Predicción de tasas de mortalidad por cáncer colorrectal y por cáncer en España
Máster en Técnicas Estadísticas.

Lara Andrea Neira González

Tutores:
José Antonio Vilar Fernández.
Salvador Pita Fernández.
Sonia Pértega Díaz.

Índice general

Introducción

Mortalidad por cáncer

La evolución de las tasas de muerte ha despertado un gran interés en las últimas décadas. El aumento histórico en la esperanza de vida muestra un pequeño signo de desacelarización, y el aumento de la supervivencia contribuye significativamente al envejecimiento de la población. En este contexto, la predicción de la mortalidad pasa a tener un papel relevante. El futuro de la mortalidad tiene interés no sólo en este contexto, sino que también en contextos económicos, sociales y de planificación de la salud, debido a la importancia de las pensiones, de los seguros de vida, ... La previsión futura de la salud y la seguridad social en poblaciones envejecidas tiene ahora un interés central en todos los países desarrolados del mundo.

En este estudio se analizan, con distintos procedimientos estadísticos, dos conjuntos de tasas de mortalidad reales. Primero se aborda el problema de predecir las tasas de mortalidad por cáncer en España. A continuación, se aborda la predicción de las tasas de mortalidad por cáncer colorrectal en España, ya que en la última década tomó una fuerza importante, sobre todo en los hombres.

El cáncer colorrectal (CCR) es uno de los de mayor impacto en la población en los países desarrollados. El CCR es una de las enfermedades más comunes, tanto en hombres como en mujeres, y es la segunda causa principal de la muerte por cáncer en USA, Australia y Europa, para ambos sexos.

Aproximadamente el 50 % de los CCR son atribuidos a factores dietéticos. La típica dieta occidental, alta en grasas y baja en fibras, se ha mostrado como un factor que contribuye al desarrollo del cáncer de colon en estudios con animales y epidemiológicos.

En la base de datos de la Organización Mundial de la Salud (WHO), fueron diagnosticados 1.023.000 nuevos casos de cáncer de colon en 2002 y se produjeron por dicha causa 529.000 muertes en todo el mundo. La supervivencia al cáncer de colon mostró una mejoría

durante las décadas recientes variando de un país a otro.

Entre 1990 y 2002, la tasa de mortalidad relativa al cáncer colorrectal disminuyó aproximadamente un 1,8 % por año en EE.UU., principalmente por la detección prematura de dicho cáncer y los avances terapéuticos. Estudios recientes han mostrado un aumento en la incidencia de cáncer de colon y presumiblemente esta tendencia se mantendrá, por el aumento de la esperanza de vida, ya que a mayor edad mayor incidencia. El exceso de riesgo de muerte entre los pacientes europeos con respecto a los pacientes de EE.UU. en el primer año después del diagnóstico fue mucho mayor que en los años siguientes. Este patrón es probable que sea atribuible a las diferencias en la etapa de detección y en la mortalidad postoperatoria. Para mejorar la eficacia del tratamiento y por lo tanto la supervivencia de los pacientes europeos con cáncer colorrectal, en comparación con la de los pacientes EE.UU., se sugiere que los países europeos presten más atención a la detección temprana, en particular mediante la aplicación de los programas de cribado poblacional, según lo recomendado por la Unión Europea.

Las tendencias de las tasas de mortalidad del CCR durante la segunda mitad del siglo XX fueron analizadas para 21 países europeos y se agrupan en 3 grandes regiones europeas. El cáncer tuvo mayor incidencia en los países nórdicos, con excepción de Dinamarca y Europa central, incidencia intermedia en el sur de Europa, salvo en el Reino Unido e Irlanda, y menor en el este de Europa. A partir de la década de los 90 la mortalidad provocada por el CCR mostró un patrón favorable en algunos países de Europa. Sin embargo, en algunos países de la Europa occidental aún se observó un incremento o un estancamiento en las tasas de mortalidad por CCR, por ejemplo en algunos países mediterráneos, entre ellos España. Si las tendencias recientes se mantienen, es probable que la mortalidad por cáncer colorrectal disminuya aún más en Europa en la década actual.

En España, la evolución de la mortalidad por cáncer en la década 1997-2006 indica cambios significativos en los tipos de tumores más frecuentes. En hombres, la mortalidad total por cáncer sufrió durante este período un descenso del 1,3 % anual, debido principalmente a la clara disminución de la mortalidad para cuatro de los cinco tipos de tumores que producen el mayor número de muertes: pulmón, próstata, vejiga y estómago. Para otras localizaciones, se observa la tendencia inversa, es decir, un ligero aumento de la mortalidad, de forma que la interpretación del descenso de la mortalidad total debe realizarse con cautela, ya que la tendencia total resulta de las tendencias de cada uno de los tumores y pueden tener sentidos opuestos.

En el caso del cáncer de colon, la tendencia de la mortalidad en España es similar a la mayoría de los países europeos. A pesar de que el cáncer colorrectal es el segundo tumor más frecuente en hombres, responsable en el año 2006 del 12,4 % de las muertes por cáncer,

las tasas de mortalidad se han mantenido estables o han aumentado ligeramente durante el periodo 1997-2006. En mujeres, la mortalidad total por cáncer descendió un 1 % anual durante el periodo 1997-2006. El descenso de la mortalidad total se atribuye, claramente, al importante descenso de la mortalidad debida a los tumores de mama, de colon y recto y de estómago, que junto con los cánceres de pulmón y páncreas fueron los más letales en mujeres españolas en el año 2006. Al contrario que en los hombres, la mortalidad por cáncer de colon en las mujeres españolas descendió en el periodo 1997-2007 en todas las comunidades autónomas en mayor o menor grado.

La historia natural del cáncer colorrectal conlleva a implementar programas de cribado poblacionales para disminuir la incidencia y la mortalidad de esta enfermedad mediante la detección temprana y la eliminación de cáncer en estadío temprano o pólipos adenomatosos precancerosos. El diagnóstico temprano y los mejores tratamientos son los principales determinantes en la mejoría de la supervivencia en dicho cáncer.

De hecho, las terapias del CCR han mejorado los resultados del paciente significativamente en las últimas décadas. Una de las explicaciones posibles son que tanto la detección precoz como el tratamiento de este tipo de cáncer han mejorado. La mejora depende de los cambios en las actitudes de los pacientes, los comportamientos de los médicos, la cobertura de seguros, la supervisión y los sistemas de recordatorio necesarios para apoyar los programas de cribado.

La U.S. Preventive Services Task Force (USPSTF) encontró una buena evidencia de que varios métodos de detección son eficaces en la reducción de la mortalidad por cáncer colorrectal.

Uno de los esfuerzos también debe ser dirigido a la reducción de la mortalidad postoperatoria con un enfoque multidisciplinario para la atención hospitalaria (médicos de gran volumen de procedimientos y mejor técnica intraoperatoria), ya que se ha demostrado que contribuyen a la mejora en la supervivencia.

Modelos estadísticos para la predicción de tasas de mortalidad

Algunos autores han propuesto modelos estadísticos específicos para predecir las tasas de mortalidad y la esperanza de vida. Muchos de ellos han seguido un trabajo clave de predicciones demográficas de Lee & Carter (1992).

Básicamente, se propone el método de componentes principales para extraer un simple

índice de la variación en el tiempo del nivel de las tasas de mortalidad, cuyas predicciones son obtenidas usando un camino aleatorio con deriva. Este método ha sido usado para predecir las tasas de mortalidad en varios países, incluyendo Australia (Booth et al. 2002, De Jong & Tickle 2006), Canadá (Lee & Nault 1993), Chile (Lee & Rofman 1994), España (Felipe et al. 2002, Debón et al. 2006), Grecia (Lundström & Qvist 2004, Tuljapurkar 2005), los 7 países más desarrollados económicamente (G-7) ... Las características más importantes del método de Lee-Carter son su sensillez y su robustez en situaciones donde las tasas logarítmicas de la mortalidad por edades específicas tienen una tendencia lineal. Aunque paulativamente se fueron desarrollando otros métodos, el método de Lee-Carter es tomado a menudo como punto de referencia.

El principio fundamental del método de Lee-Carter es la extrapolación de tendencias pasadas en las tasas de mortalidad. El método fue diseñado para predecir un periodo largo basándose en una serie de tiempo también larga de datos históricos. Sin embargo, han ocurrido cambios estructurales significativos en los patrones de la mortalidad en el siglo XX, reduciendo la relevancia de los datos pasados más distantes para predicciones a corto y medio plazo. Así que, un problema inevitablemente complejo es determinar el periodo de ajuste apropiado. Por otro lado, la fuerte demanda de datos del método de Lee-Carter puede ser una dificultad adicional. No ha sido evaluado con gran detalle si afecta significativamente la longitud del período de ajuste en la exactitud de los pronósticos.

Realmente, la evaluación no es buena a la hora de predecir a un largo horizonte de predicción. Sin embargo, para saber si afecta la longitud del período de ajuste, los pronósticos se puede evaluan usando un período pequeño de datos históticos y comporarlos con el conjunto de predicciones futuras.

En diferentes trabajos se han propuesto modificaciones del método de Lee-Carter, entre ellos Lee & Miller (2001) y Booth-Maindonald-Smith (2002). La exactitud del pronóstico del método de Lee-Carter y sus variantes fueron primero evaluadas por Booth et al. (2005) y además estudiadas por Booth et al. (2006). Además, han habido importantes extensiones del método de Lee-Carter, quedándose con algo de su esencia pero añadiendo rasgos estadísticos adicionales como la suavización no paramétrica, múltiples componentes principales. Una de las extensiones fue propuesta por Hyndman & Ullah (2007), que ha sido recibido con gran interés en el campo de la demografía y la estadística. Su método combina ideas de suavización no paramétrica, regresión por componentes principales y análisis de datos funcionales, tanto para predecir tasas de mortalidad como tasas de fertilidad.

En este estudio se presentan los resultados de la evalución de 4 métodos de la predicción de mortalidades: Lee-Carter, Lee-Miller, Booth-Maindonald-Smith y Hyndman-Ullah. Se estudia con más detalle éste último, haciendo uso de las tasas de mortalidad del cáncer

colorrectal por sexo y grupos de edad quinquenales en España del período 1980-2008 para predecir las tasas correspondientes al período 2009-2018. Se tomará como período de ajuste el 1980-1998 para predecir las tasas futuras en un periodo de 10 años, al objeto de ser comparadas con las tasas realmente observadas en ese periodo, 1999-2008, y obtener así una estimación del error de predicción. Esto último, se hará con datos reales de muerte por cáncer en España.

Adicionalmente se realizarán propuestas basadas en técnicas no paramétricas autorregresivas funcionales. Las llevaremos a cabo con la base de datos de las tasas de mortalidad por cáncer para ambos sexos en España, y evaluaremos su comportamiento de forma análoga. Notar que, estas propuestas no se tienen en cuenta para predecir las tasas de cáncer colorrectal en España para el período 2009-2018, pues nuestro objetivo es testar si pueden llegar a mejorar las predicciones de los métodos en revisión.

Siguiendo el capítulo 11 de la monografía de Ferraty y Vieu (2006), se propone un camino puramente funcional considerando un modelo no paramétrico de autoregresión funcional con respuesta escalar. Como en los trabajos de Hyndman, (Hyndman y Ullah, 2007, Hyndman y Shang, 2009, ...) se considera que los logaritmos de las tasas de mortalidad de cada año son una función continua de las edades. Pero este enfoque no requiere necesariamente un preprocesado de los datos, y el logaritmo de las tasas se estima mediante un modelo de regresión no paramétrico con variable explicativa funcioanal.

Se propone también, un modelo recursivo de regresión autorregresivo funcional con respuesta escalar. Sin aumentar el tamaño de la muestra se estiman las curvas reactualizando la variable independiente de la regresión a medida que se predice a un horizonte mayor.

Y por último, se explora un modelo de regresión autorregresivo funcional con respuesta funcional. Se ajusta mediante el estimador de Nadaraya-Watson y se seleccina la ventana de manera similar al método de validación cruzada, Antoniadis et al. (2009).

Capítulo 1

Revisión de algunos modelos estadísticos de predicción de tasas de mortalidad.

1.1. El método de Lee-Carter.

El método de Lee-Carter para predecir la mortalidad es un modelo estadístico-demográfico que permite realizar proyecciones de tasas futuras de mortalidad, combinando un modelo demográfico de mortalidad con métodos de series temporales para la predicción. En términos generales, se puede definir como un modelo de carácter extrapolativo, ya que no incorpora información acerca de efectos tecnológicos o sociales en la mortalidad, esto es, no busca incorporar información externa, ni opiniones sobre posibles acontecimientos futuros. Más bien, predice en base a la tendencia histórica observada durante un período de estudio o periodo de ajuste en el que se han registrado los datos (Lee, 2000). El método es generalmente interpretado haciendo uso de las series de tiempo más largas de los datos históricos.

La premisa básica del modelo es que existe una relación lineal entre el logaritmo de las tasas específicas de mortalidad $m_{x,t}$ y dos factores explicativos: la edad, x, y el tiempo, t. El modelo de Lee-Carter de la mortalidad es:

$$\ln m_{x,t} = a_x + b_x k_t + \epsilon_{x,t}$$

donde

- $m_{x,t}$ es la tasa de muerte a la edad x en el año t.

- k_t es un índice (no observado) de la intensidad de la mortalidad a lo largo del tiempo t. En este modelo k_t lineal.

- a_x es el patrón promedio de mortalidad para las edades a través de los años, describe el perfil general a lo largo de la edad del esquema de mortalidad.

- b_x es la velocidad relativa de cambio a cada edad. La evaluación de b_x da idea de lo rápidamente que decrecen los ratios en respuesta a cambios en k_t

$$\frac{d \ln m_{x,t}}{dt} = b_x \frac{dk_t}{dt}.$$

- $\epsilon_{x,t}$ es el residuo a la edad x y tiempo t, con $\mathbb{E}(\epsilon_{x,t}) = 0$ y $Var(\epsilon_{x,t}) = \sigma_\epsilon^2$, y refleja influencias históricas no capturadas por el modelo.

A partir de la anterior ecuación del modelo de Lee-Carter, se observa que hay una ecuación para cada edad y cada tiempo, lo cual resulta en un sistema de ecuaciones simultáneas que se tiene que resolver para encontrar los valores $\{a_x\}$, $\{b_x\}$ y $\{k_t\}$. Este sistema se puede escribir matricialmente como

$$M = A + b * k.$$

En principio, como la estructura del modelo es invariante bajo las siguientes transformaciones:

$$\{a_x, b_x, k_t\} \mapsto \{a_x, \frac{b_x}{c}, ck_t\},$$

$$\{a_x, b_x, k_t\} \mapsto \{a_x - cb_x, b_x, k_t + c\},$$

no existe solución única para este sistema.

Para obtener una solución única, se imponen las dos siguientes restricciones:

$$\sum_{t=1}^{n} k_t = 0, \qquad \sum_{x=x_1}^{x_p} b_x = 1,$$

donde n es el número de años y p es el número de edades en las observaciones del conjunto de datos en estudio.

Así a_x es calculado como la media de $\ln m_{x,t}$ sobre el tiempo y b_x y k_t se determinan de forma única. Más especificamente, los coeficientes a_x se determinan mediante la siguiente expresión

$$a_x = \frac{\sum_{t=1}^{n} \ln(m_{x,t})}{n}.$$

Una vez determinados los valores de la matriz A, este sistema se puede reescribir como

$$M' = M - A = b * k.$$

Con las restricciones ya incluidas, este sistema cuenta entonces con una solución única y sólo resta determinar los vectores b y k.

Lee-Carter estima los vectores b y k mediante el método de descomposición en valores singulares (DVS), que proporciona un ajuste exacto de mínimos cuadrados, con lo cual se obtiene una primera estimación de los parámetros del modelo. Si se deseara aumentar la precisión en el ajuste del modelo, se requeriría incluir variables adicionales, lo cual equivale a trabajar con componentes principales (Bell, 1997).

Como el procedimiento de estimación puede producir errores importantes en las predicciones, Lee-Carter propone un segundo paso deseando refinar la estimación. Específicamente, se aprovechan las estimaciones iniciales de los parámetros de a_x y b_x para producir una nueva estimación de k_t, de tal manera que, para una distribución de población específica, se produzca exactamente el número observado de muertes totales para el año en cuestión. Así, el método de Lee-Carter ajusta k_t por un reajuste del número total de muertes observadas. Esto es, se trata de encontrar los valores k_t tales que cumplan con la condición

$$D_t = \sum_{x=x_1}^{x_p} N_{x,t} exp(a_x + b_x k_t + e_{x,t}),$$

donde D_t denota el número de muertes totales en el año t, $N_{x,t}$ es el tamaño de la población con edad x en el año t y x_1 y x_p son las edades del primer y último grupo de edad, respectivamente. Al utilizar este procedimiento, es posible ajustar el modelo para que reproduzca los datos observados de las defunciones totales en cada periodo. El ajuste da pesos grandes a las edades cuyas tasas de muerte son altas, contrarrestando parcialmente el efecto de usar el logaritmo de las tasas en el modelo de Lee-Carter.

Una vez ajustado el modelo, se emplean modelos de series de tiempo ARIMA para elaborar las proyecciones. Lee-Carter propone un modelo de camino aleatorio con deriva que puede ser expresado como

$$k_t = k_{t-1} + d + e_t,$$

donde d es conocido como el parámetro de deriva y mide el promedio anual de cambio en la serie k_t, y e_t es un término de error aleatoriamente distribuido con $\mathbb{E}(e_t) = 0$ y $Var(e_t) = \sigma_e^2$, siendo estos errores mutuamente independientes.

Esto también se puede escribir como

$$\bigtriangledown k_t = d + e_t \qquad \text{con} \qquad \bigtriangledown k_t = k_t - k_{t-1}$$

Debido a los supuestos sobre los errores, las diferencias $\bigtriangledown k_t$ son independientes asumiendo normalidad, con media 0 y varianza σ_e^2.

La predicción de las tasas de muerte para edades específicas se obtienen usando extrapolación de k_t y fijando los a_x y b_x estimados.

Lee y Carter decidieron utilizar el modelo de camino aleatorio con deriva, que dió resultados satisfactorios en la mayoría de los casos, ya que no encontraron ventajas significativas al utilizar modelo más complejos. Más aún, Tuljapurkar et al. (2000) encontraron que la tasa de declive en la mortalidad fue constante para los países desarrollados, reforzando el uso del camino aleatorio con deriva como parte integral del método de Lee-Carter.

El método de Lee-Carter es un método paramétrico y poco flexible. El ajuste del modelo de Lee-Carter, ajustando k_t mediante la distribución por edad del total de muertes observadas tiene el incoveniente de que no es óptima. Bell (1997) mostró empíricamente el sesgo en la esperanza de vida que se produce en los pronósticos si se utiliza el modelo de Lee-Carter.

La principal crítica al modelo de Lee-Carter es que los parámetros a_x y b_x dependen sólo de la edad y que la predicción de futuros valores de la mortalidad se basa sólo en k_t, lo que supone admitir que no existe interacción entre la edad y el tiempo.

Su principal ventaja es la fácil interpretación de sus parámetros. El modelo goza actualmente de mucha popularidad debido a sus buenos resultados y a su simplicidad, por lo que hay una amplia literatura de su tratamiento y mejora.

1.2. El método de Lee-Miller

El método de Lee-Miller es una variante del método de Lee-Carter, del cual difiere en tres cuestiones importantes:

1. El periodo de ajuste es más reducido.

2. El ajuste de k_t consiste en hacer las segunda estimación respecto a la esperanza de vida al nacer (e(0)), en lugar de ajustar respecto al total de defunciones.

3. En el proceso de ajuste, las tasas estimadas para le último año del período de ajuste (año de cambio o de salto), son reemplzadas por las tasas reales observadas en ese año.

La expectativa de vida al nacer es una medida sintética de las tasas de mortalidad por edad que registra una población en un determinado momento en el tiempo. El pronóstico de e(0) se traduce en un conjunto de tasas de mortalidad por edad basadas en transformaciones del tipo Lee-Carter en las que se define un nivel básico de tasas de mortalidad por edad a_x y una segunda serie de factores de transformación b_x que se suman a los factores a_x en múltiplos de k_t hasta alcanzar el nivel e(0) deseado.

Se tiene así una familia de un solo parámetro de curvas de tasas de mortalidad por edad, definidas por el parámetro único k_t y por las curvas fijas según edad a_x y b_x

En su evaluación del método de Lee-Carter, Lee and Miller (2001) notaron que la predicción para los datos de los Estados Unidos fue sesgada usando el período de ajuste 1900-1989 para predecir el período 1990-1997. Encontró un desajuste entre las tasas ajustadas del último año de el período de ajuste 1989 y las tasas reales en ese año, este error de salto o sesgo aumentó en 0.6 años la esperanza de vida para hombres y para mujeres combinado. Este error de salto o sesgo fue eliminado construyendo el modelo tal que los k_t pasan por cero en los años de salto, usando así las tasas actuales en las años de salto.

El ajuste de k_t por el ajuste de la esperanza de vida e(0) fue adoptado para evitar el uso de datos funcionales requeridos para ajustar D_t (Lee and Miller, 2001).

1.3. La variante de Booth-Maindonald-Smith

La variante de Booth-Maindonal-Smith difiere del método de Lee-Carter en 3 direcciones:

1. El periodo de ajuste se determina basándose en criterios estadísticos de bondad de ajuste bajo la suposición de que k_t es linear.
2. El ajuste de k_t se desarrolla ajustando la distribución de edad de muertes.
3. Las tasas de salto son tomadas como las tasas de ajuste basadas en esta metodología de ajuste.

Booth. et al. (2002) ajustan el modelo de Lee-Carter para datos australianos de 1907-1999 y encuentran que el patrón asumido de que la mortalidad decae de modo constante, como representa el k_t lineal, no se mantiene durante el periodo de ajuste.

Más aún, encuentran problemas también en la suposición de que b_x es constante en el método fundamental de Lee-Carter.

Asumiendo la linealidad de k_t como punto de partida, la variante de Booth-Maindonald-Smith busca maximizar el grado de ajuste del modelo global, restringiendoel período de ajuste a aquel período donde se maximiza el ajuste lineal. La primera consecuencia es que la hipótesis de b_x constante es mejor refrendada.

La determinación del periodo óptimo de ajuste se realiza tomando como base la hipótesis de linealidad de k_t. Se computa primero la desviación media (DMI) del ajuste del modelo de Lee-Carter sobre todo el período observado. A continuación se trabaja con períodos más cortos, retrasando en cada paso el año de comienzo pero manteniendo siempre fijo el último año (año de salto). Para cada uno de estos períodos se ajusta de nuevo el modelo de Lee-Carter y se computa la desviación media del nuevo ajuste (DMP_i). Finalmente, el período seleccionado es aquel que minimiza el cociente $\frac{DMP_i}{DMT}$. Es significativo reseñar que el proceso puede ser computacionalmente muy intensivo si el período temporal de observación es largo.

También se modifica el procedimiento para ajustar k_t. Mejor que el ajuste total de muertes, D_t, la variante de Booth-Maindonald-Smith ajusta la distribución de muertes por edad, $D_{x,t}$, usando la distribución de Poisson para modelar el proceso de muerte y el estadístico deviance para medir la bondad de ajuste (Botth et al., 2002). Las tasas ajustadas bajo este ajuste son consideradas como tasas de salto.

En definitiva, Booth-Maindonald-Smith (2002) usan un criterio de bondad de ajuste estadístico para elegir el período de datos a considerar para estimar el modelo bilineal; y los índices temporales se ajustan para corresponder mejor a toda la distribución de los fallecimientos por edades, en vez de al número total o a la esperanza de vida.

Una ventaja de usar esta variante del modelo de Lee-Carter es que se utiliza un criterio estándar de minimización. Otra ventaja es que este método corrige principalmente en los grupos de edad donde hay una mayor discrepancia enter muertes observadas y ajustadas para cada año, ya que Lee-Carter ajusta sobre el total de muertes en cada año sin distinguir entre grupos de edad.

1.4. El método de Hyndman-Ullah

La aproximación de Hyndman y Ullah (2007) usa el paradigma de datos funcionales (Ramsay and Silverman, 2005) para modelar tasas de muerte logarítmicas. Se proponen métodos no paramétricos para ajustar un modelo conveniente que permita y predecir las tasas logarítmicas de la mortalidad. El procedimiento de Hyndman-Ullah extiende el método de Lee-Carte en las siguientes direcciones:

1. La mortalidad se asume como una función suave de la edad, que es observada con error. Bajo esta suposición las tasas de mortalidad son estimadas usando métodos de suavización no paramétricos.

2. Se usan más de un conjunto de componentes (k_t, b_x).

3. Se usan más métodos generales de series de tiempo que el camino aleatorio con deriva para predecir los coeficientes.

4. Es factible llevar a cabo una estimación robusta para permitir años atípicos debido a guerras o epidemias.

5. No se ajusta k_t.

El procedimiento puede ser esquematizado a través de los siguientes pasos:

1. Las tasas logarítmicas anuales de la mortalidad son primero suavizadas usando una regresión spline penalizada con la restricción de monotonía parcial, es decir, a partir de una cierta edad x las tasas son monótonas.

   Se asume que hay una función fundamental continua y suave $f_t(x)$ que es obserservada con error en las diferentes edades x. Se hace hincapié en que la edad x es ahora considerada como una variable continua. En lo que sigue $m_t(x)$ denotará la tasa logarítmica para la edad $x \in [x_1, x_p]$ en el año t. De este modo disponemos de un conjunto de observaciones $\{x_i, y_t(x_i)\}$ satisfaciendo:

   $$m_t(x_i) = f_t(x_i) + \sigma_t(x_i)\epsilon_{t,i}, \qquad i = 1, \ldots, p; \quad t = 1, \ldots, n, \tag{1.1}$$

   donde $m_t(x_i)$ denota el logaritmo de las tasa de mortalidad observada para la edad x_i en el año t; $\sigma_t(x_i)$ es la desviación estandar del error observacional, que mide el grado de dispersión del ruído que puede ahora variar con la edad x_i en el año t (flexibilizando así la suposición de error homocedástico del modelo de Lee-Carter); y $\epsilon_{t,i}$ es una variable aleatoria independiente e identicamente distribuida a una normal estándar.

