uc3m | Universidad Carlos III de Madrid

Máster Universitario Ciencias Actuariales y Financieras

2022-2023

Trabajo Fin de Máster

"La proyección actuarial de la edad máxima de seres humanos"

Yiran DU

Tutor/es

José Miguel Rodríguez-Pardo del Castillo

Jesús Ramón Simón del Potro

Madrid, junio de 2023

DETECCIÓN DEL PLAGIO

La Universidad utiliza el programa **Turnitin Feedback Studio** para comparar la originalidad del trabajo entregado por cada estudiante con millones de recursos electrónicos y detecta aquellas partes del texto copiadas y pegadas. Copiar o plagiar en un TFM es considerado una **Falta Grave**, y puede conllevar la expulsión definitiva de la Universidad.

Esta obra se encuentra sujeta a la licencia Creative Commons **Reconocimiento - No Comercial - Sin Obra Derivada**

RESUMEN

A medida que la esperanza de vida aumenta debido a los avances médicos y la mejora de las condiciones de vida, es esencial entender la longevidad extrema. El estudio examina la proyección actuarial de la edad máxima de los seres humanos utilizando técnicas estadísticas y de aprendizaje automático, incluyendo la distribución de Poisson, la regresión de Poisson ponderada y la regresión de vectores de soporte (SVR). Los datos utilizados son de la Base de Datos Internacional sobre Longevidad y la Lista de Supercentenarios del Grupo de Investigación en Gerontología. Tras analizar y ajustar los datos, el estudio concluye que la edad máxima potencial que los seres humanos pueden alcanzar es de aproximadamente 124 años. Esta información es crucial para las ciencias actuariales, especialmente para las compañías de seguros y los programas de pensiones.

Palabras clave:Esperanza de vida, Edad maxima, Ratio de mortalidad, Centenarios, Supercentenarios, Distribución de Poisson, Regresión de vectores de soporte (SVR), Suavización de Whittaker-Henderson

ABSTRACT

As life span increases due to medical advances and improved living conditions, it is essential to understand extreme longevity. The study examines the actuarial projection of the maximum age of humans using statistical and machine learning techniques, including Poisson distribution, weighted Poisson regression and support vector regression (SVR). The data used are from the International Longevity Database and the Gerontology Research Group's List of Supercentenarians. After analysing and adjusting the data, the study concludes that the maximum potential age that humans can reach is approximately 124 years. This information is crucial for actuarial science, especially for insurance companies and pension programmes.

Keywords: Life span, Maximum age, Mortality rate, Centenarians, Supercentenarians, Poisson distribution, Support Vector Regression (SVR), Whittaker-Henderson smoothing

DEDICATORIA

A mis abuelos, por su apoyo desde el principio. Ellos son la razón por la que elegí esta carrera y este tema en primer lugar.

ÍNDICE GENERAL

1. INTRODUCCIÓN 1
1.1. Motivación y objetivo 1
1.2. Estructura de trabajo 1
2. REVISIÓN DE LA LITERATURA 3
2.1. Investigaciones sobre la esperanza de vida y conceptos similares 3
2.2. Investigaciones sobre la esperanza de vida asociado con ciencias actuariales . 7
2.3. Revisión Literatura sobre la esperanza de vida y la ratio de mortalidad de supercentenarios 8
2.4. Vacío de investigación 8
3. MARCO TEÓRICO 10
3.1. Distribución Poisson 10
3.2. Regresión de Poisson Ponderada 11
3.3. Goodness-of- t 12
3.3.1. Prueba de Chi-cuadrado 12
3.3.2. Prueba de Kolmogorov-Smirnov 13
3.4. Regresión de Vectores de Soporte (SVR) 14
3.5. Validación cruzada K-Fold 16
3.6. Suavización por Método Whittaker-Henderson 17
3.7. Criterios de selección de modelos AIC y BIC 18
3.8. Media Móvil Simple (SMA) 19
3.9. Suavizamiento Exponencial (ES) 19
3.10. Extrapolación Lineal (LE) 20
4. DESCRIPCIÓN DE DATOS 21
4.1. Las bases de datos 21
4.2. Análisis previo de los datos de la base de datos de IDL 22
4.3. Análisis previo de los datos de la base de datos de GRG 28
5. ESTUDIO SOBRE LA EDAD MÁXIMA DE SERES HUMANOS 35
5.1. Introducción 35

5.2. Metodología 35
5.3. Resultados 37
5.3.1. Preparación de los datos 37
5.3.2. Ajuste y estimación de los parámetros de Poisson 40
5.3.3. Ajustar la curva de los parámetros lambda de cada año en función del tiempo 42
5.3.4. Predicción la media de edad máxima de seres humanos 43
5.3.5. Calculación de la edad máxima de seres humanos 45
6. ESTUDIOS SOBRE LA RATIO DE MORTALIDAD DE LOS SUPERCENTENARIOS 48
6.1. Introducción 48
6.2. Metodología 48
6.3. Resultados 49
6.3.1. Selección el intervalo adecuado para la tasa de mortalidad. 49
6.3.2. SVR 51
6.3.3. Suavización de Whittaker-Henderson 52
6.3.4. Selección de métodos 53
6.3.5. Predicción de la tasa de mortalidad de los supercentenarios 54
7. CONCLUSIONES Y FUTUROS ESTUDIOS 56
7.1. Conclusión 56
7.2. Futuros Estudios 56
BIBLIOGRAFÍA 58

ÍNDICE DE FIGURAS

2.1 Flujo de trabajo de análisis bibliométrico 3

2.2 Resumen de las estadísticas descriptivas de la bibliografía relevante . . . 5

2.3 Clustering de las palabras clave 6

3.1 Ilustración de la ecuación (3.10) 15

3.2 Ilustración de la ecuación (3.11) 16

3.3 Ilustración de Validación cruzada K-Fold 17

4.1 Análisis descriptivo de la base de datos de IDL 25

4.2 Análisis de la variable “edad” frente a otros variables de la base de datos de IDL 27

4.3 Análisis de la correlación de las variables de la base de datos de IDL . . . 28

4.4 Análisis descriptivo de la base de datos de GRG 31

4.5 Análisis de la variable “edad” frente a otros variables de la base de datos de GRG 33

4.6 Análisis de la correlación de las variables de la base de datos de GRG . . 34

5.1 Distribución de densidad de la edad de los fallecimientos en diferentes años 40

5.2 Cuantiles de la distribución comparando con la edad de fallecimiento real 42

5.3 La media de la edad máxima de seres humanos y la curva ajustada por SVR 43

5.4 La predicción media de la edad máxima de seres humanos por SVR . . . 44

5.5 La ratio de crecimiento de la media estimada de la distribución Poisson . 45

5.6 Distribución de densidad de la edad máxima de seres humanos predichos . 47

6.1 La ratio de mortalidad para los supercentenarios 51

6.2 La ratio de mortalidad y la curva de SVR 52

6.3 La ratio de mortalidad y la curva de la suavización de Whittaker-Henderson 53

6.4 Predicción de la tasa de mortalidad 55

ÍNDICE DE TABLAS

4.1 Una muestra de la base de datos de IDL 23

4.2 Una muestra de la base de datos de GRG 29

5.1 Mapa de las variables categóricas 37

5.2 Una muestra de datos después de mapear las variables categóricas 38

5.3 Resultado de análisis estadístico básico tras del ltro de los datos 39

5.4 Resultado de parámetros estimados y de las pruebas 41

5.5 Hiperparámetros determinados de SVR 42

5.6 La media predicha de la edad máxima de seres humanos y el cuantil 95 . 46

6.1 Tabla de $_nq_x$ calculado 50

6.2 Hiperparametros determinado de SVR 51

6.3 Parámetro determinado de suavización de Whittaker-Henderson 52

6.4 Comparación de AIC y BIC de SVR y Whittaker-Henderson 53

6.5 Valor suavizado por el método de Whittaker-Henderson 54

6.6 Comparación métodos de predicción 54

6.7 Predicción de la tasa de mortalidad 55

1. INTRODUCCIÓN

1.1. Motivación y objetivo

El estudio de la edad máxima de los seres humanos es importante en muchos campos, sobre todo en las ciencias actuariales. A medida que los avances médicos, los conocimientos sobre nutrición, la civilización social y la mejora de las condiciones de vida conducen a un aumento de la esperanza de vida, resulta cada vez más importante comprender cómo afectan estos cambios a la edad máxima de los seres humanos.

Por un lado, desde el punto de vista actuarial, conocer la edad máxima que probablemente alcanzarán las personas ayuda a mejorar la precisión de los cálculos y la exactitud de las proyecciones, lo cual es crucial para evaluar el riesgo, determinar las primas de los seguros de vida y determinar las prestaciones de jubilación y otras nanzas relacionadas con la vida.

Por otra parte, los estudios sobre centenarios y supercentenarios se han considerado hasta ahora sólo en determinados países y regiones. Sin embargo, la proporción de este grupo aumenta cada año a medida que aumenta el número de centenarios y supercentenarios en más países y regiones. Debido a la investigación en este ámbito todavía no está bien explorada, la investigación sobre la mortalidad y la edad máxima prevista en este ámbito puede contribuir a completar este vacío.

Por lo tanto, este estudio no se limita a un país o región concretos, sino que examina la longevidad de los centenarios y supercentenarios registrados y las tendencias de sus tasas de mortalidad a nivel mundial.

1.2. Estructura de trabajo

Este estudio contiene 7 secciones, de las cuales la parte de redes de palabras claves de análisis bibliométrico en Sección 2 se ha realizado con VOSviewer versión 1.6.19 y el resto de la programación con Python versión 3.9.2.

Sección 1 - Introducción: El objetivo de este estudio es entender la proyección actuarial de la edad máxima de los seres humanos. Con la subida de la esperanza de vida en las últimas décadas, es fundamental entender cuánto tiempo puede vivir una persona con una probabilidad extrema. Este conocimiento es crucial para las ciencias actuariales, ya que se requiere mucha atención a los eventos de poca probabilidad mientras grandes pérdidas en las compañías de seguros y los programas pensiones.

Sección 2 - Revisión de la Literatura: Se lleva a cabo una revisión exhaustiva de la literatura existente, que examina tanto los estudios generales sobre la esperanza de vida

como aquellos que se vinculan con las ciencias actuariales. También se realiza un análisis bibliométrico para identi car los temas más relevantes en este campo de estudio.

Sección 3 - Marco Teórico: En esta sección se detallan los conceptos estadísticos y matemáticos que se emplearán en el estudio. Esto incluye la distribución Poisson, la regresión de Poisson ponderada, las pruebas de bondad de ajuste, la regresión de vectores de soporte (SVR), la validación cruzada K-Fold, entre otros. Además, se discuten varios métodos de suavización y selección de modelos.

Sección 4 - Descripción de los Datos: Se describen las bases de datos que se utilizarán en el estudio, incluyendo la Base de Datos Internacional sobre Longevidad (IDL) y la Lista de Supercentenarios del Grupo de Investigación en Gerontología (GRG). Se proporciona un análisis preliminar de los datos contenidos en estas bases.

Sección 5 - Estudio sobre la Edad Máxima de los Seres Humanos: Esta sección detalla cómo se prepararon y analizaron los datos en relación con la edad máxima alcanzada por los seres humanos. Los pasos incluyen agrupación de datos, aplicación de regresiones, evaluación de bondad de ajuste y realización de proyecciones futuras utilizando SVR.

Sección 6 - Estudios sobre la Ratio de Mortalidad de los Supercentenarios: Se calculan y comparan las tasas de mortalidad para diferentes intervalos de edad utilizando los datos del GRG. Se visualiza la relación entre la mortalidad y la edad, y se seleccionan los mejores modelos de suavización y proyección mediante la comparación de varias métricas.

Sección 7 - Conclusiones y Futuros Estudios: Finalmente, se presentan los resultados clave del estudio, incluyendo la estimación de la edad máxima que los seres humanos pueden alcanzar y las tendencias de mortalidad en edades extremas. Se discuten las implicaciones de estos resultados para las ciencias actuariales y se sugieren áreas para futuras estudios.

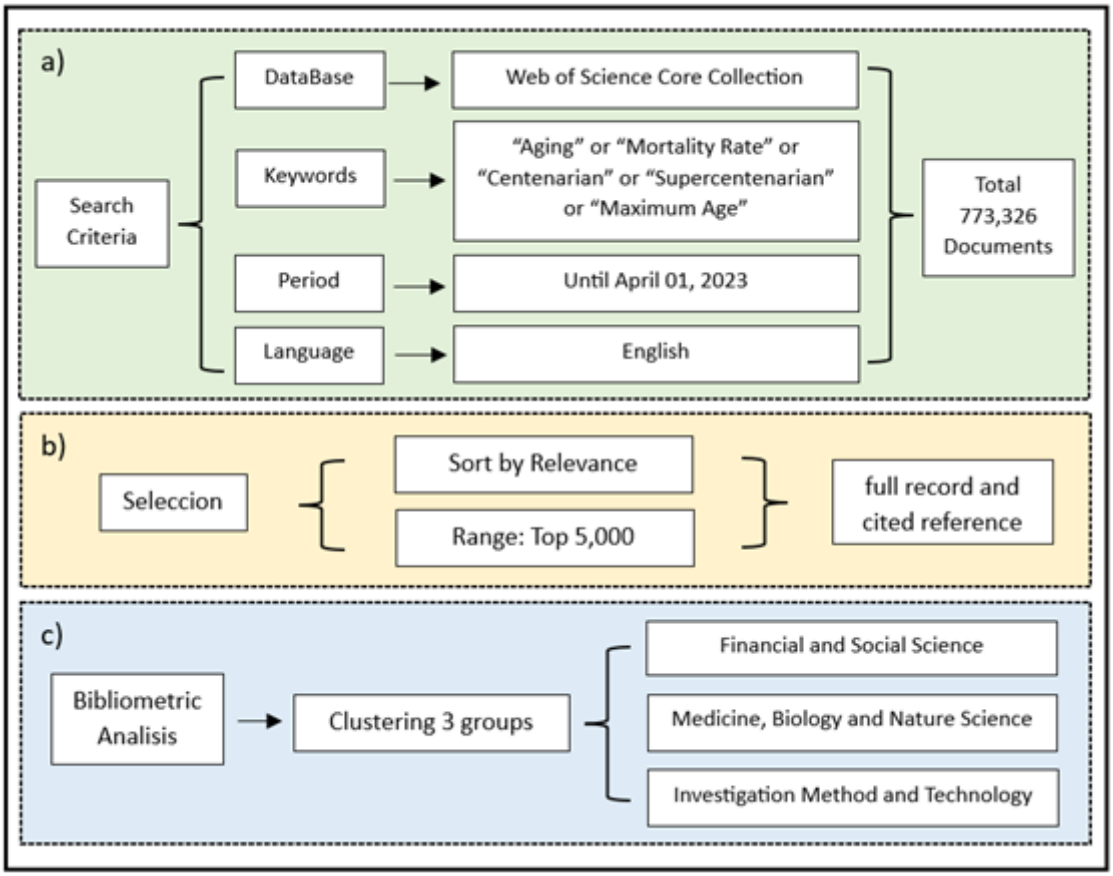
2. REVISIÓN DE LA LITERATURA

2.1. Investigaciones sobre la esperanza de vida y conceptos similares

El estudio de la esperanza de vida y conceptos similares ha sido siempre un tema de interés para la comunidad cientí ca y, en consecuencia, han publicado muchas investigaciones en estas áreas. Los conceptos clave en estos estudios incluyen el envejecimiento, la tasa de mortalidad, y las guras demográ cas conocidas como centenarios y supercentenarios.

En este estudio se realizó un análisis bibliométrico utilizando la base de datos Web of Science (WOS) para rastrear la prevalencia e in uencia de estos temas a lo largo del tiempo. Como se ha presentado en la gura 1, los criterios de búsqueda incluyeron el conjunto básico de WOS con las palabras clave “aging”, “mortality”, “centenarian”, “supercentenarian” y “maximum age”. La búsqueda se realizó únicamente en literatura en lengua inglesa, sin restricciones de periodo de tiempo.

Figura 2.1
Flujo de trabajo de análisis bibliométrico



Fuente: Elaboración Propia

En total, se encontraron 773,326 documentos que coincidían con estos criterios de búsqueda. Este gran número de estudios re eja la importancia de estos temas en una amplia gama de disciplinas, incluyendo la biología, la medicina, la demografía, la sociología, y las ciencias actuariales. Después de la búsqueda, los 5,000 documentos más relevantes han seleccionado para el análisis bibliométrico.

Una vez tener estos documentos relacionado con este campo, se puede realizar un análisis previo, y el resultado se puede encontrar en la gura 2.2.

La gura 2.2a representa las 10 categorías principales de la Web of Science por número de registros. Se puede observar una predominancia de la geriatría y la gerontología (22.3 %), seguida de las neurociencias (16.1 %) y la gerontología (14.4 %). Esta distribución re eja el interés cientí co en el estudio del envejecimiento desde una variedad de disciplinas, que van desde la biología molecular y celular hasta la medicina general y la salud ocupacional.

