TRABAJO DE FIN DE GRADO

Grado en Estadística

Título: Estudio de un modelo predictivo para la estimación de la variación de la población española, tanto en zonas rurales como zonas urbanas.

Autor: Jose Luis Mourelo Rodriguez

Tutor académico: Maricarme Riera

Departamento: Econometria, Estadistica y Economia Aplicada

Convocatoria: 1er semestre 2024/2025

:

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

Resumen del trabajo:

La variación de la población es un tema de estudio fundamental para todos los países del mundo debido a su impacto directo en la calidad de vida de sus habitantes.
Este trabajo analiza la variación poblacional anual en España, utilizando métodos estadísticos y de aprendizaje automático para entender los factores que más influyen en este fenómeno y determinar qué modelo es más adecuado para predecirla. Los principales métodos utilizados son Random Forest, XGBoost y ARIMA, que ayudan a identificar patrones temporales y relaciones entre variables relacionadas con la demografía, la economía y la sociedad.

El análisis se realiza con datos anuales del periodo 2010-2023, considerando variables como nacimientos, defunciones, PIB per cápita, tasas de empleo y paro, precios de la vivienda y matrimonios por 1000 habitantes. Entre los modelos estudiados, XGBoost se posiciona como el más eficaz, logrando un R^2 cercano a 0.38, lo que indica una capacidad moderada para explicar las variaciones en la población. También se destacan como factores clave el índice general de vivienda, las tasas de nacimientos y defunciones, y las dinámicas del empleo.

El trabajo subraya, además, el impacto de fenómenos externos como crisis económicas, cambios climáticos o movimientos migratorios en las tendencias demográficas. Por último, se concluye que los modelos predictivos son herramientas útiles para diseñar políticas públicas y estrategias de desarrollo, aunque sería importante incorporar más variables y enfoques interdisciplinarios en futuras investigaciones para mejorar la precisión y utilidad de las predicciones.

La totalidad del trabajo ha sido realizado mediante el uso del lenguaje de programación R.

Palabras clave: España, variación poblacional, demografía, ruralidad, predicción

Abstract:

Population variation is a fundamental area of study for all countries worldwide due to its direct impact on the quality of life of their inhabitants.
This work analyzes the annual population variation in Spain using statistical methods and machine learning techniques to understand the factors that most influence this phenomenon and to determine which model is best suited for prediction. The main methods employed are Random Forest, XGBoost, and ARIMA, which help identify temporal patterns and relationships between variables related to demography, economy, and society.

The analysis is based on annual data from the period 2010-2023, considering variables such as births, deaths, GDP per capita, employment and unemployment rates, housing prices, and marriages per 1000 inhabitants. Among the models studied, XGBoost emerges as the most effective, achieving an R^2 of approximately 0.38, indicating a moderate capacity to explain population variation. Key factors include the general housing index, birth and death rates, and employment dynamics.

The study also highlights the impact of external phenomena such as economic crises, climate change, or migration movements on demographic trends. Finally, it concludes that predictive models are valuable tools for designing public policies and development strategies. However, incorporating additional variables and interdisciplinary approaches in future research would be important to improve the precision and usefulness of the predictions.

The entire work has been carried out throught the use of the R programming language.

Key words: Spain, population variation, demography, rurality, prediction

AMERICAN MATHEMATICAL SOCIETY (AMS) CLASSIFICATION

62-XX - STATISTICS

62M20 – Prediction

62Jxx – Linear inference, regression

68T09 – Computacional aspects of data analysis and big data

91D20 – Mathematical geography and demography

INDICE

1. INTRODUCCIÓN

1.1. Planteamiento del problema

"El cambio es inevitable; el crecimiento es opcional." Esta cita de John C. Maxwell, escritor estadounidense, capta el espíritu de los estudios demográficos. Esta disciplina permite analizar la manera en que las sociedades cambian, evolucionan y se transforman. Las dinámicas poblacionales son el reflejo directo de complejas interacciones entre factores sociales, económicos, culturales y políticos. Comprender estas dinámicas, especialmente en contextos como el de España, no es solo un desafío académico, es una necesidad estratégica para la planificación del futuro.

España, como otros países desarrollados, se enfrenta a muchos retos demográficos que requieren un análisis exhaustivo. En los últimos años, la estructura demográfica ha cambiado fundamentalmente debido a fenómenos como el envejecimiento de la población, la continua disminución de las tasas de natalidad y las variaciones en los flujos migratorios. El Instituto Nacional de Estadística (INE) predice que a mediados del siglo XXI España verá un descenso significativo de su población activa y la proporción de su población de 65 y más años alcanzará niveles históricos (INE, 2022). Este escenario, que pone presión sobre los sistemas de pensiones y los servicios de salud, es especialmente evidente en países donde las diferencias territoriales y culturales intensifican los efectos del cambio demográfico.

Los cambios en la población española están determinados no sólo por el número de nacimientos y muertes, sino también por una red más amplia de factores interrelacionados. Por ejemplo, los flujos migratorios han demostrado ser un factor importante para la sostenibilidad demográfica, mientras que variables económicas como el Índice de Precios de Consumo (IPC) y el Índice de Precios a la Vivienda (IPV) influyen indirectamente en la capacidad de las familias para planificar la maternidad. Además, fenómenos sociales como la disminución del matrimonio, el aumento de las interrupciones voluntarias de embarazo (IVE) y la creciente influencia del turismo en los mercados inmobiliarios están transformando las dinámicas de poder en las familias y las comunidades.

El objetivo de este estudio es crear varios modelos predictivos que puedan analizar y predecir las variaciones anuales de la población española. Estos modelos incorporan una amplia gama de variables, entre ellas nacimientos, defunciones, emigración, matrimonios, IPV, PIB, censo poblacional, censo municipal, mercado laboral y exportación e importación de combustibles. Con un enfoque cuantitativo que utiliza datos tanto históricos como actuales, este análisis no solo reconoce las tendencias más importantes, sino que también ofrece ideas para intentar guiar el diseño de políticas públicas hacia un desarrollo justo y sostenible.

Varios estudios previos han proporcionado perspectivas valiosas sobre los cambios demográficos en España. Entre ellos destaca el trabajo de Goerlich y Cantarino (2015),

quienes utilizaron técnicas espaciales para modelar la densidad de población y revelar patrones de concentración urbana y despoblación rural. De manera similar, Molina (2018) realizó un análisis de la despoblación en España, destacando la pérdida constante de población en las zonas rurales. Sin embargo, aún no se han explorado enfoques que combinen estas variables tradicionales con nuevos factores como el impacto del turismo, el IPV y los patrones culturales que redefinen la estructura familiar como por ejemplo la tasa de paro o la tasa de empleo. Este tipo de patrones culturales que comentamos serían claves también para el estudio ya que son los encargados de representar las diferentes formas de vivir en cada miembro de nuestra sociedad como bien entendemos en Manzanares (2015).

Este estudio se basa en el supuesto de que los flujos migratorios son actualmente el factor más volátil y determinante de la dinámica poblacional española. En particular, la migración internacional puede desempeñar un papel compensatorio frente al descenso natural de la población, mientras que un éxodo de jóvenes puede exacerbar aún más el descenso de la población activa en determinadas comunidades autónomas. Además, factores económicos como el índice de precios de la vivienda, medido por el IPV, influyen indirectamente en las decisiones de las familias en términos de fertilidad y movilidad, especialmente en zonas con una alta proporción de vivienda turística. En este sentido, el impacto del turismo en los mercados inmobiliarios y las economías locales puede tener efectos demográficos significativos, ya sea desplazando a las poblaciones locales o creando oportunidades para atraer inmigrantes.

Por otro lado, los cambios culturales y sociales también juegan un papel importante. La disminución del matrimonio y el aumento de los abortos voluntarios reflejan cambios importantes en los patrones familiares y reproductivos. Estas tendencias no sólo tienen un impacto directo en la variación de población, sino también efectos a largo plazo en la composición social y económica de la población. Diversos estudios, como por ejemplo Cañada (2013), afirman que factores como el incremento extraordinario en el nivel de educación y una mayor estabilidad en las pautas de emparejamiento, son esenciales para comprender y predecir la dinámica demográfica de un país.

En este contexto, este estudio pretende ofrecer una primera respuesta a las siguientes preguntas importantes: ¿Cómo interactúan los factores demográficos, económicos y culturales en la dinámica de la población española? ¿Qué variables tienen mayor influencia en la dinámica demográfica y cómo influyen estas últimas por comunidades autónomas y/o por provincias? ¿Proporcionan los modelos realizados estimaciones confiables para una planificación futura viable? En base a estas preguntas, se proponen las siguientes hipótesis.

- En primer lugar, se espera que los flujos migratorios internacionales sean el factor más decisivo en el cambio de la población española en los próximos años. Esto incluye tanto la inmigración como la emigración, que están estrechamente relacionadas.
- En segundo lugar, factores económicos como el IPV y el IPC pueden tener un impacto indirecto pero significativo en la fertilidad, especialmente en áreas con una alta concentración de viviendas turísticas. Estas regiones pueden enfrentar

tensiones adicionales debido a la competencia por recursos limitados y cambios en los precios de la vivienda.

- Finalmente, otros factores económico-sociales como el PIB o las tasas de paro y empleo son destacables a primera vista ya que pueden cambiar las expectativas generales de la población con relación al hecho de tener hijos. Es lógico pensar que una mayor prosperidad económica y estabilidad en la calidad de vida de la población facilita significativamente la toma de decisiones relacionadas con la ampliación de la familia, incluida la planificación de tener hijos.

El desarrollo de este estudio pretende no sólo confirmar o rechazar estas hipótesis, sino también proporcionar una base empírica para la toma de decisiones informadas. En tiempos en que el cambio es inevitable, la capacidad de adaptarse y planificar es fundamental para garantizar el bienestar de las generaciones futuras.
La presente investigación pretende contribuir a este objetivo analizando en profundidad las tendencias demográficas y creando modelos que reflejen las realidades complejas y en constante evolución de la sociedad española.

El análisis de la variación poblacional no se limitará a identificar tendencias. Esta es una herramienta poderosa para predecir el futuro y prepararse para los desafíos que trae. Entendiendo cómo interactúan las variables demográficas, económicas y culturales, no sólo podemos predecir cambios sino también convertirlos en oportunidades de crecimiento y desarrollo sostenible en España. Un ejemplo claro de esto es anticipar el envejecimiento de la población en zonas rurales incorporando servicios geriátricos y transporte más accesible. Asimismo, si se predice un aumento de población en áreas metropolitanas, se pueden planificar nuevas infraestructuras y servicios que eviten la congestión y desigualdad social.

1.2. Objetivos generales y específicos del proyecto

El principal objetivo de este estudio es analizar y predecir cambios en la población española mediante el desarrollo de un modelo predictivo que integre tanto factores demográficos tradicionales como nuevas variables socioeconómicas y culturales. El objetivo de este enfoque es comprender de manera integral las tendencias demográficas actuales, identificar tendencias pasadas y presentes y predecir escenarios futuros para respaldar decisiones estratégicas y de políticas públicas, contribuyendo así al desarrollo sostenible.

El primer objetivo es determinar la importancia de los flujos migratorios en la composición de la población española, teniendo en cuenta tanto la inmigración como la emigración y su relación con factores económicos y sociales. En un contexto caracterizado por una disminución sostenida de las tasas de fertilidad y una aceleración del envejecimiento de la población, los movimientos migratorios son un factor clave para mitigar el impacto de estas tendencias y garantizar la sostenibilidad demográfica a

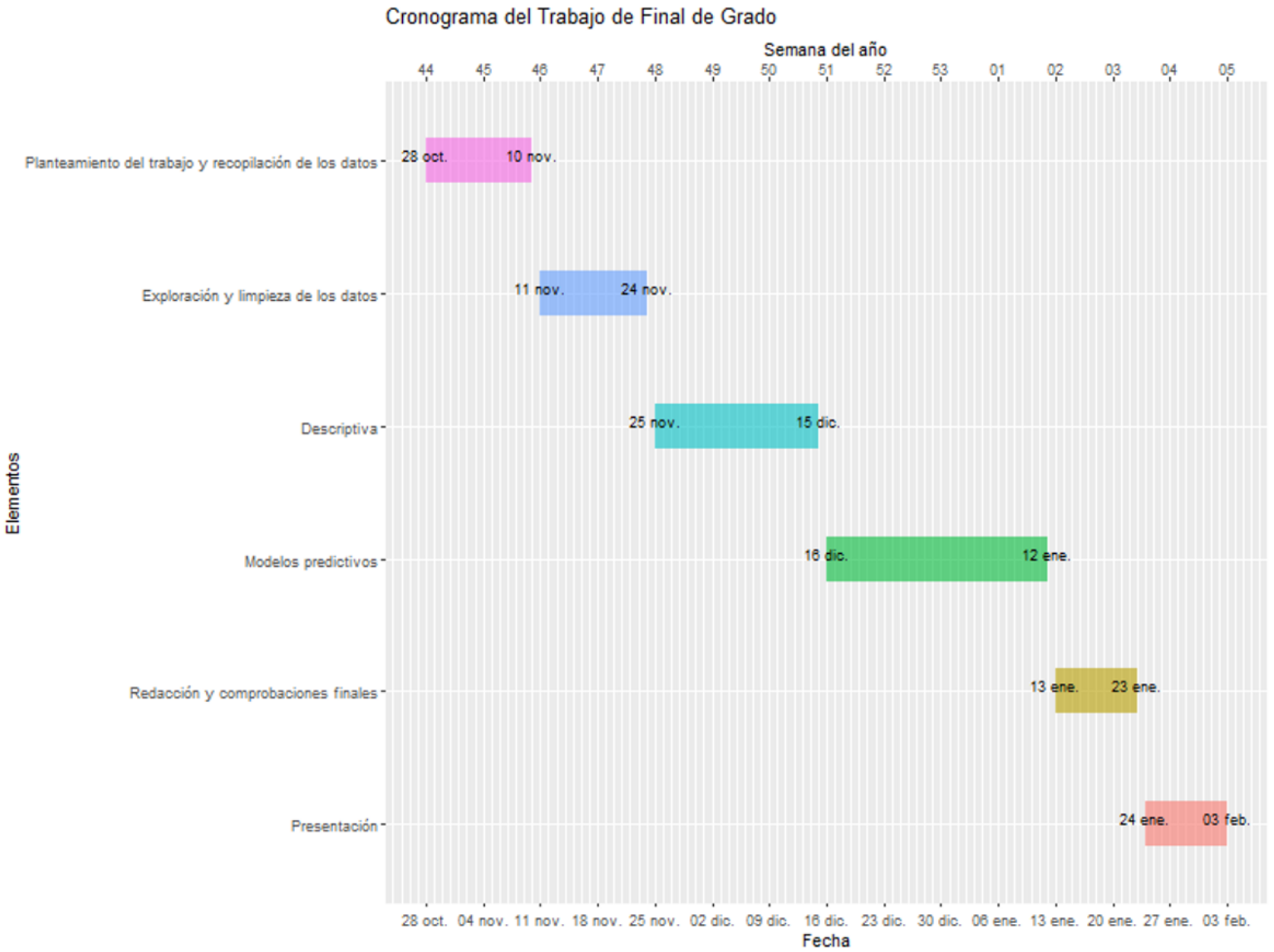
largo plazo. De manera similar, este estudio busca evaluar el impacto de factores económicos como el Índice de Precios al Consumidor (IPC), el Índice de Precios de la Vivienda (IPV), el Producto Interior Bruto (PIB) y las tasas de Paro/Empleo en las tendencias demográficas. Estas variables económicas tienen un impacto indirecto pero significativo en la fertilidad, la migración y las decisiones de planificación a largo plazo de las familias porque afectan el costo de vida, las oportunidades de empleo y la accesibilidad a la vivienda. La identificación y análisis de estas interacciones permitirá desarrollar modelos más precisos y representativos de la realidad española.

Otro objetivo importante es analizar el papel de los cambios culturales y sociales en la estructura de la población, como la disminución de las tasas de matrimonio, el aumento del aborto voluntario (IVE) y los cambios en las formas de familia. Estos fenómenos reflejan cambios culturales que están redefiniendo las relaciones familiares y afectan directamente los nacimientos y, por tanto, la estructura de la población. Finalmente, este estudio tiene como objetivo construir un modelo predictivo final que pueda predecir cambios poblacionales en diferentes escenarios temporales basándose en datos históricos y tendencias actuales. Estos modelos se desarrollan integrando variables demográficas, económicas y sociales con el fin de adaptarse a diferentes escalas regionales, teniendo en cuenta tanto el contexto nacional como las particularidades de las comunidades autónomas, o incluso de las provincias. Este enfoque de componentes múltiples tiene como objetivo crear una herramienta útil para planificadores, agentes encargados de formular las políticas públicas y otras partes interesadas en abordar los desafíos asociados con el cambio demográfico en España.

En conclusión, este estudio pretende ofrecer un análisis exhaustivo de la dinámica demográfica de España, contribuyendo al conocimiento científico y proporcionando una base sólida para planificar un futuro más sostenible y justo.

