ESTADISTICA (SANTIAGO DE CHILE), vol. 66, núm. 186, 2014, pp. 65-89.

ANÁLISIS DE LA MORTALIDAD POR EDAD Y SEXO MEDIANTE MODELOS PARA DATOS FUNCIONALES.

Andreozzi, Lucía y Blaconá María Teresa.

ESTADÍSTICA (2014), 66, 186 y 187, pp. 65-89

ANÁLISIS DE LA MORTALIDAD POR EDAD Y SEXO MEDIANTE MODELOS PARA DATOS FUNCIONALES

BLACONÁ, M.T.
Consejo de Investigaciones, Facultad de Ciencias Económicas y Estadística
Universidad Nacional de Rosario, Argentina
mblacona@fcecon.unr.edu.ar

ANDREOZZI, L.
Consejo de Investigaciones, Facultad de Ciencias Económicas y Estadística
Universidad Nacional de Rosario – CONICET, Argentina
landreozzi@fcecon.unr.edu.ar

RESUMEN

En este trabajo se realiza una breve descripción del enfoque de datos funcionales para modelar las tasas de mortalidad por edad y sexo, esta propuesta es un avance sobre el modelo tradicional de Lee-Carter y alguna de sus modificaciones. El nuevo método se aplica a las tasas de mortalidad por edad y sexo de la Argentina. Una característica de estos nuevos modelos es que permiten interpretar el comportamiento de la mortalidad a través del tiempo relacionándolo con el comportamiento por edades, esto debido especialmente a la utilización de técnicas de componentes principales sobre los datos suavizados de las tasas de mortalidad por edad y sexo. Además se puede destacar que el método permite construir intervalos de pronóstico con un nivel de incertidumbre aceptable. En el caso del estudio de mortalidad por edad y sexo en Argentina los resultados de los pronósticos que se obtienen por este método son superiores a los obtenidos por modelos ARIMA, tanto en sus valores puntuales como en la amplitud de sus intervalos de pronóstico.

Palabras Clave

Tasas de mortalidad; series de tiempo; pronósticos; datos funcionales; Argentina.

ABSTRACT

In this paper a brief description of the approach of functional data to model mortality rates by age and sex is performed, this proposal is an improvement over the traditional model of Lee-Carter and some of its modifications. The new method is applied to mortality rates by age and sex of Argentina. A feature of these new models is they allow to describe the behavior of mortality over time relating it to the behavior by age, this especially due to the use of principal components techniques on the smoothed data of mortality rates by age and sex. Also be noted that the method allows to built prediction intervals with an acceptable level of uncertainty. For the study of mortality by age and sex in Argentina the forecasting results obtained by this method are superior to those obtained by ARIMA models, both, in their specific values and in the forecast intervals width.

Keywords

Mortality rates; time series; forecasts; functional data; Argentina.

1. Introducción

Los métodos estocásticos de pronósticos han recibido una considerable atención en el área demográfica. Uno de los métodos más destacado es el modelo de Lee-Carter (Lee y Carter, 1992) el cual actualmente posee numerosas variantes y extensiones. Los autores proponen una metodología que permite modelar y extrapolar las tendencias observadas en las tasas de mortalidad a largo plazo e implementan dicha metodología para pronosticar la mortalidad de los Estados Unidos hasta el año 2065. Modificaciones al método de Lee-Carter fueron propuestas por Lee y Miller (2001) y Booth et al. (2002), estos abordan la elección del período de ajuste, el método para la estimación del parámetro de nivel y la elección de las tasas base para el pronóstico, la propuesta de los primeros es ampliamente utilizada. Sin embargo la variante propuesta por Booth et al. (2002) ha demostrado ser al menos tan precisa como la de Lee y Miller en el corto plazo (Booth et al., 2005, Booth et al. 2006a). Otros desarrollos han incorporado una estructura de error heterocedástica Poisson (Brouhns et al., 2002, Wilmoth 1993), y algunos han extendido su aplicabilidad a los factores de reducción de la mortalidad (Renshaw y Haberman, 2003a) o examinado el uso de más de un término en el modelo (Booth et al., 2002, Renshaw y Haberman, 2003b). Además han sido desarrollados enfoques paralelos en el marco de los Modelos Lineales Generalizados (MLG), (Renshaw y Haberman, 2003c). También existen dos extensiones recientes que incluyen suavizados no paramétricos en el modelo; Jong y Tickle (2006) combina un suavizado por *splines* y una estimación por medio del filtro de Kalman para

ajustar una versión generalizada del modelo de Lee y Carter. Siguiendo el paradigma de los datos funcionales Hyndman y Ullah (2007) proponen suavizar la mortalidad utilizando regresiones *spline* penalizadas para luego ajustar un modelo utilizando una descomposición en componentes principales. Estos métodos fueron comparados por Booth et al. (2006).

Como se muestra en el párrafo anterior en los últimos años se han desarrollado múltiples enfoques para pronósticos probabilísticos los más utilizados son aquellos que implican alguna forma de extrapolación, frecuentemente utilizando modelos de series de tiempo. Recientemente, se han introducido los métodos para datos funcionales (Ramsay y Silverman, 2005) que es un nuevo enfoque de análisis de series de tiempo y que han sido adoptados, entre otras finalidades, para pronósticos demográficos (Hyndman y Ullah, 2007). El trabajo de estos últimos difiere en varios aspectos importantes con respecto al modelo de Lee-Carter. En primer lugar está enmarcado en el paradigma de datos funcionales que utiliza suavizados no paramétricos con el fin de reducir la aleatoriedad inherente en los datos observados. Y en segundo lugar, para la descomposición de las componentes demográficas permite utilizar componentes principales clásicos o robustos.

Los métodos para datos funcionales tienen la ventaja de proveer un entorno flexible que puede ser utilizado para pronosticar las tres componentes demográficas. El presente trabajo aplica modelos para datos funcionales para pronosticar la mortalidad que será uno de los insumos para el pronóstico nacional de la población.

Los pronósticos de cada conjunto de tasas se combinan utilizando el método de las componentes y una simulación por Monte Carlo de pronósticos probabilísticos de población por edad y sexo. El uso de métodos de extrapolación presupone que las tendencias del pasado continuarán en el futuro. Este supuesto básico de series de tiempo ha probado ser mejor para pronosticar que los modelos estructurales que involucran variables exógenas o los métodos basados en expectativas (Booth et. al., 2006)

El método se aplica a datos de Argentina en el período 1980 a 2010, para las tasas de mortalidad de hombres, mujeres y total, con un horizonte de pronóstico de 10 años. Este artículo representa un avance sobre las técnicas presentadas en Blaconá y Andreozzi (2012), en donde se estiman las tasas de mortalidad en la República Argentina para el período 1979-2009 utilizando el modelo propuesto por Lee y Carter. Las estimaciones obtenidas por este método permiten describir la tendencia y el patrón de cambio de la mortalidad. Además, se compararon estimaciones de los parámetros del modelo para total, varones y mujeres mediante el método clásico, mínimos cuadrados ponderados (MCP) y máxima verosimilitud-modelo

log-bilineal Poisson (MV-LBP). En cambio, en este trabajo se utiliza la metodología propuesta por Hyndman y Ullah (2007), ampliada por Hyndman y Booth (2008) y se realiza una breve comparación con el modelo anteriormente propuesto.

En la sección II se presenta una breve descripción de los modelos para datos funcionales y se plantea el modelo de Hyndman y Booth (2008) para datos demográficos. En la sección III se realiza el análisis empírico para las tasas de mortalidad de Argentina utilizando el modelo descripto en la sección II y finalmente en la sección IV se presentan las conclusiones.

