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.

# 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

# ÍNDICE DE FIGURAS

# ÍNDICE DE TABLAS

# 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 finanzas 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 identificar 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

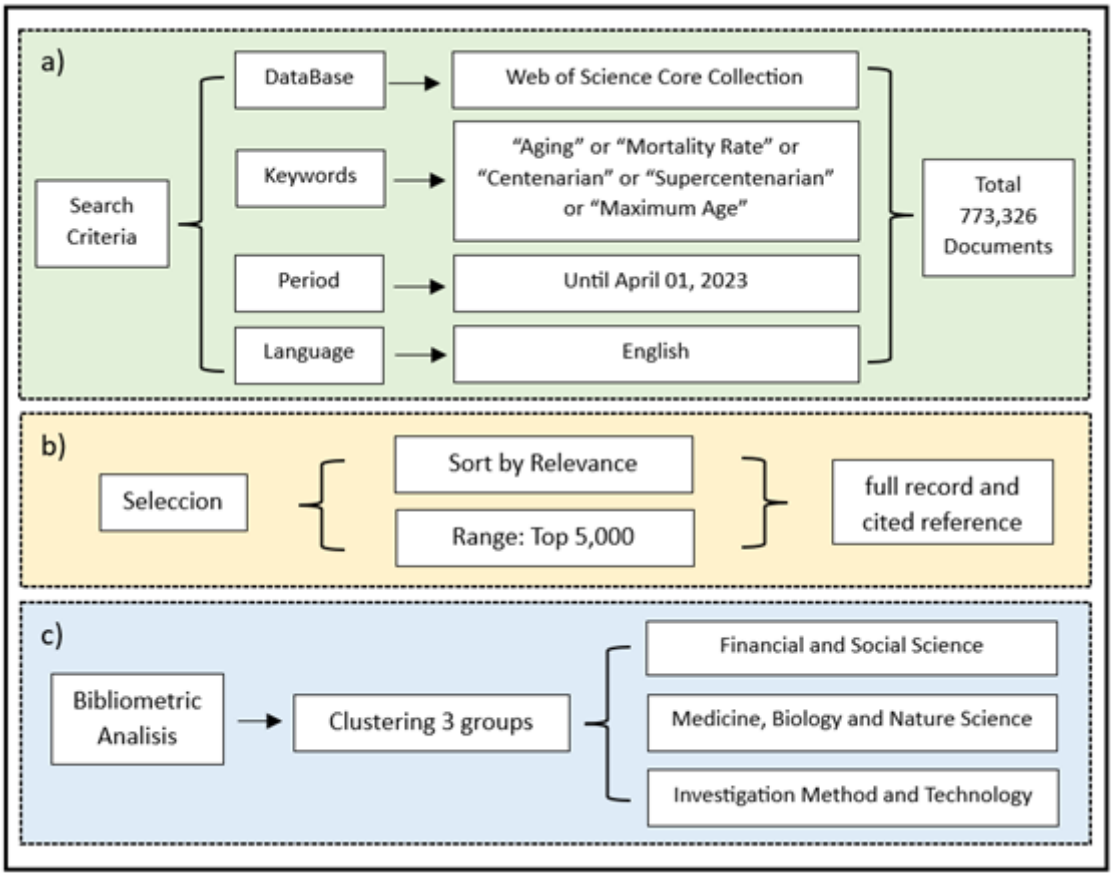
## 2.1. Investigaciones sobre la esperanza de vida y conceptos similares

El estudio de la esperanza de vida y conceptos similares ha sido siempre un tema de interés para la comunidad científica 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 figuras demográficas 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 influencia de estos temas a lo largo del tiempo. Como se ha presentado en la figura 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 refleja 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 figura 2.2.

La figura 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 refleja el interés científico 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.

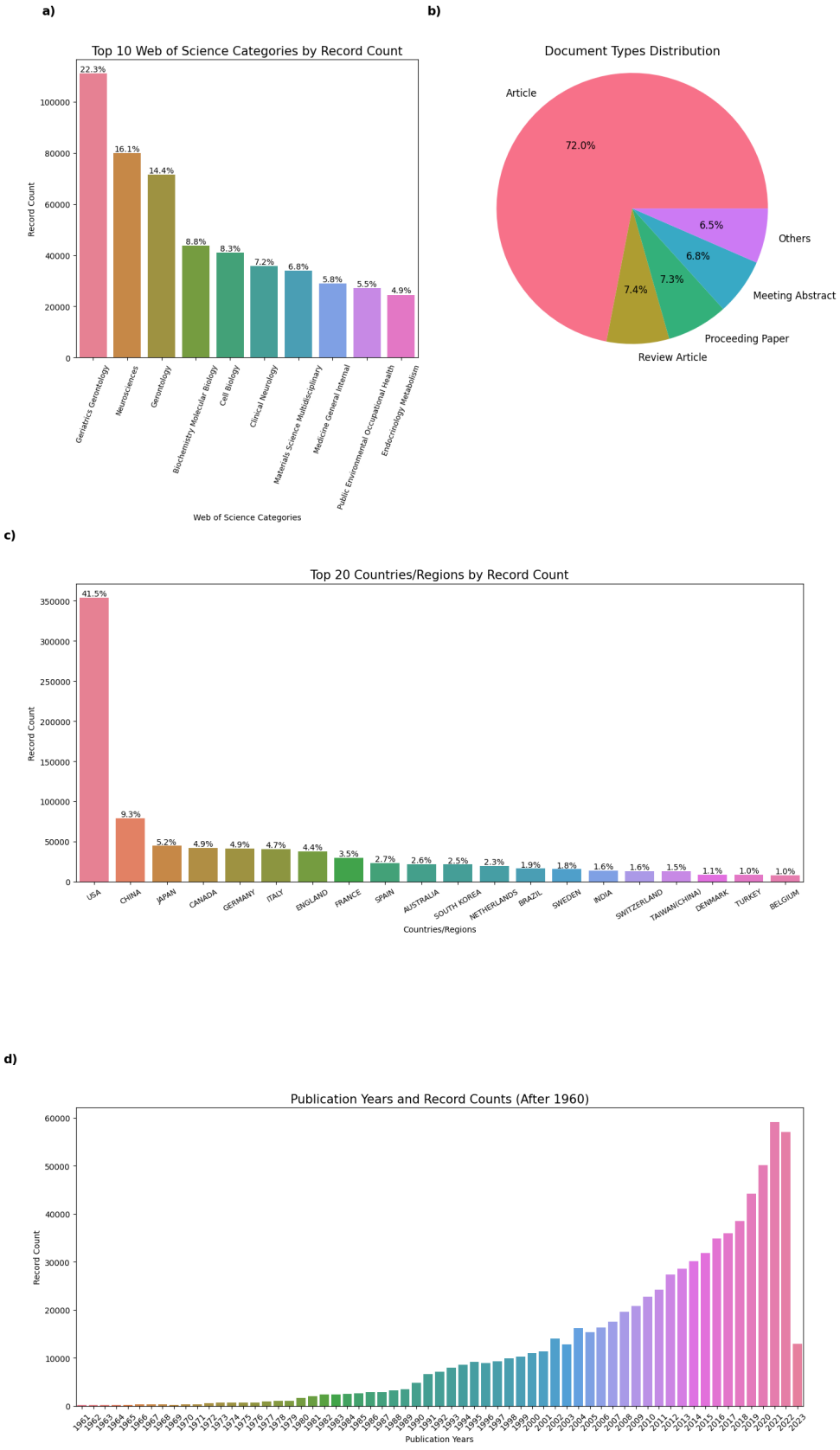
Luego, la figura 2.2b muestra la distribución de los tipos de documentos en la literatura sobre la longevidad y la mortalidad. Los artículos representan la mayoría (72 %), seguidos de las revisiones de artículos (7.4 %) y los trabajos de procedimientos (7.3 %). Esto sugiere que la mayoría de las investigaciones 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 figura 2.2c. Los Estados Unidos encabezan la lista (41.5 %), seguidos de China (9.3 %) y Japón (5.2 %). Esta distribución indica la concentración geográfica de la investigación sobre la longevidad y la mortalidad.

Finalmente, en la figura 2.2d, el grafico 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 refleja 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 clasificación de las palabras clave en tres conjuntos temáticos ilustrados en la figura 2.3.

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

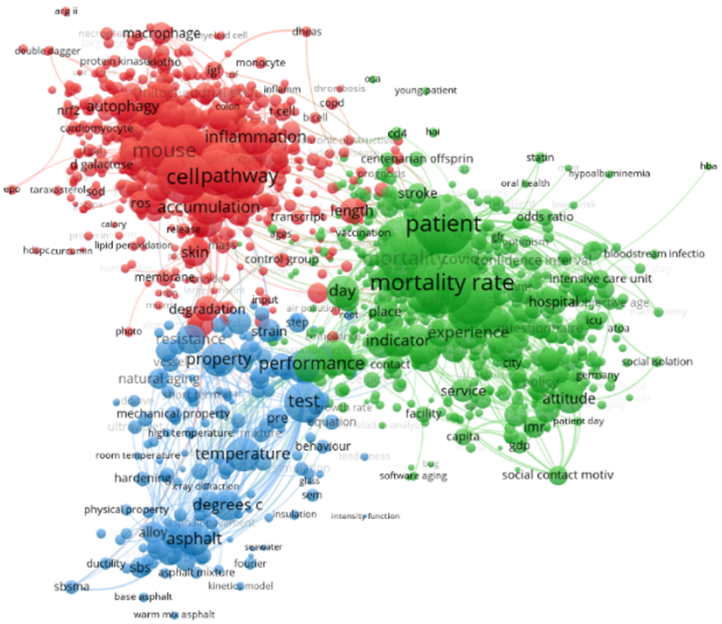
El siguiente conjunto, marcado en verde, se centra en elementos demográficos, socioeconómicos y de salud asociados al envejecimiento. Los términos “tasa de mortalidad”, “paciente”, “aislamiento social” y “envejecimiento” sugieren investigaciones que evalúan cómo factores sociales y de economía impactan 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 especificaron cuál sería este límite. Un año más tarde, Olshansky (2010) debatió la existencia de este límite superior fijo 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 afirmación se basa en análisis de datos demográficos 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 fijo. Barbi et al. (2018) desafiaron 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 fijo 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 fluctuar 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 específicas, 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.

Insuficiencia 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 verificación de la edad. Esta insuficiencia de datos puede llevar a estimaciones inexactas de la mortalidad y aumentar la incertidumbre en la predicción de la longevidad y las obligaciones financieras 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-fitting.

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:

$$N\left(x_i, y_j\right)\ Poisson\left(\lambda\left(y_j\right)\right) \tag{3.1}$$

donde:

- $X = x_1, x_2, \ldots, x_i$ sea el conjunto de edades posibles al morir en un determinado año $y_j$ , en este estudio, agrupadas por medio año, $X \in 105, 105{,}5, 106, 106{,}5, \ldots$.
- $Y = y_1, y_2, \ldots, y_j$ se presenta un conjunto de años que sea el rango de este estudio $Y \in [1987, 2018]$.
- $N\left(x_i, y_j\right)$ sea el número de individuos que mueren con edad $x_i$ $enelaoy_j$.
- Parámetro $\lambda\left(y_j\right)$ de la distribución Poisson indica el promedio del conjunto de edad $X\left(y_j\right)$.

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 significa que incluso si la tasa promedio cambia a través de los grupos de edad, variará de manera insignificante.
- La probabilidad de más de una muerte en un intervalo infinitesimalmente pequeño es insignificante: 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 influya más en las estimaciones generales así que se maneja la sobre dispersión y calcular resultados confiables.

La probabilidad de observar $z$ eventos (en nuestro caso, muertes en edad $x$ año $y$) dado un parámetro $\lambda$ en un modelo de Poisson es:

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

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

La ecuación de la verosimilitud ponderada para cada observación es:

$$L_k = \frac{w_k \, \lambda_k^{z_k} \, e^{-\lambda_k}}{z_k!} \tag{3.3}$$

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 $\frac{x}{\sum x}$ , y $z_k$ es el número observado de muertes en edad $x$ año $y$ para la observación $k$, $\lambda_k$ es el parámetro de tasa para la observación $k$ en el modelo de Poisson, y finalmente $z_k!$ es la factorial de $z_k$.

Así se define 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:

$$\log(L) = \sum_k (w_k * (z_k * \log(\lambda_k) - \lambda_k - \log(z_k!))) \tag{3.6}$$

## 3.3. Goodness-of-fit

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 $\lambda$. 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 suficiente: la prueba de chi-cuadrado asume que el tamaño de muestra es suficientemente grande.
- Distribución Teórica Especificada: La distribución teórica debe ser completamente especificada.

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 significativas 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:

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

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 fit tiene una historia muy larga (Massey, 1951), la prueba cuantifica 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 verificar que se puede aplicar la prueba de Kolmogorov-Smirnov, se utiliza la siguiente función para calcular:

$$D = \text{máx} \left| F_n(x) - F(x; \lambda) \right| \tag{3.8}$$

dónde:

- $F_n(x)$ es la función de distribución acumulada empírica
- $F(x;\lambda)$ 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:

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{3.9}$$

Sujeto a las restricciones denota como:

$$\begin{aligned} y_i - w^T\Phi(x_i) - b &\leq \epsilon + \xi \\ w^T\Phi(x_i) + b &\leq \epsilon + \xi_i^* \\ \xi_i, \xi_i^* &\geq 0, \forall i. \end{aligned} \tag{3.10}$$

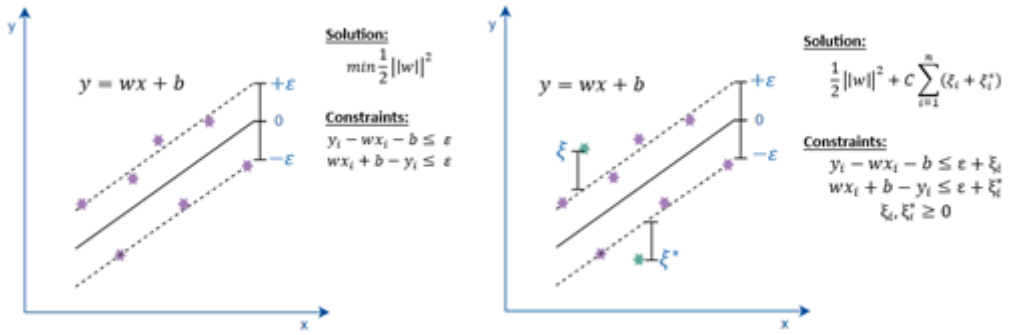
donde:

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

- $n$ es el número de ejemplos de entrenamiento
- $\phi(x_i)$ 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 predefinido $\varepsilon$ 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 superficie.