1.3. Cronograma del Trabajo

En este cronograma realizado en RStudio se muestra el tiempo dedicado a la ejecución de cada una de las tareas que constituyen este trabajo.


Cronograma del Trabajo de Final de Grado
Semana del año
44
45
46
47
48
49
50
51
52
53
01
02
03
04
05
Elementos
Planteamiento del trabajo y recopilación de los datos
28 oct.
10 nov.
Exploración y limpieza de los datos
11 nov.
24 nov.
Descriptiva
25 nov.
15 dic.
Modelos predictivos
16 dic.
12 ene.
Redacción y comprobaciones finales
13 ene.
23 ene.
Presentación
24 ene.
03 feb.
28 oct.
04 nov.
11 nov.
18 nov.
25 nov.
02 dic.
09 dic.
16 dic.
23 dic.
30 dic.
06 ene.
13 ene.
20 ene.
27 ene.
03 feb.
Fecha

2. METODOLOGÍA

En esta sección, se explicarán algunos de los conceptos integrados en este trabajo. En segundo lugar, se detallará la metodología utilizada para determinar la variación poblacional en España en los próximos años, a corto y largo plazo. Se explicarán en detalle los modelos estadísticos utilizados, su teoría y utilidad, y se proporcionarán ejemplos prácticos. Además, se incluirán las fórmulas estadísticas relevantes para cada modelo.

2.1. Definición de conceptos

IPC:
El IPC (o Índice de Precios de Consumo) es un indicador que se utiliza para medir la evolución de los precios de los bienes y servicios que consumen las familias. Al igual que la inflación, el IPC permite ver el aumento del coste de la vida en una economía.

IPV:
El IPV (o Índice de Precios de Vivienda) es un indicador que tiene como principal objetivo medir la evolución del nivel de los precios de compraventa de las viviendas de precio libre, tanto nuevas como de segunda mano, a lo largo del tiempo. Se trata, por tanto, de un indicador concebido únicamente para establecer comparaciones en el tiempo.

PIB:
El PIB (o Producto Interior Bruto) mide el valor monetario de los bienes y servicios finales, es decir, los que adquiere el consumidor final, producidos por un país en un período determinado.

IVE:
El IVE (o Interrupción Voluntaria del Embarazo) es un procedimiento clínico para finalizar un embarazo, realizado por profesionales sanitarios acreditados. Hace referencia al derecho al aborto con la solicitud como único requisito hasta la semana catorce (incluida) de gestación.

ECOICOP: es la clasificación europea de consumo, que facilita la consulta conjunta con otras estadísticas como el IPC, y a su vez permite la comparabilidad de datos con otros países de la Unión Europea (EU).
Definición del INE: Clasificación de bienes y servicios ECOICOP - INE. Instituto Nacional ...

Dos posibles cálculos para la variación anual de población:

- Variación absoluta (valores numéricos): se calcula restando el valor del censo del año anterior del valor del censo del año actual

- Variación porcentual (en términos relativos): Se calcula como la diferencia entre años consecutivos, expresada como un porcentaje del valor del año anterior.

Tipos de modelos de predicción:

- Modelos de clasificación; Los modelos de clasificación permiten predecir la pertenencia a una clase. Por ejemplo, si tratamos de clasificar entre nuestros clientes quiénes son más propensos al abandono. Los resultados del modelo son binarios, o un sí o un no (en forma de 0 y 1) con su grado de probabilidad. Es decir, nos pueden decir que un cliente nos abandonará con el 89% de probabilidad.

- Modelos de regresión; Los modelos de regresión nos permiten predecir un valor. Por ejemplo, cuál es el beneficio estimado que obtendremos de un determinado cliente en el próximo año.

A pesar de que cada modelo tiene una metodología diferente, el objetivo general de todos ellos es similar: predecir resultados futuros basándose en datos pasados.
Por lo tanto, el modelo de predicción final a utilizar siempre dependerá de los datos que tengamos y sobre todo de su buena adherencia al modelo.

En nuestro caso, como nuestra idea en predecir la variación de población, es decir predecir un valor, es recomendable directamente estudiar modelos de regresión.

2.2. Modelos Lineales

Los modelos lineales (LM) son un tipo de modelo estadístico utilizado para modelar la relación entre una variable dependiente y una o varias variables independientes. En el caso de la determinación de la variación anual poblacional, los modelos lineales pueden utilizarse para modelar la relación entre la variación anual poblacional y las características de la población, como el PIB, el IPV y la tasa de empleo, siempre y cuando exista una relación lineal entre estas variables predictoras y la variable dependiente.

Uno de los modelos lineales más utilizados es el modelo de regresión lineal. En este modelo, se asume que la relación entre la variable dependiente (esperanza de vida) y las variables independientes es lineal. La ecuación general para el modelo de regresión lineal es:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

Donde y es la variable dependiente (esperanza de vida), beta_0 es el termino constante del modelo, y β_1, β_2, …, β_n son los coeficientes de regresión que representan la contribución de cada variable independiente x_1, x_2, …, x_n a la variable dependiente (esperanza de vida).

Para que los resultados de un modelo de regresión lineal sean válidos, deben cumplirse ciertas reglas iniciales:

- Linealidad: la relación entre la variable dependiente y las variables independientes debe ser lineal.
- Independencia: las observaciones deben ser independientes entre sí.
- Homocedasticidad: la varianza de los errores debe ser constante a lo largo de todos los valores de las variables independientes.
- los errores del modelo deben seguir una distribución normal.
- No Multicolinealidad: las variables independientes no deben estar altamente correlacionadas entre sí.

La utilidad de los modelos lineales radica en su capacidad para modelar relaciones lineales entre las variables dependientes e independientes. Además, en nuestro caso, los modelos lineales proporcionan información sobre la contribución relativa de cada variable independiente a la esperanza de vida de la póliza.

2.3. Modelos Lineales Generalizados

Los modelos lineales generalizados (GLM) son una extensión de los modelos lineales que permiten modelar la relación entre una variable dependiente y una o varias variables independientes cuando la distribución de la variable dependiente no sigue una distribución normal. En el contexto de la determinación de la variación de una población a lo largo del tiempo, los GLM pueden utilizarse para modelar la relación entre la variación de la población y las características de la población en concreto, incluso cuando la distribución de la variación poblacional no sigue una distribución normal.

Un ejemplo común de GLM es el modelo de regresión logística. Este modelo se utiliza cuando la variable dependiente es binaria o categórica. La función de enlace utilizada en el modelo de regresión logística es la función logit, La ecuación general para el modelo de regresión logística es:

$$logit(p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

Donde p es la probabilidad de éxito (esperanza de vida corta o larga), β_0 es el término constante, β_1, β_2, …, β_n son los coeficientes de regresión que representan la contribución de cada variable independiente x_1, x_2, …, x_n a la probabilidad de éxito.

Validar un modelo de regresión logística implica evaluar su capacidad para predecir correctamente la variable dependiente y verificar que cumple con las premisas iniciales. A continuación, se describen algunos métodos comunes para la validación:

- División del conjunto de datos: Se divide el conjunto de datos en un conjunto de entrenamiento, el que será entrenado, y un conjunto de prueba, el que se evaluará con los resultados del entrenamiento.
- Evaluación de la bondad del ajuste: se evalúa generalmente con el criterio AIC (Criterio de Información de Akaike) o BIC (Criterio de información Bayesiano)
- Métricas de evaluación de predicción: se evalúa generalmente realizando una matriz de confusión, que evalúa la exactitud de las predicciones, o con AUC (Área Bajo la curva), que evalúa la capacidad del modelo para distinguir entre clases.

La utilidad de los GLM radica en su capacidad para modelar relaciones no lineales y distribuciones no normales. Además, los GLM permiten analizar la influencia de las variables independientes en la probabilidad de éxito, lo que proporciona información valiosa sobre la relación entre las características de la póliza y su esperanza de vida.

2.4. Modelos aditivos generalizados

Los modelos aditivos generalizados (GAM) son una extensión flexible de los modelos lineales generalizados (GLM), diseñada para capturar relaciones no lineales entre las variables independientes y la variable dependiente. Mientras que los modelos lineales tradicionales suponen que la relación entre las variables es estrictamente lineal, los GAM permiten modelar relaciones más complejas utilizando funciones suaves para cada predictor, lo que proporciona una mayor capacidad de ajuste y, al mismo tiempo mantiene interpretabilidad al modelo.

Un modelo aditivo generalizado se expresa normalmente mediante la fórmula matemática siguiente:

$$g(E[y]) = \beta_0 + f_1(x_1) + f_2(x_2) + \ldots + f_p(x_P)$$

Donde:

- $g(\cdot)$ es la función de enlace, relacionando la media de la variable dependiente y con el modelo aditivo
- E[y] es la esperanza matemática de la variable dependiente
- $f_1(x_1)$, $f_2(x_2)$, …, $f_p(x_p)$ son funciones aplicadas a cada predictor x_1, x_2, …, x_p que permiten modelar relaciones no lineales

- p es el número de predictores, variables independientes

La principal diferencia de estos modelos con los modelos lineales o los modelos lineales generalizados es el uso de las funciones $f_i(x_i)$, que, al no ser estrictamente lineales, otorgan a los GAM una flexibilidad significativa para capturar patrones en los datos.

La estimación de los parámetros del GAM implica resolver un problema de optimización que ajusta simultáneamente las funciones f(x) para cada predictor. Este proceso se realiza mediante técnicas como el método iterativo de mínimos cuadrados ponderados, similar a los GLM, pero con un enfoque adicional para manejar las funciones suaves.
La evaluación del modelo de un GAM se realiza mediante métricas como el error cuadrático medio (RMSE), el coeficiente de determinación (R^2) y el criterio de información de Akaike (AIC).

2.5. Modelos random forest

El modelo Random Forest es una técnica de aprendizaje automático que se utiliza tanto para problemas de regresión como de clasificación y está basada en la combinación de diferentes árboles de decisión durante el entrenamiento. La predicción final es el resultado de promediar (para regresión) o votar (para clasificación) sus resultados, ya que se obtiene una predicción más precisa y estable. La naturaleza no paramétrica de Random Forest lo convierte en un modelo extremadamente flexible, capaz de capturar relaciones complejas y no lineales entre las variables predictoras y la variable respuesta.

El modelo de Random Forest puede definirse como un conjunto de T arboles de decisión, donde cada árbol $h_t(x)$ se entrena de forma independiente con su subconjunto aleatorio extraído de los datos de entrenamiento:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(x)$$

Donde:

- T es el número de árboles de decisión
- $h_t(x)$ es la predicción de un árbol de decisión individual para un conjunto de predictores x
- $\hat{y}$ es la predicción final de un modelo (promedio de todos los árboles para el caso de regresión o voto mayoritario para el caso de clasificación).

La selección de un subconjunto aleatorio de variables en cada división de nodo permite a Random Forest manejar problemas de multicolinealidad y mitigar el sobreajuste. El número de variables consideradas en cada nodo es un hiperparámetro clave y generalmente se establece como sqrt(p) para problemas de regresión, en donde p es el

número total de variables predictoras.
La evaluación de un modelo Random Forest generalmente se realiza mediante métricas como el error cuadrático medio (RMSE) y el coeficiente de determinación R^2 para problemas de regresión, o métricas como el área bajo la curva (AUC) en problemas de clasificación.

2.6. Modelos XGBoost

XGBoost es una técnica de aprendizaje automático que utiliza un algoritmo de boosting para mejorar el rendimiento de otros modelos estadísticos, como los modelos lineales y los modelos de árboles de decisión. El boosting es un enfoque que combina múltiples modelos más débiles para crear un modelo más fuerte y preciso.

En el contexto de la determinación de la variación de la población, XGBoost se puede utilizar para mejorar la precisión de los modelos estadísticos existentes, como los modelos lineales y los modelos de árboles de decisión. XGBoost utiliza un enfoque de boosting iterativo para entrenar varios modelos débiles y luego combinar sus predicciones para obtener una predicción final. Esto permite capturar patrones y relaciones más complejas en los datos que pueden ser difíciles de capturar con un solo modelo.

El proceso de boosting empieza con un modelo inicial simple. El segundo paso es, para cada iteración, añadir un nuevo modelo con tal de intentar corregir los errores del conjunto de modelos anteriores. Así, con cada iteración, el modelo se vuelve cada vez más fuerte (mejor predictor). Finalmente, el último paso consiste en combinar todos los modelos para formar una predicción final.

La utilidad de XGBoost se basa en su capacidad para mejorar el rendimiento de otros modelos estadísticos existentes, especialmente en problemas donde hay relaciones no lineales o interacciones complejas entre las variables. Además, XGBoost proporciona información sobre la importancia relativa de cada variable en la predicción de la esperanza de vida, lo que puede ayudar a identificar las características más relevantes.

XGBoost prueba diferentes modelos con diferentes variables y da para cada uno un valor AIC. En principio, el modelo con menor AIC representa el mejor modelo posible para la interpretación de la variable dependiente.

A continuación, se muestran algunas de las fórmulas matemáticas utilizadas durante un XGBoost. La predicción final $\hat{Y}_i$ para una observación i es la suma de las predicciones de todos los árboles $f_t(x_i)$

$$\hat{y}_i = \sum_{t=1}^{T} f_t(x_i)$$

donde T es el número total de árboles y $f_t(x_i)$ es la predicción del árbol t para la entrada x_i .
En cada iteración t, el objetivo es minimizar la función de pérdida. Esto se hace mediante la adición de un nuevo árbol f_t que prediga los residuos. El gradiente de la función de pérdida respecto a las predicciones anteriores se utiliza para ajustar el nuevo árbol,

$$g_i = \frac{\partial L(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}}$$

donde g_i es el gradiente para la i-ésima observación.

Finalmente, la predicción se actualiza en cada iteración sumando la contribución del nuevo árbol

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(x_i)$$

donde η es la tasa de aprendizaje. Este proceso se repite hasta que se alcance el número deseado de iteraciones o se cumpla algún criterio de parada.

2.7. Modelos ARIMA

Los modelos ARIMA (*Autoregressive Integrated Moving Average*) son una metodología utilizada en análisis de series temporales para modelar y predecir datos que evolucionan con el tiempo. Son especialmente útiles cuando se dispone de una sola variable dependiente que varía en función del tiempo, lo que los hace muy útiles para el análisis y predicción de la variación poblacional. En nuestro estudio, estos modelos son interesantes porque permiten analizar patrones históricos y capturar tendencias y otros comportamientos recurrentes de los datos. Esto hace que los ARIMA sean un integrante ideal para predecir la evolución de la variación de la población española.

Un modelo ARIMA se basa en tres componentes principales:

- Autorregresivo (AR): Modela la relación entre una observación actual y sus valores pasados mediante un proceso autorregresivo de orden *p*,

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t$$

Donde

- y_t es la variable dependiente en el tiempo t,
- $Ø_i$ son los coeficientes autorregresivos,
- ϵ_t es el error aleatorio.

- Integrado (I): Representa el número de diferencias necesarias para que la serie temporal sea estacionaria. La estacionariedad implica que las propiedades estadísticas de la seria (ejemplo: media, varianza) no cambian con el tiempo. La diferencia de orden *d* se expresa:

$$y_t^{(d)} = y_t - y_{t-1}$$

- Media Móvil (MA): Modela la relación entre una observación actual y los errores pasados mediante un proceso de media móvil de orden *q*,

$$y_t = \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q}$$

Donde:

- $Ø_i$ son los coeficientes de media móvil

Por lo tanto, el modelo ARIMA se expresa como ARIMA (*p, d, q*) donde:

- *p*: número de términos autorregresivos
- *d*: número de diferencias para lograr estacionalidad,
- *q*: número de términos de media móvil

En el contexto del análisis poblacional, las predicciones del modelo ARIMA pueden compararse con datos históricos de validación. Las métricas como el RMSE y el MAE permiten evaluar la precisión de las predicciones, mientras que gráficos de predicción permiten visualizar tendencias futuras, como cambios abruptos o estabilizaciones en la población.

Por ejemplo, en estudios donde se analiza la población de diferentes provincias, un ARIMA podría predecir cómo fluctuará la población en una región específica y ayudar a identificar áreas que podrían necesitar políticas específicas de desarrollo o gestión de recursos.