2. El conjunto de curvas suavizadas son descompuestas en componentes principales funcionales ortogonales, de modo que las puntuaciones de las componentes principales son incorreladas. Esto es,

   $$f_t(x) = a(x) + \sum_{j=1}^{J} b_j(x)k_{t,j} + e_t(x), \tag{1.2}$$

   donde:

- $a(x)$ es la función media, representando el promedio de mortalidad de las edades a través de los años, que puede ser estimada por $\hat{a}(x) = \frac{1}{n}\sum_{t=1}^{n} f_t(x)$;
- $\{b_1(x), \ldots, b_J(x)\}$ denotan las funciones básicas ortonormales, i.e., conforman el conjunto de las primeras J componentes principales funcionales, con $J < n$;
- $\{k_{t,1}, \ldots, k_{t,J}\}$ es un conjunto de puntuaciones incorreladas de las componentes principales seleccionadas, tales que $\sum_{j=1}^{J} k_{t,j}^2 < \infty$ y
- $e_t(x) \sim N(0, \sigma_x^2)$ denota la función residual, de media cero.

Al usar $a(x)$ en vez de a_x obtenemos mayor versatilidad por tratarse de una función suave de la edad que ahora es una cantidad continua. Se estima aplicando la regresión por splines penalizados (Wood, 2000) para los datos de cada año y promediando los resultados.

Los pares $(k_{t,j}, b_j(x))$ para $j = 1, \ldots, J$, como ya se ha mecionado, se aproximan usando la descomposición en componentes principales.

El término error $\sigma_t(x)\epsilon_{x,t}$ mide el error observacional que varía con la edad; i.e., es aproximado mediante la diferencia entre tasas observadas y curvas spline. El término error $e_t(x)$ denota el error asociado al modelo; i.e., se aproxima mediante la diferencia entre las curvas splines y el modelo ajustado.

Tras el cómputo de las componentes pricipales, se obtenien las estimaciones de f_t, $t = 1, \ldots, n$, dadas por

$$\hat{f}_t(x) = \hat{a}(x) + \sum_{j=1}^{J} \hat{b}_j(x)\hat{k}_{t,j} + \hat{e}_t(x) \tag{1.3}$$

3. Para un mismo año t, $k_{t,j}$ y $k_{t,l}$ son incorrelados para $j \neq l$. Sin embargo, fijado j, la secuencia $\{\hat{k}_{1,j}, \hat{k}_{2,j}, \ldots, \hat{k}_{n,j}\}$ puede verse como una realización de tamaño n (años) de una serie temporal $\{k_{t,j}\}_t$. Se puede usar un amplio rango de modelos de series de tiempo univariantes para obtener predicciones de estos coeficientes para h años más adelante, $k_{t+h,l}$.

   Condicionando a los datos observados $\mathbf{I} = \{m_t(x), t = 1, \ldots, n\}$ y al conjunto de las componentes principales funcionales $\mathbf{B} = \{b_1, \ldots, b_J\}$, las h predicciones siguientes de $m_{n+h}(x)$ pueden obtenerse por:

$$\hat{m}_{n+h|n}(x) = \mathbb{E}[m_{n+h}(x)|\mathbf{I}, \mathbf{B}] = \hat{a}(x) + \sum_{j=1}^{J} b_j(x)\hat{k}_{n+h|n,j}. \tag{1.4}$$

   donde $\hat{k}_{n+h|n,j}$ denota la predicción h pasos adelante de $k_{n+h,j}$ usando un modelo univariante de series de tiempo, tal como un modelo ARIMA, (Box et al. 2008)

La varianza estimada de la predicción es:

$$\zeta_{n+h|n}(x) = \text{Var}\left[y_{n+h}(x)|\mathbf{I},\mathbf{B}\right] \approx \hat{\sigma}_a^2(x) + \sum_{j=1}^{J} u_{n+h|n,j}\hat{b}_j^2(x) + \hat{v}(x) + \hat{\sigma}_{n+h|n}^2(x) \quad (1.5)$$

que aglutina las varianzas de las 4 fuentes de error subyacentes al modelo estimado:

- La varianza del error de la suavización, i.e. $\hat{\sigma}_a^2(x) = \text{Var} = (\hat{a}(x))$, que depende del método de suavizado empleado para obtener $\hat{f}_t(x)$, $t = 1, \ldots, n$.
- La varianza del error propio de la predicción de los coeficientes $\hat{k}_{n+h,j}$, $j = 1, \ldots, J$, i.e. $u_{n+h|n,j} = \text{Var}\left[\hat{k}_{n+h,j}|\{\hat{k}_{1,j}, \ldots, \hat{k}_{n,j}\}\right]$, que depende del tipo de ajuste de la serie temporal $\left\{\hat{k}_{t,j}\right\}_{t=1}^{n}$.
- La varianza de los errores residuales generados al modelizar $f_t(x)$ usando un número finito de componentes principales, i.e. $\hat{v}(x)$. De (1.3) se sigue que:

$$\hat{v}(x) = \frac{1}{n}\sum_{t=1}^{n}\hat{e}_t^2 = \frac{1}{n}\sum_{t=1}^{n}\left(\hat{f}_t(x) - \hat{a}(x) - \sum_{j=1}^{J}\hat{k}_{t,j}\hat{b}_j(x)\right)^2 \quad (1.6)$$

- $\hat{\sigma}_{n+h|n}^2(x)$, la varianza del "error" inherente a la aleatoriedad de las tasas de mortalidad observadas (por ejemplo, con distribuciones Poisson o normal).

Por tanto, los intervalos de predicción al $100(1-\alpha)\,\%$ asumiendo normalidad:

$$\hat{m}_{n+h|n}(x) \pm z_\alpha\sqrt{\zeta_{n+h|n}(x)} \quad (1.7)$$

Así que, Hyndman y Ullah (2006) procedieron a suavizar las edades en sus estimaciones; permitieron que hubiese más de una componente principal (como Bell y sus colaboradores habían hecho antes) y usaron modelos más generales basados en series temporales para las series temporales multivariantes resultantes. Por este motivo no fue necesario ajustar los índices temporales. En comparación a Booth et al. el período de observación también fue acortado, aunque ésta no es una característica necesaria del método propuesto.

Por lo que, este enfoque combina ideas de análisis de datos funcionales, estimación no paramétrica y modelos robustos (permite RAPCA, Análisis de componentes principales robusto).

Otras variantes del método de Hyndman y Ullah son:

- Variante ponderada

- De Jong-Tickle LC (smooth) method

Sin embargo, no son incluídas en la presente revisión.

1.5. Explorando propuestas basadas en técnicas no paramétricas autorregresivas funcionales.

1.5.1. Modelos no paramétricos autorregresivos funcionales con respuesta escalar.

1. Como en los trabajos de Hyndman (Hyndman y Ullah, 2007, Hyndman y Shang, 2009, ...) se considera que los logarítmos de las tasas de mortalidad de cada año son una función continua de las edades, digamos $\chi_t \equiv \{\chi_t(x), x_1 \leq x \leq x_p\}$ para el año t.

   Ahora, se opta por un enfoque diferente del problema que no requiere necesariamente un preprocesado inicial de los datos. Específicamente, se dispone de un conjunto de n datos funcionales $\{\chi_1, \chi_2, \ldots, \chi_n\}$, realizaciones dependientes de una variable aleatoria funcional χ definida sobre un espacio $\mathcal{E}$ con una semimétrica d). En la práctica, no se observa $\chi_t \equiv \{\chi_t(x), x_1 \leq x \leq x_p\}$ sino una versión discretizada de la misma de longitud p:

   $$\chi_t \equiv \{\chi_t(x_1), \chi_t(x_2), \ldots, \chi_t(x_p)\}, \text{ con } \chi_t(x_i) = m_t(x_i), \ i = 1, \ldots, p. \tag{1.8}$$

   Nótese que ahora ni siquiera es preciso que las curvas $\chi_t(x)$ sean suaves, en función de su mayor o menor grado de suavidad se manejará la semimétrica más apropiada.

2. En este contexto, el logarítmo de la tasa de mortalidad en el año $t+h$ a la edad x_i, $m_{t+h}(x_i)$, no se modeliza como en (1.1), sino que se explica mediante un modelo de regresión no paramétrico con variable explicativa funcional χ_t, i.e., fijada una edad x_i, $1 \leq i \leq p$, se postula el modelo:

   $$m_{t+h}(x_i) = r_{h,i}(\chi_t) + \sigma_{h,i}(\chi_t)\,\varepsilon_{h,i,t}, \ \text{ para } t = 1, \ldots, n, \tag{1.9}$$

   donde $r_{h,i}(\chi) = E(m_{t+h}(x_i)|\chi)$ es un operador funcional suave de $\mathcal{E}$ en $\mathbb{R}$ y los errores $\varepsilon_{h,i,t}$ son idénticamente distribuidos y tales que $E(\varepsilon_{h,i,t}|\,\chi_t) = 0$ y $E(\varepsilon^2_{h,i,t}|\chi_t) = \sigma^2_{h,i}(\chi_t)$.

   La ecuación (1.9) define un modelo funcional no paramétrico de autoregresión de orden 1 y retardo h que permite predecir el logarítmo de la tasa de mortalidad en el

año $t+h$ a la edad x_i teniendo en cuenta la información conjunta, a lo largo de todas las edades, de los logarítmos de las tasas de mortalidad h años antes, χ_t.

Para cada edad x_i, $\sigma^2_{h,i}(\chi_t)$ mide la varianza de la predicción condicionada a las tasas observadas h años antes. Esta varianza puede variar para las distintas edades porque para cada edad tenemos una función de regresión a priori distinta.

3. De acuerdo con (1.9), para aproximar la predicción del logarítmo de la tasa de mortalidad a la edad x_i en un horizonte h años después de la última tasa observada, $m_{n+h}(x_i)$, podemos considerar la muestra de tamaño $n-h$:

$$\{(\chi_1, m_{1+h}(x_i)), (\chi_2, m_{2+h}(x_i)), \ldots, (\chi_{n-h}, m_n(x_i))\} \tag{1.10}$$

y emplearla para obtener un estimador no paramétrico $\hat{r}_{h,i}$ del operador de regresión $r_{h,i}$.

4. Estimadores funcionales de tipo núcleo:

   - **Sin selector de la ventana.**

$$\hat{r}_{h,i}(\chi) = \frac{\sum_{t=1}^{n-h} m_{t+h}(x_i) K\left(d(\chi_t, \chi)/c\right)}{\sum_{t=1}^{n-h} K\left(d(\chi_t, \chi)/c\right)}$$

   - **Con selector automático por validación cuadrada**.

$$\hat{r}_{h,i,CV}(\chi) = \frac{\sum_{t=1}^{n-h} m_{t+h}(x_i) K\left(d(\chi_t, \chi)/c_{CV}\right)}{\sum_{t=1}^{n-h} K\left(d(\chi_t, \chi)/c_{CV}\right)}$$

   - **Tipo kNN[1] con un número fijo de vecinos**.

$$\hat{r}^{kNN}_{h,i}(\chi) = \frac{\sum_{t=1}^{n-h} m_{t+h}(x_i) K\left(d(\chi_t, \chi)/c_k(\chi)\right)}{\sum_{t=1}^{n-h} K\left(d(\chi_t, \chi)/c_k(\chi)\right)}$$

   - **Tipo kNN con elección global del número de vecinos**.

$$\hat{r}^{kNN}_{h,i,GCV}(\chi) = \frac{\sum_{t=1}^{n-h} m_{t+h}(x_i) K\left(d(\chi_t, \chi)/c_{k_{opt}}(\chi)\right)}{\sum_{t=1}^{n-h} K\left(d(\chi_t, \chi)/c_{k_{opt}}(\chi)\right)}$$

   - **Tipo kNN con elección local del número de vecinos**.

$$\hat{r}^{kNN}_{h,i,LCV}(\chi) = \frac{\sum_{t=1}^{n-h} m_{t+h}(x_i) K\left(d(\chi_t, \chi)/c_{k_{opt}(\chi_0)}\right)}{\sum_{t=1}^{n-h} K\left(d(\chi_t, \chi)/c_{k_{opt}(\chi_0)}\right)}$$

[1] k-ésimo vecino más cercano

5. Los métodos de estimación anteriores son consistentes bajo ciertas condiciones (de tipo mixing) sobre la estructura de dependencia (Ferraty y Vieu (2006), Capítulo 11) de dependencia.

Se propone un método recursivo. Teniendo en cuenta el modelo autorregresivo dada en (9), i.e.,

$$\hat{m}_{n+h}(x_i) = \hat{r}_{h,i}\left(\chi_n\right),$$

ahora se considera que el estimador $\hat{r}_{h,i}$ se basa en una muestra reactualizada, sin aumentar el tamaño de la misma.

Se modifican las curvas como sigue para estimar el resultado en las edades del año $(n+1)$:

Estimación en edad i:	Usar la muestra (siempre de igual tamaño $(n-1)$):
$\hat{m}_{n+1}(x_1) = \hat{r}_{h=1,1}\left(\chi_n^{(1)}\right)$	$\left\{\left(\chi_1^{(1)}, m_2(x_1)\right), \ldots, \left(\chi_{n-1}^{(1)}, m_n(x_1)\right)\right\}$ con $\chi_t^{(1)} = \chi_t = (\chi_t(x_1), \ldots, \chi_t(x_p))$, $t = 1, \ldots, n$.
$\hat{m}_{n+1}(x_2) = \hat{r}_{h=1,2}\left(\chi_n^{(2)}\right)$	$\left\{\left(\chi_1^{(2)}, m_2(x_2)\right), \ldots, \left(\chi_{n-1}^{(2)}, m_n(x_2)\right)\right\}$ con $\begin{cases} \chi_t^{(2)} = (\chi_t(x_2), \ldots, \chi_t(x_p), \chi_{t+1}(x_1)), \ t = 1, \ldots, n-1 \\ \chi_n^{(2)} = (\chi_n(x_2), \ldots, \chi_n(x_p), \hat{m}_{n+1}(x_1)) \end{cases}$
$\hat{m}_{n+1}(x_3) = \hat{r}_{h=1,2}\left(\chi_n^{(3)}\right)$	$\left\{\left(\chi_1^{(3)}, m_2(x_3)\right), \ldots, \left(\chi_{n-1}^{(3)}, m_n(x_3)\right)\right\}$ con $\begin{cases} \chi_t^{(3)} = (\chi_t(x_3), \ldots, \chi_{t+1}(x_1), \chi_{t+1}(x_2)), \ t = 1, \ldots, n-1 \\ \chi_n^{(3)} = (\chi_n(x_3), \ldots, \hat{m}_{n+1}(x_1), \hat{m}_{n+1}(x_2)) \end{cases}$
$\vdots$	$\vdots$
$\hat{m}_{n+1}(x_p) = \hat{r}_{h=1,2}\left(\chi_n^{(p)}\right)$	$\left\{\left(\chi_1^{(p)}, y_2(x_p)\right), \ldots, \left(\chi_{n-1}^{(p)}, y_n(x_p)\right)\right\}$ con $\begin{cases} \chi_t^{(p)} = (\chi_t(x_p), \chi_{t+1}(x_1), \ldots, \chi_{t+1}(x_{p-1})), \ t = 1, \ldots, n-1 \\ \chi_n^{(p)} = (\chi_n(x_p), \hat{y}_{n+1}(x_1), \ldots, \hat{y}_{n+1}(x_{p-1})) \end{cases}$

1.5.2. Modelos no paramétricos autorregresivos funcionales con respuesta funcional.

La curva χ_t es observada en cada edad x, en p puntos equidistantes, donde $\chi_t(x_i)$, $i = 1, \ldots, p$. son las correspondientes observaciones. Se supone que χ_t satisface el siguiente modelo funcional de series de tiempo,

$$m_{t+1}(x_i) = r(\chi_t)(x_i) + \epsilon_t(x_i), i = 1, \ldots, p \tag{1.11}$$

donde

$$r(\chi)(x_i) = \mathbb{E}(r_{t+1}(x_i|\chi_t = \chi)), i = 1, \ldots, p \tag{1.12}$$

χ es una función continua y $\epsilon = (\epsilon_t), t \in \mathbb{N}$ es ruido blanco y gaussiano, i.e, es una secuencia de variables i.i.d. a una normal, con $\mathbb{E}(\epsilon) = 0$ y $\mathbb{E}(||\epsilon||^2) < \infty$. Se asume que ϵ_t es independiente de χ_s para todo $s \neq t$.

Se considera el siguiente estimador Kernel funcional de la esperanza condicional $r(\chi)(x_i)$ dada en (2.1), basado en la muestra $\chi_1, \ldots, \chi_n$,

$$\hat{r}_c(\chi)(x_i) = \frac{\sum_{t=1}^{n-1} m_{t+1}(x_i) K_c(d(\chi_t, \chi))}{\sum_{t=1}^{n-1} K_c(d(\chi_t, \chi))} \tag{1.13}$$

donde $d(x, y)$ denota la distancia (métrica o semimétrica) entre x e y; donde K es la función Kernel, y c es el parámetro ventana.

El valor de la predicción para m_{n+1} se obtiene entonces por $\hat{m}_{n+1} = \hat{r}_c(\chi_n)$, estimador funcional clásico de Nadaraya-Watson.

Los pesos asociados con los valores de $m_{t+1}(x_i)$ son mayores a menor distancia entre χ y el correspondiente χ_t. Como todo enfoque de suavización no paramétrica , el estimador (1.13) depende de la elección del parámetro c de suavización. Su elección es muy importante, pues no debe de ser muy grande (mucho sesgo) ni muy pequeño (mucha varianza).

Bajo condiciones de reguralidad, la ventana c tiene que ser de orden $(\frac{\log^2(n)}{n})^{\frac{1}{p+4}}$. Un método práctico de selecionar c es elegirlo de dentro de una malla H_n dada por

$$H_n = \frac{1}{L} Q_{Cn}, \frac{2}{L} Q_{Cn}, \ldots, \frac{L-1}{L} Q_{Cn}, \qquad L \in \mathbb{N},$$

donde $(\frac{\log^2(n)}{n})^{\frac{1}{p+4}}$.

Aquí Q es una constante positiva lo suficientemente larga para que la maya H_n cubra la ventana óptima, mientras que el valor de L controla la diferencia relativa entre 2 valores

consecutivos dentro de la maya H_n. En la práctica, Q es desconocida pero se fija para algún valor grande importante desde el punto de vista práctico.

La h es obtenida como

$$\hat{l} = arming_{c_{l,n} \in H_n} CV_l$$

donde

$$CV_l = \frac{1}{(n-1)p} \sum_{i=1}^{p} \sum_{t=2}^{n} (m_t(x_i) - \hat{m}_l^{-t}(\chi_{t-1})(x_i))^2 \tag{1.14}$$

y

$$\hat{m}_l^{-t}(\chi)(x_i) = \frac{\sum_{j=1, j\neq t}^{n-1} K_{c_{l,n}}(d(\chi_j, \chi)) m_{j+1}(x_i)}{\sum_{j=1, j\neq t}^{n-1} K_{c_{l,n}}(d(\chi_j, \chi))}, i = 1, \ldots, p. \tag{1.15}$$

Capítulo 2

Aplicación a datos reales

2.1. Descripción de las bases de datos

Con el objetivo de predecir, a partir del período de ajuste, 1980-2008, las tasas de mortalidad futuras del cáncer colorrectal por sexo y grupos de edad quinquenales para el periodo 2009-2018 de España, se extrajo la información necesaria a partir de una única fuente de datos.

El análisis comparativo de los distintos métodos revisados y la propuesta basada en técnicas no paramétricas, se realiza mediante el manejo de las tasas de mortalidad por cáncer en España, cáncer en global.

2.1.1. Tasas de mortalidad por cáncer colorrectal

Los datos de mortalidad anuales del cáncer colorrectal y los datos poblacionales, por sexo y grupos de edad quinquenales, correspondientes al período 1980-2008 en España fueron obtenidos a partir de la base de datos del Instituto Nacional de Estadística (INE).

Los grupos de edad quinquenales van desde los menores de 5 años hasta los mayores de 85 años. En nuestro estudio seleccionamos, en particular, los grupos de edad quinquenales mayores de 25 años.

Las tasas crudas de mortalidad anuales del cáncer colorrectal por sexo y grupos de edad quinquenales son calculadas dividiendo el número de defunciones correspondiente al sexo y un grupo de edad quinquenal en cada año por la población de ese sexo y ese grupo de edad en el año correspondiente.

Tanto los datos de tasas de mortalidad por cáncer colorrectal como los datos poblacionales de los hombres y mujeres españoles en el periodo 1980-2008 fueron guardados como una matriz, donde las columnas son los correspondientes años del período y las filas los distintos grupos de edad. Las tasas anuales de mortalidad del cáncer colorrectal fueron observadas anualmente como una función de edad, definida en el punto medio de los grupos de edad quinquenales, en el que cada dato es un año.

Las bases de datos anteriores se recogieron en archivos Excel, las tasas crudas de la mortalidad por cáncer colorrectal de hombres y mujeres se encuentran en los tasashombre.xls y tasasmujer.xls, respectivamente. Los datos poblacionales se guardaron en poblacionhombre.xls, para los hombres, y en poblacionmujer.xls, en el caso de las mujeres.

Cabe decir, que todos los análisis estadísticos de esta parte del estudio fueron implementados con R version 2.14.0 en un script denominado **predicporsexo.R**. Para la lectura de los datos se utulizó el paquete **xlsReadWrite**.

Para cada año, dibujamos las asociaciones entre la edad y la mortalidad, es decir, dibujamos las curvas de mortalidad-edad. Luego tomamos el logaritmo de las tasas de mortalidad, que son los datos con los que nos interesa trabajar, para cada punto medio de los grupos de edad quinquenales y para cada año.

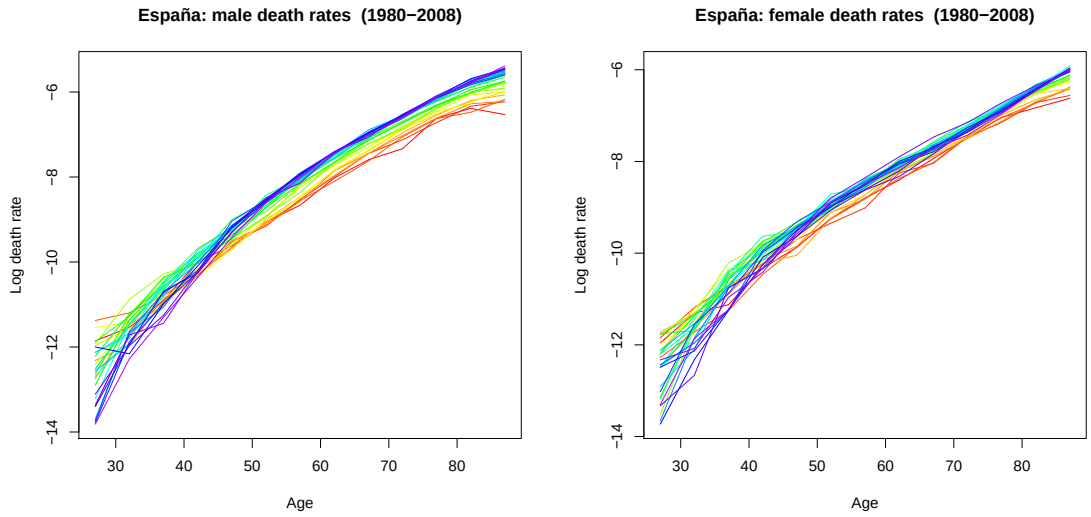

Figura 2.1: Tasas logarítmicas de la mortalidad por cáncer colorrectal en el periodo 1980-2008 en España

En la figura anterior, se muestran los datos funcionales, donde cada dato es un año

del periodo 1980-2008. Dicha figura muestra las tasas logarítmicas correspondientes a la población española, en el gráfico de la izquierda las de los hombres y en el gráfico de la derecha los de las mujeres.

La Figura 2.1 ilustra como las tasas logarítmicas de la mortalidad por cáncer colorrectal tanto en hombres como en mujeres, aumentan según avanza la edad en la mayoría de los años, aunque también se aprecia algo de variabilidad entre las edades, sobre todo en edades tempranas.

A continuación veamos los datos observados con la transformación logarítmica pero a través de los años, en el que cada curva representa las tasas de mortalidad para cada grupo de edad quinquenal (recordemos que trabajamos con los mayores de 25 años), lo que nos ayudará a ver como evolucionó la tasa de mortalidad a través de los años con respecto a las edades.

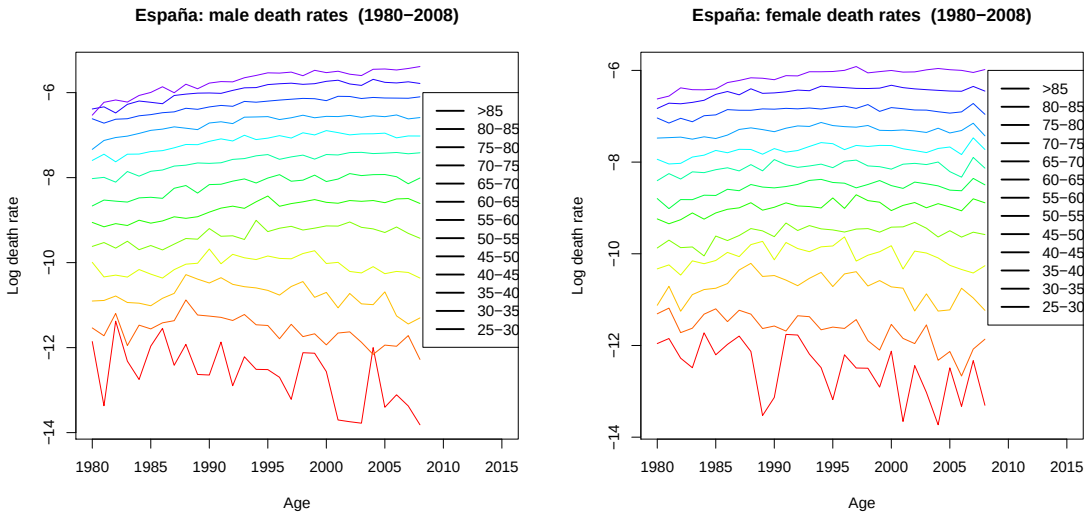

Figura 2.2: Tasas logarítmicas de la mortalidad por cáncer colorrectal a lo largo del periodo 1980-2008 en España

La Figura 2.2 muestra las tendencias de las tasas logarítmicas de la mortalidad por sexo y grupos de edad quinquenales a lo largo del periodo 1980-2008. En ella podemos observar que las tendencias de los grupos de edad más elevados aumentan y parece que se estabilizan en los 3 últimos años. Sin embargo, las relativas a las edades más jóvenes presentan una mayor variabilidad en las tasas logarítmicas durante todo el periodo; las tendencias de los hombres con estas edades, lineas amarilla, naranja y roja, presentan un descenso desde antes del comienzo de la década de los 90, aunque en los hombres con edad entre los 25 y

30 años vemos un aumento en el año 2004, incluso siendo mayor la tasa logarítmica para este grupo que para el grupo inmediatamente mayor.

