Facultad de Ciencias Económicas y Empresariales ICADE

Grado en Business Analytics

Trabajo de Fin de Grado

España vaciada: análisis del fenómeno de la despoblación rural en España mediante modelización y gestión de datos

Autor: Fernando de Dios Oubiña

Tutor: José Ramón Vallejo Rodrigo

Curso académico 2025-2026

Madrid, abril de 2026

RESUMEN

El progresivo vaciamiento demográfico del interior peninsular es uno de los grandes desequilibrios territoriales de la España contemporánea. Mientras el país ha crecido demográficamente de forma sostenida, amplias regiones rurales llevan décadas perdiendo población de manera ininterrumpida, con consecuencias que afectan no solo al número de habitantes, sino a la viabilidad de los servicios, el tejido económico y la identidad cultural de comunidades enteras.

Este trabajo analiza el fenómeno de la despoblación rural en España desde una perspectiva de Business Analytics, integrando datos estadísticos de fuentes oficiales a nivel municipal para construir una base de datos unificada que abarca dimensiones demográficas, económicas, laborales, de servicios, de conectividad y de accesibilidad. A partir de ella se desarrollan tres líneas de análisis complementarias. La primera construye proyecciones de la evolución poblacional de cada municipio hasta 2050 para identificar qué territorios se encuentran en mayor riesgo de extinción demográfica y dónde se concentran geográficamente. La segunda aplica técnicas de aprendizaje automático para explicar qué factores estructurales determinan la pérdida o recuperación de población, revelando el peso relativo de la demografía, los servicios, la conectividad y la economía en esta dinámica. La tercera segmenta los municipios españoles en perfiles estructurales homogéneos para comprender la heterogeneidad del territorio y analizar cómo la despoblación ha afectado de forma diferente a cada tipo de municipio.

Los resultados confirman patrones territoriales claros y aportan evidencia sobre el potencial alcance de las intervenciones de política pública en los territorios más vulnerables. Este trabajo no solo demuestra la aplicabilidad de técnicas avanzadas de ciencia de datos al análisis territorial, sino que ofrece una herramienta analítica con potencial para apoyar la toma de decisiones en materia de reto demográfico.

Palabras clave:

despoblación rural · España vaciada · reto demográfico · ciencia de datos · aprendizaje automático · Random Forest · clustering · proyección de población · análisis territorial · envejecimiento · Business Analytics · política pública · vulnerabilidad demográfica · extinción municipal · desequilibrio territorial ·

ABSTRACT

The progressive demographic emptying of Spain's interior is one of the most significant territorial imbalances in the country today. While Spain's overall population has grown steadily, vast rural regions have been losing inhabitants continuously for decades, with consequences that extend beyond population figures to affect the viability of services, local economies, and the cultural identity of entire communities.

This paper analyses rural depopulation in Spain from a Business Analytics perspective, integrating official statistical data at the municipal level to build a unified dataset covering demographic, economic, labour, services, connectivity, and accessibility dimensions. Three complementary lines of analysis are developed. The first builds population projections for each municipality up to 2050 to identify which territories face the greatest risk of demographic extinction and where they are geographically concentrated. The second applies machine learning techniques to explain which structural factors drive population loss or recovery, revealing the relative weight of demographics, services, connectivity, and the economy in this dynamic. The third segments Spanish municipalities into homogeneous structural profiles to understand territorial heterogeneity and analyse how depopulation has affected different types of municipality in different ways.

The findings confirm clear territorial patterns and provide evidence on the potential reach of public policy interventions in the most vulnerable territories. This work not only demonstrates the applicability of advanced data science techniques to territorial analysis, but offers an analytical tool with potential to support evidence-based decision-making on Spain's demographic challenge.

Key words:

rural depopulation · empty Spain · demographic challenge · data science · machine learning · Random Forest · clustering · population projection · territorial analysis · ageing · Business Analytics · public policy · demographic vulnerability · municipal extinction · territorial imbalance

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

Declaración de Uso de Herramientas de Inteligencia Artificial Generativa en Trabajos Fin de Grado

ADVERTENCIA: Desde la Universidad consideramos que ChatGPT u otras herramientas similares son herramientas muy útiles en la vida académica, aunque su uso queda siempre bajo la responsabilidad del alumno, puesto que las respuestas que proporciona pueden no ser veraces.