3. TRABAJO DE CAMPO

3.1. Extracción, manipulación y preprocesamiento de datos

3.1.1. Datos y archivos

En primer lugar, toda la extracción de datos ha sido basada en la página oficial del Instituto Nacional de Estadística (INE), una fuente oficial y confiable que proporciona información detallada y actualizada sobre diversas variables demográficas, económicas y sociales en España.
Los archivos originales se obtuvieron en formato Excel desde la plataforma del INE, donde están organizados por indicadores específicos como nacimientos, defunciones, emigraciones, inmigraciones, matrimonios, Índice de Precios al Consumo (IPC), Índice de Precios de Vivienda (IPV), Producto Interior bruto (PIB), viviendas turísticas y tasas de interrupciones voluntarias del embarazo (IVE).
Para entender mejor los datos escogidos, explicamos brevemente los archivos escogidos, que son los siguientes:

- Nacimientos anuales por comunidad autónoma y sexo. Datos desde 1975 hasta 2022, ambos incluidos.
- Defunciones anuales por comunidad autónoma y sexo. Datos desde 1975 hasta 2022, ambos incluidos.
- Matrimonios anuales de diferente sexo por comunidad autónoma. Datos desde 1975 hasta 2022, ambos incluidos.
- Emigraciones con destino al extranjero por año y sexo. Datos desde 2021 hasta 2022, ambos incluidos.
- Inmigraciones procedentes del extranjero por año y sexo. Datos desde 2021 hasta 2022, ambos incluidos.
- Censo poblacional y Censo municipal (número de municipios) por año y comunidad autónoma. Datos desde 2010 hasta 2024.
- Índices de Precios al Consumo (IPC) mensuales, por comunidad autónoma y grupo ECOICOP. Datos desde enero 2002 hasta octubre 2024, ambos incluidos.
- Índices de Precios de Vivienda (IPV) trimestrales por comunidad autónoma. Índices por CCAA: general, vivienda nueva y vivienda de segunda mano. Datos desde el primer trimestre de 2007 hasta el segundo trimestre de 2024, ambos incluidos.
- Producto Interior Bruto (PIB) anuales por comunidad autónoma. Datos desde 2008 hasta 2023.
- Tasas de Paro y Empleo según Genero y año por comunidad autónoma. Datos desde 2002 hasta 2024.
- Porcentajes semestrales de Viviendas turísticas por comunidad autónoma. Datos desde 2020 hasta 2024, ambos incluidos.
- Tasas anuales de Interrupciones Voluntarias del Embarazo (IVE) por mil mujeres entre 15 y 44 años según comunidad autónoma. Datos desde 2014 hasta 2023,

ambos incluidos.
- Cantidad importada y exportada de combustibles por año y comunidad autónoma. Datos desde 2008 hasta 2023.

Para preparar estos datos para el análisis, se desarrolló un proceso en dos etapas. En esta primera etapa, los archivos Excel fueron cargados en R tras una primera reestructuración de los datos. Este caso permitió asignar nombres representativos a cada conjunto de datos y estructurarlos en tablas compatibles con el entorno de análisis estadístico.
La segunda etapa se basa en una limpieza de datos exhaustiva, explicada en parte a continuación, efectuada para una mejor realización del análisis posterior de los datos.

3.1.2. Limpieza de datos

En esta segunda etapa de preparación para el análisis, los datos se sometieron a un proceso de limpieza y transformación de los conjuntos de datos cargados en R anteriormente.
Este proceso incluye la eliminación de valores inconsistentes o faltantes ("NA"), la reestructuración de tablas a fin de obtener un formato uniforme y cómodo para realizar el análisis posterior, y finalmente la separación de los datos en categorías específicas (totales, hombres y mujeres) en el caso de las variables desagregadas por género. Además, se generaron listas organizadas que integran las diferentes dimensiones de los datos, garantizando su utilidad y accesibilidad para el análisis posterior.

Este enfoque sistemático asegura la calidad y la coherencia de los datos utilizados, proporcionando una base sólida para los modelos predictivos y las interpretaciones que se desarrollarán en este estudio.

En primer lugar, los datos extraídos sobre los nacimientos anuales a través del INE (y muchos otros) vienen dados únicamente o según las comunidades autónomas o según las provincias. Por lo tanto, como queremos una tabla que contenga tanto las comunidades autónomas como las provincias, cogemos la tabla de las provincias y le añadimos una tabla nuestra asociando cada provincia a su comunidad autónoma correspondiente.
Es decir, de tener una tabla como la siguiente: (Para reducir el tamaño y explicación de las siguientes tablas, presentamos únicamente los nacimientos de las Comunidades de Cataluña y de la Comunidad Valenciana para el año 2022).

Provincia	Año	Total
Barcelona	2022	41 166
Girona	2022	5 926
Lleida	2022	3 230
Tarragona	2022	6 022

Alacant	2022	13 262
Castelló	2022	4 178
València	2022	18 164

Tabla 2.1.1. Nacimientos anuales por provincia, extraída del INE

Por lo tanto, la tabla final se convierte en la siguiente:

Comunidad_Autonoma	Provincia	Año	Total
Catalunya	Barcelona	2022	41 166
Catalunya	Girona	2022	5 926
Catalunya	Lleida	2022	3 230
Catalunya	Tarragona	2022	6 022
Comunitat Valenciana	Alacant	2022	13 262
Comunitat Valenciana	Castelló	2022	4 178
Comunitat Valenciana	València	2022	18 164

Tabla 2.1.1.bis. Nacimientos anuales por provincia, datos modificados

Estos cambios también se realizan para las tablas conteniendo los datos de defunciones, del mercado laboral, del censo de municipios, inmigraciones, emigraciones, matrimonios, población y viviendas turísticas.

Seguidamente, algunos de los archivos provenientes del INE contenían únicamente los datos por comunidades autónomas, es decir que no se disponía de los datos concretos para las provincias. Para poder conectar toda la información de todas las tablas en una sola y única tabla, se procedió a desglosar algunos de los datos presentados por comunidades autónomas a provincias.
Este problema se presentaba por ejemplo para los datos relacionados con el producto interior bruto (PIB). A continuación, mostramos los datos recogidos en la web del INE sobre el PIB. Para no mostrar la tabla oficial de 60 líneas, decidimos mostrar únicamente 3 comunidades autónomas (Nacional, Andalucía y Aragón) para exclusivamente el año 2023.

Comunidad_Autonoma	Año	PIB a precios de mercado
Nacional	2023	1 498 324 000
Andalucía	2023	199 951 793
Aragón	2023	46 673 641

Tabla 2.1.2. PIB anual por comunidad autónoma, datos extraídos del INE

En este caso concreto, queremos desglosar los datos del PIB de las comunidades autónomas en 2023. Para acercarnos lo máximo a la realidad, debemos hacer un desglosamiento ponderado, es decir, desglosar el PIB total por el porcentaje de población de cada provincia de la esa comunidad autónoma en concreto en 2023.

Tras desglosar por provincias, obtenemos lo siguiente para nuestros datos del PIB. Para no mostrar una tabla de 159 líneas, hemos decidido presentar los mismos datos presentados en la tabla anterior (tabla 2.1.2).

Comunidad_Autonoma	Provincia	Porcentaje_Poblacional	PIB
Nacional	Nacional	1.00000000	1 498 324 000
Andalucía	Almería	0.08776224	17 548 218
Andalucía	Cádiz	0.14618412	29 229 777
Andalucía	Córdoba	0.09028294	18 052 235
Andalucía	Granada	0.10881361	21 757 476
Andalucía	Huelva	0.06221748	12 440 497
Andalucía	Jaén	0.07230037	14 456 589
Andalucía	Málaga	0.20418196	40 826 550
Andalucía	Sevilla	0.22825727	45 640 451
Aragón	Huesca	0.16914923	7 894 810
Aragón	Teruel	0.10068375	4 699 277
Aragón	Zaragoza	0.73016703	34 079 554

Tabla 2.1.2.bis. PIB por comunidad autónoma y provincia, datos modificados

Realizamos la misma idea para los datos del IVE (Interrupción Voluntaria del Embarazo) y del IPV (Índice de Precios de Vivienda) ya que estos datos vienen ambos presentados por comunidades autónomas y no provincias.

De una manera similar a las dos señaladas previamente, algunos de nuestros datos del INE están presentados en periodos trimestrales, es decir 4 trimestres por año, en vez de anuales, como es el caso para nuestros datos sobre la tasa de empleo. Apreciamos a continuación una representación de la tabla comentada, con únicamente la tasa de empleo de las provincias de Bizkaia y Burgos en el año 2021.

Provincia	Periodo	Tasa de empleo
Bizkaia	2021T4	51,34
Bizkaia	2021T3	50,27
Bizkaia	2021T2	48,63
Bizkaia	2021T1	46,89
Burgos	2021T4	52,21
Burgos	2021T3	52,68
Burgos	2021T2	51,07
Burgos	2021T1	50,99

Tabla 2.1.3. Tasa de empleo por trimestre y provincia, datos extraídos del INE

Por lo tanto, para obtener un conjunto de datos correcto debemos convertir el periodo trimestral de la tabla a un periodo anual, realizando una media de los 4 trimestres para obtener el valor anual correspondiente. De este modo, obtenemos una tabla de la siguiente forma:

Provincia	Año	Tasa de empleo
Bizkaia	2021	49,28
Burgos	2021	51,74

Tabla 2.1.3.bis. Tasa de empleo por año y provincia, datos modificados

Se gestiona exactamente igual el mismo problema que aparece para el índice de Precios de Vivienda (IPV), el IPC y las viviendas turísticas.

Por último, nuestro estudio se basa en la intención de predecir la variación de población. Actualmente, no tenemos ninguna variable proveniente del INE que contenga esta información, por lo tanto, tendremos que crearla para poder utilizar modelos que predigan lo que queremos.

La variación anual de la población se puede expresar principalmente de dos maneras:

- Variación absoluta (valores numéricos): se calcula restando el valor del censo del año anterior del valor del censo del año actual

$$\Delta Y_t = Y_t - Y_{t-1}$$

Donde:

- Δ representa la variación o cambio absoluto entre dos valores consecutivos de la serie temporal
- Y es la variable analizada a lo largo del tiempo
- t representa el tiempo asociado a la observación de Y

- Variación porcentual (en términos relativos): Se calcula como la diferencia entre años consecutivos, expresada como un porcentaje del valor del año anterior.

$$m_t = \frac{Y_t - Y_{t-1}}{Y_{t-1}} \cdot 100, \quad t = 2, \ldots, T$$

Donde:

- m_t representa la variación porcentual en el tiempo t
- Y_t es el valor de la seria en el periodo actual t
- T_{t-1} es el valor de la serie en el periodo anterior t-1
- T representa el número total de periodos

Nos parece más interesante obtener los resultados de las predicciones de nuestros modelos en términos relativos, es decir, en porcentajes de aumento o disminución de la población en comparación con el año anterior.
En nuestro caso, existen dos maneras de crear esta variable. El primer método, más concreto y preciso, sería de calcular la variación de población según la siguiente formula:

Variación anual de Población = Crecimiento Natural + Saldo Migratorio

Siendo Crecimiento Natural = Nacimientos – Defunciones,
Saldo Migratorio = Inmigrantes – Emigrantes.

La segunda manera, más rápida y útil por la utilización de una sola variable previa, es realizando directamente la variación anual del Censo poblacional anual.
Sería recomendable utilizar el primer método ya que parece algo más preciso. No obstante, al no tener suficientes datos recogidos sobre las inmigraciones y emigraciones, no podemos obtener valores del saldo migratorio para suficientes años. Por lo tanto, optamos por utilizar el segundo método para calcular los valores de la variación anual poblacional porcentual.

Finalmente, tenemos unos datos relacionados con el censo en número municipal, es decir el número de municipios que albergan una cantidad concreta de habitantes. Estos datos están separados entre 11 modalidades diferentes de municipios ("Menos de 101" […*personas*], "Entre 101 y 500", …, "Más de 500.000") seguida de la separación por provincias y años. En nuestro caso, queremos estudiar más concretamente el aumento de la ruralidad o urbanización, por lo que nos parece más adecuado juntar estas modalidades en 2 únicos tipos diferentes "Zona_rural" y "Zona_urbana".

Ahora bien, sabemos que la ruralidad hace referencia al conjunto de los fenómenos sociales que se desarrollan en un entorno rural. ¿Pero en qué momento un entorno rural se convierte en entorno urbano? Para marcar una línea que separe estas dos zonas es importante entender que no existe una única definición de ruralidad, ya que la noción suele estar en debate, y por lo tanto esta línea separadora entre zonas puede variar considerablemente en diferentes estudios y definiciones.
En nuestro caso, basándonos en las definiciones del ministerio de medio ambiente y medio rural y marino[1] tenemos las siguientes definiciones:

- Rural: aglomeración de <2.000 habitantes
- Semi rural: <=10.000 habitantes
- Urbano >10.000 habitantes

Por lo tanto, tenemos dos opciones viables; utilizar un limitador de zonas con valor de 2.000 habitantes o con valor de 10.000 habitantes.
Debido a que este límite entre zonas se deja bastante a la decisión del investigador, decidimos realizar los modelos con los dos límites diferentes para visualizar si hay diferencias.

3.2. Descriptiva de la base de datos

En este apartado se realiza una descripción exhaustiva de los datos de cada tabla, siguiendo un enfoque estructurado que combina análisis estadístico y visualización. Esto permitirá comprender tanto la distribución de los datos como sus patrones principales.

Como hemos explicado al inicio de este trabajo, el tema de estudio que tocamos es muy amplio debido a la innumerable cantidad de variables y datos que pueden provocar un

cambio significativo en la variación de la población a lo largo de un tiempo específico.

Está claro que lo primero que nos imaginamos a la hora de pensar en cómo evolucionara la población es en sí aumentan los nacimientos o no. Sin realizar ningún análisis ni contemplar otros trabajos, es lógico pensar que un crecimiento de los nacimientos provoca una variación positiva anual de la población, es decir un aumento de la población.

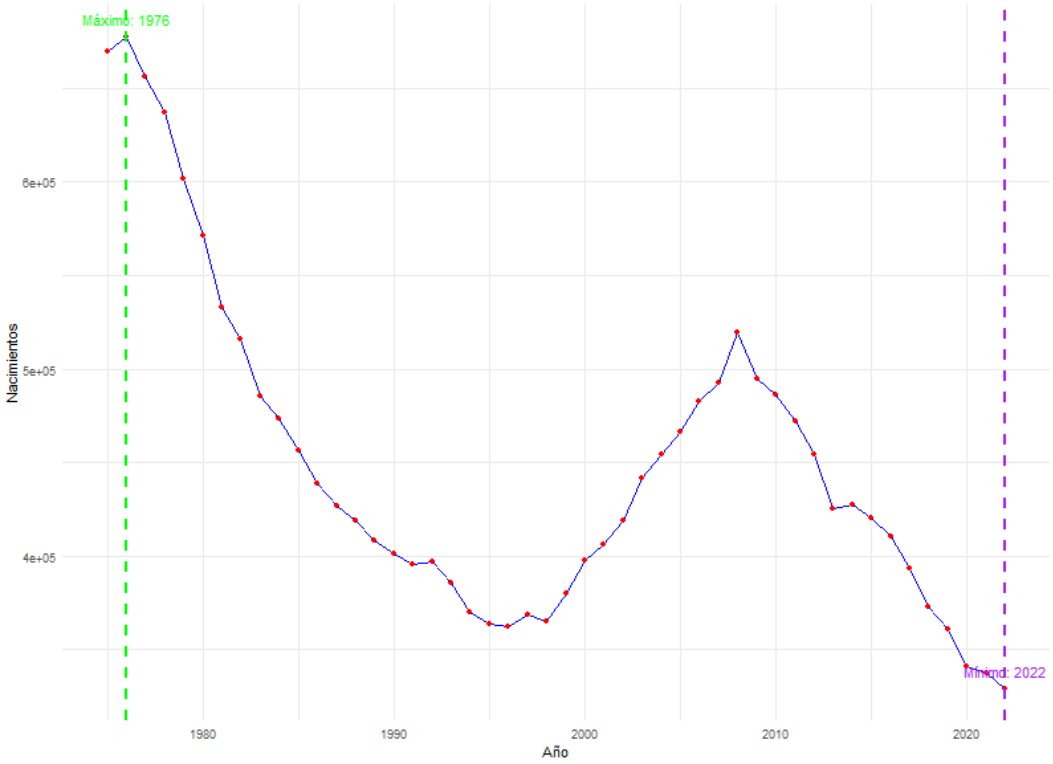


Figura 3.2.1. Tendencia de nacimientos nacionales a lo largo del tiempo

Como vemos en la figura 3.2.1, la tendencia de nacimientos anuales en España ha sido decreciente desde el inicio de los datos en los años 70. Únicamente se vio contrarrestada por una subida durante los años 2000 generada por el optimismo social gracias a la mejora de la situación económica del país. Ya que, seguidamente, debido posiblemente en gran parte a la crisis de 2008, los nacimientos volvieron a decaer considerablemente hasta llegar a su punto más bajo alcanzado hoy en día con alrededor de 500 nacimientos en 2023 en la provincia de Soria.

En 2010, los nacimientos por 1000 habitantes alcanzaban valores cercanos a 10, es decir casi un 1% de nacidos sobre la población total mientras que, en los datos más recientes, estas cifras han caído considerablemente hasta los 6 nacidos por 1000 habitantes. Este patrón se refleja claramente en la figura 3.2.2 que muestra una tendencia descendente continua.