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 define como:

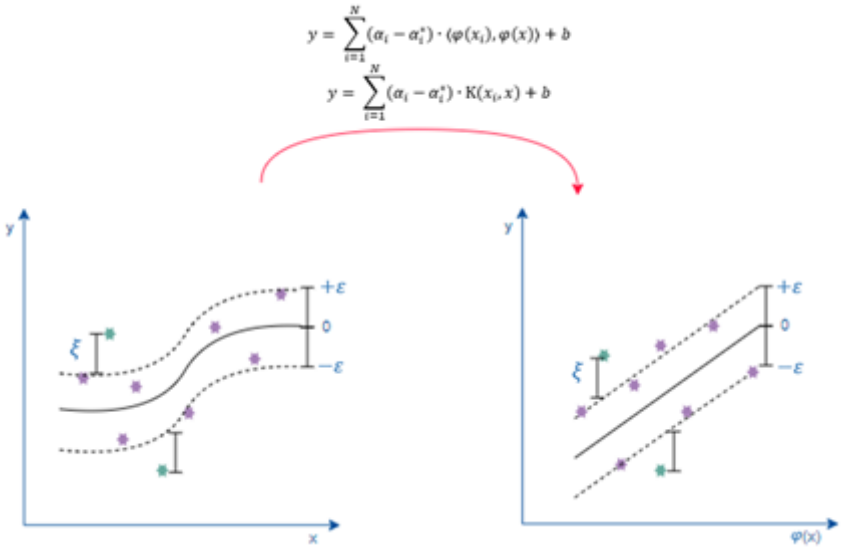
$$K(x, x') = exp\left(-||x - x'||^2\right) \tag{3.11}$$

donde:

- $x$ y $x'$ son dos vectores de características de entrada.
- $||x - x'||^2$ es la distancia euclidiana al cuadrado entre los vectores $x$ y $x'$.

- $\gamma$ es un hiperparámetro que controla la influencia de cada punto en el cálculo del núcleo. Un valor más alto de $\gamma$ 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 $k$ subconjuntos de aproximadamente el mismo tamaño.
- Para cada $i$ en $[1, 2, ..., k]$:
- Entrenar el modelo con los $k - 1$ subconjuntos que no incluyen $i$.
- Evaluar el modelo con el subconjunto $i$ y registrar el resultado.

- Calcular la media de los $k$ 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.

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

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 $z_i$ son los datos suavizados.
- $\lambda$ 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.
- Eficiencia computacional: A pesar de que el problema de optimización puede parecer complejo, en realidad se puede resolver de manera eficiente utilizando técnicas de álgebra lineal.

También es importante considerar las siguientes desventajas:

- Selección del parámetro de suavización: La elección del parámetro $\lambda$ puede ser complicada. Un valor de $\lambda$ demasiado alto puede resultar en una serie temporal excesivamente suave que ignora aspectos importantes de los datos, mientras que un valor de $\lambda$ demasiado bajo puede no suavizar adecuadamente el ruido.
- Bordes de la serie temporal: El método Whittaker-Henderson puede presentar dificultades 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:

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

donde:

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

El BIC se calcula como:

$$BIC = k \ln(n) - 2 \ln(L) \tag{3.14}$$

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 confiabilidad.

## 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 financieros. Este método calcula el promedio de un conjunto de datos durante un período de tiempo especificado, que se mueve a lo largo del tiempo. El SMA se utiliza para suavizar las fluctuaciones 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 financieros, 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:

$$ES(t) = \alpha * X(t) + (1 - \alpha) * ES(t-1) \tag{3.15}$$

donde:

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

## 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 Demográfica, el Instituto Francés de Estudios Demográficos (INED) y el Instituto Francés de Salud e Investigación Médica (INSERM). Su principal objetivo es proporcionar datos verificables 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 confirmar 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 confidencialidad de los individuos, evitando la divulgación de información personal identificable.

Por otro lado, el GRG se centra específicamente 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 significativamente 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 verificación 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 eficientemente 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 demográfico 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.

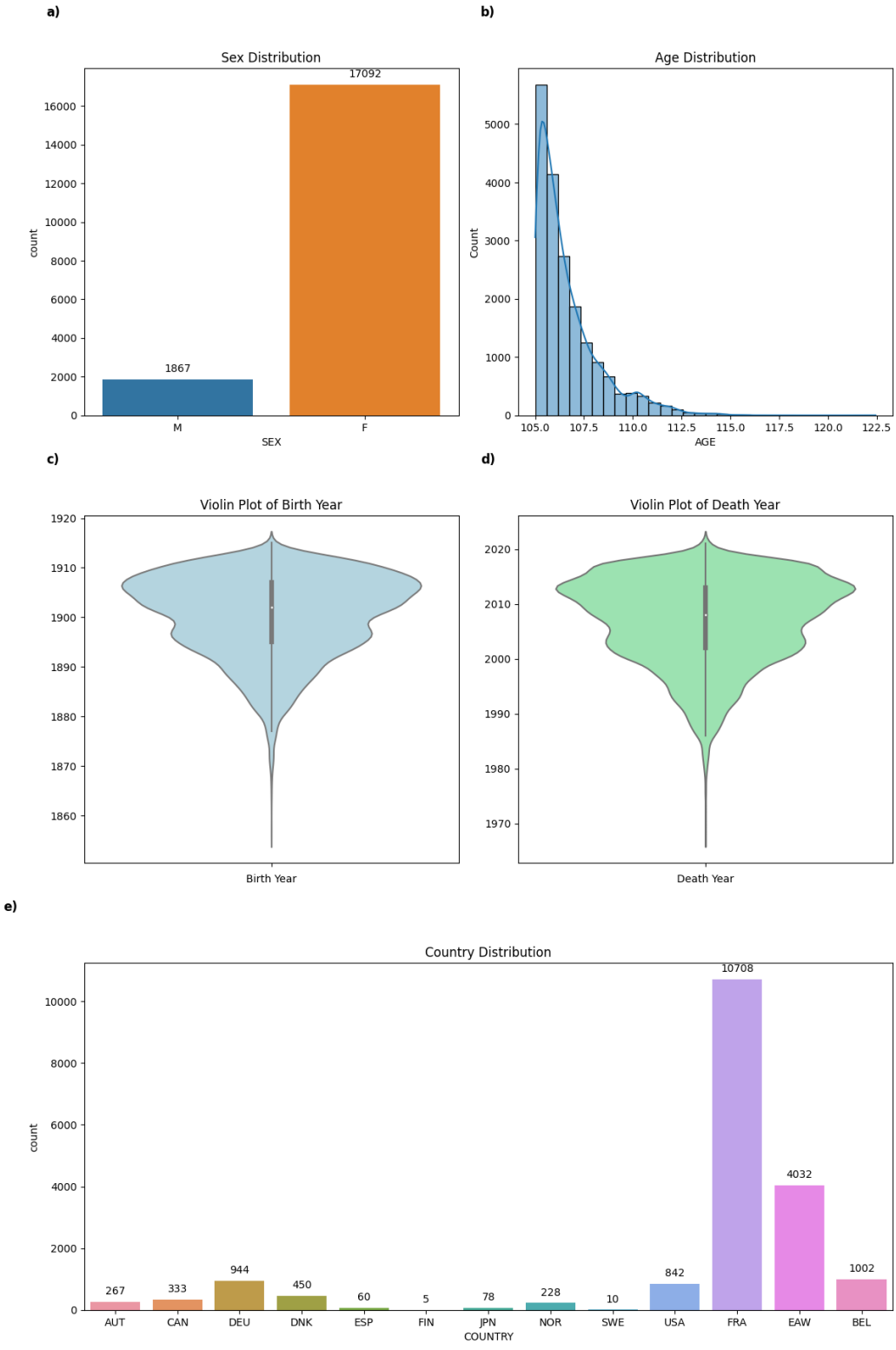
De los 18,959 incidentes registrados, 17,092 corresponden a mujeres y 1,867 a hombres (Figura 4.1a). Además, se observa que el rango de edades varía entre 105 y 122.5 años, y se observa una tendencia decreciente en el número de incidentes a medida que aumenta la edad (Figura 4.1b).

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.

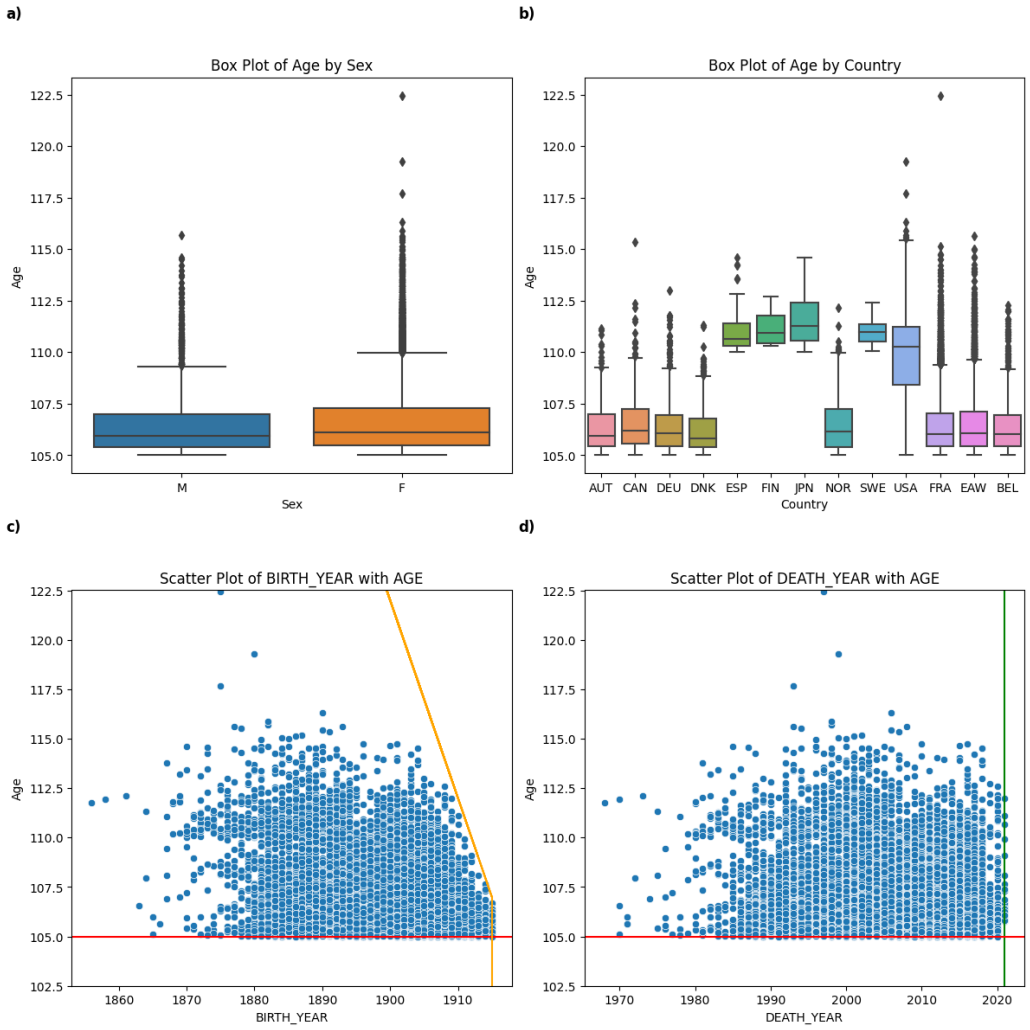
Mediante un diagrama de caja (Figura 4.2a), se puede constatar que la media de la edad de fallecimiento es similar en hombres y mujeres. Sin embargo, es notable que la distribución de la edad de fallecimiento en mujeres presenta una cola más prolongada. Esto implica que existe una mayor proporción de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

**Figura 4.2**

*Análisis de la variable "edad" frente a otros variables de la base de datos de IDL*




Fuente: Elaboración Propia

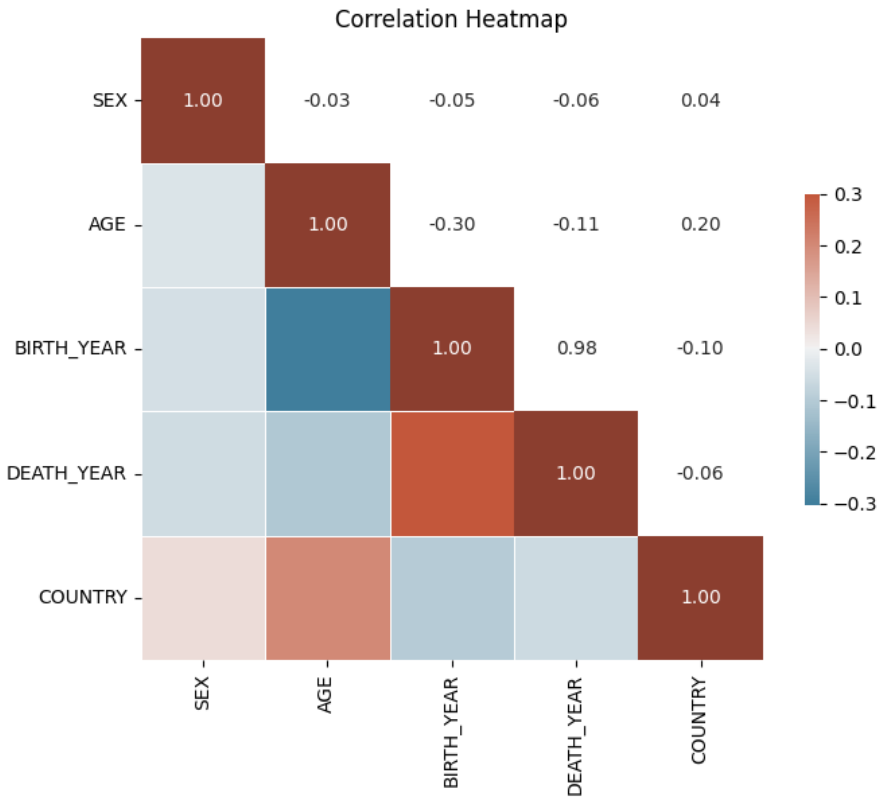
La siguiente figura (Figura 4.3) expone la correlación existente entre las variables anteriormente mencionadas. Se destacan los siguientes aspectos:

1. Existe una correlación positiva y significativa 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.

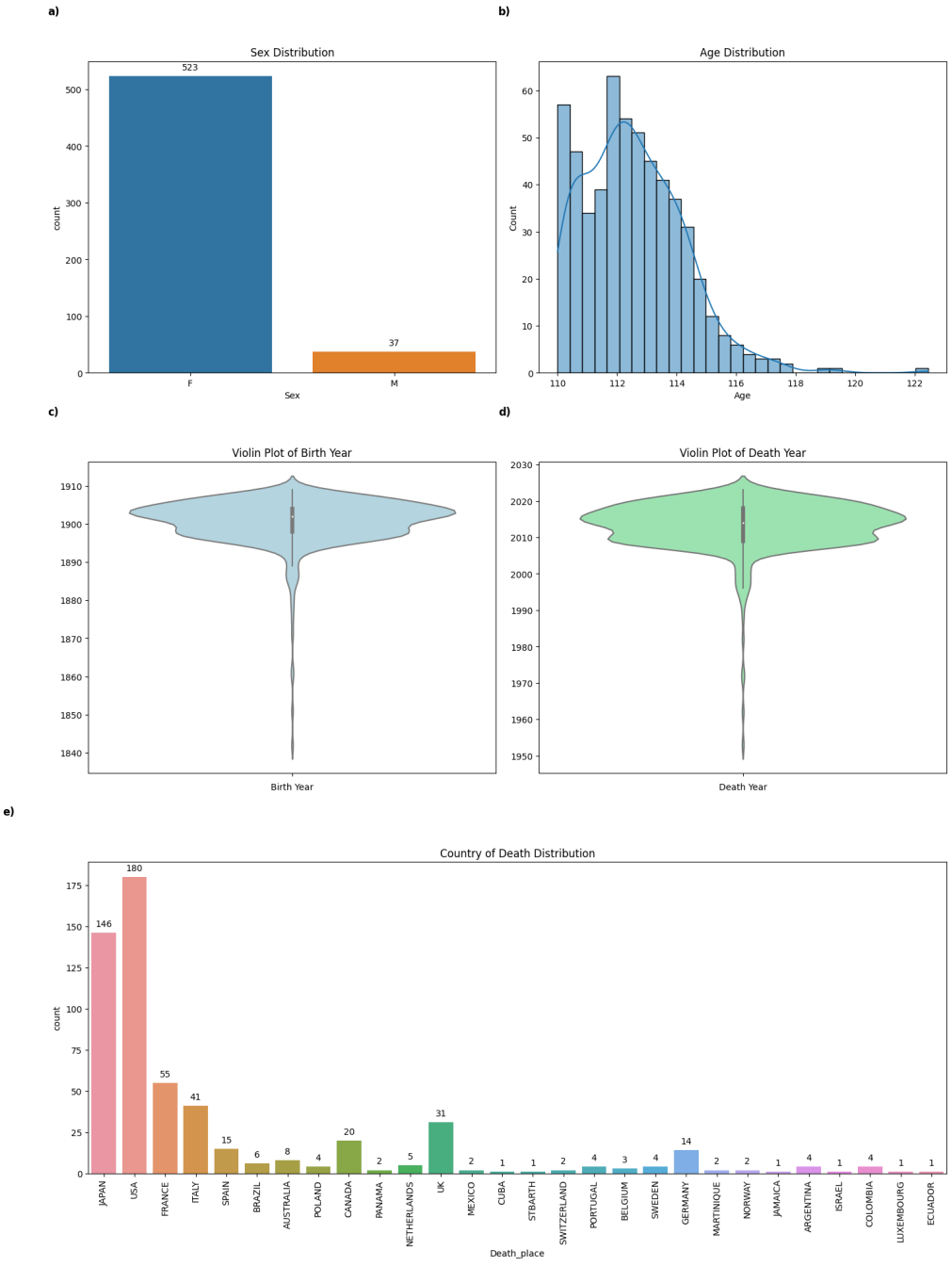
De los 560 incidentes registrados, un total de 523 corresponden a sujetos de sexo femenino y 37 a sujetos de sexo masculino (Figura 4.4a). En cuanto a la distribución por edades, esta oscila entre 110 y 122.5 años. Es notable una disminución en el número de incidentes a medida que la edad supera los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

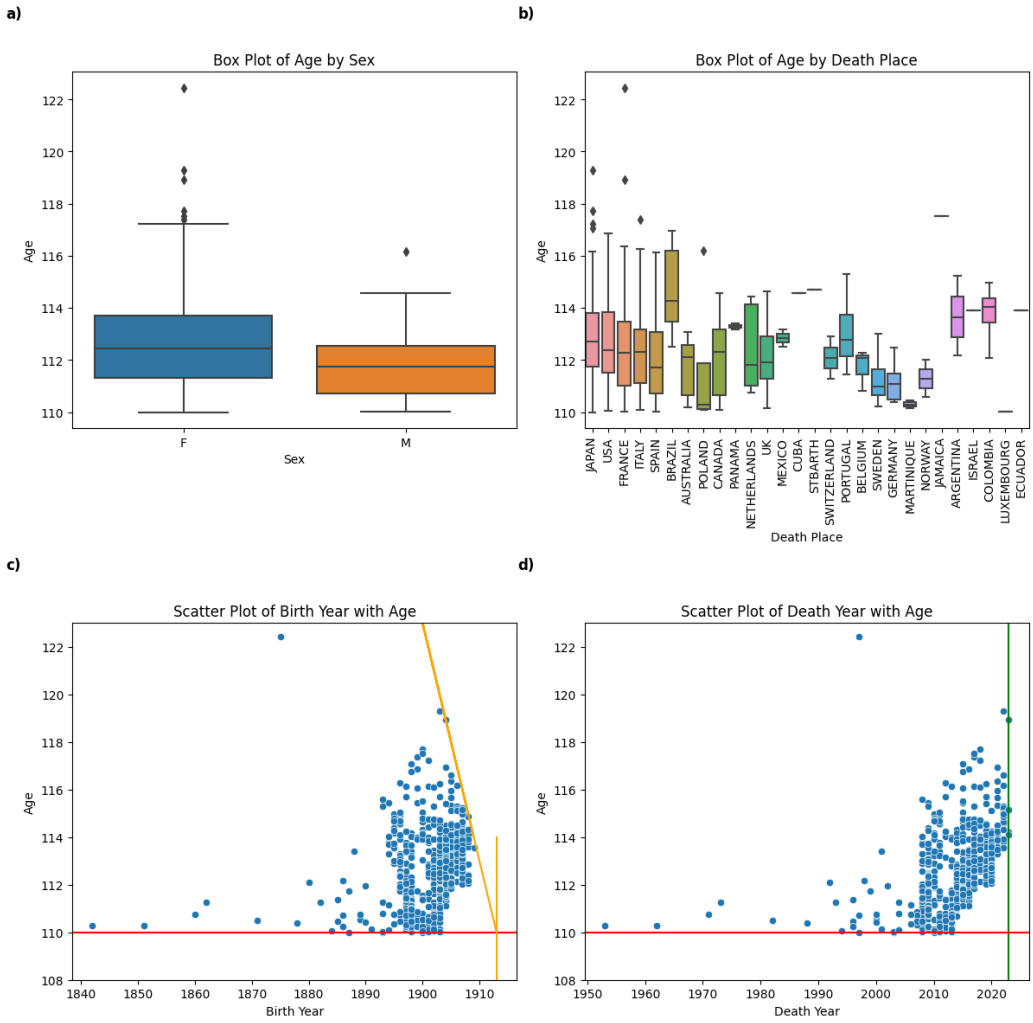
A través de un diagrama de caja (Figura 4.5a), se evidencia que la media de la edad de fallecimiento de las mujeres supera a la de los hombres. Además, la distribución de la edad de fallecimiento en mujeres exhibe una cola más extensa. Este patrón sugiere una proporción mayor de mujeres que alcanzan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

La Figura 4.5b ofrece una visión general de la edad de fallecimiento de los supercentenarios en distintos países y regiones. Aunque los valores promedio fluctúan 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 figuras.

**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 identifica una correlación positiva significativa entre el año de nacimiento ("Birth_year") y el año de fallecimiento ("Death_year"). Este significa 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.

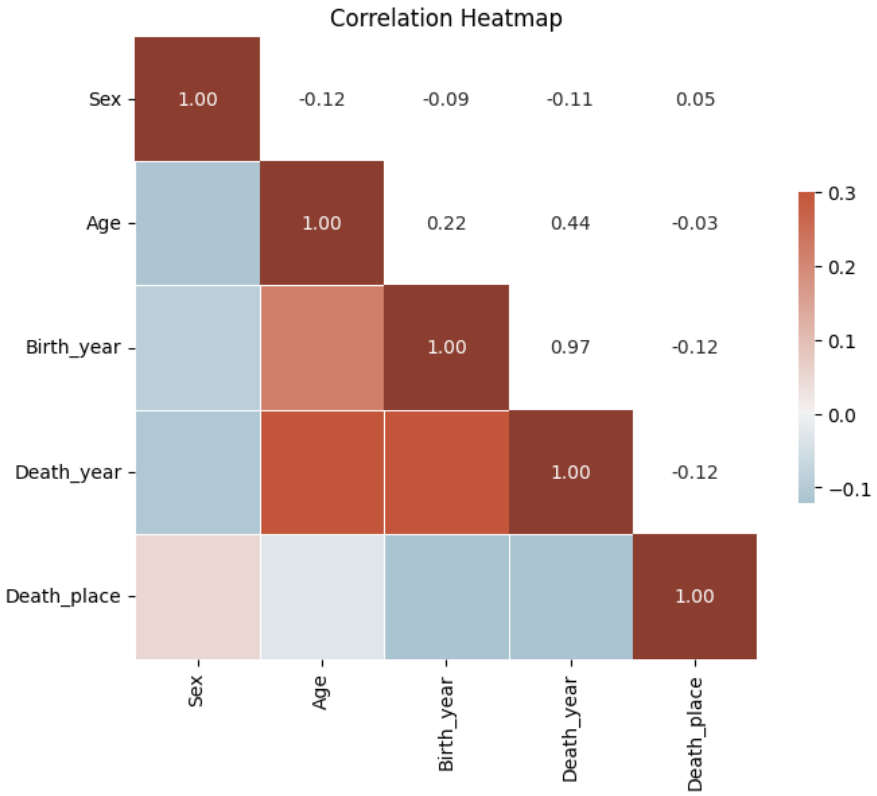
En segunda instancia, se aprecia una correlación positiva entre la edad ("Age") y el año de nacimiento ("Birth_year"). Esta correlación indica que a medida que el año de nacimiento se aproxima al presente, la edad registrada suele ser más alta. Esto podría ser interpretado como una señal de que la edad 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 especifico. 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 significativa 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 confiables 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-fit) para verificar 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 predefinidos para un algoritmo de aprendizaje automático. El objetivo de este enfoque es identificar la combinación de hiperparámetros que proporciona el mejor rendimiento del modelo, medido según una métrica de evaluación predefinida.

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 filtra 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 filtro 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 |

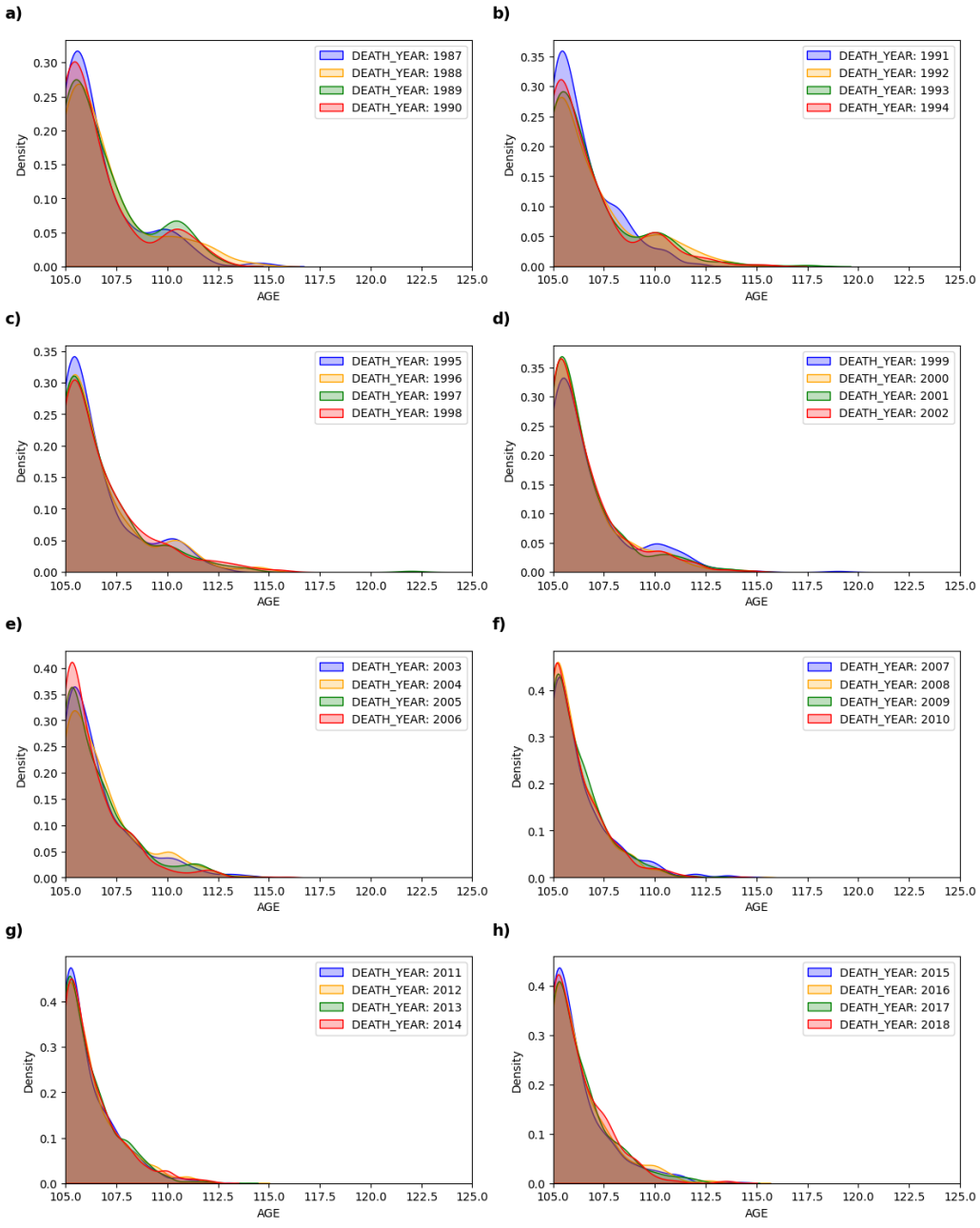
Fuente: Elaboración Propia

### 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 subgráfico 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*

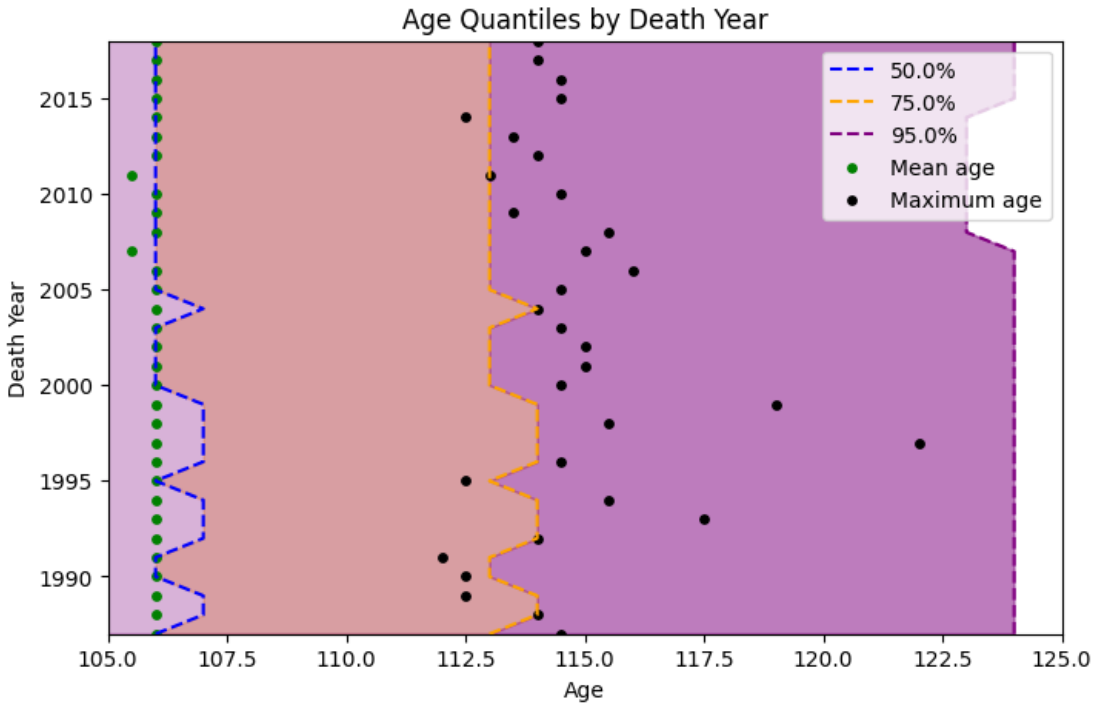
| | DEATH_YEAR | Estimated $\lambda$ | KS P-value | CHI2 P-value |
|---|---|---|---|---|
| 1 | 1987 | 106.610693 | 8.74E-21 | 8.37E-15 |
| 2 | 1988 | 106.871206 | 2.62E-19 | 7.40E-14 |
| 3 | 1989 | 106.85374 | 6.63E-27 | 3.45E-19 |
| 4 | 1990 | 106.625955 | 8.19E-27 | 3.46E-33 |
| 5 | 1991 | 106.416051 | 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 | 0.00E+00 |
| 25 | 2011 | 106.19585 | 7.43E-192 | 1.93E-298 |
| 26 | 2012 | 106.268402 | 1.74E-227 | 0.00E+00 |
| 27 | 2013 | 106.227568 | 1.49E-234 | 0.00E+00 |
| 28 | 2014 | 106.266782 | 1.98E-229 | 0.00E+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 figura (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*