Por la presente, yo, Fernando de Dios Oubiña, estudiante de Doble Grado de Administración y Dirección de Empresas y Business Analytics de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado *"España vaciada: análisis del fenómeno de la despoblación rural en España mediante modelización y gestión de datos"*, declaro que he utilizado herramientas de Inteligencia Artificial Generativa — principalmente Claude (Anthropic) y ChatGPT (OpenAI)— únicamente en el contexto de las actividades descritas a continuación:

1. Corrector de estilo literario y de lenguaje: para mejorar la calidad lingüística y estilística del texto.
2. Revisor: para recibir sugerencias sobre cómo mejorar y perfeccionar el trabajo con diferentes niveles de exigencia.
3. Traductor: para traducir textos de un idioma a otro.
4. Apoyo en la confección y revisión del código de modelización. En todo momento la lógica analítica, las decisiones metodológicas y la interpretación de los resultados han sido responsabilidad del autor.

Afirmo que toda la información y contenido presentados en este trabajo son producto de mi investigación y esfuerzo individual, excepto donde se ha indicado lo contrario y se han dado los créditos correspondientes. Soy consciente de las implicaciones académicas y éticas de presentar un trabajo no original y acepto las consecuencias de cualquier violación a esta declaración.

Fecha: 19 de abril de 2026

Firma: 

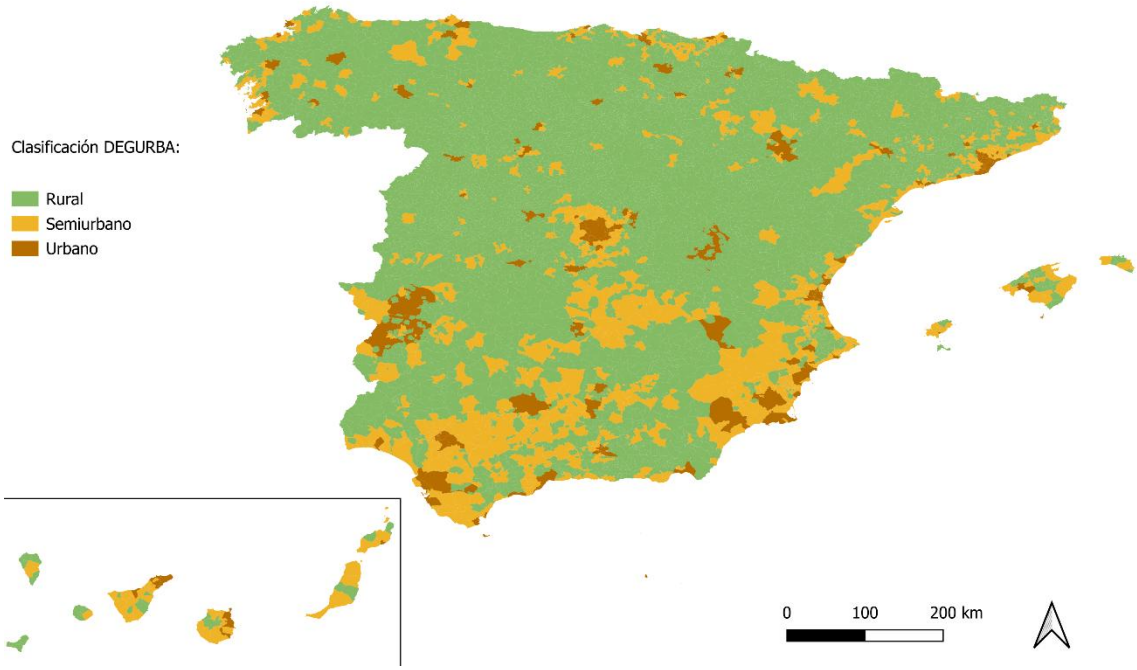
Introducción y justificación del tema

Contexto general del problema

Desde 1975 hasta 2025, la población española ha aumentado de los 35,6 millones de habitantes hasta los 49,1 millones (INE, s.f.). ¿Cómo es posible entonces que hablemos de despoblación?