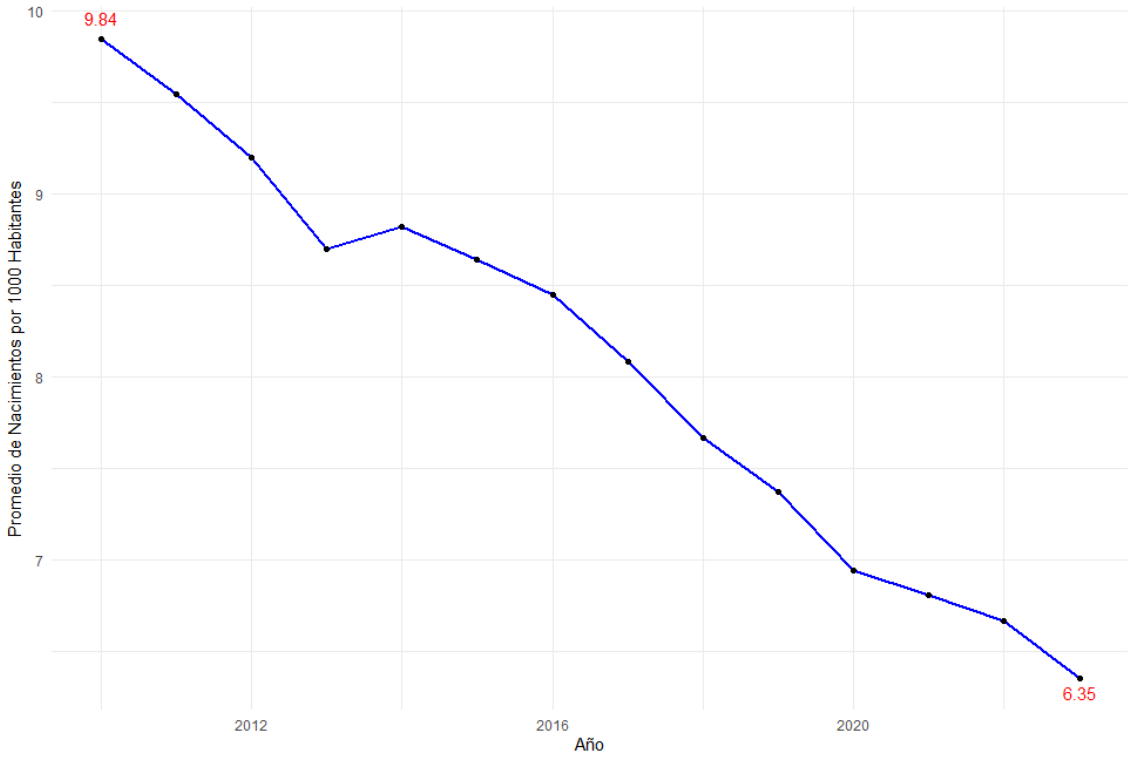


Figura 3.2.2: Evolución temporal del promedio de nacimientos por 1000 habitantes en España

La disminución en los nacimientos está asociada a varios factores, como el retraso en la edad de maternidad, la incertidumbre económica y cambios culturales que priorizan otras metas personales antes que formar una familia. Además, esta dinámica varía según la región. La figura 3.2.3 mostrada a continuación evidencia diferencias significativas entre comunidades, donde provincias como Lugo o Zamora presentan tasas más altas en comparación con otras como Almería. Estas disparidades pueden estar influenciadas por factores como el acceso a recursos, políticas familiares y dinámicas culturales regionales.

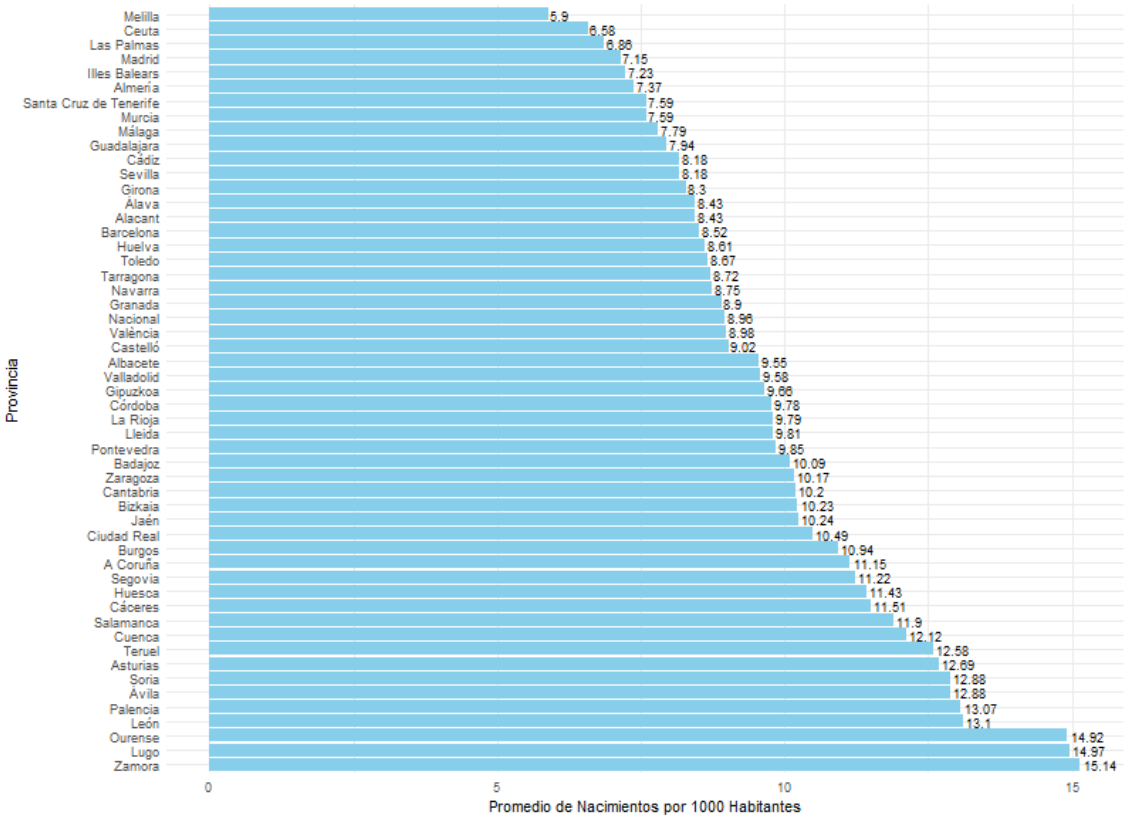


Figura 3.2.3. Promedio de Nacimientos por 1000 habitantes por provincia

¿Aun así, significa por consiguiente que la población decrece también? ¿Y si lo hace, decrece de la misma manera?
Debemos tener en cuenta que al igual que los nacimientos son muy importantes para el calculo de una población, las defunciones también tienen su punto a decir ya que son un factor crucial en el balance demográfico.

A medida que las tasas de natalidad disminuyen, las defunciones tienden a aumentar, especialmente en un país con una de las esperanzas de vida más altas del mundo. España enfrenta un envejecimiento poblacional significativo, particularmente en regiones como Galicia y Castilla y León, donde una gran parte de la población supera los 65 años.

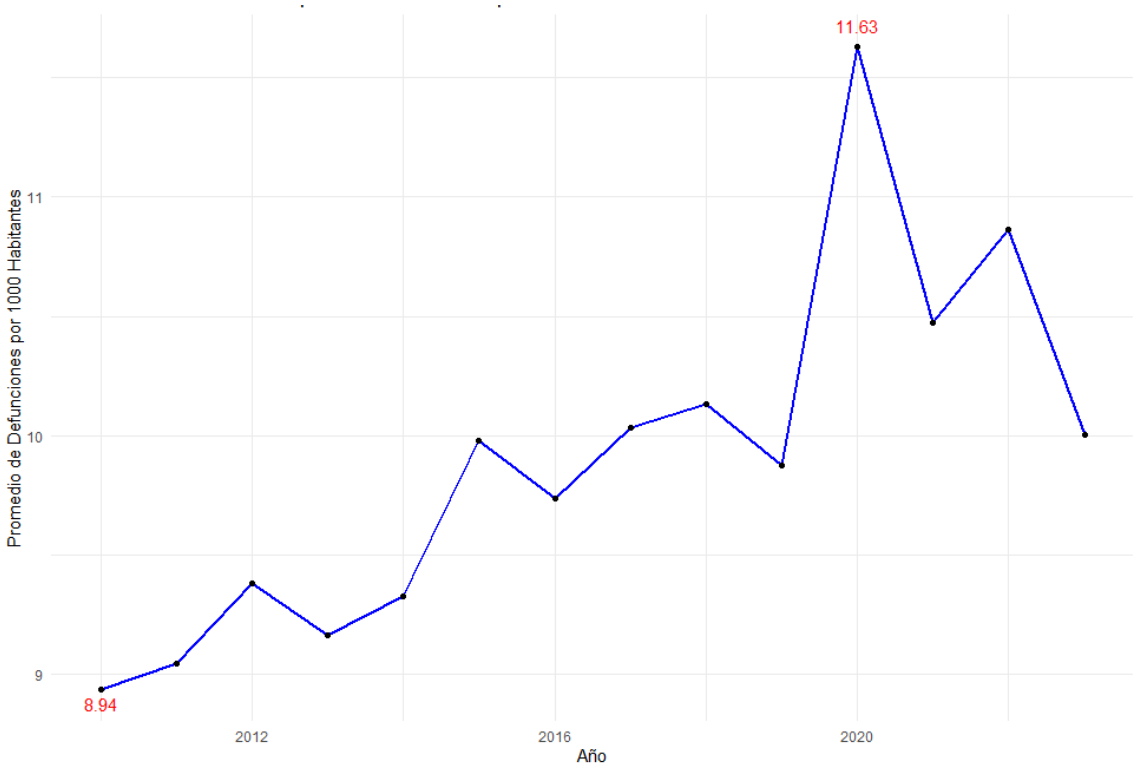


Figura 3.2.4. Promedio de defunciones por 1000 habitantes por año

Vemos en la figura 3.2.4 que al contrario de lo observado con los nacimientos previamente, las defunciones han aumentado factualmente desde 2010 hasta 2023. Es muy probable que el pico elevado en 2020 sea debido a la pandemia acontecida durante ese año a nivel mundial. Aunque el envejecimiento es un signo de desarrollo y mejores condiciones de vida, también plantea retos en términos de recursos y atención sanitaria.

Un factor clave en el crecimiento natural, valor de las defunciones restado al de los nacimientos, es que los nacimientos anuales superen a las defunciones anuales, ya que se considera que el crecimiento natural es uno de los factores e indicadores más importantes de la variación de población.
Ya que tenemos datos desde 1975 hasta 2023 miremos si en algún momento el crecimiento natural (o saldo natural) pasa a ser negativo.

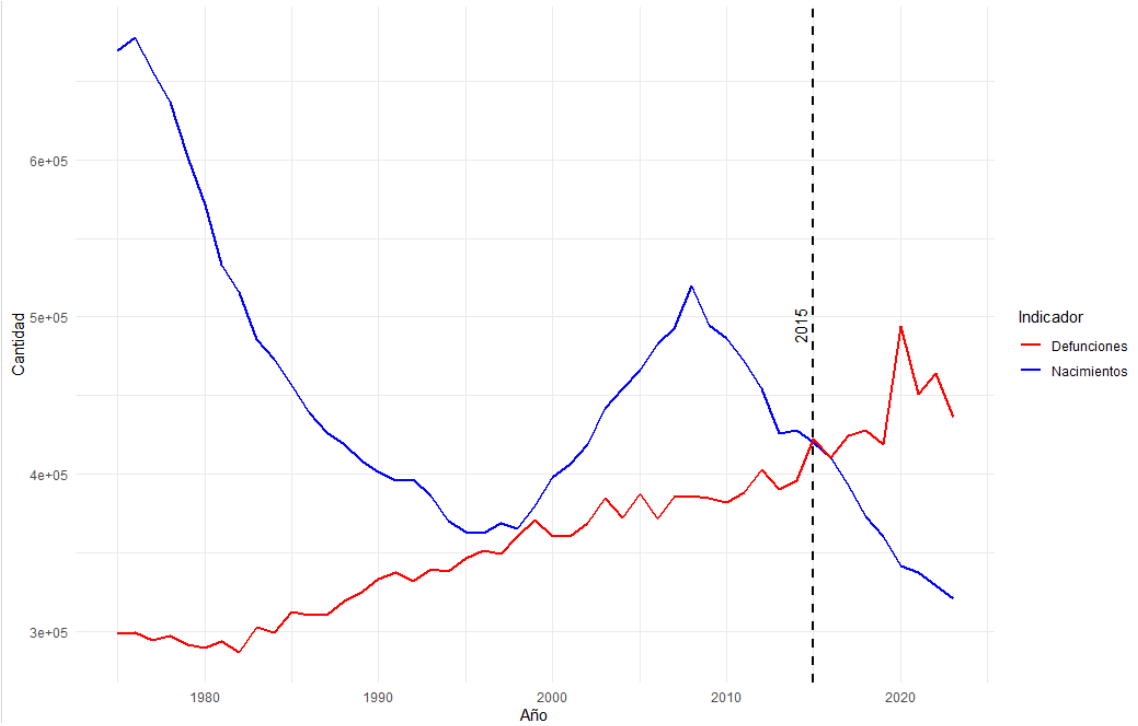


Figura 3.2.5. Tendencia de nacimientos y defunciones nacionales anuales

La figura 3.2.5 nos indica que en 2015 se produce por primera vez un alcance de los fallecimientos a los nacimientos y que desde entonces el crecimiento natural ha sido negativo, alcanzando su nivel más bajo en 2020. Que el saldo natural sea ahora negativo en muchas provincias significa que, sin migración, estas regiones pierden población cada año. Esto refleja un desequilibrio demográfico que puede amplificarse si no se toman medidas entre otras como fomentar la natalidad o atraer inmigrantes.

Es evidente que las nuevas costumbres y la evolución social también son en parte causas de la nueva situación respecto al saldo natural. Años atrás, era muy común y aceptado por la sociedad casarse a muy temprana edad y por consiguiente tener hijos. Hoy en día, el matrimonio también ha experimentado transformaciones significativas. Los datos muestran en la figura 3.3.6 una disminución considerable en el número de matrimonios nacionales por año.

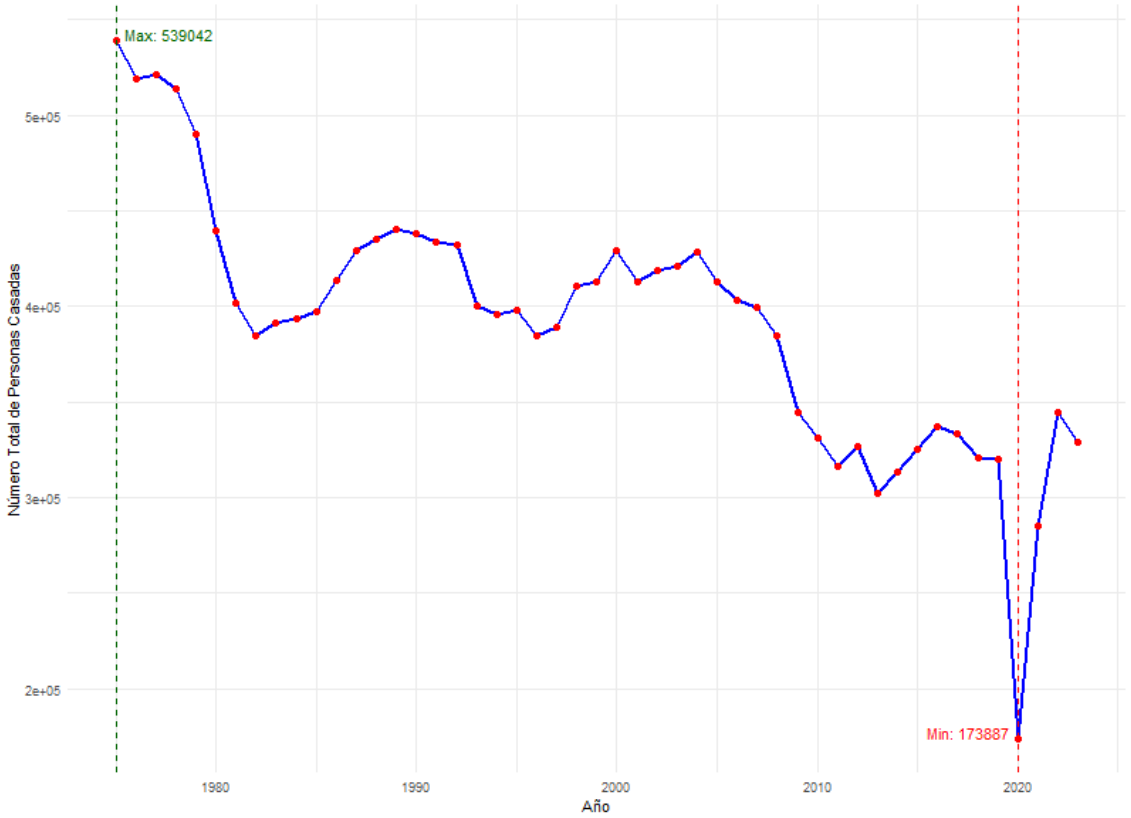


Figura 3.2.6. Evolución temporal nacional del número de personas casadas anualmente