La mortalidad en las mujeres con edades comprendidas entre los 30 y los 40 evolucionó hasta finales de la década de los 80 y después descenció levemente. La tendencia de las tasas logarítmicas de la mortalidad para las mujeres entre los 40 y 45 años también presenta un aumento en los primeros años, hasta el año 1996, a partir del cual se produce también un descenso. Para el grupo de edad quinquenal más joven tenemos que la mortalidad por cáncer colorrectal tuvo su punto más elevado en el año 1990, y a partir de éste fue descendiendo, pero con cierta variabilidad entre los años, hasta el año 2008.

Además de ser conscientes de que en general el leve descenso de la mortalidad se produjo antes en mujeres que en hombres, también se puede observar que la tasa de mortalidad en hombres es mayor que en mujeres.

2.1.2. Tasas de mortalidad por cáncer

Los datos de mortalidad anuales por cáncer por sexo y por edad, correspondientes al período 1960-2009 en España fueron obtenidos a partir de la base de datos de "The Human Mortality Database", http://www.mortality.org, al igual que los datos poblacionales.

Al igual que con la base de datos del cáncer colorrectal, tanto las tasas de mortalidad por cáncer como los datos poblacionales del período 1960-2009 fueron guardados como una matriz, donde las columnas son los años y las filas las distintas edades, desde los 25 a los 100 años. Ahora la función observada para cada año está definida en cada edad, en vez en el punto medio de los grupos de edad quinquenales, Figuras 2.3 y 2.4.

Las tasas de mortalidad por cáncer de hombres y mujeres se guardaron en los archivos Excel, espanahombres.xls y espanamujeres.xls, respectivamente, pero resaltar que se trabajará con el logarítmo de estas tasas. Y los datos poblacionales se encuentran en pobhombresespana.xls.xls y pobmujerespana.xls, para los hombres y las mujeres respectivamente.

Estes datos se utilizarán para el análisis comparativo de los diferentes métodos en revisión y la nueva propuesta, el cual se realiza en el script denominado **comparativo.R**, y al igual que con los datos del cáncer colorrectal para la lectura de datos se utilizó el paquete **xlsReadWrite**.

Dado que se utilizarán para el análisis comparativo, tomaremos una muestra de entrenamiento y otra de test, esto es, tomaremos como muestra de entrenamiento los datos relativos al período 1960-1999, y como muestra test los años desde el 2000 hasta el 2009

para calcular el error cuadrático medio (ECM) cometido al predecir para un horizonte de hasta 10 años.

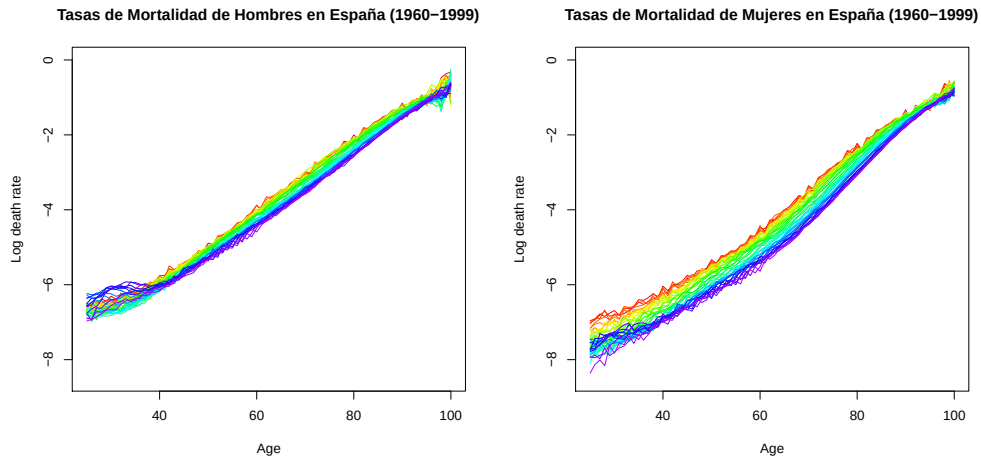

Figura 2.3: Tasas logarítmicas de la mortalidad por cáncer de la población española 1960-1999

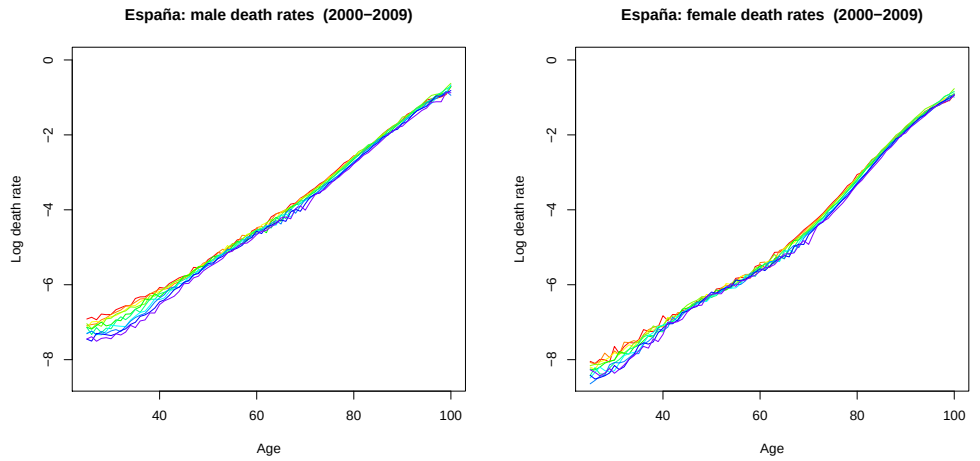

Figura 2.4: Tasas logarítmicas de la mortalidad por cáncer en la población 2000-2009

En la Figura 2.3, y teniendo en cuenta que trabajamos con la transformación logarítmica de las tasas, se presentan los datos observados por sexo. Y los datos de test, período 2000-2009, por sexo se muestran en la Figura 2.4.

En ambas figuras se puede ver que, en general, la mortalidad por cáncer en España es mayor en hombres que en mujeres. Las tasas logarítmicas presentan una variabilidad mayor

en las edades más avanzadas, sobre todo en los hombres, y en las más bajas, en especial en las mujeres.

Las Figuras 2.5 y 2.6 muestran la evolución en las tasas logarítmicas de mortalidad a lo largo del período de entrenamiento y el periodo de test 2000-2009, respectivamente.

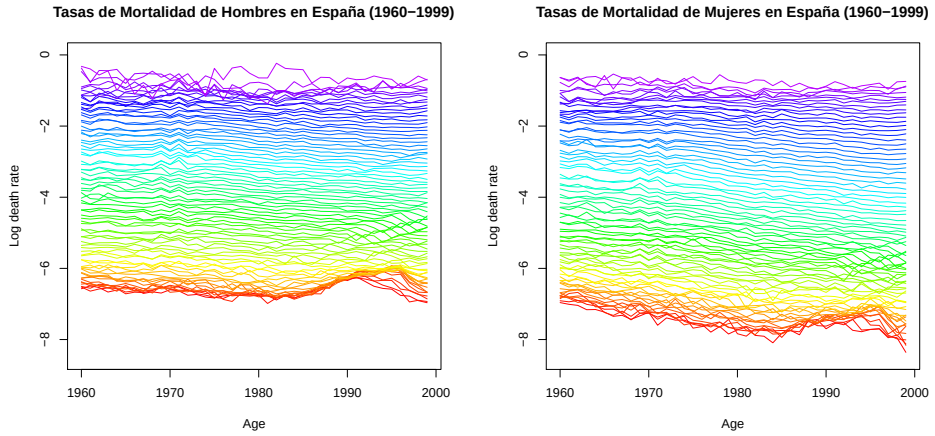

Figura 2.5: Tasas logarítmicas de la mortalidad por cáncer de la población española 1960-1999

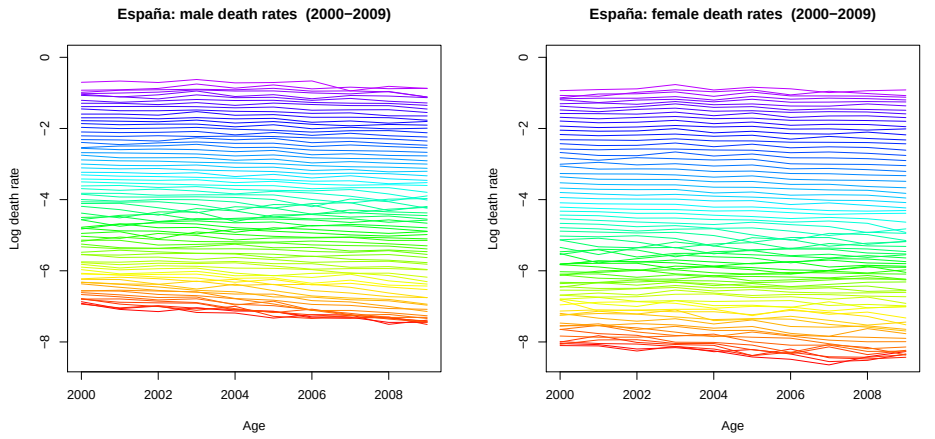

Figura 2.6: Tasas logarítmicas de la mortalidad por cáncer de la población española 1960-1999

Se puede decir que del 2000 al 2009 se reduce la variabilidad en general, sobretodo en las edades más altas. Y también que los gráficos del período de test muestran que la mortalidad en hombres es más elevada que en las mujeres.

2.2. Modelización y predicción.

En esta sección se procede a modelizar las tasas logarítmicas de la mortalidad del cáncer colorectal por sexo y edad de la población española para el periodo 1980-2008, mediante los métodos revisidos en la capítulo 1 para predecir tasas de mortalidad a un horizonte de 10 años, período 2009-2018. El paquete empleado para realizar los siguientes estudios es el **demography**, que contiene funciones para el análisis demográfico incluyendo cálculos para tablas de superviviencia o esperanza de vida, para la modelización de Lee-Carter; también para análisis de datos funcionales para tasas de mortalidad, tasas de fertilidad, números de migración neta, y proyecciones de población estocástica.

Notar que los datos no son directamente de naturaleza funcional, pero asumimos que hay series de tiempo funcionales fundamentales las cuales son observadas con error en los puntos discretizados, intervalos de edad de longitud 5 años de edad.

2.2.1. Método de Lee-Carter.

Recordemos que este método es parámetrico y poco flexible. Realiza el ajuste del modelo mediante la descomposición en un valor singular y después ajusta k_t mediante un reajuste del número total de muertes padecidas en cada año. Y , por último, para predecir, pronostica el coeficiente k_t mediante un melelo de series de tiempo ARIMA.

Para realizar el ajuste del modelo con el Método de Lee-Carter requerimos de la función **lca**, a la cual le pasamos los datos reales y el método mediante el cual queremos ajustar k_t, que por defecto, utiliza el realizado por Lee-Carter.

La Figura 2.7 muestra el ajuste del modelo Lee-Carter para las tasas de logarítmicas del cáncer de colorrectal de hombres españoles mayores de 25 y por grupos de edad quinquenales en el periodo 1980-2008.

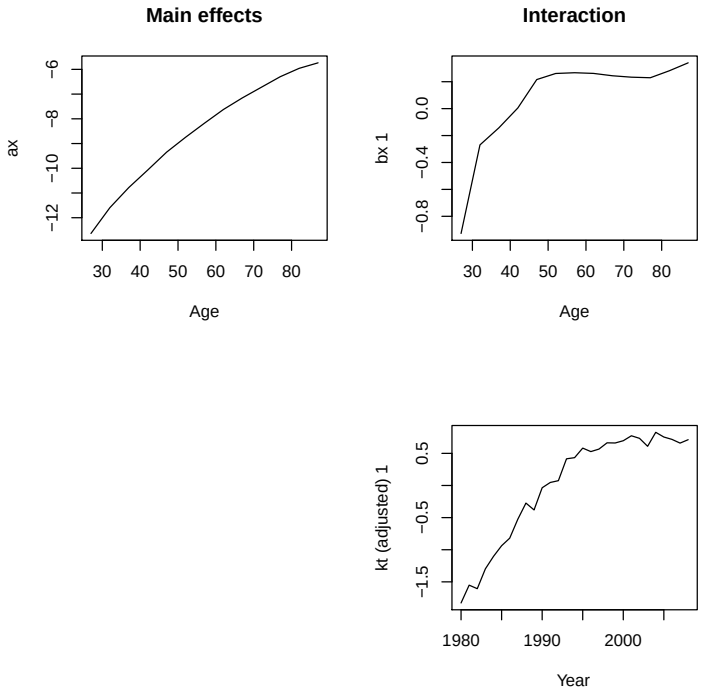

Figura 2.7: Ajuste del modelo de Lee-Carter para el cáncer colorrectal de los hombres españoles en el periodo 1980-2008.

El parámetro b_x explica un 61,2 % de la variabilidad en torno a la media, y para los grupos de edad quinquenales mayores o iguales a los de $[40, 45]$, k_t muestra un incremento en la mortalidad logarítmica en general, aunque realmente sufre un máximo en el año 2004, y después un descenso leve.

En la Figura 2.8 se puede ver que los pronósticos de k_t, realizados a partir de la función **forecast.lca**, muestran un aumento de las tasas logarítmicas de mortalidad a lo largo de todo el periodo de predicción y los correspondientes intervalos de predicción muestran que las estimaciones poseen mayor incertidumbre a lo largo que predecimos un paso más, ya que la amplitud aumenta año a año.

Las predicciones estimadas se presentan en la Figura 2.9, y muestran un ascenso en la mortalidad para los mayores o iguales del grupo quinquenales de los 40 a los 45 , sin embargo para los menores presentan un descenso leve en la mortalidad.

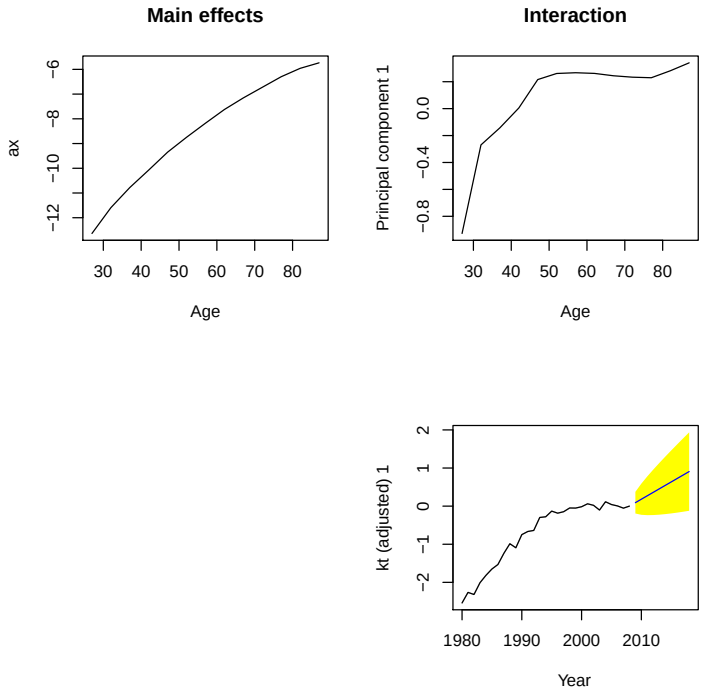

Figura 2.8: Modelo de predicción de Lee-Carter por cáncer colorrectal de los hombres españoles.

A continuación , se realiza la modelización y la predicción en el caso de las mujeres españolas. El modelo de Figura 2.10 explica un 56,6 % de la variabilidad. Se estima, para los grupos quinquenales menores de los 45 años, un descenso brusco hasta el año 1988 junto con un aumento leve en el 1990 y con un descenso también leve hasta el año 1995, a partir del cual ascienden un poco exceptuando en el año 2007, año en el que menos tasas de mortalidad se producen.

En la Figura 2.11 se muestra el modelo de predicción a partir del modelo ajustado, y al contrario de lo predicho con los hombres, se pronostica un descenso paulatino de las tasas de mortalidad, junto con unos intervalos con una amplitud más grande a mayor horizonte de predicción.

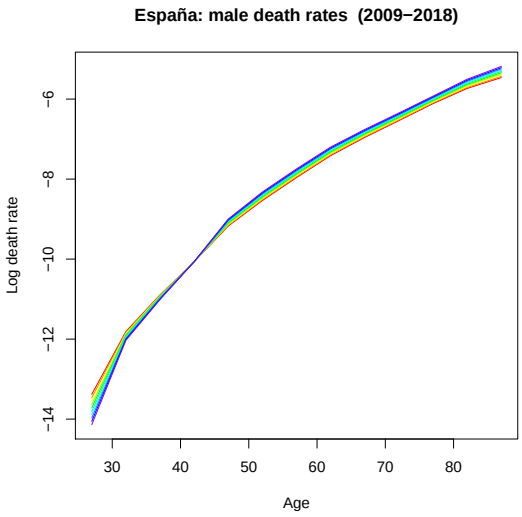

Figura 2.9: Predicciones para el periodo 2009-2018 por cáncer colorrectal de los hombres españoles.

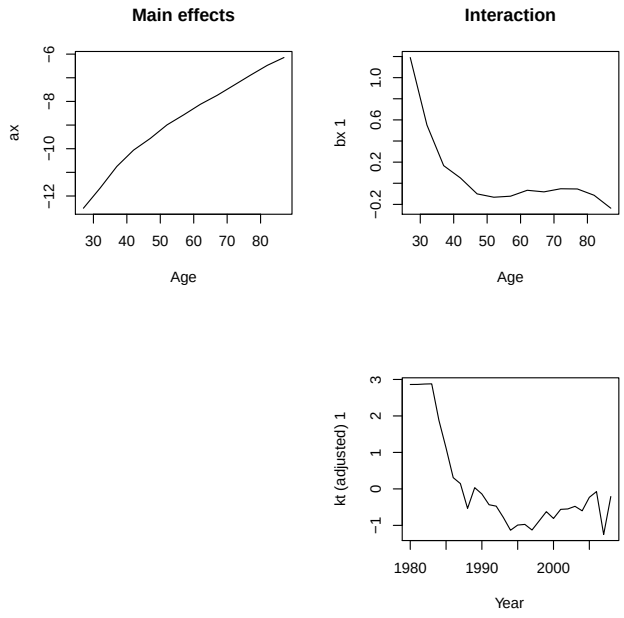

Figura 2.10: Modelización de Lee-Carter para el cáncer colorrectal de las mujeres españolas.

Se obtienen las tasas de predicción, Figura 2.12, y vemos que para los mayores de 45 años no se produce un descenso sino un aumento de la mortalidad.

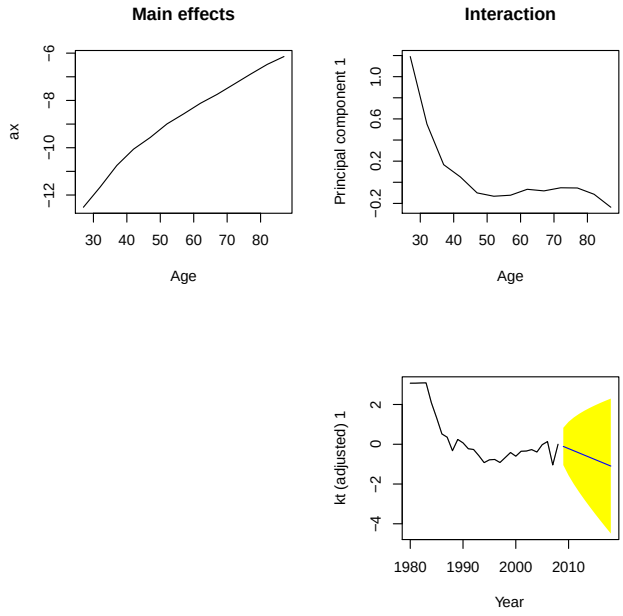

Figura 2.11: Modelo de predicción de Lee-Carter para el cáncer colorrectal en las mujeres españolas.

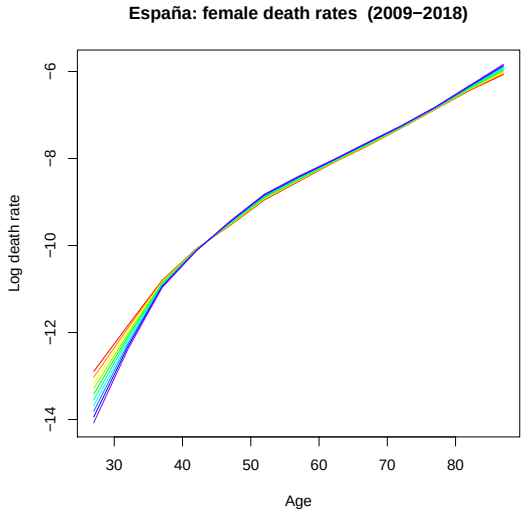

Figura 2.12: Predicciones de Lee-Carter para el cáncer colorrectal en las mujeres españolas.

2.2.2. Método Lee-Miller.

Este difiere del anterior básicamente en que realiza la segunda estimación de k_t respecto a la esperanza de vida, en lugar de ajustar repecto al total de defunciones; y en que como

tasas de salto toma las tasas actuales en vez de las de ajuste.

La Figura 2.13 muestra el modelo de Lee-Miller de la mortalidad para los hombres, y sugiere una tendencia ascendente de la mortalidad durante todo el período, para los grupos de edad quinquenales menores de los 40 años.

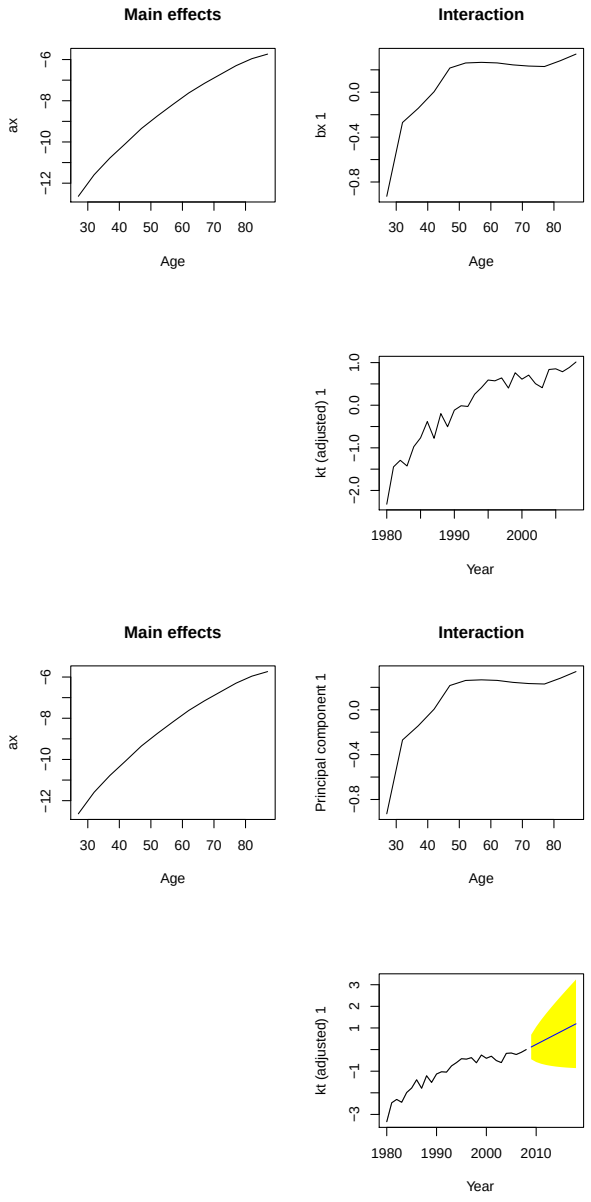

Figura 2.13: Ajuste del modelo de Lee-Miller junto con su modelo de predicción para el cáncer el cáncer colorrectal de los hombres españoles.

La variabilidad explicada por el modelo es de un 61,2 %, una proporción baja, la misma que con Lee-Carter. El modelo se realiza también mediante la función **lca**, pero ahora le decimos que me reajuste k_t mediante la esperanza de vida.

Tras realizar modelos de series de tiempo ARIMA para estimar los pronósticos de k_t a 10 pasos futuros, se obtiene el modelo de predicción de la Figura 2.13, el cual también se realiza con las misma función que con Lee-Carter pero ahora pidiéndole que tome como tasas de salto las actuales.

Las proyecciones de k_t sugieren, al igual que las de Lee-Carter, un aumento continuo en la mortalidad a lo largo de las 10 proyecciones futuras, y los intervalos de predicción también reflejan que las estimaciones más futuras poseen mayor incertidumbre pues la cobertura de los intervalos es mayor.

A continuación, Figura 2.14, mostramos las proyecciones realizadas para los hombres por dicho método, que además de predecir un aumento en la mortalidad para los grupos quinquenales mayores de 40, predicen un descenso de la mortalidad para los menores, al igual que lo esperado por el modelo de Lee-Carter.

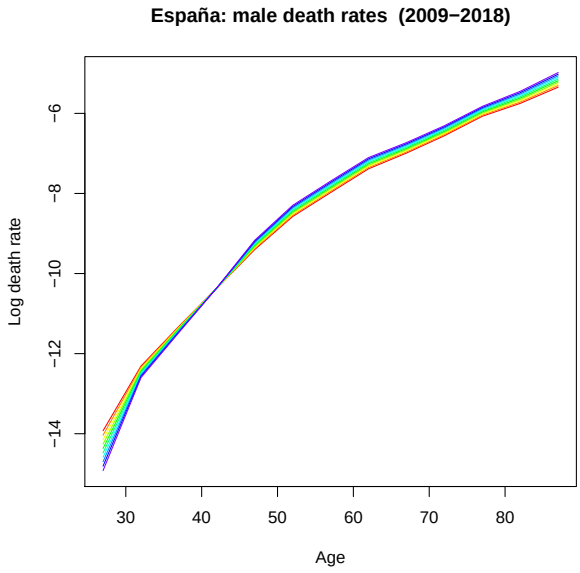

Figura 2.14: Predicciones de Lee-Miller por cáncer colorrectal para los hombres españoles.