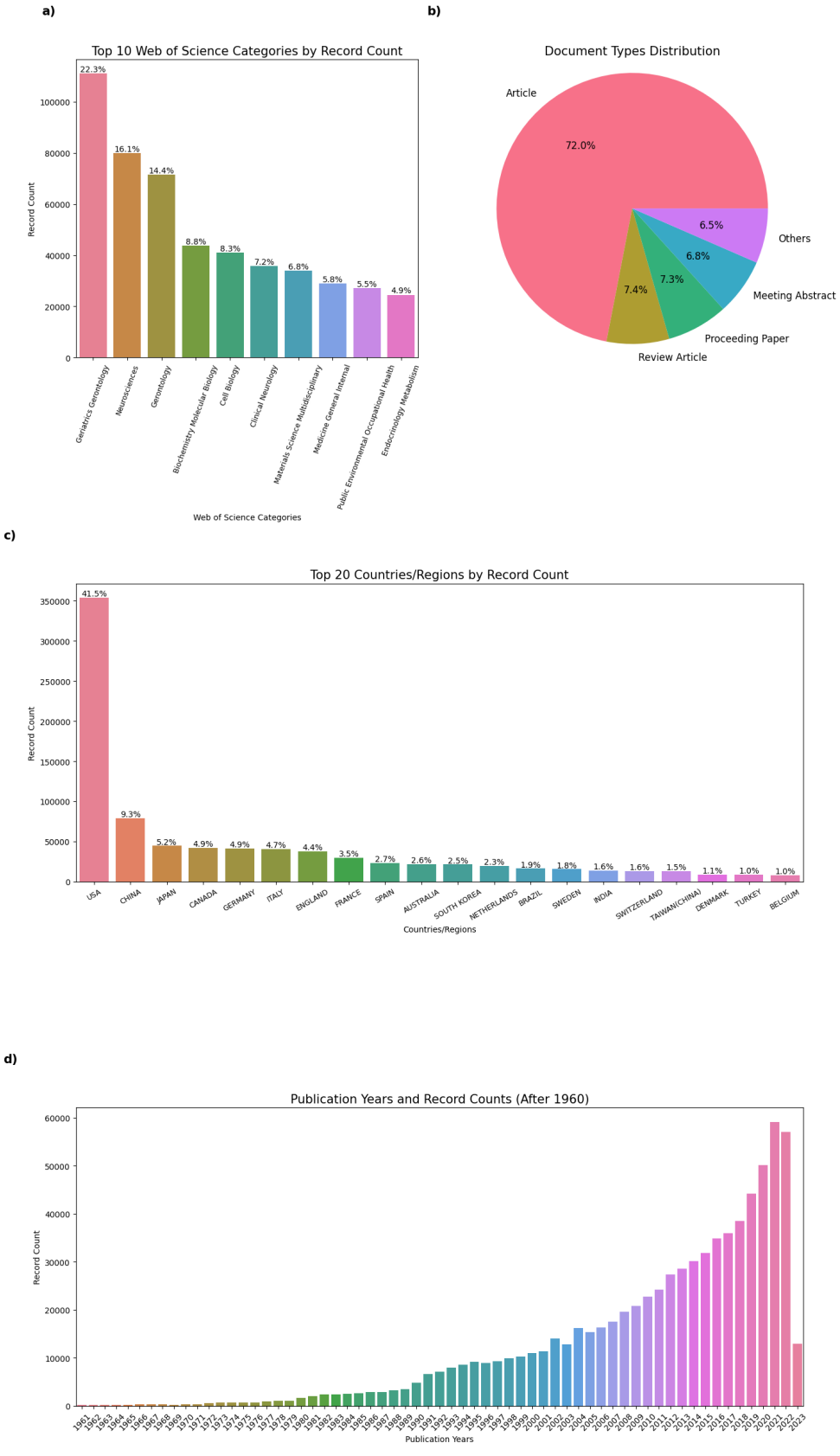
Luego, la gura 2.2b muestra la distribución de los tipos de documentos en la literatura sobre la longevidad y la mortalidad. Los artículos representan la mayoría (72 %), seguidos de las revisiones de artículos (7.4 %) y los trabajos de procedimientos (7.3 %). Esto sugiere que la mayoría de las investigaciones en este campo se publican en forma de artículos originales.

Además, los 20 países o regiones principales por número de registros han presentado en la gura 2.2c. Los Estados Unidos encabezan la lista (41.5 %), seguidos de China (9.3 %) y Japón (5.2 %). Esta distribución indica la concentración geográ ca de la investigación sobre la longevidad y la mortalidad.

Finalmente, en la gura 2.2d, el gra co muestra las tendencias de publicación a lo largo del tiempo, comenzando alrededor de 1960. Se observa un crecimiento exponencial en el número de registros, llegando a aproximadamente 60,000 en 2022. Esto re eja el creciente interés y la rápida expansión de la investigación en este campo durante las últimas seis décadas.

Figura 2.2

Resumen de las estadísticas descriptivas de la bibliografía relevante



Fuente: Elaboración Propia

La revisión bibliométrica de la literatura ha utilizado VOSviewer, y ha dado un resultado de clasi cación de las palabras clave en tres conjuntos temáticos ilustrados en la gura 2.3.

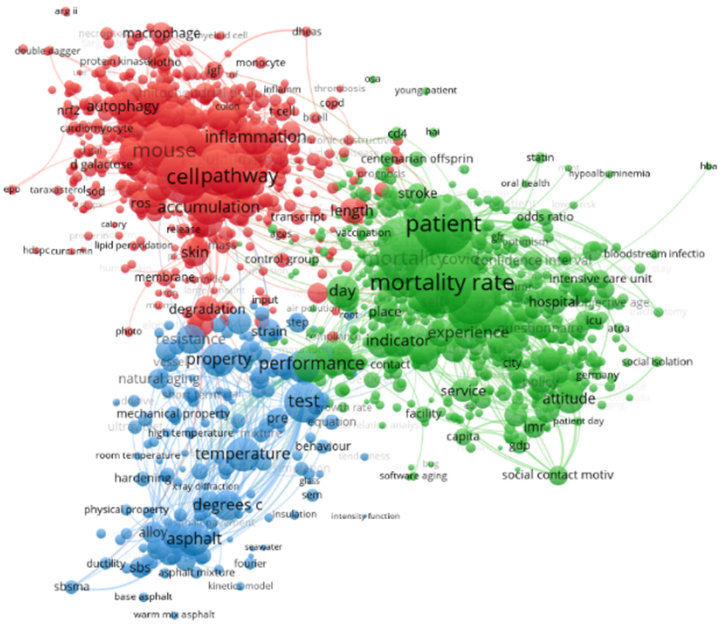
El primer conjunto, representado en rojo, aborda aspectos biológicos y medicinas. Palabras como “in amación”, “acumulación”, “degradación” y “macrófago” implican estudios sobre los procesos biológicos y moleculares del envejecimiento y su in uencia en la longevidad y la mortalidad.

El siguiente conjunto, marcado en verde, se centra en elementos demográ cos, socioeconómicos y de salud asociados al envejecimiento. Los términos “tasa de mortalidad”, “paciente”, “aislamiento social” y “envejecimiento” sugieren investigaciones que evalúan cómo factores sociales y de economía impactan en la longevidad y la mortalidad.

Finalmente, el grupo azul se aparta de los enfoques anteriores, centrando en características físicas y campos de investigación menos relevante con la longevidad y la mortalidad comparando con los anteriores como asfalto, propiedades mecánicas y propiedades físicas. Este conjunto tiene una mayor relación con investigaciones de materiales y su resistencia frente al envejecimiento y degradación, y también se asocia a tecnologías y metodologías.

Estos conjuntos representan la variedad de disciplinas y métodos adoptados para examinar la mortalidad y longevidad en ciencias biológicas, sociales y físicas. Esta diversidad de investigación genera una base sólida para el estudio en este ámbito.

Figura 2.3
Clustering de las palabras clave



Fuente: Elaboración Propia

2.2. Investigaciones sobre la esperanza de vida asociado con ciencias actuariales

Las investigaciones asociadas con la ciencia actuarial han evolucionado en las últimas décadas. Antes del siglo XX, las técnicas actuariales se basaban principalmente en la estadística descriptiva como la ley de los grandes números (Bernoulli y Sylla, 2006) y el teorema central del límite (Fischer, 2011). Anderson y Rosenberg (Anderson y Rosenberg, 1998) discutieron la diferencia entre la población y los diferentes subgrupos de edad y el ajuste estadístico de la ratio de mortalidad. Después de la década de 1970, los científicos empezaron a incorporar más teoría sobre probabilidad y estadística inferencial, tanto la inferencia bayesiana (Reinhardt, 1987) como la teoría de la credibilidad (Kakar, 2007).

Desde principios de los años 2000, la ciencia actuarial ha estado utilizando cada vez más modelos estocásticos y simulaciones de Monte Carlo (Eckhardt, 1987). Los modelos de mortalidad de Lee-Carter (Lee, 2000) y Cairns-Blake-Dowd (Cairns et al., 2008) se han convertido en estándares de la industria para modelar la evolución futura de las tasas de mortalidad.

En la década de 1980 y 1990, los estudios se centraron en los seguros de vida, ya que era el producto de seguro más común. Durante este tiempo, la investigación mostró una fuerte correlación entre la esperanza de vida y la tasa de interés utilizada para descontar los beneficios de los seguros de vida (Yaari, 1965).

A finales de los años 90 y principios de los 2000, los estudios actuariales comenzaron a centrarse más en los seguros de salud y las pensiones, debido al envejecimiento de la población y la creciente importancia de la seguridad del ingreso en la jubilación. Los estudios durante este período se centraron en el impacto de la mejora de la esperanza de vida en la solvencia de los sistemas de pensiones y los seguros de salud (Bowers et al., 1987).

Además, las investigaciones relacionadas con la esperanza de vida en la última década ya han movido desde analizar las potencias factores que afectan la esperanza de vida (Christensen et al., 2009) hasta las proyecciones de la esperanza de vida en diferentes situaciones, tanto la ubicación como el entorno socioeconómico (Ho y Hendi, 2018).

J. M. R. Del Castillo y López-Farré (2017) señalan que “En el siglo XII la esperanza de vida va en aumento, y actualmente en España supera los 80 años. Pero lo que puede cambiar el panorama actual es que gracias a los conocimientos científicos y tecnológicos que se desarrollan a gran velocidad, no es improbable que el ser humano pueda superar frecuentemente la barrera de los 100 o 120 años a finales del siglo XXI.” En un estudio más reciente, Gimeno-Miguel et al. (2019) examinaron la salud de las personas centenarias en España. Descubrieron que la edad media de este grupo demográfico en crecimiento es de 101.6 años. A pesar de la alta prevalencia de multimorbilidad y polifarmacia en este grupo, su existencia representa un fenómeno de longevidad significativo.

2.3. Revisión Literatura sobre la esperanza de vida y la ratio de mortalidad de supercentenarios

En el campo de la gerontología, una cuestión recurrente es si existe un límite superior para la vida humana. Varias teorías y estudios han surgido a lo largo de los años para tratar de responder a esta pregunta.

Weon y Je (2009) utilizaron un enfoque teórico para argumentar que la vida humana tiene un límite máximo, aunque no especi caron cuál sería este límite. Un año más tarde, Olshansky (2010) debatió la existencia de este límite superior jo para la vida humana.

Posteriormente, Dong et al. (2016) presentaron pruebas de un límite para la vida humana, situándolo alrededor de los 115 años. Esta a rmación se basa en análisis de datos demográ cos y en tendencias observadas en la duración de la vida de las personas con edad extrema.

Sin embargo, no todos los investigadores concuerdan con la existencia de un límite máximo jo. Barbi et al. (2018) desa aron esta idea, sugiriendo que la mortalidad humana se estabiliza después de los 105 años. Esto implicaría que, una vez alcanzada esta edad, la probabilidad de morir en un año determinado permanecería constante.

Añadiendo otra perspectiva, Hughes y Hekimi (2017) discutieron la idea de un límite de la vida humana, sugiriendo que existen múltiples trayectorias posibles para la longevidad máxima. Este estudio plantea la posibilidad de que no exista un límite jo y uniforme para todos los individuos, sino que la longevidad máxima puede variar según una serie de factores.

En cuanto a la mortalidad de los supercentenarios, existe una considerable variabilidad en las conclusiones de diferentes estudios. Fries (2002) propuso que la tasa de mortalidad aumenta hasta cierto punto y luego se estabiliza. Esta teoría es respaldada por otros estudios como el de Hammond (2000), quien propone un patrón complejo de incremento, estabilización y posible disminución en las tasas de mortalidad a edades extremas.

Estudios más recientes, como el de Gampe (2010), sugieren que las tasas de mortalidad se estabilizan a una edad avanzada y pueden incluso disminuir después de los 110 años. Esta idea es apoyada por el análisis exhaustivo de Maier et al. (2012) sobre los supercentenarios, que también discute cómo la mortalidad puede uctuar dependiendo de varios factores. Gavrilov y Gavrilova (2011) proporcionan un enfoque complementario, examinando las tendencias de mortalidad en las edades avanzadas usando datos de la Administración de Seguridad Social de EE. UU.

2.4. Vacío de investigación

El campo de la ciencia actuarial y la demografía se enfrenta a varios desafíos al tratar de entender la esperanza de vida y la mortalidad en las edades más avanzadas.

Falta de investigaciones de vista global: Muchos estudios sobre la longevidad humana y la tasa de mortalidad de supercentenarios se centran en un único país o región (es decir, de poblaciones especí cas, a menudo en países desarrollados con registros de nacimiento y muerte). Esta falta de vista global puede limitar nuestra comprensión de la longevidad humana, ya que los factores culturales, socioeconómicos y ambientales pueden variar considerablemente en diferentes partes del mundo.

Investigación estática versus dinámica: Muchos estudios actuales ofrecen instantáneas de la mortalidad y la esperanza de vida en un momento dado. Sin embargo, la dinámica de la mortalidad y la longevidad puede cambiar con el tiempo debido a avances en la medicina, cambios en las condiciones de vida, y otros factores. Por lo tanto, es importante que la investigación sea dinámica y se actualice regularmente.

Insu ciencia de datos: Las tasas de mortalidad para las edades extremas pueden ser difíciles de estimar con precisión debido a la escasez de datos. En las edades más avanzadas, hay relativamente pocas personas, y los datos pueden ser inciertos debido a problemas como la veri cación de la edad. Esta insu ciencia de datos puede llevar a estimaciones inexactas de la mortalidad y aumentar la incertidumbre en la predicción de la longevidad y las obligaciones nancieras futuras.

Estos desafíos subrayan la necesidad de más y mejor investigación en este campo, especialmente estudios que sean globales, que se actualicen regularmente y que utilicen datos de alta calidad.

3. MARCO TEÓRICO

3.1. Distribución Poisson

Utilizar un método estadístico en lugar de un método empírico para investigar la mortalidad tiene una historia larga, se ha producido mucha literatura analizar la mortalidad utilizando diferentes distribuciones (Brillinger, 1986 ; Janssen y Kunst, 2007 ; ...). Dado que el carácter de pocos datos de este estudio, se requiere un modelo no complicado para analizar los datos para evitar el efecto de over- tting.

Al considerar la edad promedia de las personas mayores de 105 años que fallecieron en los últimos años en todo el mundo, se propone la siguiente ecuación:

$$N\ x_i; y_j \quad Poisson \quad y_j \qquad (3.1)$$

donde:

- $X = x_1; x_2; : : : ; x_i$ sea el conjunto de edades posibles al morir en un determinado añoy_j , en este estudio, agrupadas por medio aX 2 105; 105:5; 106; 106:5; : : :.
- $Y = y_1; y_2; : : : ; y_j$ se presenta un conjunto de años que sea el rango de este estudio Y 2 [1987; 2018].
- $N\ x_i; y_j$ sea el número de individuos que mueren con e$x_i$ad delañoy_j.
- Parámetro y_j de la distribución Poisson indica el promedio del conjunto de edad $X\ y_j$.

Para que se cumpla la ecuación anterior, se deben cumplir las siguientes suposiciones:

- Las muertes deberían ser independientes: la muerte de un individuo no afecta la probabilidad de que otro individuo muera.
- La tasa promedio de muertes es constante para cada grupo de edad: a través de la revisión de la literatura, ya sabemos que las tasas de mortalidad se estabilizan o disminuyen en edades extremadamente avanzadas, lo que signi ca que incluso si la tasa promedio cambia a través de los grupos de edad, variará de manera insigni - cante.
- La probabilidad de más de una muerte en un intervalo in nitesimalmente pequeño es insigni cante: esta suposición sería razonable ya que es poco probable que dos personas de la misma edad mueran simultáneamente.