Estos valores, aunque reflejan una sociedad aun en transición, parecen asentarse desde la década de los 2010. Las parejas de hoy en día optan más por convivencias informales o por retrasar el matrimonio debido a factores económicos y sociales. Además, el concepto de la maternidad joven ha cambiado, sobre todo en los países desarrollados como España. La edad promedio para contraer matrimonio ha aumentado, lo que también afecta la edad a la que las parejas deciden tener hijos. Este fenómeno se evidencia en la figura 3.2.7 siguiente:

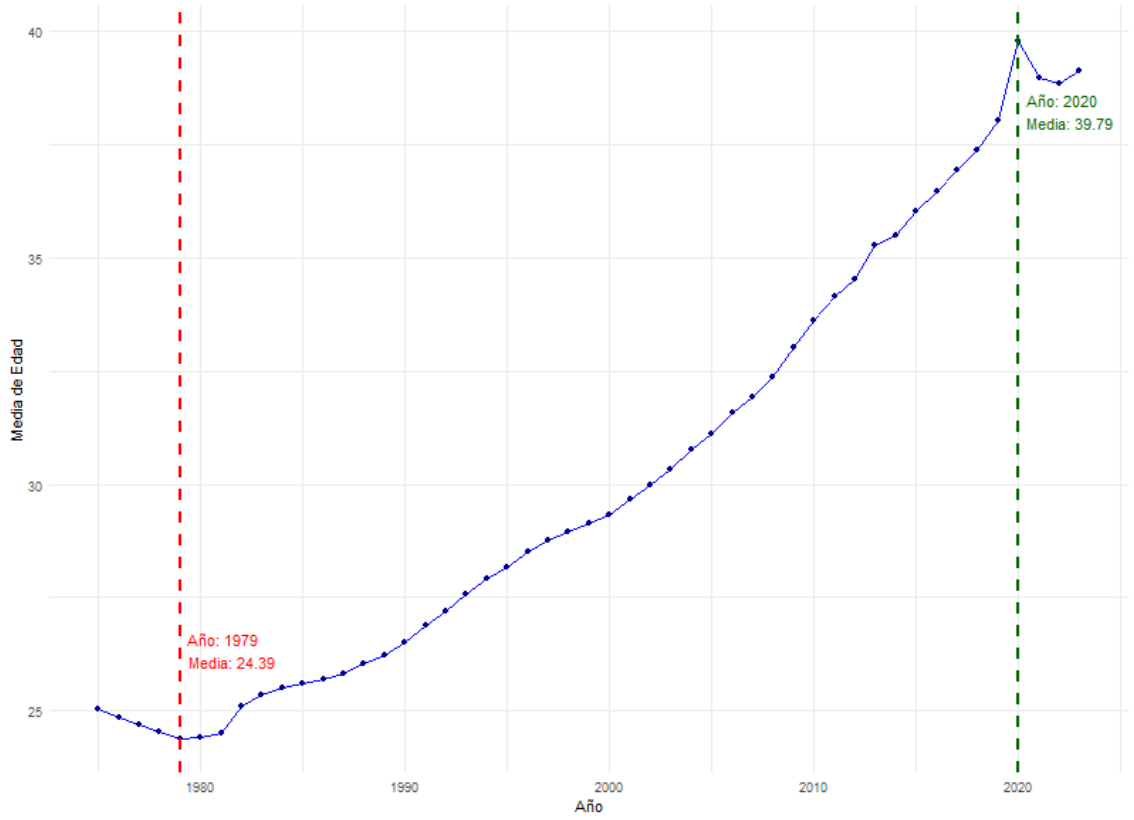


Figura 3.2.7. Evolución de la media de edad anual de personas casadas

Como bien hemos comentado, una de las grandes razones por la que parecen haber sucedido estos cambios a nivel nacional es la economía. Esta es un motor clave que afecta a diario la dinámica poblacional y por ende cada una de las decisiones de los individuos del país. El PIB, por ejemplo, no solo mide la capacidad productiva de una región, sino que también afecta las decisiones de migración interna y externa.

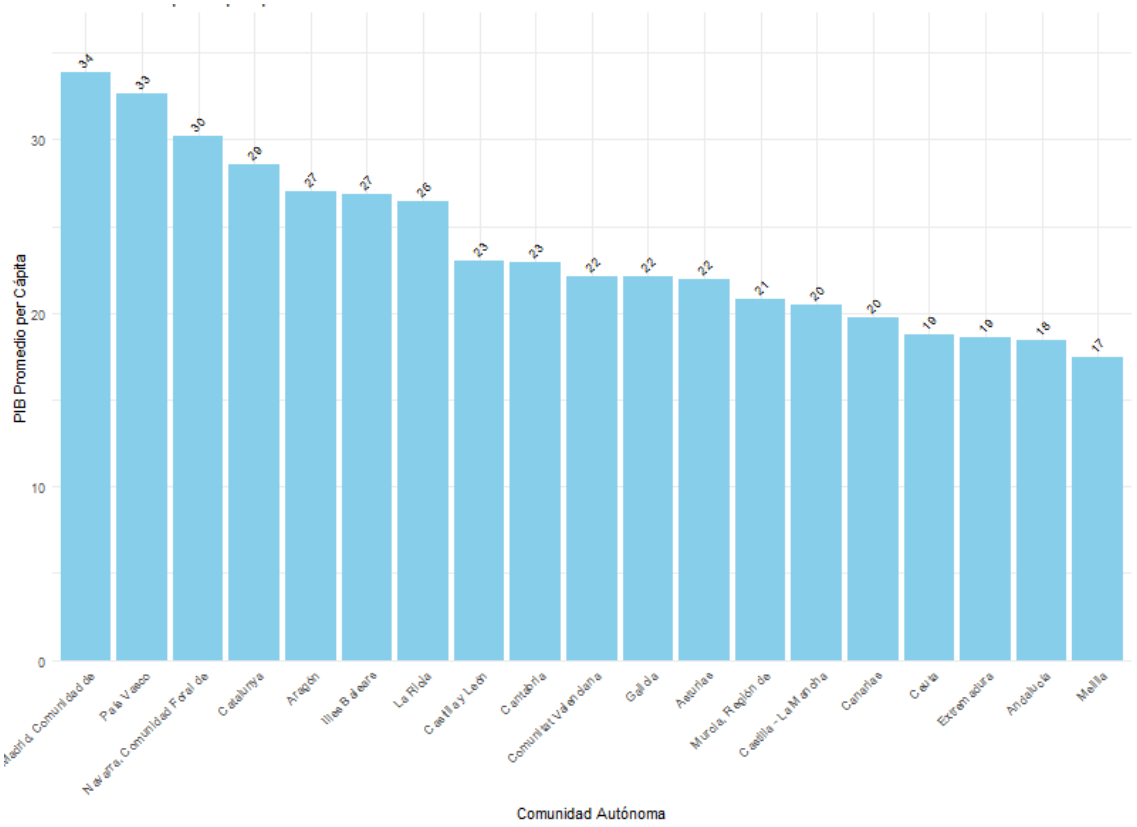


Figura 3.2.8. PIB promedio per cápita por comunidad autónoma

La figura 3.2.8 revela disparidades significativas entre regiones. Comunidades como Madrid y País Vasco tienen un PIB per cápita mucho más alto que regiones como Extremadura o Andalucía. Estas diferencias económicas atraen población hacia áreas más prósperas, mientras que las regiones menos desarrolladas enfrentan despoblación.

Además, el costo de la vivienda, reflejado en el Índice de Precios de Vivienda (IPV), tiene un impacto significativo en la movilidad y en las decisiones de residencia.
La figura 3.2.9 muestra cómo las comunidades con mayores precios y variabilidad, como Madrid y las Islas Baleares, enfrentan desafíos para retener a los residentes jóvenes. Esto es debido a los jóvenes, generalmente con ingresos más bajos y una menor estabilidad económica, tienden a tenerlo más complicado en estas regiones. Siguiendo una idea similar, una menor accesibilidad de los jóvenes a una vivienda tiende a provocar un aumento de la migración, sea tanto dentro de España como en otros países, en busca de lugares más económicos. Otra consecuencia, que afecta directamente al caso de estudio, es que las dificultades para acceder a la vivienda afectan o retrasan a la decisión de los jóvenes de formar una familia debido a la inseguridad económica, lo que contribuye al descenso de la tasa de natalidad.

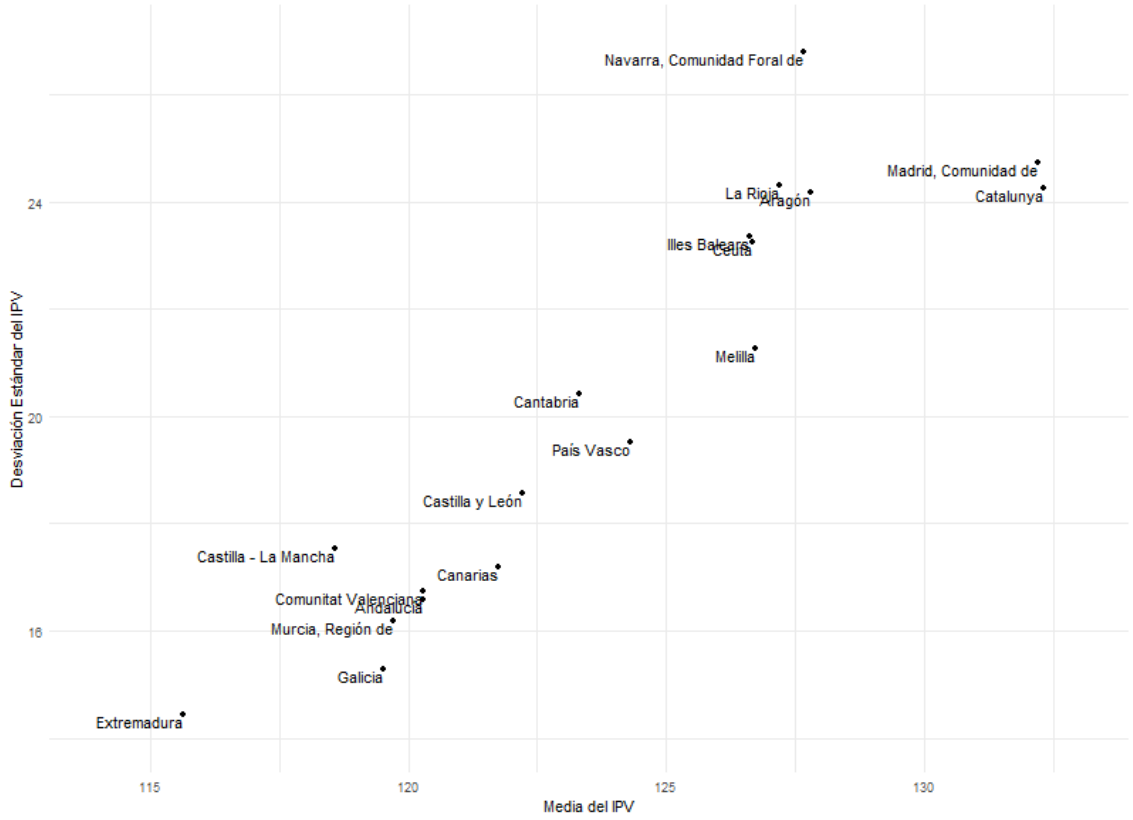


Figura 3.2.9. Relación entre la media y la variabilidad del IPV por comunidad autónoma

La distribución de la población en el territorio español también está cambiando. El número de municipios rurales ha disminuido, mientras que las áreas urbanas continúan expandiéndose. Desde principios del siglo XX, en gran parte de Europa y por consiguiente en España, la idea de una vida en el campo empieza a desaparecer gradualmente. Según Delgado (2019), a nivel regional, en la última década se ha producido un notable descenso en la población rural por dos causas fundamentales: saldos naturales negativos y migraciones interiores hacia las urbes de la misma región.
En la segunda década del siglo XXI, especialmente a partir del año 2008, con el inicio de la crisis económica global, comenzaron a notarse de manera gradual los problemas financieros y laborales en la población. La crisis financiera tuvo un impacto significativo en la forma en que las personas vivían y trabajaban, lo que llevó a muchas a mudarse hacia ciudades medianas o grandes capitales de provincia. Este movimiento hacia áreas urbanas se debía principalmente a la búsqueda de mejores oportunidades laborales y económicas, ya que en los pueblos y zonas rurales las posibilidades se redujeron considerablemente durante este periodo de incertidumbre.

No obstante, a través de nuestros datos recopilados desde 2010 obtenemos las siguientes figuras. En estos gráficos, el delimitador entre zona rural y zona urbana es de 2000 habitantes.