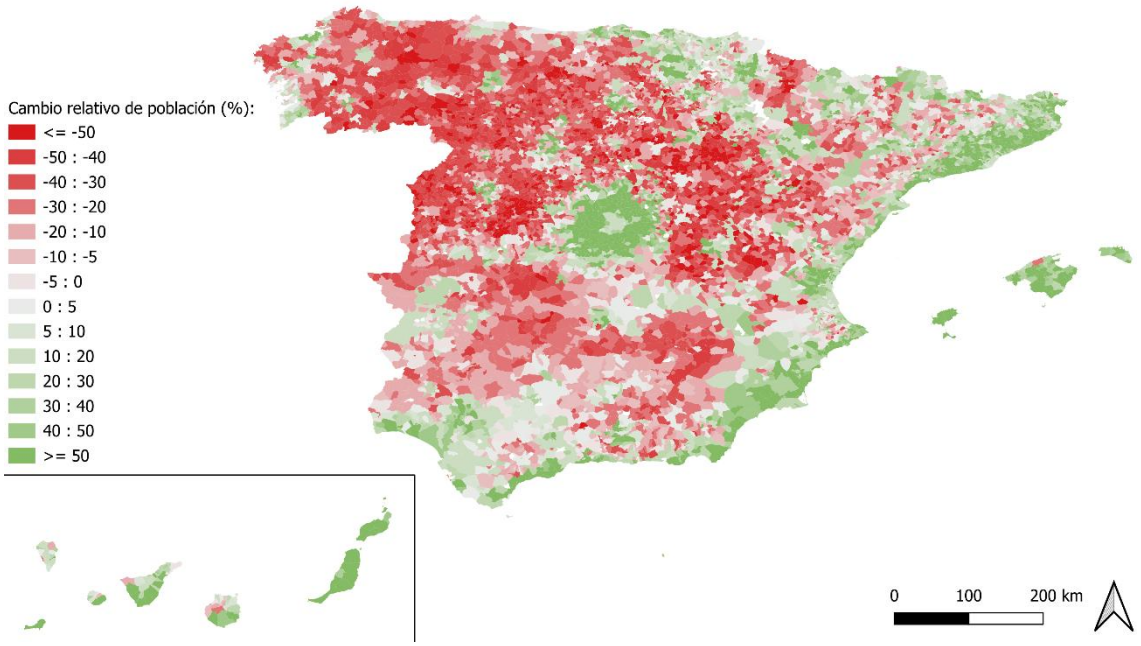
Esto es debido a que este incremento demográfico, de más de 13 millones de personas, no se ha notado por igual en todas las partes de España. Durante las últimas décadas, el país ha sufrido una revolución económica donde amplias regiones se han visto afectadas por movimientos migratorios de gran calado desde las zonas rurales hacia las grandes ciudades. La terciarización de la economía ha ido acumulando las oportunidades laborales en torno a grandes urbes y corredores dinámicos, a diferencia de las zonas rurales, donde no se han generado debido a una mecanización del sector primario desde mediados del siglo XX que redujo la necesidad de mano de obra y la desindustrialización del sector secundario sufrida desde la entrada de España en la Comunidad Europea en 1985. A toda esta falta de oportunidades en las zonas rurales, se suma el deterioro y desaparición de centros educativos, servicios básicos y actividades de ocio, además de una búsqueda de progreso motivada por la promesa de una vida mejor en las ciudades. En las siguientes figuras, se puede observar: primero, un mapa de España municipal actual clasificado según el grado de urbanización, para conocer cuáles son las zonas rurales y cuáles las urbanas; y segundo, la ganancia o pérdida relativa de población por municipio en España desde 1998 hasta 2024.

Figura 1. *Clasificación DEGURBA de los municipios de España (2023)*



Fuente: Eurostat (DEGURBA). Elaboración propia.

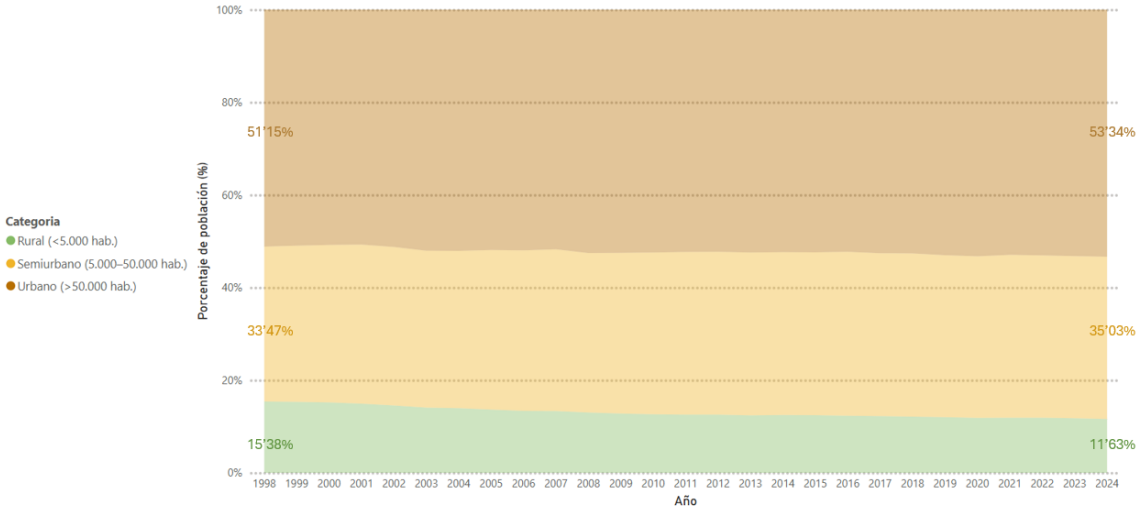
Figura 2. *Variación relativa de la población municipal en España (1998–2024)*



Fuente: Instituto Nacional de Estadística (Padrón municipal). Elaboración propia.

