

C smoothing: a web-tool for controlled smoothing by segments of mortality data

Eliud Silva, Víctor M. Guerrero, Yhael Jacinto Cruz & Emanuel Rodriguez Soler

To cite this article: Eliud Silva, Víctor M. Guerrero, Yhael Jacinto Cruz & Emanuel Rodriguez Soler (2022): C smoothing: a web-tool for controlled smoothing by segments of mortality data, Communications in Statistics - Simulation and Computation, DOI: [10.1080/03610918.2022.2154794](https://doi.org/10.1080/03610918.2022.2154794)

To link to this article: <https://doi.org/10.1080/03610918.2022.2154794>



Published online: 09 Dec 2022.



Submit your article to this journal [↗](#)







View related articles [↗](#)



View Crossmark data [↗](#)



CSmoothing: a web-tool for controlled smoothing by segments of mortality data

Eliud Silva^a , Víctor M. Guerrero^b , Yhael Jacinto Cruz^a , and Emanuel Rodriguez Soler^a 

^aFaculty of Actuarial Sciences, Universidad Anáhuac México, Naucalpan, Mexico; ^bDepartment of Statistics, Instituto Tecnológico Autónomo de México (ITAM), Mexico City, México

ABSTRACT

Csmoothing allows an analyst to use the so-called Controlled Smoothing technique to estimate trends in a time series framework. In this Web-tool (Shiny), the analyst may apply the methodology to at most 3 mortality time series simultaneously, as well as to other kind of time series individually. Likewise, this smoothing approach allows the analyst to establish one, two or three segments in order to take into account possible changes in variance regimes. For estimating trends it uses different amounts of smoothness, both globally for the total data set and through some partial indices for each selected segment. It is also possible to endogenously fix the points where the segments start and end (the cutoff points) with continuous joints. Additionally, intervals of different standard deviations for their respective trends are given. Particular emphasis is placed on a big data set of log mortality rates, $\log(qx)$, taken from period life tables of the Human Mortality Database (HMD) (University of California Berkeley (USA) and Max Planck Institute for Demographic Research (Germany)), (2021). In all cases, dynamic graphs and several statistics related to the Controlled Smoothing technique are illustrated.

ARTICLE HISTORY

Received 7 December 2021
Accepted 3 September 2022

KEYWORDS

Life expectancy;
Mortality; Smoothness

1. Introduction

In a time series framework, one basic objective is to identify underlying trends in a data set being analyzed. In fact, this objective can be thought as a first step in descriptive terms when applying smoothing techniques. Sometimes it is important to have a clear perspective of the trend without the effect of extreme observations. This is a key point to facilitate decision-making in several contexts, for example in epidemiology as well as in demography. For this purpose, the proposal of Guerrero and Silva (2015) provides a non-parametric alternative. Likewise, it is well-known that data smoothing can be done using parametric or non-parametric methods (Bowman and Evers 2013). The last strategy is more flexible, in the sense that it does not require verifying distributional assumptions and it is more robust. In accordance with Hyndman (2008), among the non-parametric techniques are: Density estimation, Kernel regression, Additive models, Functional data analysis and Splines.

In particular, we will briefly mention Splines and some of its extensions available in the software environment R (R Core Team 2021) since it is a tool frequently applied in the time series context. In addition, the aforementioned statistical method Guerrero and Silva (2015), and its implementation (CSmoothing) is presented through a Web-tool, also so-called Shiny (Chang et al. 2017). With this approach, the analyst can control and measure the induced smoothness on

a data set through indices, both globally and by segments, as well as generate estimation intervals. It should be noted that CSmoothing is more than a graph maker, because it estimates trends through a statistical method that provides measures of statistical performance.

CSmoothing uses one, two or three smoothness indices corresponding to the number of segments chosen, as well as a global smoothness index. The beginning and ending of every segment can be established exogenously in accordance with the potential existence of different variability regimes or any other theoretical information as well. CSmoothing is based on the Hodrick-Prescott (HP) filter (Hodrick and Prescott 1997), and, it could be decoded as an extension of the original proposal of Guerrero (2008). In general, the indices depend on a smoothing parameter λ and the time series size N .

Controlled Smoothing, as a non-parametric technique, has several advantages. For instance, it allows getting estimation intervals of trends, with plus or minus a certain number of standard deviations (sd) chosen by the analyst (± 1 , ± 2 , ± 3). There are also no distributional assumptions, so that their verification is not required when producing estimates. For the estimated trend that covers more than one segment, the method provides continuity, even when there is a markedly different variance throughout the data range, that is, the joints of the estimated trend are continuous. Thereby, this tool is ideal for making valid comparisons between trend estimates having the same smoothing indices.

CSmoothing is a user-friendly Web-tool based on the RStudio language (RStudio Team 2021) for smoothing by segments, without needing to use any additional code in that environment. Thus, our approach is such that the analyst can use it via any device with an internet connection, independently of the operating system. It is worth mentioning that the potential benefits of this kind of tools have been explored previously by employing them in both demographic research and higher demographic education (Devedzic and Devedzić 2003, 2010). Additionally, we believe that this Web-tool is in line with the huge need to develop new tools that facilitate the implementation and understanding of different analysis in the digital era. This is particularly due to the fact that data sets in demography are currently showing an outstanding growth, such as it has been pointed out by Alburez-Gutierrez et al. (2019).

Nowadays, there are several web-tools focused on mortality. One of them, published online by the New Zealand Ministry of Health (2021), is a Mortality web-tool (<https://minhealthnz.shinyapps.io/mortality-web-tool/>). In this tool, domestic mortality and demographic data from 1948 through 2019 is available by selected causes. Other web-tool consists in quantifying potential gains in life expectancy for selected neighborhoods in Baltimore, given public health interventions (Chandran et al. 2021). There are also other web-tools to provide estimates in order to analyze specific causes of death such as Cancer (Rosenberg, Check, and Anderson 2014) and COVID-19 (Das, Mishra, and Gopalan 2020 and Noll et al. 2020).

We believe that CSmoothing has advantages over the aforementioned web-tools, in terms of a global mortality estimates over time. In particular, with data taken from the HMD, CSmoothing presents significant diversity of countries and time periods, allowing the possibility of comparing them and offering also the possibility of using it with own data. It also makes it possible to apply a statistical method to measure and control the smoothing by segments to estimate mortality rates, life expectancies, confidence intervals and so on. Additionally, two alternatives for analyzing data sets, as exposed below, could be considered an asset.

There are statistics that allow to quantify the standard error generated as a companion of an estimate in CSmoothing. Those statistics provide a guide for the analyst when facing the need to segment or not the series of interest, in order to obtain a better estimate. When mortality rates are employed, the life expectancy at birth and its interval is also provided for every data set. Given the estimation intervals associated to the estimated trend, a coverage percentage is suggested to decide the percentage of smoothness that goes hand in hand with the selected smoothing parameter(s).

This document is organized as follows. In the following section, general concepts of Splines and their extensions available in R, considered as a smoothing method, are explained. Then, the Controlled Smoothing method for one and two segments and without loss of generality for 3 or more segments, is expositied. Next, CSmoothing is exemplified with data sets taken from the Human Mortality Database (HMD) (University of California Berkeley (USA) and Max Planck Institute for Demographic Research (Germany)), (2021). Finally, main conclusions are presented.