2. Metodología

2.1. Modelos para Datos Funcionales

En primer lugar se realiza una breve descripción del enfoque de datos funcionales en demografía (Hyndman y Ullah, 2007, Hyndman y Booth, 2008). Para ello se definen los datos necesarios para estimar la mortalidad, como sigue:

$D_t(x)$: Muertes en el año calendario t de la población de edad x,
$E_t(x)$: Población de edad x expuesta al riesgo al 30 de Junio del año t,

donde $x = 1, 2, \ldots, p^+$ y $t = 1, 2, \ldots, n$. Con p^+ se indica el último grupo de edad abierto. La tasa de mortalidad de la edad x en el año calendario t se define como:

$$m_t(x) = \frac{D_t(x)}{E_t(x)}$$

Se denota con y_t^* a la cantidad a ser modelada, (es posible modelar la mortalidad, la fecundidad o la migración neta) para la edad x en el año t. Primero se plantea una transformación de Box-Cox de y_t^* para permitir una variación que aumente a medida que crece el valor de la variable, es decir la variabilidad de las tasas es mayor para las edades más avanzadas.

$$y_t(x) \begin{cases} \frac{1}{\lambda}\left([y_t^*]^\lambda - 1\right) \text{ si } 0 < \lambda \leq 1 \\ \log(y_t^*) \text{ si } \lambda = 0 \end{cases}. \quad (1)$$

El valor de λ determina la intensidad de la transformación, no se aplica ninguna cuando $\lambda = 1$.Luego se supone el siguiente modelo para la cantidad transformada $y_t(x)$

$$y_t(x) = s_t(x) + \sigma_t(x)\varepsilon_{t,x}$$
$$s_t(x) = \mu(x) + \sum_{k=1}^{K} \beta_{t,k}\phi_k(x) + e_t(x), \quad (2)$$

donde $s_t(x)$ es una función suave subyacente de x, $\varepsilon_{t,x}$ son variables aleatorias gausseanas, independientes e idénticamente distribuidas y $\sigma_t(x)$es la variancia que puede variar con la edad y con el tiempo. Es posible implementar el enfoque para años y edades simples como así también para grupos quinquenales. Esto significa que $s_t(x)$es una función suave de la edad que se observa con error. La segunda ecuación describe la dinámica de $s_t(x)$ a través del tiempo. En esta ecuación, $\mu(x)$es la media de $s_t(x)$a través de los años, $\{\phi_k(x)\}$es un conjunto de funciones base ortogonales calculadas utilizando una descomposición en componentes principales, $e_t(x)$ es el error del modelo, el cual se supone no correlacionado serialmente. La dinámica del proceso está controlada por los coeficientes$\{\beta_{t,k}\}$, los cuales tienen un comportamiento independiente uno de otro (por propiedades del método de componentes principales). Existen tres fuentes de variación en el modelo: $\varepsilon_{t,x}$ representa la variación aleatoria con respecto a la distribución relevante, para los nacimientos, muertes y migrantes (Poisson o Normal); $e_t(x)$representa el modelo de los residuos que surge al modelar $s_t(x)$utilizando un conjunto de funciones bases; además existe una aleatoriedad inherente al modelo de series de tiempo para cada $\{\beta_{t,k}\}$que ejerce los cambios en la dinámica de la curva suave $\{s_t(x)\}$.

Este modelo fue propuesto inicialmente por Hyndman y Ullah, (2007) para modelar tasas de mortalidad y fecundidad con una transformación logaritmo en lugar de la transformación general de Box-Cox. También ha sido utilizado por Erbas et. al. (2007) para pronosticar tasas de mortalidad por cáncer de mama. Como señalan los primeros el modelo es una generalización del conocido modelo de Lee y Carter (1992) para pronosticar tasas de mortalidad. En este enfoque, y_t^* representa a la tasa de mortalidad y $\lambda = 0$, por ello $y_t(x)$es el logaritmo de la mortalidad para el año t y la edad x. En el modelo de Lee-Carter no se realiza ningún tipo de suavizado, por ello $\sigma_t(x) = 0, y_t(x) = s_t(x)$ y a$\mu(x)$se la estima como el promedio de $y_t(x)$ a través de los años. Para $K = 1$, $\beta_{t,1}$se obtiene a partir de la primera componente principal de la matriz$[y_t(x) - \hat{\mu}(x),]$. Los pronósticos se obtienen ajustando un modelo de serie de tiempo a $\beta_{t,1}$; en la práctica el modelo que se obtiene resulta generalmente un paseo aleatorio con pendiente.

Hyndman y Booth (2008) extienden el método de Hyndman y Ullah (2007) utilizando una transformación más general y modificando el método de cálculo de la variancia de pronóstico para permitir una mejor calibración con los datos observados. Los pasos a seguir para la estimación del modelo son:

a) Estimar las funciones suaves $s_t(x)$utilizando regresión no paramétrica aplicada sobre $y_t(x)$ para cada año t,
b) Estimar $\mu(x)$ como la media de $s_t(x)$ a través de los años,
c) Estimar $\beta_{t,k}\, y \phi_k(x)$, con $k = 1, \ldots, K$ por medio de análisis de componentes principales sobre la matriz $[y_t(x) - \hat{\mu}(x)]$,
d) Estimar un modelo de series de tiempo para cada $\beta_{t,k}$ donde $k = 1, \ldots, K$. Para ello es posible utilizar modelos ARIMA (Box y Jenkins, 1976) o modelos de espacio de estado de innovaciones (Hyndman et al., 2008).

Se debe especificar el valor de K, Hyndman y Ullah (2007) sostienen que el método es insensible al valor elegido siempre y cuando sea lo suficientemente grande. Esto significa, que el costo al elegir K grande es pequeño (más allá del tiempo computacional), mientras que seleccionar un K pequeño puede producir menor exactitud en los pronósticos. Hyndman y Booth (2008) utilizan $K = 6$ para todos los componentes demográficos; esta cantidad puede ser mayor a la que realmente se requiere.

La variancia observacional, depende de la naturaleza de los datos. Para las muertes se estima a partir de $y_t^* = m_t(x)$suponiendo que las muertes se distribuyen Poisson (λ) (Brillinger, 1986) con media $m_t(x)E_t(x)$. Luego y_t^* tiene una variancia aproximada $E_t(x)^{-1}m_t(x)$ y la variancia de y_t (por aproximación de Taylor) es

$$\sigma_t^2 \approx [m_t(x)]^{-2\lambda-1}E_t(x)^{-1}. \qquad (3)$$

El suavizado no paramétrico puede ser llevado a cabo utilizando alguno de los múltiples métodos de suavizados existentes, ver Ruppert et al. (1976) y Simono (1976). Pero, puntualmente, Hyndman y Booth (2008) sugieren utilizar regresión *spline* penalizada con restricciones para la mortalidad, de modo que los pesos contemplen la heterogeneidad presente en este tipo de datos. Se impone además una restricción de monotonía para la mortalidad, más específicamente, se definen pesos iguales a la inversa de la variancia teórica (derivada del supuesto de la distribución de Poisson) y se utiliza una regresión *spline* penalizada, ver Wood (2003) y He y Hg (1999), para estimar las curvas $y_t(x)$ que representan a la tasas $m_t(x)$ luego de la transformación de Box-Cox. Se impone además una restricción; las curvas deben ser monótonamente crecientes para $x > c$ es decir para una edad

determinada, lo que permite reducir el ruido en las curvas estimadas para edades avanzadas. La imposición resulta lógica dado que cuando más anciana es una persona, la probabilidad de muerte es mayor.

Para suavizar las tasas de mortalidad Hyndman y Ullah (2007) utilizan el enfoque de modelos aditivos generalizados (*gam*), que son modelos lineales generalizados con un predictor lineal que involucra una suma de funciones suavizadas de las covariables, y en los que el grado del suavizado se estima como parte del proceso de estimación. Las funciones de suavizado se representan a través de regresión *spline* penalizada, Wood (2003), ya que ésta permite incluir de un modo sencillo restricciones de monotonía creciente para edades mayores a los 65 años. Para realizar el suavizado es posible utilizar la función *gam* del paquete *mgvc* de R.

La adecuación del modelo depende de la bondad del ajuste de la superficie bivariada $\{s_t(x) - \mu_t(x)\}$, la cual puede ser aproximada por la suma de unos pocos productos de funciones univariadas del tiempo (t) y la edad (x). Hyndman y Ullah (2007) sugieren que el modelo que se obtiene es adecuado para producir pronósticos, pero no lo es tanto para estimar la variancia de pronóstico, por lo que proponen un ajuste para su cálculo.

