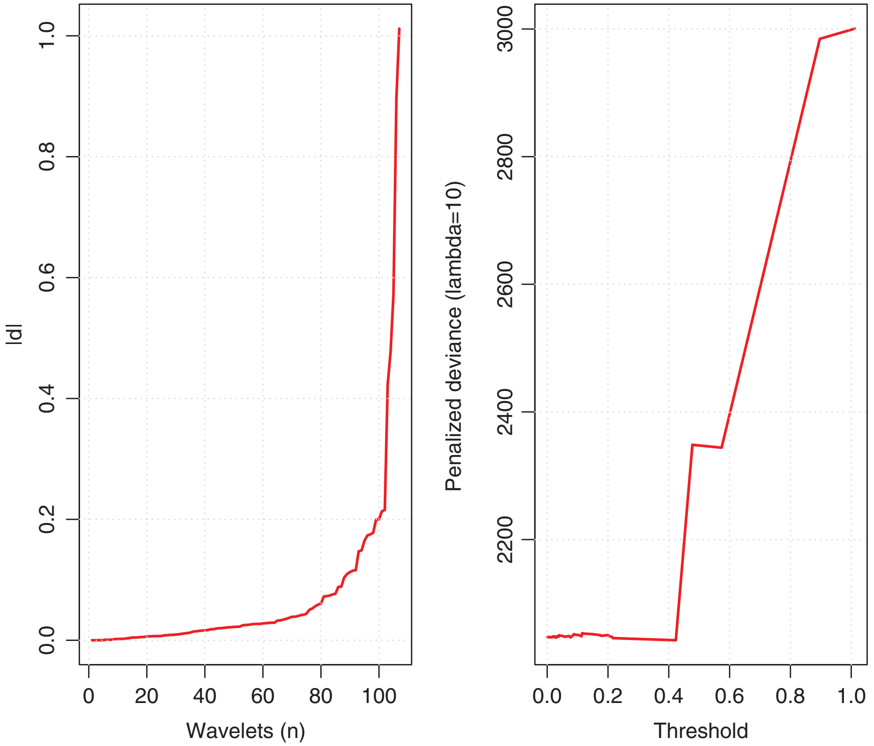

FIGURE 4: Smoothing with Lasso, year 2015. Left plot: series of thresholds, $|\boldsymbol{d}|$, sorted by ascending order. Right graph penalized deviances. Lasso parameter: $\lambda = 10$.

the wavelet coefficients exhibit some regular trend, extrapolating these trends would allow us to forecast the evolution of mortality rates.

To answer these questions, we perform a wavelet decomposition of Belgian mortality curves (both genders) from 1965 to 2015 for ages ranging from x_m=0 to x_M=109 years. Figure 6 shows the surface of the 128 wavelet coefficients for the period 1965–2015. On this graph, wavelet coefficients are sorted according

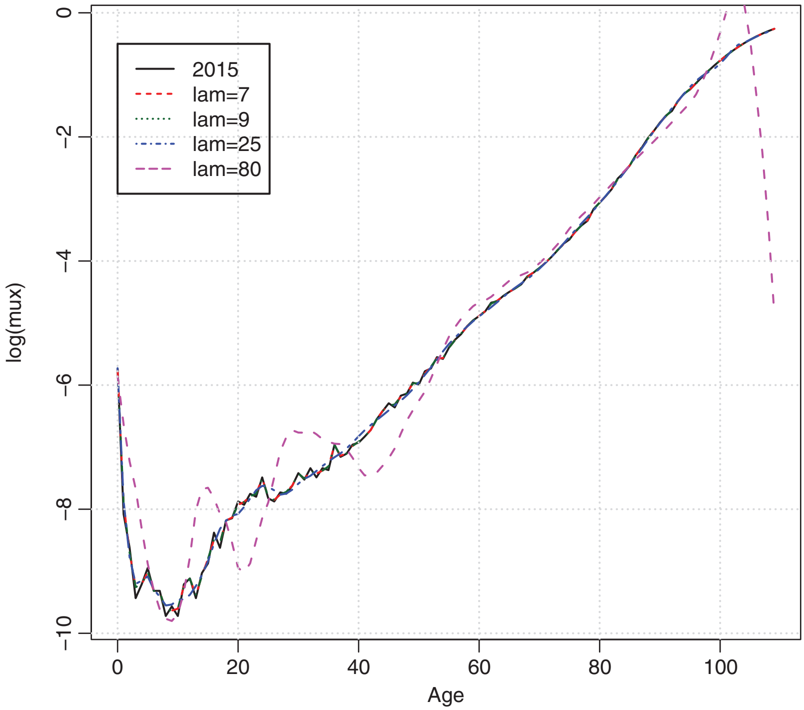

FIGURE 5: Smoothing with Lasso, year 2015. Comparison of smoothed curves for different penalty weights.

to their average value. Figure 6 reveals that most wavelet coefficients are null or close to zero. Another remarkable observation is that wavelets with coefficients close to 0 are the same for every calendar year. In other words, a small number of wavelets can be used for reconstructing all curves of mortality between 1965 and 2015.

Most of wavelet coefficients are null for two reasons. The first one is related to the intrinsic feature of wavelet functions. The wavelet analysis converts a signal, here the curve of log forces of mortality, into a linear combination of orthogonal and compactly supported functions. Compactness implies that wavelet functions are non-null only on a small subdomain of $\mathbb{R}$. This feature is visible in Figure A.1 of Appendix A.1. Therefore, we can decompose locally a signal over a time interval into a finite number of wavelets, and these wavelets will not interfere with the decomposition of the same signal to some later time interval. This is a great advantage of wavelet transform over Fourier's transform. In a Fourier's transform, we project the signal over orthogonal sinusoidal functions. These basis functions are non-null over $\mathbb{R}$, excepted for a countable number of points. Therefore, the local decomposition of a signal in a Fourier's basis over a time interval is pertubated by the projection of the signal over the entire timeline. The consequence of this lack of compactness is that we need more Fourier than wavelet basis functions to explain the same signal. The

https://doi.org/10.1017/asb.2020.18 Published online by Cambridge University Press

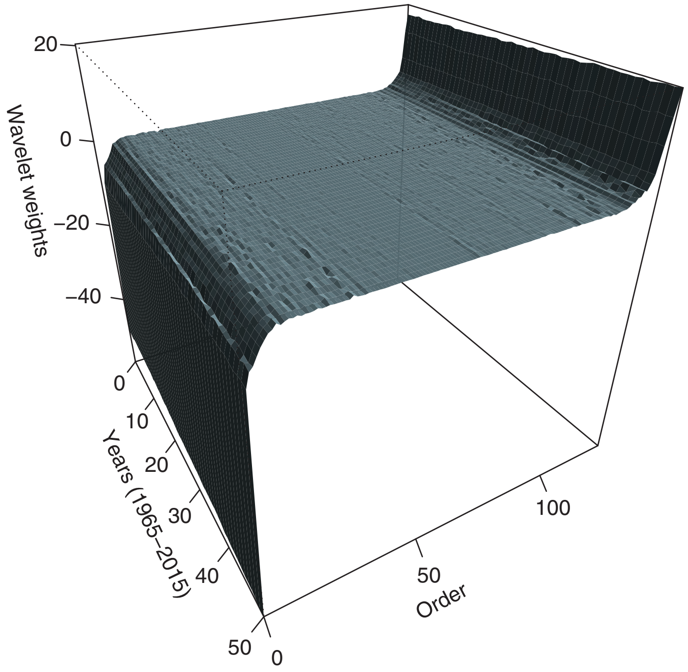

FIGURE 6: Surface of wavelet coefficients, period: 1965–2015.

second reason is related to the properties of the Daubechies wavelet. These wavelets have vanishing moments up to order $M=4$. It offers then a sparse representation of polynomials up to the fourth degree. Given that the curve of log forces of mortality for ages between 30 and 80 years old may be fitted by a polynomial function of low order, most of wavelets coefficients explaining mortality trends over this range of ages are null.