2. Splines and their extensions in R

According to Wold (1974) Spline functions can be seen as piecewise polynomials with continuity conditions. As a particular case, the Cubic Spline is a set of third degree polynomials generated from a set of points. Hence, both of them can be used with several advantages instead of other functions to smooth data sets. A Splines extension for smoothing can be represented by B-Splines (De Boor 1978). Let us consider a B-Spline function of degree q and let $x = (x_0, x_1, \dots, x_n)$ be a vector of equally spaced knots, so that

$$B_{i,1}(x) = \begin{cases} 1 & \text{for } x \in [x_i, x_{i+1}) \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

and

$$B_{i,q+1}(x) = \frac{x - x_i}{x_{i+q} - x_i} B_{i,q}(x) + \frac{x_{i+q+1} - x}{x_{i+q+1} - x_{i+1}} B_{i+1,q}(x) \quad (2)$$

where $B_{i,q+1}(x) > 0$ for $x_i \leq x \leq x_{i+k+1}$ and $B_{i,q+1}(x) = 0$ for $x_0 \leq x \leq x_i$ and $x_{i+k+1} \leq x \leq x_{n+k+1}$. Now let us consider a set of B-Splines that form a basis and act as predictors in a regression model. Given m pairs of observations (x_i, y_i) to estimate the regression of y on x by Least Squares the P-Spline method reduces the number of Splines by placing a smoothness penalty on the differences of coefficients for adjacent B-Splines (Eilers and Marx 1996). This approach seeks to minimize the function

$$F = \sum_{i=1}^m \left\{ y_i - \sum_{j=1}^n \alpha_j B_{j,q}(x_i) \right\}^2 + \lambda \sum_{j=k+1}^n \left(\Delta^2 \alpha_j \right)^2 \quad (3)$$

where p is the number of B-Splines, the α_j 's are constant coefficients and the $B_{j,q}(x)$ form a basis of degree q . Likewise, the parameter λ trades off fit against smoothness induced by the second order difference $\Delta^2 \alpha_j = \alpha_j - 2\alpha_{j-1} + \alpha_{j-2}$.

Nowadays, there are different packages in R that can be used for non-parametric smoothing (Perperoglou et al. 2019), where reference is made to functions for modeling with Splines in the regression framework. Some of their features are given to provide an overview of the most widely used Spline-based techniques and their implementation in R. Additionally, the MortalitySmooth library (Camarda 2012) is notable for smoothing through P-Splines with mortality data, which is the type of information with which our Shiny is exemplified. Unlike previous libraries, CSmoothing allows the analyst to smooth a time series in a controlled way, both for one, two or more segments, generate estimation intervals, impose the beginning and end of segments and have a standard error measurement when considering different possibilities of smoothing.

A distinctive aspect of the above methods with respect to Controlled Smoothing, is the way of choosing the smoothing parameter λ . In general, for the aforementioned methods, the optimal value of λ is found through automatic criteria such as: Akaike information criterion, cross-validation, generalized cross-validation or Bayesian information criterion. In contrast, with CSmoothing there is no optimality criterion for the smoothing parameter, but it is the analyst

judgment that leads to the selection of λ that makes sense for descriptive and/or comparability purposes.

In the mortality framework, we advise to employ the approach of fixing a global smoothness index of 100S% = 75%, such as it is expositied in [sec. 4](#). We notice that, with this Shiny the estimated trends obtained with a pure statistical criterion are equivalent to fit mortality curves in demography. Besides, the smoothness achieved when the λ value is chosen through automatic criteria can also be measured (Cortés-Toto, Guerrero, and Reyes 2017). In such cases it is evidenced that the smoothness induced in the estimates are beyond the analyst's decision. On other hand, it is clear that when using automatic criteria with mortality data, the resulting λ values generate trends that typically lack of demographic sense and comparability (Silva and Ovin 2018).

3. Controlled smoothing methods

3.1. One segment

Let us assume that the observed data can be expressed as a signal plus noise model (Guerrero 2008) that is, $y_i = \tau_i + \eta_i$, where $\{\tau_i\}$ is the trend (or signal) and $\{\eta_i\}$ is the noise of $\{y_i\}$, for $i = 1, 2, \dots, N$. The target is to estimate the underlying trend by solving the following problem

$$\text{Min}_{\{\tau_i\}} \left\{ \sum_{i=1}^N \frac{1}{\sigma_1^2} (y_i - \tau_i)^2 + \sum_{i=3}^N \frac{1}{\sigma_0^2} (\nabla^2 \tau_i)^2 \right\} \quad (4)$$

where σ_1^2 is the variance of the deviation in relation to trend $\{y_i - \tau_i\}$ and σ_0^2 is the variance of the differenced trend $\{\nabla^2 \tau_i\}$, with $\nabla^2 \tau_i = \tau_i - 2\tau_{i-1} + \tau_{i-2}$. The ratio of variances $\lambda = \sigma_1^2 / \sigma_0^2$ is the smoothing parameter which is used to balance the smoothness of the trend against its fidelity to the original data. It should be noticed that when $\lambda \rightarrow \infty$, the trend approaches a first degree polynomial, and as $\lambda \rightarrow 0$, it gets closer to the original data.

When solving the minimization problem (4) we obtain the HP filter for the entire range of observations ($i = 1, 2, \dots, N$). Using a statistical model or performing formal statistical inference is not required in this context; thus, instead of estimating the λ parameter through a statistical procedure, its value is calibrated, and this is the most important practical decision that needs to be made when using this method. The approach proposed by Guerrero (2008), through Generalized Least Squares (GLS), is applied to solve (4). Then, it can be shown that the Best Linear Unbiased Estimator (BLUE) for the trend and its variance are given by and

$$\hat{\tau} = (I_N + \lambda K'K)^{-1} y \quad (5)$$

$$\Gamma = \text{Var}(\hat{\tau}) = \left(\sigma_\eta^{-2} I_N + \sigma_\varepsilon^{-2} K'K \right)^{-1} \quad (6)$$

where I_N is the N -dimensional identity matrix and K is an $(N-2) \times N$ matrix that represents the matrix difference operator ∇^2 , whose i, j -th entry is the binomial coefficient

$$K(i, j) = \frac{(-1)^{2+i-j} 2!}{(j-i)!(2-j+i)!}$$

for $i = 1, 2, \dots, N-2$ and $j = 1, 2, \dots, N$, with $K(i, j) = 0$ if $j < i$ or $j > 2 + i$. Further, it is proposed to use the following index that lets the analyst to control the smoothness level

$$S(\lambda; N) = 1 - \text{tr}[(I_N + \lambda K'K)^{-1}] / N \quad (7)$$

where tr is the trace function. It is worth mentioning that this bounded index depends only on the values λ and N , since the matrix K is a function of these two values. Likewise, it is clear that $S(\lambda; N) \rightarrow 0$ as $\lambda \rightarrow 0$ and $S(\lambda; N) \rightarrow 1$ as $\lambda \rightarrow \infty$. Thereby, this index can be interpreted as the

proportion of smoothness achieved with this technique. It is important to keep in mind that the amount of smoothness that can be attained globally is bounded by $1 - 2/N$ as $\lambda \rightarrow \infty$. This fact is because the trace that appears in S can be written in terms of the $N - 2$ non-zero eigenvalues of $K'K$, that is, $e_1, \dots, e_{N-2} > 0$, as $tr(I_N + \lambda K'K)^{-1} = (1 + \lambda e_1)^{-1} + \dots + (1 + \lambda e_{N-2})^{-1} + 2$ so that the trace tends to 2 as $\lambda \rightarrow \infty$.