Este método modeliza también un descenso pronunciado en las mujeres menores de 45 años, aunque menos brusco que Lee-Carter, pero hasta el año 1997, Figura 2.15. Esta misma figura (derecha) se muestra también que las predicciones de k_t monótonas decrecientes a lo

largo del horizonte de predicción para las anteriores edades. El modelo sugiere también un

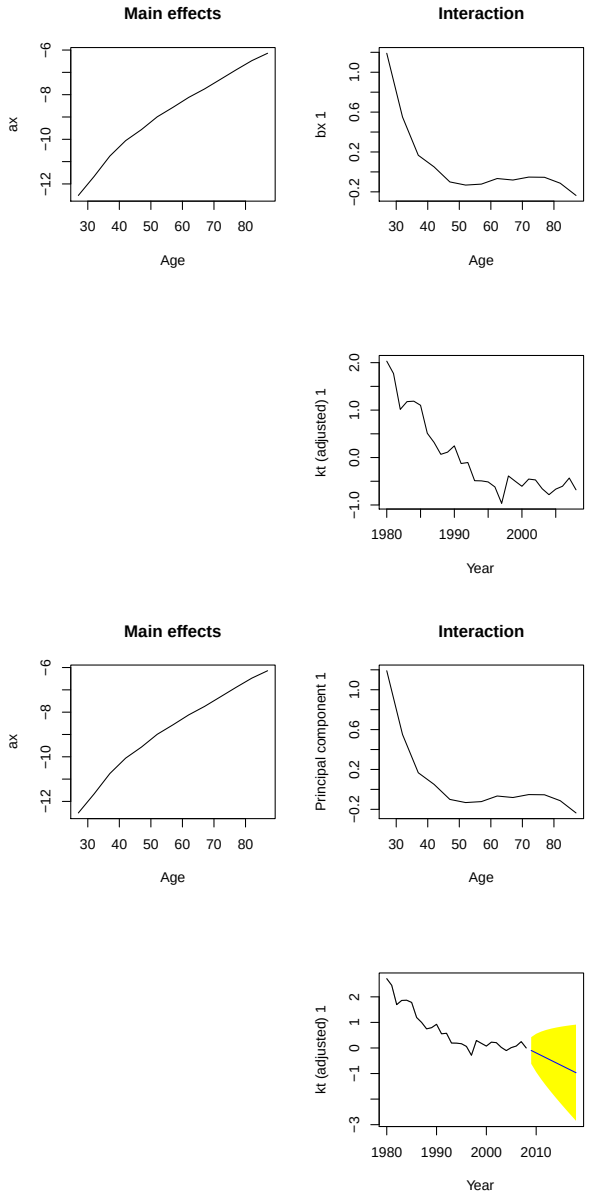

Figura 2.15: Ajuste y modelo de predicción de Lee-Miller por cáncer colorrectal para las mujeres españolas.

aumento en edades mayores de 45 años, Figura 2.16.

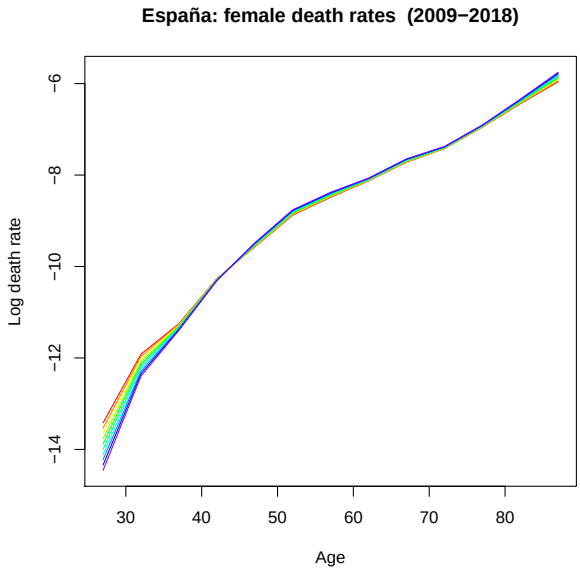

Figura 2.16: Predicciones de Lee-Miller por cáncer colorrectal para las mujeres españolas.

2.2.3. Método de Booth-Maindonald-Smith (BMS)

Estos autores propusieron ajustar k_t mediante un modelo de regresión de Poisson sobre el número de muertes en cada año $\mathbf{D_t}$ suponiendo k_t lineal, para lo cual dejaron los parámetros a_x y b_x tal como se obtuviron en la primera estimación.

El modelo de ajuste resultante por el método BMS para los hombres, que en R se realizó mediante la función **bms** pidiéndole que las tasas de salto se tomen según técnicas de bondad de ajuste, explica un 72 % de la variabilidad, un 10 % más que con los anteriores. Y capta, para los grupos quinquenales menores de 45 años, un descenso de la mortalidad a lo largo del período de ajuste, excepto en el año 2003 que tiene un pequeño aumento en las tasas, Figura 2.17. Por lo que cabe esperar una tendencia decreciente en las tasas logarítmicas de la mortalidad para el período 2009-2018, tal y como muestra el modelo de predicción, Figura 2.17.

Las predicciones se representan en la Figura 2.18 , en la que vemos que en edades altas se espera un aumento de las tasas logarítmicas.

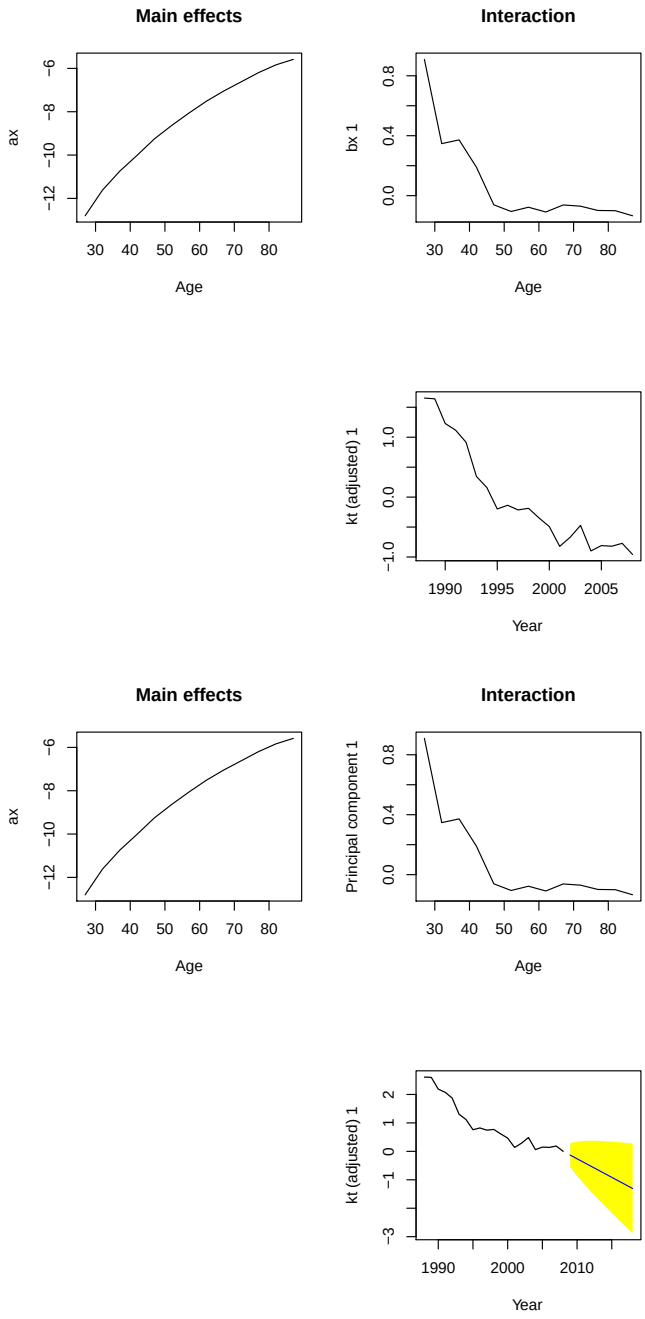

Figura 2.17: Modelo de ajuste y predicción de BMS para el cáncer colorrectal de los hombres españoles.

Para las mujeres el modelo BMS explica casi un 7 % más de la variabilidad existente que en los anteriores métodos, un 63,3 %. Y, observando k_t en la Figura 2.19, ajusta para

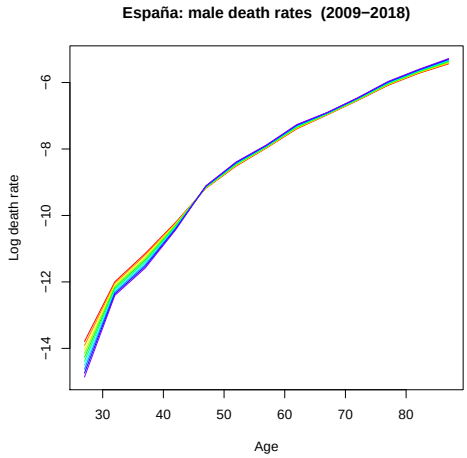

Figura 2.18: Predicciones BMS para el cáncer colorrectal de los hombres españoles.

las mujeres menores de 45 años, una tendencia en global decreciente a lo largo de los años 1980-2008, con mayor variabilidad que en el caso de los hombres, sobre todo en la última década.

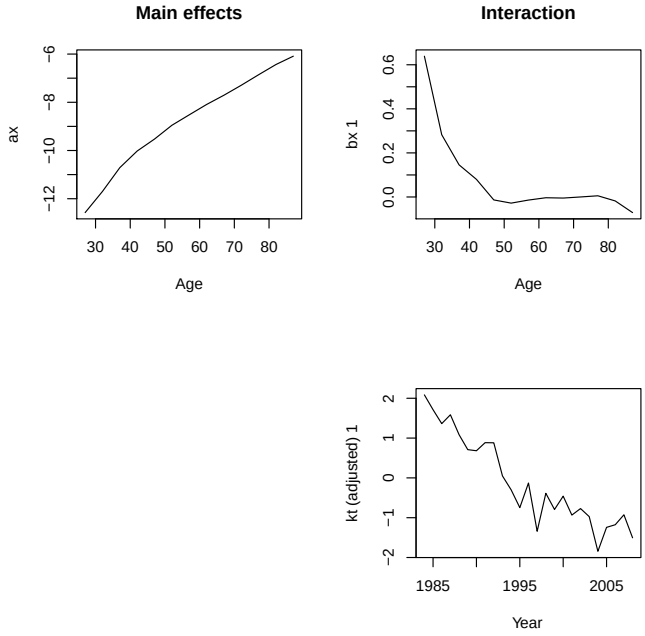

Figura 2.19: Modelo de estimación y predicción BMS para el cáncer colorrectal de las mujeres españolas.

Tras proyectar k_t mediante modelos de series de tiempo ARIMA, y pidiéndole a la función que ajuste las tasas de salto según este método, se obtiene un descenso continuo en las predicciones a medida que avanzamos en horizontes de predicción, Figura 2.20.

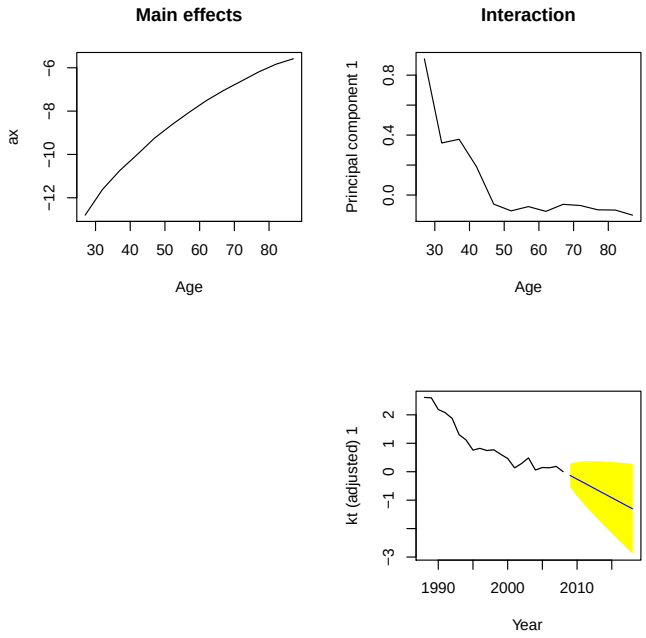

Figura 2.20: Modelo de predicción BMS para el cáncer colorrectal de las mujeres españolas.

Obteniendo así, las predicciones de la Figura 2.21, en la que se observa que, al igual que los anteriores, se pronóstican mayores tasas para edades elevadas a mayor horizonte de predicción.

2.2.4. Método de Hyndman

Sea $m_t(x_i)$ el logaritmo de las tasas de mortalidad observadas para la edad x en el año t. Asumimos que hay una función suavizada fundamental $f_t(x_i)$ que observamos con error y discretizada en los puntos de x, de la edad. Nuestras observaciones son $\{x_i, m_t(x_i)\}$ con $i = 1, \ldots, p, t = 1, \ldots, n$, entonces podemos escribir

$$m_t(x_i) = f_t(x_i) + \sigma_t(x_i)\epsilon_{t,i}$$

donde $\epsilon_{t,i}$ es una variable independiente e identicamente distribuida a una normal estándar, y $\sigma_t(x_i)$ permite contar el ruido que varía con x, con la edad.

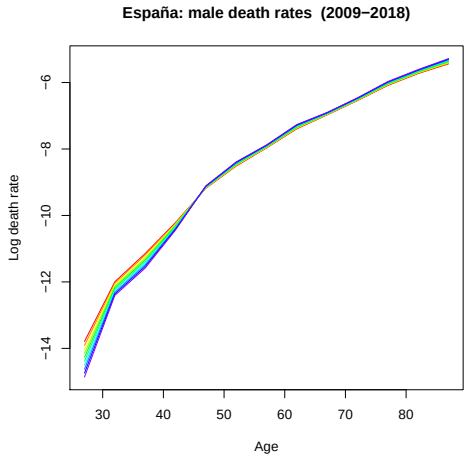

Figura 2.21: Predicciones BMS para el cáncer colorrectal de las mujeres españolas.

Particularmente $\{x_1, \ldots, x_{18}\}$ son las edades, en nuestro caso, los grupos de edad quinquenales ($x_1 = [25, 30]$, $x_2 = [30, 35]$,...). Estamos interesados en predecir $m_t(x_i)$ para $x \in [x_1, x_{18}]$ y $t = 2009, \ldots, 2018$, donde h es el horizonte que queremos predecir.

Notar que los datos no son directamente de naturaleza funcional, pero asumimos que hay series de tiempo funcionales fundamentales las cuales son observadas con error en los puntos discretizados, intervalos de edad de longitud 5 años de edad.

1. Se realiza un suavizado de los datos de cada año t, $\{x_i, m_t(x_i), quadi = 1, \ldots, 18\}$, al objeto de obtener una aproximación no paramétrica a la curva de mortalidad subyacente $f_t(x)$, con $x \in [x_1, x_{18}]$.

   En este estudio el suavizado se lleva a cabo usando la regresión por splines penalizados ponderados con pesos inversamente proporcionales a la $\sigma_t^2(x_i)$, mediante la función **smooth.demogdata**.

   Definamos $N_t(x)$ como la población total de la edad x en el año t. Entonces, $m_{x,t}$ es distribuida aproximadamente como una binomial con varianza estimada $N_t^{-1}(x)m_{x,t}[1 - m_{x,t}]$. Así que un estimador de la varianza $\sigma_t^2(x_i)$, varianza de $m_t(x) = ln[m_{x,t}]$, se obtiene mediante

   $$\hat{\sigma}^2 \approx (1 - m_{x,t})^{-1} N_t(x) m_{x,t}^{-1}.$$

   Los pesos son definidos igual a la inversa de la varianza empírica de las tasas logarítmicas de la mortalidad:

   $$w_t(x) = \frac{N_t(x)m_{x,t}}{(1 - m_{x,t})}$$

Se decide usar los anteriores pesos ya que le dan menos peso a las edades grandes y un peso mayor a las edades jóvenes. Cabe decir que la ponderación se hace cargo de la heterogeneidad debido a $\sigma_t(x)$.

Para obtener mejores estimaciones de $f_t(x)$ aplicamos una restricción, especialmente cuando $\sigma_t(x)$ es grande. Como estamos trabajando con datos de mortalidad, el criterio que se aplica es suponer que $f_t(x)$ aumenta de forma monótona para las edades mayores de una edad c, en nuestro estudio, para las edades mayores de 50 años, $x > 50$. Esta restricción monótona nos permite reducir el ruido en las curvas estimadas para las edades más avanzadas.

Se obtiene así, $\left\{\hat{f}_1(x), \ldots, \hat{f}_n(x)\right\}$ (una por año), y se presentan, tanto para los hombres como para las mujeres, en la Figura 2.22 junto con los datos originales. Las n curvas suavizadas, concretamente 29, son nuestras observaciones funcionales, y se puede decir que aproximan bien a los datos observados.

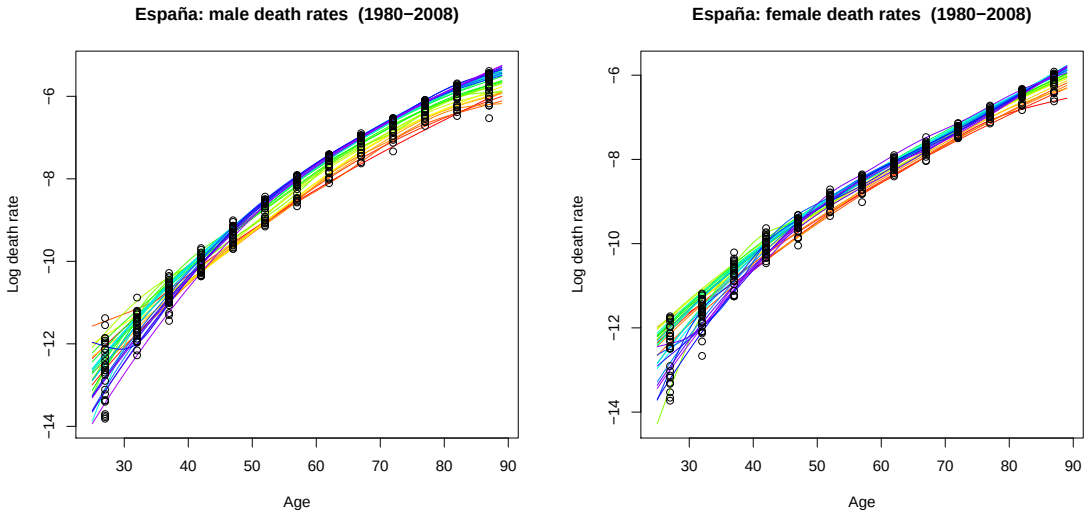

Figura 2.22: Tasas logarítmicas suavizadas de la mortalidad por cáncer colorrectal junto con los datos observados a lo largo del periodo 1980-2008 en España.

Al suavizar y estimar $f_t(x)$ para la edad, tenemos una curva para cada una de las edades de entre 25 años y los 89 años. Las tendencias de las tasas logarítmicas suavizadas de la mortalidad por cáncer colorrectal a lo largo del periodo 1980-2008 para las diferentes edades, que se observan en la Figura 2.23, muestran, para los hombres más jóvenes, un descenso desde principios del periodo 1980-2008, en concreto desde el año 1982, aunque en el año 2004 presentan un pico para edades hasta los 30

años aproximadamente. Vemos también que para los mayores de 33 aproximadanete el descenso es da a partir de el año 1988, y que para los hombres con edades mayores de los 38 años aproximadamente y cercanas a esta edad las tasas logaritmicas suavizadas de la mortalidad disminuyen levemente desde el año 1990.

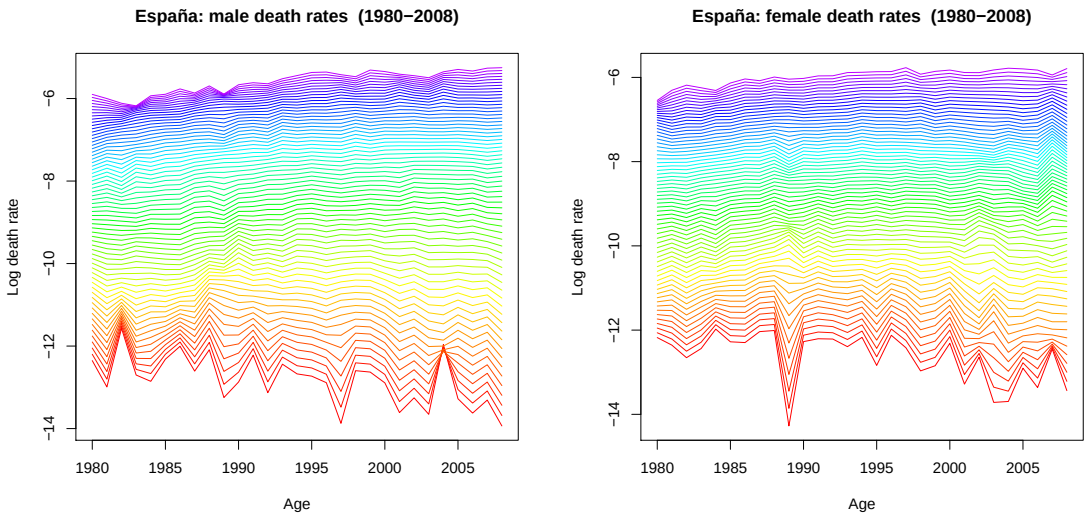

Figura 2.23: Tendencias de las tasas logarítmicas suavizadas de la mortalidad por cáncer colorrectal a lo largo del periodo 1980-2008 en España para todas las edades.

Las tendencias de las observaciones funcionales de las mujeres, presentan en general un aumento hasta la segunda mitad de la década de los 90 y a partir de estos años muestran un leve descenso con un aumento 2007 y disminuyen en el 2008. Pero para las mujeres más envejecidas, al revés que lo anterior, disminuyen en el 2007 y aumentan en el 2008.

2. Pasamos al ajuste del modelo, que se realiza con la función **fdm**. Para ello, se descomponen las curvas ajustadas a través de la expansión de una función básica, descomposición en componentes principales funcionales.

   Notar que, en lugar de la suavización de las componentes principales directamente, se prefiere la suavización de los primeros datos observados, ya que permite colocar restricciones relevantes en la suavización con mayor facilidad, como las del punto anterior.

   Si en la base de datos hay años atípicos, se debe de usar una estimación robusta para las funciones de edad $a(x)$ y $b_j(x)$. Las series de tiempo $k_{t,j}$ no se estimarían de forma robusta, de modo que cualquier año atípico sería modelado como atípico en las series

de tiempo. Esto permite que los años atípicos se identifiquen claramente, y los modelos de series de tiempo usados para predecir serían seleccionados para permitir los datos atípicos. Cabe decir, que la existencia de años atípicos podrían afectar negativamente al análisis de componentes principales, ya que los atípicos pueden incluso llegar a provocar que las componentes principales cambien de dirección; por eso es importante que, cuando existen, que se realice una estimación robusta.

En nuestro caso , tras hacer un estudio en R para detectar datos atípicos, no encontramos datos atípicos. El estudio para la detección de datos atípicos se realizó en un script denominado **predicfda** se utiliza la función **outliers.depth.trim** que se encuentra en el paquete **fda.usc**. Esta función realiza la deteccción mediante un Bootstrap suavizado basado en recorte.

Este método consiste en 3 pasos:

a) Se calcula la profundidad de las curvas de la muestra original. En nuestro caso, le pasamos como muestra original los datos suavizados, ya los datos que se usan para ajustar el modelo.

b) Se obtienen b muestras por bootstrap suavizado del conjunto de datos recortado obtenido después de descartar el $\alpha\,\%$ de las curvas más profundas. En nuestro caso le decimos que tome un valor de suavizado de 0,10 que haga 200 réplicas. Sean $f_t^b(x)$, donde $t = 1, \ldots, n$,$b = 1, \ldots, B$, estas muestras.

c) Para cada b, se obtiene un valor C^b como el cuantil empírico $1\,\%$ de la distribución de las profundidades, $D(f_t^b(x))$. El C final, el corte final, es la mediana de los valores C^b, $b = 1, \ldots, B$. Así que un atípico vendrá determinado por su baja profundidad.

Tras aplicar el anterior método la función nos devuelve el corte final, todos los datos con profundidades menores a ese corte son los datos atípicos. El corte que nos devuelve en nuestro caso es 0, por lo que es obvio que no existen atípicos en los datos, ya que no hay profundidades más bajas de 0.

Por lo que, las curvas ajustadas y centradas, $\hat{f}_t^{\star}(x) = \hat{f}_t(x) - \hat{a}(x)$ se descomponen en una base ortonormal de funciones usando análisis de componentes principales para datos funcionales clásico, de modo que

$$\hat{f}_t(x) = \hat{a}(x) + \sum_{k=1}^{J} \hat{k}_{t,j}\hat{b}_j(x) + \hat{e}_t(x), \tag{2.1}$$

donde $b_j(x)$ las funciones básicas ortonormales, donde $a(x)$ es la estimación de la media de localización de $m_t(x)$, el promedio de las tasas logarítmicas de la mortalidad

a través de los años y $e_t(x) \sim N(0, \sigma_x^2)$ está modelando el error, es decir, la diferencia entre las curvas splines y las curvas ajustadas por el modelo .

En este estudio, la elección del parámetro J, del número de funciones básicas, se determina minimizando el error cuadrático medio integrado de predicción (MISFE).

$$MISFE(h) = \int_x e_{n,h}^2(x)dx$$

donde $e_{n+h} = m_{n+h}(x) - \hat{m}_{n+h|n}(x)$, y $\hat{m}_{n+h|n}(x)$ son las h predicciones futuras de $m_{n+h}(x)$.

Cabe decir, que Hyndman & Booth (2008) encontraron que las predicciones son insensibles a elecciones de J, cuando J es suficientemente largo.

En nuestro estudio calculamos el error cuadrático de predicción integrado, el ISFE, utilizando una descomposición en una base ortonormal de diferentes ordenes, para diferentes valores del parámetro J, mediante la función **isfe** que se encuentra en el paquete demography de R, luego se calcula el MISFE haciendo el promedio de todas las predicciones para cada orden.