3.2. Regresión de Poisson Ponderada

Para minimizar el efecto del cambio de la tasa promedio de muertes en la segunda suposición mencionado anteriormente, en este estudio, se utiliza modelo de Poisson ponderado (J. Del Castillo y Pérez-Casany, 1998) para ajustar, cuyas ventajas son siguientes:

- Equilibrio entre grupos de edad: el método de ajustar con Poisson ponderada puede ayudar a corregir los desequilibrios de representación de datos entre diferentes grupos de edad, fomentando una distribución de datos más equilibrada.
- Gestión de la heterogeneidad en las tasas de mortalidad: Este modelo permite adaptar las variaciones intrínsecas en las tasas de mortalidad entre diferentes grupos de edad, así como mejorar las estimaciones de la tasa de mortalidad.
- Abordar la sobre dispersión: al ponderar diferentes grupos, garantiza que ningún grupo in uya más en las estimaciones generales así que se maneja la sobre dispersión y calcular resultados con ables.

La probabilidad de observar eventos (en nuestro caso, muertes en edad y) dado un parámetro en un modelo de Poisson es:

$$P(Z = z) = \frac{^{z}e}{z!} \tag{3.2}$$

Para ajustar nuestros datos con el modelo de Poisson ponderado, y estimar el valor para cada año, se utiliza la estimación de máximo verosimilitud.

La ecuación de la verosimilitud ponderada para cada observación es:

$$L_k = \frac{w_k \; {}^{z_k}_{k} e \; {}^{k}}{z_k!} \tag{3.3}$$

Siendo L_k es la verosimilitud de la observación k, w_k es el peso de dicha observación que en este estudio ha sido calculado por p_x, y z_k es el número observado de muertes en edad x año y para la observación k, $_k$ es el parámetro de tasa para la observación k en el modelo de Poisson, y nalmente $z_k!$ es la factorial de z_k.

Así se de ne la función de verosimilitud global como:

$$L = \prod_{k} L_k \tag{3.4}$$

Donde L es la función de verosimilitud global, y el producto es sobre todas las observaciones k.

Por lo tanto, el log-verosimilitud se puede escribir como:

$$\log(L) = \sum_{k} \log(L_k) \tag{3.5}$$

De allí que según la función (3.3) y (3.5), se pueda tener la siguiente ecuación:

$$\log(L) = \sum_{k} w_k \quad z_k \quad \log(\ _k) \quad _k \quad \log(z_k!) \tag{3.6}$$

3.3. Goodness-of- t

Una vez obtenido el vector de los parámetros de lambda, es necesario probar la Goodness-of-Fit de nuestros datos observados (casos de fallecimiento) y la distribución Poisson con parámetro . En este estudio, elegimos dos métodos para probar la Goodness-of-Fit.

3.3.1. Prueba de Chi-cuadrado

La prueba de Chi-cuadrado (Cochran, 1952) es un método clásico de estadística para chequear la similitud de las frecuencias esperadas y observadas. Dicha prueba tiene los siguientes supuestos y limitaciones:

Supuestos:

- Independencia de las observaciones: chi-cuadrado asume que cada observación en el conjunto de datos es independiente.
- Tamaño de muestra su ciente: la prueba de chi-cuadrado asume que el tamaño de muestra es su cientemente grande.
- Distribución Teórica Especi cada: La distribución teórica debe ser completamente especi cada.

Limitaciones:

- Sensibilidad a muestras grandes: la prueba Chi-Square es sensible a muestras de gran tamaño. A medida que aumenta el tamaño de la muestra, las pequeñas diferencias entre las frecuencias esperadas y observadas de cada caso pueden resultar en el rechazo de la hipótesis nula, incluso cuando estas diferencias no sean signi cativas en la práctica. Como en este estudio los subconjuntos de los eventos no tienen un tamaño relevante, está a favor a este estudio para distinguir la diferencia entre los datos y la distribución.
- Arbitrariedad del agrupamiento: la prueba requiere la creación de “contenedores” o “clases” cuando se trata de datos continuos, lo que puede conducir a resultados de prueba diferentes según cómo se agrupan los datos.

Una vez cumple los supuestos y que las limitaciones no afectan este estudio, se puede aplicar la prueba con la siguiente función matemática.

La estadística Chi-Cuadrado se calcula como:

$$X^2 = \sum \frac{(O_i \quad E_i)^2}{E_i} \tag{3.7}$$

dónde:

- O_i sea la frecuencia observada,
- E_i sea la frecuencia esperada bajo el supuesto de la distribución teórica.

3.3.2. Prueba de Kolmogorov-Smirnov

Utilizar la prueba de Kolmogorov-Smirnov para el Goodness of t tiene una historia muy larga (Massey, 1951), la prueba cuanti ca la diferencia entre la distribución empírico y el modelo estimado para detectar la probabilidad de que los datos o muestras vienen de dicho modelo o distribución.

Supuestos:

- Independencia de las observaciones: Al igual que la prueba Chi-cuadrado, la prueba Kolmogorov-Smirnov también asume que cada observación en el conjunto de datos es independiente.
- Observaciones distribuidas de forma idéntica: Las observaciones están distribuidas de forma idéntica, lo que implica que cada observación está sujeta a la misma distribución de probabilidad.

Limitaciones:

- Sensibilidad a las desviaciones centrales: la prueba Kolmogorov-Smirnov es más sensible a las desviaciones de la distribución teórica en el centro de la distribución en comparación con las colas.
- Potencia reducida con parámetros estimados: La potencia de la prueba es menor cuando los parámetros de la distribución teórica se estiman a partir de los datos.

Después de veri car que se puede aplicar la prueba de Kolmogorov-Smirnov, se utiliza la siguiente función para calcular:

$$D = \text{máx} j F_n(x) \quad F(x; \ ) j \tag{3.8}$$

dónde:

- $F_n(x)$ es la función de distribución acumulada empírica
- $F(x; \ )$ es la función de distribución acumulativa de la distribución teórica.

3.4. Regresión de Vectores de Soporte (SVR)

Después de ajustar las distribuciones de Poisson, se pasa a modelar las tendencias temporales en los parámetros de lambda estimados. Utilizamos una regresión de vectores de soporte (SVR) (Drucker et al., 1997) con un núcleo RBF para modelar la relación entre el año de fallecimiento y el parámetro lambda.

La regresión de vectores de soporte (SVR), una extensión de la máquina de vectores de soporte (SVM), es un potente algoritmo para modelar relaciones tanto lineales como no lineales entre variables.

Conceptualmente, SVR opera bajo un paradigma distintivo, divergente de los métodos tradicionales de regresión estadística. En lugar de minimizar el error de entrenamiento observado, SVR intenta minimizar el error de generalización, con el objetivo de lograr un modelo que funcione bien en datos no observados. Esto se consigue introduciendo la llamada función de pérdida insensible a , que ignora los errores dentro de un cierto umbral , centrándose así en los errores más grandes fuera de este rango aceptable.

La función de objetivo denota como:

$$\frac{1}{2} jjwjj^2 + C \sum_{i=1}^{n} \quad _i + \ _i \tag{3.9}$$

Sujeto a las restricciones denota como:

$$\begin{aligned} y_i \quad & w^T \ (x_i) \quad b \quad + \\ & w^T \ (x_i) + b \quad + \ _i \\ & \quad _i; \ _i \quad 0; 8i: \end{aligned} \tag{3.10}$$

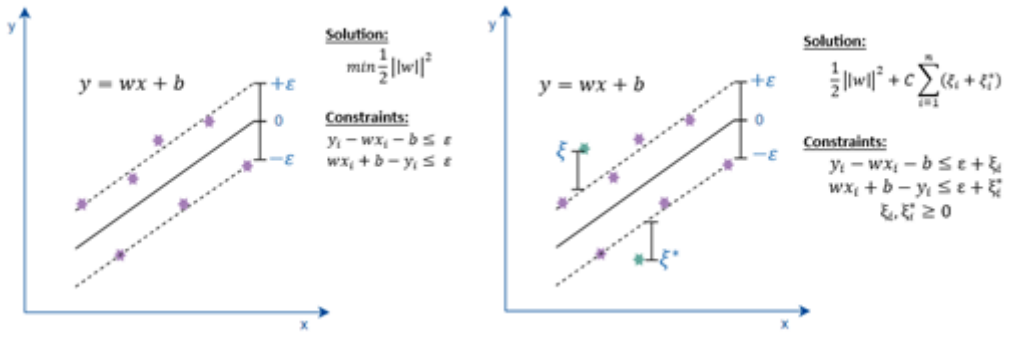
donde:

- w es el vector de pesos
- x_i son los vectores de entrada
- b es el sesgo
- C es el parámetro de regularización
- $"$ es el error máximo permitido
- $_i$ y $_i$ son variables de holgura que permiten errores mayores$"$ que

- n es el número de ejemplos de entrenamiento
- (x_i) es la transformación de x_i en un espacio dimensional superior realizada por la función de kernel

En este marco, el objetivo principal de SVR es determinar una función de estimación que se desvíe de las verdaderas respuestas y_i en no más del umbral predenido ε para todos los datos de entrenamiento, al tiempo que se esfuerza por ser lo más plana posible. Esto se consigue minimizando su norma $\|w\|$ (la medida de planitud), lo que conduce a un modelo escaso y simple.

Figura 3.1
Ilustración de la ecuación (3.10)



Fuente: Elaboración Propia

En SVR, el reto clave es encontrar esta función "más plana". Cuando los datos están relacionados linealmente, la función es un hiperplano en el espacio de características. Sin embargo, cuando los datos muestran una relación no lineal, el espacio de características se transforma utilizando una función de Kernel y la función "más plana" en este nuevo espacio es una curva o super cie.

La función de núcleo de base radial (RBF) es un tipo de función de Kernel más usada en algoritmos el SVR que permite modelar relaciones no lineales entre variables y transformar el espacio de características para hacerlo más expresivo.

La función de núcleo RBF se de ne como:

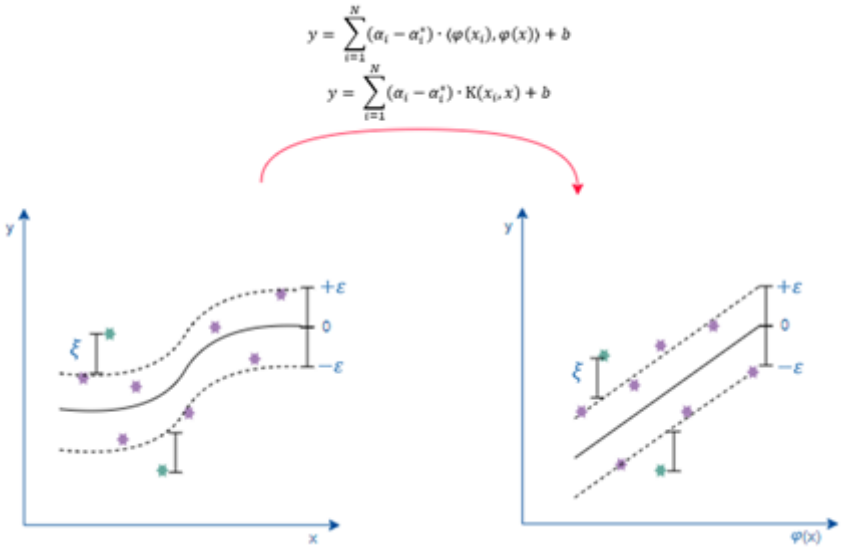
$$K(x; x^0) = \exp \quad kx \quad x^0k^2 \qquad (3.11)$$

donde:

- x y x^0 son dos vectores de características de entrada.
- $jjx \quad x^0jj^2$ es la distancia euclidiana al cuadrado entre los vectores x y x^0.

- es un hiperparámetro que controla la in uencia de cada punto en el cálculo del núcleo. Un valor más alto de hace que los puntos cercanos tengan un mayor impacto.

Figura 3.2
Ilustración de la ecuación (3.11)



Fuente: Elaboración Propia

En general, SVR ofrece un método de regresión robusto que hace hincapié en la generalización del modelo, por lo que es una herramienta poderosa para tareas de regresión lineal y no lineal.

3.5. Validación cruzada K-Fold

La validación cruzada K-Fold es un procedimiento utilizado en aprendizaje automático para evaluar la habilidad de un modelo en datos independientes y para ajustar el modelo (tunear sus hiperparámetros) si es necesario. Es particularmente útil en situaciones en las que los datos son limitados.

El procedimiento general se puede describir de la siguiente manera:

- Dividir el conjunto de datos en subconjuntos de aproximadamente el mismo tamaño.
- Para cada en [1; 2; :::; k]:
- Entrenar el modelo con los 1 subconjuntos que no incluyen .
- Evaluar el modelo con el subconjunto y registrar el resultado.

- Calcular la media de los resultados para obtener el rendimiento final del modelo.

Figura 3.3
Ilustración de Validación cruzada K-Fold

Fuente: Elaboración Propia

3.6. Suavización por Método Whittaker-Henderson

El método de suavización de Whittaker-Henderson (Joseph, 1952) es una técnica de suavización de series de tiempo que se basa en un enfoque de penalización de diferencias. Esta técnica es útil para tratar con ruido y outliers, y utiliza una función objetivo que equilibra la fidelidad a los datos con la suavidad de la serie de tiempo. La suavidad se logra minimizando las diferencias de las diferencias sucesivas de los datos suavizados.

$$L(y) = \sum_i w_i (y_i - z_i)^2 + \sum_i (z_{i+1} - 2z_i + z_{i-1})^2 \qquad (3.12)$$

donde:

- $L(y)$ es la función objetivo que queremos minimizar.
- w_i son los pesos asociados a cada observación.
- y_i son los datos originales y z son los datos suavizados.
- λ es el parámetro de suavizado.
- La expresión $(z_{i+1} - 2z_i + z_{i-1})^2$ es una diferencia de segundo orden que mide la curvatura de la serie de tiempo en el punto z_i.

El método Whittaker Henderson tiene las siguientes ventajas:

- Flexibilidad: El método Whittaker-Henderson es notablemente versátil, capaz de manejar distintos tipos de ruido y valores atípicos, además de permitir diversos grados de suavización.
- E ciencia computacional: A pesar de que el problema de optimización puede parecer complejo, en realidad se puede resolver de manera e ciente utilizando técnicas de álgebra lineal.

También es importante considerar las siguientes desventajas:

- Selección del parámetro de suavización: La elección del parámetro puede ser complicada. Un valor de demasiado alto puede resultar en una serie temporal excesivamente suave que ignora aspectos importantes de los datos, mientras que un valor de demasiado bajo puede no suavizar adecuadamente el ruido.
- Bordes de la serie temporal: El método Whittaker-Henderson puede presentar dicultades para manejar los bordes de la serie temporal, donde las diferencias de segundo orden no están disponibles.
- Ausencia de una sólida teoría estadística: A diferencia de otros métodos de suavizado que están fundamentados en modelos estadísticos, el método Whittaker-Henderson es más empírico y carece de una sólida teoría estadística que respalde su aplicación.

3.7. Criterios de selección de modelos AIC y BIC

Los criterios de selección de modelos AIC (Akaike Information Criterion) (Akaike, 1987) y BIC (Bayesian Information Criterion) (Neath y Cavanaugh, 2012) son dos métodos comúnmente utilizados en estadística y aprendizaje automático para la selección de modelos. Ambos criterios tienen como objetivo equilibrar la bondad del ajuste del modelo con la complejidad del modelo para evitar el sobreajuste.

El AIC se calcula como:

$$AIC = 2k \quad 2\ln(L) \tag{3.13}$$

donde:

- k es el número de parámetros en el modelo
- L es la verosimilitud máxima del modelo

El BIC se calcula como:

$$BIC = k \ln(n) \quad 2 \ln(L) \qquad (3.14)$$

donde:

- n es el número de observaciones.
- k es el número de parámetros en el modelo.
- L es la verosimilitud máxima del modelo

Si el objetivo principal es la precisión de las predicciones y se presupone que el modelo "verdadero"no se encuentra entre los modelos candidatos, el Criterio de Información de Akaike (AIC) puede ser la opción más apropiada. Por otro lado, si existe la suposición de que el modelo "verdadero"sí se encuentra entre los modelos candidatos, y el interés se centra más en la inferencia que en la predicción, el Criterio de Información Bayesiano (BIC) podría ser la alternativa más conveniente.

Adicionalmente, es importante considerar el tamaño de la muestra. Cuando se trata de muestras grandes, el BIC tiende a seleccionar modelos más parsimoniosos en comparación con el AIC. En contraste, cuando se trabaja con muestras pequeñas, el AIC puede ofrecer mayor con abilidad.

3.8. Media Móvil Simple (SMA)

La Media Móvil Simple (SMA) es un indicador comúnmente usado en análisis de series temporales y particularmente en el análisis de mercados nancieros. Este método calcula el promedio de un conjunto de datos durante un período de tiempo especi cado, que se mueve a lo largo del tiempo. El SMA se utiliza para suavizar las uctuaciones a corto plazo y destacar tendencias a largo plazo.

La principal suposición al usar la SMA es que los movimientos de precios pasados predicen los movimientos de precios futuros. Esta es una suposición básica en el análisis técnico, pero no siempre es válida en los mercados nancieros, donde el precio puede verse afectado por una variedad de factores imprevistos.