3.2. Two or more segments

Sometimes the trend changes due to changes in the variance of the series, a fact reflected in the smoothness of the trend. Thus, an extension of the minimization problem (4) allows for different trend behaviors for segments of the data range, which are in turn linked to different variances, one for each regime. In this case, we pose a minimization problem that accounts for different trends in adjacent segments and hence different λ values must be calibrated. For simplicity, we present the method for the case of just two segments, which can easily be generalized to more than two segments (see for details Guerrero and Silva 2015). The two-segment problem is formulated as follows

$$\text{Min}_{\{\tau_i\}} \left\{ \sum_{i=1}^{N_1} \frac{1}{\sigma_1^2} (y_i - \tau_i)^2 + \sum_{i=N_1+1}^N \frac{1}{\sigma_2^2} (y_i - \tau_i)^2 + \sum_{i=3}^N \frac{1}{\sigma_0^2} (\nabla^2 \tau_i)^2 \right\} \quad (8)$$

where σ_1^2 and σ_2^2 are the variances of the first and second data segments, with N_1 and $N_2 = N - N_1$ observations, respectively; $\{y_i\}$ are the observed mortality rates and $\{\tau_i\}$ is the trend. The unobservable component model that underlies the minimization problem posed by (8) can be written as follows,

$$y_i = \tau_i + \eta_{1,i} \text{ with } \eta_{1,i} \sim (0, \sigma_1^2) \text{ for } i = 1, 2, \dots, N_1 \quad (9)$$

$$y_i = \tau_i + \eta_{2,i} \text{ with } \eta_{2,i} \sim (0, \sigma_2^2) \text{ for } i = N_1 + 1, \dots, N \quad (10)$$

$$\nabla^2 \tau_i = \varepsilon_i \text{ with } \varepsilon_i \sim (0, \sigma_0^2) \text{ for } i = 3, \dots, N \quad (11)$$

where we write $\eta \sim (0, \sigma_\eta^2)$ to say that the random variable η has mean 0 and variance σ_η^2 . The sequences of random errors $\{\eta_{j,i}\}$ for $j = 1, 2$, are serially uncorrelated, and $\{\varepsilon_i\}$ is another sequence of serially uncorrelated random errors that is also mutually uncorrelated with the previous two sequences. The matrix representation of model (9)-(11) becomes

$$\mathbf{y} = \boldsymbol{\tau} + \begin{pmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \end{pmatrix} \quad (12)$$

and

$$K\boldsymbol{\tau} = \boldsymbol{\varepsilon} \quad (13)$$

Then, an application of GLS produces the BLUE of the vector of trends, that can be expressed as

$$\begin{pmatrix} \hat{\boldsymbol{\tau}}_1 \\ \hat{\boldsymbol{\tau}}_2 \end{pmatrix} = \left[\begin{pmatrix} I_{N_1} & 0 \\ 0 & I_{N_2} \end{pmatrix} + \begin{pmatrix} \lambda_1 I_{N_1} & 0 \\ 0 & \lambda_2 I_{N_2} \end{pmatrix} K'K \right]^{-1} \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} \quad (14)$$

where I_{N_j} is the N_j -dimensional identity matrix and $\lambda_j = \sigma_j^2 / \sigma_0^2$, for $j = 1, 2$. The matrices K_i are of size $(N_j - 2) \times N_j$ for $j = 1, 2$ and they have the same shape as matrix K . In fact, we have that

$$K = \begin{pmatrix} K_1 & 0 \\ k_1 & k_2 \\ 0 & K_2 \end{pmatrix} \quad (15)$$

with $k_1 = \begin{pmatrix} 1 & -2 \\ 0_{2 \times (N_1-2)} & 0 \end{pmatrix}$ and $k_2 = \begin{pmatrix} 1 & 0 \\ -2 & 1 \end{pmatrix} 0_{2 \times (N_2-2)}$, with k_1 being of size $2 \times N_1$ and k_2 being of size $2 \times N_2$.

The BLUE given by (14) is also be called predictor, since we are in fact estimating the realization of a random vector rather than a vector of constants (a justification for the use of GLS in this context was presented by Guerrero 2007). In addition, the variance-covariance matrix of the GLS estimator is given by

$$\Gamma = \text{Var}(\hat{\tau}) = \left[\begin{pmatrix} \sigma_1^{-2} I_{N_1} & 0 \\ 0 & \sigma_2^{-2} I_{N_2} \end{pmatrix} + \sigma_0^{-2} \begin{pmatrix} K_1' K_1 + k_1' k_1 & k_1' k_2 \\ k_2' k_1 & K_2' K_2 + k_2' k_2 \end{pmatrix} \right]^{-1} \quad (16)$$

and unbiased estimators of the error variances are given by $\hat{\sigma}_0^2 = \frac{RSS}{N-2}$ and $\hat{\sigma}_j^2 = \lambda_j \hat{\sigma}_0^2$ for $j = 1, 2$, with *RSS* the Residual Sum of Squares (see Guerrero and Silva 2015 for more details).

To be able to apply (14) - (16) values for λ_1 and λ_2 must be provided, as well as the cutoff point N_1 . The smoothing parameters will be chosen by applying the controlled smoothing approach proposed by Guerrero (2007), which is based on measuring the relative precision attributable to the smoothness specification, that is, the second equation of model (13). To this end we should notice that the total precision accomplished by estimating the trend is the inverse of the variance-covariance matrix given by (16), that is, it is provided by Γ^{-1} . Therefore, the amount of smoothness to be attained globally by the trend for the unsegmented time series is $S = 1 - \text{tr}(I_N + \lambda K \hat{K})^{-1} / N$. This index is a reparameterization of the well-known statistic called degrees of freedom (*df*), which is reported by CSmoothing.

We now measure the precision share for each segment by means of the following indices

$$S_j(\lambda_1, \lambda_2; N) = \frac{\text{tr} \left[B_{j,\lambda} (I_N + B_1 + B_2)^{-1} \right]}{N} \quad (17)$$

for $j = 1, 2$ that quantify the smoothness attained by smoothing segment j of data, where

$$B_{1,\lambda} = B_1 B_0^{-1} = \frac{N_1}{N} \begin{bmatrix} \lambda_1 \left(\frac{N}{N_1} K_1' K_1 + k_1' k_1 \right) & \lambda_2 k_1' k_2 \\ \lambda_1 k_2' k_1 & \lambda_2 k_2' k_2 \end{bmatrix} \quad (18)$$

$$B_{2,\lambda} = B_2 B_0^{-1} = \frac{N_2}{N} \begin{bmatrix} \lambda_1 k_1' k_1 & \lambda_2 k_1' k_2 \\ \lambda_1 k_2' k_1 & \lambda_2 \left(\frac{N}{N_2} K_2' K_2 + k_2' k_2 \right) \end{bmatrix}. \quad (19)$$

Thus, when assigning the values of the indices $S_j(\lambda_1, \lambda_2; N)$ for $j = 1, 2$, or likewise, the percentages of smoothness (obtained by multiplying the indices by 100), the smoothing parameters are obtained by solving expression (14) numerically for λ_1 and λ_2 .