A la anterior función le decimos con qué método queremos realizar el ajuste del modelo, un análisis clásico de componentes principales funcionales de orden 0 a un máximo orden de 9, también el método para proyectar los $k_{t,j}$, para lo que usamos modelos de series de tiempo univariantes, modelos ARIMA; también le decimos que utilice como mínimo los datos disponibles de los 15 primeros años para obtener predicciones a horizontes $h = 1, 2, \ldots, 10$.

En el caso de la mortalidad por cáncer colorectal de los hombres españoles, un adecuado ajuste del modelo, determinado por el MISFE, es un análisis funcional de componentes principales con 4 funciones básicas:

J	h=1	h=2	h=3	h=4	h=5
0	4.557833	4.665397	5.757542	5.640925	6.936790
1	1.153478	1.439918	2.250522	2.117522	4.175689
2	1.153339	1.439383	2.250380	2.117437	4.175436
3	1.163164	1.406845	2.216845	1.999217	3.973785
4	1.180846	1.494386	2.284718	2.043064	4.099120
5	1.180840	1.494388	2.284718	2.043061	4.099118
6	1.180841	1.494388	2.284718	2.043062	4.099118
7	1.181377	1.495120	2.285542	2.042814	4.097177
8	1.181377	1.495120	2.285542	2.042814	4.097178
9	1.181378	1.495120	2.285542	2.042814	4.097178

J	h=6	h=7	h=8	h=9	h=10	MISFE
0	7.697928	8.290362	8.497790	8.496017	9.702295	7.024288
1	6.359976	8.260845	9.104182	10.836421	13.420984	5.911954
2	6.359265	8.260340	9.103274	10.836094	13.420400	5.911535
3	6.063169	8.014476	8.615167	10.369443	12.929226	**5.675134**
4	6.229530	8.172864	8.765770	10.428822	13.243936	5.794306
5	6.229530	8.172863	8.765771	10.428818	13.243930	5.794304
6	6.229530	8.172863	8.765770	10.428817	13.243930	5.794304
7	6.229488	8.174098	8.765476	10.429070	13.244792	5.794495
8	6.229489	8.174098	8.765476	10.429070	13.244793	5.794496
9	6.229488	8.174099	8.765476	10.429071	13.244792	5.794496

Obtenemos entonces una descomposición en una base ortonormal de 4 funciones usando análisis de componentes principales para datos funcionales, Figura 2.24. Al mismo tiempo, se obtiene una estimación de la tendencia central.

Las funciones estimadas explican un 98,8 % de la variabilidad total. La componente principal explica un 65,0 % la 2ª un 25,4 % y la 3ª un 7,1 %.

La primera función básica muestra un incremento en la mortalidad durante el periodo 1980-2008 para las edades mayores de los 40 años, con los cambios más fuertes en los niveles de edad más altos. La segunda función básica muestra un descenso en la mortalidad desde el año 2000 para todas las edades, ocurriendo los cambios más fuertes en los niveles de edad más bajos. La tercera describe la variación en la mortalidad logarítmicas en hombres menores de 30 años y mayores de 50 comparandolo con los

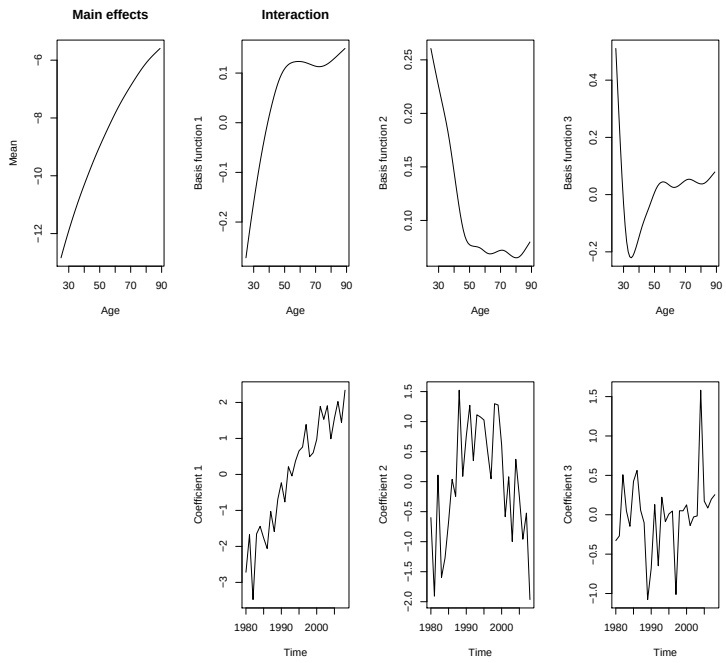

Figura 2.24: Modelo con 4 funciones básicas para los hombres españoles para el periodo 1980-2008.

que tienen edades intermedias a éstas.

Pasemos entonces a estudiar el ajuste del modelo para las tasas logarítmicas de la mortalidad de las mujeres españolas, a continuación mostramos el ISFE con los distintos valores del parámetro J en las diferentes h predicciones futuras junto con el MISFE cometido con cada valor del parámetro.

El mínimo MISFE cometido se alcanza con J=0, esto es, solamente tenemos en cuenta la media, esto implicaría que las tasas ajustadas fuesen la media en todos los años, es decir, no existiría el efecto año. Esto produciría que todo es constante y no se podrían realizar predicciones ya que no tendríamos coeficientes los cuales poder proyectar, así que tanto las tasas ajustadas como las predicciones serían la media. Lo anterior es producido, como veremos más adelante a que las predicciones convergen a la media muy rápido, en un plazo muy corto de tiempo.

Elegimos entonces el siguiente valor del parámetro J que minimiza el MISFE, que es J=5.

Ajustando un modelo,figura 2.25, de regresión funcional con 5 funciones básicas ortonormales explicamos un 99,1 % de la variación alrededor de la curva media de

J	h=1	h=2	h=3	h=4	h=5
0	1.789991	1.555779	2.119611	2.877476	3.401605
1	1.311778	1.096037	2.134190	3.791837	4.714510
2	1.311853	1.096195	2.134474	3.792033	4.714841
3	1.233349	1.042467	1.963256	3.522177	4.584538
4	1.233322	1.042473	1.963228	3.522154	4.584498
5	1.233322	1.042472	1.963227	3.522154	4.584498
6	1.233322	1.042472	1.963228	3.522154	4.584498
7	1.233326	1.042478	1.963231	3.522162	4.584516
8	1.233325	1.042478	1.963232	3.522153	4.584523
9	1.233328	1.042480	1.963230	3.522162	4.584369

J	h=6	h=7	h=8	h=9	h=10	MISFE
0	5.135011	6.161986	6.457596	6.536589	7.024761	4.30604
1	6.862845	9.360023	10.114718	10.862065	12.317147	6.256515
2	6.863259	9.360547	10.115344	10.862615	12.317698	6.256886
3	6.561418	8.944006	9.859381	10.394190	11.887745	5.999253
4	6.561410	8.943964	9.859250	10.394145	11.887677	5.999212
5	6.561409	8.943962	9.859248	10.394147	11.887676	**5.999211**
6	6.561409	8.943961	9.859249	10.394146	11.887675	5.999211
7	6.561423	8.943962	9.859248	10.394152	11.887685	5.999218
8	6.561423	8.943958	9.859232	10.394148	11.887685	5.999216
9	6.561378	8.943954	9.859205	10.394153	11.887729	5.999199

la mortalidad logarítmica. La proporción explicada por cada función básica es de un 58,8 %, 27,9 %, 9,8 %, 1,6 % y 1,0 % para $J = 1, \ldots, 5$, respectivamente.

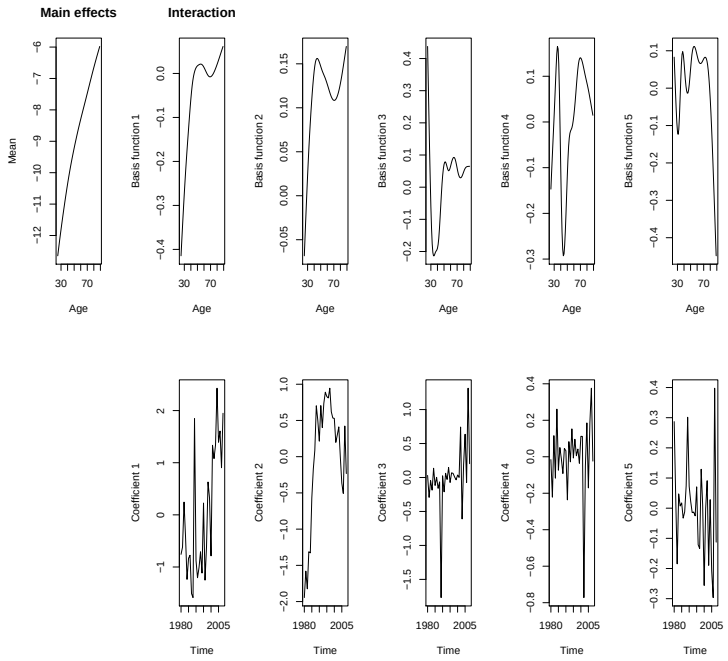

Figura 2.25: Modelo con 5 funciones básicas para las mujeres españolas para el periodo 1980-2008.

La primera componente principal describe la variación en la mortalidad logarítmica en mujeres entre los 45 y los 64, y para las mayores de 75 (incluyendo estas últimas), comparado las mujeres jóvenes (menores de 45) y las que tienen edades entre los 65 y los 74, y muestra que para ellas la mortalidad logarítmica aumentó hasta el año 2004. La segunda función básica modela la mortalidad logarítmica de las mujeres con edades mayores de los 28 años, sin incluir éstas últimas. La tercera explica la variabilidad de la mortalidad logarítmica para las mujeres más jóvenes, menores de 30 años (sin incluir éstas), y para las mayores de 46 años (sin contar con las de 46). La cuarta es compleja y contrasta dicha mortalidad en mujeres con edades entre los 30 y los 38, ambos inclusive, y las mayores de 58 años, éstas últimas sin incluir, con las otras edades. La quinta función básica es bastante compleja, explica la variabilidad para las mujeres de 25 y 26 años y con niveles de edad intermedios (hasta los 79 incluidos), salvo las comprendidas entre los 27 y 35, y las de 45, 46 y 47 años.

3. A continuación, se ajustan modelos de series de tiempo univariante para cada uno de los coeficientes $k_{t,j}$, $j = 1, \ldots, J$.

   Los modelos de series de tiempo univariantes realizados en este estudio, para estimar los coeficientes o puntuaciones de las componentes principales $k_{t,j}$, fueron los modelos

ARIMA.

4. Se pronostican los coeficientes $k_{t,j}$, $j = 1, \ldots, J$ para $t = n+1, \ldots, n+h$ usando los modelos de series de tiempo ajustados.

5. Usa los coeficientes de predicción para obtener predicciones de $f_t(x)$, $t = n+1, \ldots, n+h$. Por la ecuación 1.1 tenemos que las predicciones de $f_t(x)$ son tmabién predicciones de $m_t(x)$.

   Proyectamos a 10 años futuros, por lo tanto, las tasas logarítmicas de la mortalidad a 10 años para ambos sexos tomando como periodo de ajuste los datos observados del periodo 1980-2008, y se calculan los intervalos de predicción correspondientes al 95 %, mediante la función **forecast**.

   Para las tasas logarítmicas de los hombres se proyectan cada uno de los 3 $k_{t,j}$ del ajuste del modelo del punto 2 mediante modelos ARIMA. Para los pronósticos del primero se estiman mediante un modelo ARIMA(0,1,1) con deriva, los pronósticos del segundo mediante un ARIMA(1,0,1) con media cero, los del $k_{t,3}$ mediante ARIMA(0,0,0) con media cero. En el siguiente modelo, figura 2.26, vemos los pronósticos de cada uno de los coeficientes junto con sus intervalos deconfianza al 95 %.

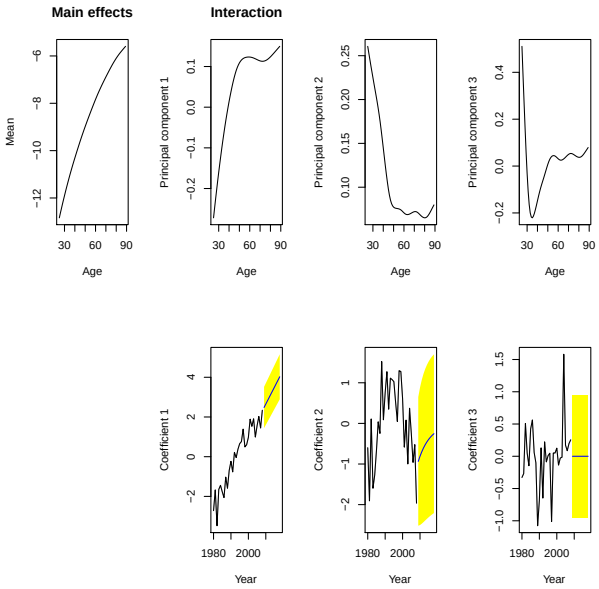

Figura 2.26: Modelo de predicción para los hombres españoles.

Los pronósticos del $k_{t,1}$ controlan, para los hombres mayores de 38 años un incremento continuo en la mortalidad en todo el periodo 2009-2018, y los intervalos de confian-

za presentan cierta incertidumbre permitiendo que las tasas aún sea más altas. Los pronósticos del $k_{t,2}$ muestran una tendencia ascendente para todas las edades en el periodo de predicción, con los cambios más fuertes producidos en los niveles de edad más jóvenes. Los correspondientes al tercer coeficiente, $k_{t,3}$ presentan un estacamiento durante todo el periodo ya que convergen a la media desde el año 2009. La amplitud de los intervalos de predicción de los 2 últimos es bastante amplia, lo que no nos asegura que las predicciones estimadas vayan a ser lo bastante precisas con lo que ocurra en un futuro.

Las predicciones resultantes para los años del periodo 1980-2008 se muestran en la figura 2.27. Éstas presentan un descenso leve a lo largo de todo el periodo de predicción para las edades entre los 25 y 33 años, ambos inclusive, sin embargo presentan un aumento también leve a lo largo de todo el periodo 2009-2018 para los hombres con edad mayor o igual a los 37. Para las edades 34, 35 y 36 las predicciones no presentan un aumento ni un descenso continuo a lo largo del periodo, sino que para los hombres de 34 años las predicciones aumentan hasta el año 2013 y disminuyen a partir de éste, las correspondientes a los hombres de 35 años aumentan hasta el año 2015 y después disminuyen, y las predicciones para los españoles de 36 años presentan un incremento hasta el 2016 y después un descenso.

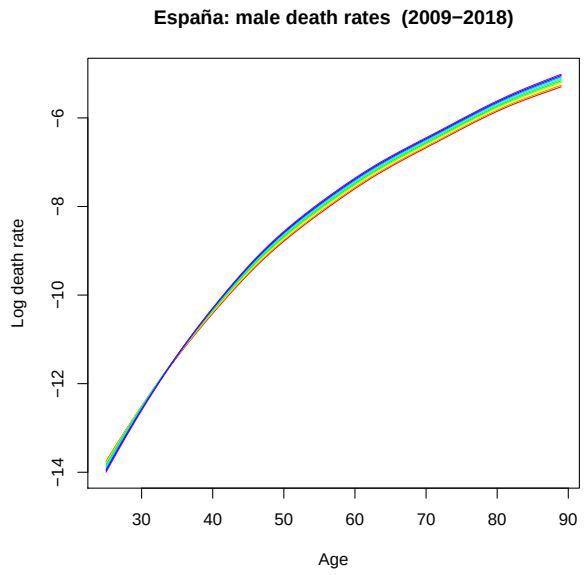

Figura 2.27: Predicciones de las tasas logarítmicas de la mortalidad para el periodo 2009-2018.

Se pronostican las tasas de mortalidad de las mujeres. Para ello se pronostican cada uno de los coeficientes del modelo, las puntuaciones de las 5 componentes principales, para datos funcionales mediante modelos de series de tiempo univariantes ARIMA. Tras ajustarle al $k_{t,1}$ un ARIMA(0,1,1), al $k_{t,2}$ un ARIMA(1,2,1), al $k_{t,3}$ un ARIMA(2,1,0), y a los correspondientes a los $k_{t,4}$ y $k_{t,5}$, un ARIMA(0,0,0) y un ARIMA(2,0,2) respectivamente, se obtiene el siguiente modelo, figura 2.28, con los pronósticos de los coeficientes $k_{t,j}$ para $j = 1, \ldots, 5$. Las previsiones del coeficiente

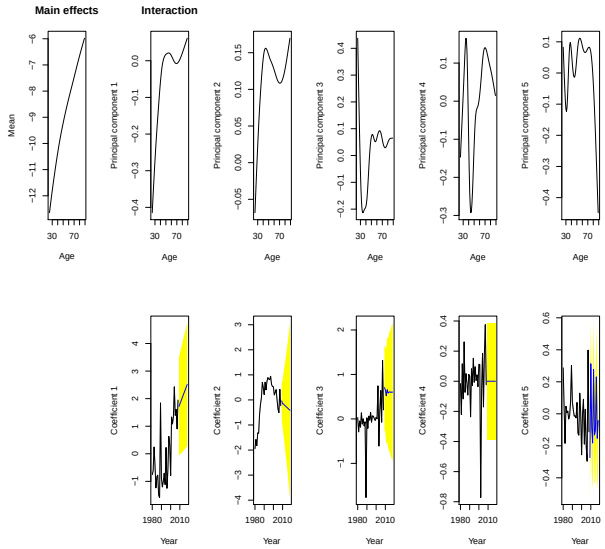

Figura 2.28: Modelo de predicción para las mujeres españolas.

$k_{t,1}$ para el periodo 2009-2018 muestran, para las mujeres con edades entre los 45 y los 64 ambos incluido y las de 75 y mayores, un incremento leve en la mortalidad a lo largo del periodo, y el ancho de los intervalos es bastante amplio. Los pronósticos del $k_{t,2}$ muestran un leve descenso en la mortalidad para las mujeres mayores de 28 años y los intervalos sugieren uan incertidumbre cada vez más alta a medida que avanzamos en las estimaciones de pasos futuros. Las previsiones del tercer coeficiente sugieren un oscilamiento en las tasas con un estancamiento posteriori, sobres todo en los niveles de edad más bajos, y el ancho de los intervalos son cada vez más amplios. El $k_{t,4}$ prevee un estacamiento en la media durante todo el periodo y unos intervalos de predicción también con amplitud constante, al contrario que los pronósticos y que los intervalos del ceficiente $k_{t,j}$, el cual es complejo y presenta un oscilamiento en las tasas y en los límites de los intervalos.

Las predicciones realizadas mediante el anterior modelo se muestran en la Figura 2.29:

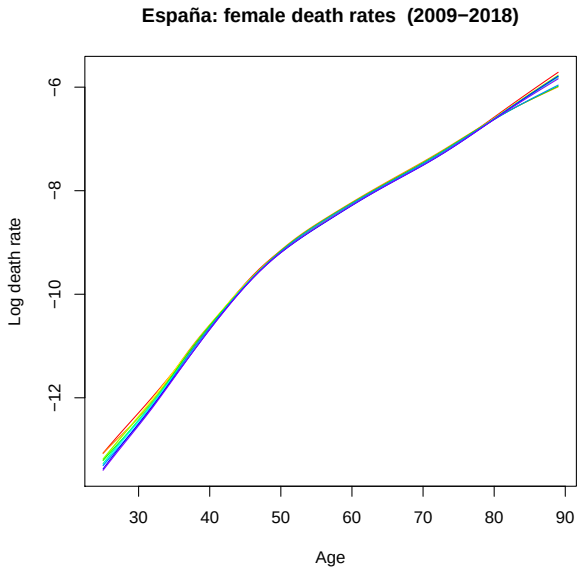

Figura 2.29: Predicciones para las mujeres españolas.

Las anteriores predicciones presentan un oscilamiento en la mortalidad en las tasas de mortalidad a medida que avanzamos a años futuros en todas las edades pero comportándose distinto según la edad. Por ejemplo, para los niveles de edad más alto tenemos que, para las españolas de 25, 26, 27 y 28 años disminuyen hasta el año 2012, en el cual aumentan y a partir de éste vuelven a disminuir. Para las de 29 años disminuyen durante todo el periodo excepto en los año 2011 y 2014 que padecen un leve aumento respecto al año anterior, para las edades entre los 30 y 33 ocurres lo mismo pero el aumento es los años 2011, 2014 y 2017. Para las de 35, 36 y 37 también ocurre lo mismo, disminuyen pero presentan un aumento, como un piquito en los años 2011 y 2015, para las de 35, en 2010 y 2016 para las de 36 y años y para las de 37 años en el 2010, 2013 y 2017. Para los grupos de edad más altos las subidas y bajadas continuas en las predicciones se pueden observar en el siguiente gráfico, en el cual se muestran junto con las reales del periodo 1980-2008 por grupo de edad quinquenales.

Vemos entonces que las predicciones para la población española se comportan distinto según el sexo, ya que en las correspondientes a los hombres sugiren un aumento en la mayoría de las edades, y sin embargo en las de las mujeres sugieren un oscilamiento y son menores aunque con una pequeña diferencia las del año 2018 que las del 2009.

2.3. Análisis comporativo.

2.3.1. Introducción.

Con la bases de datos de las tasas de mortalidad por cáncer para los hombres y mujeres en España, se evalúa cada uno de los métodos del capítulo con el objetivo de realizar un análisis comparativo. Para ello, se toma una base de datos de entrenamiento para predecir a 10 horizontes temporales. En particular, tal y como se presentó en la sección 3.1.2 del capítulo 3, se toman como entrenamiento los datos correspondientes al periodo 1960-2008 y como muestra test los correspondientes al 2000-2009, para las edades mayores de 25 años.

Para determinar que método predice mejor se comparan las predicciones con los datos observados, y se toma como mejor método aquel que posee menor error cuadrático medio, ECM, en todo el período de predicción.

$$ECM = \frac{1}{h}\sum_{t=1}^{h}\left(\frac{1}{p}\sum_{i=1}^{p}\left((m_{n+h}(x_i) - \hat{m}_{n+h|n}(x_i)\right)^2\right)$$

donde $\hat{m}_{n+h|n}(x_i)$ es la predicción de la tasa de mortalidad para la edad i en el horizonte h, y $m_{n+h}(x_i)$ es tasa observada en la edad i en el año $n+h$.

No se presentan los gráficos del modelo ni los del modelo de predicción, puesto que lo que interesa es un análisis comparativo. Notar que, en esta parte del estudio también trabajaremos con la transformación logarítmica de las tasas.

2.3.2. Resultados

El ajuste del modelo Lee-Carter para el caso de las tasas logarítmicas de la mortalidad por cáncer de los hombres españoles mayores de 25 años en el periodo 1980-2008 explica un 64,9 % de la variabilidad en torno a la media y muestra un descenso de las tasas logarítmicas para los hombres mayores de 36 años a lo largo de todo el período de entrenamiento.

Para obtener el camino de predicción sobre el periodo 2000-2009, se ajusta la evolución del índice pronostica kt mediante un camino aleatorio con deriva. Este ajuste conduce a un modelo de predicción que, como era de esperar, prevé un moderado descenso continuado de la mortalidad. La amplitud creciente de los intervalos de predicción al 95 % delata como crece el grado de incertidumbre de las predicciones a medida que se prolonga el horizonte de predicción.

La Figura 2.30, proporcionan una representación simultánea de los valores realmente

observados y de las predicciones del modelo de Lee-Carter con sus intervalos de confianza al 95 % para los años 2000, 2005 y 2009.

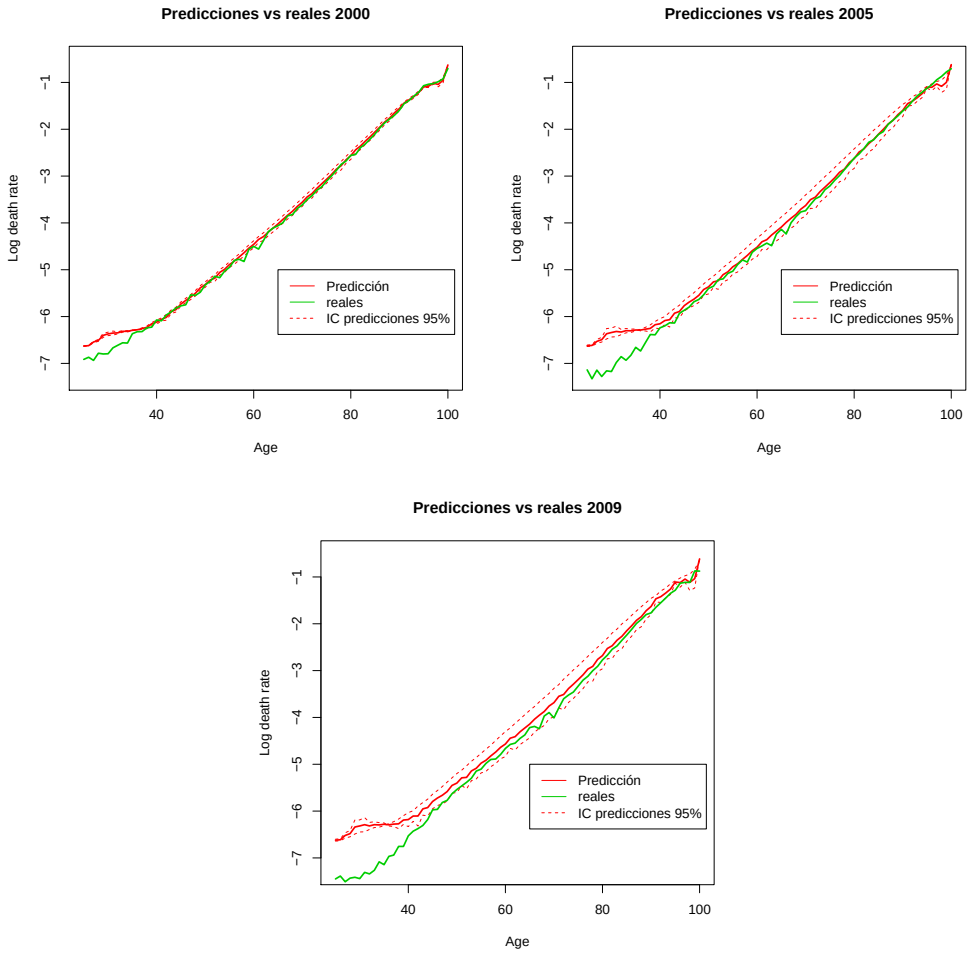

Figura 2.30: Predicciones e IC al 95 %de Lee-Carter para el cáncer de los hombres junto con los datos observados de los años 2000,2005,2009.