3.9. Suavizamiento Exponencial (ES)

El Suavizamiento Exponencial (ES) es una técnica de pronóstico de series de tiempo que utiliza un promedio ponderado de observaciones pasadas. A diferencia de la Media Móvil Simple, el Suavizamiento Exponencial da más peso a los datos más recientes, basándose en la premisa de que estos datos son más relevantes para los pronósticos futuros.

El modelo más simple de Suavizamiento Exponencial, también conocido como Suavizamiento Exponencial Simple, se puede expresar con la siguiente fórmula:

$$ES(t) = \quad X(t) + (1 \quad) \quad ES(t \quad 1) \qquad (3.15)$$

donde:

- ES(t) es el valor suavizado en el tiempo
- es el factor de suavizamiento, un valor entre 0 y 1.
- X(t) es el valor de la serie de tiempo en el tiempo
- ES(t 1) es el valor suavizado en el tiempo 1

3.10. Extrapolación Lineal (LE)

La Extrapolación Lineal (LE) es una técnica de estimación que se utiliza para predecir el valor de una variable en base a su relación lineal con otra variable. La extrapolación lineal se utiliza a menudo en el análisis de datos cuando se desea estimar un valor que está más allá del rango de los datos observados. Es un método sencillo y fácil de entender. Es muy útil para hacer predicciones cuando se tiene muy pocos datos y se requiere un modelo simple para predecir.

Es probable que las predicciones sean inexactas si la relación entre las variables no es realmente lineal, o si la relación cambia en los valores que se están extrapolando. Además, las predicciones son cada vez menos precisas cuanto más lejos se esté del rango de los datos originales. Por lo tanto, se requiere revisión frecuentemente o algún tipo de control en la práctica.

4. DESCRIPCIÓN DE DATOS

4.1. Las bases de datos

En este estudio, se utilizan dos bases de datos para la investigación de la edad máxima de seres humanos: la Base de Datos Internacional sobre Longevidad (IDL, French Institute for Demographic Studies, 2023) y la Lista de Supercentenarios del Grupo de Investigación en Gerontología (GRG, GERONTOLOGY RESEARCH GROUP, 2023).

La IDL es una colaboración internacional que incluye organismos como el Instituto Max Planck de Investigación Demográ ca, el Instituto Francés de Estudios Demográ cos (INED) y el Instituto Francés de Salud e Investigación Médica (INSERM). Su principal objetivo es proporcionar datos veri cables y precisos sobre la longevidad humana extrema, documentando tanto a semi-supercentenarios (individuos de entre 105-109 años) como a supercentenarios (individuos mayores de 110 años). La recopilación de datos de la IDL se realiza a nivel nacional, con instituciones locales encargadas de recolectar, validar y mantener los datos. Estas organizaciones utilizan múltiples fuentes de documentación, como registros de nacimiento y matrimonio, datos censales y registros de salud, para con- rmar la edad de cada individuo. Los datos validados se envían posteriormente al equipo de la IDL para su inclusión en la base de datos. La IDL mantiene la con dencialidad de los individuos, evitando la divulgación de información personal identi cable.

Por otro lado, el GRG se centra especí camente en los "supercentenarios", es decir, individuos que han alcanzado los 110 años o más. Fundado en 1990 por el Dr. L. Stephen Coles junto con un equipo de académicos multidisciplinarios, el GRG ha contribuido signi cativamente a la Lista de Ranking Mundial de Supercentenarios. Esta lista exhaustivamente actualizada incluye a supercentenarios vivos validados a nivel mundial, así como a aquellos que han fallecido. Para la validación, se exige la veri cación cruzada de registros de nacimiento y otros documentos relevantes para corroborar la edad de los individuos. Esta lista sirve como un referente reputado para aquellos interesados en las personas más longevas del mundo.

Las distintas metodologías y enfoques de la IDL y el GRG las hacen adecuadas para diferentes aspectos de la investigación sobre longevidad. En este estudio, se utiliza un enfoque de los dos bases de datos, aprovechando la IDL para estimar la edad máxima humana y el GRG para examinar las tasas de mortalidad de los individuos mayores de 110 años. Este método robusto permite aprovechar e cientemente las características de ambas bases de datos.

El extenso alcance y el anonimato de la IDL permiten una visión amplia y un análisis estadístico del límite de la esperanza de vida. Simultáneamente, la concentración del GRG en los supercentenarios proporciona una visión más detallada, lo que resulta particular-

mente útil para estudiar las tasas de mortalidad de este grupo demográ co y deducir la edad en la que la tasa de mortalidad se aproxima a uno. Esta edad proporciona una medida relevante de las posibles limitaciones de la esperanza de vida y de la tasa de mortalidad de los supercentenarios.

4.2. Análisis previo de los datos de la base de datos de IDL

La base de datos de IDL contiene información de 18,959 individuos mayores de 105 años. En la Tabla 1 se presenta una muestra de dicha base de datos, en la cual se pueden observar los datos relacionados con el género, lugar de residencia, año de nacimiento y fallecimiento y la edad al momento del fallecimiento de cada persona.

Tabla 4.1
Una muestra de la base de datos de IDL

	SEX	BIRTH_YEAR	DEATH_YEAR	COUNTRY	AGE
1	M	1897	2003	AUT	106.499
2	M	1898	2003	AUT	105.277
3	M	1898	2004	AUT	105.501
4	M	1899	2004	AUT	105.134
5	M	1899	2005	AUT	105.219
6	M	1900	2005	AUT	105.403
7	M	1900	2005	AUT	105.008
8	M	1900	2008	AUT	107.745
9	M	1900	2006	AUT	105.542
10	M	1900	2006	AUT	105.877
11	M	1901	2007	AUT	105.707
12	M	1902	2007	AUT	105.170
13	M	1902	2008	AUT	105.929
14	M	1902	2008	AUT	105.970
15	M	1903	2008	AUT	105.581
16	M	1903	2008	AUT	105.452
17	M	1903	2010	AUT	106.855
18	M	1904	2009	AUT	105.170
19	M	1904	2011	AUT	106.827
20	M	1905	2013	AUT	108.033
21	M	1905	2010	AUT	105.392
22	M	1905	2013	AUT	107.745
23	M	1905	2010	AUT	105.132
24	M	1906	2011	AUT	105.321
25	M	1906	2013	AUT	107.252
26	M	1906	2014	AUT	107.767
27	M	1906	2012	AUT	106.140
28	M	1907	2013	AUT	106.778
29	M	1907	2013	AUT	106.129
30	M	1907	2012	AUT	105.162

Fuente: Elaboración Propia

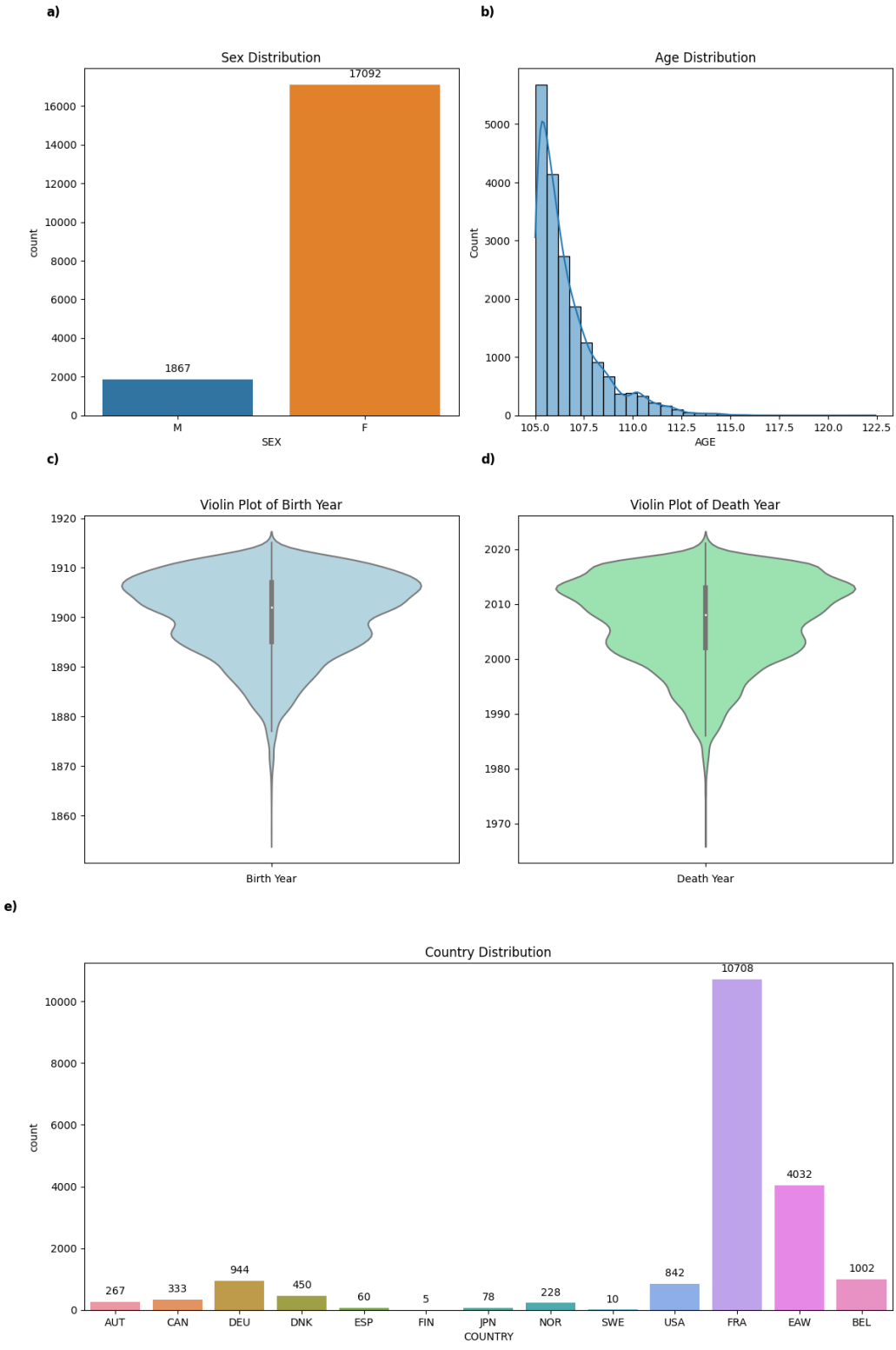
Para obtener una visión más detallada de la base de datos, se realizó un análisis previo de los datos. Los resultados se presentan en la Figura 4.1.

De los 18,959 incidentes registrados, 17,092 corresponden a mujeres y 1,867 a hombres (Figura 4.1a). Además, se observa que el rango de edades varía entre 105 y 122.5 años, y se observa una tendencia decreciente en el número de incidentes a medida que aumenta la edad (Figura 4.1b).

Asimismo, se realizó un análisis de la función de densidad de los incidentes en función del año de nacimiento y del año de fallecimiento. Se encontró que ambas funciones presentaron patrones similares, dado que la base de datos se encuentra truncada y solo incluye a personas con edades superiores a 105 años (Figuras 4.1c y 4.1d).

La mayoría de los eventos registrados en la base de datos corresponden a individuos de Francia (código “FRA”), con un total de 10,708 incidentes. En segundo lugar, se encuentran los incidentes de Inglaterra y Gales (código “EAW”), seguidos por Bélgica (código “BEL”) y Alemania (código “DEU”) con 4,032, 1,002 y 944 incidentes respectivamente (Figura 4.1e).

Figura 4.1
Análisis descriptivo de la base de datos de IDL



Fuente: Elaboración Propia

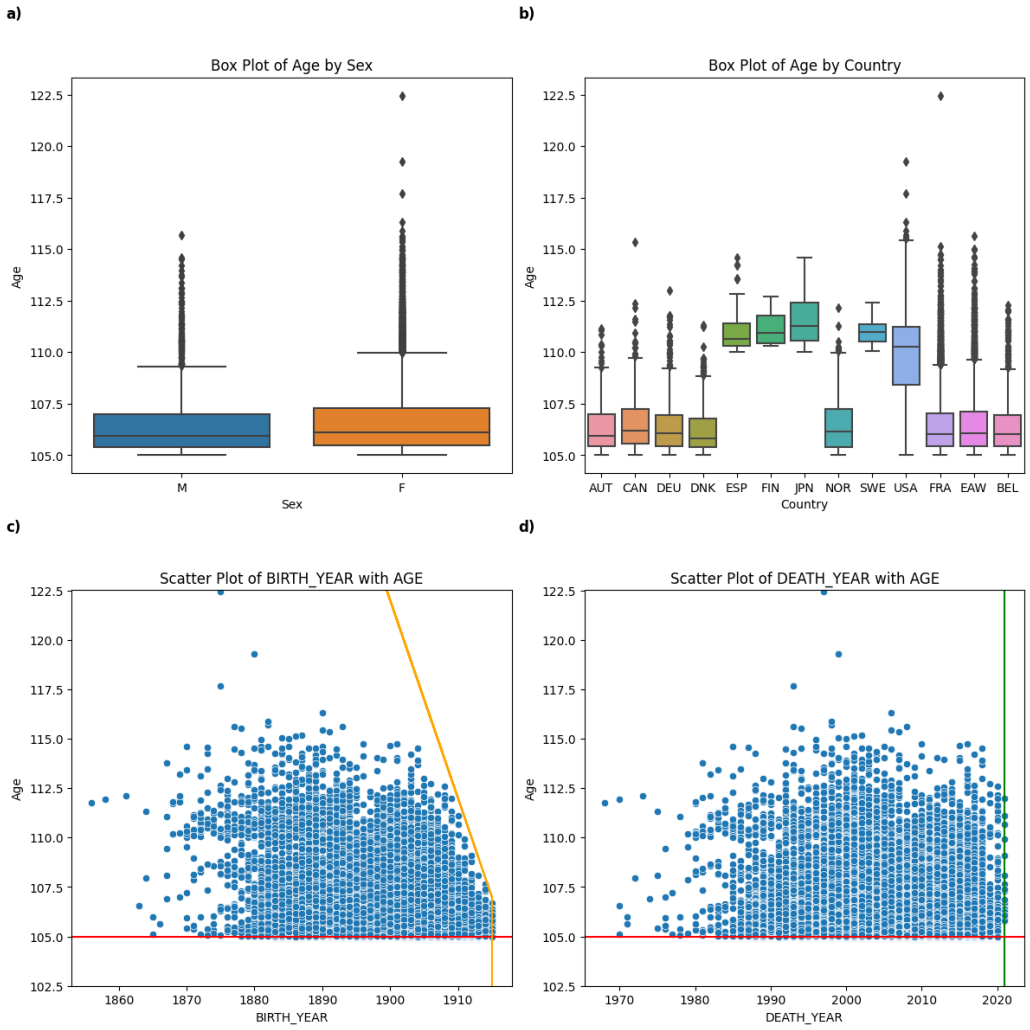
Tras realizar un análisis de la relación entre la edad de fallecimiento y diversas variables (género, lugar de residencia, año de nacimiento y de fallecimiento), los resultados obtenidos se presentan en la Figura 4.2.

Mediante un diagrama de caja (Figura 4.2a), se puede constatar que la media de la edad de fallecimiento es similar en hombres y mujeres. Sin embargo, es notable que la distribución de la edad de fallecimiento en mujeres presenta una cola más prolongada. Esto implica que existe una mayor proporción de mujeres que alcanzan edades considerablemente avanzadas en comparación con los hombres.

La Figura 4.2b proporciona un resumen de la edad de fallecimiento de semi-supercentenarios y supercentenarios en diversos países y regiones. Debido a los criterios de colleccionar los datos, Hay ciertas áreas que han incluido los semi-supercentenarios, individuos entre las edades de 105 y 109 años, y otros solo incluye los supercentenarios, aquellos de 110 años o más. Se destacan los datos de Francia y de Inglaterra y Gales, integrantes del Reino Unido, que exhiben una notable concentración de ambos grupos de edad. En contraste, Finlandia, Japón, España y Suecia no registraron individuos en la categoría de semi-supercentenarios, pero sí presentaron población dentro de los supercentenarios. Austria, Bélgica, Canadá, Dinamarca, Alemania, Noruega y Estados Unidos registraron individuos en ambas categorías.

Es importante señalar que los datos contenidos en esta base de datos son de carácter truncado, considerando únicamente a aquellos individuos cuya edad supera los 105 años. Este hecho implica necesariamente que dichos sujetos debieron haber nacido previamente al año 1917. Para aquellos nacidos antes del año 1917, la máxima edad observable se puede calcular como la diferencia entre el año actual y su año de nacimiento. Esta relación se encuentra representada por la línea naranja en la Figura 4.2c. Asimismo, cabe destacar que los eventos que se pueden observar tanto en la Figura 4.2c como en la Figura 4.2d corresponden a individuos que han superado la edad de 105 años antes del año actual. Los límites de esta situación se han marcado con líneas rojas y verdes en ambas guras.