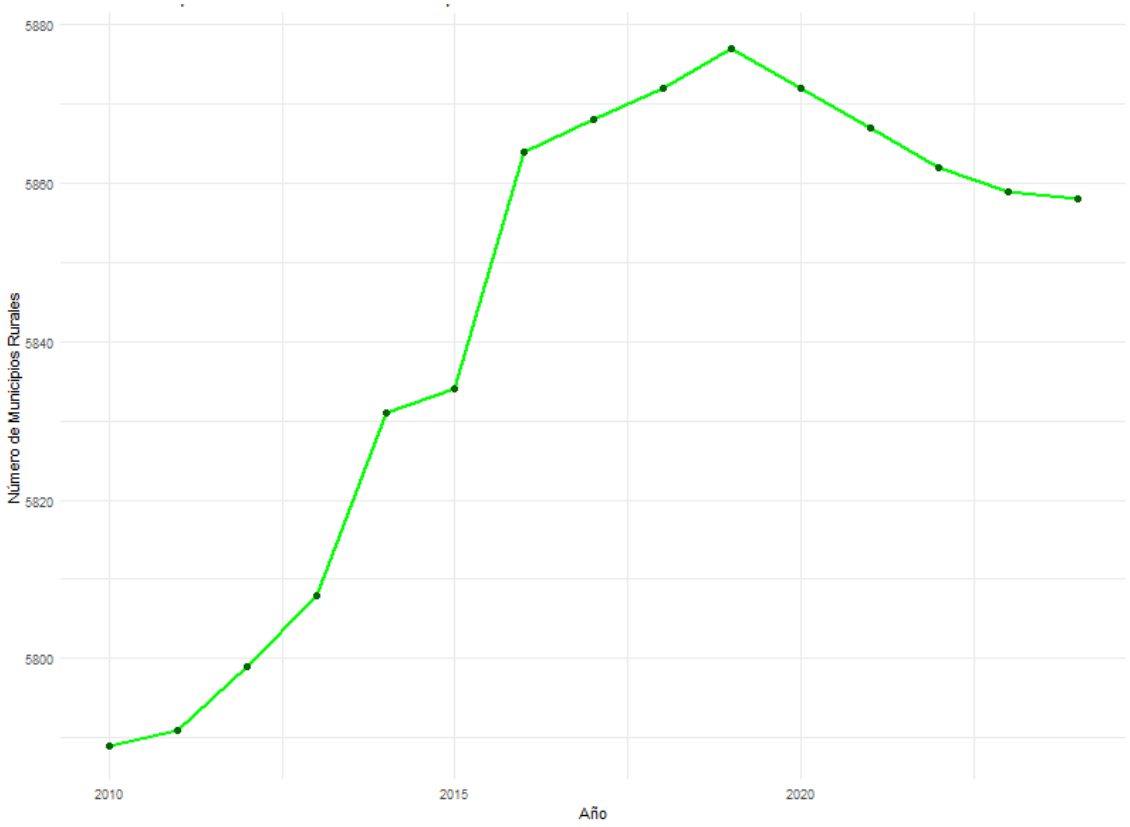


Figura 3.2.10. Evolución temporal del número total de municipios rurales

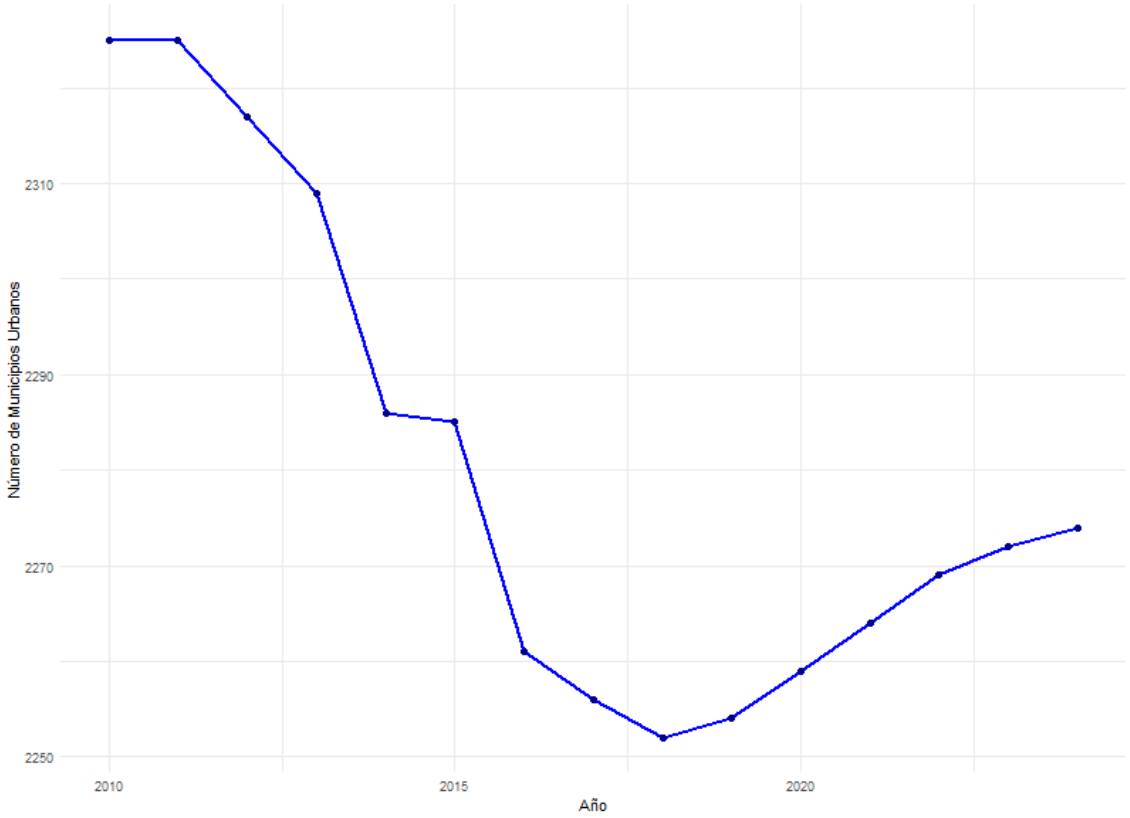


Figura 3.2.11. Evolución temporal del número total de municipios urbanos

Analizando estas figuras conjuntamente podemos decir que se observa una dinámica clara y relacionada de transición entre lo rural y lo urbano. En esta última década, los municipios rurales han crecido en número. Mientras tanto, los municipios urbanos, inversamente a los rurales, han comenzado a caer en números. Este caso se debe posiblemente a la despoblación y envejecimiento poblacional comentado anteriormente.
Por otro lado, a partir de 2019 aproximadamente, estas trayectorias se han invertido y no parecen estabilizarse hoy en día. Según datos del Eurostat, en 2022 el 26% de la población europea residía en zonas rurales, un porcentaje que se reduce hasta el 13% en España, lo que la sitúa en el tercer puesto por la cola, solo por debajo de Países Bajos (11%) y de Malta (3%).

Estos cambios en el número de municipios rurales y urbanos también se reflejan en la distribución de la población por provincia, como muestran las siguientes figuras.

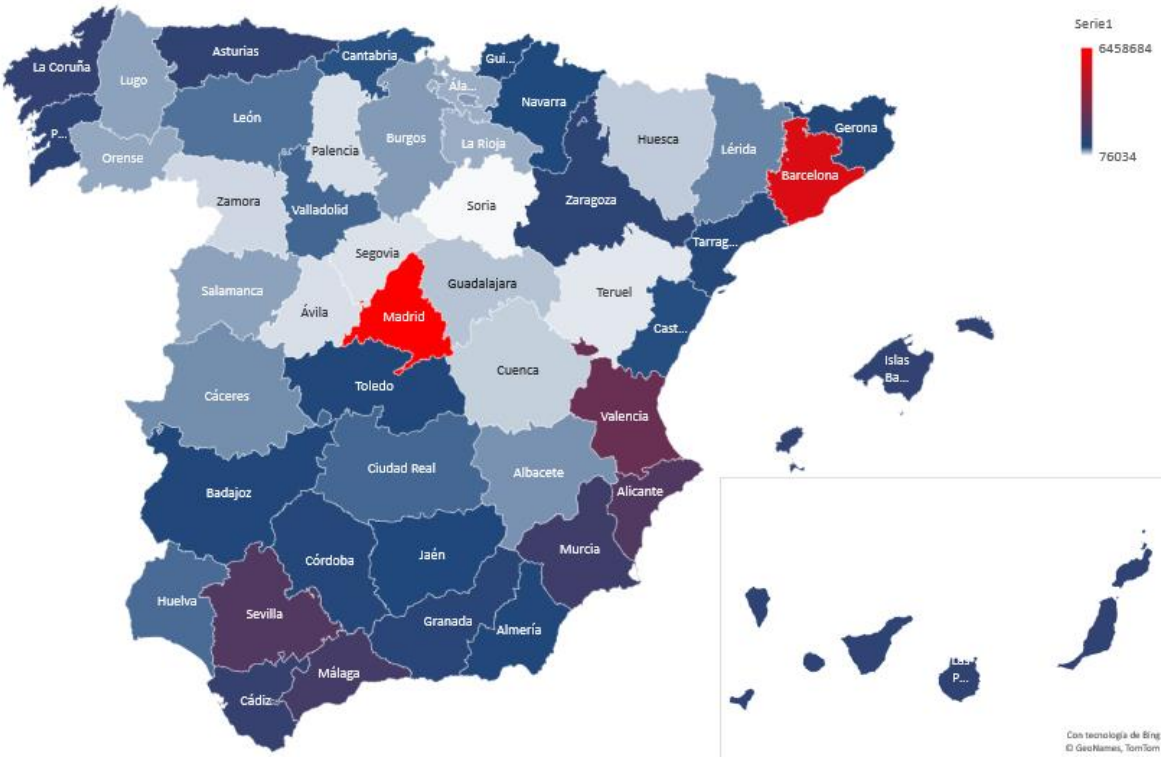


Figura 3.2.12. Distribución de la población española por provincia (2010)

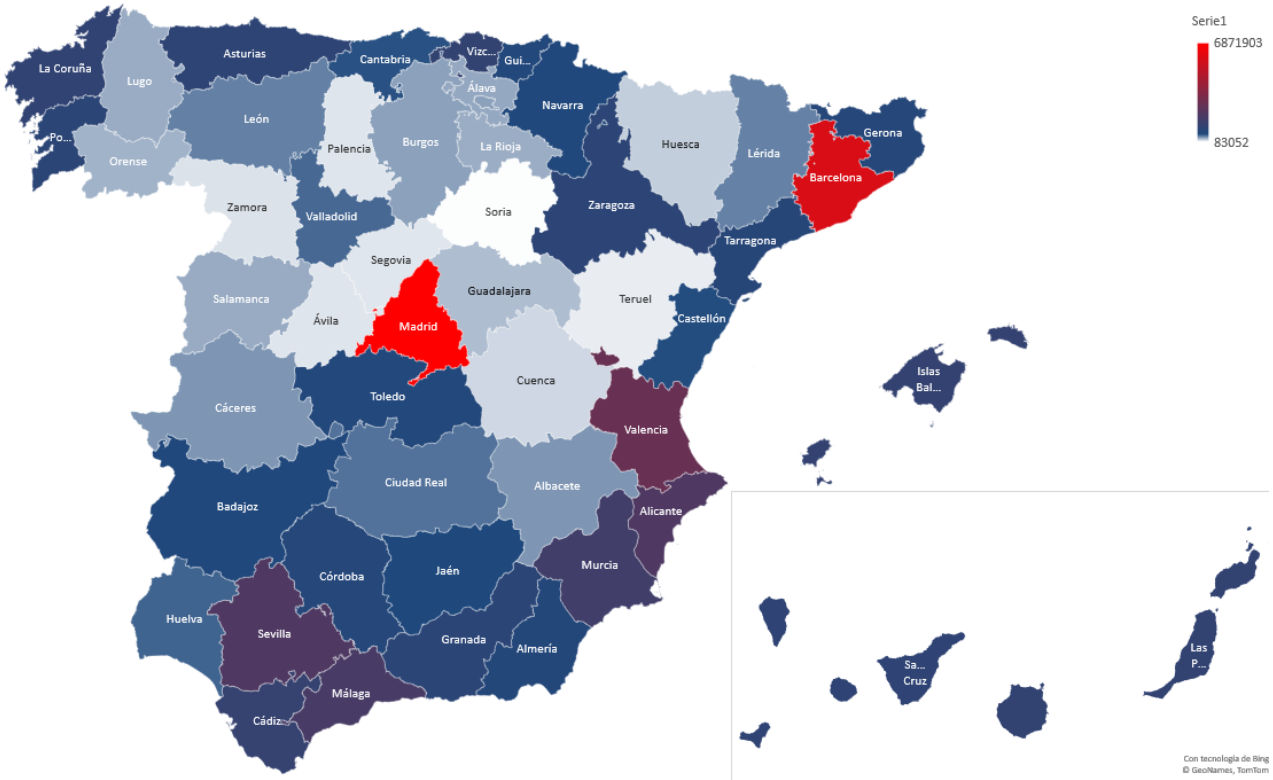


Figura 3.2.13. Distribución de la población española por provincia (2023)

A primera vista, se observa que las provincias con mayor densidad de población, como Madrid y Barcelona, siguen destacando en ambos años, con tonos rojos más intensos en el mapa. Esto refleja su rol como principales centros económicos, culturales y laborales del país, que continúan atrayendo población. Sin embargo, provincias rurales o menos densamente pobladas, como Soria, Teruel y Cuenca, mantienen sus niveles más bajos, evidenciando la persistencia de la despoblación en estas áreas.

En términos de cambios, aunque los patrones generales de concentración poblacional se mantienen, se observa un ligero aumento de población en provincias como Madrid, lo cual refuerza la tendencia hacia la urbanización y el crecimiento en grandes ciudades. En contraste, algunas provincias rurales parecen haber perdido habitantes, reflejando el impacto de factores como el envejecimiento poblacional, la migración interna hacia zonas urbanas y la baja natalidad en estas áreas.

Finalmente, el empleo es quizás uno de los factores más determinantes para la variación de la población. Los cambios en la tasa de empleo y de paro impactan directamente en las decisiones de residencia y ampliación familiar.

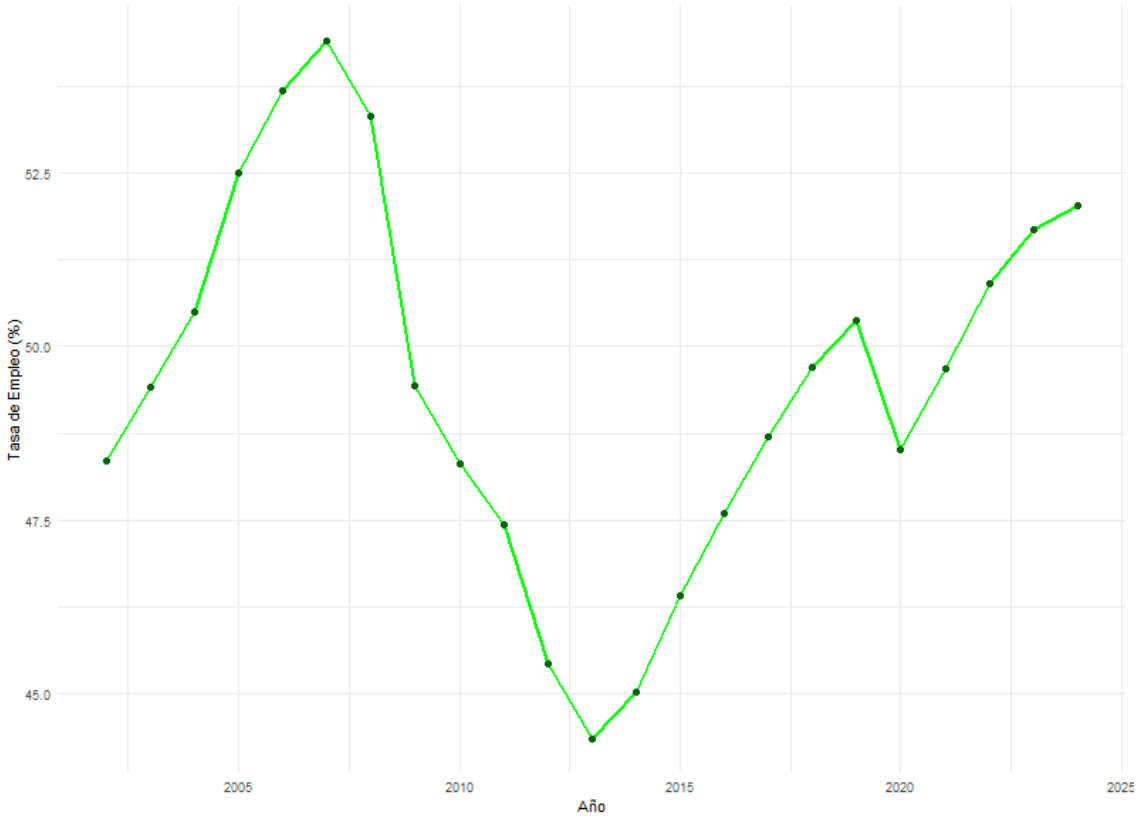


Figura 3.2.14. Evolución de la tasa de empleo a lo largo del tiempo

La figura 3.2.14, mostrando la evolución de la tasa de empleo a lo largo del tiempo, muestra un claro pico antes de la crisis financiera de 2008, cuando la tasa tocó su punto mas alto alrededor del 54%. Sin embargo, a medida que la crisis golpeaba la economía global, la tasa de empleo presenció una caída drástica, alcanzando su nivel mas bajo alrededor del 44.5% en 2013. Desde entonces, se ha observado una recuperación sostenida de la tasa de empleo, coincidiendo con la recuperación económica del país y por lo tanto de la estabilización del mercado laboral, únicamente alteradas en 2020 por la pandemia mundial.
Por otro lado, la tasa de paro sigue una evolución inversa a la del empleo como se puede prever. Cuando la tasa de empleo desciende, la tasa de para aumenta, y viceversa.
Este análisis no solo refleja cómo el mercado laboral responde a los ciclos económicos, sino también cómo las tasas de empleo y paro impactan directamente en la calidad de vida de la población y en su comportamiento demográfico. Por ejemplo, niveles altos de desempleo pueden retrasar decisiones como el matrimonio o la formación de familias, afectando indirectamente la tasa de natalidad y la dinámica poblacional.

Finalmente, tras realizar una limpieza exhaustiva de los datos iniciales y un análisis amplio y variado de estos, consideramos que los datos finales que estarán presentes en el estudio de los modelos predictivos posteriores son los siguientes:

Nombre de variable	Descripción
Comunidad_Autonoma	Nombre de la comunidad autónoma. Variable categórica: 20 categorías de las cuales 19 comunidades autónomas de España y una llamado “Nacional” representando el conjunto de todas las comunidades.
Provincia	Nombre de la provincia. Variable categórica: 53 categorías, de las cuales 52 provincias de España y una llamado “Nacional” representando el conjunto de las provincias.
Año	Año en el que se registra la observación. Variable numérica discreta representada como un entero.
Variacion_Poblacion_Porcentual	Porcentaje de cambio de población respecto al año anterior. Variable numérica continua y variable respuesta buscada en nuestros modelos. Variable respuesta.
Nacimientos_por_1000_habitantes	Número de nacimientos por cada mil habitantes. Variable numérica continua.
Defunciones_por_1000_habitantes	Número de defunciones por cada mil habitantes. Variable numérica continua.
Matrimonios_por_1000_habitantes	Número de matrimonios registrados en la provincia por cada mil habitantes
Censo_Total	Población total censada en la provincia durante el año. Variable numérica discreta.
Indice_General_Vivienda	Índice midiendo la evolución anual de los precios de las viviendas en las provincias. Variable numérica continua.
Tasa_paro_poblacion	Porcentaje de la población en edad laboral que está desempleada. Variable numérica continua.
Tasa_empleo_poblacion	Porcentaje de la población en edad laboral que esta empleada. Variable numérica continua.
PIB_per_capita	Producto Interior Bruto anual per cápita provincial a precios de mercado. Variable numérica continua.
Zona_rural	Número de municipios clasificados como rurales en la provincia. Variable numérica discreta.
Zona_urbana	Número de municipios clasificados como urbanos en la provincia. Variable numérica discreta.

Importaciones_combustibles_per_capita	Cantidad (en toneladas) de combustibles importados por la provincia durante el año por persona. Variable numérica continua.
Exportaciones_combustibles_per_capita	Cantidad (en toneladas) de combustibles exportados por la provincia durante el año por persona. Variable numérica continua.

Tabla 3.2.15. Tabla de las variables finales de nuestra base de datos y sus respectivas descripciones

3.3. Estudio de modelos predictivos

En este apartado del trabajo presentamos diferentes modelos predictivos, estudiamos su utilidad para nuestros datos y nos quedamos, es decir utilizamos, aquellos que son más apropiados en nuestro caso.
Tras ello, compararemos las predicciones de los modelos utilizados entre ellos y tomaremos una decisión sobre la importancia y diferencia de cada uno, y sobre todo buscaremos elegir el mejor de los modelos para la predicción de nuestros datos.