To select the optimal number of wavelets, we apply a hard thresholding based on the average of wavelet coefficients. Let us, respectively, denote by T and d^*, the number of years in the data set and the threshold. If

$$\frac{1}{T}\left|\sum_{t=1}^{T} d_{j,k}(t)\right| < d^* \qquad j \in \{0, ..., J-1\}, k \in \{0, ..., 2^j - 1\},$$

then we set $\widehat{d}_{j,k}(t)=0$ for all $t \in \{1, ..., T\}$. The statistical significance of each smoothed mortality curves could be checked with the help of the chi-square testing procedure described above. However, in practice, this procedure fails to produce a parsimonious representation for all years. Whatever the level of thresholding, there is always a limited number of smoothed curves that fail the chi-square test due to abnormal shocks on mortality curves like heatwaves, flu epidemics, or simply because of the volatility of estimators of log forces of mortality. Explaining these temporary perturbations requires to add a few wavelets

https://doi.org/10.1017/asb.2020.18 Published online by Cambridge University Press

that are useless for explaining mortality observed during normal years. Here, we opt for an alternative method based on BIC and AIC that is more tolerant with respect to deviations between smoothed and original log forces of mortality.

This approach is based on the Poisson regression model introduced in Section 2.5. The Poisson log-likelihood denoted by $\ln \mathcal{L}_{Pois}(d^*, \boldsymbol{\mu}^S)$ is defined in (2.5) for a single year. Here, this log-likelihood is summed up over all years to get

$$\ln \mathcal{L}_{Pois}(d^*, \boldsymbol{\mu}^S) = \sum_{t=1}^{T} \sum_{x=x_m}^{x_M} \left(n_{t,x} \ln \left(\mu^S(t,x) E_{t,x} \right) - \mu^S(t,x) E_{t,x} - \ln \left(n_{t,x}! \right) \right).$$

The scaled deviance D^* is defined in the same way:

$$D^*(\widehat{\boldsymbol{\mu}}, \boldsymbol{\mu}^S) = 2 \sum_{t=1}^{T} \sum_{x=x_m}^{x_M} E_{t,x} \left(\mu(t,x) \ln \left(\frac{\mu(t,x)}{\mu^S(t,x)} \right) + \mu^S(t,x) - \mu(t,x) \right).$$

This Poisson log-likelihood is used to calculate the AIC and BIC:

$$\mathrm{AIC} = 2(T\,p) - 2 \ln \mathcal{L}_{Poiss}(d^*, \boldsymbol{\mu}^S),$$
$$\mathrm{BIC} = \ln(T\,n)\,(T\,p) - 2 \ln \mathcal{L}_{Poiss}(d^*, \boldsymbol{\mu}^S).$$

The optimal threshold level d^* is either the one minimizing the AIC or the BIC, depending upon the sought level of sparsity. Table 3 contains the AIC, BIC, and log-likelihood for a range of thresholds. The lowest AIC and BIC are obtained with $d^* \in [0.12 \;\; 0.24]$, and the number of non-null wavelet coefficients is 24. We have also computed the chi-square statistics for each year between 1965 and 2015. This statistics is computed with $x_{min} = 0$ and $x_{max} = 109$. The third column of Table 3 reports the percentage of smoothed curves that pass the chi-square test of Section 2.3. For a threshold level in [0.12, 0.24], 59% of smoothed curves are accepted. Table 3 reveals that thresholding wavelet coefficients with an absolute value higher than 0.30 cause a jump in the deviance. Figure 7 shows the heat map of differences between the smoothed and observed log-mortality rates. We observe in the upper part of this map an increasing straight line. This line corresponds to cohorts born around the second world war. These cohorts experience a higher mortality rates than the other ones.

3.2. Mortality forecasting

In the previous section, we have seen that the information contained in the mortality surface from 1965 to 2015 may be summarized by a surface of p =24 wavelet coefficients observed over 51 years. Figures 8, 9, 10, and 11 show the evolution over time of these 24 time series of wavelets coefficients. For most of them, we observe a clear linear increasing or decreasing trend. Based on this remarkable observation, we regress linearly these coefficients with respect to

https://doi.org/10.1017/asb.2020.18 Published online by Cambridge University Press

TABLE 3

GOODNESS-OF-FIT STATISTICS OF GOODNESS OF FIT FOR SHRINKED MORTALITY CURVES, PERIOD 1965–2015.

	Threshold d^*	$p=\#$ of $\widehat{d}_{j,k}\neq 0$	Percentage of chi-square tests passed	$\ln\mathcal{L}(d^*,\boldsymbol{\mu}^S)$	$D^*(\boldsymbol{\mu},\boldsymbol{\mu}^S)$	AIC	BIC
1	0.03	33	0.76	$-23,103.12$	4920.3	49,572.24	60,734.41
2	0.06	30	0.76	$-23,167.18$	5048.34	49,394.35	59,541.78
3	0.09	28	0.76	$-23,221.25$	5157.44	49,298.5	58,769.44
4	0.12	24	0.59	$-23,785.23$	6283.42	50,018.46	58,136.4
5	0.15	24	0.59	$-23,785.23$	6283.42	50,018.46	58,136.4
6	0.18	24	0.59	$-23,785.23$	6283.42	50,018.46	58,136.4
7	0.21	24	0.59	$-23,785.23$	6283.42	50,018.46	58,136.4
8	0.24	24	0.59	$-23,785.23$	6283.42	50,018.46	58,136.4
9	0.27	22	0.31	$-24,318.95$	7350.23	50,881.89	58,323.34
10	0.3	21	0	$-29,637.46$	17,987.36	61,416.93	68,520.13

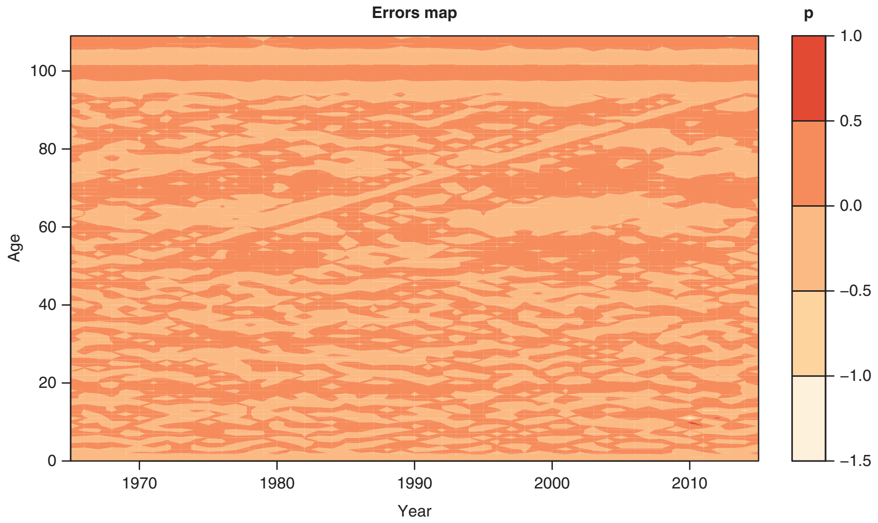

FIGURE 7: Heat map of differences between the smoothed and observed log-mortality rates.

time. Precisely, all non-null wavelet coefficients $\widehat{d}_{j,k}(t)\in\widehat{\boldsymbol{d}}(t)$ obey the following dynamics:

$$\widehat{\boldsymbol{D}}(t)=\boldsymbol{\alpha}+\boldsymbol{\beta}t+\boldsymbol{\mathcal{E}}(t),$$

where $\boldsymbol{\mathcal{E}}(t)$ is mutually independent, multivariate normal random vectors with a zero mean vector variance–covariance matrix $\boldsymbol{\Sigma}$ of dimension $p\times p$. The p vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ contain the intercepts and regression factors. Table 4 reports the statistics of goodness of fit for residuals. The Shapiro–Wilk test validates the assumption of normality (at 5% confidence level) for 22 time series out of 24.

https://doi.org/10.1017/asb.2020.18 Published online by Cambridge University Press

TABLE 4

STATISTICS ABOUT LINEAR REGRESSIONS OF THE 24 WAVELET COEFFICIENTS, PERIOD 1965–2015.