Nótese que las predicciones no son buenas para edades inferiores a los 42 años dado que en ninguno de los tres años considerados en la Figura 2.30 las tasas logarítmicas predichas por el modelo caen dentro del intervalo de predicción al 95 %.

También se observa que a medida que se predice a horizontes más largos, los pronósticos son peores ya que se alejan más de lo real. Se comete entonces un ECM global al predecir de $ECM = 0{,}07609832$

En el caso de las mujeres el modelo ajustado de Lee-Carter explica un $89,9\,\%$ de la

variabilidad existente en los datos. EL estimador de b_x capta lo que ocurre en todas la edades. Y, al igual que con los hombres, la evolución de los índices ajustados para k_t es monótono decreciente durante todo el período de predicción.

La Figura 2.31 muestra como, en este caso, los intervalos de predicción al 95 % de confianza envuelven en general a las tasas reales, a diferencia de lo que ocurría con los intervalos obtenidos para los hombres, que mostraban ciertamente resultados muy pobres para edades bajas y en los tres horizontes de predicción considerados en la Figura 2.30.

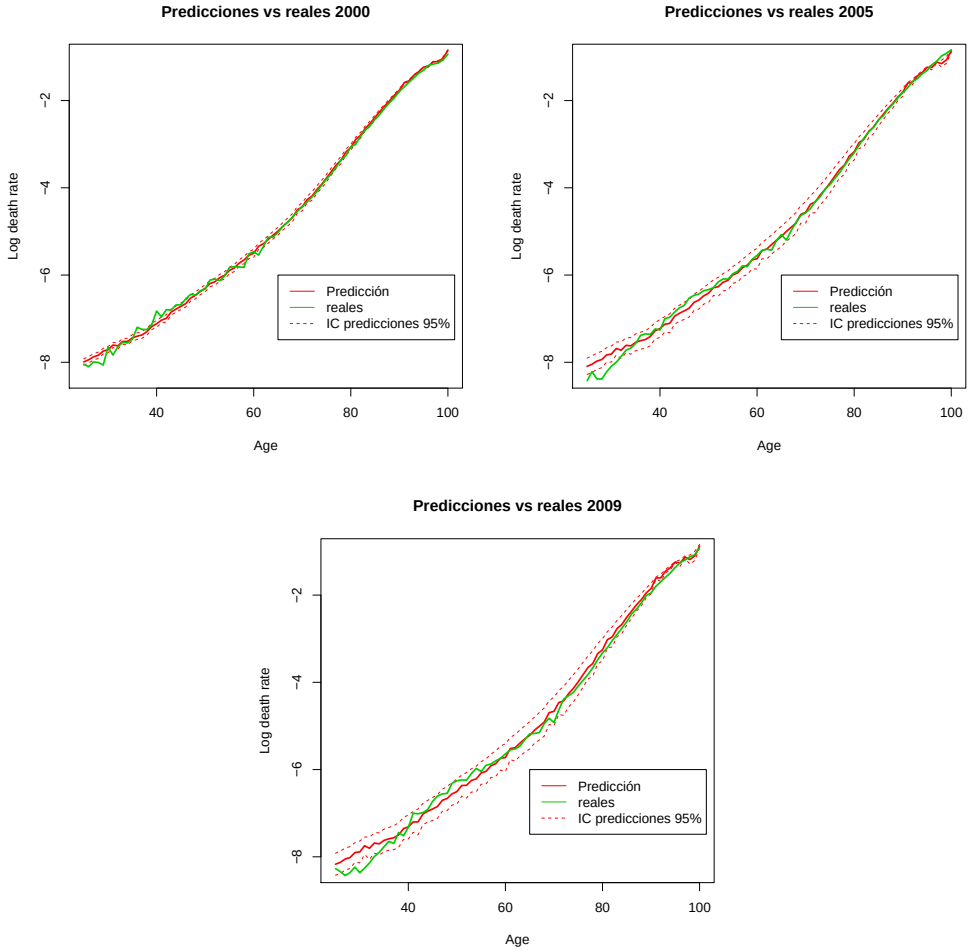

Figura 2.31: Predicciones e IC al 95 % de Lee-Carter de las mujeres junto con los datos observados de los años 2000,2005,2009.

El ECM global cometido al predecir las tasas logarítmicas de las mujeres para el período 2000-2009 es $ECM = 0{,}01788829$, menor que en el caso de los hombres, demostrando lo

mencionado anteriormente.

Al estimar el modelo mediante Lee-Miller en el caso de los hombres, al igual que con Lee-Carter, se explica un 64,9 % de la variabilidad existente en las tasas logarítmicas, se predice el coeficiente k_t, y se obtiene un k_t monótono decreciente a lo largo de todo el período de predicción, lo que provoca que las predicciones de las tasas logarímicas sean cada vez menores, y como es obvio, con mayor incertidumbre a medida que pasan los años.

A la vista de la Figuras 2.32, el modelo de Lee-Miller para las hombres predice mejor que el método de Lee-Carter, ya que como se puede ver, las tasas predichas para las edades inferiores están más próximas a la reales. Las tasas predichas para las edades inferiores

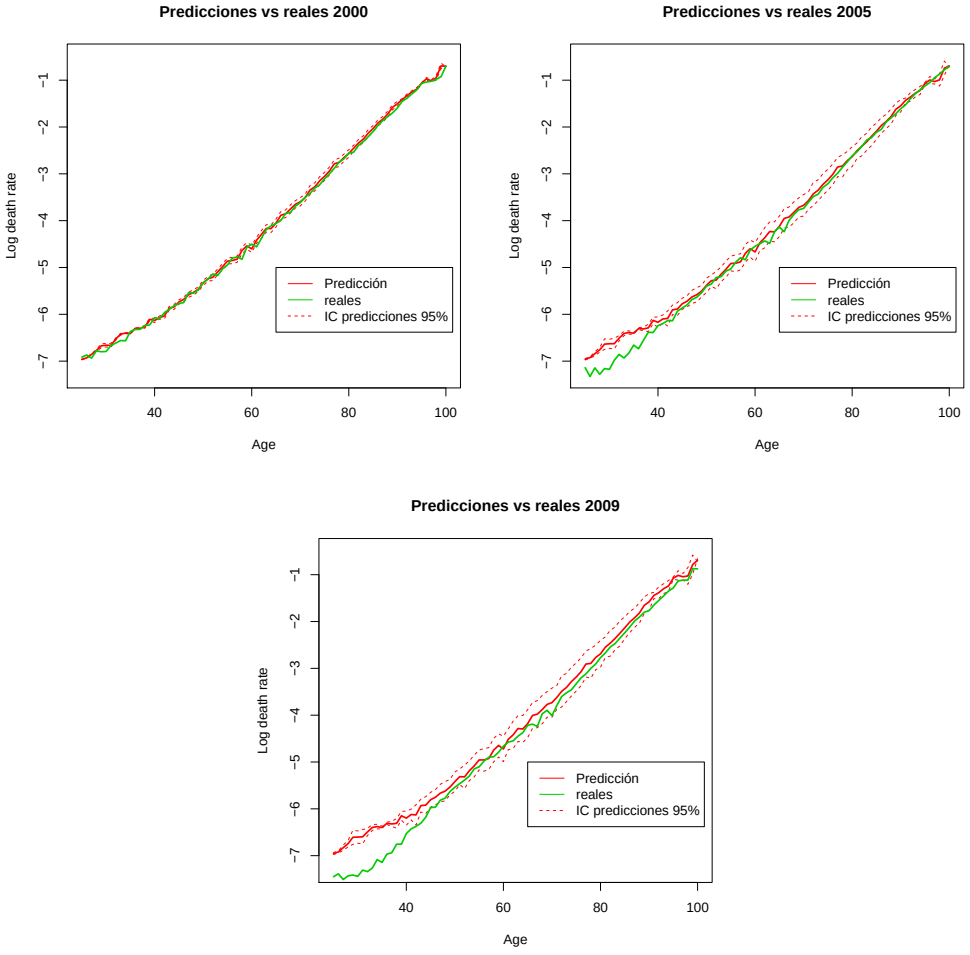

Figura 2.32: Predicciones e IC al 95 %de Lee-Miller para el cáncer de los hombres españoles junto con los datos observados de los años 2000,2005,2009.

siguen lejos de las reales pero mejoran sustancialmente su comportamiento con respecto a las predicciones del método de Lee-Carter.

Los errores cuadráticos medios de predicción con el método de Lee-Miller se recogen en la segunda fila de las Tabla 2.1 y 2.2. Corroboran la clara mejoría que el método de Lee-Miller reporta en este caso con respecto al método de Lee-Carter. Sin embargo, es claro que el modelo predictivo sigue sin funcionar adecuadamente para el medio plazo, como delata el mal comportamiento de los intervalos de predicción para los años 2005 y 2009. Más aún, la mejora del ECM respecto a Lee-Carter se aminora sensiblemente a medida que crece el horizonte de predicción.

El ECM global para el período 2000-2009 con el método de Lee-Miller es de un $ECM = 0{,}04186873$, inferior al 0,07609832 que se obtuvo con el método de Lee-Carter.

Al igual que Lee-Carter, el ajuste del modelo para las tasas de las mujeres explica un 89,9 % de la variabilidad, y los pronósticos de k_t, tras ajustar un ARIMA, controlan un descenso continuo de las tasas logarítmicas en todas las edades. Se obtienen entonces las predicciones a 10 horizontes futuros.

A la vista de la figura 2.33 Lee-Miller predice mejor las tasas de las mujeres que las de los hombres, y esto puede ser debido a que el modelo explica un 25 % más de la variabilidad existente en las tasas, siendo los pronósticos mejores para las mayores de 60 años aproximadamente. Además se puede intuir que, a diferencia de Lee-Carter y de que lo que ocurría con los hombres, las predicciones no empeoran año a año.

El método de Lee-Miller vuelve a mejorar al de Lee-Carter a la hora de predecir la mortalidad de las mujeres españolas, ya que se comente un $ECM = 0{,}01667496$ al pronósticar en todo el período 2000-2009, auque en este caso lo mejora en menor medida ya que la diferencia entre los ECM es muy pequeña, de un 0,001213323.

Se procede a ajustar el modelo mediante el método de Método de Booth-Maindonald-Smith, a las tasas por cáncer de los hombres. El método BMS consigue explicar un 63,3 % de la variabilidad, un 1,6 % menos que los métodos anteriores, y mencionar que el K_t controla una tendencia descendente de las tasas en el período 1980-2008 correspondientes a los hombres mayores de 30 años y para los que no tienen 79 años.

Los pronósticos de k_t mediante un modelo ARIMA sugieren un decrecimiento continuo en las tasas para los hombres con las anteriores edades, se obtienen así las predicciones. Se muestran las correspondientes a los años 2000, 2005 y 2009 en la Figura 2.34 que, comparándolas con las tasas reales, las predicciones presentan un distanciamento destacado en las tasas de las edades más bajas, aproximadamente las correspondientes a los menores

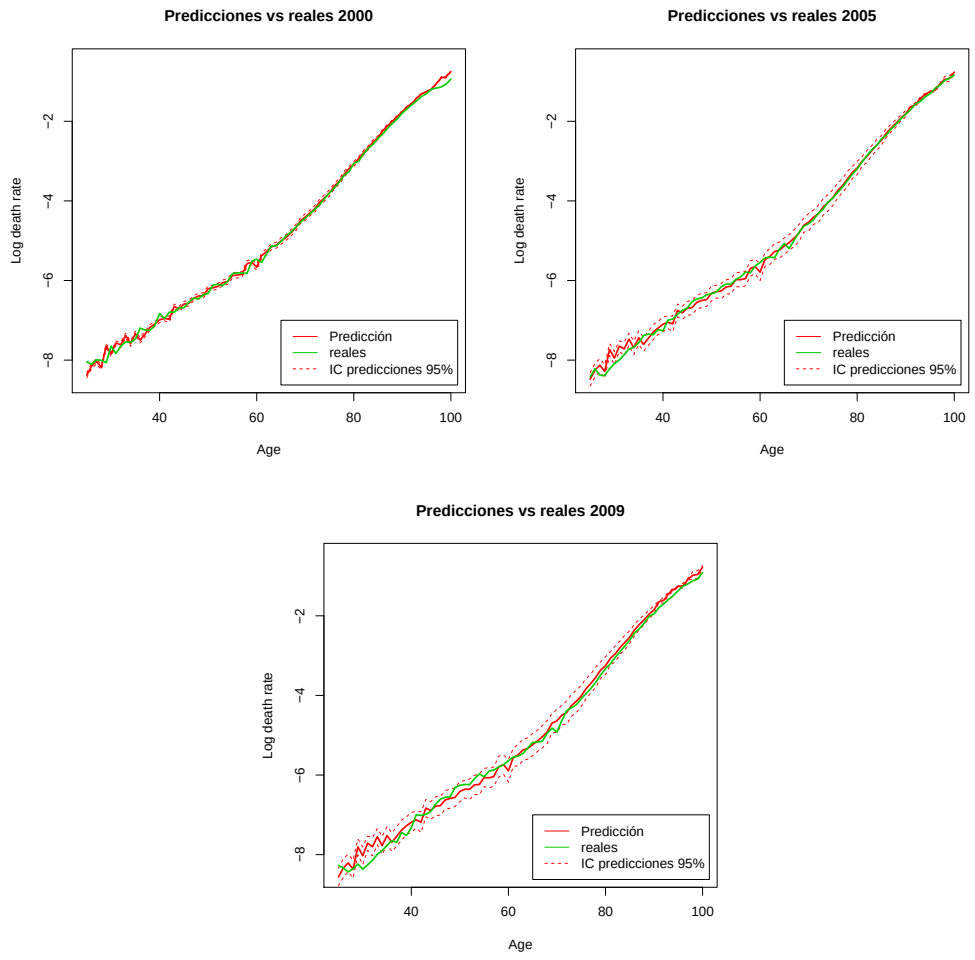

Figura 2.33: Predicciones e IC al 95 %de Lee-Miller para el cáncer de las mujeres junto con los datos observadas de los años 2000,2005,2009.

de 40 años.

Las predicciones cometen un mayor error a medida que se avanza en el período 2000-2009, siendo éste más pronunciado en las edades más bajas, ya que las tasas logarítmicas de los hombres de estas edades se salen de los intervalos de predicción al 95 %, en los 3 años siendo mayor la distancia en el año 2009. Y se puede ver que son sustancialmente peores que el anterior modelo de acuerdo con los ECM que se muestran en la tercera fila de las tablas 2.1 y 2.2. El ECM cometido es elevado, $ECM = 0{,}1185866$ y es considerablemente mayor que el comotido con Lee-Miller de 0,04186873

En el caso de las mujeres , tras ajustar el modelo, que explica menos variabilidad que los

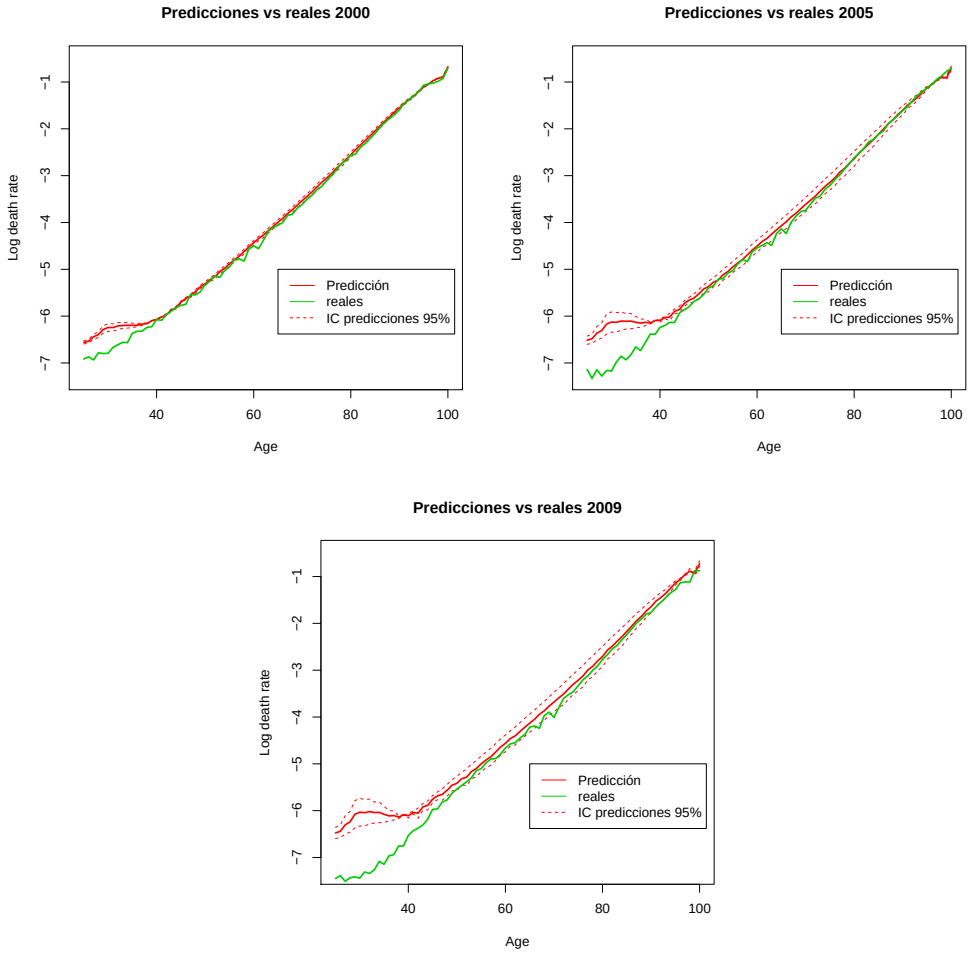

Figura 2.34: Predicciones e IC al 95 %de BMS para el cáncer de los hombres españoles junto con los datos observados de los años 2000,2005,2009.

anteriores métodos un 67,7 %, y tras proyectar los coeficientes k_t con un modelo ARIMA, los coeficientes pronosticados presentan, para las mujeres mayores de los 33 años, excepto para las de 97, 98 y 99 años, una tendencia descendente para el logaritmo de las tasas.

En las Figuras 2.35, se muestran las comparaciones entre las predicciones y las tasas en los años 2000, 2005 y 2009. Los intervalos de predicción para el año 2000, son muy pequeños, de tal manera que no se llegan casi ni apreciar. También se puede ver que en los 3 años, los intervalos son muy estrechos sobretodo en los extremos de las edades, sin embargo las predicciones son malas aunque mejores que las de los hombres.

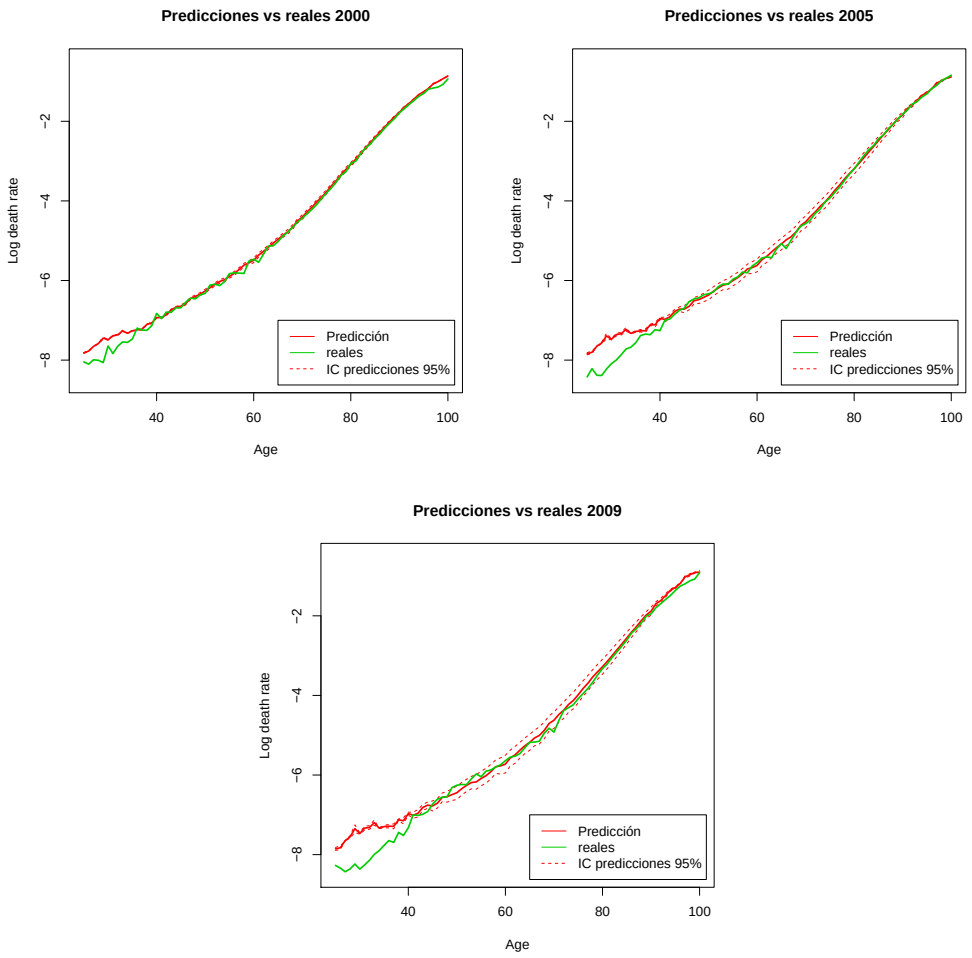

Figura 2.35: Predicciones e IC al 95 %de BMS para las mujeres españolas junto con los datos observados de los años 2000,2005,2009.

Efectivamente, el método predice bastante mejor las tasas por cáncer de las mujeres ya que el $ECM = 0{,}05426091$, un 0,06432565 menor que antes. Pero, al igual que lo ocurrido con los hombres, el método BMS sigue siendo que los anteriores.

Para la evaluación del método, primero se suavizan las tasas por cáncer para ambos sexos mediante una suavización spline penalizada y ponderada, tomando como criterio que las tasas son monótonas a partir de los 50 años, y como pesos las inversa de la varianza empírica.

Para poder realizar el ajuste del modelo, tenemos que seleccionar el número de componentes principales funcionales a utilizar. Se eligen 3 componentes principales ya que con

esta número se minimiza el MISFE.

Se ajusta entonces el modelo con 3 componentes principales obteniendo que la 1ª, la 2ª y la 3ª componentes principales explican un 74,2 %,22,2 % y 2,3 % de la variabilidad existente en los datos, respectivamente, llegando a explicar juntas un 98,7 %, un porcentaje mucho más elevado que con los demás métodos anteriores.

Se pronostican los coeficientes de las 3 componentes $k_{t,j}$ para $j = 1, 2, 3$ con modelos ARIMA, concretamente con un ARIMA(1,1,0), un ARIMA(1,0,2) y un ARIMA(2,0,2) respectivamente, y se obtiene que los pronósticos de las puntuaciones de la primera componente pricipal, que controla las tasas logarímicas para los hombres mayores de 36 años, sufren un descenso a lo largo de todo el periodo 2000-2009. Sin embargo, el correspondiente a la segunda sugiere un aumento continuo de las tasas para edades comprendidas entre los 76 y los 89, ambos inclusive. Y las predicciones del coeficiente de la tercera, la cual modeliza las tasas para los hombres con edades entre los 25 y los 30, entre los 47 y 71, y entre 77 y 87; sugiere un aumento hasta el año 2007 junto con un posterior descenso.

Se obtienen entonces las predicciones, y se presentan en las Figuras 2.36 junto con las tasas reales, no con las suavizadas, las correspondientes a los años 2000, 2005 y 2009.

El método de Hyndman también comete mayor error al predecir las tasas de los hombres más jóvenes. Se puede ver como las prediciones de Hyndman pierden credibilidad a mediano-largo plazo. De hecho, se puede ver que en el año 2000 produce buenas predicciones y en el año 2005 se escapan de lo real en edades inferiores. Obteniendo así, un ECM global de $ECM = 0{,}06825377$, un 0,02638504 menor que el de Lee-Miller.

En el caso de las mujeres se ajusta el modelo de Hyndman-Ullah con 4 componentes principales, número de funciones básicas que minimizan el MISFE, y se proyectan a 10 horizontes los coeficientes de dichas componentes. Obteniendo que las predicciones del coeficiente de la primera componente son monótonas decrecientes a lo largo del período 2000-2009 para todas las edades. Las puntuaciones de la segunda, que controla las tasas para las mujeres con edades comprendidas ente los 45 y los 97, predicen un aumento en las tasas hasta el año 2004 junto con un decremento hasta el año 2009. Las predicciones del coeficiente de la tercera componente presentan un aumento a lo largo de todo el período de predicción, en edades comprendidas entre los 50 y los 94, ambos inclusive. Y los de la cuarta componente principal producen un aumento del 2000 al 2001, con un posterior descenso continuo, para las edades más jóvenes y más mayores, y para los 50 y 51 años.

Para los años 2000, 2005 y 2009 se presentan, Figura 2.37, las predicciones frente a los datos reales. Éstas se distancian menos, que las de los hombres, de las reales, en especial para los mujeres más jóvenes.