Para determinar la utilidad de cada modelo utilizaremos métricas de evaluación que sirven comúnmente en modelos de regresión para medir qué tan bien un modelo está prediciendo la variable dependiente.
Para evaluar el rendimiento de los modelos, se utilizarán las tres métricas de regresión siguientes:

- **RMSE** (Root Mean Squared Error), que indica la diferencia, en promedio, entre los valores reales y las predicciones del modelo. Esta métrica, muy similar a MAE, es más sensible a los errores grandes por lo que penaliza más los errores significativos. Cuanto más cercano a 0 sea el RMSE mejor rendimiento del modelo.

- **MAE** (Mean Absolute Error), que cuantifica la diferencia entre las predicciones del modelo y los valores reales, sin tener en cuenta la dirección de la diferencia (es decir, si la predicción es mayor o menor que el valor real). Esta métrica, similar a RMSE, es más equilibrada y menos influenciada por los valores atípicos. Al igual que le RMSE, un valor más bajo de MAE indica un mejor rendimiento del modelo.

- **R^2** (Coeficiente de Determinación): es un estadístico que proporciona una medida indicando que tan bien las predicciones del modelo de regresión se ajustan a los datos reales. Su valor varía entre 0 y 1, donde un valor más cercano a 1 indica un mejor ajuste del modelo.

Muchos modelos son posibles a primera vista. Vamos a estudiar en primer lugar la linealidad de nuestros datos.

Nombre de variable	Correlación
Indice_General_Vivienda	0,39
Tasa_empleo_poblacion	0,30
Año	0,09
Tasa_paro_poblacion	-0,08
PIB	0,08
Censo_Total	0,07
Nacimientos_Total	0,06
Defunciones_Total	0,06
Zona_urbana	0,05
Total_Casados	0,04

Tabla 3.3.1. Tabla de las correlaciones de las variables independientes con la variable respuesta

La tabla muestra las correlaciones entre las variables explicativas y la variable respuesta. Viendo que no hay linealidad entre muchas de nuestras variables, no deberíamos utilizar modelos que asumen estrictamente linealidad entre las variables predictoras y la variable respuesta. Por lo tanto, modelos como modelo lineal (LM) y modelo lineal generalizado (GLM) son descartables.
De manera rápida y para más concretización, si intentamos realizar un modelo ML obtenemos las siguientes métricas:

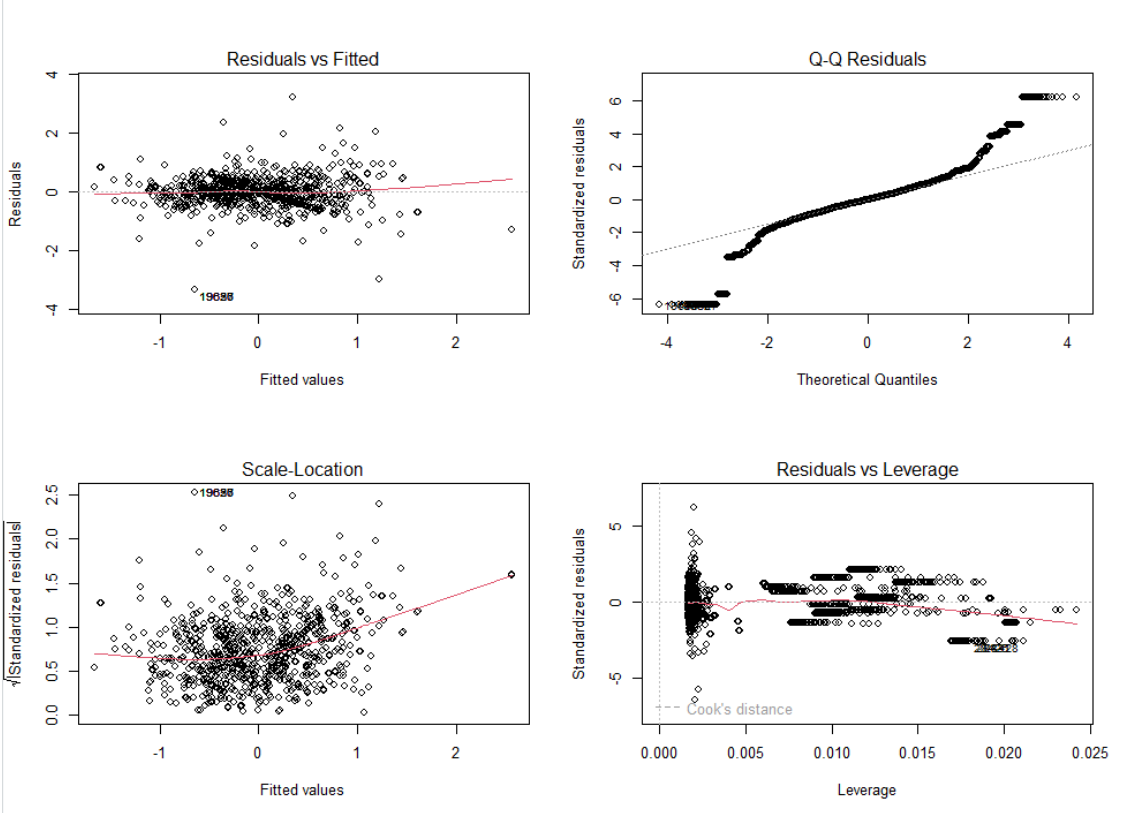


Figura 3.3.2. Métricas obtenidas para el modelo lineal

Estas nos indican que hay heterocedasticidad, es decir, varianza no constante, observado gracias al grafico de '*residuals vs fitted*' y además no se observa normalidad en los residuos lo que es clave en los datos si queremos realizar un modelo lineal.

3.3.1. RANDOM FOREST

Procedemos a realizar un análisis y modelo predictivo utilizando el modelo Random Forest en nuestra base de datos final obtenida previamente.

El modelo Random Forest ha sido implementado para analizar y predecir la variación porcentual anual de la población utilizando datos de entrenamiento y prueba.
Los denominados datos de entrenamiento, que suelen representar en estos modelos un 80% de los datos, están compuestos por todos nuestros datos desde el año 2010 hasta el año 2020, ambos incluidos. El segundo grupo, denominado datos de prueba, que representan generalmente el otro 20% de los datos, se compone de los datos restantes, es decir de los datos desde 2021 hasta 2023, ambos incluidos.

Este modelo es especialmente útil para detectar relaciones complejas y no lineales entre las variables predictoras y la variable objetivo. Además, permite evaluar la importancia relativa de las diferentes variables incluidas en el modelo. A continuación, se presentan los resultados obtenidos en distintas configuraciones, evaluando el impacto de incluir o excluir una variable dummy para el año 2022, un período excepcional marcado por las alteraciones causadas por la pandemia del COVID-19, y explorando diferentes límites para la clasificación entre zonas rurales y urbanas. Aunque tengamos la impresión de que el año 2022 debería ser extraído de los datos del modelo debido a que es un gran outlier en muchas de las variables, realizamos los modelos con este año incluido para estar seguros de si es mejor añadir esta variable dummy sobre el año 2022.
Además, en este estudio hemos decidido comparar los resultados predictivos de los modelos realizados según si marcábamos nuestro limitador entre zona rural y zona urbana en 2000 habitantes o bien en 10000 habitantes.

Métricas	Separación zonas = 2.000	Separación zonas = 10.000
RMSE	0.632	0.635
MAE	0.480	0.481
R^2	0.334	0.329

Tabla 3.3.3. Tabla comparando las métricas obtenidas por el modelo Random Forest con los datos finales sin modificar

Limitador de zonas	2.000		10.000	
Años de datos de prueba	2	3	2	3
RMSE	0.612	0.720	0.612	0.725

MAE	0.468	0.578	0.476	0.597
R^2	0.364	0.117	0.355	0.088

Tabla 3.3.3.bis Tabla comparando las métricas obtenidas por el modelo Random Forest con la variable dummy_2022 incluida de los datos

Las dos tablas 3.3.3 y 3.3.3.bis presentadas permiten comparar los resultados obtenidos por el modelo Random Forest en diferentes configuraciones, analizando cómo varían las métricas de rendimiento (RMSE, MAE y R^2) dependiendo de si se incluye o no la variable *dummy_2022* y utilizando dos límites diferentes para la separación entre zonas rurales y urbanas (2,000 y 10,000 habitantes).

Una de las conclusiones más claras es que el limitador de zonas no tiene un impacto significativo en las métricas del modelo. Al comparar los resultados entre los límites de 2,000 y 10,000 habitantes, las diferencias en los valores de RMSE y MAE son mínimas. Por ejemplo, con 2 años de datos de prueba, los valores de RMSE son prácticamente idénticos en ambas configuraciones (0.612 en ambos casos cuando se incluye la variable *dummy_2022*).
De manera similar, los valores de R^2 también son muy cercanos, lo que indica que, independientemente del limitador, el modelo es capaz de capturar patrones similares en los datos. Esto sugiere que el impacto de este cambio en el modelo es relativamente bajo y que otros factores, y que la inclusión de una variable (*dummy_2022*) que sirve para captar la presencia de un año atípico en el conjunto de datos, tiene mayor relevancia en el desempeño del modelo.

Con relación a esto último, otra observación importante es que el modelo mejora ligeramente cuando se incluye esta variable relacionando la atipicidad del año 2022 en los datos. Esto puede observarse en los valores de R^2, que representan la capacidad explicativa del modelo. Con los datos finales sin modificar, es decir sin añadir esta variable dummy, el mejor valor de R^2 se encuentra alrededor de 0.33 (por ejemplo, 0.334 para el limitador de 2,000 habitantes y 3 años de prueba).
Sin embargo, al incluir la variable *dummy_2022*, el mejor valor de R^2 aumenta a aproximadamente 0.36 (0.364 con 2 años de prueba y 2,000 habitantes). Este aumento, aunque pequeño, sugiere que la inclusión de una variable que identifique la atipicidad del año 2022 permite al modelo capturar tendencias más consistentes en los datos. Esto es lógico y se debe probablemente a que 2022, año muy atípico debido a la pandemia de COVID-19, introduce variabilidad que dificulta la capacidad del modelo para ajustarse a los datos.

En resumen y como acabaremos de concluir más adelante, estos resultados indican que el modelo de Random Forest es moderadamente eficaz para predecir la variación anual poblacional. El limitador de zonas rurales y urbanas no tiene un impacto considerable en el desempeño del modelo, mientras que la inclusión de *dummy_2022* parece mejorar ligeramente la precisión de las predicciones.

3.3.2. XGBOOST

Anteriormente hemos realizado un modelado predictivo con el modelo Random Forest. Ahora, realizaremos exactamente lo mismo utilizando el modelo XGBoost para nuestra base de datos.

Cada uno de los modelos ejecutados se guarda en archivos RDS para posteriormente evitarnos, cada vez que entremos al script de R, una pérdida de tiempo elevada a la hora de ejecutar de nuevo los modelos.

En primer lugar, en nuestros datos tenemos 14 variables numéricas, de las cuales una (el año) está representada como un entero, y 2 variables categóricas, que son la variable de la comunidad autónoma y la variable de la provincia.
El modelo XGBoost es ejecutable únicamente con una base de datos plenamente numérica, por lo que, para poder utilizar estas dos últimas variables en el modelo, debemos convertirlas en variables numéricas. Para ello, convertimos nuestras variables categóricas en dummies y eliminamos las variables de base.
Ahora sí, sin haber perdido información y teniendo una base de datos final completamente numérica, podemos realizar el modelo XGBoost.

Por otro lado, siguiendo la misma idea que el modelo anterior, obtenemos nuestra base de datos final con dummies separadas entre unos datos de entrenamiento y unos datos de prueba.
Se convierten el set de entrenamiento y el set de prueba en matrices DMatrix requeridas por XGBoost. Una DMatrix es una estructura de datos de la biblioteca XGBoost utilizada para almacenar grandes conjuntos de datos y procesar eficientemente los datos durante el entrenamiento y la predicción de los modelos.
A continuación, mostramos un comando de R con el modelo que se ha realizado para predecir mediante un XGBoost:

```
modelo_xgb <- xgboost( data = dtrain, max_depth = 6, eta = 0.1,
 nrounds = 1000, objective = "reg:squarederror"  )
```

- objective: Tipo de clasificación (en este caso “reg:squarederror” para regresión)
- nrounds: Número de iteraciones (10)
- max.depth: Profundidad máxima de los árboles (2)
- eta: Tasa de aprendizaje (0.3)

En segundo lugar, generamos las predicciones utilizando el modelo entrenado (`modelo_xgb`) sobre el conjunto de datos de prueba. Las predicciones generadas son los valores estimados por el modelo para la variable objetivo de la variación de la población (*Variación_Poblacion_Porcentual*) en el conjunto de datos de prueba.

A continuación, evaluamos el rendimiento del modelo utilizando las tres métricas de regresión explicadas anteriormente.
Se presentan los resultados obtenidos en distintas configuraciones, evaluando el

impacto de incluir o no la variable *dummy_2022* en los datos y de elegir un limitador de zonas de 2,000 habitantes o de 10,000 habitantes, tal y como hemos hecho también para el modelo Random Forest anteriormente.

Métricas	Separación zonas = 2.000	Separación zonas = 10.000
RMSE	0.651	0.653
MAE	0.489	0.491
R^2	0.294	0.338

Tabla 3.3.4. Comparación de las métricas obtenidas en el modelo XGBoost con los datos finales sin modificar

Limitador de zonas	2.000		10.000	
Años de datos de prueba	2	3	2	3
RMSE	0.657	0.695	0.917	0.643
MAE	0.508	0.553	0.746	0.489
R^2	0.349	0.217	-0.260	0.379

Tabla 3.3.4.bis Comparación de las métricas obtenidas en el modelo XGBoost con la variable dummy_2022 incluida en los datos

En primer lugar, al analizar los resultados con los datos finales sin modificar (Tabla 3.3.4), se observa que el rendimiento del modelo mejora ligeramente al utilizar un limitador de 10,000 habitantes en comparación con 2,000 habitantes. Por ejemplo, el valor de R^2 es más alto con el limitador de 10,000 (0.338 frente a 0.294). Además, las métricas de error, como el RMSE (0.653 frente a 0.651) y el MAE (0.491 frente a 0.489), son apenas superiores, pero las diferencias son marginales. Esto sugiere que, al igual que en el modelo de Random Forest, el limitador de zonas no tiene un impacto significativo en el rendimiento del modelo cuando se incluyen todos los datos.

Sin embargo, al incluir la variable mostrando la atipicidad el año 2022, los resultados se vuelven más interesantes (Tabla 3.3.4.bis). Aquí, se destaca una clara mejora en el modelo con el limitador de 10,000 habitantes y 2 años de prueba, donde el R^2 alcanza su valor más alto (0.379), lo que indica que el modelo es más explicativo en esta configuración. Este valor indica que el modelo XGBoost explica aproximadamente el 38% de la variabilidad en los datos. Este comportamiento está alineado con las observaciones realizadas para el modelo de Random Forest, donde la inclusión de esta variable *dummy_2022* permite al modelo ajustarse mejor a los datos históricos, posiblemente entendiendo mejor el ruido introducido por los efectos atípicos de ese año.

Por lo tanto, tanto para el modelo Random Forest como el modelo XGBoost, los datos finales a utilizar son aquellos incluyendo la variable *dummy_2022* en los datos, sin importar si hemos escogido un limitador de zona rural/urbana de 2,000 habitantes o de 10,000 habitantes.

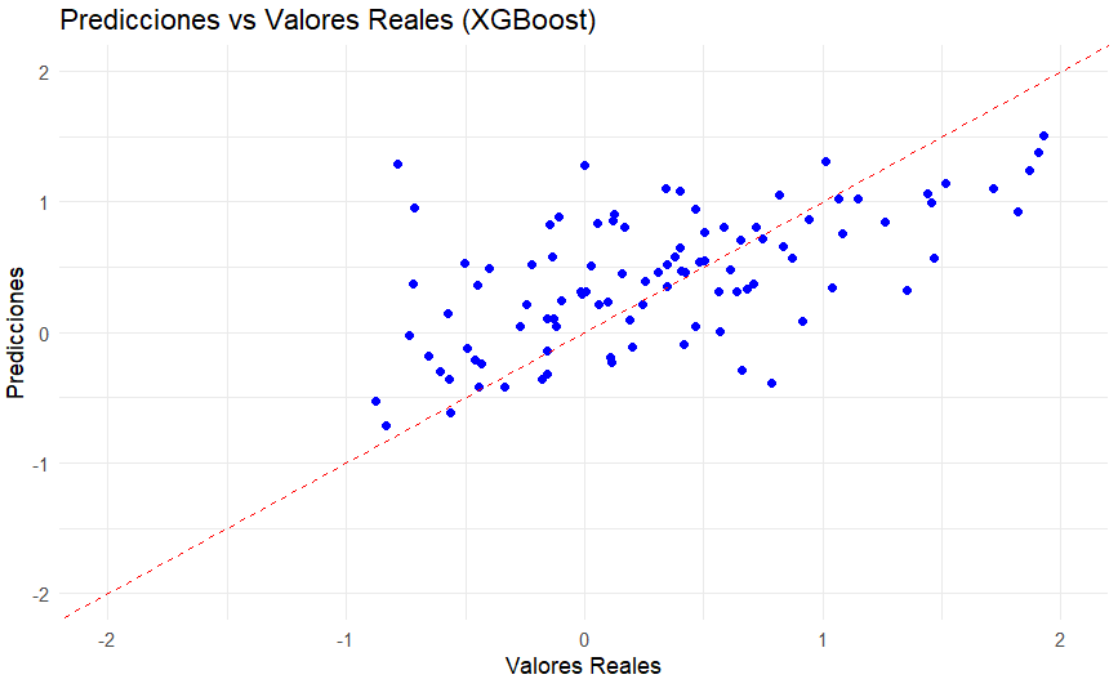


Figura 3.3.5. Gráfico de las predicciones vs valores reales del modelo XGBoost

En la figura 3.3.5 se capta la idea de que, aunque el modelo XGBoost no sea perfecto para la predicción de nuestros datos, sigue capturando patrones generales.