Wavelets n°	p-values Shapiro	Gaussian residuals	R^2	p-values Llung–Box	Independent increments
1	0.06	Yes	0.98	0	No
2	0.83	Yes	0.85	0	No
3	0.46	Yes	0.77	0	No
4	0.62	Yes	0.97	0.23	Yes
5	0.64	Yes	0.69	0	No
6	0.03	No	0.14	0.04	No
7	0.21	Yes	0.76	0	No
8	0.72	Yes	0.98	0.02	No
9	0.53	Yes	0.84	0.92	Yes
10	0.37	Yes	0.83	0.14	Yes
11	0.36	Yes	0.77	0.22	Yes
12	0.17	Yes	0.27	0	No
13	0	No	0.72	0	No
14	0.6	Yes	0.97	0.06	Yes
15	0.8	Yes	0.81	0.19	Yes
16	0.9	Yes	0.36	0.5	Yes
17	0.5	Yes	0.05	0.18	Yes
18	0.82	Yes	0.96	0	No
19	0.64	Yes	0.81	0.31	Yes
20	0.99	Yes	0.04	0.23	Yes
21	0.1	Yes	0.11	0.34	Yes
22	0.85	Yes	0.97	0	No
23	0.47	Yes	0.97	0	No
24	0.18	Yes	0.74	0.8	Yes

Furthermore, 62.5% of linear regressions have an R^2 above 75%, while 16.66% of regressions have an R^2 smaller than 25%. For 12 regressions, the Llung–Box test validates the assumption of independent increments (5% confidence level). These statistics confirm that linear regressions succeed to explain the evolution of wavelets coefficients.

Remark. A better fit can be achieved with an auto-regressive model of the form

$$\widehat{\boldsymbol{d}}(t) = \boldsymbol{\alpha} + \boldsymbol{\beta}^1\, t + \boldsymbol{\beta}^2 \odot \widehat{\boldsymbol{d}}(t) + \epsilon(t),$$

where $\odot$ is the Hadamard product and $\boldsymbol{\alpha}$, $\boldsymbol{\beta}^1$, and $\boldsymbol{\beta}^2$ are the p vectors. For 95.83% of wavelets, the Shapiro–Wilk test does not reject the assumption of normality. We validate in 95.83% of cases, the assumption of independent increments. However, forecasting life expectancy with this model slightly decreases from 2015 to 2020 before increasing again, and the auto-regressive model has been discarded for this reason.

https://doi.org/10.1017/asb.2020.18 Published online by Cambridge University Press

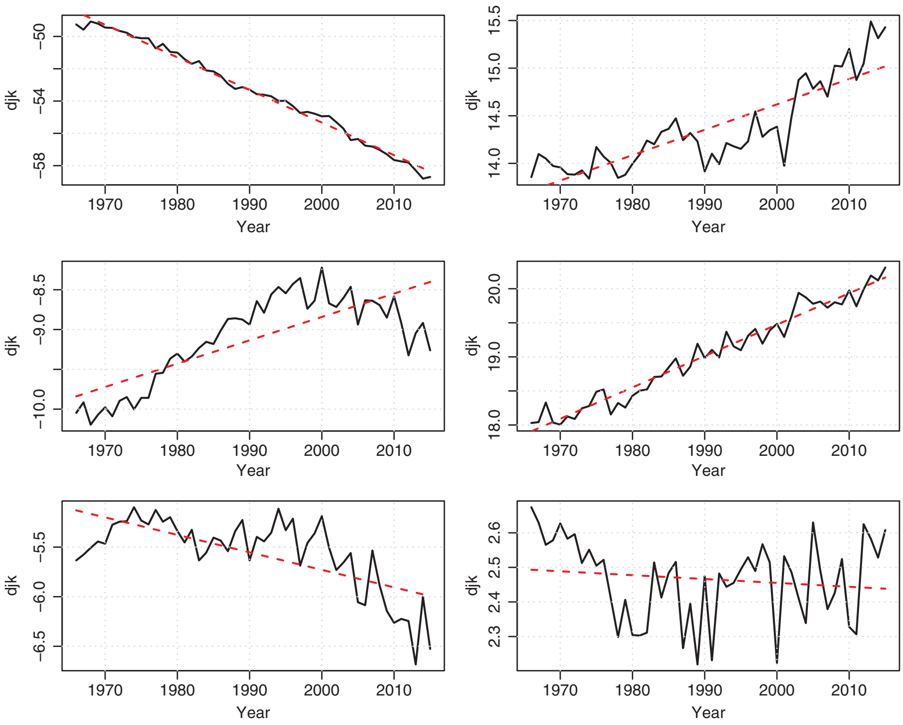

FIGURE 8: Time series of coefficients of wavelets 1–6. The red dotted line is the linear regression.

By linear extrapolation, we can simulate future log-mortality rates. The left plot of Figure 12 compares 2015 log forces of mortality to the average simulated mortality rates in 2046, computed with 1000 simulations. The mid plot shows the evolution of average future log-mortality rates from 2016 to 2046. These two graphs clearly emphasize that mortality rates will continue to decline according to the wavelet model. The right panel of Figure 12 displays 1000 simulated curves of mortality rates for the year 2046. This underlines the ability of our model to generate various scenarios of mortality. To better grasp the amplitude of this reduction of mortality, we have computed cross-sectional life expectancy from 2016 to 2045. Table 5 reports these statistics at birth and at ages 20, 40, 60, and 80. The wavelet model forecast an increase of the life expectancy at birth from 80.61 in 2016 up to 85.37 years in 2045. At age 80, the average gain of longevity over this period is around 2 years.

3.3. Validation by backtesting

In order to benchmark the predictive power of the wavelet approach, we compare it to competitors proposed (Lee and Carter, 1992; Renshaw and

https://doi.org/10.1017/asb.2020.18 Published online by Cambridge University Press

TABLE 5

EVOLUTION OF FORECAST CROSS-SECTIONAL LIFE EXPECTANCIES FROM 2016 TO 2045.

	Cross-sectional life expectancy				
Year	At birth	Age 20	Age 40	Age 60	Age 80
2016	80.61	61.03	41.69	23.58	8.62
2020	81.34	61.7	42.33	24.14	8.9
2025	82.2	62.51	43.1	24.81	9.25
2030	83.05	63.31	43.86	25.47	9.61
2035	83.84	64.06	44.58	26.1	9.94
2040	84.62	64.81	45.3	26.72	10.29
2045	85.37	65.53	45.99	27.33	10.63

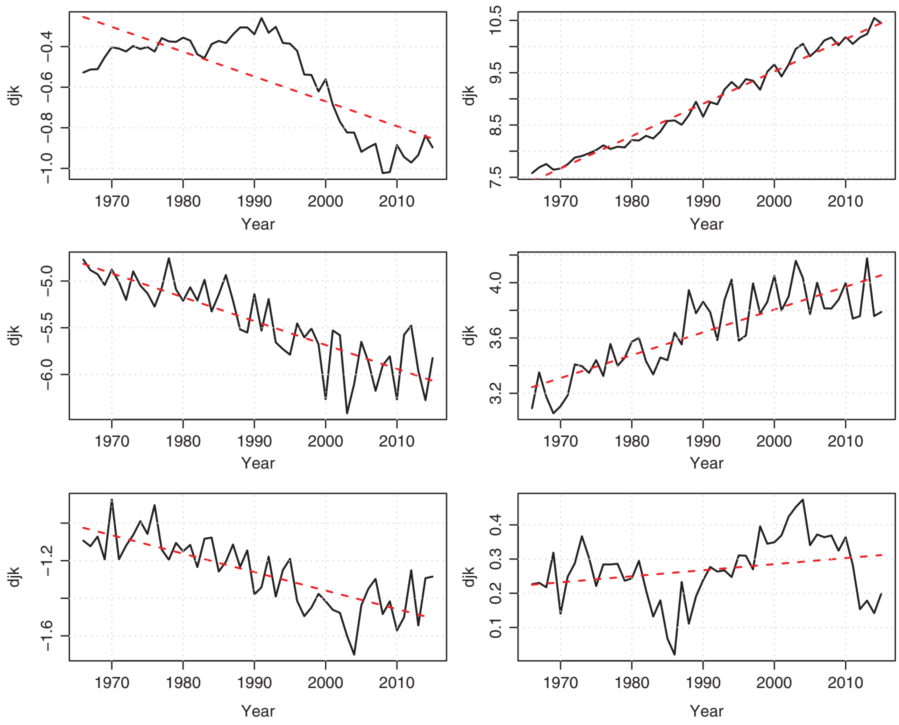

FIGURE 9: Time series of coefficients of wavelets 7–12. The red dotted line is the linear regression.