The cutoff points could be chosen using the search procedure suggested by Guerrero and Silva (2015). To do that, we first decide the amount of smoothness to be attained globally by the trend. Then, we choose the percentage of smoothness for the first segment, keeping in mind that the higher the data variability, the more smoothness will be required for the trend. Finally, the smoothness for the second segment gets fixed by means of $S_2 = (NS - N_1 S_1) / (N - N_1)$ and the optimal cutoff point is the value of N_1 that minimizes the estimated error variance σ_0^2 . In practice, there is no guarantee that such cutoff point exists, and when this happens, the segmentation can be deemed to be appropriate in a statistical sense. However, the conditions of the application may lead to an exogenous selection of the cutoff point, as is the case in the application shown here, where two cutoff points are chosen based on subject matter knowledge. Another

possibility is to take into account the proposal suggested by Killick and Eckley (2014) for seeking variance changes in time series.

4. Obtaining numerical results

CSSmoothing has two alternatives for analyzing data sets. The first one is to employ those taken from the HMD, consisting of 4,762 period life tables, and the second one is for other kind of time series imported by the analyst. When data sets from HMD are used, there are also three alternatives: Comparing by Country (not more than 3), Comparing by sex (Female, Male, Total) and Comparing by Years (according to the available dates for every country). The available mortality data sets from HMD can be listed by clicking on the upper right corner button of CSSmoothing called "See Data". As a matter of fact, they are illustrated in the Appendix's Table. Meanwhile, when data sets are provided by the analyst the format should be a CSV file with only two columns. The first with increasing and equally spaced ages (0, 1, 2, ...), and the second one for specific mortality rates (qx or log(qx)).

In all cases, it is possible to download data and estimates into a CSV or Excel file. That file contains both the original and the estimated time series, the upper and lower bounds as well as the smoothing parameters, in accordance with the options imposed by the analyst. A sample sheet of a downloaded Excel data is displayed in Figure 1. CSSmoothing can be used with different kinds of time series data. Here the analyst can upload her/his own data and explore controlled smoothing with or without segmentation. There are no limitations regarding the size of the time series, nor its periodicity. However, for the time series to be analyzed no missing values are allowed and just one time series can be analyzed at once.

When CSSmoothing is employed, it should be remembered that looking for a minimum standard error is a key statistical idea, but also the resulting trend estimate should make sense in the corresponding field of study. As a basic rule, greater variance in the data implies that a larger smoothness index is required. Finally, it could be necessary to try several choices of λ values in order to decide which one provides the best estimate. If the uploaded data corresponds to $\log(qx)$ with its ages (x), then the life expectancy is estimated with its interval (see Figure 2) and the results could be compared with those coming from the HMD.

To facilitate the use of CSSmoothing for studying log mortality rates, $\log(qx)$, we have coded initial values for the smoothing parameters and cutoff points to achieve the suggested 75% global

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	x	data	estimates	Upper/Bou	Lower/Bou	returns	parameters																
2	1	1	-3.486826	-4.141585	-3.844114	-4.439057	λ1	5.8															
3	2	2	-5.369497	-5.030338	-4.818093	-5.242582	Global Sm	0.750378761516248															
4	3	3	-6.304805	-5.805201	-5.614464	-5.907937	df	21.467405596026															
5	4	4	-6.571304	-6.414760	-6.224549	-6.665371	Standard	0.0794678711421389															
6	5	5	-6.901979	-6.853085	-6.662885	-7.043285	Life Expect	70.04															
7	6	6	-7.179276	-7.145251	-6.956376	-7.334126	Life expect	67.4															
8	7	7	-7.275902	-7.327651	-7.136001	-7.513526		72.37															
9	8	8	-7.366255	-7.420996	-7.239781	-7.614206	Coverage	177.9009767441861															
10	9	9	-7.394449	-7.485052	-7.298028	-7.672095	Elaborated																
11	10	10	-7.43286	-7.517615	-7.330616	-7.704613																	
12	11	11	-7.514331	-7.528679	-7.341680	-7.715673																	
13	12	12	-7.598117	-7.587659	-7.320663	-7.694655																	
14	13	13	-7.458815	-7.441500	-7.254506	-7.628494																	
15	14	14	-7.394741	-7.332733	-7.145741	-7.519726																	
16	15	15	-7.154243	-7.188878	-6.999887	-7.373870																	
17	16	16	-7.030425	-7.020146	-6.833155	-7.207137																	
18	17	17	-6.910276	-6.843135	-6.656128	-7.030111																	
19	18	18	-6.604652	-6.668154	-6.481163	-6.855145																	
20	19	19	-6.401901	-6.519149	-6.332158	-6.706141																	
21	20	20	-6.378727	-6.409057	-6.222066	-6.596049																	
22	21	21	-6.28415	-6.338615	-6.145629	-6.517604																	
23	22	22	-6.447656	-6.271323	-6.084331	-6.458314																	
24	23	23	-6.140742	-6.210678	-6.023687	-6.397670																	
25	24	24	-6.148767	-6.158575	-5.975183	-6.345566																	
26	25	25	-6.166854	-6.112847	-5.929856	-6.299838																	
27	26	26	-5.992743	-6.009141	-5.882409	-6.244632																	

Figure 1. Data exported after using CSSmoothing.

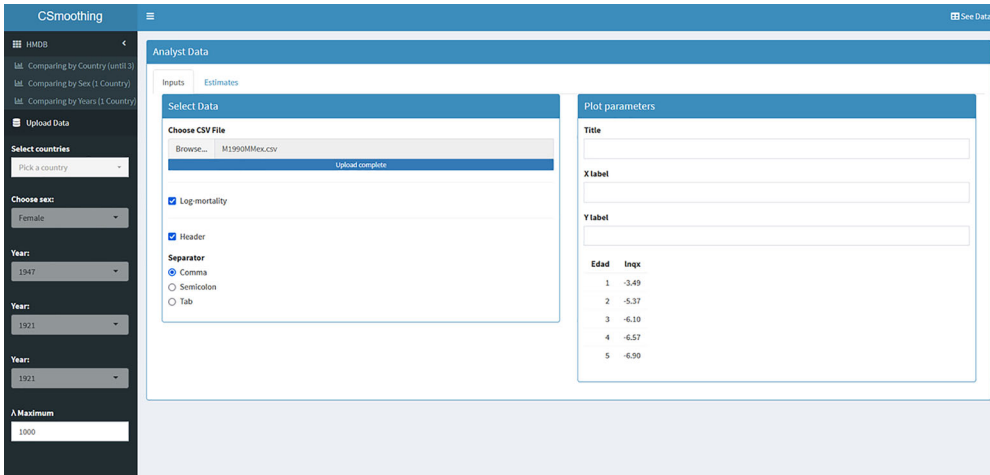


Figure 2. Tab to upload data.

smoothness index. They are as follows: a) for one segment $\lambda_1 = 5.6$ (no cutoff point is required); b) for two segments $\lambda_1 = 29.5$, $\lambda_2 = 3.9$ and cutoff point $x = 25$; c) for three segments $\lambda_1 = 25.4$, $\lambda_2 = 5$, $\lambda_3 = 3$, cutoff points $x = 20$ and $x = 65$. It should be noted that these are just a reference values recommended by the authors, but they can be changed according to the analyst criterion.