2.2. Pronósticos funcionales

Se tienen datos hasta un tiempo $t = n$ y se desean estimar valores futuros de $y_t(x)$ para $t = n + 1, \ldots, n + h$ para todo x. Si con $\hat{\beta}_{n,k,h}$ se denota el pronóstico h pasos hacia adelante de $\beta_{n+h,k}$, y siendo $\hat{y}_{n+h}(x)$ el pronóstico h pasos hacia adelante de $y_{n+h}(x)$, $\hat{s}_{n,h}(x)$ es el pronóstico h pasos hacia delante de $s_{n+h}(x)$. Luego,

$$\hat{y}_{n+h}(x) = \hat{s}_{n,h}(x) = \hat{\mu}(x) + \sum_{k=1}^{K} \hat{\beta}_{n,k,h}\hat{\phi}_k(x), \qquad (4)$$

un pronóstico de y_t^* se encuentra a través de la transformación inversa.

Según Hyndman y Ullah (2007) se puede expresar la variancia del pronóstico como,

$$V_h(x) = \mathrm{Var}[s_{n+h}(x)|\mathcal{I}, \Phi] = \hat{\sigma}_\mu + \sum_{k=1}^{K} u_{n+h,k}\hat{\phi}_k(x) + v(x)\ , \qquad (5)$$

Donde $\mathcal{I} = \{y_t(x_i)\}$ denota todos los datos observados, $u_{n+h,k} = \mathrm{Var}(\beta_{n+h,k}|\beta_{1,k}, \ldots, \beta_{n,k})$, se puede obtener a partir del modelo de series de tiempo, y la $\hat{\sigma}_\mu^2(x)$ (la variancia del estimador suave $\hat{\mu}(x)$), a través del método de suavizado empleado, y $v(x)$ se estima promediando $\hat{e}_t^2(x)$ para cada x. El error

de suavizado está dado por el primer término, el error debido a la predicción de la dinámica está dado por el segundo término, y el tercero es el error debido a la variación dinámica sin explicar. Luego de que se incluye el error observacional, se obtiene,

$$\mathrm{Var}[y_{n+h}(x)|\mathcal{J},\Phi] = V_h(x) + \sigma_t^2(x). \quad (6)$$

Se destaca que en esta formulación se contemplan las correlaciones entre las edades, a través de las funciones suavizadas de la edad x. Asimismo las correlaciones entre años se tienen en cuenta a través del modelo de series de tiempo planteado para los coeficientes $\beta_{t,1},\ldots,\beta_{t,K}$. Para valores bajos de h es posible probar la validez de $V_h(x)$ calculando la variancia de pronóstico empírica dentro de la muestra

$$W_h(x) = \frac{1}{n-h-m+1}\sum_{t=m}^{n-h}\left[s_{n+h} - \hat{s}_{t,h}\right]^2, \quad (7)$$

donde m es el menor número de observaciones utilizadas para ajustar el modelo. En la práctica, pueden existir considerables diferencias entre $W_h(x)$ y $V_h(x)$. En consecuencia se utiliza la siguiente expresión de la variancia ajustada:

$$\mathrm{Var}[y_{n+h}(x)|\mathcal{J},\Phi] = V_h(x)W_1(x)\sigma_t^2(x). \quad (8)$$

Este ajuste hace que la variancia del pronóstico un paso hacia adelante coincida con la variancia del pronóstico empírica un paso hacia adelante dentro de la muestra. Se supone que el mismo ajuste multiplicativo se aplica a mayores horizontes de pronóstico.

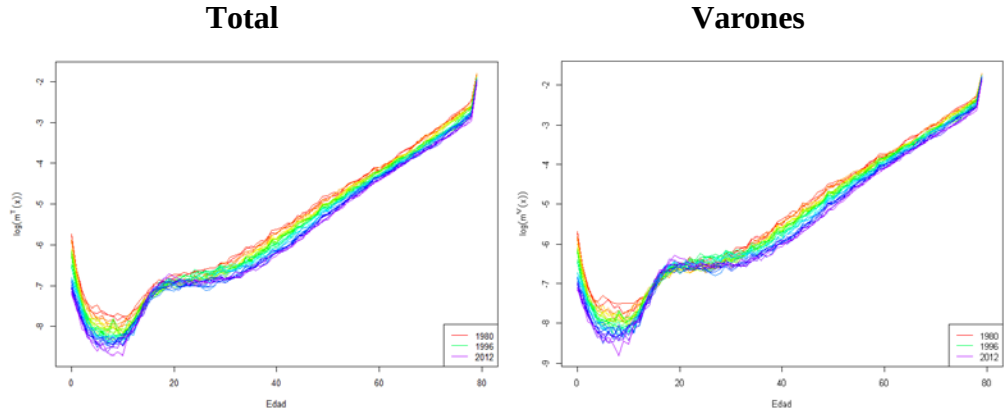
3. Análisis empírico de la mortalidad en Argentina

Los datos relativos a la mortalidad anual consisten en las tasas de mortalidad específicas por edad simple y sexo de la República Argentina durante el período 1980 a 2012. Las mismas se construyen en base a datos de la Dirección de Estadísticas e Información de Salud, que proporciona el número de defunciones y el Instituto Nacional de Estadística y Censo (INDEC), que brinda los datos de población. Los datos se encuentran desagregados en edades simples con un grupo abierto final de 80 años y más. Se utiliza la transformación logarítmica de las tasas.

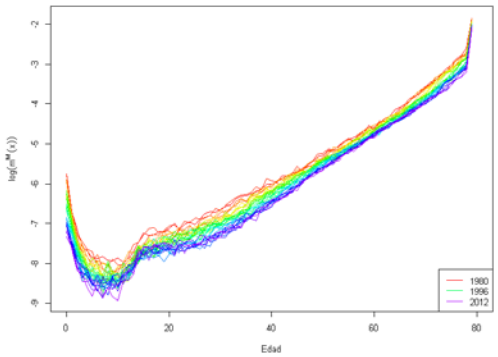
La figura 1 en la que se representan los logaritmos de las tasas observadas para el total de la población, presenta la forma típica del patrón de la mortalidad, alta del inicio de la vida, luego un descenso hasta antes de los 10 años seguido de un aumento hasta su pico alrededor de los 20 años, fenómeno presente principalmente

en varones y en mucho menor medida en las mujeres. Este pico está vinculado, en ésta como en la mayoría de las poblaciones, a accidentes de tránsito, muertes relacionadas al consumo de drogas y muertes violentas en general. Luego de este pico, las tasas presentan un leve descenso para volver a crecer en forma sostenida hasta las edades avanzadas, en este punto se observa un pico final que no se debe a la naturaleza de los datos observados sino a la agrupación forzada en mayores de 80 años (por falta de información desagregada para este grupo de edad), cuando sería de sumo interés que fueran analizadas como edades simples, dada la importante información que se obtendría en relación al fenómeno del envejecimiento poblacional. Un aspecto que se destaca visualmente es la caída en los niveles a través del tiempo, que se presenta en todas las edades, excepto para el pico alrededor de los 20 años, franja en la que parece observarse una caída más leve o fluctuante. La caída en general, en los niveles de la mortalidad se atribuye principalmente a las mejoras en la medicina y salud pública, en especial cuando se evalúan dinámicas de largo plazo (períodos de estudio de 100 años o más), por lo cual el descenso que se observa en el presente período de tiempo puede obedecer en parte a estas causas y quizás a causas tales como una mayor educación, condiciones de higiene y otras.

Figura 1. Logaritmos de las tasas de mortalidad observadas, para el total, varones y mujeres. Argentina 1980-2012.