Haberman, 2003, 2006; Cairns *et al.*, 2006, 2009). In the Lee–Carter (henceforth referred to as LC) model, the log mortality rates are related to ages as follows:

$$\ln \mu(t, x) = \alpha_x + \beta_x \kappa_t, \tag{3.1}$$

where $\alpha_x \in \mathbb{R}^{x_{max}}$ is the constant vector representing the permanent impact of age on mortality. Whereas $\beta_x \in \mathbb{R}^{x_{max}}$ is the constant vector that quantifies the

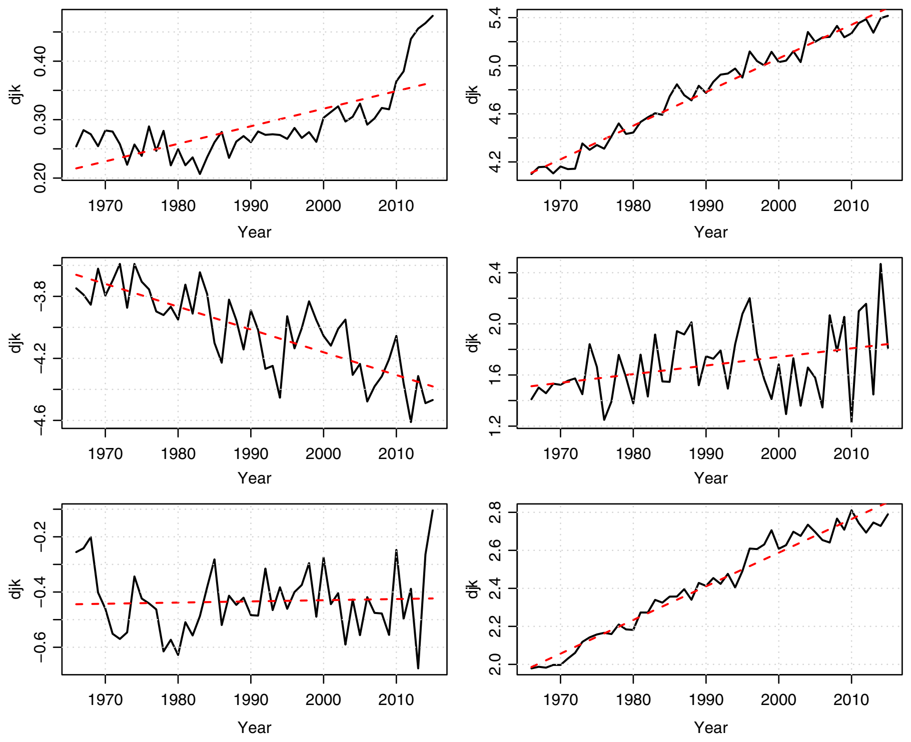

FIGURE 10: Time series of coefficients of wavelets 13–18. The red dotted line is the linear regression.

marginal effect of the latent factor κ_t on mortality at each age. κ_t is the latent process that describes the evolution of mortality over time. We consider a bivariate extension (henceforth referred to as RH 2D) as proposed by Renshaw and Haberman (2006). In the RH 2D, the log force of mortality is a linear combination of two time latent factors, κ_t^1 and κ_t^2, with covariates that depend on the age as follows:

$$\ln \mu(t, x) = \alpha_x + \beta_x^1 \kappa_t^1 + \beta_x^2 \kappa_t^2, \tag{3.2}$$

where β_x^1 and $\beta_x^2 \in \mathbb{R}^{x_{max}}$, and κ_t^1 and κ_t^2 are the latent processes. The next model that we consider (henceforth referred to as RH coh) adds a cohort effect in the dynamic of log force of mortality:

$$\ln \mu(t, x) = \alpha_x + \beta_x^1 \kappa_t + \beta^2 \gamma_{t-x}, \tag{3.3}$$

where $\beta^2 \in \mathbb{R}$ represents the marginal effect of a generation factor γ_{t-x} on mortality. In a fourth test, we fit the CBD (Cairns, Blake and Dowd, 2006) model for which:

$$\text{logit}\, q(t, x) = \kappa_t^{(1)} + (x - \bar{x})\, \kappa_t^{(2)}, \tag{3.4}$$

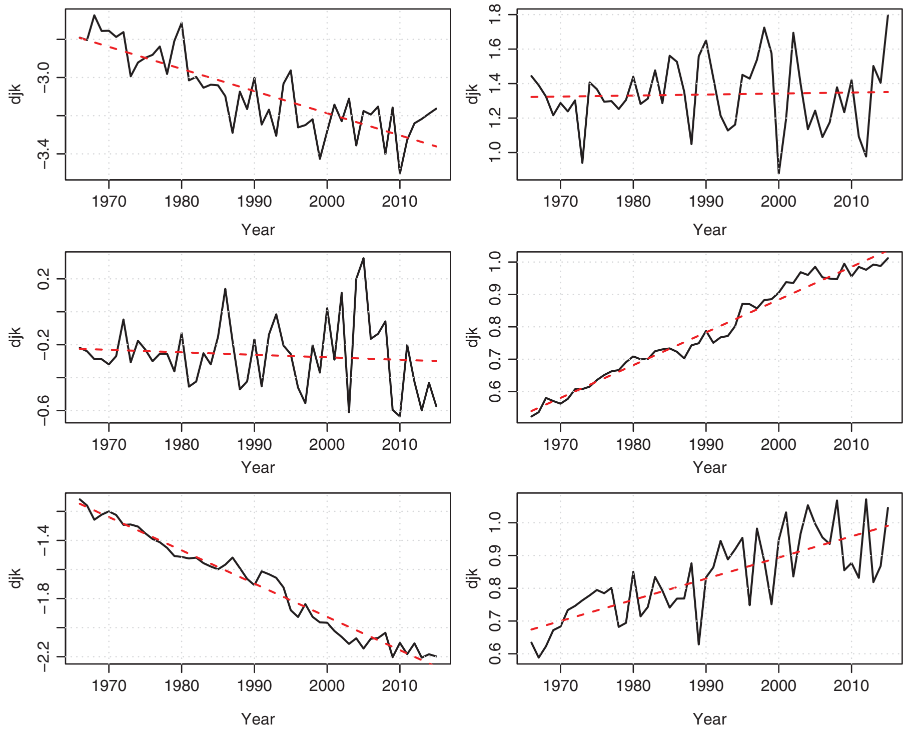

FIGURE 11: Time series of coefficients of wavelets 19–24. The red dotted line is the linear regression.

where $\bar{x}$ is the average of ages. The last model, proposed by Cairns *et al.* (2009), adds a cohort effect to the CBD model and is referred to as M7:

$$\operatorname{logit} q(t, x) = \kappa_t^{(1)} + (x - \bar{x})\, \kappa_t^{(2)} + \left((x - \bar{x})^2 - \widehat{\sigma}_x^2\right) \kappa_t^{(3)} + \gamma_{t-x}. \tag{3.5}$$

In this equation, $\widehat{\sigma}_x^2$ is the variance of ages and γ_{t-x} is the generation factor. The LC, RH 2D, RH coh, CBD, and M7 models are estimated by log-likelihood maximization with the R package StMoMo.

The six models are fitted to Belgian mortality curves (both genders) from 1965 to 2005 for ages ranging from 0 to 90 years (including older ages causes numerical instabilities for the RH coh model). Next, we forecast log forces of mortality for years 2006 to 2015 and compare with the observed ones. For the RH coh model, we exclude all cohorts that have fewer than three observations. Table 6 reports the sum of squared errors between the predicted and observed log-mortality rates. Whatever the year, the wavelet model outperforms its three competitors. The RH coh model has the lowest performance due to the difficulty to extrapolate the time series of γ_{t-x}, which is nonlinear over the considered period.

https://doi.org/10.1017/asb.2020.18 Published online by Cambridge University Press

TABLE 6

SUM OF SQUARED ERRORS BETWEEN THE FORECAST AND REAL LOG-MORTALITY RATES.