There are two ways to execute CSmoothing. The first one is to access it directly at <https://ana-huac.shinyapps.io/CSmoothing/>. The second alternative could be to upload the software to any personal computer with RStudio installed and upload the following files into the same directory where CSmoothing is going to be run: the selected HMD sets (mortalitydata_1x1_2021-10-19.RDS), as well as the code (Silvaetal.R). Both of them are available online at <http://shorturl.at/loHM6> Then open the R file and just click the “Run App” button in RStudio.

In accordance with information provided by shinyapps.io, where CSmoothing was published, a default configuration for a single web-tool allows 150 simultaneous connections. For CSmoothing we have done tests with up to 50 concurrent users (university students), and it does not present any performance problems. It is worth mentioning that to overcome the potential risk of online ineffectiveness, we have developed the second alternative for employing it.

4.1. Comparing by country

In this example three scandinavian countries were selected (Finland, Norway and Sweden) to compare their total mortality rates at the end of the First World War in 1918. It should be remember *e d* that Finland declared its independence in 1917. Although they already had a long tradition to produce good vital statistics, smoothing still seems necessary (see Figure 3). In order to produce the trends, let us consider standard deviation (sd) equal to 0, that implies just to estimate the trends without any sd intervals, one segment and $\lambda_1 \approx 0$. The oldest and most recent common years for all of them are 1878 and 2020 respectively. Note that for a fast knowledge of this, it is enough to click again the "See Data" button, where the period of time available for every country can also be identified, something equivalent to Barbieri et al. (2015).

Now a global smoothness index $100S\% = 75\%$ is imposed as it is suggested by Guerrero and Silva (2015) so that the three trends are comparable (see Figure 4). Thereby, $\lambda_1 = 5.6$ and the estimated standard error of the trend σ^2 were as follows: 0.0302 (Finland); 0.0366 (Norway) and 0.0372 (Sweden). It is possible to say that the mortality trends for Norway and Sweden were very similar. In fact, the most substantive differences appear around the child mortality. It is clear that Finland had higher mortality for all ages except around the hump. This conclusion could have been obtained

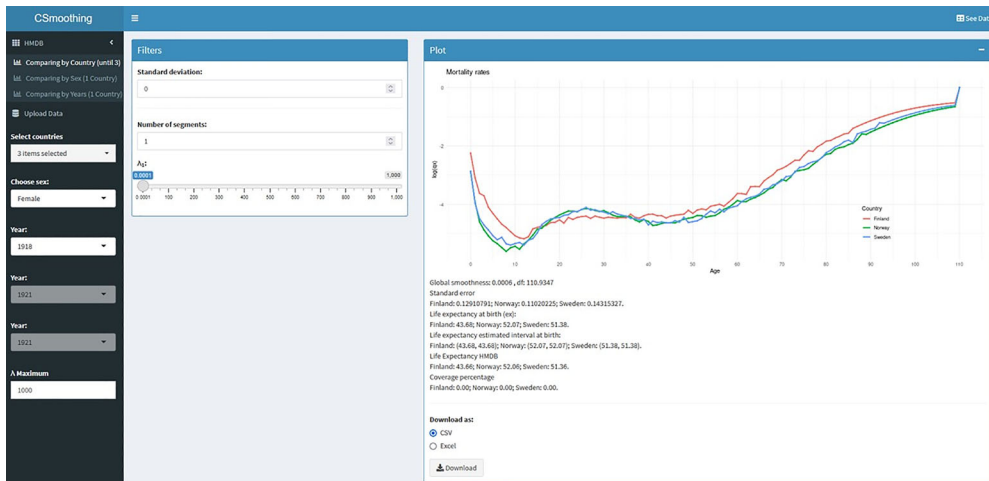


Figure 3. Comparing three scandinavian countries.

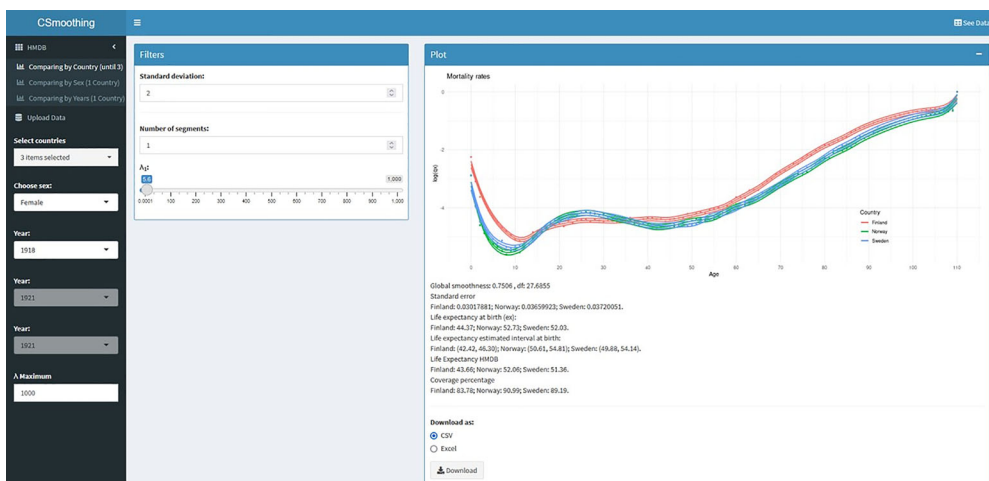


Figure 4. Comparing three scandinavian countries with 100% = 75% and one segment.

without smoothing the data, however to summarize the mortality level it is preferable to use an indicator such as the life expectancy with smoothed and comparable mortality rates.

The estimated life expectancies depend on the smoothness attained, the number of segments and their intervals on the sd chosen. In summary, the life expectancies for one segment (estimated intervals with ± 2 sd) were the following: Finland 44.37 (42.42, 46.30), Norway 52.73 (50.61, 54.81) and Sweden 52.03 (49.88, 54.14). The estimates taken directly from the HMD are: Finland 43.66, Norway 52.06 and Sweden 51.36, all these figures are contained in our estimated intervals. It could be shown that employing an automatic criterion, for instance cross-validation, through the smooth.Pspline (Ramsey and Ripley 2017), the λ_1 value is different for each country and as a consequence the resulting trends are not comparable. Further, the segmentation could be justified in statistical terms because the estimated error variance σ^2 is reduced and also because the estimate makes demographic sense. Based on exogenous information, the first and second selected cutoff points are $x = 20$ and $x = 65$ (see Figure 5).