Figura 4.2
Análisis de la variable “edad” frente a otros variables de la base de datos de IDL



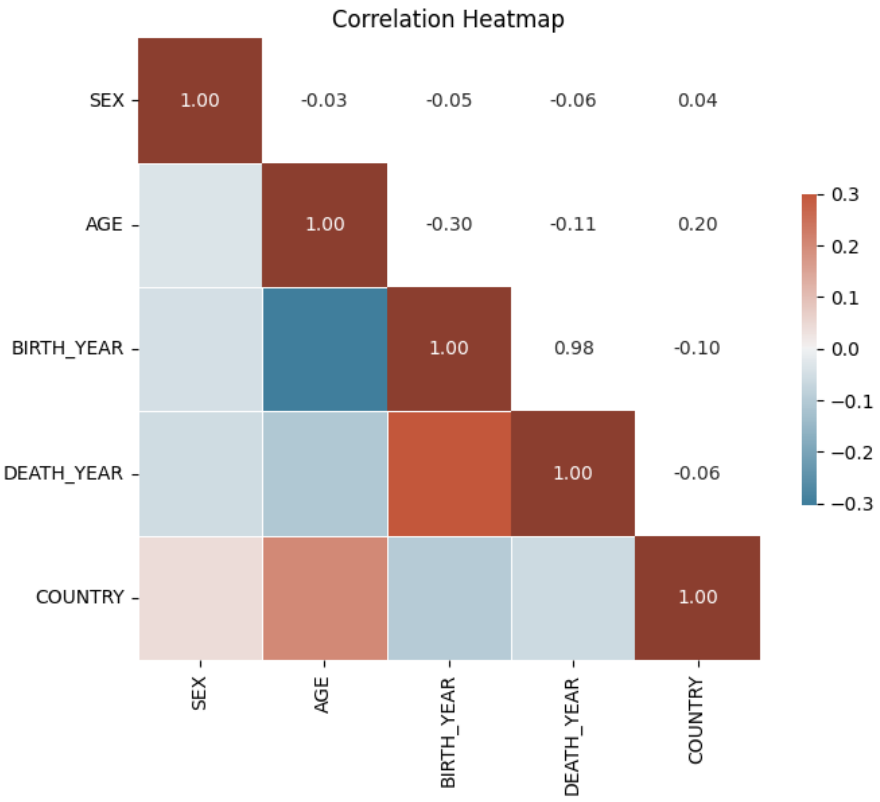
Fuente: Elaboración Propia

La siguiente gura (Figura 4.3) expone la correlación existente entre las variables anteriormente mencionadas. Se destacan los siguientes aspectos:

1. Existe una correlación positiva y signi cativa entre el año de nacimiento (“BIRTH_ YEAR”) y el año de fallecimiento (“DEATH_YEAR”). Esto sugiere que a medida que el año de nacimiento aumenta, también lo hace el año de fallecimiento.

2. Se observa una correlación negativa entre la edad (“AGE”) y el año de nacimiento (“BIRTH_YEAR”). Esto indica que a medida que el año de nacimiento se aproxima a la actualidad, la edad registrada tiende a ser menor. Este comportamiento es esperado debido a la naturaleza truncada de los datos, que solo incluyen a individuos con una edad superior a 105 años.

Figura 4.3
Análisis de la correlación de las variables de la base de datos de IDL



Fuente: Elaboración Propia

4.3. Análisis previo de los datos de la base de datos de GRG

Al comparar con la base de datos del IDL, se observa que la base de datos del GRG registra un número inferior de incidentes. Este fenómeno se debe a que el GRG únicamente incluye a los individuos que han superado los 110 años y también su forma de coleccionar los datos. En total, el registro del GRG engloba a 560 individuos. La Tabla 2 presenta una muestra de esta base de datos, donde se encuentran detallados datos relevantes sobre el lugar de fallecimiento, género, año de nacimiento y de fallecimiento y edad al momento del fallecimiento de cada persona.

Tabla 4.2
Una muestra de la base de datos de GRG

	NAME	D_PLACE	SEX	B_YEAR	D_YEAR	AGE
1	N*** T****	JAPAN	F	1900	2018	117.712
2	M*** N****	JAPAN	M	1905	2019	113.490
3	L*** D***	USA	F	1905	2018	112.564
4	J***** N***	USA	F	1886	1998	112.156
5	I*** W****	USA	F	1905	2021	115.351
6	O*** S***	JAPAN	F	1905	2019	114.227
7	A**** M***	USA	F	1905	2019	114.384
8	H***** F***	USA	F	1905	2021	115.671
9	T**** K**	JAPAN	F	1905	2019	113.548
10	M**** L****	FRANCE	F	1905	2018	113.000
11	M*** S***	JAPAN	F	1905	2017	112.712
12	L*** T**	USA	F	1905	2018	112.868
13	M** G***	ITALY	F	1905	2019	113.214
14	K** B****	JAPAN	F	1905	2019	114.030
15	D*** C*****	ITALY	F	1905	2019	113.644
16	M**** K****	JAPAN	F	1905	2020	115.126
17	K** A****	JAPAN	F	1905	2019	113.800
18	M*** T****	JAPAN	F	1896	2008	111.945
19	F**** N**-O**	SPAIN	M	1904	2018	113.129
20	G**** A******	ITALY	F	1906	2019	113.079
21	J*** B**	FRANCE	F	1905	2021	116.351
22	M** G*** V****	BRAZIL	F	1896	2011	114.951
23	A*** B*****-C**	ITALY	F	1906	2019	113.723
24	S*** A***	JAPAN	F	1903	2019	115.699
25	M**** V****	AUSTRALIA	F	1906	2018	112.567
26	G*** V***R***	FRANCE	F	1904	2018	114.499
27	J. L* S****	USA	F	1906	2018	112.129
28	R*** O****	USA	M	1906	2018	112.630
29	T**** J***	POLAND	F	1906	2022	116.192
30	J** M**	FRANCE	F	1906	2019	113.252

Fuente: Elaboración Propia

Como se realizó en la sección previa (Sección 4.2), se efectuó un análisis previo de los datos, cuyos resultados se muestran en la Figura 4.4.

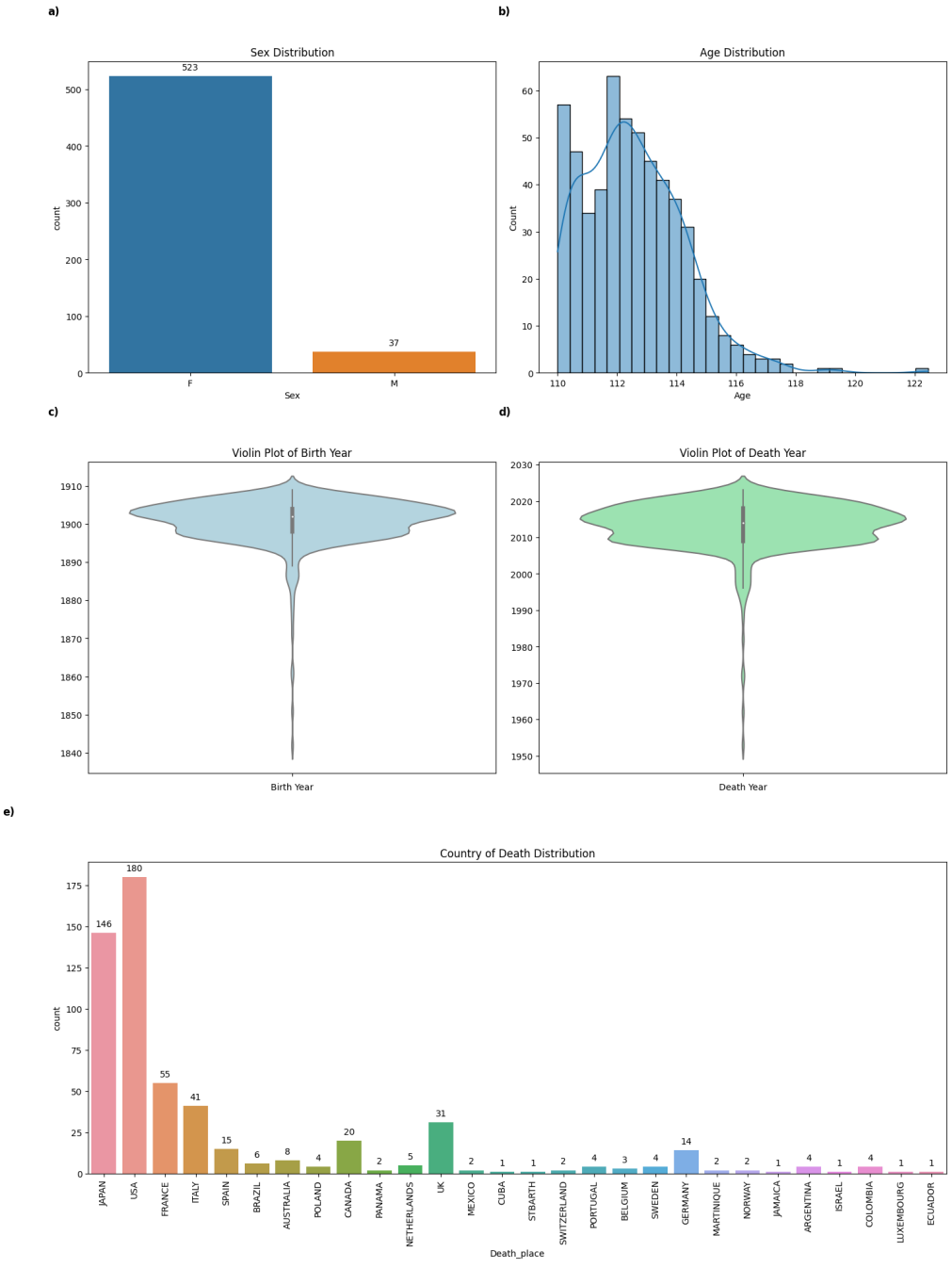
De los 560 incidentes registrados, un total de 523 corresponden a sujetos de sexo femenino y 37 a sujetos de sexo masculino (Figura 4.4a). En cuanto a la distribución por edades, esta oscila entre 110 y 122.5 años. Es notable una disminución en el número de incidentes a medida que la edad supera los 112 años (Figura 4.4b).

Adicionalmente, se desarrolló un análisis de la función de densidad de los incidentes en relación con el año de nacimiento y al año de deceso. Resulta evidente que ambas funciones presentan patrones similares. Este hecho probablemente se debe a que la base de datos se encuentra truncada, ya que únicamente incluye a individuos con edades superiores a los 110 años (Figuras 4.4c y 4.4d).

Respecto a la distribución geográ ca de los incidentes registrados en la base de datos, la mayoría corresponde a individuos procedentes de Estados Unidos (código “USA”), con un total de 180 incidentes. A continuación, se ubican los incidentes de Japón (código “JAPAN”), seguidos por los de Francia (código “FRANCE”) e Italia (código “ITALY”), con 146, 55 y 41 incidentes respectivamente (Figura 4.4e).

Figura 4.4

Análisis descriptivo de la base de datos de GRG



Fuente: Elaboración Propia

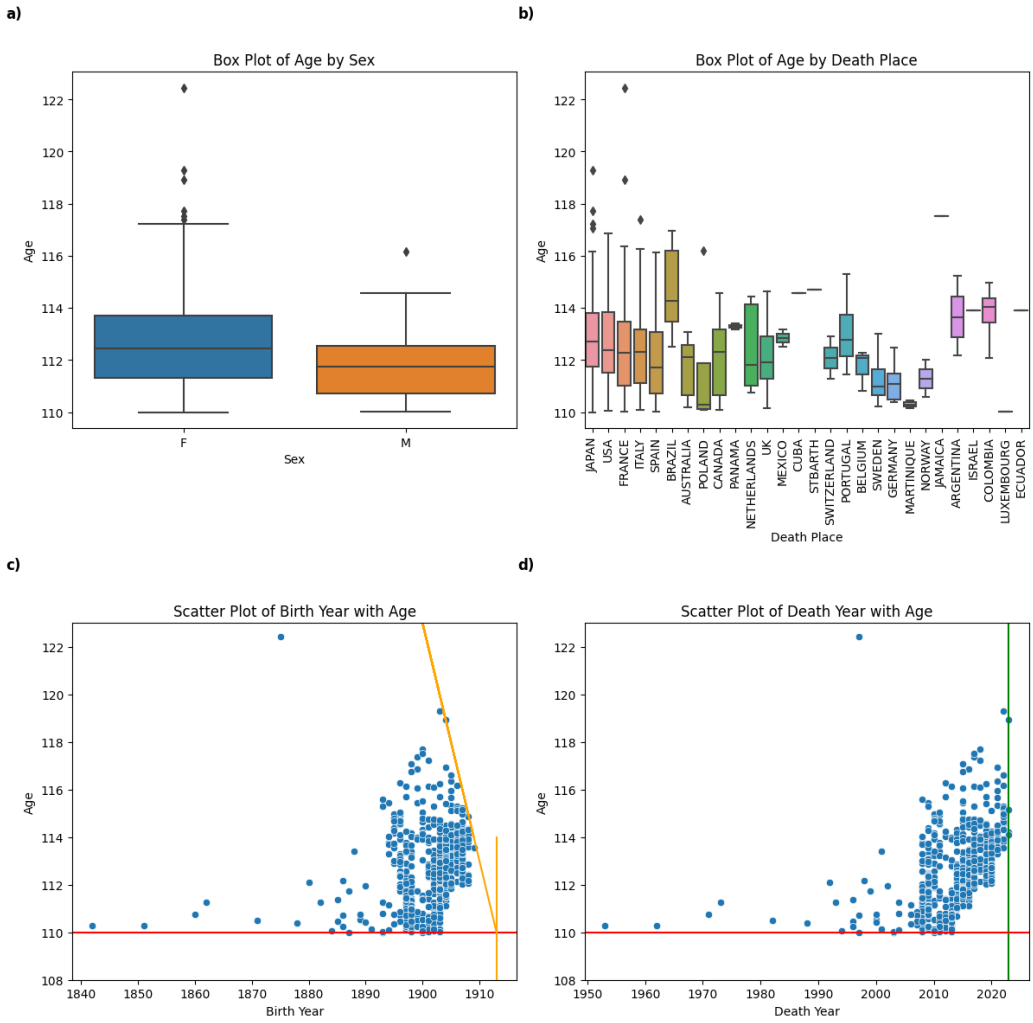
Se llevó a cabo un análisis de la relación existente entre la edad de fallecimiento y diversas variables como el género, el lugar de fallecimiento, el año de nacimiento y de fallecimiento. Los resultados de este análisis se ilustran en la Figura 4.5.

A través de un diagrama de caja (Figura 4.5a), se evidencia que la media de la edad de fallecimiento de las mujeres supera a la de los hombres. Además, la distribución de la edad de fallecimiento en mujeres exhibe una cola más extensa. Este patrón sugiere una proporción mayor de mujeres que alcanzan edades signi cativamente avanzadas en comparación con los hombres.

La Figura 4.5b ofrece una visión general de la edad de fallecimiento de los super-centenarios en distintos países y regiones. Aunque los valores promedio uctúan de una región a otra, el rango intercuartílico se ubica entre los 110 y 116 años.

Es necesario indicar que los datos contenidos en esta base de datos son de naturaleza truncada, dado que se enfocan exclusivamente en individuos con edades superiores a los 110 años. Este hecho implica que estos sujetos deben haber nacido antes del año 1913. Para aquellos que nacieron antes de 1913, la máxima edad observable puede calcularse como la diferencia entre el año actual y su año de nacimiento. Esta relación se encuentra representada por la línea naranja en la Figura 4.5c. Además, es importante resaltar que los sucesos observables tanto en la Figura 4.5c como en la Figura 4.5d corresponden a individuos que han superado los 110 años antes del año presente. Los límites de esta condición se han demarcado con líneas rojas y verdes en ambas guras.

Figura 4.5
Análisis de la variable “edad” frente a otros variables de la base de datos de GRG



Fuente: Elaboración Propia

La Figura 4.6 presenta la correlación existente entre las variables previamente mencionadas. Se pueden destacar algunos puntos relevantes en este análisis. Primero, se identi ca una correlación positiva signi cativa entre el año de nacimiento (“Birth_year”) y el año de fallecimiento (“Death_year”). Este signi ca que, conforme avanza el año de nacimiento, también tiende a progresar el año de fallecimiento.

Este vínculo positivo insinúa una tendencia en la que las personas nacidas en años más recientes también tienden a fallecer en años posteriores.