	Wavelet (24 coefficients)	LC	RH 2D	RH coh	CBD	M7
2006	2.11	2.59	1.96	4.32	64.57	41.83
2007	1.14	1.43	1.37	12.51	67.44	48.87
2008	1.29	1.62	1.21	29.54	65.32	56.02
2009	1.77	2.77	2.59	53.47	67.54	59.41
2010	4.01	5.12	4.91	87.54	69.31	76.11
2011	3.01	3.77	4.03	124.45	72.75	71.38
2012	5.18	6.3	6.4	179.46	74.49	78.79
2013	3.81	5.42	4.86	235.79	71.57	97.82
2014	4.00	5.49	5.48	294.46	70.09	105.42
2015	4.69	5.95	5.41	380.99	71.2	107.2

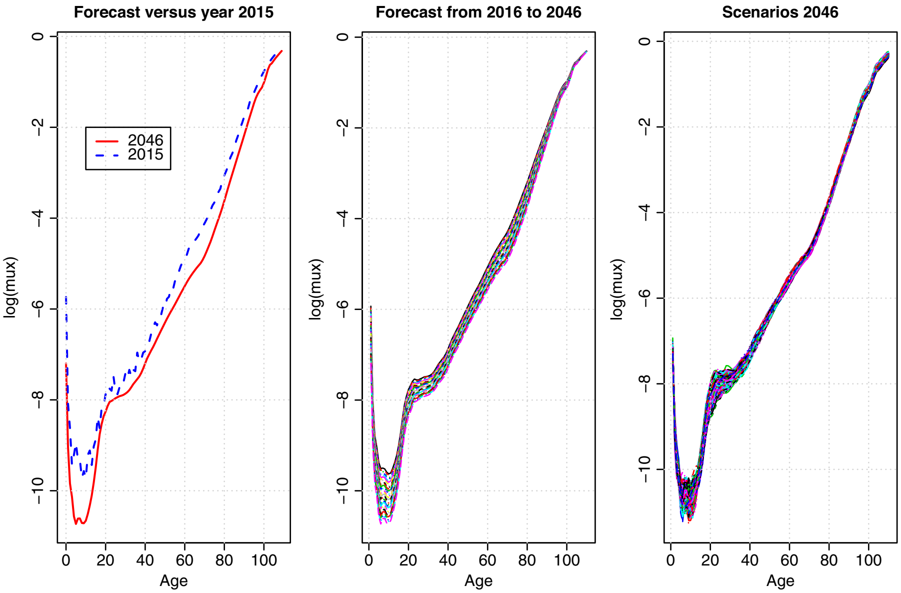

FIGURE 12: Left plot: log forces of mortality in 2015 and 2046. Mid plot: average log forces of mortality from 2016 to 2046. Right plot: 1000 scenarios of log-mortality rates, year 2046.

Table 7 presents the log-likelihoods, deviances, and AIC of the six models. The lowest deviance is obtained with the wavelet model. Notice, however, that due to its high number of parameters, the best AIC is obtained with an RH 2D model. These results confirm that the wavelet model is a reliable alternative to existing approaches for modeling log-mortality rates.

https://doi.org/10.1017/asb.2020.18 Published online by Cambridge University Press

TABLE 7

GOODNESS-OF-FIT STATISTICS, PERIOD 1965–2005.

	$\ln \mathcal{L}(d^*, \boldsymbol{\mu}^S)$	$D^*(\boldsymbol{\mu}, \boldsymbol{\mu}^S)$	AIC	p: # of parameters
Wavelets	$-16,636.68$	4734.78	35,241.35	984
LC	$-17,182.62$	5729.67	34,807.24	221
RH 2D	$-16,716$	4863.88	34,130.00	349
Rensh. Haber.	$-248,575.96$	5855.26	497,841.93	345
CBD	$-263,059.84$	492,684.73	526,283.69	82
M7	$-415,382.74$	166,349.05	831,255.47	245

TABLE 8

US POPULATION. SUM OF SQUARED ERRORS BETWEEN THE FORECAST AND REAL LOG-MORTALITY RATES.

	Wavelet (24 coefficients)	LC	RH 2D	RH coh	CBD	M7
2006	0.26	0.52	0.46	36.52	43.62	48.69
2007	0.25	0.54	0.46	65.87	44.73	57.31
2008	0.29	0.71	0.51	107.93	44.57	70.68
2009	0.54	0.93	0.7	162.91	44.46	81.73
2010	0.62	1.1	0.79	233.54	44.47	97.75
2011	0.6	1.05	0.83	320.87	46.39	109.47
2012	0.68	1.14	0.96	427.44	47.85	124.44
2013	0.8	1.2	1.1	553.83	49.36	136.64
2014	1.03	1.23	1.23	694.87	51.32	151.17
2015	1.44	1.32	1.64	861.92	57.15	160.82

3.4. US and UK populations

We fit a model with 24 wavelets to the US and UK populations and benchmark it to LC, RH 2D, RH coh, CBD, and M7 models. As for the Belgian population, the χ^2 statistics does not reject most of smoothed curves obtained with this approach. The six models are fitted to mortality rates (both genders) from 1965 to 2005 for ages ranging from 0 to 90 years. Next, we forecast log forces of mortality for years 2006 to 2015. Tables 8 and 9 report the sum of squared errors between the modeled and observed log-mortality rates. The best performance is achieved by the Wavelet model. Surprisingly, the cohort model , RH coh, has an excellent predictive power for the UK population compared to its performance with the Belgian and US datasets.

As in Section 3.2, we forecast log-mortality rates from 2016 to 2046 for the US and UK populations. For this purpose, the wavelets model is estimated with data from 1965 to 2015 and for ages from 0 to 109 years. Tables 10 and 11 report the forecast cross-sectional life expectancies. UK figures are comparable

TABLE 9

UK POPULATION. SUM OF SQUARED ERRORS BETWEEN THE FORECAST AND REAL LOG-MORTALITY RATES.

	Wavelet (24 coefficients)	LC	RH 2D	RH coh	CBD	M7
2006	0.45	1.58	0.64	0.53	74.86	42.67
2007	0.37	1.53	0.58	0.55	76.79	49.9
2008	0.76	2.18	0.88	0.79	78.88	58.97
2009	0.63	1.61	0.79	0.65	76.38	72.51
2010	1.14	2.17	1.31	0.85	76.56	80.78
2011	1.78	2.5	1.95	0.96	74.75	97.33
2012	2.15	2.98	2.36	1.74	75.64	103.88
2013	1.76	2.77	2.19	2.19	77.16	118.45
2014	1.8	3.12	2.21	2.98	82.1	124.52
2015	1.52	3	1.95	4.68	85.23	138.8

TABLE 10

US POPULATION: EVOLUTION OF FORECAST CROSS-SECTIONAL LIFE EXPECTANCIES FROM 2016 TO 2045.

	Cross-sectional life expectancy (US)				
Year	At birth	Age 20	Age 40	Age 60	Age 80
2016	79.01	59.71	40.76	23.09	8.9
2020	79.61	60.25	41.26	23.5	9.08
2025	80.32	60.89	41.86	23.98	9.29
2030	81.01	61.51	42.45	24.47	9.51
2035	81.68	62.13	43.03	24.94	9.73
2040	82.32	62.71	43.58	25.39	9.94
2045	82.95	63.3	44.13	25.84	10.15

TABLE 11

UK POPULATION: EVOLUTION OF FORECAST CROSS-SECTIONAL LIFE EXPECTANCIES FROM 2016 TO 2045.

	Cross-sectional life expectancy (UK)				
Year	At birth	Age 20	Age 40	Age 60	Age 80
2016	80.47	60.93	41.56	23.25	8.55
2020	81.16	61.57	42.19	23.78	8.79
2025	81.96	62.31	42.91	24.39	9.08
2030	82.73	63.04	43.62	25	9.38
2035	83.46	63.73	44.29	25.57	9.66
2040	84.21	64.44	44.98	26.17	9.97
2045	84.87	65.07	45.59	26.7	10.24

https://doi.org/10.1017/asb.2020.18 Published online by Cambridge University Press

with Belgian life expectations of Table 5. Whereas the wavelets model predicts a lower increase of life expectancy at birth for the US than for UK or Belgium. In 2045, the life expectancy at birth in Belgium is around 85 years, whereas it is only 83 years for US.

4. CONCLUSIONS

The numerical illustrations performed in this paper suggest that wavelets are powerful tools for analyzing mortality trends. The first part of this work proposes two alternative approaches to optimize the smoothing of log-mortality rates by wavelets shrinkage. The first method is based on a chi-square test built with a Gaussian approximation of log forces of mortality. The second one uses a penalized Poisson likelihood to find a compromise between statistical relevance and sparsity. Both approaches reveal that a mortality curve with 110 death rates may be summarized by around 20 wavelet coefficients.

The second part of this article focuses on mortality forecasting. We show that the set of significant wavelet coefficients is stable over the last half century. A small number of wavelets can be used for reconstructing all curves of mortality between 1965 and 2015. Furthermore, most of these coefficients exhibit clear trends that can be extrapolated with a basic multivariate linear regression. This technique allows us to predict that the cross-sectional life expectancy for the Belgian population (both gender) will increase on average up to 85.37 years in 2045. In the last section of this work, we have demonstrated that the wavelet model widely outperforms other popular actuarial models fitted to Belgian, US, and UK populations, both in terms of goodness of fit and predictive errors.