It should be noticed that with three segments, the estimated error is reduced for all trends. Now they are 0.0276 (Finland); 0.0328 (Norway) and 0.0327 (Sweden), while the global smoothness index

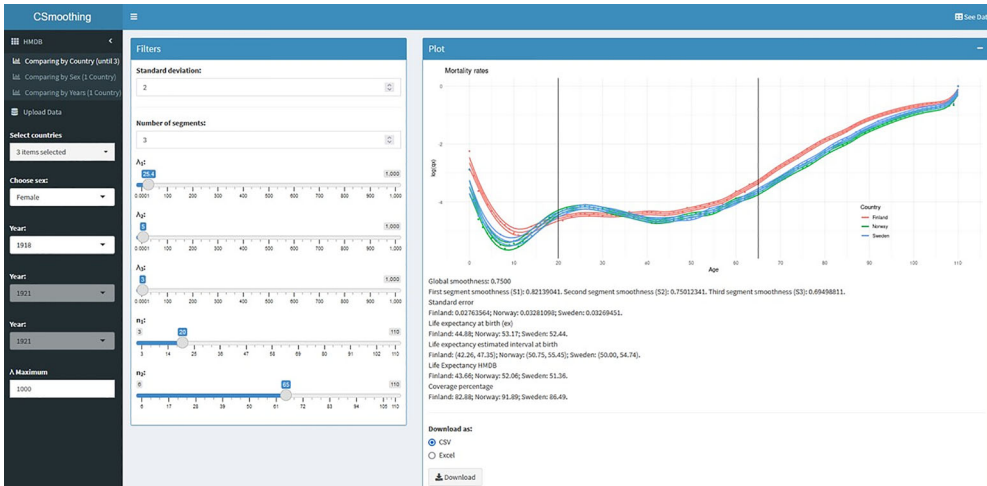


Figure 5. Comparing three Scandinavian countries with 100% = 75% and three segments.

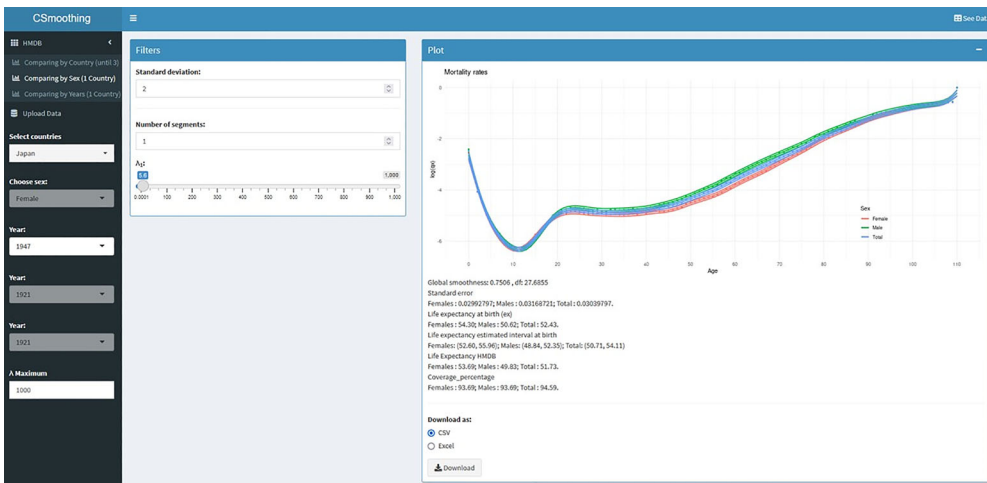


Figure 6. Comparing Japan mortality curves by sex, 1947 with 100% = 75% and one segment.

remains as 100% = 75%. The smoothing parameters and their induced smoothness are $\lambda_1 = 25.4$ ($S_1 = 0.8214$), $\lambda_2 = 5$ ($S_2 = 0.75$) and $\lambda_3 = 3$ ($S_3 = 0.6950$). Likewise, the corresponding life expectancies are Finland 44.88 (42.26, 47.35), Norway 53.17 (50.75, 55.45) and Sweden 52.44 (50.00, 54.74). The estimates coming from the HMD belong to the estimated intervals.

4.2. Comparing by sex

In this exercise we compare the Japan mortality level by sex after the Second World War in 1947. First, it can be appreciated that the mortality rates by sex seem similar until about age 20 as well as beyond 90 years of age and over (see Figure 6). Again, the global smoothness index is fixed in 100% = 75%. This way we get $\lambda_1 = 5.6$ and the standard errors are 0.0299 (Females), 0.0317 (Males) and 0.0304 (Total) respectively. It is possible to estimate the life expectancy with just one segment, nevertheless smoothing could be enhanced if segmentation is applied. Sometimes using two segments is enough to reduce the standard errors as it happens in this example.

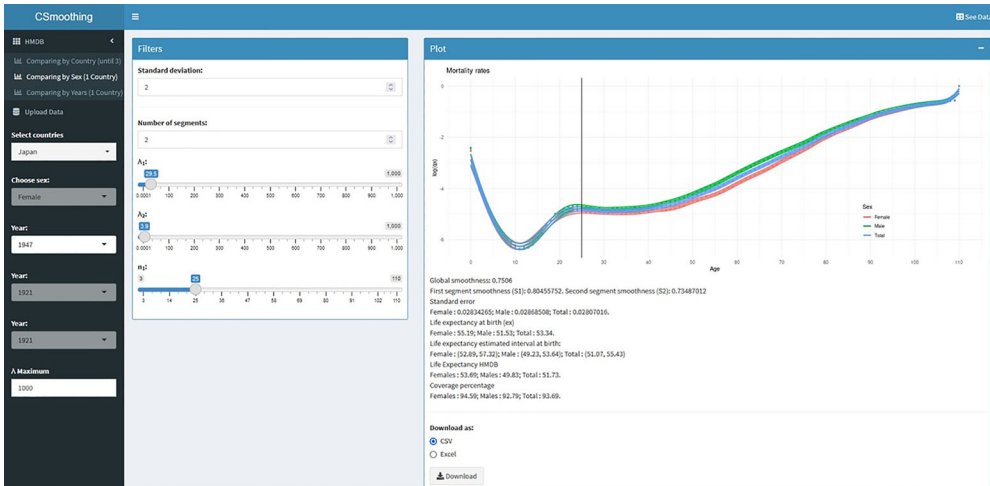


Figure 7. Comparing Japan mortality by sex with $100S\% = 75\%$ and two segments.

For this case, using two segments serves to reduce the standard error of the estimated trend and produces a result that makes demographic sense. In order to get that, one possibility is to fix the cutoff point in $x=25$ (approximately half of the male life expectancy). Then, the global smoothness index remains in the same level, that is $100S\% = 75\%$, and with the parameters $\lambda_1 = 29.5$ ($S_1 = 0.8050$) and $\lambda_2 = 3.9$ ($S_2 = 0.7350$). So, given this segmentation the standard errors decrease and they are 0.0283 (Female), 0.0287 (Male) and 0.0281 (Total). The corresponding estimates of life expectancy at birth with their intervals by sex are 55.19 (52.89, 57.32), 51.53 (49.23, 53.64) and 53.34 (51.07, 55.43) respectively.

4.3. Comparing by years

In this application we analyze the Denmark case for the last three years: 2018, 2019 and 2020, where the aim is to identify changes in life expectancy by sex. We want to dimension the impact by direct and/or indirect deaths happened amid the COVID-19 pandemic in 2020 to be of major concern. A summary indicator of mortality such as the life expectancy at birth is useful to that end. As a matter of fact, it could be possible to do that for every country once the data are available, see for instance the approach made by Andrasfay and Goldman (2021; Figure 7). So, the global smoothness index chosen is $100S\% = 75\%$ that corresponds to the smoothing parameter $\lambda_1 = 5.6$, then the point indicator is estimated with two standard deviations by sex (see Table 1).

It is worth saying that life expectancies taken from HMD are very close to ours, and all those figures are contained in our estimated intervals. If we go further, using CSmoothering we can infer what the λ_1 value that reproduces life expectancies is. In other words, with $\lambda_1 = 0.0001$ - the minimum allowed in CSmoothering - we are able to get the HMD estimates. So, it is evident that smoothing is required to estimate an underlying trend (approximating mortality curves) and with them the life expectancies.