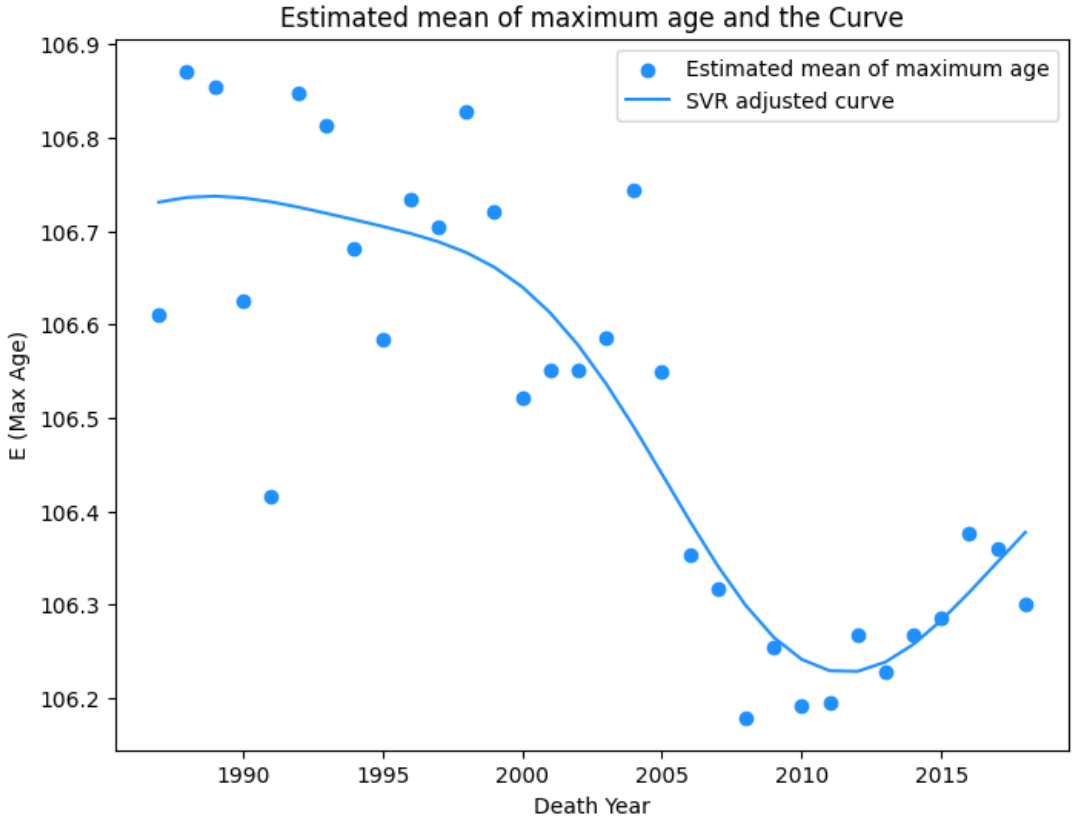
| **Best Hyperparameters** | | | |
|---|---|---|---|
| C | 0.76 | Epsilon | 0.12 |
| **Mean Squared Error** | | | |
| 0.013166848 | | | |

Fuente: Elaboración Propia

Con estos hiperparámetros se puede encontrar la curva fijada por el método SVR (Figura 5.3). En esta figura, 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 deficiencia 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*



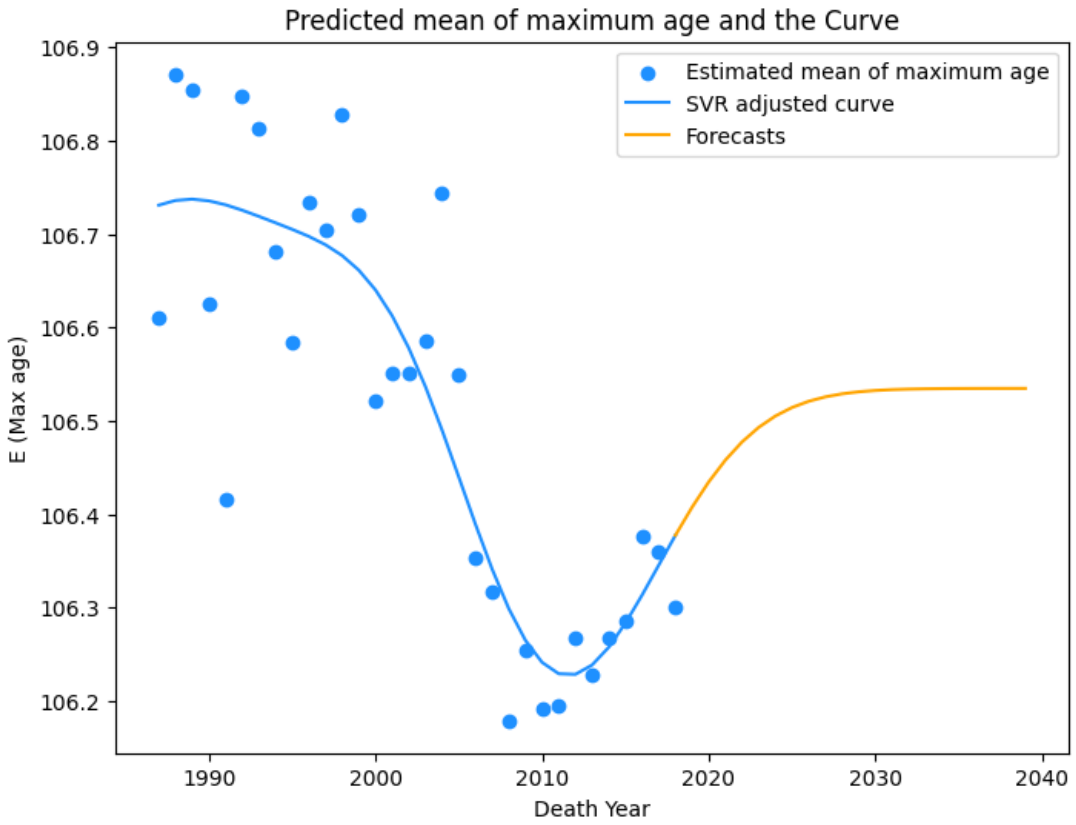

Fuente: Elaboración Propia

### 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 figura un punto de inflexión, donde la tasa de crecimiento del valor predicho se ralentiza con el tiempo y finalmente tiende a ser constante.

**Figura 5.4**

*La predicción media de la edad máxima de seres humanos por SVR*

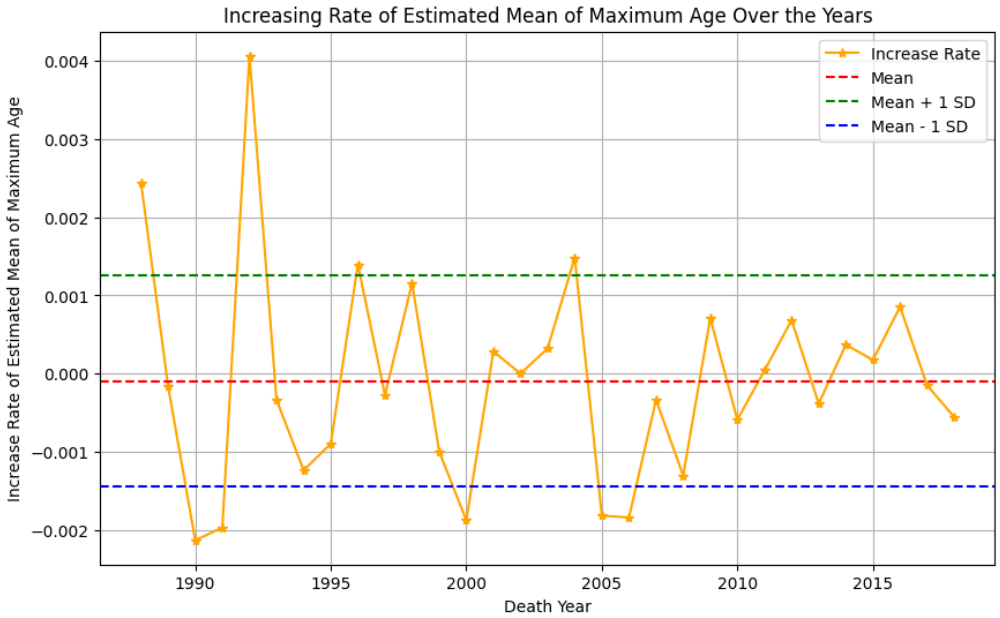



Fuente: Elaboración Propia

Para investigar la potencial explicación de ese punto de inflexión, se genera el Figura 5.5 sobre la ratio de crecimiento de la media estimada de la edad máxima. Se puede observar en esta figura que la varianza de dicha ratio disminuye con el tiempo, y la línea naranja en la figura converge entre las líneas punteada verde y azul con el tiempo. Además, la línea punteada roja es ligeramente negativa, lo que significa 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*

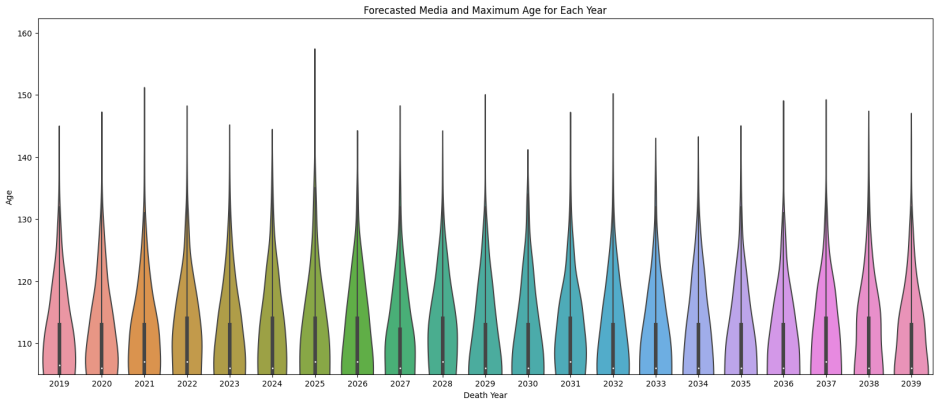
| YEAR | Forecasted Media | Quantile .95 |
|---|---|---|
| 2019 | 106.3774757 | 123 |
| 2020 | 106.4075658 | 125 |
| 2021 | 106.4346435 | 124 |
| 2022 | 106.4580256 | 123 |
| 2023 | 106.4774702 | 125 |
| 2024 | 106.4930829 | 124 |
| 2025 | 106.50521 | 124 |
| 2026 | 106.514336 | 123 |
| 2027 | 106.5209974 | 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 figura 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 dificulta sacar conclusiones lo más fiables 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 defina ${}_nq_x$, donde $n$ 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 específicas 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 qué ${}_nq_x$ 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**

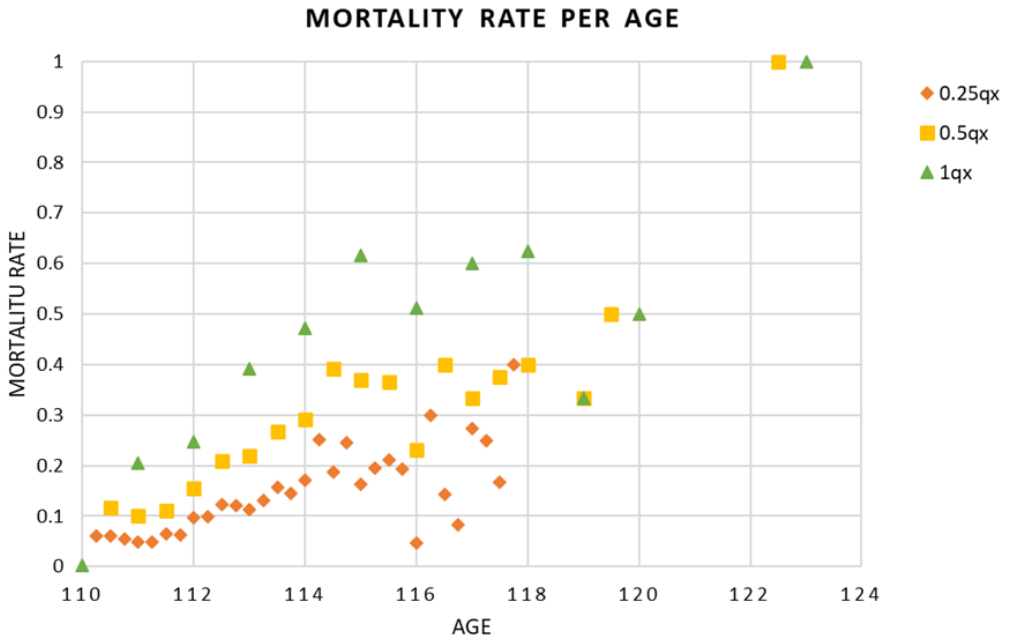
*Tabla de ${}_nq_x$ calculado*

| 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*

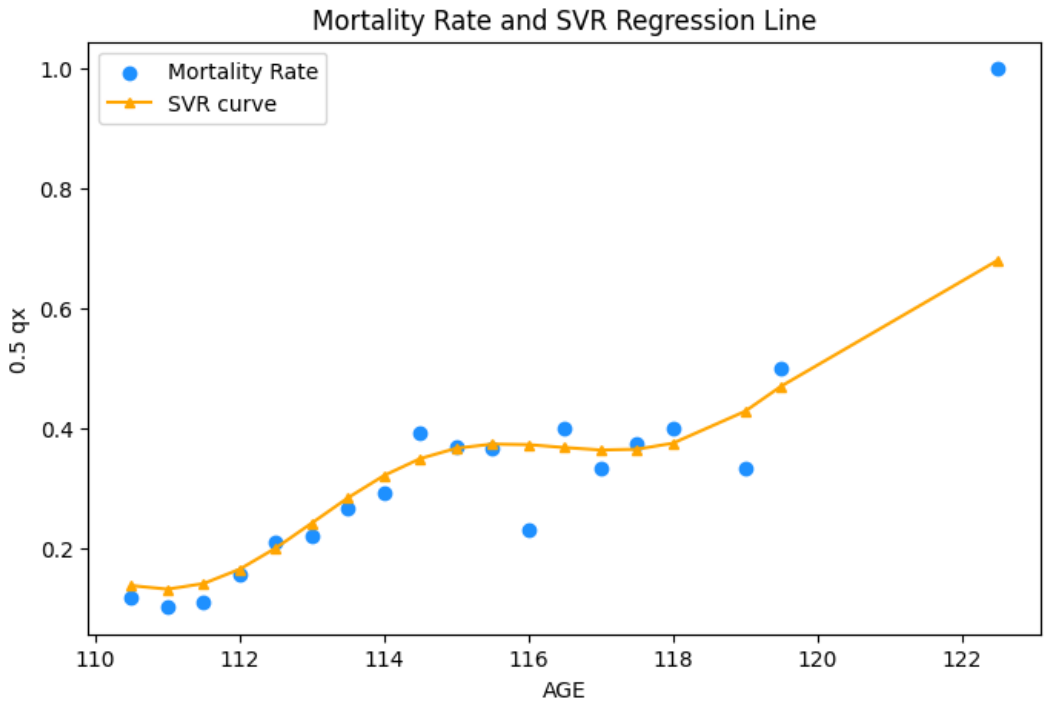
| **Best Hyperparameters** | | | |
|---|---|---|---|
| C | 0.27 | Epsilon | 0.03 |
| **Mean Squared Error** | | | |
| 0.029422933629131565 | | | |