ACKNOWLEDGMENT

The first author is grateful to the “Fonds de la Recherche Scientifique - FNRS” for financial support under Grant number 33658713.

REFERENCES

BESBEAS, P., DE FEIS, I. and SAPATINAS T. (2004) A comparative simulation study of wavelet shrinkage estimators for Poisson counts. *International Statistical Review*, **72**, 209–237.

BROUHNS, N., DENUIT, M. and VERMUNT, J.K. (2002) A Poisson log-bilinear regression approach to the construction of projected lifetables. *Insurance: Mathematics and Economics*, **31**, 373–393.

CAIRNS, A.J.G., BLAKE, D. and DOWD, K. (2006) A two-factor model for stochastic mortality with parameter uncertainty: Theory and calibration. *Journal of Risk and Insurance*, **73**(4), 687–718.

CAIRNS, A.J.G., BLAKE, D., DOWD, K., COUGHLAN, G.D., EPSTEIN, D., ONG, A. and BALEVICH, I. (2009) A quantitative comparison of stochastic mortality models using data from England and Wales and the United States. *North American Actuarial Journal*, **13**(1), 1–35.

https://doi.org/10.1017/asb.2020.18 Published online by Cambridge University Press

COCHRAN, W.G. (1952) The χ^2 test of goodness of fit. *Annals of Mathematical Statistics*, **23**, 315–345.

DENUIT, M., HAINAUT, D. and TRUFIN J. (2019a) *Effective Statistical Learning Methods for Actuaries – Volume 1: GLM and Extensions*. Springer Actuarial Lecture Notes Series. Springer Nature Switzerland.

DENUIT, M., HAINAUT, D. and TRUFIN, J. (2019b) *Effective Statistical Learning Methods for Actuaries – Volume 3: Neural Networks and Extensions*. Springer Actuarial Lecture Notes Series. Springer Nature Switzerland.

DENUIT, M. and LEGRAND, C. (2018) Risk classification in life and health insurance: Extension to continuous covariates. *European Actuarial Journal* **8**, 245–255.

DONOHO, D.L. and JOHNSTONE, I.M. (1994) Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425–455.

DONOHO, D.L. and JOHNSTONE, I.M. (1995) Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Assocation* **90**, 1200–1224.

GBARI, S., POULAIN, M., DAL, L. and DENUIT, M. (2017) Extreme value analysis of mortality at the oldest ages: A case study based on individual ages at death. *North American Actuarial Journal*, **21**(3), 397–416.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2016) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer-Verlag New York.

HYNDMAN, R.J. AND ULLAH, MD.S. (2007) Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics and Data Analysis* **51**, 4942–4956.

JURADO, F.M. and SAMPERE, I.B. (2019) Using wavelet techniques to approximate the subjacent risk of death. In: *Modern Mathematics and Mechanics* (eds. V.A. Sadovnichiy and M.Z. Zgurovsky), Chapter 28, pp. 541–557. Springer International Publishing.

LEE, R.D. and CARTER, L. (1992) Modelling and forecasting the time series of US mortality. *Journal of the American Statistical Association* **87**, 659–671.

MALLAT, S.G. (1989a) Multiresolution approximations and wavelet orthonormal bases of $L_2(\mathbb{R})$. *Transactions of the American Mathematical Society* **315**, 69–87.

MALLAT, S.G. (1989b) A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **11**, 674–693.

MORILLAS, F., BAEZA, I. and PAVIA, J.M. (2016) Risk of death: A two-step method using wavelets and piecewise harmonic interpolation. *Estadistica Espanola* **58**, 245–264.

NICKOLAS, P. (2017) *Wavelets: A Student Guide*. Cambridge University Press.

PITACCO, E., DENUIT, M., HABERMAN, S. and OLIVIERI, A. (2009) *Modelling Longevity Dynamics for Pensions and Annuity Business*. New York: Oxford University Press.

RENSHAW, A.E. and HABERMAN, S. (2003) Lee-Carter mortality forecasting with age-specific enhancement. *Insurance: Mathematics and Economics* **33**, 255–272.

RENSHAW, A.E. and HABERMAN, S. (2006) A cohort-based extension to the Lee-Carter model for mortality reduction factors. *Insurance: Mathematics and Economics* **38**, 556–570.

RENSHAW, A.E., HABERMAN, S. and HATZOUPOULOS, P. (1996) The modelling of recent mortality trends in United Kingdom male assured lives. *British Actuarial Journal* **2**, 449–477.

STRANG, G. and NGUYEN, T. (1996) *Wavelets and Filter Banks*. Wellesley, MA.

TIBSHIRANI, R. (1996) Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society – Series B* **58**, 267–288.

WILMOTH, J.R., ANDREEV, K., JDANOV, D., GLEI, D.A. and RIE, T. (2019) Methods Protocol for the Human Mortality Database. https://www.mortality.org/Public/Docs/MethodsProtocol.pdf

DONATIEN HAINAUT (Corresponding author) and MICHEL DENUIT
Institute of Statistics, Biostatistis and Actuarial Science - ISBA
Louvain Institute of Data Analysis and Modeling - LIDAM
UCLouvain
B-1348 Louvain-la-Neuve, Belgium
E-Mails: donatien.hainaut@uclouvain.be; michel.denuit@uclouvain.be

https://doi.org/10.1017/asb.2020.18 Published online by Cambridge University Press

APPENDIX A. WAVELETS IN A NUTSHELL

A.1. Definition

The wavelets analysis consists to project a function on an orthornormal basis. We denote by $L^2(\mathbb{R})$ the space of square-integrable functions equipped with the inner product and the norm, respectively, defined by

$$\langle f, g\rangle = \int_{-\infty}^{+\infty} f(x)g(x)dx \qquad ||f|| = \sqrt{\langle f, f\rangle}\,,$$

for $f, g \in L^2$ ($\mathbb{R}$). Functions f and g are orthonormal if they are orthogonal $\langle f, g\rangle = 0$ and of norm equal to 1. $L^2(\mathbb{R})$ is an Hilbert space for this inner product. Recall that a Hilbert space is a complete inner product space, that is, all Cauchy sequences converge to a limit in this space. An orthornormal basis of $V \subset L^2(\mathbb{R})$ is a maximal subset $B = (f_k)_{k\in\mathbb{Z}}$ of orthonormal functions such that if $g \in V$ with $\langle g, f_n\rangle = 0$ for all $f_n \in B$, then $g = 0$.

Definition. *The multiresolution analysis consists of a collection of closed subspace of $L^2(\mathbb{R})$, noted $\left(V_j\right)_{j\in\mathbb{Z}}$ and of a scaling function, also referred to as father wavelet, $\phi \in V_0$, satisfying the following conditions. First, the function ϕ forms an orthonormal basis of V_0 by translation. Every function $f \in V_0$ can then be rewritten as an infinite sum:*

$$f(x) = \sum_{k\in\mathbb{Z}} \langle \phi(x-k), f(x)\rangle\, \phi(x-k).$$

Second, the spaces V_j are nested in the sense that $V_{j-1} \subset V_j$ for all $j \in \mathbb{Z}$ and is dense in $L^2(\mathbb{R})$. The only function belonging to all $\left(V_j\right)_{j\in\mathbb{Z}}$ is the null function. Finally, we have the following properties for all $j \in \mathbb{Z}$:

$$\begin{aligned} f(x) \in V_0 &\Longleftrightarrow f(2^j x) \in V_j\,, \\ f(x) \in V_0 &\Longleftrightarrow f(x-k) \in V_0. \end{aligned} \tag{A.1}$$

If sets $\left(V_j\right)_{j\in\mathbb{Z}}$ and scaling function ϕ satisfy the conditions of the multiresolution analysis, the norms of the projection of any function $f \in L^2(\mathbb{R})$ on V_j, noted P_jf for $j \in \mathbb{Z}$, satisfy the relation:

$$||P_jf|| \leq ||P_{j+1}f||\,,$$

and $\lim_{j\to\infty} P_jf = f$ given that $\left(V_j\right)_{j\in\mathbb{Z}}$ are dense. We can also show that $\lim_{j\to-\infty} P_jf = 0$. We refer the reader to Nickolas (2017) for details. Furthermore, conditions (A.1) imply that

$$\phi_{j,k}(x) = 2^{j/2}\phi\left(2^j x - k\right) \quad k \in \mathbb{Z}$$

are in V_j and forms an orthonormal basis of this set.