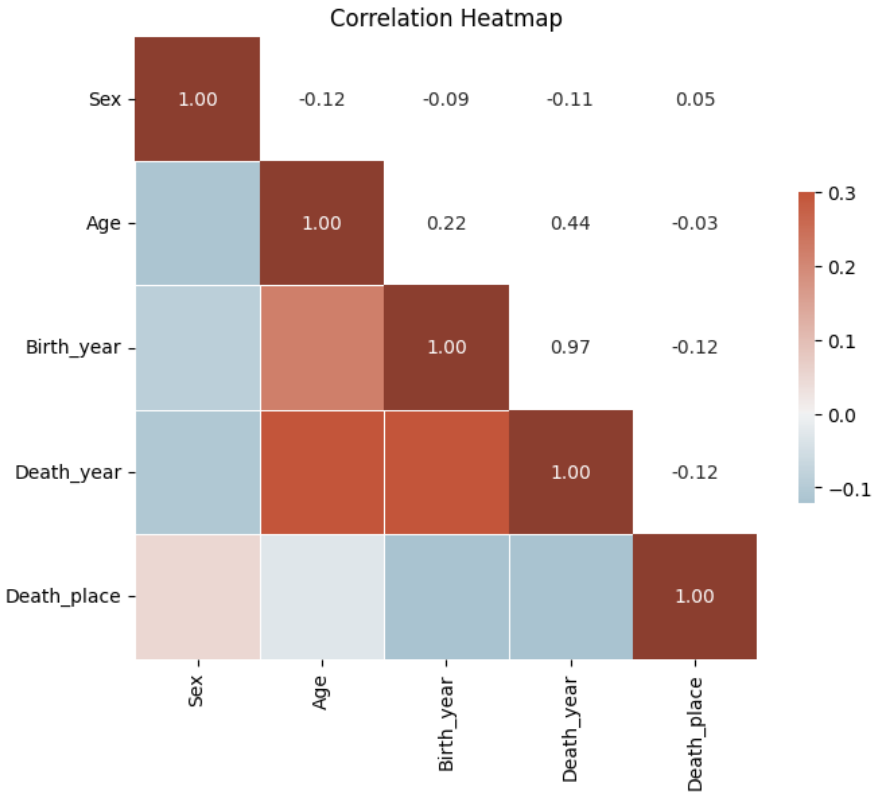
En segunda instancia, se aprecia una correlación positiva entre la edad (“Age”) y el año de nacimiento (“Birth_year”). Esta correlación indica que a medida que el año de nacimiento se aproxima al presente, la edad registrada suele ser más alta. Esto podría ser interpretado como una señal de que la edad máxima esperada para los seres humanos se encuentra en aumento a lo largo del tiempo.

Sin embargo, es crucial mencionar que esta relación positiva entre la edad y el año de

nacimiento únicamente se evidencia en la presente base de datos, y no en la de IDL. Esta discrepancia puede atribuirse a una mayor presencia del efecto de datos truncados en la base de datos de IDL, como se puede observar al comparar la Figura 4.2c con la Figura 4.5c.

Figura 4.6

Análisis de la correlación de las variables de la base de datos de GRG



Fuente: Elaboración Propia

5. ESTUDIO SOBRE LA EDAD MÁXIMA DE SERES HUMANOS

5.1. Introducción

Como se ha mencionado anteriormente, el enfoque principal de nuestro estudio se centra en evaluar la máxima edad que los seres humanos pueden llegar a vivir. Una pregunta sobre los estudios anteriores es que se formula en base a información limitada, lo que podría conducir a conclusiones incompletas o potencialmente incorrectas. Así que, en el siguiente análisis, vamos a utilizar los datos de la base de datos de IDL primero para analizar si existe la edad máxima de seres humanos o no y en el caso sí existe, si lo hemos llegado o no.

En nuestro estudio, cada caso registrado en la base de datos se trata como un evento individual de llegar a su edad máxima en un año especi co. Esta aproximación nos permite considerar las variaciones a través del tiempo, lo que podría revelar patrones que no son evidentes en un análisis más general.

Entonces, utilizando estos eventos individuales, se puede estudiar la tendencia de la edad máxima a lo largo de los últimos años. Este análisis nos permite observar cómo ha cambiado la edad máxima a lo largo del tiempo y luego predecir la edad máxima potencial de los seres humanos en los próximos años a partir de este patrón.

En esta sección, vamos a analizar la distribución de la edad máxima de los eventos de cada año y luego estudiar el cambio de distribuciones a través del tiempo. Por lo tanto, no sólo vamos a analizar la distribución de la edad máxima de los eventos de cada año, sino también a estudiar cómo esta distribución ha cambiado a lo largo del tiempo. Esto nos permite observar si las edades máximas están aumentando, o manteniendo constante, lo que a su vez nos dará más información sobre si hemos alcanzado el límite de la edad humana.

5.2. Metodología

Para iniciar el análisis, es esencial efectuar una adecuada preparación de los datos, lo cual facilitará los siguientes pasos de este estudio. Después de observar en la Tabla 1 de la sección 4.2, se han implementado las siguientes medidas de preparación de datos:

1. Se realiza un mapeo de las variables categóricas.
2. Se agrupa la edad de fallecimiento en intervalos de medio año, dado que nuestro objetivo de análisis es calcular la edad máxima alcanzada por los seres humanos. Este agrupamiento de datos con edades similares permite reducir de manera signi cativa la complejidad computacional.

3. Se analizan las características de los datos de acuerdo con el año de fallecimiento.

4. Se mantienen aquellos años de fallecimiento en los que se registra el fallecimiento de más de 100 personas, con el propósito de analizar su distribución en la etapa subsiguiente. La elección de años con una mayor cantidad de datos (recuentos de personas) puede brindar resultados más con ables y generalizables, conforme a un principio básico de estadística conocido como la Ley de los Grandes Números. Este principio establece que, a medida que se incrementa el tamaño de la muestra, su media se aproxima cada vez más al promedio de la población total. Por lo tanto, seleccionar años con un mayor número de personas podría ofrecer una visión más precisa de la distribución de las edades de fallecimiento.

Tras la obtención de los datos preparados, se puede recurrir a la Poisson ponderada para ajustar los datos correspondientes a cada año. La razón para emplear la distribución Poisson y la regresión Poisson ponderada se explicó previamente en las secciones 3.1 y 3.2.

Además, una vez que se haya estimado el parámetro lambda de la Poisson, tal como se mencionó en la sección 3.3, es necesario evaluar el ajuste del modelo (goodness-of- t) para veri car si los parámetros estimados son aceptables.

Después de obtener los parámetros lambda estimados y aceptados, se puede emplear el método SVR (sección 3.4) para ajustar la curva de los parámetros lambda de cada año en función del tiempo. En este caso, los hiperparámetros se han determinado mediante el método de GridSearch con validación cruzada de 10 particiones (sección 3.5). GridSearch es un método de optimización de hiperparámetros que examina de manera exhaustiva todas las combinaciones posibles de hiperparámetros prede nidos para un algoritmo de aprendizaje automático. El objetivo de este enfoque es identi car la combinación de hiperparámetros que proporciona el mejor rendimiento del modelo, medido según una métrica de evaluación prede nida.

Finalmente, se puede predecir las lambdas estimadas para los años futuros. Es importante comprender que, como se mencionó anteriormente, en este estudio las lambdas estimadas representan las medias de la edad máxima de los seres humanos. Por tanto, el percentil 95 puede considerarse como la edad máxima que los seres humanos pueden alcanzar.

5.3. Resultados

5.3.1. Preparación de los datos

Como se ha explicado en la sección anterior, los primeros resultados son del proceso de preparación de los datos. Después de mapear las variables categóricas según la Tabla 3 y agrupar la edad de fallecimiento, se puede tener el conjunto de datos como se muestra en la Tabla 5.1.

Tabla 5.1
Mapa de las variables categóricas

SEX			
0	F	1	M
COUNTRY			
0	AUT	1	BEL
2	CAN	3	DEU
4	DNK	5	EAW
6	ESP	7	FIN
8	FRA	9	JPN
10	NOR	11	SWE
12	USA		

Fuente: Elaboración Propia

Tabla 5.2
Una muestra de datos después de mapear las variables categóricas

	SEX	BIRTH_YEAR	DEATH_YEAR	COUNTRY	AGE
1	1	1897	2003	0	106.0
2	1	1898	2003	0	105.0
3	1	1898	2004	0	105.5
4	1	1899	2004	0	105.0
5	1	1899	2005	0	105.0
6	1	1900	2005	0	105.0
7	1	1900	2005	0	105.0
8	1	1900	2008	0	107.5
9	1	1900	2006	0	105.5
10	1	1900	2006	0	105.5
11	1	1901	2007	0	105.5
12	1	1902	2007	0	105.0
13	1	1902	2008	0	105.5
14	1	1902	2008	0	105.5
15	1	1903	2008	0	105.5
16	1	1903	2008	0	105.0
17	1	1903	2010	0	106.5
18	1	1904	2009	0	105.0
19	1	1904	2011	0	106.5
20	1	1905	2013	0	108.0
21	1	1905	2010	0	105.0
22	1	1905	2013	0	107.5
23	1	1905	2010	0	105.0
24	1	1906	2011	0	105.0
25	1	1906	2013	0	107.0
26	1	1906	2014	0	107.5
27	1	1906	2012	0	106.0
28	1	1907	2013	0	106.0
29	1	1907	2012	0	105.0

Fuente: Elaboración Propia

Luego se ltra los datos según los numero de muertes de cada año, solo conservar los años en lo que hay más de 100 fallecimientos y se realiza un análisis estadístico básico para obtener la información como la media, desviación típica, los cuantiles, etc. (Tabla 5.3).

Tabla 5.3
Resultado de análisis estadístico básico tras del ltro de los datos

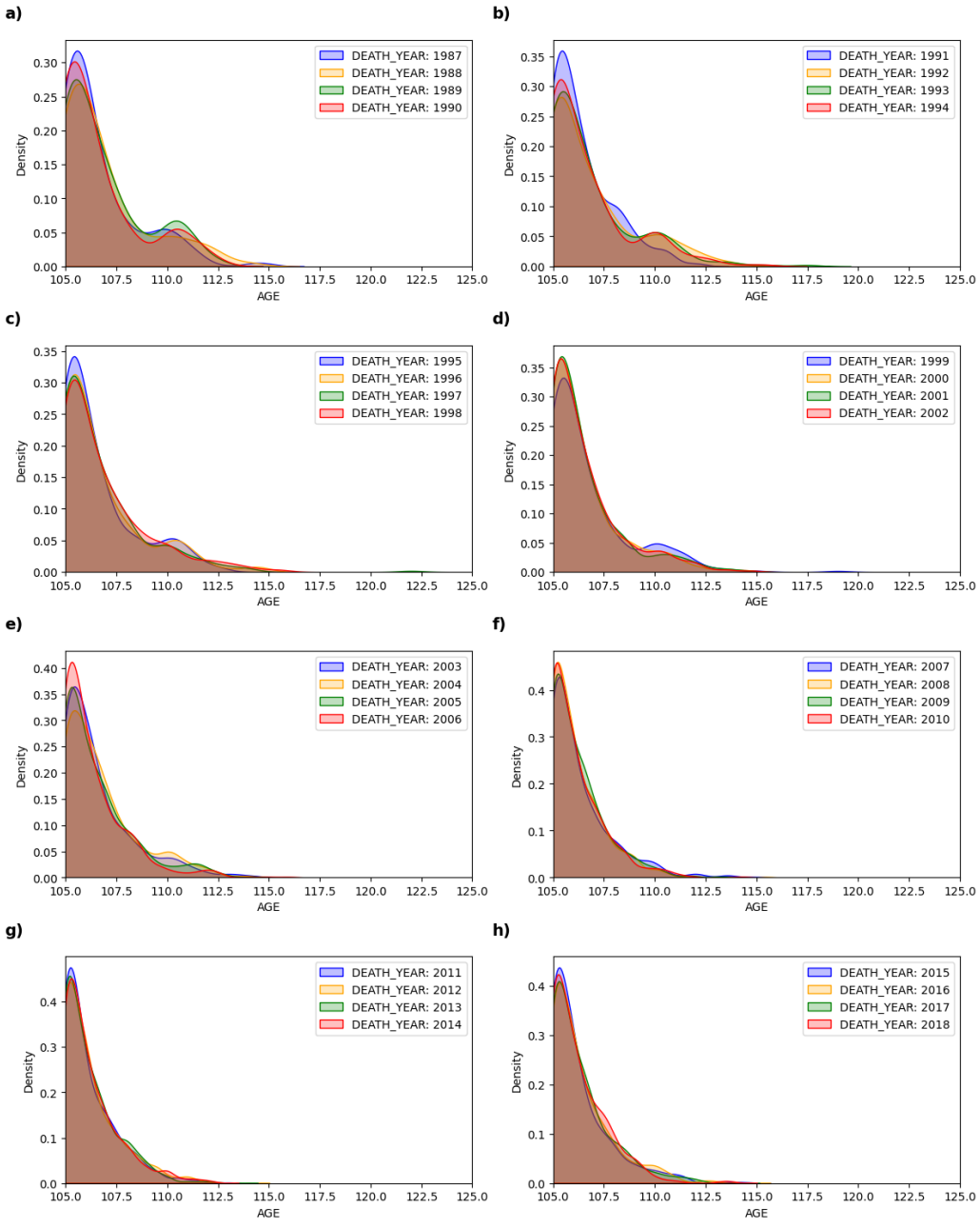
	D_Year	count	mean	std	median	min	max	Q .25	Q .75
1	1987	103	106.577	1.878	106	105	114.5	105.5	107
2	1988	100	106.830	2.104	106	105	114	105.5	107.5
3	1989	140	106.817	1.976	106	105	112.5	105.5	107.5
4	1990	134	106.589	1.967	106	105	112.5	105	107
5	1991	158	106.395	1.492	106	105	112	105.5	107
6	1992	217	106.806	2.119	106	105	114	105	107.5
7	1993	247	106.769	2.150	106	105	117.5	105	107.5
8	1994	271	106.642	2.030	106	105	115.5	105	107.5
9	1995	249	106.554	1.823	106	105	112.5	105	107
10	1996	291	106.694	2.044	106	105	114.5	105	107.5
11	1997	351	106.664	2.088	106	105	122	105	107.5
12	1998	408	106.787	2.104	106	105	115.5	105.375	107.5
13	1999	382	106.681	2.048	106	105	119	105.5	107
14	2000	596	106.492	1.773	106	105	114.5	105	107
15	2001	673	106.520	1.814	106	105	115	105.5	107
16	2002	663	106.521	1.786	106	105	115	105	107
17	2003	771	106.556	1.757	106	105	114.5	105.5	107
18	2004	687	106.713	1.803	106	105	114	105.5	107.5
19	2005	653	106.520	1.752	106	105	114.5	105	107
20	2006	657	106.328	1.625	106	105	116	105	107
21	2007	719	106.294	1.568	105.5	105	115	105	107
22	2008	826	106.163	1.323	106	105	115.5	105	107
23	2009	867	106.237	1.345	106	105	113.5	105	107
24	2010	875	106.173	1.360	106	105	114.5	105	107
25	2011	905	106.179	1.330	105.5	105	113	105	107
26	2012	1087	106.25	1.405	106	105	114	105	107
27	2013	1113	106.212	1.318	106	105	113.5	105	107
28	2014	1096	106.249	1.390	106	105	112.5	105	107
29	2015	790	106.264	1.502	106	105	114.5	105	107
30	2016	813	106.354	1.521	106	105	114.5	105	107
31	2017	815	106.339	1.480	106	105	114	105	107
32	2018	922	106.283	1.371	106	105	114	105	107

Fuente: Elaboración Propia

5.3.2. Ajuste y estimación de los parámetros de Poisson

De allí, se puede tener la distribución de densidad de los números de fallecimientos cada año corresponde a su edad. En la Figura 5.1, cada subgrá co se presenta cuatro de los años de fallecimiento desde el año 1987 hasta el año 2018.

Figura 5.1
Distribución de densidad de la edad de los fallecimientos en diferentes años



Fuente: Elaboración Propia

Luego se realiza la estimación del parámetro Lambda de la distribución de Poisson ajustando los datos de los fallecimientos de cada año usando máximo verosimilitud. Y se

utiliza la prueba de Chi-Cuadrado y de Kolmogorov-Smirnov para evaluar el ajuste del modelo. La siguiente tabla (Tabla 5.4) se presenta el resultado de los parámetros estimados y de las pruebas.