Mujeres

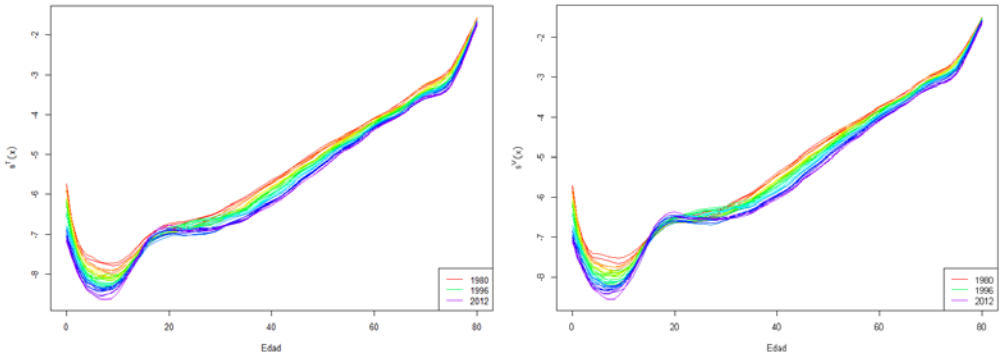


Fuente: Elaboración propia en base a datos de la DEIS e INDEC.

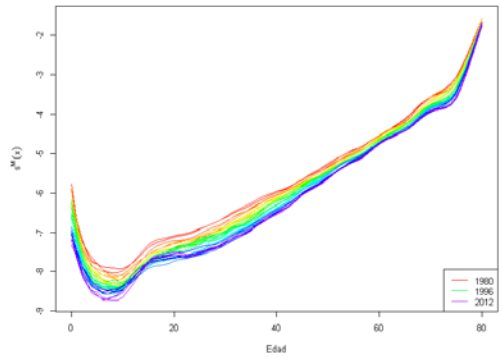
Para convertir las tasas observadas en datos funcionales (figura 2) se estiman a partir de ellas curvas suavizadas. Para este paso se aplica una regresión *spline* penalizada, dado que la metodología permite incorporar la restricción de monotonía: se supone que la función suavizada es monótona creciente para alguna edad $x > c$ (en esta aplicación se especifica $c = 65$ años, edad usada por Hyndman y Ullah (2007) y que además es la edad tope en Argentina para pasar a retiro). Esta restricción permite disminuir el ruido en las curvas estimadas para edades avanzadas lo que resulta razonable en este contexto. La implementación se lleva a cabo mediante una modificación de la versión propuesta por Wood (2006) a fin de incluir la restricción. Todo el análisis se realiza usando el paquete *demography* de R, desarrollado por Rob J. Hyndman.

Figura 2. Logaritmos de las tasas de mortalidad suavizados, para el total, varones y mujeres. Argentina 1980-2012.

Total **Varones**



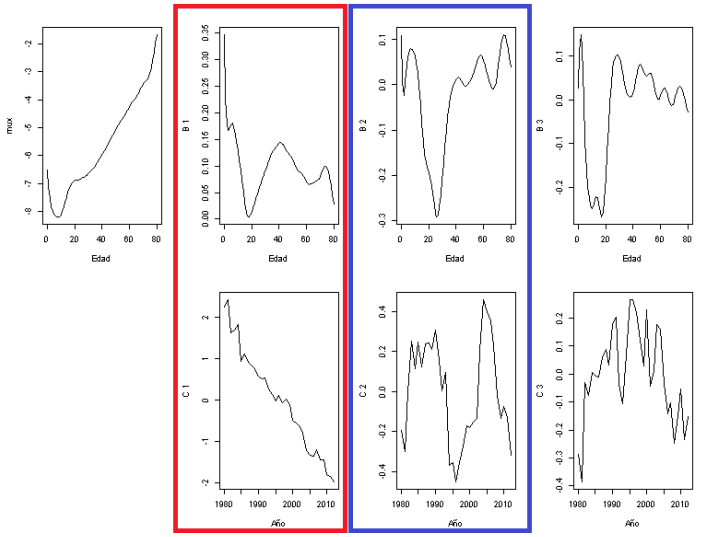
Mujeres



Fuente: Elaboración propia en base a datos de la DEIS e INDEC.

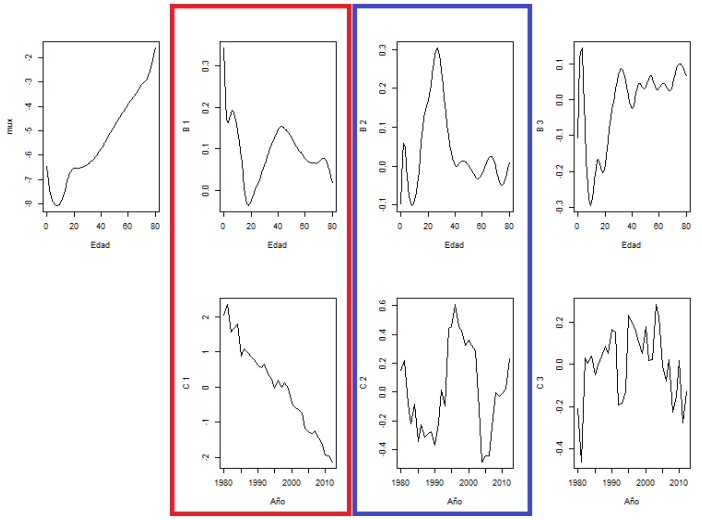
La siguiente etapa consiste en realizar una descomposición mediante un análisis de componentes principales. De acuerdo a lo recomendado por Hyndman y Ullah (2007) se ajustan seis bases ($K = 6$) y se obtienen sus correspondientes coeficientes asociados. No obstante, Hyndman y Booth (2008) señalan que no es posible interpretar más allá de la segunda base. Los resultados para el total, los varones y las mujeres se presentan en las figuras 3, 4 y 5.Estas figuras deben interpretarse de la siguiente manera: el primer recuadro de la primera fila representa el comportamiento promedio de la mortalidad a través de las edades. A partir del segundo recuadro la primera fila (base) se debe interpretar en forma conjunta con el correspondiente recuadro de la segunda fila (coeficientes). Por ejemplo, en el caso de la mortalidad total de la población, para la primera componente el recuadro de la segunda fila muestra un decrecimiento en la mortalidad a través del tiempo, que al ponerlo en correspondencia con el recuadro de la primera fila éste muestra que el comportamiento de ese decrecimiento es especialmente en los primeros años de vida (mortalidad infantil y primeros años), en menor medida para los mayores de 40 años, pero ese decaimiento no es tan notorio en las personas de alrededor de 20 años (pico hacia abajo en esas edades).

Figura 3. Bases y Coeficientes del Modelo de datos funcionales para el Total.



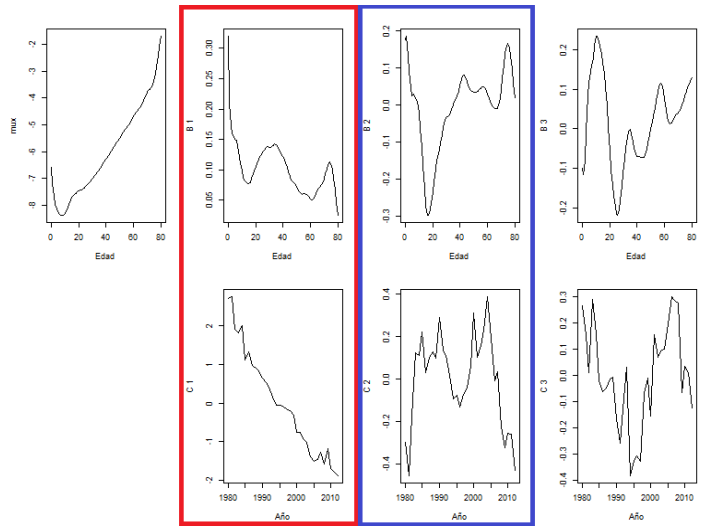
Fuente: Elaboración propia en base a datos de la DEIS e INDEC.

Figura 4. Bases y Coeficientes del Modelo de datos funcionales para los Varones.



Fuente: Elaboración propia en base a datos de la DEIS e INDEC

Figura 5. Bases y Coeficientes del Modelo de datos funcionales para las Mujeres.



Fuente: Elaboración propia en base a datos de la DEIS e INDEC

En el comportamiento de la mortalidad del total de la población se solapan las tendencias de ambos géneros. Las bases explican el 91.6%, 3.8%, 1.6%, 0.7%, 0.6% y 0.4% de la variación en los datos respectivamente, dejando un 1.3% sin explicar. La primera base presenta una estructura similar a la de los varones, pero más suave mientras que la segunda base presenta un comportamiento similar a la misma base en mujeres.