An example of father wavelet ϕ satisfying the conditions of the multiresolution analysis is

$$\phi(x) = \begin{cases} 1 & 0 \leq x < 1 \\ 0 & \text{otherwise.} \end{cases} \tag{A.2}$$

https://doi.org/10.1017/asb.2020.18 Published online by Cambridge University Press

Since $\left(\phi_{0,k}(x)\right)_{k\in\mathbb{Z}}$ and $\left(\phi_{j,k}(x)\right)_{k\in\mathbb{Z}}$, respectively, form an orthonormal basis of V_0 and V_j, V_j is the set of piecewise constant functions:

$$V_j = \left\{ f \in L^2\left(\mathbb{R}\right) : f \text{ is constant on } \left[\frac{k}{2^j}, \frac{k+1}{2^j}\right) \text{ for all } k \in \mathbb{Z} \right\}.$$

The projection of any function f of $L^2(\mathbb{R})$ on V_j is then its piecewise constant approximation. For this reason, the spaces V_j can be referred to as approximation spaces.

In the remainder, we denote by $W_j = V_j^{\perp}$ the orthogonal complement of V_j in V_{j+1}. Every function f of V_{j+1} admits therefore a unique representation as

$$f = f_j + f_j^{\perp},$$

where $f_j = P_j f \in V_j$ and $f_j^{\perp} \in W_j$. By induction, we can show that for each $i \in \mathbb{Z}$, $f \in V_i$ is expressible as a convergent series $f = \sum_{-\infty}^{i} w_j$, where $w_j \in W_j$. We say that V_i is the direct sum of W_j and denote it by

$$V_i = \overset{[}{j} = -\infty]i \oplus W_j.$$

We now define a wavelet as follows.

Definition. *A wavelet is a function* ψ *of* $L^2(\mathbb{R})$ *whose scaled, dilated, and translated copies*

$$\psi_{j,k}(x) = 2^{j/2}\psi\left(2^j x - k\right) \quad k, j \in \mathbb{Z} \tag{A.3}$$

form an orthornormal basis of $L^2(\mathbb{R})$.

The link between the multiresolution analysis and wavelets comes from the following theorem. For a proof, see Theorem 6.6 in Nickolas (2017).

Theorem. *Let us consider spaces* $\left(V_j\right)_{j\in\mathbb{Z}}$ *and a scaling function* $\phi \in V_0$ *that form a multiresolution analysis. If* $\left(\psi_{j,k}\right)_{k\in\mathbb{Z}} \in V_{j+1}$ *is a family of functions such that*

$$\{\phi_{j,k} : k \in \mathbb{Z}\} \cup \{\psi_{j,k} : k \in \mathbb{Z}\}$$

is an orthonormal basis for V_{j+1}*, then* ψ *is a wavelet, referred as mother wavelet.*

In the multiresolution analysis, we know that the collection $\left(\phi_{j,k}\right)_{k\in\mathbb{Z}}$ forms an orthonormal basis of V_j. Hence, for any function $f \in L^2(\mathbb{R})$, the series

$$f_j = P_j f = \sum_{k\in\mathbb{Z}} \alpha_{j,k}\phi_{j,k},$$

where $\alpha_{j,k} = \left\langle f, \phi_{j,k}\right\rangle$ is the projection of f into V_j. It is the best approximation to f in V_j and V_j is, for this reason, referred to as an approximation space. On the other hand, the projection of f on V_{j+1} admits an unique decomposition as

$$f_{j+1} = f_j + f_j^{\perp},$$

where $f_j \in V_j$ and $f_j^{\perp} \in W_j$. Given that $f_j^{\perp}$ admits the decomposition

$$\sum_{k=-\infty}^{\infty} \beta_{j,k}\psi_{j,k} = f_j^{\perp},$$

where $\beta_{j,k} = \langle f, \psi_{j,k} \rangle$, and $\psi_{j,k}$ is an orthonormal basis of $W_j = V_j^{\perp}$. The series $\sum_{k=-\infty}^{\infty} \beta_{j,k} \psi_{j,k}$ provides then the extra details which when added to the best approximation of f in V_j provide the best approximation in V_{j+1}. The space W_j is for this reason referred to as the detail space. The fact that ψ is a wavelet is a direct consequence of $\psi_{j,k} \in V_{j+1}$ and of conditions (A.1).

When the scaling function ϕ is defined by (A.2), the associated wavelets are called Haar wavelets. These wavelets are obtained by translating, scaling, and dilating a function H :

$$H(x) = \begin{cases} 1 & 0 \leq x < \frac{1}{2} \\ -1 & \frac{1}{2} \leq x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

The functions $H_{j,k}$ for $k \in \mathbb{Z}$ are then defined by

$$H_{j,k}(x) = 2^{j/2} H(2^j x - k),$$

for $j, k \in \mathbb{Z}$. The function $H_{j,k}$ is null outside its support $\left[\frac{k}{2^j}, \frac{k+1}{2^j}\right)$. The $H_{j,k}$ forms an orthornormal basis of W_j, and any function $f \in L^2(\mathbb{R})$ admits the projection into V_{j+1}

$$f_{j+1} = \sum_{k \in \mathbb{Z}} c_{j,k} \phi_{j,k} + \sum_{k \in \mathbb{Z}} d_{j,k} H_{j,k}.$$

By recurrence, we also have that

$$f_{j+1} = \sum_{k \in \mathbb{Z}} c_{0,k} \phi_{0,k} + \sum_{j'=0}^{j} \sum_{k \in \mathbb{Z}} d_{j',k} H_{j',k}.$$

If the support of the function f is bounded, for example, [0, 1], then

$$f_{j+1}(x) = c_0 \phi(x) + \sum_{j'=0}^{j} \sum_{k=0}^{2^j - 1} d_{j',k} H_{j',k}(x). \tag{A.4}$$

where $c_0 \phi = c_0 = \int_0^1 f(x) dx$ can be interpreted as the average of $f(x)$ over [0, 1].

The Haar wavelet is nevertheless not the only function compatible with the multiresolution analysis. From the ladder of subspaces, $V_{j-1} \subset V_j$ for all $j \in \mathbb{Z}$, the space V_0 is a subspace of V_1. Since $\left(\phi_{1,k}\right)_{k \in \mathbb{Z}}$ is a basis for V_1 and $\phi \in V_0 \subset V_1$, any eligible scaling function is the solution of the dilation or scaling equation

$$\phi\left(\frac{x}{2}\right) = \sum_{k \in \mathbb{Z}} a_k \phi(x - k), \tag{A.5}$$

where $a_k = \left\langle \phi\left(\frac{x}{2}\right), \phi\,(x - k) \right\rangle$. Wavelets are orthogonal to the subspaces generated by scaled, translated scaling functions. Therefore, the wavelets $\psi(.)$ associated to any scaling function satisfying (A.5) are defined by

$$\psi(x) = \sum_{k \in \mathbb{Z}} (-1)^k a_{1-k} \phi(2x - k). \tag{A.6}$$

https://doi.org/10.1017/asb.2020.18 Published online by Cambridge University Press

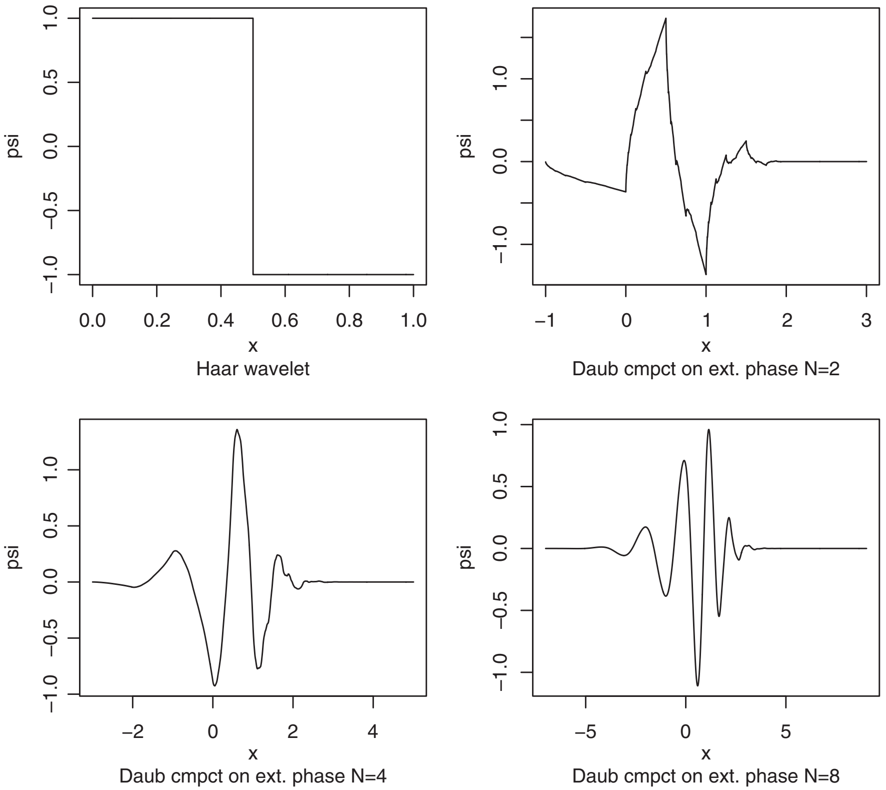

FIGURE A.1: Daubechies wavelets with vanishing moments up to M=1, 2, 4, and 8. For M=1, we retrieve the Haar wavelet.