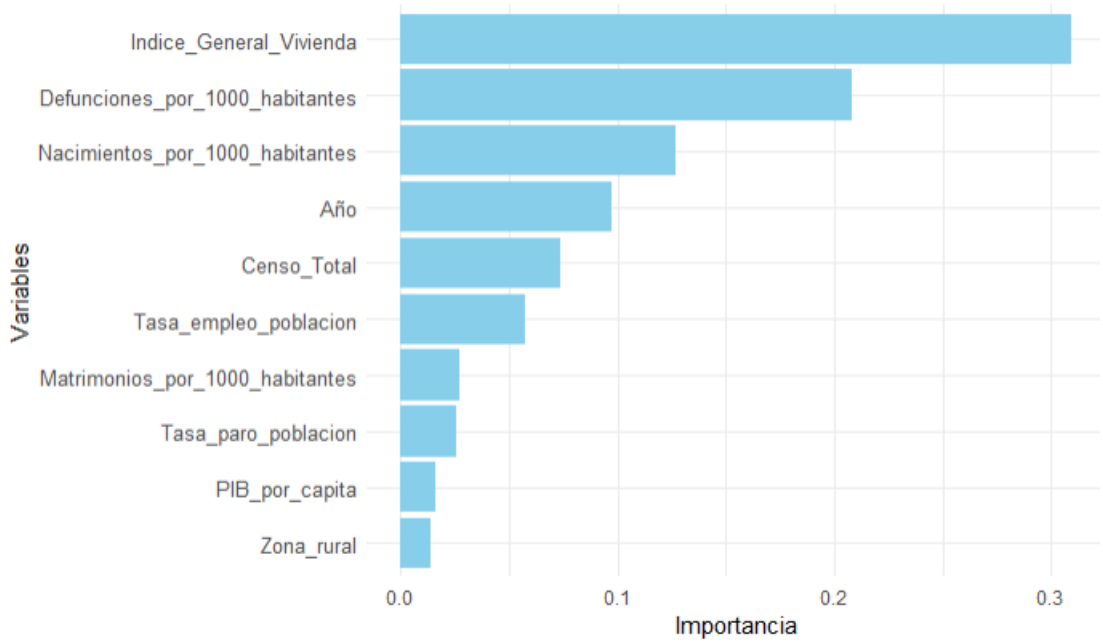


Tabla 3.3.6. Top 10 variables más importantes en XGBoost

Finalmente, este modelo estadístico estudiado nos revela nuestras diez variables más relevantes e importantes en la predicción de la variación poblacional española.

3.3.3. ARIMA

En último lugar, tras realizar dos modelos de predicciones a corto plazo, buscamos estudiar si es posible un modelo que prediga de manera valorable a un futuro más lejano que un año, es decir a más de un año vista. Por esta razón, se decide probar a realizar un modelo ARIMA.

Sabemos que para poder realizar un modelo ARIMA, la serie temporal debe ser estacionaria, es decir que sus medias, varianzas y autocorrelaciones permanezcan constantes en el tiempo. Para comprobar esta estacionalidad en nuestros datos se realiza una prueba de Dickey-Fuller en donde se rechaza la hipótesis de estacionalidad en caso de obtener un p-value>0.05, tal y como es nuestro caso, ya que obtenemos un 0,34.
Después, identificamos los parámetros del modelo con la función "auto.arima". Esta función detecta un modelo ARIMA(0,0,0) que sugiere que no se han detectado patrones temporales significativos en la serie diferenciada.
Al encontrarnos con este problema, la única solución viable a la que recurrimos es la de forzar componentes AR o MA para intentar captar relaciones más complejas. Por lo tanto, haciendo esto decidimos incorporar al modelo un término autorregresivo (AR) y un término de media móvil (MA) para un ARIMA(1,0,1).

RMSE	MAE
0.790	0.678

Tabla 3.3.7. Métricas obtenidas en el modelo ARIMA

Aunque estos valores son razonables, el RMSE no es particularmente bajo, y el modelo todavía muestra limitaciones para capturar toda la complejidad de los datos.

Por lo tanto, por lo que hemos estudiado, en el contexto de la variación poblacional, el hecho de obtener un modelo ARIMA (0,0,0) sugiere que los patrones históricos en la serie no son lo suficientemente significativos como para predecir con precisión cambios a futuro lejano.

4. CONCLUSIONES

4.3. Conclusiones del trabajo

Este estudio nos revela que los modelos basados en árboles, particularmente XGBoost y Random Forest, ofrecen un desempeño robusto al predecir la variación anual de la población. Cabe decir que generalmente estos modelos suelen darse como buenos cuando más de un 80% de la variabilidad de la variable dependiente puede explicarse mediante las variables predictoras del modelo, es decir que sus métricas de predicción R^2 superen el valor 0,8. En nuestro caso, para los modelos Random Forest y XGBoost obtenemos unos R^2 de aproximadamente 0,35 y 0,38 respectivamente. Aunque parezca a primera vista que estos resultados son insuficientes, otros estudios de características similares nos dicen todo lo contrario. Para ser más concretos, un estudio de BMC Medical Research Methodology (2010) sobre un modelado de regresión adaptativa (RandomForest) de biomarcadores de daño potencial en una población de adultos de EE. UU. fumadores y no fumadores de cigarrillos nos indica que sus resultados de R^2, obtenidos entre 0,29 y 0,41, son valores muy consistentes e interesantes de los que poder extraer conclusiones científicamente correctas. Esta idea se expresa también en el estudio de Huang, Tsai, Wu, Lien, Yi Chien, Kuo, Hung, Chen y Kuo (2020) donde se obtiene para el modelo XGBoost que un 41% de la variabilidad de la variable dependiente se explica gracias a las variables predictoras.

Por lo tanto, los resultados obtenidos para dos de nuestros modelos estudiados, Random Forest y XGBoost, son estadísticamente correctos y útiles. Por otro lado, XGBoost demuestra ser ligeramente mejor para la predicción de nuestros datos, ya que

con un coeficiente de determinación (R^2) de hasta 0.379 en las mejores configuraciones, es el modelo que mejor indica una mayor capacidad para capturar las relaciones complejas entre las variables. Por su parte, Random Forest también muestra resultados competitivos y cercanos a XGboost, alcanzando un R^2 cercano a 0.35 en configuraciones óptimas. Estas métricas, junto con valores aceptables de RMSE y MAE, sugieren que ambos modelos son herramientas efectivas para analizar los patrones históricos de variación poblacional, siendo XGBoost el modelo que ligeramente mejor se adapta a las particularidades de los datos.

En contraste, el modelo ARIMA, diseñado específicamente para series temporales y predicciones a largo plazo, presenta limitaciones significativas en este contexto. Aunque la serie diferenciada permitió realizar el ajuste necesario para garantizar la estacionariedad, el modelo identificado como ARIMA(1,0,1) no logró capturar patrones relevantes en los datos, lo que se reflejó en valores de RMSE y MAE más altos en comparación con los modelos basados en árboles. Este resultado muestra lo difícil que es predecir cambios en la población cuando los datos no siguen patrones claros o consistentes a lo largo del tiempo, especialmente cuando intentamos hacer estimaciones a largo plazo.

Uno de los hallazgos más relevantes de este análisis es la influencia marginal del limitador de zonas rurales y urbanas (2,000 frente a 10,000 habitantes) en el rendimiento de los modelos. Las métricas obtenidas para ambos delimitadores son prácticamente equivalentes, lo que sugiere que este limitador no es un factor determinante en la capacidad predictiva de los modelos.
Por otro lado, la inclusión de la variable *dummy_2022* hace entender mejor la atipicidad del año 2022 y resultó en una mejora consistente en el desempeño de los modelos. Esto destaca la importancia de no considerar directamente eventos tan extraordinarios al entrenar modelos predictivos, ya que estos pueden introducir ruido y sesgos en los resultados.

Además de evaluar el desempeño de los modelos, este estudio ha identificado las variables más relevantes para explicar las dinámicas poblacionales. Factores económicos como el Índice de Precios de Vivienda (IPV) y las tasas de empleo y desempleo emergen como determinantes clave. Las regiones con mayores precios de vivienda y variabilidad en estos tienden a experimentar mayores dificultades para atraer y retener residentes jóvenes, lo que impacta negativamente en la natalidad y, por ende, en la variación poblacional. Asimismo, las tasas de empleo y paro se relacionan directamente con la estabilidad económica de una región, lo que influye en las decisiones de residencia, formación de familias y movilidad geográfica. Las migraciones, tanto dentro del país como desde el extranjero, son clave porque ayudan a equilibrar el descenso natural de la población en áreas donde la población está envejeciendo. Por otro lado, como era de imaginar, el crecimiento natural es un gran determinante para la variación de la población, como lo es también, en una medida más pequeña, el número de matrimonios por cada mil habitantes.

Finalmente, este análisis destaca los desafíos y oportunidades que enfrenta España en términos de población. Factores como el envejecimiento de la población, el aumento de

la migración hacia las ciudades y las diferencias económicas entre regiones seguirán influyendo en la dinámica demográfica en los próximos años. Los modelos como XGBoost, al ofrecer predicciones más acertadas, pueden ser herramientas útiles para diseñar políticas públicas y estrategias que ayuden a mantener el equilibrio poblacional. Según los resultados obtenidos, este último modelo se presenta como la mejor opción para predecir la variación poblacional en España. Además, las variables relacionadas con la economía, el empleo y las migraciones deben ser el centro de atención para comprender mejor y responder a las tendencias demográficas del país.

4.4. Lineas futuras

Como sucede en cualquier investigación, han surgido líneas futuras que podrían complementar y enriquecer este estudio. Estas áreas de trabajo adicional al nuestro no solo permitirían un análisis más completo, sino que también ofrecerían herramientas más robustas para la toma de decisiones gubernamentales, en caso de utilización de estos modelos.

En primer lugar, sería fundamental incluir en los modelos una mayor cantidad de variables que puedan tener un impacto directo o indirecto en la variación poblacional española. Por ejemplo, el saldo migratorio, tanto interno como externo, es decir las inmigraciones y las emigraciones, representa un factor clave que podría proporcionar una visión más detallada de las dinámicas demográficas. Asimismo, variables como la calidad de los servicios públicos y los cambios en los patrones educativos y culturales podrían ayudar a entender mejor las dinámicas de la población. Incorporar estas y muchas más variables permitiría a los modelos capturar con mayor precisión la complejidad del contexto socioeconómico y ofrecer predicciones más útiles para la planificación gubernamental.

Otra línea interesante para investigar sería analizar cómo las crisis económicas, sanitarias o climáticas afectan directamente las tendencias demográficas. Por ejemplo, estudiar más a fondo los efectos a largo plazo de eventos como la pandemia de COVID-19 o los desplazamientos causados por el cambio climático podría ayudar a anticipar escenarios extremos y a desarrollar políticas que hagan frente a estos desafíos. Este tipo de análisis podría combinar enfoques económicos, sociales y ambientales para obtener una visión más completa.

Por último, sería interesante llevar a cabo un análisis más detallado que no solo compare comunidades autónomas o provincias, sino que también examine dinámicas locales a nivel municipal. Esto ayudaría a detectar patrones más específicos, como zonas que están perdiendo población lentamente o áreas con un crecimiento inesperado. Además, incorporar variables como la cercanía a ciudades grandes y pobladas o la densidad de infraestructuras podría mejorar las conclusiones y aportar datos más útiles para planificar el desarrollo territorial.

Resumiendo, estas líneas futuras muestran la importancia de seguir indagando en esta

investigación para entender y afrontar mejor los futuros desafíos demográficos que enfrenta España.

5. BIBLIOGRAFIA

Referencias bibliográficas de los datos extraídos del Instituto Nacional de Estadística:

- Índice de Precios de Consumo (IPC). Base 2021.
  https://www.ine.es/jaxiT3/Tabla.htm?t=50913
- Índice de Precios de Vivienda (IPV). Base 2015.
  https://www.ine.es/jaxiT3/Tabla.htm?t=25171
- Censo anual de población 2021-2024. Distribución del número de municipios según comunidad autónoma y provincia y tamaño del municipio.
  https://www.ine.es/jaxiT3/Tabla.htm?t=68201&L=0
  https://www.ine.es/jaxiT3/Datos.htm?t=2913
- Censo anual de población 2021-2024. Distribución del número de personas según provincia, comunidad autónoma y sexo.
  https://www.ine.es/jaxiT3/Tabla.htm?t=67988&L=0
- Nacimientos (cifras anuales). Por lugar de residencia de la madre y sexo. Total nacional y provincias. Añadidas las comunidades autónomas correspondientes.
  https://www.ine.es/jaxiT3/Tabla.htm?t=6506&L=0
- Defunciones (cifras anuales). Por lugar de residencia de la madre y sexo. Total nacional y provincias. Añadidas las comunidades autónomas correspondientes.
  https://www.ine.es/jaxiT3/Tabla.htm?t=6545&L=0
- Interrupciones voluntarias del embarazo (IVE). Tabla 3: “Tasas por 1.000 mujeres entre 15 y 44 años según Comunidad Autónoma de residencia. Total Nacional”.
  https://www.sanidad.gob.es/areas/promocionPrevencion/embarazo/
- Mercado Laboral (Tasas de actividad, empleo y paro) por provincia y sexo.
  https://www.ine.es/jaxiT3/Datos.htm?t=72989

https://www.ine.es/jaxiT3/Tabla.htm?t=3996

- Matrimonios de diferente sexo (cifras anuales).
  https://www.ine.es/jaxiT3/Tabla.htm?t=6532
- Porcentaje de viviendas turísticas sobre el total de viviendas censadas. Total nacional. Comunidades autónomas y provincias.
  https://www.ine.es/jaxiT3/Datos.htm?t=39365
- Inmigraciones procedentes del extranjero por comunidad autónoma, provincia y año (2021 a 2023).
  https://www.ine.es/jaxiT3/Datos.htm?t=69693
- Emigraciones con destino al extranjero por comunidad autónoma, provincia y año (2021 a 2023).
  https://www.ine.es/jaxiT3/Tabla.htm?t=69708&L=0
- Producto interior bruto (PIB) a precios de mercado, por comunidad autónoma, provincias y año.
  https://www.ine.es/jaxi/Tabla.htm?tpx=72943&L=0
  https://www.ine.es/jaxi/Tabla.htm?tpx=72943&L=0
- Importación y Exportación de Combustibles por año (2008 a 2023).
  https://www.ine.es/jaxi/Tabla.htm?tpx=33388
  https://www.ine.es/jaxi/Tabla.htm?tpx=33389

Definición de Ruralidad extraída de
https://www.mapa.gob.es/en/ministerio/servicios/analisis-y-prospectiva/Agrinfo12_tcm38-88390.pdf

Referencias a artículos o trabajos con relación a este tema de estudio:

- Goerlich y Cantarino (2015). *Estimaciones de la población rural y urbana a nivel municipal*.
- Manzanares (2015). *Patrones culturales de la población del Área Urbana Waslala*.
- Molina de la Torre (2018). *La despoblación en España, un análisis de la situación*.
- Hernández y Cruz (2020). *Evolución de la distribución de la población urbana y rural: un retrato de la España vaciada*.
- H. Warner, Liang, Sarkar, E. Mendes, J.Roethig. *Biomarkers of potential harm in a population of U.S. adult cigarette smokers and nonsmokers*. BMC Medical Research Methodology (2010).
- Everingham (2016). *Accurate prediction of sugarcane yield using a random forest algorithm.*
- Huang, Tsai, Wu, Lien, Yi Chien, Kuo, Hung, Chen y Kuo (2020). *Predictive modeling of blood pressure during hemodialysis: a comparison of linear model,*

random forest, support vector regression, XGBoost, LASSO regression and ensemble method. Computer Methods and Programs in Biomedicine, Vol. 195. Obtenido de https://www.sciencedirect.com/science/article/abs/pii/S0169260720301206