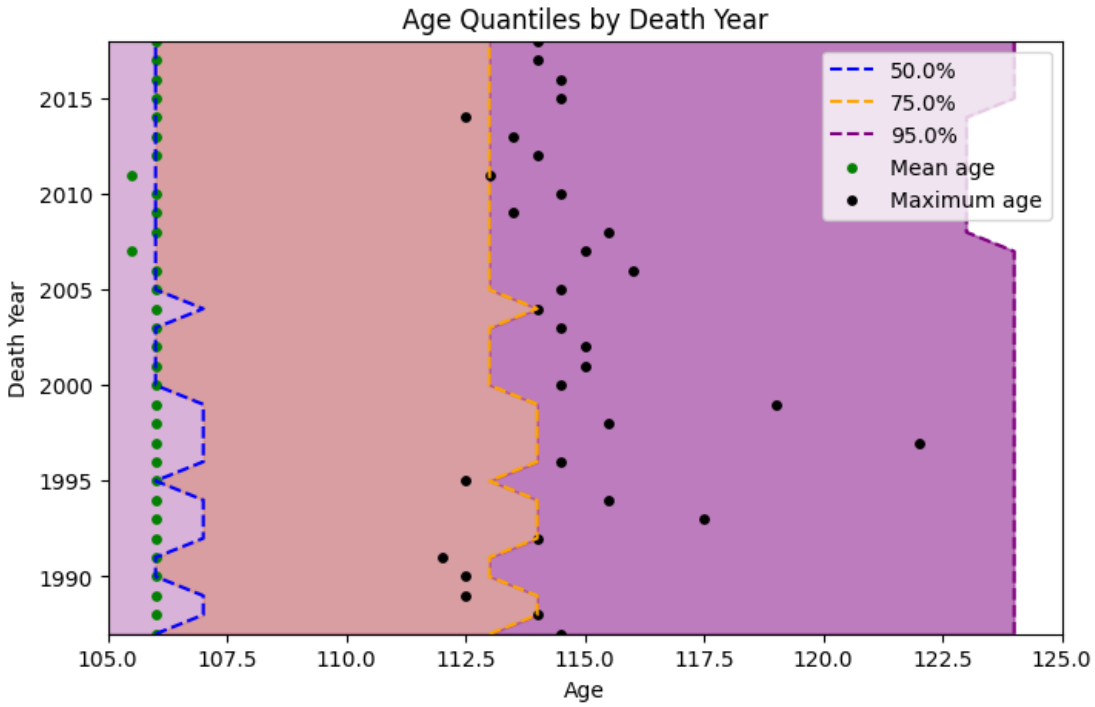
Tabla 5.4
Resultado de parámetros estimados y de las pruebas

	DEATH_YEAR	Estimated	KS P-value	CHI2 P-value
1	1987	106.610693	8.74E-21	8.37E-15
2	1988	106.871206	2.62E-19	7.40E-14
3	1989	106.85374	6.63E-27	3.45E-19
4	1990	106.625955	8.19E-27	3.46E-33
5	1991	106.416051	1.11E-32	2.77E-20
6	1992	106.848432	1.87E-41	2.09E-58
7	1993	106.812519	2.06E-47	7.40E-63
8	1994	106.680495	3.48E-53	1.25E-86
9	1995	106.584807	9.89E-50	2.31E-59
10	1996	106.733478	1.52E-56	3.48E-82
11	1997	106.704056	1.89E-68	3.28E-108
12	1998	106.82746	8.99E-78	3.03E-119
13	1999	106.720865	3.14E-74	1.89E-123
14	2000	106.520814	1.23E-119	5.16E-181
15	2001	106.550918	2.56E-134	7.29E-228
16	2002	106.55088	2.49E-132	3.21E-204
17	2003	106.585142	6.94E-153	1.11E-241
18	2004	106.743209	1.47E-132	7.41E-164
19	2005	106.549415	2.24E-130	1.38E-203
20	2006	106.353303	1.05E-135	2.14E-241
21	2007	106.317601	2.22E-149	1.04E-276
22	2008	106.179071	1.18E-175	5.14E-287
23	2009	106.254294	5.45E-182	1.97E-276
24	2010	106.191524	1.19E-185	0.0000E+00
25	2011	106.19585	7.43E-192	1.93E-298
26	2012	106.268402	1.74E-227	0.0000E+00
27	2013	106.227568	1.49E-234	0.0000E+00
28	2014	106.266782	1.98E-229	0.0000E+00
29	2015	106.285296	5.41E-165	3.76E-262
30	2016	106.375893	3.75E-167	2.57E-236
31	2017	106.359899	5.02E-168	2.06E-239
32	2018	106.300694	5.76E-192	6.17E-274

Fuente: Elaboración Propia

La siguiente gura (Figura 5.2) ha presentado una vista muy clara comparando la lambda estimada y sus corresponde cuantiles con los datos. Se puede observar que los puntos verdes que se presenta la edad media de los fallecimientos de cada edad se sitúan en la línea azul que el valor de Lambda estimada. Además, los mayores puntos negros que indican la edad máxima de fallecimiento observado en la vida real están por el rango del cuantil 75 y cuantil 95.

Figura 5.2
Cuantiles de la distribución comparando con la edad de fallecimiento real



Fuente: Elaboración Propia

5.3.3. Ajustar la curva de los parámetros lambda de cada año en función del tiempo

Los mejores hiperparámetros determinado mediante el método de GridSearch con validación cruzada de 10 particiones se muestran en la siguiente tabla (Tabla 5.5).

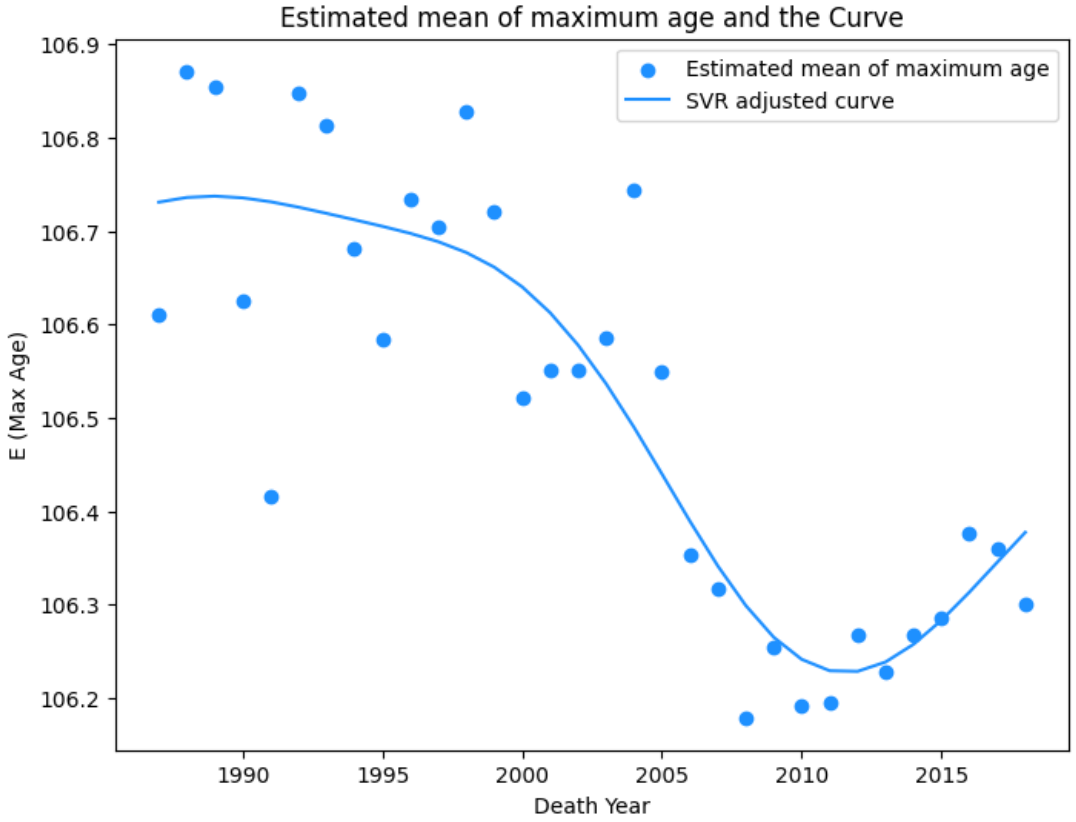
Tabla 5.5
Hiperparámetros determinados de SVR

Best Hyperparameters			
C	0.76	Epsilon	0.12
Mean Squared Error			
0.013166848			

Fuente: Elaboración Propia

Con estos hiperparámetros se puede encontrar la curva jada por el método SVR (Figura 5.3). En esta gura, se puede observar que la curva tiene un mínimo valor por el año 2011, luego se va aumentando hasta el año 2018. Una posible explicación sobre esta bajada desde el año 1987 hasta el año 2011 es la data original tienen de ciencia dado que durante esa época el sistema de registración de nacimiento y fallecimiento en algunos países o regiones no está completo. De allí, ha perdido de esta parte de la información y los datos tiene más varianza en esta época, eso se puede observar en la Tabla 5.3. Luego la subida desde el año 2011 indica que la media de la edad máxima de seres humanos está aumentando en función del tiempo.

Figura 5.3
La media de la edad máxima de seres humanos y la curva ajustada por SVR

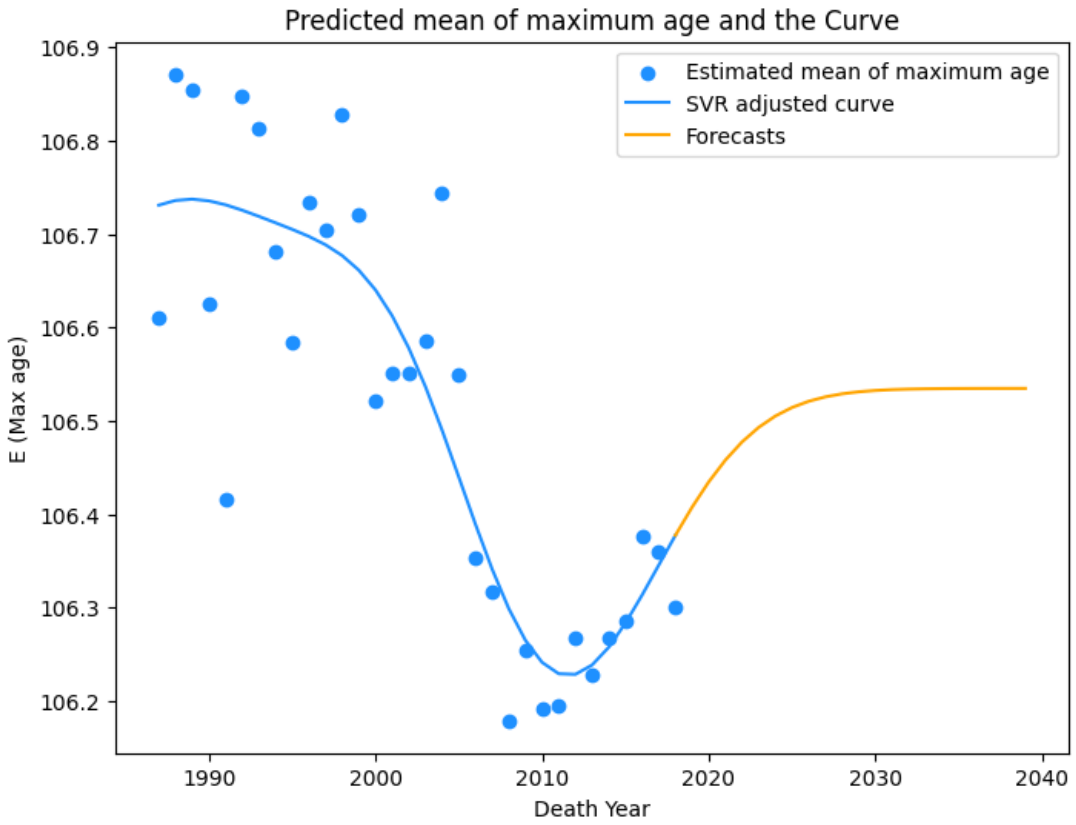


Fuente: Elaboración Propia

5.3.4. Predicción la media de edad máxima de seres humanos

Una vez se obtiene la curva de SVR, se puede predecir la media de edad máxima de seres humanos en los próximos años (Figura 5.4). Se puede observar en la gura un punto de in exión, donde la tasa de crecimiento del valor predicho se ralentiza con el tiempo y nalmente tiende a ser constante.

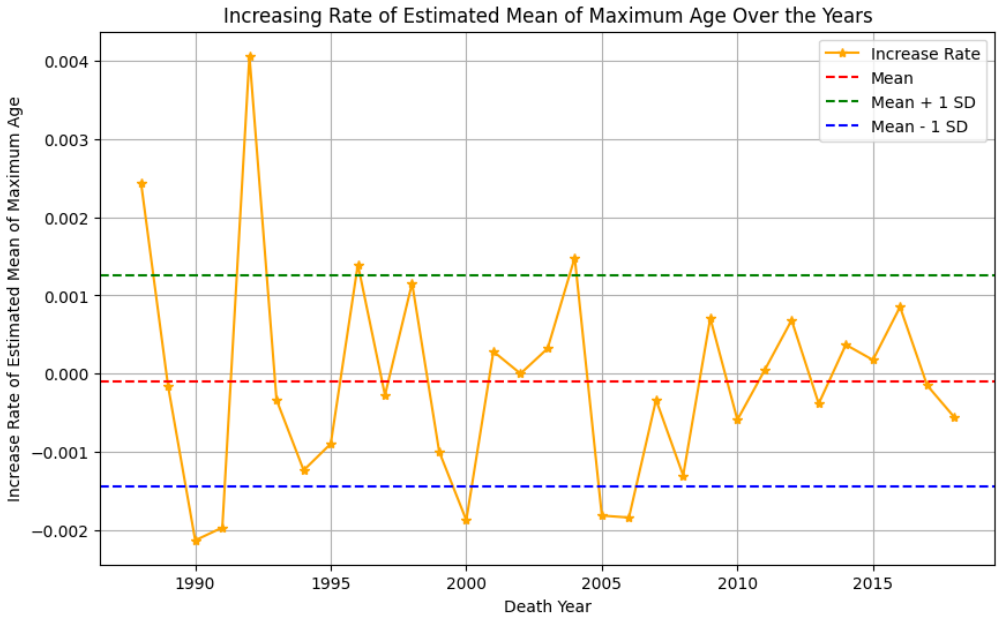
Figura 5.4
La predicción media de la edad máxima de seres humanos por SVR



Fuente: Elaboración Propia

Para investigar la potencial explicación de ese punto de in exión, se genera el Figura 5.5 sobre la ratio de crecimiento de la media estimada de la edad máxima. Se puede observar en esta gura que la varianza de dicha ratio disminuye con el tiempo, y la línea naranja en la gura converge entre las líneas punteada verde y azul con el tiempo. Además, la línea punteada roja es ligeramente negativa, lo que signi ca que la tasa de aumento de la media estimada de la edad máxima se ralentizará gradual y eventualmente hará que tienda a un valor constante.

Figura 5.5
La ratio de crecimiento de la media estimada de la distribución Poisson



Fuente: Elaboración Propia

5.3.5. Calculación de la edad máxima de seres humanos

En el último paso, se calcula el cuantil 95 de la distribución de la media de la edad máxima de seres humanos para tener la potencial edad máxima de seres humanos (Tabla 5.6). Se puede observar en la tabla que después del año 2030 la media de la edad máxima ha aumenta muy poco. Por tanto, se puede considerar que el valor medio se ha estabilizado en este intervalo. Así se puede calcular la media del cuantil 95 de los años después de 2030, se puede llegar a la conclusión que la edad máxima potencial de seres humanos seria 124 años aproximadamente.

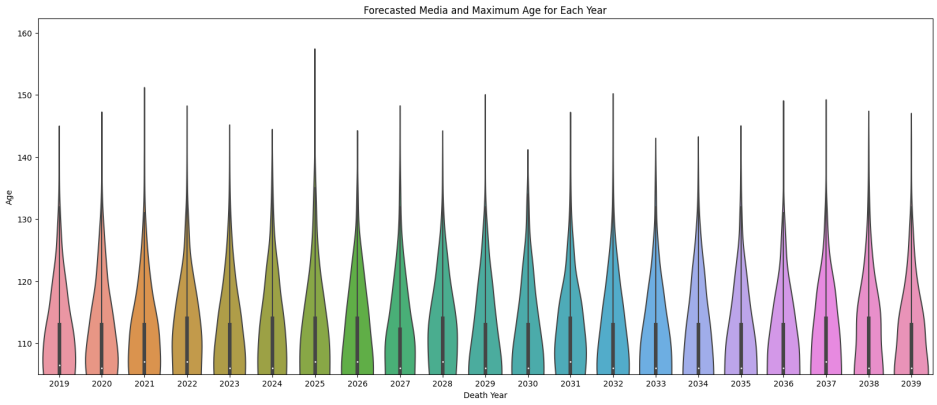
Tabla 5.6
La media predicha de la edad máxima de seres humanos y el cuantil 95

YEAR	Forecasted Media	Quantile .95
2019	106.3774757	123
2020	106.4075658	125
2021	106.4346435	124
2022	106.4580256	123
2023	106.4774702	125
2024	106.4930829	124
2025	106.50521	124
2026	106.514336	123
2027	106.5209974	125
2028	106.525718	124
2029	106.5289684	123.05
2030	106.5311443	123
2031	106.5325612	124
2032	106.5334591	123
2033	106.534013	125
2034	106.5343459	124
2035	106.5345407	125
2036	106.5346518	123
2037	106.5347135	124
2038	106.534747	124
2039	106.5347647	124

Fuente: Elaboración Propia

Luego también se puede observar la distribución de densidad de la edad máxima de seres humanos predichos en los próximos años según una gura de violín (Figura 5.6).