Al analizar los resultados obtenidos para varones, las bases explican el 90.3%, 5.0%, 1.4%, 0.8%, 0.7% y 0.5% de la variación en los datos respectivamente, dejando un 1.3% sin explicar. El primer coeficiente representa la tendencia general, que desciende a través del tiempo, con un pico en el año 1982, el cual podría deberse al conflicto bélico por la soberanía de las Islas Malvinas en abril de ese año. La base nos indica de qué modo este descenso mostrado por el coeficiente se manifiesta para las distintas edades; es decir, el pico en las edades iniciales indica el descenso en la mortalidad infantil, seguido de valores altos para la niñez y las edades entre 40 y 50 años, en esta última franja presenta descensos a nivel mundial, los mismos podrían atribuirse por ejemplo al descenso en la muertes por afecciones cardíacas u otras patologías debido al avance de la medicina. La segunda componente indicaría que el coeficiente fluctúa con períodos de subas moderadas alrededor de 1980 y 2010 y una suba marcada a mediados de la década de los 90, menores bajas se presentan en los otros períodos. La base que se corresponde a este

coeficiente indica que este comportamiento se manifiesta en la franja de entre 18 y 30 años aproximadamente, es decir, que se relaciona con el fenómeno que caracteriza a las edades alrededor de los 20 años, si bien se presenta un leve corrimiento hacia edades superiores.

Si se analizan los resultados de las mujeres las bases explican el 92.2%, 2.2%, 1.8%, 0.9%, 0.6% y 0.5% de la variación en los datos respectivamente, dejando un 1.8% sin explicar. En este caso, la primera base se refiere a niñas, mujeres de 30 años y con un peso menor a mujeres de alrededor de 75 años, y como indica el coeficiente asociado a esta base, existiría una tendencia decreciente a lo largo del tiempo para estos grupos. También se detecta una leve fluctuación durante el período 2005-2010.

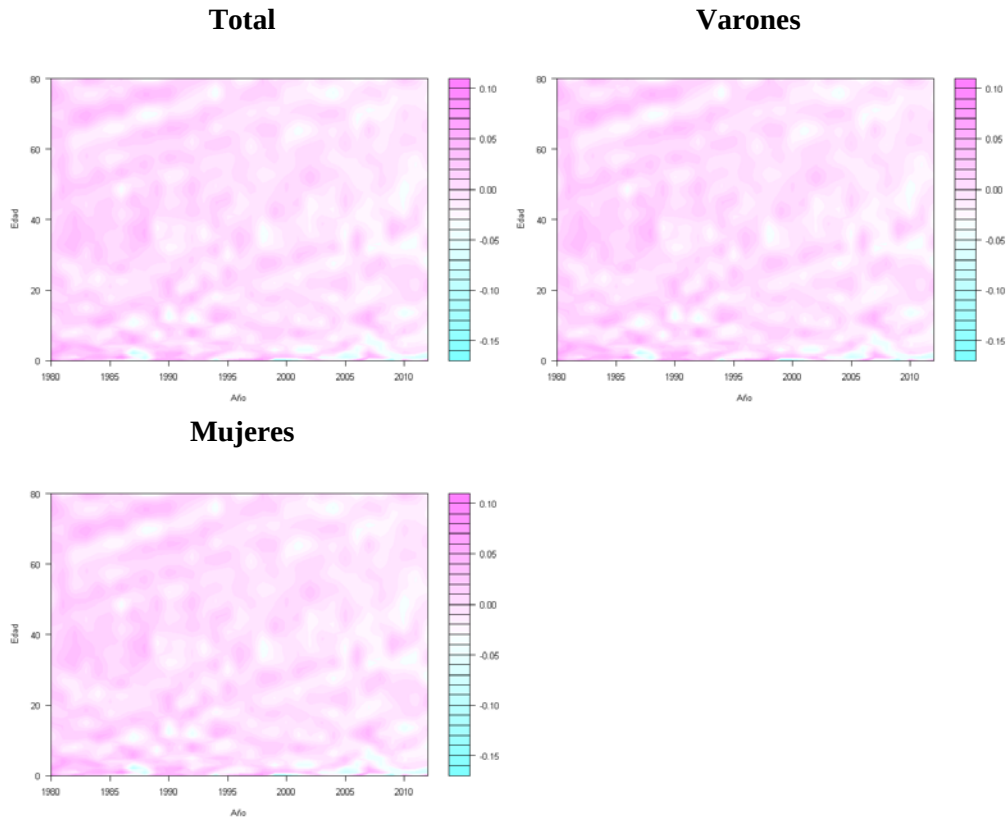
La segunda componente se refiere a la diferencia entre la mortalidad de las mujeres de alrededor de 20 años y el resto de los grupos etarios. El coeficiente asociado mostraría alzas marcadas en el periodo 1985-1990 y 2000-2005.

Para evaluar la bondad del ajuste del modelo propuesto se utiliza un gráfico de contorno para verificar la independencia de los residuos. Bajo independencia se debería esperar zonas pequeñas y mezcladas de colores blanco, rosa y celeste. En los residuos observados se presentan grupos o bandas rosa y blanco, este último indica que los residuos son cercanos a cero, en cambio los primeros reflejarían subestimación. Para las edades cercanas a cero se detectan manchas de color celeste, que son indicio de una leve sobrestimación.

En los tres casos analizados, los residuos son cercanos a cero y alternan valores positivos en su mayoría y negativos en edades cercanas al cero, aunque la alternancia debería ser más marcada y la presencia de colores más equilibrada.

Luego con el fin de pronosticar las tasas de mortalidad con un horizonte de pronóstico fuera de la muestra de diez años (h = 10), se realizan los pronósticos ARIMA de los coeficientes β que intervienen en el modelo de datos funcionales, para estimar los valores futuros de las tasas de mortalidad ($y_{n+h}(x)$).

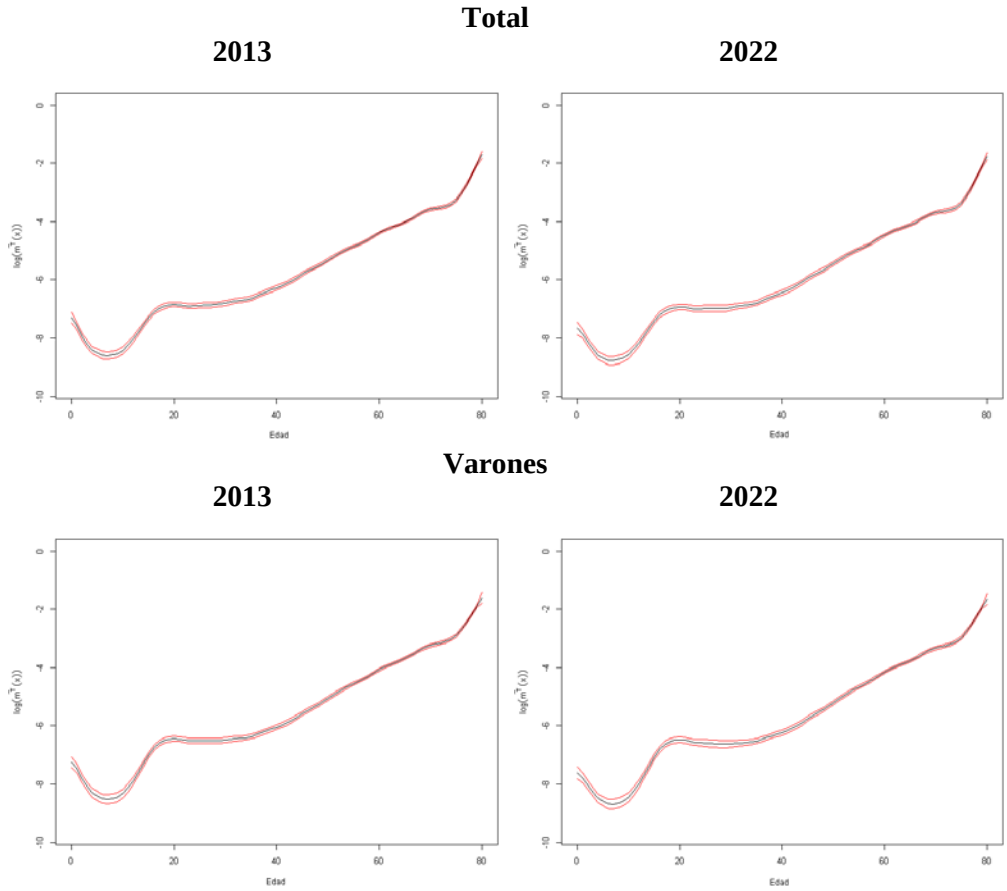
Figura 6. Residuos de los modelos de datos funcionales, para el total, varones y mujeres. Argentina 1980-2012.