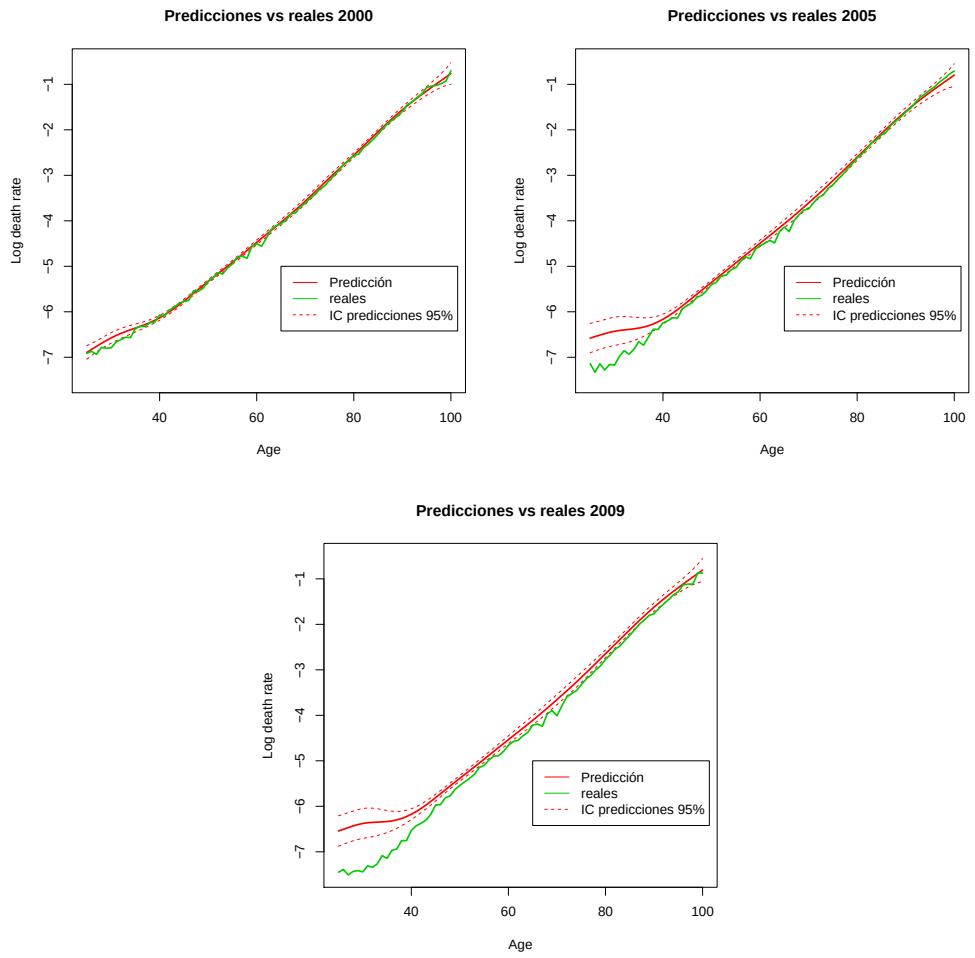

Figura 2.36: Predicciones e IC al 95 %de Hyndman-Ullah para hombres junto con los datos observados de los años 2000,2005,2009.

El método proporciona buenas estimaciones a corto plazo, llegando incluso, a ser el que mejor predicciones aporta a largo plazo. Por lo que el ECM global no es de 0,01240688, el más bajo hasta ahora.

A continuación, se evalúan los métodos no paramétricos autorregresivos funcionales.

Esta parte del estudio se realiza con funciones que se encuentran en la librería **fda.usc**, la cual contiene funciones que llevan a cabo un análisis exploratorio y descriptivo de los datos funcionales explorando en su mayor parte características importantes, tales como mediciones de la profundidad o de detección de valores atípicos funcionales. También contiene funciones para ajustar modelos de regresión, para hacer predicciones, para métodos

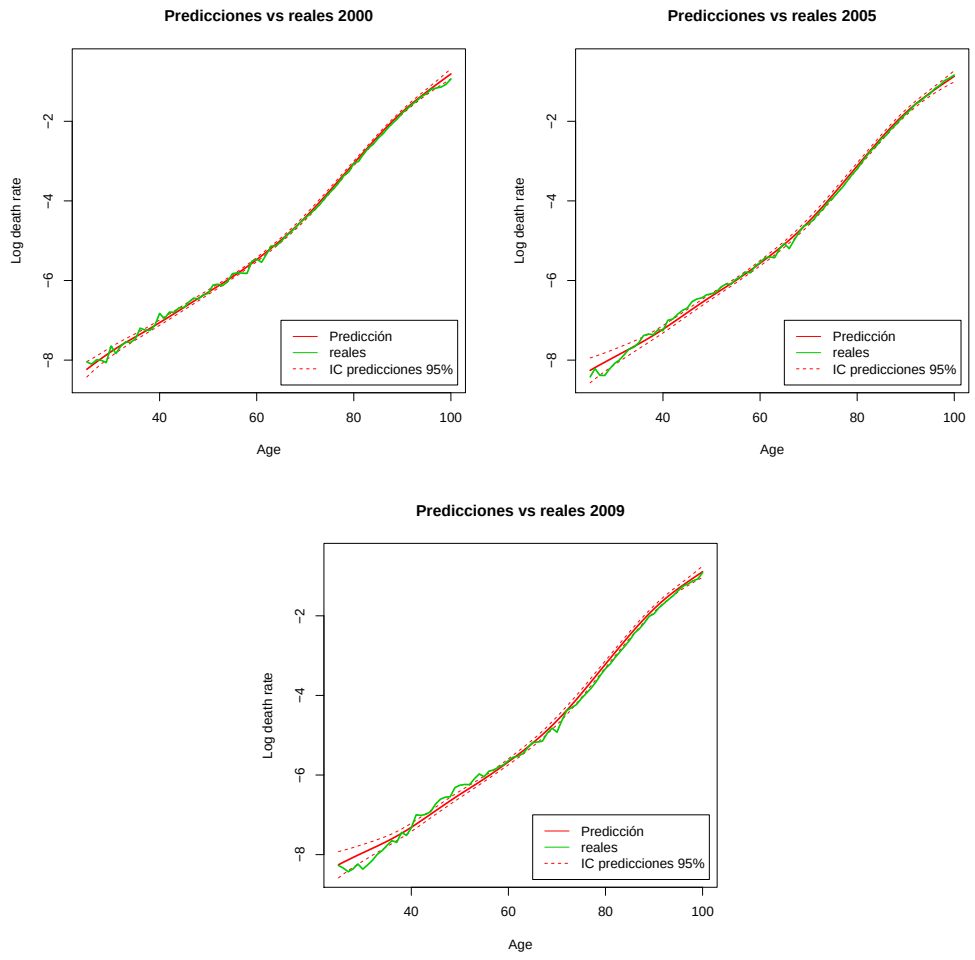

Figura 2.37: Predicciones e IC al 95 %de Hyndman-Ullah las mujeres españolas junto con los datos observados de los años 2000,2005,2009.

para la clasificación, entre otras.

Aunque en esta parte del estudio no es necesario que las curvas sean suaves, tal y como se mencionó en la subsección 3.2.1., se sigue trabajando con la transformación logarítmica de las tasas de la mortalidad.

Primero se estima un modelo de regresión no paramétrico con variable explicativa funcional, las tasas para el período 1980-1998, y con variable independiente escalar las tasas a la edad x para los años 1981-1999. Se ajustan tantos modelos como edades, en nuestro caso, 76 regresiones. Se obtiene así una estimación de la regresión para cada edad, que se realiza con la función **fregre.np.cv**, mediante el estimador de Nadaraya-Watson, se elige

como Kernel el normal asimétrico y la ventana se selecciona mediante validación cruzada.

Ahora, en base a las regresiones estimadas, se obtienen predicciones para un año más y para cada edad. Por lo que también se tienen que realizar tantas predicciones como edades. Y como se quiere predecir a 10 horizontes futuros tendremos que realizar 10∗76 predicciones, que se computan mediante la función **predict.fregre.fd**.

Para predecir a un primer horizonte, esto es, obtener pronósticos para el año 2000, a la regresión estimada se le mete como variable explicativa los datos reales del año 1999, obteniendo entonces las predicciones para el siguiente año.

Para predecir tasas logarítmicas para el segundo año del período de predicción, se realizan otra vez tantas regresiones como edades, pues aunque se mantiene la variable explicativa, la dependiente escalar varía, ahora son las tasas correspondientes para cada edad x pero para los años 1962-1999.

Para predecir, como se tiene estimada una regresión no paramétrica autorregresiva, evaluamos en la regresión los datos predichos en el horizonte anterior, esto es, los predichos para el año 2000, y se obtienen así predicciones para un año más, año 2001.

Siguiendo el anterior proceso se obtienen estimaciones de los 10 datos futuros. La comparación de las tasas reales con las predicciones para el período 2000-2009 para los hombres, se presenta en la figura 2.38.

Las diferencias entre las proyecciones y los datos reales son considerables, siendo mayor en las edades más jóvenes. Las predicciones para el año 2005 como las del 2009 son prácticamente las mismas excepto en edades bajas. Se ve que el modelo no es bueno a largo plazo.

Las prediciones son bastante malas, ya a corto plazo, véase fila 5 de la tablas 2.1 y 2.2, por lo que se tiene un $ECM = 0{,}1163177$ global muy elevado. Lo que indica que este método no es lo bueno para compertir con anteriores excepto con el modelo de BMS.

A continuación, se predicen las tasas logarítmicas para las mujeres. Se sigue el mismo procedimiento que con los hombres, y también se elige un Kernel asimétrico y el método de validación cruzada como selector de la ventana c.

Las predicciones para las mujeres son mejores que las de los hombres, véase Figura 2.39, y como ocurría con ellos se comete un error mayor en los valores más pequeños del rango de la edad, y aunque no se aprecie en el gráfico, las predicciones para estas edades son cada vez peores, pues las reales presentan un descenso y el método prevee un aumento, aunque ligero.

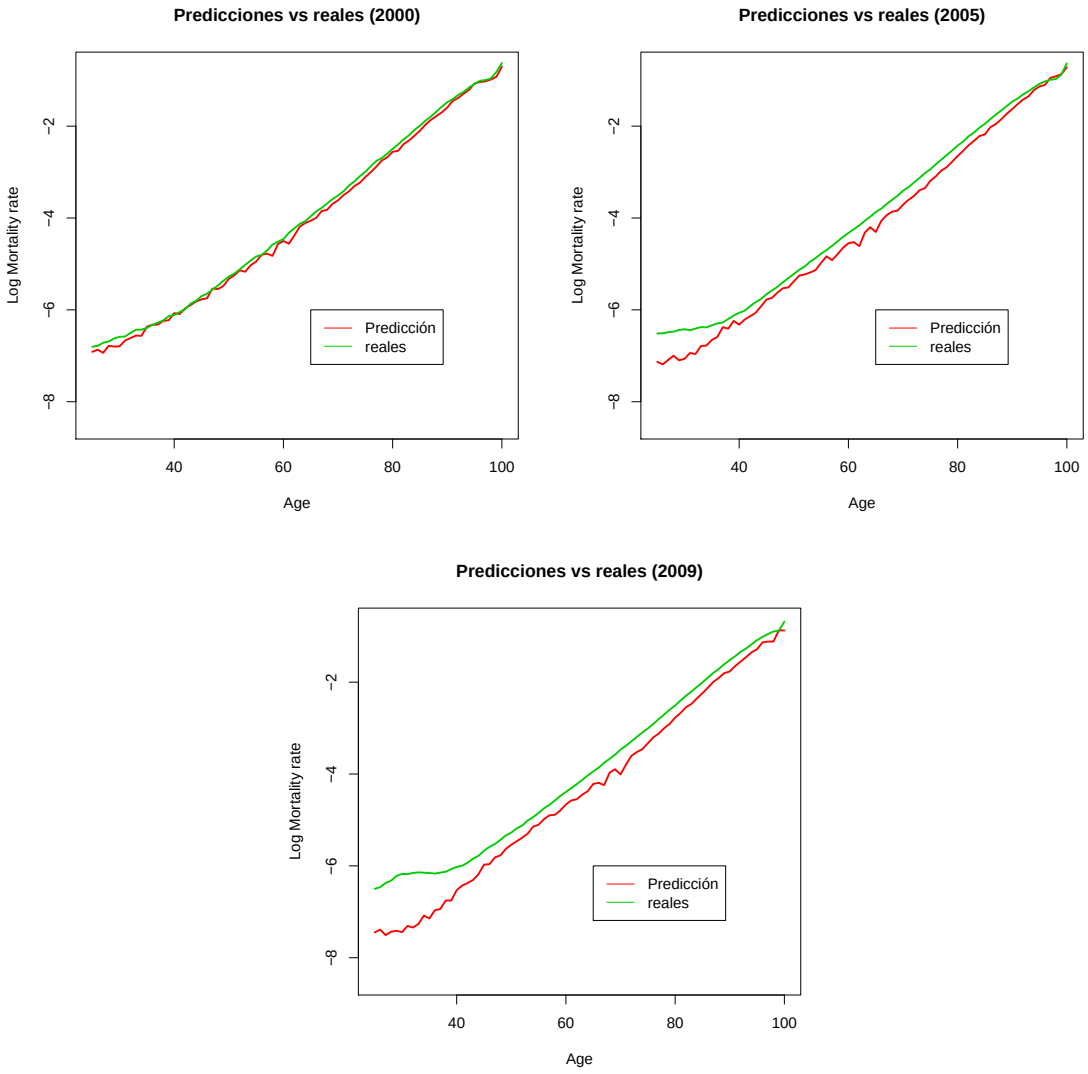

Figura 2.38: Predicciones mediante un método no paramétrico autorregresivo funcional con respuesta escalar para el cáncer de los hombres españoles junto con los datos observados de los años 2000,2005,2009.

Las predicciones excepto en edades pequeñas tienden a ser constantes, convergen a la media, provocando así mayores errores a largo plazo.

Se comete un $ECM = 0{,}08604693$ global, así que el método comete menor error al predecir las tasas de la mujeres. No obstante al comporarlo con Hyndman-Ullah, de un 0,01240688, se puede apreciar que tiene una exactitud muy inferior.

Se estudia el modelo recursivo de regresión no paramétrico autorregresivo funcional con

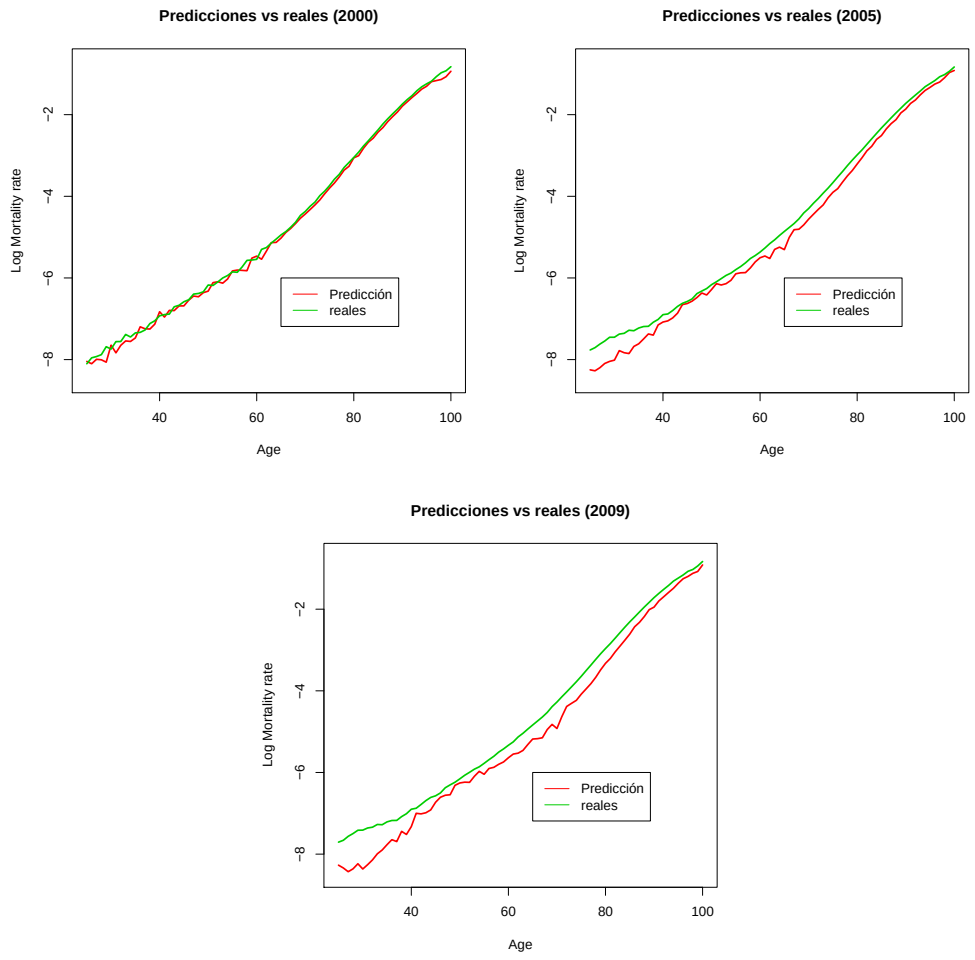

Figura 2.39: Predicciones para el cáncer de las mujeres junto con los datos reales de los años 2000, 2005, 2009.

respuesta respuesta escalar. La diferencia con respecto, al anterior método, es que la variable independiente del modelo de regresión se actualiza a medida que se avanza en período de predicción.

Por tanto, en este estudio, para predecir a un año, se estima de manera no paramétrica la regresión funcional como el anterior método. Pero para predecir a 2 años, se actualiza la muestra. Se prescinde del año 1960, y se introducen las tasas del año 1999, y se toma como variable respuesta las tasas predichas mediante la anterior regresión. Se obtiene, así las tasas para el año 2001.

Ahora la variable independiente será los datos observados del 1963-1999 y se le añade

la predicción para el año 2000, y se toma como dependiente para estimar la regresión las predicciones para el año 2001.

Y así sucesivamente hasta predecir los h años futuros. La Figura 2.40 muestra un sesgo en las predicciones de los hombres, siendo más preciso en edades próximas a los 100 años. Las tasas predichas tienden a ser constantes, y mejoran a las realizadas con el método

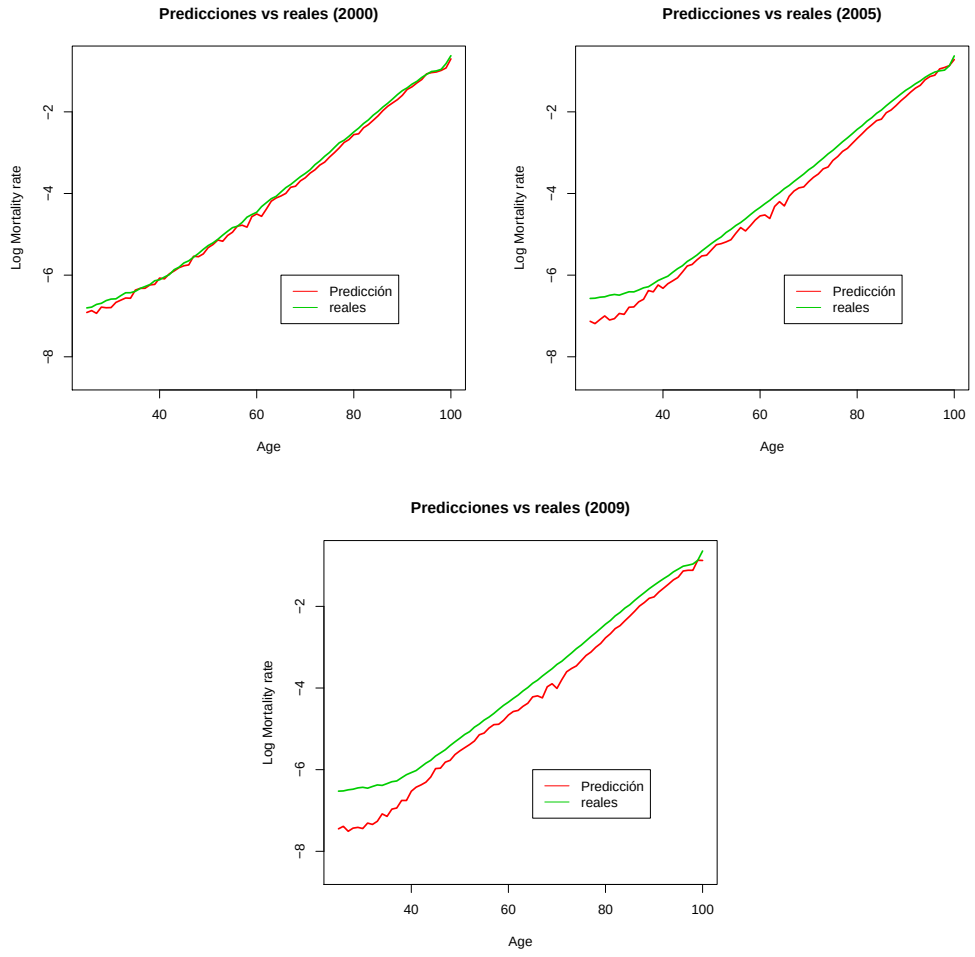

Figura 2.40: Predicciones mediante un método recursivo no paramétrico autorregresivo funcional con respuesta escalar para los hombres junto con los datos observados de los años 2000,2005,2009.

anterior. Fijándose en la penúltima fila de la tabla, se ve que las predicciones son bastante malas en general, lo que provoca que el ECM global sea de 0,05216445, mejorando en un 0,06415328 el ECM respecto al no recursivo.

Se estudian a continuación las predicciones relativas a las mujeres. Y se muestra que casi no se mueven, inclusive en edades bajas, Figura 2.41. Las predicciones se comportan de la

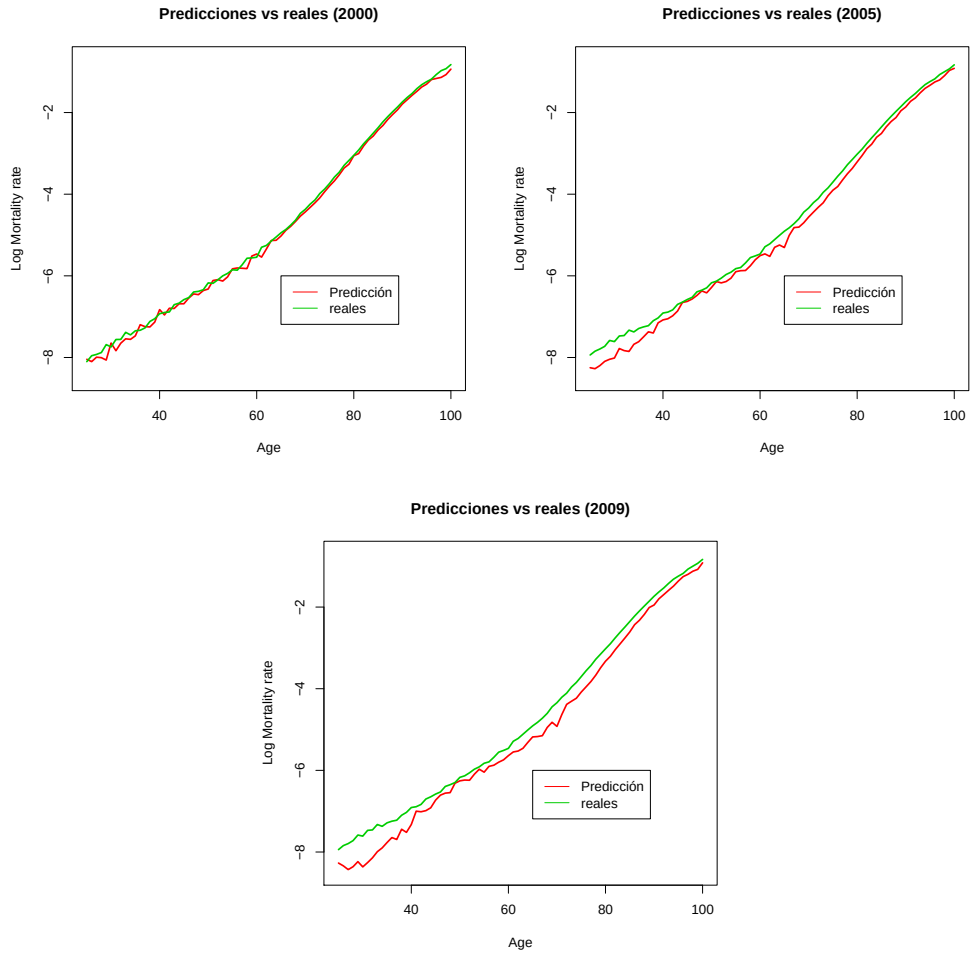

Figura 2.41: Predicciones mediante un método recursivo no paramétrico autorregresivo funcional con respuesta escalar por cáncer para mujeres españolas junto con los datos observados de los años 2000,2005,2009.

misma manera para todas las edades tanto en el año 2005 como en el 2009, y se diferencian poco de las del año 2000, sugiriendo un estancamiento.

La tabla anterior muestra que las predicciones son mejores que las de los hombres, aunque tampoco son buenas en este caso, cometiendo un ECM global de 0,03373216, más que con los método paramátricos excepto Bootstrap.

A continuación se estudia un modelo no paramétrico autorregresivo funcional con re-

spuesta funcional. Para ello se ajusta una regresión no paramétrica con variable independiente funcional, los tasas logarítmicas de mortalidad para el período 1960-1998, y con variable dependiente funcional, los datos correspondientes a los años 1961-1999. Se realiza mediante el estimador de Nadaraya-Watson y se selecciona de manera local el número óptimo de vecinos mediante un procedimiento de validación cruzada. Se elige la semimétrica de la derivada y el Kernel cuadrático.

Como el número de vecinos se selecciona de manera local, para elegir el número óptimo, la función tiene como argumento el número de años para el que se realiza el ajuste, dejando el resto para testar. Notar que, que el número de años debe de ser elevado, ya que no tiene sentido tomar pocos datos cuando se espera estimar un número mucho mayor, como por ejemplo, tomar 20 años (se tienen 40 datos.)

Para realizar el estudio se tiene que crear una muetra de entrenamiento y una muestra test. La muestra de entrenamiento contiene en la primeras 76 columnas (edades en estudio), la variable respuesta, los datos 1961-1999, y en las siguientes 76 columnas la varible independiente, los datos 1960-1998, y es con la que se estima la regresión.

La muestra test recoge en las primeras 76 columnas los datos del período 1999-2008, a partir de los cuales, mediante la regresión estimada, se obtienen las predicciones para el período 2000-2009; y en las siguientes 76 columnas las tasas reales del período de predicción para testar las predicciones.

Las predicciones para las tasas logarítmicas por cáncer de los hombres son mayores que las tasas reales para todas las edades, y al igual que los métodos con respuesta escalar muestran un estacamiento, Figura 2.42. Las predicciones, no se mueven, mientras que las tasas reales muestran un descenso, así que a medio plazo ya se comportan mal. Se comete un ECM global de 0,05216445. Lo que indica que es el segundo que mejor va, después de Lee-Miller.