Fuente: Elaboración Propia

Con estos hiperparámetros se puede encontrar la curva fijada por el método SVR. La Figura 6.2 muestra los resultados del ajuste para el SVR y puede observarse que en el gráfico 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 $\lambda$, 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*

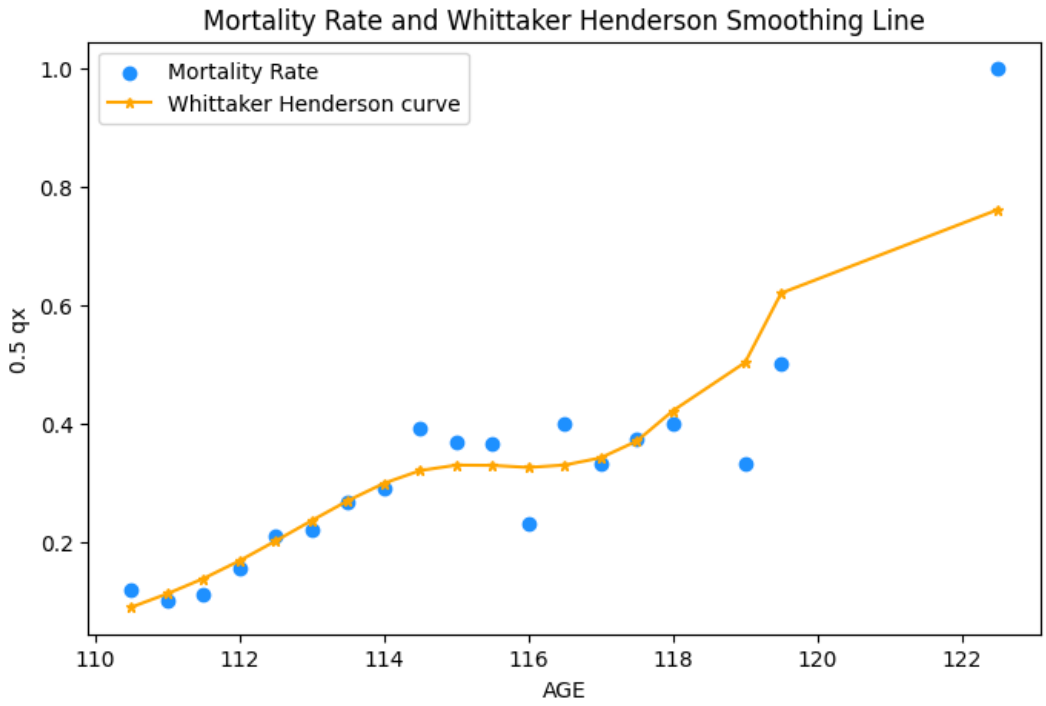
| **Best Hyperparameters** | |
|---|---|
| $\lambda$ | 10 |
| **Root mean squared error** | |
| 0.0811505826615225 | |

Fuente: Elaboración Propia

Con este parámetro estimado se puede fijar 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 gráfico 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

Una vez se decide el mejor método de predicción, se puede predecir la tasa de mortalidad de las próximas edades (Tabla 6.7). Aquí es necesita mencionar que el cálculo termina hasta que el valor acercándose a 1 (en este estudio, se fija en 0.95), dado que la tasa de mortalidad es una probabilidad así tiene que ser entre 0 y 1. Además, Se puede visualizar los valores predichos en la Figura 6.4.

**Tabla 6.7**

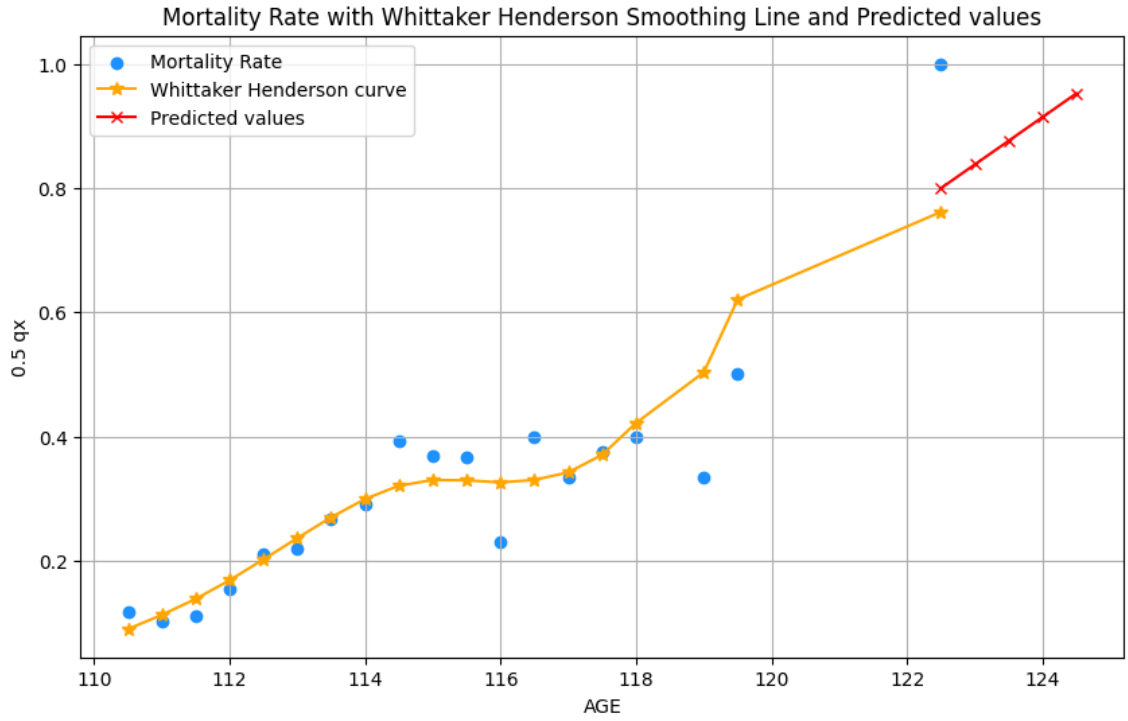
*Predicción de la tasa de mortalidad*

| AGE | PREDICTED ${}_{0,5}q_x$ |
|---|---|
| 122.5 | 0.799852 |
| 123.0 | 0.838035 |
| 123.5 | 0.876218 |
| 124.0 | 0.914401 |
| 124.5 | 0.952584 |

Fuente: Elaboración Propia

**Figura 6.4**

*Predicción de la tasa de mortalidad*




Fuente: Elaboración Propia

Así se puede llegar a la conclusión que la edad máxima potencial de seres humanos seria 124.5 años aproximadamente.

# 7. CONCLUSIONES Y FUTUROS ESTUDIOS

## 7.1. Conclusión

Nuestros resultados se obtuvieron aplicando los dos métodos a dos bases de datos y los resultados muestran una concordancia significativa, lo que pone de relieve la solidez de nuestras conclusiones. La coherencia entre estos resultados pone de relieve la fiabilidad de nuestras conclusiones y reafirma la validez de nuestro método y enfoque.

En primer lugar, se ha demostrado que la edad máxima alcanzable para los seres humanos se sitúa entre 124 y 124.5 años. Esto establece un límite superior aproximado de la esperanza de vida humana. Esta constante, que persiste con independencia de la base de datos utilizada, revela las limitaciones biológicas inherentes a la longevidad humana y proporciona un claro punto de referencia para futuras investigaciones en el campo de la gerontología.

En segundo lugar, nuestro estudio revela pautas interesantes en la mortalidad entre los 115 y los 117 años. A esta edad, la mortalidad tiende a estabilizarse y se mantiene en un nivel constante. Sin embargo, a partir de esa edad, la mortalidad empieza a aumentar hasta alcanzar una probabilidad de 1. El aumento de la mortalidad a partir de los 117 años sugiere que, aunque es posible vivir hasta una edad muy avanzada, las probabilidades de hacerlo descienden drásticamente a partir de ese momento.

Estos resultados representan un paso importante en la comprensión de los límites de la longevidad humana y de los patrones de mortalidad por edades extremas. Al comprender mejor estos fenómenos, nuestra investigación puede ayudar a orientar futuras investigaciones relacionadas con el envejecimiento y la longevidad y a mejorar los modelos de riesgo.

## 7.2. Futuros Estudios

En primer lugar, en términos de ampliación y profundización del estudio, se podrían explorar mejoras de los modelos de Poisson existentes o compararlos con otros modelos. Este enfoque podría ofrecer una imagen más clara de las tendencias de la mortalidad en las edades extremas. Diversificando los modelos analíticos utilizados, podemos obtener nuevos resultados y confirmar la validez de nuestras conclusiones actuales. Por tanto, es necesario seguir desarrollando y aplicando distintos métodos estadísticos y matemáticos para estudiar la mortalidad.

En segundo lugar, cabe destacar la oportunidad de combinar el modelo Gompertz de mortalidad de menores de 80 años con la observación de los supercentenarios. Tal vinculación podría proporcionar una visión más completa de la mortalidad humana y, por

tanto, una comprensión cabal de la transición de la vejez "típica.ª la edad extrema. Este tipo de investigación interdisciplinar podría arrojar luz sobre el misterio de la longevidad y sobre cómo varían nuestras tasas de mortalidad a medida que nos acercamos a los límites de la longevidad humana.

Además, si ampliamos el alcance del estudio a personas de 80 o 90 años, podremos obtener una imagen más completa de la longevidad. Este enfoque permitiría un análisis más detallado de la etapa intermedia entre la vejez típica y la supervivencia extrema, enriqueciendo nuestra comprensión de cómo evoluciona la mortalidad en esta etapa de la vida. De este modo, sería posible identificar y analizar los factores clave que determinan la transición de la vejez al supercentenario.

Por último, una línea de investigación prometedora consiste en analizar los factores y características individuales que pueden influir en la mortalidad y la esperanza de vida. Este enfoque implica realizar análisis de subdivisión exhaustivos para identificar subgrupos específicos basados en características individuales (por ejemplo, genéticas, demográficas, socioeconómicas, etc.). Por ejemplo, determinadas variantes genéticas pueden afectar a la resistencia a las enfermedades o al ritmo de envejecimiento. Las condiciones socioeconómicas, como la calidad de la atención sanitaria, el nivel de educación y el acceso a una dieta sana, también pueden desempeñar un papel crucial en la longevidad. Analizando cómo afectan estas y otras características a la mortalidad, podemos conocer con más detalle y matices los factores que influyen en la longevidad humana. Este enfoque de segmentación tiene importantes implicaciones en el campo de la ciencia actuarial, sobre todo en lo que respecta a la segmentación empresarial. Los resultados de estos estudios pueden utilizarse para mejorar los modelos de riesgo y crear estimaciones más precisas y justas de las primas de seguros.

# BIBLIOGRAFÍA


Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, *52*(3). https://doi.org/10.1007/BF02294359

Anderson, R. N., y Rosenberg, H. M. (1998). Age standardization of death rates: implementation of the year 2000 standard.

Barbi, E., Lagona, F., Marsili, M., Vaupel, J. W., y Wachter, K. W. (2018). The plateau of human mortality: Demography of longevity pioneers. *Science*, *360*(6396). https://doi.org/10.1126/science.aat3119

Bernoulli, J., y Sylla, E. D. (2006). The art of conjecturing, together with Letter to a friend on sets in court tennis, 430.

Bowers, N., Gerber, H., Hickman, J., Jones, D., y Nesbitt, C. (1987). Actuarial mathematics. . Actuarial Mathematics by Bowers, Hickman, Gerber, Jones and Nesbitt [Published in 1986 by The Society of Actuaries]. *Transactions of the Faculty of Actuaries*, *41*. https://doi.org/10.1017/s0071368600009812

Brillinger, D. R. (1986). A Biometrics Invited Paper with Discussion: The Natural Variability of Vital Rates and Associated Statistics. *Biometrics*, *42*(4). https://doi.org/10.2307/2530689

Cairns, A. J., Blake, D., y Dowd, K. (2008). Modelling and management of mortality risk: A review. *Scandinavian Actuarial Journal*, (2-3). https://doi.org/10.1080/03461230802173608

Christensen, K., Doblhammer, G., Rau, R., y Vaupel, J. W. (2009). Ageing populations: the challenges ahead. https://doi.org/10.1016/S0140-6736(09)61460-4

Cochran, W. D. (1952). The $\chi 2$ Test of Goodness of Fit on JSTOR. *Annals of Mathematical Statistics*, *23*(3), 315-345. https://www.jstor.org/stable/2236678?origin=JSTOR-pdf

Del Castillo, J. M. R., y López-Farré, A. (2017). Longevidad y envejecimiento en el tercer milenio. Consultado el 24 de junio de 2023, desde https://www.fundacionmapfre.org/publicaciones/todas/informe-longevidad-envejecimiento-tercer-milenio/

Del Castillo, J., y Pérez-Casany, M. (1998). Weighted Poisson distributions for overdispersion and underdispersion situations. *Annals of the Institute of Statistical Mathematics*, *50*(3). https://doi.org/10.1023/A:1003585714207

Dong, X., Milholland, B., y Vijg, J. (2016). Evidence for a limit to human lifespan. *Nature*, *538*(7624). https://doi.org/10.1038/nature19793

Drucker, H., Surges, C. J., Kaufman, L., Smola, A., y Vapnik, V. (1997). Support vector regression machines. *Advances in Neural Information Processing Systems*, *1*, 155-161. https://papers.nips.cc/paper/1238-support-vector-regression-machines.pdf

Eckhardt, R. (1987). Stan Ulam, John von Neumann, and the Monte Carlo Method. *Los Alamos Science*, *Special Is*.