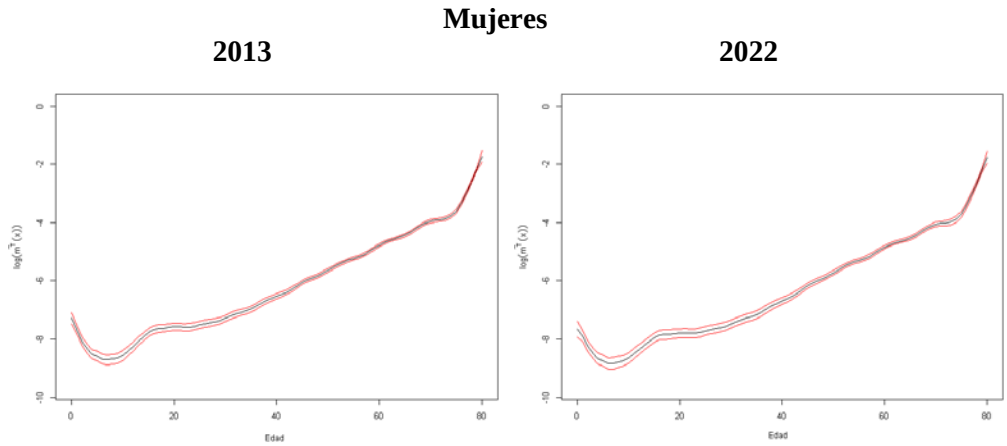


Fuente: Elaboración propia en base a datos de la DEIS e INDEC.

Los pronósticos de las tasas de mortalidad para el total, varones y mujeres, por edad, con sus respectivos intervalos de predicción del año 2013 y 2022 se presentan en la Figura 7. Esto se realiza en forma similar para los años intermedios. En la tabla A (Anexo) se presentan los valores de las tasas de mortalidad para el total, varones y mujeres con sus respectivo intervalos de pronóstico del año 2022. Una característica a destacar es la estrechez de dichos intervalos de pronóstico. Otra característica es que los intervalos correspondientes a las tasas de mortalidad de las mujeres son levemente más amplios que los de los varones.

Figura 7. Pronósticos funcionales e intervalos de pronóstico del 95%, para el total, varones y mujeres. Argentina 2013 y 2023.

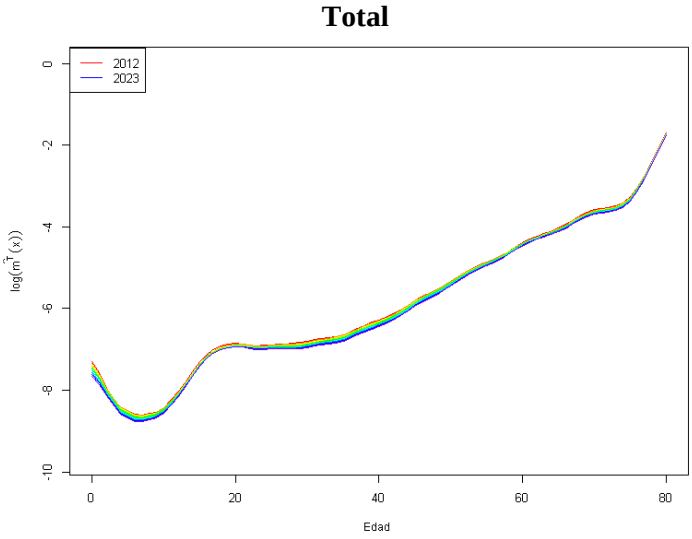




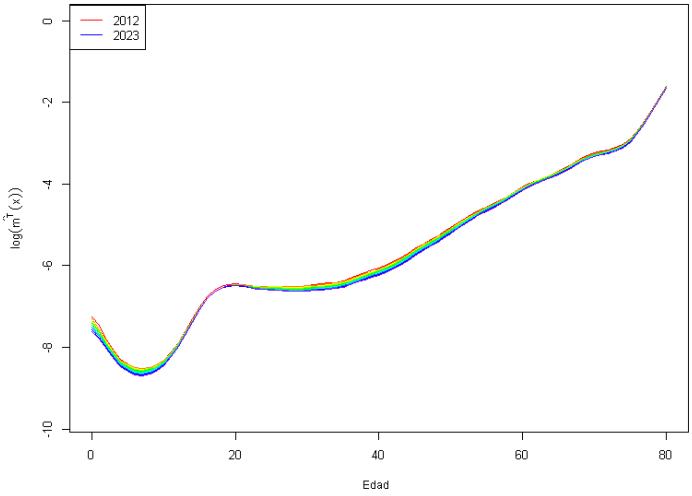
Fuente: Elaboración propia en base a datos de la DEIS e INDEC.

En la figura 8 se representan los logaritmos de las tasas de mortalidad para el total, varones y mujeres, por edad a, través de todos los años pronosticados. La reducción de las tasas de mortalidad a través de los años se presenta especialmente en los primeros años de vida y a partir de los 40 años. La menor reducción se manifiesta alrededor de los 20 años especialmente para las tasas de los varones. Estos fenómenos también fueron advertidos en la descomposición por componentes principales.

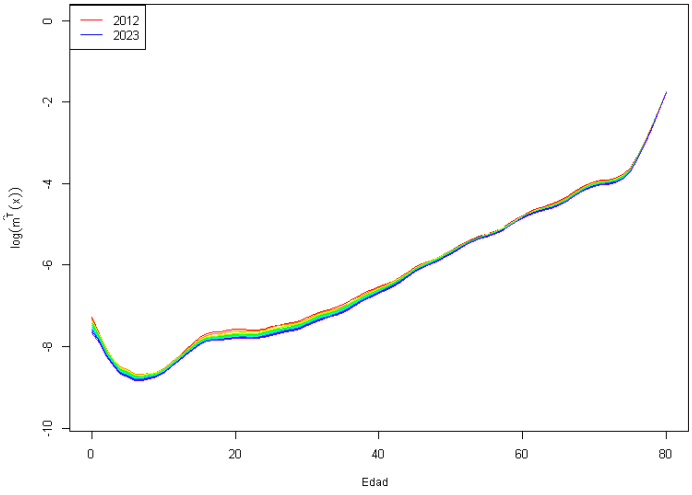
Figura 8. Pronósticos funcionales, para el total, varones y mujeres. Argentina 2013 al 2023.



Varones



Mujeres



Fuente: Elaboración propia en base a datos de la DEIS e INDEC.

Para evaluar la bondad de pronóstico fuera de la muestra los resultados obtenidos mediante Modelos para Datos Funcionales (MDF), se calculan los pronósticos mediante modelos ARIMA univariados para cada edad. Tanto para modelos ARIMA como el enfoque funcional se ajustan los modelos dejando los cuatro últimos valores (2009-2012) para evaluar la calidad de los mismos. Los pronósticos MDF evaluados a través del MAPE son levemente superiores a los obtenidos por

modelos ARIMA, especialmente para el total y los varones, y resultando similares o algo inferiores en el caso de las mujeres. Además, por lo general, los pronósticos demográficos y particularmente los referidos a mortalidad, se necesitan a largo plazo, la comprobación empírica de este aspecto no es posible debido al relativamente corto período para el cual se dispone de información, pero, es de esperar, que los pronósticos obtenidos por MDF brinden información que se comporte mejor a largo plazo dado que reduce la variabilidad mediante el análisis funcional, esto se ve reflejado en la amplitud de los intervalos de pronóstico, siendo estos, en promedio, un 47,9% más estrechos que los obtenidos mediante modelos ARIMA, con una variación entre el 24,7 y 79,9%. Si bien los intervalos para MDF tienen una menor amplitud, en la mayoría de las categorías (sexo-edad) dichos intervalos cubren el valor observado de cada tasa. (Ver Tablas B y C del Anexo).