According to Figure 8, the total mortality seems very similar for the three years, with greater variability found for the children's data; likewise, there is a little increase of mortality among people aged 20 and 30 years old in 2020. Something similar occurs with the male mortality trend, however there are several outliers that are ignored in order to get the estimated trends. For female data it is notorious the high mortality around people aged 10 and 20 years old in 2020. Overall, opposite to what was expected in many countries amid the COVID-19 pandemic, the life expectancies by sex have had a little increase in Denmark in 2020. The increases were 0.16, 0.20 and 0.13 for men, women and total respectively between 2019 and 2020. It is worth mentioning that our results are in line with those taken from Aburto et al. (2022).

Table 1. Life expectancy for Denmark with CSmoothing vs that provided by HMD, 2sd interval below with estimated standard error of estimated trend. Note: * denotes HMD life expectancy.

Year/Sex	Male	Female	Total
2018	79.30 vs 79.02*	83.16 vs 82.96*	81.19 vs 80.99*
	(75.08, 83.46)	(79.60, 86.68)	(79.89, 82.48)
2019	79.65 vs 79.44*	83.58 vs 83.42*	81.60 vs 81.43*
	(76.57, 82.69)	(82.15, 85.02)	(80.39, 82.80)
2020	79.81 vs 79.58*	83.78 vs 83.51*	81.73 vs 81.55*
	(75.49, 84.03)	(79.89, 87.67)	(80.30, 83.16)
	0.3659	0.3626	0.1252

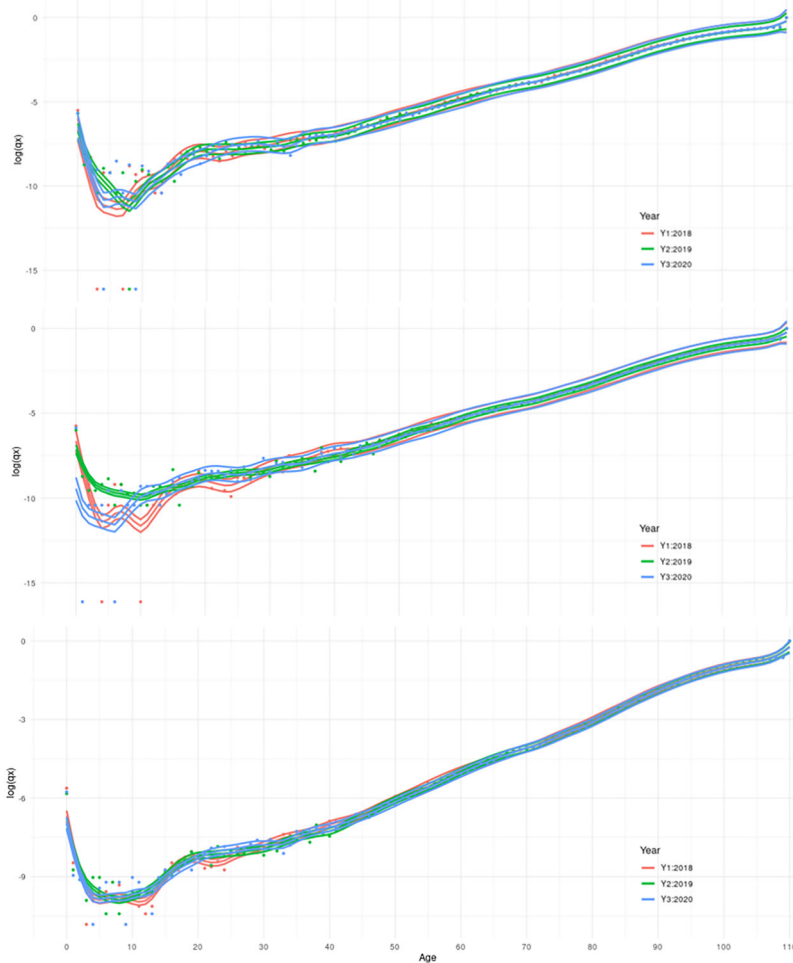


Figure 8. Comparing Danish mortality by years with 100% = 75%. Top: Male, Middle: Female, Down: Total.

5. Conclusion

CSmoothing is a Web-tool for Controlled Smoothing by segments suitable for analyzing mortality data as shown in the examples with cutoffs on selected ages. This tool can also be applied to other types of data sets, such as economic as well as other demographic time series. In particular, it is intended to be an easy-to-use tool for the analysis of mortality data. Some examples of the use of CSmoothing together with the underlying model are provided. Those examples include a

short interpretation of results from both a demographic and a statistical perspective. When smoothing by segments, the application allows users to produce estimates with the HMD or upload their own data, and then easily generate smoothed curves with exploratory character. The analyst may choose and control the degree of smoothness employing to that end an appropriate index.

Appendix

Table A1. Available data sets by country and years in CSmoothing from HMD.

Number	Country (CODE)	From	To	Total
1	Australia (AUS)	1921	2018	98
2	Austria (AUT)	1947	2019	73
3	Belarus (BLR)	1959	2018	60
4	Belgium (BEL)	1841	2020	180
5	Bulgaria (BGR)	1947	2017	71
6	Canada (CAN)	1921	2019	99
7	Chile (CHL)	1992	2017	26
8	Croatia (HRV)	2001	2019	19
9	Czechia (CZE)	1950	2019	70
10	Denmark (DNK)	1835	2020	186
11	Estonia (EST)	1959	2019	61
12	Finland (FIN)	1878	2020	143
13	France total population (FRATNP)	1816	2018	203
14	France civilian population (FRACNP)	1816	2018	203
15	Germany total population (DEUTNP)	1990	2017	28
16	East Germany (DEUTE)	1956	2017	62
17	West Germany (DEUTW)	1956	2017	62
18	Greece (GRC)	1981	2017	37
19	Hungary (HUN)	1950	2017	68
20	Iceland (ISL)	1838	2018	181
21	Ireland (IRL)	1950	2017	68
22	Northern Ireland (GBR NIR)	1922	2018	97
23	Israel (ISR)	1983	2016	34
24	Italy (ITA)	1872	2018	147
25	Japan (JPN)	1947	2019	73
26	Latvia (LVA)	1959	2019	61
27	Lithuania (LTU)	1959	2019	61
28	Luxembourg (LUX)	1960	2019	60
29	Netherlands (NLD)	1850	2019	170
30	New Zealand total population (NZL NP)	1948	2013	66
31	New Zealand Maori (NZL MA)	1948	2008	61
32	New Zealand non-Maori (NZL NM)	1901	2008	108
33	Norway (NOR)	1846	2020	175
34	Poland (POL)	1958	2019	62
35	Portugal (PRT)	1940	2020	81
36	Republic of Korea (KOR)	2003	2018	16
37	Russia (RUS)	1959	2014	56
38	Scotland (GBR SCO)	1855	2018	164
39	Slovakia (SVK)	1950	2017	68
40	Slovenia (SVN)	1983	2017	35
41	Spain (ESP)	1908	2018	111
42	Sweden (SWE)	1751	2020	270
43	Switzerland (CHE)	1876	2018	143
44	United Kingdom total population (GBR NP)	1922	2018	97
45	England and Wales total population (GBR TENW)	1841	2018	178
46	England and Wales civilian population (GBR CENW)	1841	2018	178
47	Taiwan (TWN)	1970	2019	50
48	USA (USA)	1933	2019	87
49	Ukraine (UKR)	1959	2013	55

Acknowledgements

The authors gratefully acknowledge the comments and suggestions from an anonymous reviewer and the editor of this journal. Eliud Silva dedicates this work to the memory of one of his best german friends, Francis (Pancho), who died surprisingly amid the COVID-19 pandemic in Mexico. Víctor M. Guerrero thanks the financial support provided by Asociación Mexicana de Cultura, A. C. to carry out this work.