En las Figura 2.43, se presentan la comparación de las predicciones con las tasas reales de las mujeres para el año 2000, 2005 y 2009. En la misma vemos que el método no comete mucho sesgo al predecir a un horizonte, mientras que para los otros 2 años no proporciona buenos pronósticos. Las predicciones tienden a ser constantes. El ECM global es de 0,03512471.

Para facilitar el análisis comparitivo se muestran a continuación las Figuras 2.44 y 2.45 que recogen la información conjunta, comparan las predicciones de todos los métodos frente a los datos reales para los hombres y las mujeres, respectivamente.

La Figura 2.44 muestra como todos los métodos van mal a largo plazo. En el año 2000

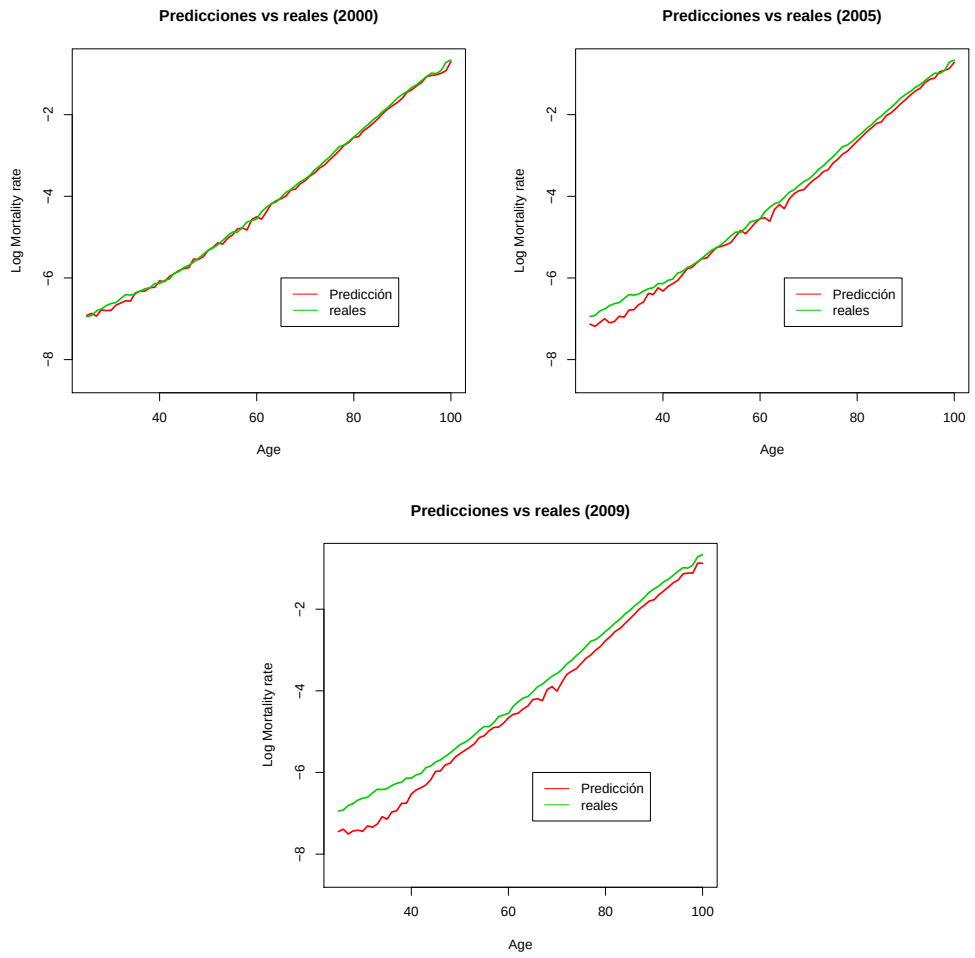

Figura 2.42: Predicciones junto con los datos observados de los hombres en los años 2000,2005,2009.

predicen valores similares excepto el método de Lee-Carter y el BMS que ya cometen un sesgo considerable en edades inferiores. Los métodos predicen tasas mayores que las reales, siendo la diferencia más alta a largo plazo, donde todos los métodos cometen un error elevado para edades bajas. Esto puede ser debido a que la estimación de b_x no es buena en estas edades.

Los métodos que proporcionan mejores resultados son el no paramétrico funcional con respuesta funcional y el método de Lee-Miller. Las tasas predichas por estos métodos tienden a ser constantes desde un corto plazo.

Mientras las correspondientes a los métodos BMS y autorregresivo con respuesta escalar,

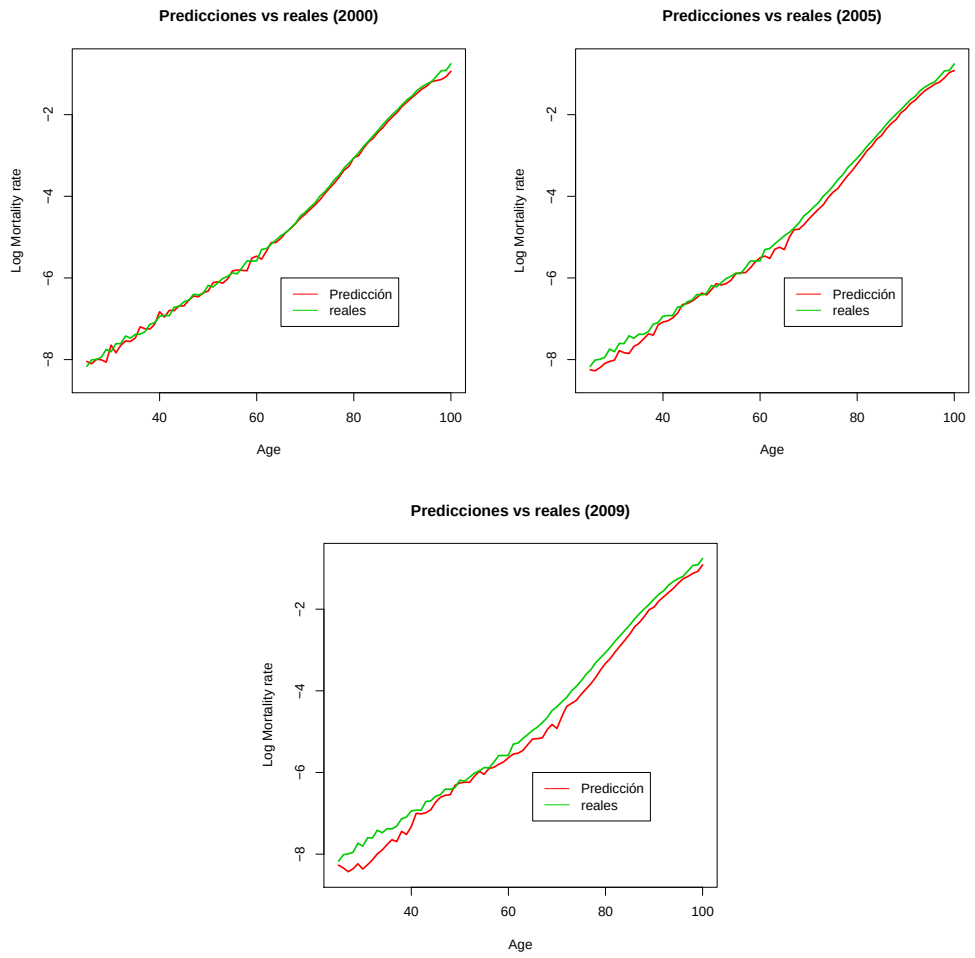

Figura 2.43: Predicciones junto con los datos observados de los mujeres en los años 2000,2005,2009.

para edades bajas, son los que muestran los resultados más insatisfactorios.

En la Figura 2.45 se presenta la comparación para el caso de las mujeres. Los métodos, como ya se ha mencionado, predicen mejor las tasas que en el caso de hombres, especialmente en edades inferiores. Los métodos no paramétricos junto con BMS son los que muestran peores resultados. A medio-corto plazo producen estimaciones insactisfactorias. La diferencia de las curvas respecto a las de los método no paramétricos, exceptuando BMS, es clara, y la predicciones tienden a estancarse.

En las siguientes tablas se presentan los ECM cometidos al predecir por cada método cada uno de los años del período de predicción. Las 2.1 y 2.2 se corresponden con los

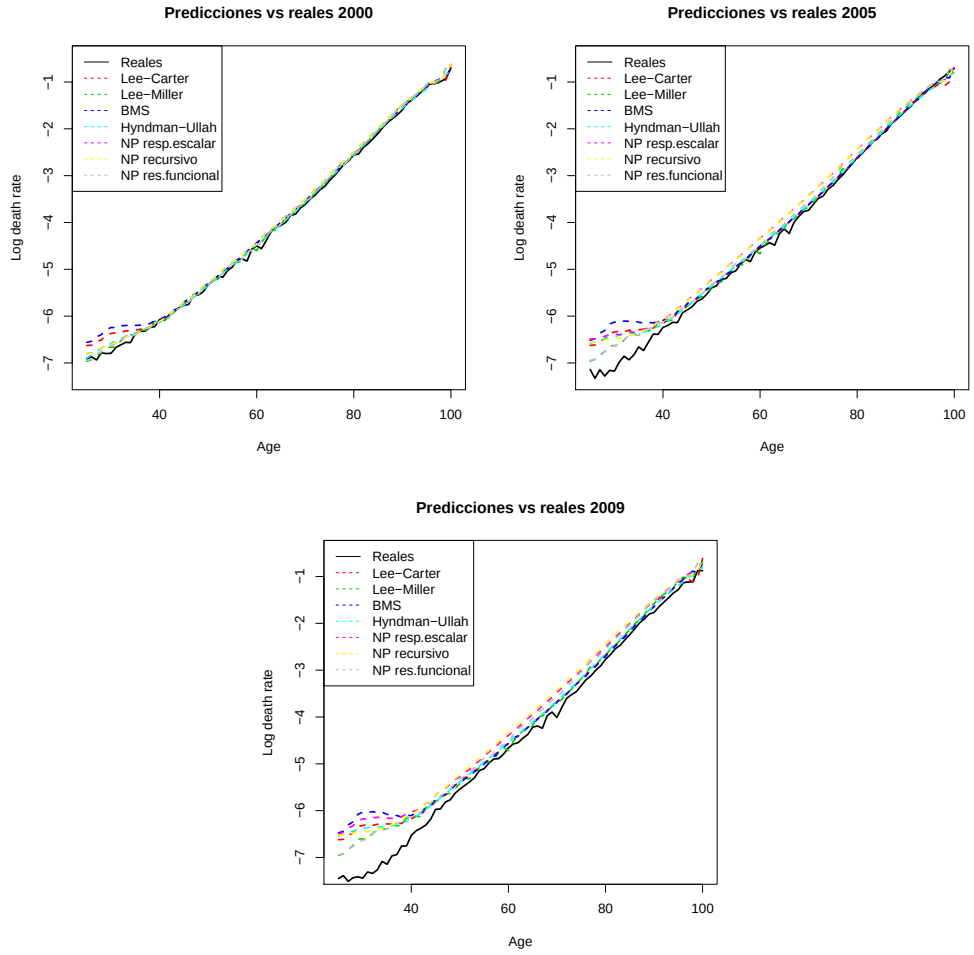

Figura 2.44: Comparación de todos los métodos junto con los datos observados de los hombres en los años 2000,2005,2009.

hombres y las 2.2 y 2.3 se corresponden con las mujeres.

Las tablas 2.1 y 2.2 muestra que los métodos cometen mayor error a medida que se predice un horizonte más, y corrobora que los métodos con mejores estimaciones de las tasas reales son el Lee-Miller y el autorregresivo con respuesta funcional, aunque en el año 2000 Hyndman les gana.

También se puede observar que las predicciones para el año 2000 para el método de Lee-Carter, BMS y los no paramétricos con respuesta no escalar son sustancialmente peores que los otros, el ECM en Hyndman es por lo menos la mitad.

Los modelo no paramétricos y el BMS cometen un mayor sesgo a lo largo de los años,

	Años de predicción				
Método	2000	2001	2002	2003	2004
Lee-Carter	0.01618596	0.02590977	0.03122607	0.03770282	0.05496022
Lee-Miller	0.00533646	**0.00888713**	**0.01008844**	**0.01198005**	**0.02445117**
BMS	0.02868044	0.04344859	0.05265925	0.06324223	0.08911678
Hyndman-Ullah	**0.00503639**	0.01285026	0.01894009	0.02774152	0.04624461
NP escalar	0.01050215	0.02955122	0.04460778	0.05523461	0.08509021
NP recursivo	0.01050215	0.02707784	0.03932756	0.04841150	0.07539750
NP funcional	0.00595259	0.01013548	0.01288690	0.01533101	0.03193327

Cuadro 2.1: Valores del ECMP definido en 2.3.1 según los diferentes métodos de predicción examinados, para los hombres.

	Años de predicción				
Método	2005	2006	2007	2008	2009
Lee-Carter	0.07176972	0.09861564	0.11186546	0.14035053	0.17239699
Lee-Miller	**0.03513974**	**0.05636371**	**0.06467191**	**0.08732117**	**0.11444744**
BMS	0.11376103	0.15239035	0.17273722	0.21240014	0.25742965
Hyndman-Ullah	0.06434209	0.09366390	0.10631030	0.13726348	0.17014500
NP escalar	0.10628987	0.15513904	0.17788551	0.22534381	0.27353244
NP recursivo	0.09223862	0.13283653	0.14662801	0.18646900	0.23035594
NP funcional	0.04219788	0.07103544	0.08037268	0.10838293	0.14341627

Cuadro 2.2: Valores del ECMP definido en 2.3.1 según los diferentes métodos de predicción examinados para los hombres.

	Años de predicción				
Método	2000	2001	2002	2003	2004
Lee-Carter	0.00855270	0.00885607	0.00818311	0.01263536	0.01195526
Lee-Miller	0.01377050	0.01460995	0.00878945	0.01138037	0.01199466
BMS	0.02084576	0.02016228	0.02633782	0.03207631	0.04202278
Hyndman-Ullah	**0.00617420**	**0.00768693**	**0.00509214**	**0.00769789**	**0.00823523**
NP escalar	0.01103872	0.01649173	0.02837690	0.04017274	0.07107508
NP recursivo	0.01103872	0.01552983	0.02231426	0.02646462	0.04592342
NP funcional	0.00989142	0.01046855	0.01097613	0.01176473	0.02406901

Cuadro 2.3: Valores del ECMP definido en 2.3.1 según los diferentes métodos de predicción examinados para las mujeres.

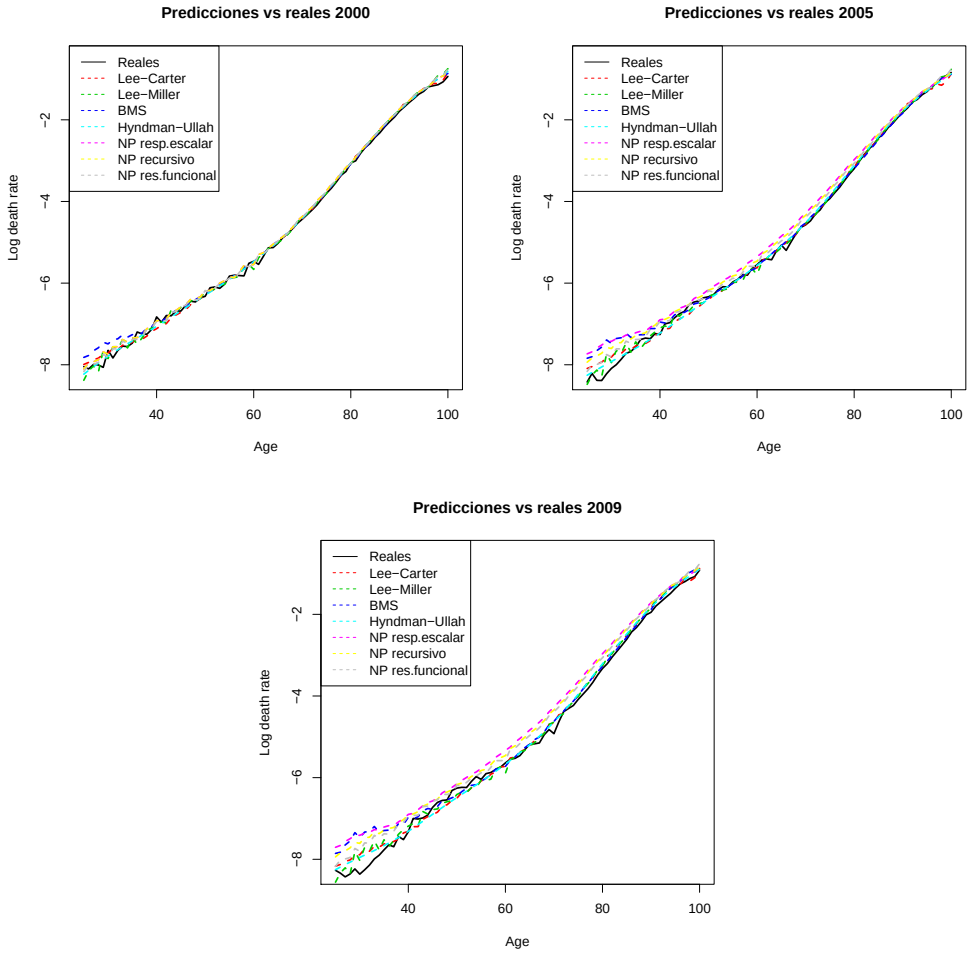

Figura 2.45: Predicciones de todos los métodos junto con los datos observados de los mujeres en los años 2000,2005,2009.

mientras que los otros en el año 2003 acortan distancias respecto a las tasas.

El método no paramétrico funcional en el primer año está próximo a Hyndman mientras que a partir del segundo comienza a dar resultados malos. Los no paramétricos con respuesta escalar no son competitivos ni a corto plazo, pues en el año 2000 cometen ya un error considerablemente mayor que Hyndman, que es el que da mejores resultados en este caso.

Se obtienen mejores resultados, aunque no buenos, con el modelo no paramétrico recursivo con respuesta escalar, que con el de respuesta escalar sin modificar la muestra. Esto no tiene porque ser simempre así, ya que si en algún momente se comete un error relevante al predecir, vamos a incluirlo en la muestra, lo que conllevará a resultados mucho peores.

	Años de predicción				
Métodos	2005	2006	2007	2008	2009
Lee-Carter	0.01729567	0.02411068	0.02746536	0.03127826	0.02855038
Lee-Miller	0.01262944	0.01923620	0.01966660	0.02661089	0.02806155
BMS	0.05631052	0.06988591	0.08352592	0.09535298	0.09608886
Hyndman-Ullah	**0.00947872**	**0.01678481**	**0.01839501**	**0.02242640**	**0.02209744**
NP escalar	0.09272837	0.12476428	0.14187039	0.15864885	0.17530223
NP recursivo	0.05802288	0.08241229	0.09457766	0.10750790	0.12118109
NP funcional	0.02933924	0.04771811	0.05662892	0.06871677	0.08167421

Cuadro 2.4: Valores del ECMP definido en 2.3.1 según los diferentes métodos de predicción examinados para las mujeres.

2.3.3. Discusión.

En este estudio se consideran distintos métodos utilizando técnicas paramétricas y no paramétricas para predecir las tasas logarítmicas de la mortalidad de cáncer por sexo en España. Aunque las primeras se encuentran en Bibliografía, hemos explorado propuestas no paramétricas funcionales con respuesta escalar y con respuesta funcional.

Las técnicas paramétricas que se consideran se basan en la descomposición en un valor singular, pudiendo en Hyndman-Ullah seleccionar más componentes en función del MISFE, obteniendo las predicciones mediante la proyeccion del coeficiente K_t según modelos de series de tiempo ARIMA. Notar que, el enfoque de Hyndman-Ullah conlleva el estudio de técnicas funcionales.

Las técnicas no paramétricas con respuesta escalar y funcional consisten en estimar la regresión mediante un AR(1), para predecir valores futuros en función de la estimación de la misma mediante el estimador de Nadaraya-Watson y la selección de la ventana por validación cruzada.

Los procedimientos de predicción se realizan con datos reales tanto del cáncer colorrectal como del cáncer en general, diferenciando entre las tasas de los hombres y mujeres de España. Los datos son las curvas discretizadas en las edades o grupo de edad quinquenales mayores de 25 años, las asociaciones entre la edad y la mortalidad en cada año.

Las tasas son más altas en hombres que en mujeres. En el caso de las tasas por cáncer las curvas presentan mayor variabilidad en los extremos de la variable edad, esta variabilidad posee gran importancia en los resultados a la hora de computar los intervalos de confianza

de las predicciones de los diferentes modelos paramétricos.

Las mujeres poseen mayor amplitud en los intervalos de predicción en edades bajas y esto se debe a que hay mayor variabilidad de las curvas para estas edades que en el caso de los hombres.

Los métodos paramétricos para los que se calcularon los intervalos de predicción al 95 % asumen normalidad. Sin embargo, en las propuestas no paramétricas no se realizan, pues ante el incumplimiento de normalidad de los residuos éstes no tendrían validez.

Se mostró que las predicciones para los hombres con edades bajas son insactisfactorias, dan mejores resultados en las mujeres, pues los ECM de predicción son menores en todos. Esto se debe a que la estimación de b_x es peor en el caso de los hombres. El método Lee-Miller y el no paramétrico con respuesta funcional son los que mejor predicen las tasas de los hombres, mientras que en las mujeres van mejor los métodos paramétricos, excepto BMS, y en especial el Método de Hyndman-Ullah. El método BMS es el peor a la hora de predecir, y esto puede deberse a la restricción de que k_t es lineal. El modelo de Hyndman explica un 99,8 % de la variabilidad existente, y dado que las estimaciones de las puntuaciones y de las componentes principales son buenas, se tiene que es el que menor sesgo comete.

Los métodos paramétricos son poco flexibles pero en general proporcionan mejores resultados. Sin embargo, los que proponemos como título exploratorio son flexibles pero al usar un AR(1), tienden a la media a medio y largo plazo, motivo por lo que a partir de cierta predicción van mal. Están sujetos entonces a las restricciones autorregresivas, por lo que hay que pensarse el usarlos a la hora de predecir a un medio-largo horizonte futuro.

Las tasas predichas para el cáncer son mayores que las reales tanto en hombres como en mujeres, y como estos modelos no tienen en cuanta ninguna variable externa, se puede pensar que los sistemas de cribado y los avances médicos están sirviendo de ayuda para combatir la mortalidad por dicha enfermedad.

A la hora de predecir las tasas de mortalidad para el período 2009-2018 mediante los modelos paramétricos muestran en general un incremento en los hombres mayores de los 45 años, y un descenso de la mortalidad en las mujeres de la misma edad.

Como se ha mencionado, estes modelos no tienen en cuenta factores externos, por lo que se proponen nuevas vías para poder conseguir mejores resultados a la hora de predecir, dada la importancia del cáncer en la sociedad.

Por ello se sugiere, para los métodos paramétricos, modelos parcialmente lineales que den cabida a variables externas. En el caso de los modelo no paramétricos se proponen modelos autorregresivos pero de mayores órdenes.

Por último, se plantea calcular, tanto en los modelos paramétricos como no paramétricos, los intervalos de predicción al 95 % mediante Bootstrap, pues de esta manera no es necesario que los residuos sean gaussianos.

Bibliografía

[1] Booth H, Hyndman R J, Tickle L, de Jong P. (2006). Lee-Carter mortality forecasting: A multi-country comparison of variants and extensions. *Demographic Research*, 15 (9), 289-310.

[2] Shang L H, Hyndman R J, Booth H. (2010). A comparison of ten principal component methods for forecasting mortality rates. *J. Epidemiol*, 20(2): 159-165.

[3] Erbas B, Akram M, Gertig D M, English D, Hopper J L, Kavanagh A M, Hyndman R J. (2010). Using functional data analysis models to estimate future time trends in age-specific breast cancer mortality for the United States and England-Wales. *J. Epidemiol*, 20(2), 159-165.

[4] Hyndman R J, Shahid Ullah Md. (2006). Robust forecasting of mortality and fertility rates: A functional data approach. *Monash University.*

[5] Hyndman R J, Shang L H. (2009). Forecasting functional time series. *Journal of the Korean Statistics Society*, 2009, 199-211.

[6] Erbas B, Hyndman R J, Gerting D M. (2007). Forecasting age-specific breast cancer mortality using functional data models. *Statistics in Medicine,*, 2007, 458-470.

[7] Ferraty, F. and Vieu, P. (2006). Nonparametric modelling for functional data. *Springer.*

[8] Ramsay, J.O. and Silverman, B.W. (2005). Applied functional data analysis.*Springer.*

[9] Ramsay, J.O. and Silverman, B.W. (1997). Functional data analysis. *Springer.*

[10] Anestis Antoniadis, Efstathios Paparoditis, Theofanis Sapatinas. (2009). Bandwidth selection for functional time series prediction*Statistics and Probability Letters*, 79, 733-740.

[11] Borrás J.M., Pareja L., Peris M., Espinás J.A. (2008). Análisis de la incidencia, la supervivencia y la mortalidad según las principales localizaciones tumorales, 1985-2019: cáncer colorrectal. *Med CLin (Barc.)*,131(supl 1), 58-62.

[12] Fernández E, La Vecchia C, González JR, Lucchini F, Negri E, Levi F. (2005). Converging patterns of colorectal cancer mortality in Europe. *Eur J Cancer.*, 41(3), 430-7.

[13] Berrino F, De Angelis R, Sant M, Rosso S, Lasota MB, Coebergh JW, et al. (2007). Survival for eight major cancers and all cancers combined for European adults diagnosed in 1995-99: results of the EUROCARE-4 study. *Lancet Oncol*,8, 773-783.

[14] Catenacci DV, Kozloff M, Kindler HL, Polite B. (2011). Personalized colon cancer care in 2010. *Semin Oncol*, 38(2),284-308.

[15] Ciccolallo L, Capocaccia R, Coleman MP, Berrino F, Coebergh JW, Damhuis RA, et al. (2005). Survival differences between European and US patients with colorrectal cancer: role of stage at diagnosis and surgery. *Gut*, 54, 268-273.

[16] González Pérez C.Y., Guerrero Guzmán V. M. (2007). Pronósticos estadísticos de mortalidad y su impacto sobre el Sistema de Pensiones de México *Actuarios pronosticadores*.

[17] U.S. Preventive Services Task Force. (2002). Screening for colorectal cancer: recommendation and rationale. *Ann Intern Med.*, 137(2), 129-31.