Fischer, H. (2011). A History of the Central Limit Theorem. *A History of the Central Limit Theorem*. https://doi.org/10.1007/978-0-387-87857-7

French Institute for Demographic Studies. (2023). IDL. Consultado el 1 de abril de 2023, desde https://www.supercentenarians.org/en/

Fries, J. F. (2002). Aging, natural death, and the compression of morbidity. https://doi.org/10.1056/nejm198007173030304

Gampe, J. (2010). Human mortality beyond age 110. En *Demographic Research Monographs*. https://doi.org/10.1007/978-3-642-11520-2_13

Gavrilov, L. A., y Gavrilova, N. S. (2011). Mortality Measurement at Advanced Ages. *North American Actuarial Journal*, *15*(3). https://doi.org/10.1080/10920277.2011.10597629

GERONTOLOGY RESEARCH GROUP. (2023). Supercentenarian Data – Table E. Consultado el 1 de abril de 2023, desde https://grg.org/WSRL/TableE.aspx

Gimeno-Miguel, A., Clerencia-Sierra, M., Ioakeim, I., Poblador-Plou, B., Aza-Pascual-Salcedo, M., González-Rubio, F., Rodríguez Herrero, R., y Prados-Torres, A. (2019). Health of Spanish centenarians: A cross-sectional study based on electronic health records. *BMC Geriatrics*, *19*(1). https://doi.org/10.1186/s12877-019-1235-7

Hammond, M. (2000). The Forces of Mortality at Ages 80 to 120. *International Journal of Epidemiology*, *29*(2). https://doi.org/10.1093/ije/29.2.384-a

Ho, J. Y., y Hendi, A. S. (2018). Recent trends in life expectancy across high income countries: Retrospective observational study. *BMJ (Online)*, *362*. https://doi.org/10.1136/bmj.k2562

Hughes, B. G., y Hekimi, S. (2017). Many possible maximum lifespan trajectories. https://doi.org/10.1038/nature22786

Janssen, F., y Kunst, A. (2007). The choice among past trends as a basis for the prediction of future trends in old-age mortality. *Population Studies*, *61*(3). https://doi.org/10.1080/00324720701571632

Joseph, A. W. (1952). The Whittaker-Henderson Method of Graduation. *Journal of the Institute of Actuaries*, *78*(1). https://doi.org/10.1017/s0020268100052495

Kakar, G. (2007). A Course in Credibility Theory and its Applications. By H. Bühlmann A. Gisler (Springer, 2005). *Annals of Actuarial Science*, *2*(2). https://doi.org/10.1017/s1748499500000415

Lee, R. (2000). The lee-carter method for forecasting mortality, with various extensions and applications. *North American Actuarial Journal*, *4*(1). https://doi.org/10.1080/10920277.2000.10595882

Maier, H., Gampe, J., Jeune, B., Robine, J.-M., Vaupel, J., y Brown, R. L. (2012). Supercentenarians. *Canadian Studies in Population [ARCHIVES]*, *39*(1-2), 135-140. https://doi.org/10.25336/P6XS56

Massey, F. J. (1951). The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the American Statistical Association*, *46*(253). https://doi.org/10.1080/01621459.1951.10500769

Neath, A. A., y Cavanaugh, J. E. (2012). The Bayesian information criterion: Background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics*, *4*(2). https://doi.org/10.1002/wics.199

Olshansky, S. J. (2010). The Law of Mortality Revisited: Interspecies Comparisons of Mortality. *Journal of Comparative Pathology*, *142*(SUPPL. 1). https://doi.org/10.1016/j.jcpa.2009.10.016

Reinhardt, H. E. (1987). Statistical Decision Theory and Bayesian Analysis. Second Edition (James O. Berger). *SIAM Review*, *29*(3). https://doi.org/10.1137/1029095

Weon, B. M., y Je, J. H. (2009). Theoretical estimation of maximum human lifespan. *Biogerontology*, *10*(1). https://doi.org/10.1007/s10522-008-9156-4

Yaari, M. E. (1965). Uncertain lifetime, life insurance, and the theory of the consumer. *Review of Economic Studies*, *32*(2). https://doi.org/10.2307/2296058

# ANEXO SCRIPTS DE PYTHON

## A. Revisión de la Literatura

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from matplotlib.gridspec import GridSpec
import warnings
warnings.filterwarnings("ignore")

# Create an irregular grid of subplots
fig = plt.figure(figsize=(15, 25))
gs = GridSpec(3, 2, figure=fig)
ax1 = fig.add_subplot(gs[0, 0])
ax3 = fig.add_subplot(gs[0, 1])
ax2 = fig.add_subplot(gs[1, :])
ax4 = fig.add_subplot(gs[2, :])

# Set spacing between subplots
fig.subplots_adjust(wspace=0.4, hspace=0.4)

# Read data and plot subplot 1
data1 = pd.read_csv("WoS Categories.txt", sep='\t', header=0)
data1.columns = ['Web of Science Categories', 'Record Count', '
    Percentage of 773,326']
data1 = data1.sort_values('Record Count', ascending=False).head(10)
     # Here we change to top 10
sns.barplot(y='Record Count', x='Web of Science Categories', data=
    data1, ax=ax1, palette='husl')
total = data1['Record Count'].sum()
data1['Percentage'] = data1['Record Count'] / total * 100
for index, row in data1.iterrows():
    ax1.text(row.name, row['Record Count'], f"{row['Percentage']:.1f
        }%", ha='center', va='bottom')
plt.setp(ax1.get_xticklabels(), rotation=70, fontsize=9)
ax1.set_title("Top 10 Web of Science Categories by Record Count",
    fontsize=15)
ax1.set_ylabel("Record Count")
ax1.set_xlabel("Web of Science Categories")

# Read data and plot subplot 2
data2 = pd.read_csv("Countries_Regions.txt", sep='\t', header=0)
data2 = data2.sort_values(by='Record Count', ascending=False)
top_20 = data2.head(20)  # Here we change to top 20
total_records = top_20['Record Count'].sum()
top_20['Percentage'] = (top_20['Record Count'] / total_records) *
```

```
    100
sns.barplot(x='Countries/Regions', y='Record Count', data=top_20, ax
    =ax2, palette='husl')
plt.setp(ax2.get_xticklabels(), rotation=30, fontsize=9)
ax2.set_title("Top 20 Countries/Regions by Record Count", fontsize
    =15)
ax2.set_xlabel("Countries/Regions")
ax2.set_ylabel("Record Count")
for index, row in top_20.iterrows():
    ax2.text(index, row['Record Count']+2000, f"{row['Percentage
        ']:.1f}%", ha='center', fontsize=10)

# Read data and plot subplot 3
data3 = pd.read_csv("Document Types.txt", sep='\t', header=0)
data3.columns = ['Document Types', 'Record Count', 'Percentage of
    773,326']
others = data3[data3['Percentage of 773,326'] < 5].sum(numeric_only=
    True)
others['Document Types'] = 'Others'
data3 = data3[data3['Percentage of 773,326'] >= 5].append(others,
    ignore_index=True)
# Generate a 'husl' color palette
colors = sns.color_palette('husl', len(data3))
pie_plot = ax3.pie(data3['Record Count'], labels=data3['Document
    Types'], autopct='%1.1f%%', colors=colors, textprops={'fontsize
    ': 12})
ax3.set_title("Document Types Distribution", fontsize=15)
ax3.axis('equal')


# Read data and plot subplot 4
data4 = pd.read_csv("Publication Years.txt", sep='\t', header=0)
data4.columns = ['Publication Years', 'Record Count', 'Percentage of
    773,326']
data4_filtered = data4[data4['Publication Years'] > 1960]
sns.barplot(x='Publication Years', y='Record Count', data=
    data4_filtered, ax=ax4, palette='husl')
ax4.set_title("Publication Years and Record Counts (After 1960)",
    fontsize=15)
ax4.set_xlabel("Publication Years")
ax4.set_ylabel("Record Count")
plt.setp(ax4.get_xticklabels(), rotation=45)

# Add subfigure labels
ax1.text(-0.1, 1.15, "a)", transform=ax1.transAxes, size=15, weight
    ='bold')
ax3.text(-0.1, 1.15, "b)", transform=ax3.transAxes, size=15, weight
    ='bold')
ax2.text(-0.1, 1.15, "c)", transform=ax2.transAxes, size=15, weight
    ='bold')
ax4.text(-0.1, 1.15, "d)", transform=ax4.transAxes, size=15, weight
```

```
    ='bold')

# Tight layout to prevent overlap
plt.tight_layout()

# Show plot
plt.show()
```

## B. Descripción de Datos de IDL

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import matplotlib.gridspec as gridspec
import warnings
warnings.filterwarnings("ignore")

data_idl = pd.read_excel('data_idl.xlsx',usecols=[1,2,3,4,5])
data_idl.head(30)

# Summary statistics
print(data_idl.describe())

# Define the grid
gs = gridspec.GridSpec(3, 2)

fig = plt.figure(figsize=(12, 18))

# Subplot for sex
ax0 = plt.subplot(gs[0, 0])
sex_plot = sns.countplot(x="SEX", data=data_idl, ax=ax0)
ax0.set_title("Sex Distribution")
for p in sex_plot.patches:
    sex_plot.annotate(format(p.get_height(), '.0f'),
                      (p.get_x() + p.get_width() / 2., p.get_height()),
                      ha='center', va='center',
                      xytext=(0, 10),
                      textcoords='offset points')

# Subplot for age
ax1 = plt.subplot(gs[0, 1])
sns.histplot(data=data_idl, x="AGE", bins=30, kde=True, ax=ax1)
ax1.set_title("Age Distribution")

# Subplot for birth year
ax2 = plt.subplot(gs[1, 0])
sns.violinplot(data=data_idl, y="BIRTH_YEAR", ax=ax2, inner='box',
```

```
    palette=[(0.678, 0.847, 0.902)])
ax2.set_title("Violin Plot of Birth Year")
ax2.set_ylabel('')
ax2.set_xlabel("Birth Year")

# Subplot for death year
ax3 = plt.subplot(gs[1, 1])
sns.violinplot(data=data_idl, y="DEATH_YEAR", ax=ax3, inner='box',
    palette=[(0.565, 0.933, 0.678)])
ax3.set_title("Violin Plot of Death Year")
ax3.set_ylabel('')
ax3.set_xlabel("Death Year")

# Large subplot for country
ax4 = plt.subplot(gs[2, :])
country_plot = sns.countplot(x="COUNTRY", data=data_idl, ax=ax4)
ax4.set_title("Country Distribution")
for p in country_plot.patches:
    country_plot.annotate(format(p.get_height(), '.0f'),
                   (p.get_x() + p.get_width() / 2., p.get_height()),
                   ha='center', va='center',
                   xytext=(0, 10),
                   textcoords='offset points')

# Add subfigure labels
ax0.text(-0.1, 1.15, "a)", transform=ax0.transAxes, size=12, weight
    ='bold')
ax1.text(-0.1, 1.15, "b)", transform=ax1.transAxes, size=12, weight
    ='bold')
ax2.text(-0.1, 1.15, "c)", transform=ax2.transAxes, size=12, weight
    ='bold')
ax3.text(-0.1, 1.15, "d)", transform=ax3.transAxes, size=12, weight
    ='bold')
ax4.text(-0.1, 1.15, "e)", transform=ax4.transAxes, size=12, weight
    ='bold')

# Tight layout to prevent overlap
plt.tight_layout()

# Show the plot
plt.show()

# Create a figure with subplots
fig, axes = plt.subplots(2, 2, figsize=(12, 12))

# Box plot for AGE by SEX
sns.boxplot(data=data_idl, x="SEX", y="AGE", ax=axes[0, 0])
axes[0, 0].set_title("Box Plot of Age by Sex")
axes[0, 0].set_xlabel("Sex")
axes[0, 0].set_ylabel("Age")

```

```
# Box plot for AGE by COUNTRY
sns.boxplot(data=data_idl, x="COUNTRY", y="AGE", ax=axes[0, 1])
axes[0, 1].set_title("Box Plot of Age by Country")
axes[0, 1].set_xlabel("Country")
axes[0, 1].set_ylabel("Age")

# Scatter plot for BIRTH_YEAR with AGE
sns.scatterplot(data=data_idl, x="BIRTH_YEAR", y="AGE", ax=axes[1, 0])
axes[1, 0].axhline(y=105, color='red')
axes[1, 0].plot(data_idl['BIRTH_YEAR'], 2022 - data_idl['BIRTH_YEAR'], color='orange')
ymax_relative = (107 - 102.5) / (122.5 - 102.5)
axes[1, 0].axvline(x=1915, ymax=ymax_relative, color='orange')
axes[1, 0].set_ylim(102.5, 122.5)
axes[1, 0].set_title("Scatter Plot of BIRTH_YEAR with AGE")
axes[1, 0].set_xlabel("BIRTH_YEAR")
axes[1, 0].set_ylabel("Age")

# Scatter plot for DEATH_YEAR with AGE
sns.scatterplot(data=data_idl, x="DEATH_YEAR", y="AGE", ax=axes[1, 1])
axes[1, 1].axvline(x=2021, color='green')
axes[1, 1].axhline(y=105, color='red')
axes[1, 1].set_ylim(102.5, 122.5)
axes[1, 1].set_title("Scatter Plot of DEATH_YEAR with AGE")
axes[1, 1].set_xlabel("DEATH_YEAR")
axes[1, 1].set_ylabel("Age")

# Add subfigure labels
axes[0, 0].text(-0.15, 1.15, "a)", transform=axes[0, 0].transAxes, size=12, weight='bold')
axes[0, 1].text(-0.15, 1.15, "b)", transform=axes[0, 1].transAxes, size=12, weight='bold')
axes[1, 0].text(-0.15, 1.15, "c)", transform=axes[1, 0].transAxes, size=12, weight='bold')
axes[1, 1].text(-0.15, 1.15, "d)", transform=axes[1, 1].transAxes, size=12, weight='bold')

# Adjust spacing between subplots
plt.tight_layout()

# Show the plot
plt.show()

# Convert categorical columns to category codes
for col in ['SEX', 'COUNTRY']:
    data_idl[col] = data_idl[col].astype('category').cat.codes

# Compute the correlation matrix
corr = data_idl[['SEX', 'AGE', 'BIRTH_YEAR', 'DEATH_YEAR', 'COUNTRY
```

```
    ']].corr()

# Create masks for the upper, lower triangles, and diagonal
mask_lower = np.tril(np.ones_like(corr, dtype=np.bool_))
mask_upper = np.triu(np.ones_like(corr, dtype=np.bool_))
mask_diag = np.eye(*corr.shape).astype(np.bool_)

# Generate a custom diverging colormap
cmap = sns.diverging_palette(230, 20, as_cmap=True)

# Create a figure and axes
fig, ax = plt.subplots(figsize=(8, 6))

# Draw the heatmap for the lower triangle with the colormap and without numbers
sns.heatmap(corr, mask=mask_upper, cmap=cmap, vmax=.3, center=0,
            square=True, linewidths=.5, cbar_kws={"shrink": .5}, annot=False, ax=ax)

# Draw the heatmap for the upper triangle without the colormap and with numbers
sns.heatmap(corr, mask=mask_lower, cmap=['white'], cbar=False, annot=True, fmt=".2f", ax=ax)

# Draw the diagonal with both color and value
sns.heatmap(corr, mask=~mask_diag, cmap=['#8B3E2F'], cbar=False, annot=True, fmt=".2f", ax=ax)

# Set the title of the correlation heatmap
ax.set_title("Correlation Heatmap")

# Rotate axis labels by 90 degrees
ax.set_xticklabels(ax.get_xticklabels(), rotation=90)
ax.set_yticklabels(ax.get_yticklabels(), rotation=0)
# Show the plot
plt.show()
```

## C. Descripción de Datos de GRG

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import matplotlib.gridspec as gridspec
import warnings
warnings.filterwarnings("ignore")

data_grg = pd.read_excel('data_grg.xlsx')
```

```
data_grg.head(30)

# Summary statistics
print(data_grg.describe())

data_grg = pd.read_excel('data_grg.xlsx')

# Define the grid
gs = gridspec.GridSpec(3, 2)

fig = plt.figure(figsize=(15, 20))

# Subplot for sex
ax0 = plt.subplot(gs[0, 0])
sex_plot = sns.countplot(x="Sex", data=data_grg, ax=ax0)
ax0.set_title("Sex Distribution")
for p in sex_plot.patches:
    sex_plot.annotate(format(p.get_height(), '.0f'),
                      (p.get_x() + p.get_width() / 2., p.get_height()),
                      ha='center', va='center',
                      xytext=(0, 10),
                      textcoords='offset points')