ORCID

Eliud Silva  <http://orcid.org/0000-0003-0499-0446>
Víctor M. Guerrero  <http://orcid.org/0000-0003-2184-5216>
Yhael Jacinto Cruz  <http://orcid.org/0000-0003-3640-1415>
Emanuel Rodriguez Soler  <http://orcid.org/0000-0003-4166-3292>

References

- Aburto, J. M., J. Schöley, I. Kashnitsky, L. Zhang, C. Rahal, T. I. Missov, M. C. Mills, J. B. Dowd, and R. Kashyap. 2022. Quantifying impacts of the COVID-19 pandemic through life expectancy losses: A population-level study of 29 countries. *International Journal of Epidemiology* 51 (1):63–74. doi:10.1093/ije/dyab207.
- Alburez-Gutierrez, D., E. Zagheni, S. Aref, S. Gil-Clavel, A. Grow, and D. V. Negraia. 2019. Demography in the digital era: New data sources for population research. SocArXiv.
- Andrasfay, T., and N. Goldman. 2021. Reductions in 2020 US life expectancy due to COVID-19 and the disproportionate impact on the Black and Latino populations. *Proceedings of the National Academy of Sciences* 118 (5):1–6. doi:10.1073/pnas.2014746118.
- Barbieri, M., J. R. Wilmoth, V. M. Shkolnikov, D. Gleij, D. Jasilionis, D. Jdanov, C. Boe, T. Riffe, P. Grigoriev, and C. Winant. 2015. Data resource profile: The human mortality database (HMD). *International Journal of Epidemiology* 44 (5):1549–56. doi:10.1093/ije/dyv105.
- Bowman, A., and L. Evers. 2013. Lecture notes: Nonparametric smoothing. University of Glasgow, School of Mathematics Statistics, Glasgow. www2.warwick.ac.uk/fac/sci/statistics/apts/students/resources/1617/smoothing-notes.pdf.
- Camarda, C. G. 2012. MortalitySmooth: An R package for smoothing poisson counts with P-Splines. *Journal of Statistical Software* 50 (1):1–24. doi:10.18637/jss.v050.i01.
- Chandran, A., C. Xu, J. Gross, K. M. Leifheit, D. Phelan-Emrick, S. Helleringer, and K. N. Althoff. 2021. A Web-based tool for quantification of potential gains in life expectancy by preventing cause-specific mortality. *Frontiers in Public Health* 9:663825.
- Chang, W., J. Cheng, J. Allaire, Y. Xie, and J. McPherson. 2017. Shiny: Web application framework for R. *R Package Version 1* (5):2017.
- Cortés-Toto, D., H. J. Guerrero, and Reyes, V. M. 2017. Trend smoothness achieved by penalized least squares with the smoothing parameter chosen by optimality criteria. *Communications in Statistics - Simulation and Computation* 46 (2):1492–507. doi:10.1080/03610918.2015.1005236.
- Das, A. K., S. Mishra, and S. Gopalan. 2020. Predicting COVID-19 community mortality risk using machine learning and development of an online prognostic tool. *PeerJ* 8:e10083. doi:10.7717/peerj.10083.
- De Boor, C. 1978. *A practical guide to splines*. New York: Springer Verlag.
- Devedžić, M., and V. Devedžić. 2003. Towards web-based education of demography. *Informatics in Education* 2 (2):201–10., and doi:10.15388/infedu.2003.15.
- Devedžić, M., and V. Devedžić. 2010. Imagine: Using new web technologies in demography. *Social Science Computer Review* 28 (2):206–31.
- Eilers, P. H., and B. D. Marx. 1996. Flexible smoothing with B-splines and penalties. *Statistical Science* 11 (2): 89–121. doi:10.1214/ss/1038425655.
- Guerrero, V. M., and E. Silva. 2015. Smoothing a time series by segments of the data range. *Communications in Statistics - Theory and Methods* 44 (21):4568–85. doi:10.1080/03610926.2014.901372.
- Guerrero, V. M. 2007. Time series smoothing by penalized least squares. *Statistics & Probability Letters* 77 (12): 1225–34. doi:10.1016/j.spl.2007.03.006.
- Guerrero, V. M. 2008. Estimating trends with percentage of smoothness chosen by the user. *International Statistical Review* 76 (2):187–202. doi:10.1111/j.1751-5823.2008.00047.x.
- Hodrick, R. J., and E. C. Prescott. 1997. Postwar U.S. business cycles: An empirical investigation. *Journal of Money, Credit and Banking* 29 (1):1–16. doi:10.2307/2953682.
- Hyndman, R. J. 2008. ETC5410: Nonparametric smoothing methods. Accessed December 3, 2020. <https://www.robjhyndman.com/etc5410/fda.pdf>.

- Killick, R., and I. Eckley. 2014. Changepoint: An R package for changepoint analysis. *Journal of Statistical Software* 58 (3):1–19. doi:10.18637/jss.v058.i03.
- Noll, N. B., I. Aksamentov, V. Druelle, A. Badenhorst, B. Ronzani, G. Jefferies, J. Albert, and R. A. Neher. 2020. COVID-19 scenarios: An interactive tool to explore the spread and associated morbidity and mortality of SARS-CoV-2. MedRxiv. doi:10.1101/2020.05.05.20091363.
- Perperoglou, A., M. Sauerbrei, M. Abrahamowicz, and Schmid, W. 2019. A review of spline function procedures in R. *BMC Medical Research Methodology* 19 (1):1–16. doi:10.1186/s12874-019-0666-3.
- R Core Team. 2021. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Ramsey, J., and, and B. Ripley. 2017. P-spline: Penalized smoothing splines. R Package Version 1.0-18.
- Rosenberg, P. S., D. P. Check, and W. F. Anderson. 2014. A Web tool for age–period–cohort analysis of cancer incidence and mortality rates: Software for cancer rates and trends. *Cancer Epidemiology, Biomarkers & Prevention : A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology* 23 (11):2296–302. doi:10.1158/1055-9965.EPI-14-0300.
- RStudio Team. 2021. RStudio: Integrated development environment for R. RStudio, PBC, Boston, MA. <http://www.rstudio.com/>.
- Silva, E., and A. Ovin. 2018. Aproximación a curvas de mortalidad a través de una propuesta no paramétrica el caso del modelo de Heligman y Pollard. *Estudios Demográficos y Urbanos* 34 (1):129–67. doi:10.24201/edu.v34i1.1805.
- University of California Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). 2021. Human mortality database (HMD). www.mortality.org; www.humanmortality.de, data downloaded on 10/10/21.
- Wold, S. 1974. Spline functions in data analysis. *Technometrics* 16 (1):1–11. doi:10.1080/00401706.1974.10489142.