Figura 5.6
Distribución de densidad de la edad máxima de seres humanos predichos



Fuente: Elaboración Propia

6. ESTUDIOS SOBRE LA RATIO DE MORTALIDAD DE LOS SUPERCENTENARIOS

6.1. Introducción

Como ya se ha mencionado, el carácter incompleto y truncado de los datos di culta sacar conclusiones lo más ables posible. Por lo tanto, en el siguiente estudio, adoptamos un enfoque diferente para inferir la edad máxima de los seres humanos a partir de los datos de otra fuente de datos (la base de datos GRG).

En esta sección se centra en primer lugar en las tasas de mortalidad de las personas mayores de 110 años, y dado que la edad máxima en los registros actuales es de unos 122 años, se puede obtener unos 12 valores si se calcula la tasa de mortalidad de la siguiente unidad de edad para cada grupo de edad de los supercentenarios, utilizando un año como unidad de cálculo. Una cantidad menor tiene un mayor potencial de sesgar el cálculo, por lo que calculamos la tasa de mortalidad de cada semestre para los supercentenarios utilizando una unidad de cálculo de seis meses para aumentar el número de valores calculables.

Cabe mencionar el efecto de las generaciones que este estudio ignora, esto se debe a que a través de la observación de los datos en la sección 4, un gran número de muertes ocurren entre 2008-2022, por lo que, para minimizar el efecto del número de variables en el sobreajuste del modelo, no se toma en cuenta las generaciones.

Tras obtener las tasas de mortalidad de cada semestre, se realiza el ajuste y luego predecir la tasa de mortalidad al modelo, y cuando la tasa de mortalidad alcanza alrededor del 95 % e puede considerar aproximadamente que esa edad es la edad máxima alcanzable para los seres humanos, ya que existe una probabilidad del 95 % de que una persona que alcance esa edad muera en el siguiente semestre.

6.2. Metodología

Dado que la base de datos GRG proporciona datos sobre la mortalidad evento por evento, las tasas de mortalidad deben calcularse antes de iniciar la modelización. Como ya se ha mencionado, se debe comparar la viabilidad de los distintos intervalos de tiempo para determinar su aplicación en los cálculos posteriores. En primer lugar, se de na donde $_nl_x$, n es el intervalo de tiempo utilizado en el modelo, x denota la edad y $_nq_x$ denota la probabilidad de que una persona de edad x muera en los próximos n periodos de tiempo. Por lo tanto, cotejamos nuestros datos y utilizamos:

$$ {}_nq_x = \frac{d_x}{l_x} \tag{6.1} $$

Para calcular la tasa de mortalidad para ese intervalo de edad, donde d_x denota el número de personas que mueren en ese intervalo de edad y l_x denota el número de personas que están vivas al principio de ese intervalo de edad. En esta sección, se seleccionan ${}_1q_x$, ${}_{0,5}q_x$ y ${}_{0,25}q_x$ para comparar, y finalmente se elige el conjunto de valores calculados más apropiado para el debate posterior.

Una vez seleccionado un ${}_nq_x$ adecuado, se puede visualizar la relación entre mortalidad y edad para encontrar una curva de suavizado adecuada que muestre cómo varía la mortalidad con la edad. En esta sección se utilizan dos métodos, el SVR utilizado anteriormente y el método de suavizado de Whittaker Henderson (Sección 3.6), que se comparan luego mediante AIC, BIC (Sección 3.7) para seleccionar el método óptimo.

Una vez determinada la variación de la mortalidad con la edad, se pueden predecir las futuras tasas de mortalidad especí cas por edad, determinando si las tasas de mortalidad aumentarán, disminuirán o permanecerán invariables. Para esta predicción, se comparan el método de Media Móvil Simple (SMA, Sección 3.8), Suavizamiento Exponencial (ES, Sección 3.9), y Extrapolación Lineal (LE, Sección 3.10). Luego se selecciona el mejor método con el Error cuadrático medio (MSE).

En el caso de que la mortalidad sigue aumentando, podemos suponer provisionalmente que la esperanza de vida humana alcanza su límite a esta edad, cuando la mortalidad alcanza el 95 %.

6.3. Resultados

6.3.1. Selección el intervalo adecuado para la tasa de mortalidad

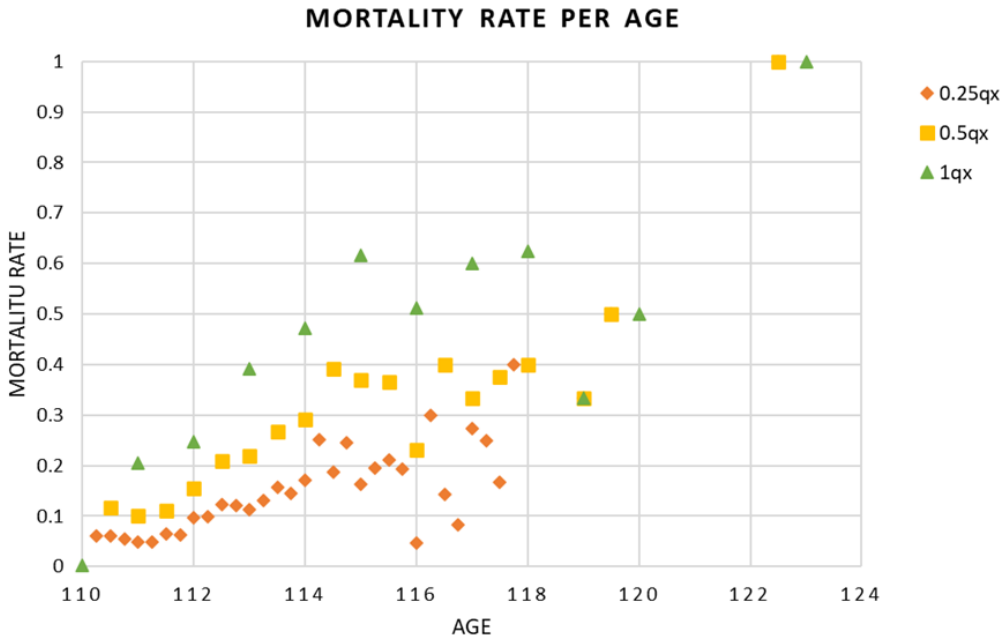
El primer resultado que obtenemos es sobre ${}_nq_x$ a aplicar. Comparando los valores dentro de la Tabla 6.1, como se puede visualizar en la Figura 6.1, la tendencia de ${}_1q_x$ y ${}_{0,5}q_x$ que se indican por los puntos verdes y amarillos respectivamente son similares, pero el ${}_{0,5}q_x$ tiene mas puntos que el otro. La tendencia de ${}_{0,25}q_x$ varia más, por ejemplo, en el rango de la edad 115-118 años dado que cuando más intervalos tienen, los ruidos de diferente momento en un año van a ser más altos. Así, se elige ${}_{0,5}q_x$ para el siguiente paso de estudio.

Tabla 6.1
Tabla de ${}_nq_x$ calculado

AGE	${}_{0;25}q_x$	${}_{0;5}q_x$	${}_1q_x$
110	0.001786		
110.25	0.060714		
110.5	0.060837	0.117857	
110.75	0.054656		
111	0.049251	0.101215	0.205725
111.25	0.04955		
111.5	0.063981	0.11036	
111.75	0.063291		
112	0.097297	0.15443	0.247748
112.25	0.098802		
112.5	0.122924	0.209581	
112.75	0.121212		
113	0.112069	0.219697	0.392216
113.25	0.131068		
113.5	0.156425	0.26699	
113.75	0.145695		
114	0.170543	0.291391	0.472906
114.25	0.252336		
114.5	0.1875	0.392523	
114.75	0.246154		
115	0.163265	0.369231	0.616822
115.25	0.195122		
115.5	0.212121	0.365854	
115.75	0.192308		
116	0.047619	0.230769	0.512195
116.25	0.3		
116.5	0.142857	0.4	
116.75	0.083333		
117	0.272727	0.333333	0.6
117.25	0.25		
117.5	0.166667	0.375	
117.75	0.4		
118	0.4	0.625	
119	0.333333	0.333333	0.333333
119.5	0.5	0.5	
120		0.5	
122.5	1	1	
123		1	

Fuente: Elaboración Propia

Figura 6.1
La ratio de mortalidad para los supercenternarios



Fuente: Elaboración Propia

Una vez determinado $_{0,5}q_x$ como base para los cálculos, se elige SVR y la suavización de Whittaker-Henderson para ajustar las curvas por separado.

6.3.2. SVR

Como antes, para obtener los hiperparámetros óptimos determinados mediante el método de GridSearch con validación cruzada de 10 particiones se muestran en la siguiente tabla (Tabla 6.2).

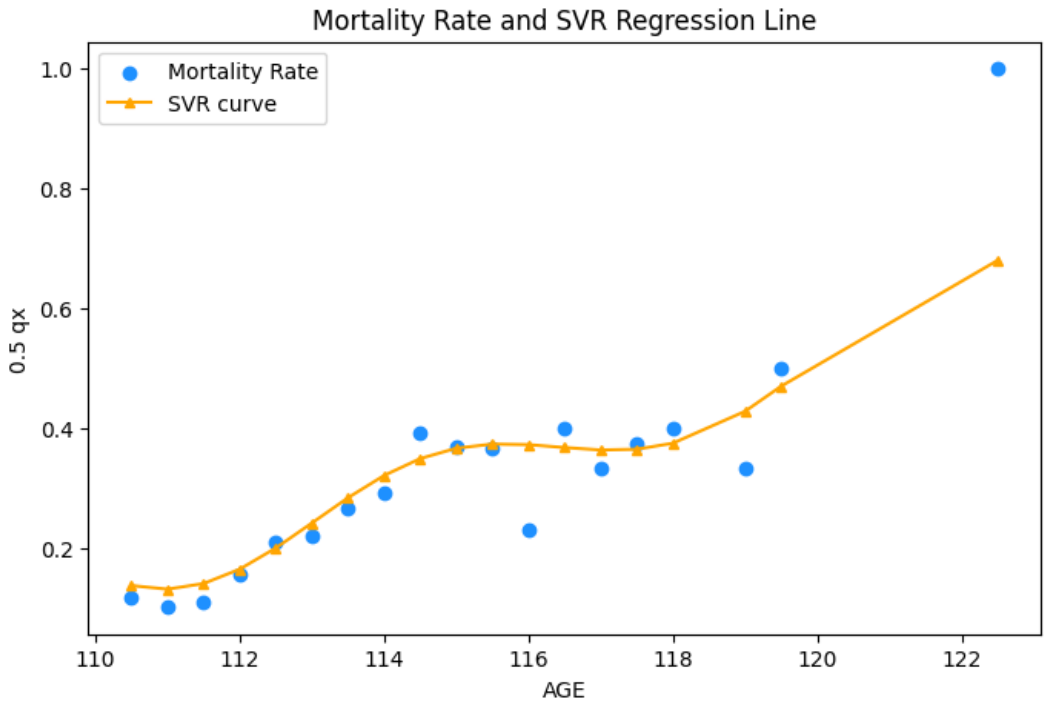
Tabla 6.2
Hiperparametros determinado de SVR

Best Hyperparameters			
C	0.27	Epsilon	0.03
Mean Squared Error			
0.029422933629131565			

Fuente: Elaboración Propia

Con estos hiperparámetros se puede encontrar la curva jada por el método SVR. La Figura 6.2 muestra los resultados del ajuste para el SVR y puede observarse que en el grá co aparecen dos intervalos relativamente planos, edad de 110-112 años y 115-118 años.

Figura 6.2
La ratio de mortalidad y la curva de SVR



Fuente: Elaboración Propia

6.3.3. Suavización de Whittaker-Henderson

El método de suavización de Whittaker-Henderson tiene un parámetro de suavizado , y el valor optimizado de dicho parámetro según el método de validación cruzada de 10 particiones se presenta en la Tabla 6.3.

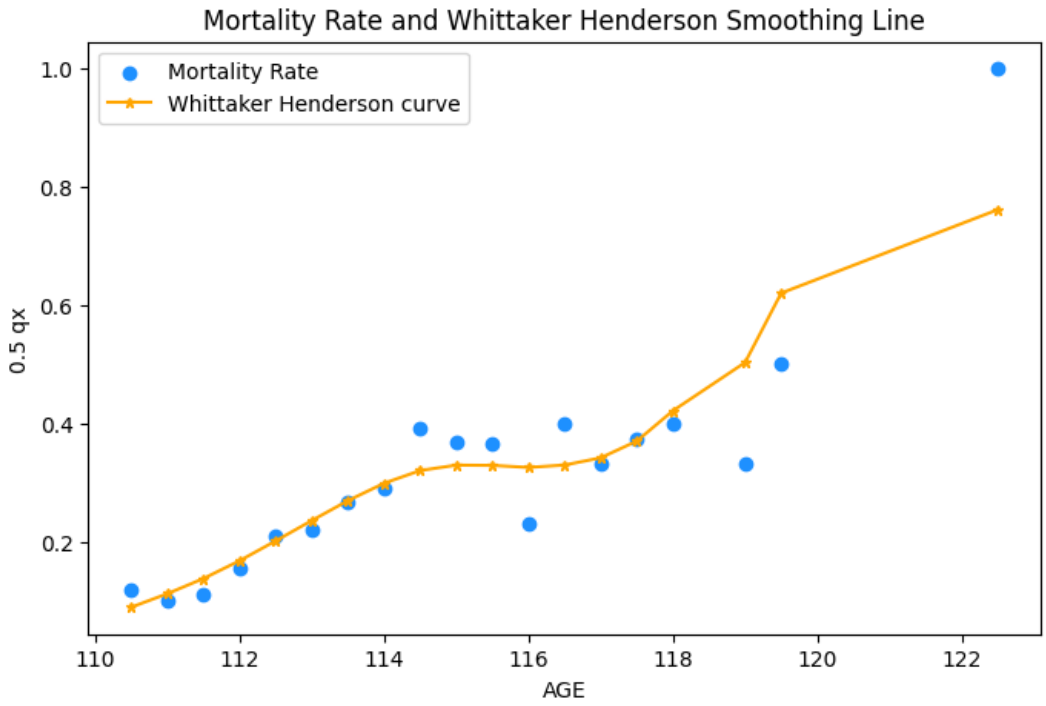
Tabla 6.3
Parámetro determinado de suavización de Whittaker-Henderson

Best Hyperparameters
10
Root mean squared error
0.0811505826615225

Fuente: Elaboración Propia

Con este parámetro estimado se puede jar la curva por el método de suavización de Whittaker-Henderson. La Figura 6.3 presenta los resultados del ajuste para el Whittaker-Henderson y puede observarse que en el grá co aparece solo un intervalo relativamente plano, edad de 115-118 años.

Figura 6.3
La ratio de mortalidad y la curva de la suavización de Whittaker-Henderson



Fuente: Elaboración Propia

6.3.4. Selección de métodos

Una vez determina las dos curvas suavizadas por diferentes métodos, es importante seleccionar una mejor para representar la tasa de mortalidad de los supersentenarios. Los resultados de AIC y BIC se presentan en la siguiente tabla (Tabla 6.4), en lo cual se puede encontrar que el método de suavización de Whittaker-Henderson es mejor que el método SVR.

Tabla 6.4
Comparación de AIC y BIC de SVR y Whittaker-Henderson

	AIC	BIC
SVR	-87.1804	-84.3470
WHITTAKER-HENDERSON	-93.4351	-92.4906

Fuente: Elaboración Propia

Así que, en el siguiente paso del estudio de la tasa de mortalidad, se utilizar dicho método para predecir. La siguiente tabla (Tabla 6.5) se presenta los valores suavizando de la tasa de mortalidad de los supercentenarios.

Tabla 6.5
Valor suavizado por el método de Whittaker-Henderson

AGE	SMOOTHED VALUE
110.5	0.089727
111.0	0.112559
111.5	0.138204
112.0	0.168341
112.5	0.201863
113.0	0.236274
113.5	0.269849
114.0	0.299204
114.5	0.320670
115.0	0.329799
115.5	0.329324
116.0	0.325925
116.5	0.329932
117.0	0.342162
117.5	0.370437
118.0	0.421698
119.0	0.503339
119.5	0.620588
122.5	0.761669

Fuente: Elaboración Propia

6.3.5. Predicción de la tasa de mortalidad de los supercentenarios

Después de obtener la curva de suavización, el próximo paso es predecir los futuros valores de la tasa de mortalidad. La siguiente tabla (Tabla 6.6) se presenta los parámetros óptimos de cada método y su error cuadrático medio (MSE) calculado. Así se puede llegar la conclusión que el mejor método en este análisis es extrapolación lineal con un rango de 18.

Tabla 6.6
Comparación métodos de predicción

	OPTIMAL PARAMETER			MSE
Simple Moving Average (SMA)	Range	:	4	0.0031
Exponential Smoothing (ES)	Smoothing Level	:	0.9999	0.0028
Linear Extrapolation (LE)	Range	:	18	0.0015

Fuente: Elaboración Propia