4. Conclusiones

En este trabajo se realiza una breve descripción del enfoque de datos funcionales para modelar la tasa de mortalidad por edad y sexo, esta propuesta es un avance sobre los tradicionales modelos de Lee y Carter y algunas de sus modificaciones. El nuevo método se aplica a la tasa de mortalidad por edad y sexo de la Argentina.

Una característica de estos nuevos modelos es que permiten interpretar el comportamiento de la mortalidad a través del tiempo puesto en correspondencia con el comportamiento por edades, esto debido especialmente a la utilización de técnicas de componentes principales sobre los datos suavizados de las tasas de mortalidad por edades y sexo. Es importante también, señalar que este método permite construir intervalos de pronóstico con un nivel de incertidumbre aceptable.

Los pronósticos fuera de la muestra (con h=4) para MDF son superiores a los obtenido por modelos ARIMA en el caso de total y varones, y en las mujeres prácticamente coinciden. No obstante la superioridad de los MDF reside en la precisión de los intervalos de pronóstico, cuya amplitud resulta sustancialmente menor a la de los obtenidos por modelos ARIMA. Si además se considera que por lo general los pronósticos demográficos se hacen para horizontes mucho más extensos, es de esperar que los pronósticos e intervalos de pronóstico del enfoque funcional brinden información más precisa sobre valores futuros.

El paso siguiente consiste en aplicar este método a datos funcionales de las tasas de fecundidad y con un tratamiento especial sobre migraciones, para luego calcular pronósticos de la población con sus respectivos intervalos de pronóstico y por lo tanto de otras variables demográficas como por ejemplo la esperanza de vida.

Anexo

Tabla A. Tasas de mortalidad por mil pronosticadas para el 2022 y sus intervalos de pronóstico del 95%. Total, varones y mujeres.

Edad	Total	LI 95%	LS 95%	Varo-nes	LI 95%	LS 95%	Muje-res	LI 95%	LS 95%
0	0,47	0,38	0,58	0,49	0,4	0,6	0,47	0,36	0,62
1	0,39	0,33	0,45	0,41	0,35	0,49	0,37	0,3	0,45
2	0,29	0,26	0,33	0,32	0,28	0,37	0,26	0,22	0,31
3	0,23	0,2	0,26	0,25	0,22	0,29	0,2	0,17	0,25
4	0,19	0,16	0,22	0,21	0,18	0,24	0,17	0,14	0,21
5	0,17	0,15	0,2	0,19	0,16	0,22	0,16	0,13	0,19
6	0,15	0,13	0,18	0,17	0,14	0,2	0,15	0,12	0,18
7	0,15	0,13	0,18	0,17	0,14	0,2	0,15	0,12	0,18
8	0,16	0,14	0,19	0,18	0,15	0,21	0,15	0,13	0,19
9	0,17	0,15	0,2	0,19	0,16	0,22	0,16	0,13	0,19
10	0,19	0,17	0,22	0,22	0,18	0,25	0,18	0,15	0,21
11	0,23	0,21	0,26	0,27	0,23	0,31	0,2	0,17	0,24
12	0,29	0,26	0,32	0,34	0,3	0,39	0,24	0,2	0,28
13	0,36	0,33	0,4	0,45	0,4	0,51	0,27	0,23	0,31
14	0,47	0,43	0,51	0,62	0,56	0,69	0,31	0,26	0,36
15	0,61	0,55	0,67	0,85	0,77	0,94	0,35	0,3	0,41
16	0,74	0,67	0,82	1,1	1	1,22	0,38	0,32	0,45
17	0,84	0,76	0,93	1,31	1,18	1,45	0,39	0,33	0,46
18	0,91	0,82	1	1,44	1,29	1,6	0,39	0,34	0,46
19	0,95	0,86	1,05	1,51	1,37	1,68	0,4	0,34	0,47
20	0,97	0,88	1,07	1,53	1,39	1,7	0,41	0,35	0,48
21	0,96	0,88	1,06	1,51	1,36	1,66	0,41	0,35	0,48
22	0,94	0,85	1,03	1,45	1,31	1,61	0,41	0,35	0,47
23	0,92	0,83	1,01	1,41	1,26	1,57	0,41	0,35	0,47
24	0,91	0,82	1,02	1,38	1,23	1,55	0,42	0,36	0,49
25	0,92	0,82	1,03	1,37	1,21	1,55	0,44	0,38	0,52
26	0,93	0,83	1,04	1,36	1,19	1,54	0,46	0,4	0,54
27	0,92	0,82	1,04	1,34	1,18	1,52	0,48	0,41	0,56

Continuación

Edad	Total	LI 95%	LS 95%	Varo-nes	LI 95%	LS 95%	Muje-res	LI 95%	LS 95%
28	0,92	0,82	1,03	1,32	1,17	1,49	0,49	0,43	0,57
29	0,93	0,84	1,04	1,32	1,17	1,48	0,52	0,45	0,6
30	0,96	0,87	1,06	1,33	1,19	1,49	0,56	0,49	0,65
31	0,99	0,9	1,09	1,35	1,22	1,5	0,61	0,53	0,7
32	1,02	0,93	1,12	1,37	1,24	1,52	0,65	0,57	0,74
33	1,04	0,96	1,14	1,39	1,27	1,53	0,68	0,6	0,78
34	1,07	0,99	1,17	1,42	1,3	1,55	0,72	0,63	0,82
35	1,12	1,03	1,22	1,47	1,35	1,6	0,77	0,68	0,89
36	1,2	1,1	1,3	1,55	1,42	1,69	0,85	0,75	0,97
37	1,3	1,19	1,41	1,65	1,52	1,8	0,95	0,84	1,08
38	1,4	1,29	1,53	1,76	1,61	1,92	1,05	0,93	1,19
39	1,5	1,38	1,64	1,86	1,71	2,03	1,15	1,02	1,29
40	1,59	1,46	1,74	1,97	1,8	2,15	1,24	1,1	1,39
41	1,71	1,57	1,86	2,1	1,92	2,29	1,34	1,19	1,5
42	1,86	1,71	2,02	2,27	2,08	2,48	1,47	1,32	1,64
43	2,07	1,9	2,25	2,51	2,3	2,74	1,65	1,48	1,83
44	2,32	2,14	2,52	2,81	2,59	3,06	1,86	1,69	2,06
45	2,61	2,41	2,82	3,15	2,9	3,42	2,09	1,91	2,29
46	2,89	2,68	3,11	3,52	3,25	3,81	2,3	2,11	2,52
47	3,17	2,95	3,41	3,9	3,61	4,21	2,5	2,3	2,72
48	3,47	3,23	3,72	4,32	4	4,66	2,7	2,49	2,92
49	3,83	3,57	4,1	4,82	4,48	5,18	2,94	2,72	3,17
50	4,28	4	4,57	5,42	5,05	5,82	3,24	3,01	3,5
51	4,82	4,52	5,13	6,14	5,74	6,58	3,62	3,37	3,89
52	5,41	5,09	5,75	6,94	6,5	7,42	4,02	3,74	4,31
53	5,99	5,65	6,35	7,77	7,29	8,28	4,39	4,09	4,7
54	6,53	6,16	6,91	8,57	8,07	9,11	4,69	4,38	5,03
55	7,04	6,66	7,44	9,38	8,85	9,94	4,97	4,63	5,32
56	7,6	7,19	8,03	10,25	9,69	10,84	5,27	4,92	5,66
57	8,29	7,86	8,76	11,28	10,68	11,91	5,7	5,31	6,11
58	9,19	8,72	9,68	12,53	11,88	13,21	6,28	5,87	6,73
59	10,27	9,77	10,79	13,98	13,29	14,71	7,02	6,58	7,49