For a proof, see Theorem 6.9 in Nickolas (2017). Scaling and wavelet functions, solution of Equations (A.5) and (A.6), may be used for constructing by recursion of the projection of f on V_{j+1}:

$$f_{j+1}(x)=P_{j+1}f=\sum_{k\in\mathbb{Z}} c_{j,k}\phi_{j,k}(x)+\sum_{k\in\mathbb{Z}} d_{j,k}\psi_{j,k}(x). \tag{A.7}$$

In the next section, we will see how to infer scaling and wavelets coefficients.

The Daubechies Wavelets ψ_M is a family of functions satisfying Equations (A.5) and (A.6) with vanishing moments up to a certain order M, that is,

$$\int_{-\infty}^{+\infty} x^m \psi_M(x)dx = 0 \text{ for } m=0, 1, ..., M. \tag{A.8}$$

Using wavelets with vanishing moments allows a sparse representation of piecewise polynomial functions of $L^2(\mathbb{R})$. Unfortunately, there does not exist any closed-form expression for the Daubechies wavelets and its scaling function. However, they are easily numerically computable because coefficients a_k in Equation (A.5) are known. The Daubechies scaling function is largely asymmetric. As alternative, Daubechies has proposed the least asymmetric wavelet that has vanishing moments with better symmetry. Coiflets have similar properties to Daubechies wavelets except the scaling function is also chosen so that it has vanishing moments. In other words, the scaling function satisfies (A.8) with ϕ instead of ψ. There exist many others wavelets, for instance, in the complex space, that are out of the scope of this article.

A.2. Discrete wavelet transform

Let us assume that we observe the values $y_i = f(x_i)$ of an $L^2(\mathbb{R})$ function, for an equispaced sequence $\{x_1, \ldots, x_n\}$ of length 2^J. The DWT computes a vector of parameters as in Equation (A.3), consisting of the last, most coarse, scaling coefficient c_0, and the wavelet coefficients $d_{j,k}$ for $j = 0, ..., J-1$ and $k = 0, ..., 2^j - 1$. These coefficients allow us to construct an approximation f_J of f:

$$f_J(x) = c_0\phi\left(\frac{x - x_m}{x_M - x_m}\right) + \sum_{j=0}^{J-1}\sum_{k=0}^{2^j-1} d_{j,k}\psi_{j,k}\left(\frac{x - x_m}{x_M - x_m}\right), \tag{A.9}$$

with $f_J(x_i) = y_i$ for $i = 1, ..., n$.

If we denote $h_k = 2^{-1/2}a_k$, the scaling Equation (A.5) may be rewritten as

$$\phi\,(x) = \sum_{k\in\mathbb{Z}} h_k\phi_{1,k}(x). \tag{A.10}$$

If we define $g_k = (-1)^k 2^{-1/2} a_{1-k}$, the wavelet Equation (A.6) becomes

$$\psi(x) = \sum_{k\in\mathbb{Z}} g_k\phi_{1,k}(x). \tag{A.11}$$

Even if the wavelets do not have any closed-form expression, the value of coefficients $(h_k)_{k\in\mathbb{Z}}$ and $(g_k)_{k\in\mathbb{Z}}$ are computable and are the only information needed for estimating scaling and wavelets coefficients in the decomposition (A.9).

We first show that it is possible to obtain coarser-level wavelet coefficients in Equation (A.7) from finer ones (level $j-1$ from j). As $\left(\phi_{j-1,k}\right)_{k\in\mathbb{Z}}$ is an orthonormal basis for V_{j-1}, we have that

$$c_{j-1,k} = \int_{\mathbb{R}} f(x)\phi_{j-1,k}(x)dx\,. \tag{A.12}$$

On the other hand, from Equation (A.10), we can develop $\phi_{j-1,k}$ as follows:

$$\begin{aligned}\phi_{j-1,k}(x) &= 2^{(j-1)/2}\phi\left(2^{j-1}x - k\right)\\ &= 2^{(j-1)/2}\sum_{n\in\mathbb{Z}} h_n\phi_{1,n}\left(2^{j-1}x - k\right)\\ &= 2^{j/2}\sum_{n\in\mathbb{Z}} h_n\phi\left(2^j x - 2k - n\right)\\ &= \sum_{n\in\mathbb{Z}} h_n\phi_{j,n+2k}\,(x).\end{aligned} \tag{A.13}$$

https://doi.org/10.1017/asb.2020.18 Published online by Cambridge University Press

If we substitute (A.13) into (A.12), then we infer that

$$\begin{aligned} c_{j-1,k} &= \sum_{n\in\mathbb{Z}} h_n \int_{\mathbb{R}} f(x)\phi_{j,n+2k}(x)\,dx \\ &= \sum_{n\in\mathbb{Z}} h_n c_{j,n+2k}. \end{aligned}$$

Or after rearrangement,

$$c_{j-1,k} = \sum_{n\in\mathbb{Z}} h_{n-2k} c_{j,n}. \tag{A.14}$$

In a similar manner, we can prove that

$$d_{j-1,k} = \sum_{n\in\mathbb{Z}} g_{n-2k} c_{j,n}. \tag{A.15}$$

Computing with accuracy, the initial fine father coefficient (layer J) is a hard task. However in practice, the wavelet transform is initialized using the original function samples, that is,

$$c_{J,k} = y_{k+1} \quad k = 0, 1, \ldots, 2^J. \tag{A.16}$$

This approach is, for example, implemented in the R package wavetresh and is satisfactory for our analysis. We also use a technical assumption of periodicity for computing coefficients at the beginning and end of the series. Notice, however, that this method of initialization is sometimes called the wavelet crime (see Strang and Nguyen, 1996). Equations (A.14), (A.15), and (A.16) allow us to calculate recursively the vector of coefficients

$$\boldsymbol{d} = \left(c_0, \left(d_{j,k}\right)_{j\in\{0,\ldots,J-1\},k\in\{0,\ldots,2^j-1\}}\right).$$

Notice also that these coefficients are obtained by linear transformation of y_k for $i \in \{1, ..., 2^J\}$. If the vector of $(y_i)_{i\in\{1,\ldots,2^k\}}$ is denoted by $\boldsymbol{y}$, then we can show that

$$\boldsymbol{d} = T\boldsymbol{y},$$

where T is an orthogonal matrix, that is, $TT^\top = I$. Another consequence is that

$$\|\boldsymbol{d}\|^2 = \|\boldsymbol{y}\|^2 .$$

Given that the vector $\boldsymbol{d}$ is sparse, the information carried by $\boldsymbol{y}$, and measured $\|\boldsymbol{y}\|^2$, referred to as the "power," is redistributed among a smaller number of coefficients, significantly different from zero.

Finally, Mallat (1989a,b) has shown that we can retrieve scaling coefficients of level j from those of level $j-1$ by inverting Equations (A.14) and (A.15):

$$c_{j,n} = \sum_{k\in\mathbb{Z}} h_{n-2k} c_{j-1,k} + \sum_{k\in\mathbb{Z}} g_{n-2k} d_{j-1,k}. \tag{A.17}$$

This relation is often referred to as the Mallat's pyramid. In this framework, the function $f(x_k)$ is approached by $f_J(x) = \sum_k c_{J,k}\mathbf{1}_{\{x=x_k\}}$.

https://doi.org/10.1017/asb.2020.18 Published online by Cambridge University Press