# Subplot for age
ax1 = plt.subplot(gs[0, 1])
sns.histplot(data=data_grg, x="Age", bins=30, kde=True, ax=ax1)
ax1.set_title("Age Distribution")

# Subplot for birth year
ax2 = plt.subplot(gs[1, 0])
sns.violinplot(data=data_grg, y="Birth_year", ax=ax2, inner='box',
   palette=[(0.678, 0.847, 0.902)])
ax2.set_title("Violin Plot of Birth Year")
ax2.set_ylabel('')
ax2.set_xlabel("Birth Year")

# Subplot for death year
ax3 = plt.subplot(gs[1, 1])
sns.violinplot(data=data_grg, y="Death_year", ax=ax3, inner='box',
   palette=[(0.565, 0.933, 0.678)])
ax3.set_title("Violin Plot of Death Year")
ax3.set_ylabel('')
ax3.set_xlabel("Death Year")

# Large subplot for death country
ax4 = plt.subplot(gs[2, :])
country_plot = sns.countplot(x="Death_place", data=data_grg, ax=ax4)
ax4.set_title("Country of Death Distribution")
for p in country_plot.patches:
    country_plot.annotate(format(p.get_height(), '.0f'),
                      (p.get_x() + p.get_width() / 2., p.get_height()),
```

```
                    ha='center', va='center',
                    xytext=(0, 10),
                    textcoords='offset points')
plt.xticks(rotation=90)

# Add subfigure labels
ax0.text(-0.1, 1.15, "a)", transform=ax0.transAxes, size=12, weight='bold')
ax1.text(-0.1, 1.15, "b)", transform=ax1.transAxes, size=12, weight='bold')
ax2.text(-0.1, 1.15, "c)", transform=ax2.transAxes, size=12, weight='bold')
ax3.text(-0.1, 1.15, "d)", transform=ax3.transAxes, size=12, weight='bold')
ax4.text(-0.1, 1.15, "e)", transform=ax4.transAxes, size=12, weight='bold')

# Tight layout to prevent overlap
plt.tight_layout()

# Show the plot
plt.show()

# Create a figure with subplots
fig, axes = plt.subplots(2, 2, figsize=(12, 12))

# Box plot for Age by Sex
sns.boxplot(data=data_grg, x="Sex", y="Age", ax=axes[0, 0])
axes[0, 0].set_title("Box Plot of Age by Sex")
axes[0, 0].set_xlabel("Sex")
axes[0, 0].set_ylabel("Age")

# Box plot for Age by Death Place
sns.boxplot(data=data_grg, x="Death_place", y="Age", ax=axes[0, 1])
axes[0, 1].set_title("Box Plot of Age by Death Place")
axes[0, 1].set_xlabel("Death Place")
axes[0, 1].set_ylabel("Age")
axes[0, 1].set_xticklabels(axes[0, 1].get_xticklabels(), rotation=90)  # rotate x-axis labels 90 degrees

# Scatter plot for Birth_year with Age
sns.scatterplot(data=data_grg, x="Birth_year", y="Age", ax=axes[1, 0])
axes[1, 0].axhline(y=110, color='red')
axes[1, 0].plot(data_grg['Birth_year'], 2023 - data_grg['Birth_year'], color='orange')
ymax_relative = (114 - 108) / (123 - 108)
axes[1, 0].axvline(x=1913, ymax=ymax_relative, color='orange')
# Create an array for the x values (birth years)
x = np.arange(1907, 1914)  # 1907 to 1913 inclusive
# Compute the corresponding y values
```

```
y = 2023 - x
# Plot the orange line
axes[1, 0].plot(x, y, color='orange')
axes[1, 0].set_ylim(108, 123)
axes[1, 0].set_title("Scatter Plot of Birth Year with Age")
axes[1, 0].set_xlabel("Birth Year")
axes[1, 0].set_ylabel("Age")

# Scatter plot for Death_year with Age
sns.scatterplot(data=data_grg, x="Death_year", y="Age", ax=axes[1, 1])
axes[1, 1].axvline(x=2023, color='green')
axes[1, 1].axhline(y=110, color='red')
axes[1, 1].set_ylim(108, 123)
axes[1, 1].set_title("Scatter Plot of Death Year with Age")
axes[1, 1].set_xlabel("Death Year")
axes[1, 1].set_ylabel("Age")

# Add subfigure labels
axes[0, 0].text(-0.15, 1.15, "a)", transform=axes[0, 0].transAxes, size=12, weight='bold')
axes[0, 1].text(-0.15, 1.15, "b)", transform=axes[0, 1].transAxes, size=12, weight='bold')
axes[1, 0].text(-0.15, 1.15, "c)", transform=axes[1, 0].transAxes, size=12, weight='bold')
axes[1, 1].text(-0.15, 1.15, "d)", transform=axes[1, 1].transAxes, size=12, weight='bold')

# Adjust spacing between subplots
plt.tight_layout()

# Show the plot
plt.show()

# Convert categorical columns to category codes
for col in ['Sex', 'Death_place']:
    data_grg[col] = data_grg[col].astype('category').cat.codes

# Compute the correlation matrix
corr = data_grg[['Sex', 'Age', 'Birth_year', 'Death_year', 'Death_place']].corr()

# Create masks for the upper, lower triangles, and diagonal
mask_lower = np.tril(np.ones_like(corr, dtype=np.bool_))
mask_upper = np.triu(np.ones_like(corr, dtype=np.bool_))
mask_diag = np.eye(*corr.shape).astype(np.bool_)

# Generate a custom diverging colormap
cmap = sns.diverging_palette(230, 20, as_cmap=True)

# Create a figure and axes
```

```
fig, ax = plt.subplots(figsize=(8, 6))

# Draw the heatmap for the lower triangle with the colormap and without numbers
sns.heatmap(corr, mask=mask_upper, cmap=cmap, vmax=.3, center=0,
            square=True, linewidths=.5, cbar_kws={"shrink": .5},
                annot=False, ax=ax)

# Draw the heatmap for the upper triangle without the colormap and with numbers
sns.heatmap(corr, mask=mask_lower, cmap=['white'], cbar=False, annot=True, fmt=".2f", ax=ax)

# Draw the diagonal with both color and value
sns.heatmap(corr, mask=~mask_diag, cmap=['#8B3E2F'], cbar=False, annot=True, fmt=".2f", ax=ax)

# Set the title of the correlation heatmap
ax.set_title("Correlation Heatmap")

# Rotate axis labels by 90 degrees
ax.set_xticklabels(ax.get_xticklabels(), rotation=90)
ax.set_yticklabels(ax.get_yticklabels(), rotation=0)
# Show the plot
plt.show()
```

## D. Estudio de la Edad Máxima de Seres Humanos

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from scipy.optimize import minimize
from scipy.special import factorial
from scipy.stats import kstest
from scipy.stats import chisquare
from scipy.stats import poisson
from sklearn.svm import SVR
from sklearn.model_selection import GridSearchCV, KFold
from sklearn.metrics import mean_squared_error
from statsmodels.tsa.holtwinters import SimpleExpSmoothing
import warnings
warnings.filterwarnings("ignore")

# Load the datasets
data_idl = pd.read_excel('data_idl.xlsx',usecols=[1,2,3,4,5])

# Print the code mappings for each column
```

```
for col in ['SEX', 'COUNTRY']:
    print(f'{col} categories:')
    category_codes = {code: category for code, category in enumerate
       (sorted(data_idl[col].unique()))}
    print(category_codes)
    print()

# Convert categorical columns to category codes
for col in ['SEX', 'COUNTRY']:
    data_idl[col] = data_idl[col].astype('category').cat.codes

data_idl

data_idl.loc[data_idl['AGE'] < data_idl['AGE'].astype(int)+0.5, 'AGE
   '] = data_idl['AGE'].astype(int)
data_idl.loc[data_idl['AGE'] >= data_idl['AGE'].astype(int)+0.5, '
   AGE'] = data_idl['AGE'].astype(int)+0.5
data_idl.head(30)

grouped_stats = data_idl.groupby(['DEATH_YEAR',]).agg(
    {
        'AGE': [
            ('count', 'count'),
            ('mean', np.mean),
            ('std', np.std),
            ('median', np.median),
            ('min', np.min),
            ('max', np.max),
            ('25% quantile', lambda x: x.quantile(0.25)),
            ('75% quantile', lambda x: x.quantile(0.75))
        ]
    }
)

# Reset index to display 'group1' and 'group2' in each line
grouped_stats = grouped_stats.reset_index()

# Filter lines with count > 100
filtered_stats = grouped_stats.loc[grouped_stats[('AGE', 'count')]
   >= 100]

filtered_stats

# Convert the 'DEATH_YEAR' values in filtered_stats to a NumPy array
filtered_death_years = filtered_stats['DEATH_YEAR'].to_numpy()

# Filter the data_idl DataFrame
filtered_data_idl = data_idl[data_idl['DEATH_YEAR'].isin(
   filtered_death_years)]
filtered_data_idl.head(30)

```

```
# Get unique and sorted DEATH_YEAR values
death_years = np.sort(filtered_data_idl['DEATH_YEAR'].unique())

# Split the DEATH_YEAR values into 8 groups of 4
death_year_groups = np.array_split(death_years, 8)

# Define a list of colors for the plots
colors = ['blue', 'orange', 'green', 'red']

# Create a figure with 8 subplots, arranged in a 2x4 grid
fig, axs = plt.subplots(4, 2, figsize=(12, 15))

# Flatten the axs array for easy iteration
axs = axs.flatten()

# Create a list of labels
labels = ['a)', 'b)', 'c)', 'd)', 'e)', 'f)', 'g)', 'h)']

# Iterate over each group and each subplot axis
for i, (group, ax) in enumerate(zip(death_year_groups, axs)):
    # Iterate over each year in the group
    for j, year in enumerate(group):
        # Filter the data for the current year
        year_data = filtered_data_idl[filtered_data_idl['DEATH_YEAR
            '] == year]

        # Create a density plot for the current year in the current
            subplot
        sns.kdeplot(data=year_data, x='AGE', fill=True, color=colors
            [j], label=f'DEATH_YEAR: {year}', ax=ax)

    # Add a label to the subplot
    ax.text(-0.15, 1.15, labels[i], transform=ax.transAxes, fontsize
        =14, fontweight='bold', va='top')

    ax.set_xlim(105, 125)
    ax.legend()

# Add padding to prevent labels from overlapping
fig.tight_layout()

# Show the figure
plt.show()

# Assuming 'filtered_data_idl' is a pandas DataFrame with columns '
    DEATH_YEAR' and 'AGE'
df = filtered_data_idl[['DEATH_YEAR', 'AGE']]

def weighted_poisson_neg_log_likelihood(lmbda, x):
    weights = x / np.sum(x)
    log_likelihood = np.sum(weights * (x * np.log(lmbda) - lmbda -
```

```
        np.log(factorial(x))))
    return -log_likelihood

# Find unique death years
unique_death_years = np.arange(1987, 2019)

# Create an empty DataFrame to store the estimated lambda values for each death year
lambdas = pd.DataFrame(columns=['DEATH_YEAR', 'Estimated_Lambda'])

# Loop through each unique death year
for death_year in unique_death_years:
    # Get the ages for the current death year
    ages = df[df['DEATH_YEAR'] == death_year]['AGE'].to_numpy()

    # Estimate lambda for the current death year
    result = minimize(weighted_poisson_neg_log_likelihood, x0=1, args=(ages,), method='L-BFGS-B')
    estimated_lambda = result.x[0]

    # KS test
    ks_result = kstest(ages, 'poisson', args=(estimated_lambda,))

    # Get the observed and expected frequencies
    unique_ages, observed_freq = np.unique(ages, return_counts=True)
    expected_freq = np.exp(-estimated_lambda) * np.power(estimated_lambda, unique_ages) / factorial(unique_ages)

    # Normalize the expected frequencies so they sum to the total count of ages
    expected_freq = expected_freq / np.sum(expected_freq) * len(ages)

    # Perform the Chi-Square test
    chi2_result = chisquare(observed_freq, f_exp=expected_freq)


    # Append the estimated lambda to the lambdas DataFrame
    lambdas = lambdas.append({'DEATH_YEAR': death_year, 'Estimated_Lambda': estimated_lambda, 'KS Test Pvalue': ks_result.pvalue, 'Chi2 P-value': chi2_result.pvalue}, ignore_index=True)

lambdas

# Set parameters
quantiles = [0.5, 0.75, 0.95]
colors = ['blue', 'orange', 'purple']

# Create a DataFrame to store quantiles for each year
quantile_df = pd.DataFrame(columns=['DEATH_YEAR'] + [f'{q*100}%' for
```

```
    q in quantiles])

# Calculate and store the quantiles for each year
for _, row in lambdas.iterrows():
    quantile_values = poisson.ppf(quantiles, mu=row['Estimated_Lambda'])
    quantile_df = quantile_df.append({'DEATH_YEAR': row['DEATH_YEAR'], **{f'{q*100}%': v for q, v in zip(quantiles, quantile_values)}}, ignore_index=True)

# Quantiles to calculate for empirical data
emp_quantiles = [0.5, 1]

# Create a DataFrame to store quantiles for empirical data
empirical_quantile_df = pd.DataFrame(columns=['DEATH_YEAR'] + [f'{q*100}%' for q in emp_quantiles])

# Get unique death years
unique_death_years = filtered_data_idl['DEATH_YEAR'].unique()

# Calculate and store the quantiles for each unique death year
for death_year in unique_death_years:
    ages = filtered_data_idl[filtered_data_idl['DEATH_YEAR'] == death_year]['AGE']
    quantile_values = ages.quantile(emp_quantiles)
    empirical_quantile_df = empirical_quantile_df.append({'DEATH_YEAR': death_year, **{f'{q*100}%': v for q, v in zip(emp_quantiles, quantile_values)}}, ignore_index=True)

# Sort DataFrame by DEATH_YEAR
empirical_quantile_df = empirical_quantile_df.sort_values(by='DEATH_YEAR')

# Set plot size
plt.figure(figsize=(8, 5))

# Plot the quantiles for each year and fill the areas between them with dashed lines
for i, q in enumerate(quantiles):
    plt.plot(quantile_df[f'{q*100}%'], quantile_df['DEATH_YEAR'], label=f'{q*100}%', color=colors[i], linestyle='--')
    if i > 0:
        plt.fill_betweenx(quantile_df['DEATH_YEAR'], quantile_df[f'{quantiles[i-1]*100}%'], quantile_df[f'{q*100}%'], color=colors[i], alpha=0.3)

# Fill the area above the highest quantile
plt.fill_betweenx(quantile_df['DEATH_YEAR'], quantile_df[f'{quantiles[-1]*100}%'], color=colors[-1], alpha=0.3)

# Plot empirical quantiles
```

```
for q in emp_quantiles:
    if q == 0.5:
        plt.scatter(empirical_quantile_df[f'{q*100}%'],
            empirical_quantile_df['DEATH_YEAR'], marker='o', color='
            green', s=15, label=f'Mean age')  # color changed to
            green and size increased
    else:
        plt.scatter(empirical_quantile_df[f'{q*100}%'],
            empirical_quantile_df['DEATH_YEAR'], marker='o', color='
            black', s=15, label=f'Maximum age')

# Configure plot
plt.ylabel('Death Year')
plt.xlabel('Age')
plt.title('Age Quantiles by Death Year')
plt.xlim(105, 125)
plt.ylim(1987,2018)
plt.legend(loc='upper right')

# Show plot
plt.show()

death_years = np.arange(1987, 2019)
lambda_vals = lambdas[['Estimated_Lambda']].to_numpy()
# Perform 10-fold cross-validation to find the optimal
    hyperparameters
X = death_years.reshape(-1, 1)
y = lambda_vals
param_grid = {'C': np.arange(0.01, 1.51, 0.01), 'epsilon': np.arange
    (0.01, 1.51, 0.01)}
kfold = KFold(n_splits=10, shuffle=True, random_state=1)
svr_model = SVR(kernel='rbf')
grid_search = GridSearchCV(svr_model, param_grid, cv=kfold, scoring
    ='neg_mean_squared_error')
grid_search.fit(X, y)

# Print the best hyperparameters and mean squared error
print("Best hyperparameters:", grid_search.best_params_)
print("Mean squared error:", -grid_search.best_score_)

svr_model = SVR(kernel='rbf', C=0.76,epsilon=0.12)
svr_model.fit(X, y)

# Plot the lambda values and SVR predictions against death year
plt.figure(figsize = (8,6))
plt.scatter(death_years, lambda_vals, label='Estimated mean of
    maximum age', color = 'dodgerblue')
plt.plot(death_years, svr_model.predict(X), label='SVR adjusted
    curve',color = 'dodgerblue')
plt.xlabel('Death Year')
plt.ylabel('E (Max Age)')
```

```
plt.title('Estimated mean of maximum age and the Curve')
plt.legend()
plt.show()

# Calculate the relative increase of 'Estimated_Lambda' from the previous year
lambdas['Lambda_Increase_Rate'] = lambdas['Estimated_Lambda'].pct_change()

# Calculate the mean and standard deviation of Lambda_Increase_Rate
mean_increase_rate = lambdas['Lambda_Increase_Rate'].mean()
std_increase_rate = lambdas['Lambda_Increase_Rate'].std()

# Plot the increasing rate
plt.figure(figsize=(10, 6))
plt.plot(lambdas['DEATH_YEAR'], lambdas['Lambda_Increase_Rate'], marker = '*', color = 'orange', label='Increase Rate')

# Add a line for the mean
plt.axhline(mean_increase_rate, color='red', linestyle='--', label='Mean')

# Add lines for one standard deviation above and below the mean
plt.axhline(mean_increase_rate + std_increase_rate, color='green', linestyle='--', label='Mean + 1 SD')
plt.axhline(mean_increase_rate - std_increase_rate, color='blue', linestyle='--', label='Mean - 1 SD')

plt.title('Increasing Rate of Estimated Mean of Maximum Age Over the Years')
plt.xlabel('Death Year')
plt.ylabel('Increase Rate of Estimated Mean of Maximum Age')
plt.grid()
plt.legend()
plt.show()

# Define the next 10 years
future_years = np.arange(2018, 2040).reshape(-1, 1)

# Use the trained SVR model to predict the lambda values for the next 10 years
future_lambda_vals = svr_model.predict(future_years)

# Plot the lambda values, SVR predictions, and forecasts against death year
plt.figure(figsize = (8,6))
plt.scatter(death_years, lambda_vals, label='Estimated mean of maximum age', color = 'dodgerblue')
plt.plot(death_years, svr_model.predict(X), label='SVR adjusted curve', color = 'dodgerblue')
plt.plot(future_years, future_lambda_vals, label='Forecasts', color
```

```
    = 'orange')
plt.xlabel('Death Year')
plt.ylabel('E (Max age)')
plt.title('Predicted mean of maximum age and the Curve')
plt.legend()
plt.show()

# Print the forecasted lambda values and the 95th quantile for the
    next 10 years
for year, lmbda in zip(range(2019, 2040), future_lambda_vals):
    # Generate a Poisson distribution with the current lambda
    poisson_dist = np.random.poisson(lmbda, 1000)

    # Calculate the 95th quantile
    quantile_95 = np.percentile(poisson_dist, 95)

    print(f"Year: {year}, Forecasted Lambda: {lmbda}, 95th Quantile:
        {round(quantile_95, 2)}")

# Create a DataFrame to hold the Poisson distributions for each year
poisson_df = pd.DataFrame()

# Loop through the lambdas DataFrame for death years from 1990 to
    2015 every year
for year, lmbda in zip(range(2019, 2040), future_lambda_vals):
    # Generate a Poisson distribution with the current lambda
    poisson_dist = np.random.poisson(lmbda, 1000)

    # Add the Poisson distribution to the DataFrame
    poisson_df = pd.concat([poisson_df, pd.DataFrame({'Death Year':
        year, 'Age': poisson_dist})])

# Create a violin plot of the Poisson distributions
plt.figure(figsize=(20, 8))
sns.violinplot(data=poisson_df, x='Death Year', y='Age')

# Set the lower limit of the y-axis to 105
plt.ylim(105, None)

plt.title('Forecasted Media and Maximum Age for Each Year')
plt.show()
```