Continuación

Edad	Total	LI 95%	LS 95%	Varo-nes	LI 95%	LS 95%	Muje-res	LI 95%	LS 95%
60	11,44	10,9	11,99	15,56	14,8	16,36	7,82	7,34	8,32
61	12,55	11,98	13,15	17,12	16,29	17,99	8,55	8,04	9,09
62	13,5	12,9	14,13	18,56	17,67	19,49	9,12	8,59	9,69
63	14,32	13,7	14,98	19,91	18,99	20,88	9,57	9,01	10,15
64	15,16	14,5	15,85	21,32	20,36	22,33	10,01	9,42	10,63
65	16,21	15,49	16,97	23	21,95	24,09	10,64	9,99	11,33
66	17,64	16,84	18,48	25,12	23,98	26,32	11,6	10,87	12,38
67	19,48	18,57	20,42	27,71	26,44	29,03	12,93	12,09	13,83
68	21,53	20,5	22,62	30,57	29,14	32,06	14,49	13,51	15,54
69	23,46	22,26	24,73	33,32	31,69	35,04	15,98	14,83	17,22
70	24,89	23,53	26,34	35,61	33,77	37,54	17,08	15,77	18,5
71	25,72	24,24	27,3	37,3	35,33	39,37	17,66	16,21	19,24
72	26,27	24,67	27,96	38,74	36,68	40,92	17,98	16,4	19,72
73	27,22	25,46	29,1	40,71	38,47	43,08	18,62	16,85	20,57
74	29,5	27,46	31,69	44,27	41,7	47	20,36	18,31	22,63
75	34,25	31,81	36,87	50,71	47,67	53,95	24,24	21,76	26,99
76	43,02	40,06	46,2	61,64	58,06	65,45	31,82	28,73	35,23
77	58,01	54,4	61,86	79,04	74,89	83,42	45,64	41,73	49,92
78	82,11	77,65	86,82	105,08	100,13	110,27	69,54	64,54	74,93
79	118,65	112,99	124,59	141,87	135,84	148,17	108,77	102,43	115,5
<80	171,61	150,86	195,21	191,65	159,84	229,78	170,34	139,79	207,55

Tabla B. MAPE obtenido para 4 valores fuera de la muestra (2009-2012) por MDF y ARIMA.

	MAPE											
	Total				**Varones**				**Mujeres**			
ARIMA	**2009**	**2010**	**2011**	**2012**	**2009**	**2010**	**2011**	**2012**	**2009**	**2010**	**2011**	**2012**
	0,0430	0,0562	0,0607	0,0905	0,0537	0,0777	0,0784	0,1047	0,0748	0,0707	0,0787	0,0949
		09-10	**09-11**	**09-12**		**09-10**	**09-11**	**09-12**		**09-10**	**09-11**	**09-12**
		0,0496	0,0533	0,0626		0,0657	0,0699	0,0786		0,0728	0,0747	0,0798
FDM	**2009**	**2010**	**2011**	**2012**	**2009**	**2010**	**2011**	**2012**	**2009**	**2010**	**2011**	**2012**
	0,0541	0,0570	0,0573	0,0812	0,0601	0,0746	0,0656	0,0789	0,0879	0,0756	0,0883	0,1061
		09-10	**09-11**	**09-12**		**09-10**	**09-11**	**09-12**		**09-10**	**09-11**	**09-12**
		0,0555	0,0561	0,0624		0,0673	0,0667	0,0698		0,0817	0,0839	0,0894

Tabla C. Amplitud promedio de los Intervalos de Pronóstico y cantidad de valores no cubiertos (N), obtenidos para 4 valores fuera de la muestra (2009-2012) por MDF y ARIMA.

	Amplitud promedio											
	Total				**Varones**				**Mujeres**			
	2009	**2010**	**2011**	**2012**	**2009**	**2010**	**2011**	**2012**	**2009**	**2010**	**2011**	**2012**
ARIMA	0,0023	0,0031	0,0037	0,0042	0,0032	0,0040	0,0046	0,0051	0,0019	0,0025	0,0030	0,0033
FDM	0,0016	0,0017	0,0018	0,0008	0,0011	0,0023	0,0024	0,0024	0,0015	0,0015	0,0016	0,0016

Bibliografía

BLACONÁ, M.T y ANDREOZZI, L. (2012). "Comparación de métodos de estimación del modelo de Lee y Carter". *Estadística.* **64** (182 y 183):57-84.

BOOTH, H., HYNDMAN, R., TICKLE, L., y de JONG, P. (2006). "Lee-Carter mortality forecasting: a multi-country comparison of variants and extensions". *Demographic Research.* **15**(9): 289-310.

BOOTH, H., MAINDONALD, J., y SMITH, L. (2002). "Applying Lee-Carter under conditions of variable mortality decline". *Population Studies.* **56**(3):325-336.87

BOOTH, H., TICKLE, L., y SMITH, L. (2005). "Evaluation of the variants of the Lee-Carter method of forecasting mortality: A multi-country comparison". *New Zealand Population Review*. **31**(1):13-34.

BOX, G. y JENKINS, G. (1976). *Time Series Analysis: Forecasting and Control*. Holden-Day.

BRILLINGER, D. (1986). "The natural variability of vital rates and associated statistics". *Biometrics*. **42**:693-734.

BROUHNS, N., DENUIT, M., y VERMUNT, J. (2002). "A Poisson log-bilinear regression approach to the construction of projected life tables". *Insurance*: *Mathematics and Economics*. **1**(3):373-393.

ERBAS, B., HYNDMAN, R., y GERTIG, D. (2007). "Forecasting age-specific breast cancer mortality using functional data models". *Statistics in Medicine*. **2**(26):458-470.

HE, X. y HG, P. (1999). "Cobs: qualitatively constrained smoothing via linear programming". *Computational Statistics*. **14**:315-337.

HYNDMAN, R y BOOTH, H (2008). "Stochastic population forecast using functional data models for mortality, fertility and migration". *International Journal of Forecasting*. **24**:323-342.

HYNDMAN, R.J., KOHELER, A.B., ORD, J.K., SNYDER, R.D. (2008). *Forecasting with exponential smoothing, the State space approach*. Springer.

HYNDMAN, R. y ULLAH, M. (2007). "Robust forecasting of mortality and fertility rates: A functional data approach". *Computational Statistics and Data Analysis*. **51**:4942-4956.

JONG, P. D. y TICKLE, L. (2006). "Extending lee-carter mortality forecasting". Mathematical Population Studies, **13**(1):1-18.

LEE, R. y CARTER, L. (1992). "Modeling and forecasting U. S. mortality". *Journal of the American Statistical Association*. **87**:659-671.

LEE, R. y MILLER, T. (2001). "Evaluating the performance of the Lee-Carter method for forecasting mortality". *Demography*. **38**:537-549.

RAMSAY, J. O. y SILVERMAN, B. (2005). *Functional data analysis* 2nd ed. Springer, NewYork.

RENSHAW, A. y HABERMAN, S. (2003a). "Lee-carter mortality forecasting: A parallel generalized linear modelling approach for England and Wales mortality projections". *Applied Statistics*, **52**(1):119-137.

RENSHAW, A. y HABERMAN, S. (2003b). "Lee-carter mortality forecasting with age-specific enhancement". *Insurance: Mathematics and Economics*. **33**(2):255-272.

RENSHAW, A. y HABERMAN, S. (2003c). "On the forecasting of mortality reduction factors". *Insurance: Mathematics and Economics*,. **32**(3):379-401.

RUPPERT, D., WAND, M., y CARROL, R. (1976). *Semiparametric regression*. Oxford University Press, New York.

SIMONO, J. (1976). *Smoothing methods in statistics*. Springer-Verlag, New York.

WILMOTH, J.R. (1993*).* "Computational Methods for Fitting and Extrapolating the Lee-Carter Model of Mortality change." *Technical Report, Department of Demography*. University of California, Berkeley.

WOOD, S. (2003). "Thin plate regression splines". *Journal of the Royal Statistical Society*, Serie B. **65**(1):95-114.

WOOD, S. (2006). "*Generalized Additive Models: An Introduction with R"*. Chapman &Hall/CRC Texts in Statistical Science. Taylor & Francis.

Invited Paper
Received December 2014
Revised December 2014