## E. Estudio de la Ratio de Mortalidad de los Supercentenarios

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

```
from scipy.optimize import minimize
from scipy.special import factorial
from scipy.stats import kstest
from scipy.stats import chisquare
from scipy.stats import poisson
from sklearn.svm import SVR
from sklearn.model_selection import GridSearchCV, KFold
from sklearn.metrics import mean_squared_error
from statsmodels.tsa.holtwinters import SimpleExpSmoothing
import warnings
warnings.filterwarnings("ignore")

data = pd.read_excel("qx_calculation.xlsx")
df1qx = data[['Agebar.','1qx']].dropna()
df5qx = data[['Agebar.5','.5qx']].dropna()
df25qx = data[['Agebar.25','.25qx']].dropna()
print(df5qx.head())
plt.scatter(df5qx['Agebar.5'],df5qx['.5qx'])
plt.show()

X = df5qx['Agebar.5'].values.reshape(-1, 1)
y = df5qx['.5qx']
param_grid = {'C': np.arange(0.01, 0.5, 0.01), 'epsilon': np.arange
    (0.01, 0.5, 0.01)}
kfold = KFold(n_splits=10, shuffle=True, random_state=1)
svr_model = SVR(kernel='rbf')
grid_search = GridSearchCV(svr_model, param_grid, cv=kfold, scoring
    ='neg_mean_squared_error')
grid_search.fit(X, y)

# Print the best hyperparameters and mean squared error
print("Best hyperparameters:", grid_search.best_params_)
print("Mean squared error:", -grid_search.best_score_)

svr_model = SVR(kernel='rbf', C=0.27,epsilon=0.03)
svr_model.fit(X, y)
# Get the SVR predictions
svr_preds = svr_model.predict(X)
# Calculate the root mean squared error (RMSE)
rmse = np.sqrt(mean_squared_error(df5qx['.5qx'], svr_preds))
# Print the RMSE
print('RMSE:', rmse)

# Plot the lambda values and SVR predictions against death year
plt.figure(figsize=(8, 5))
plt.scatter(df5qx['Agebar.5'],df5qx['.5qx'], label='Mortality Rate',
      color = 'dodgerblue')
plt.plot(df5qx['Agebar.5'], svr_preds, color='orange', marker = '^',
     markersize=5, label='SVR curve')
plt.xlabel('AGE')
plt.ylabel('0.5 qx')
```

```
plt.title('Mortality Rate and SVR Regression Line')
plt.legend()
plt.show()

# Functions to calculate RSS, AIC, BIC
def calculate_rss(y, smoothed_y):
    return np.sum((y - smoothed_y)**2)

def calculate_aic(rss, n, k):
    return n * np.log(rss/n) + 2 * k

def calculate_bic(rss, n, k):
    return n * np.log(rss/n) + np.log(n) * k

n = len(y)  # n mero de observaciones
k = 3  # n mero de par metros en SVR con kernel RBF (C, epsilon y el par metro de escala del kernel)

# Calculate the RSS for the SVR model
rss_svr = calculate_rss(y, svr_preds)

n = len(y)
# Number of parameters in the SVR model (kernel, C, epsilon)
k_svr = 3

# Calculate AIC and BIC for the SVR model
aic_svr = calculate_aic(rss_svr, n, k_svr)
bic_svr = calculate_bic(rss_svr, n, k_svr)

print(f"SVR AIC: {aic_svr}")
print(f"SVR BIC: {bic_svr}")

x = df5qx['Agebar.5']
y = df5qx['.5qx']

def whittaker_henderson(y, lambda_, d=2):
    m = len(y)
    E = np.identity(m)
    D = np.diff(E, n=d, axis=0)

    penalty_matrix = D.T @ D
    smoothed_y = np.linalg.inv(E + lambda_ * penalty_matrix) @ y

    return smoothed_y

# Functions to calculate RSS, AIC, BIC
def calculate_rss(y, smoothed_y):
    return np.sum((y - smoothed_y)**2)

def calculate_aic(rss, n, k):
    return n * np.log(rss/n) + 2 * k
```

```
102
103 def calculate_bic(rss, n, k):
104     return n * np.log(rss/n) + np.log(n) * k
105
106 n = len(y)                          # number of data points
107 k = 1                               # number of estimated parameters
108 lambda_values = np.logspace(1, 10, 10)   # range of lambda values
109 cv_scores = []
110
111 kf = KFold(n_splits=10, shuffle=True, random_state=42)
112
113 for lambda_ in lambda_values:
114     smoothed_y = whittaker_henderson(y, lambda_)
115
116     # 10-fold cross-validation
117     cv_rss = []
118     for train_index, val_index in kf.split(y):
119         y_train, y_val = y[train_index], y[val_index]
120         smoothed_y_train = whittaker_henderson(y_train, lambda_)
121         rss_train = calculate_rss(y_train, smoothed_y_train)
122         cv_rss.append(rss_train)
123     cv_score = np.mean(cv_rss)
124     cv_scores.append(cv_score)
125
126 # Find the optimal lambda value based on 10-fold cross-validation
127 optimal_lambda_cv = lambda_values[np.argmin(cv_scores)]
128 print(f"Optimal lambda value based on 10-fold cross-validation: {
    optimal_lambda_cv}")
129
130 # Calculate the smoothed_y, rss, aic, and bic at the optimal lambda
131 smoothed_y_opt = whittaker_henderson(y, optimal_lambda_cv)
132 rss_opt = calculate_rss(y, smoothed_y_opt)
133
134 aic_opt = calculate_aic(rss_opt, n, k)
135 bic_opt = calculate_bic(rss_opt, n, k)
136
137 print(f"Corresponding RSS: {rss_opt}")
138 print(f"Corresponding AIC: {aic_opt}")
139 print(f"Corresponding BIC: {bic_opt}")
140
141 # Use the optimal lambda found using cross-validation
142 lambda_ = optimal_lambda_cv
143
144 smoothed_y = whittaker_henderson(y, lambda_)
145
146 # Calculate the root mean squared error (RMSE)
147 rmse = np.sqrt(mean_squared_error(df5qx['.5qx'], smoothed_y))
148 # Print the RMSE
149 print('RMSE:', rmse)
150
151 # Display smoothed values
```

```
for age, smoothed_value in zip(x, smoothed_y):
    print(f"Age: {age}, Smoothed value: {smoothed_value:.6f}")

# Plot the lambda values and smoothed predictions against age
plt.figure(figsize=(8, 5))
plt.scatter(df5qx['Agebar.5'],df5qx['.5qx'], label='Mortality Rate', color = 'dodgerblue')
plt.plot(df5qx['Agebar.5'],smoothed_y, color='orange', marker = '*', markersize=5, label='Whittaker Henderson curve')
plt.xlabel('AGE')
plt.ylabel('0.5 qx')
plt.title('Mortality Rate and Whittaker Henderson Smoothing Line')
plt.legend()
plt.show()

# SMA
best_mse_sma = float('inf')
best_sma_range = None
smoothed_y_series = pd.Series(smoothed_y)

for r in range(4, len(smoothed_y_series)):
    sma_forecast = smoothed_y_series.rolling(r).mean()
    sma_forecast = sma_forecast[r-1:]
    mse = mean_squared_error(smoothed_y_series[r-1:-1], sma_forecast[:-1])
    if mse < best_mse_sma:
        best_mse_sma = mse
        best_sma_range = r

# ES
model = SimpleExpSmoothing(smoothed_y)
model_fit = model.fit(optimized=True, use_brute=True)
fitted_values = model_fit.fittedvalues
best_mse_es = mean_squared_error(smoothed_y[1:], fitted_values[1:])

# LE
best_mse_le = float('inf')
best_le_range = None
for r in range(2, len(smoothed_y_series)):
    last_r_smoothed_y = smoothed_y[-r:]
    differences = np.diff(last_r_smoothed_y)
    mean_difference = np.mean(differences)
    predicted_values_le = smoothed_y + mean_difference
    mse = mean_squared_error(smoothed_y[:-1], predicted_values_le[:-1])
    if mse < best_mse_le:
        best_mse_le = mse
        best_le_range = r

# Compare the three models
print(f'Best MSE for SMA: {best_mse_sma} with range: {best_sma_range
```

```
    }')
print(f'Best MSE for ES: {best_mse_es} with smoothing level: {
    model_fit.model.params["smoothing_level"]}')
print(f'Best MSE for LE: {best_mse_le} with range: {best_le_range}')

# Choose the best model
if min(best_mse_sma, best_mse_es, best_mse_le) == best_mse_sma:
    # Use the best range to compute the SMA forecast
    predicted_values = [smoothed_y_series.rolling(best_sma_range).
        mean().iloc[-1]]
    print("SMA chosen as best model")
elif min(best_mse_sma, best_mse_es, best_mse_le) == best_mse_es:
    # Use the best model to compute the ES forecast
    predicted_values = [fitted_values[-1]]
    print("ES chosen as best model")
else:
    # Get the differences for the best range
    last_r_smoothed_y = smoothed_y[-best_le_range:]
    differences = np.diff(last_r_smoothed_y)
    mean_difference = np.mean(differences)
    predicted_values = [smoothed_y_series.iloc[-1] + mean_difference
        ]
    print("LE chosen as best model")

# Predict future values until they reach 1
predicted_ages = [last_age]
for i in range(10):
    predicted_age = predicted_ages[-1] + 0.5
    if min(best_mse_sma, best_mse_es, best_mse_le) == best_mse_sma:
        predicted_value = np.mean(predicted_values[-best_sma_range
            :])
    elif min(best_mse_sma, best_mse_es, best_mse_le) == best_mse_es:
        predicted_value = model_fit.predict(start=len(smoothed_y) +
            i, end=len(smoothed_y) + i)[0]
    else:
        predicted_value = predicted_values[-1] + mean_difference
    if predicted_value >= 0.96:
        break
    predicted_ages.append(predicted_age)
    predicted_values.append(predicted_value)

# Print the predicted ages and values
for age, value in zip(predicted_ages, predicted_values):
    print(f"Age: {age}, Predicted .5qx: {value:.6f}")

# Plot the original data, the smoothed values, and the predicted
    values
plt.figure(figsize=(10, 6))
plt.scatter(df5qx['Agebar.5'], df5qx['.5qx'], label='Mortality Rate
    ', color='dodgerblue')
plt.plot(df5qx['Agebar.5'], smoothed_y, color='orange', marker='*',
```

```
    markersize=7, label='Whittaker Henderson curve')
plt.plot(predicted_ages, predicted_values, color='red', marker='x',
    markersize=6, label='Predicted values')
plt.xlabel('AGE')
plt.ylabel('0.5 qx')
plt.title('Mortality Rate with Whittaker Henderson Smoothing Line
    and Predicted values')
plt.legend()
plt.grid(True)
plt.show()
```