Como se puede apreciar visualmente, las zonas más castigadas demográficamente coinciden con las áreas rurales, mientras que las más beneficiadas coinciden con zonas urbanas y semiurbanas. De hecho, resulta destacable el aumento en grandes ciudades, los territorios insulares, el sur peninsular y el corredor levantino. Profundizando un poco más en este fenómeno, en la siguiente figura se aprecia la evolución desde 1998 hasta 2024 de la distribución del porcentaje de población que vive en zonas rurales, semiurbanas y urbanas. Para la construcción de este gráfico, se ha tomado como inspiración los criterios DEGURBA, clasificación que realiza el Eurostat sobre el nivel de urbanización, pero con datos del INE, para disponer de un rango temporal mayor que el disponible por la fuente original.

Figura 3. *Evolución de la distribución de la población en España según tipo de zona (1998–2024)*



Fuente: Instituto Nacional de Estadística (Padrón municipal). Elaboración propia.

Mientras que el porcentaje de población que vive en zonas urbanas y semiurbanas ha crecido, el porcentaje de población que vive en zonas rurales ha caído del 15,4% al 11,6% en el último cuarto de siglo.

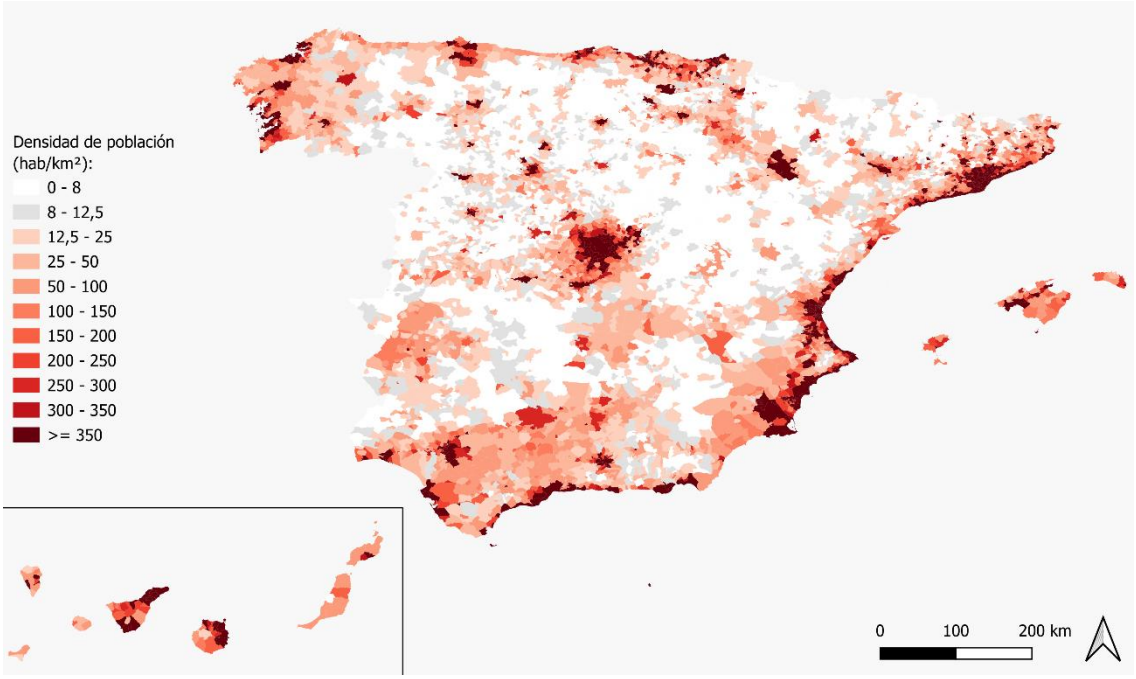
Llegados a este punto, cabría preguntarse: ¿es realmente tan grave que haya menos población en las zonas rurales si, en conjunto, el país ha seguido creciendo y la caída del 4% no parece extrema? La respuesta exige mirar más allá y tener en cuenta tres dimensiones clave: cuántos municipios están afectados y con qué intensidad pierden población, cómo de homogéneo es el reparto de población en el territorio nacional, y cuál es su estructura demográfica.

En primer lugar, esta variación se ha dado en apenas 30 años, continuando la ininterrumpida, larga y abrupta despoblación que los pueblos llevan sufriendo desde mediados del siglo XX por las citadas causas mencionadas anteriormente. Según información del Instituto para la Diversificación y Ahorro de la Energía, dependiente del Ministerio para la Transición Ecológica y el Reto Demográfico, en España existen 6.827 municipios que no superan los 5.000 habitantes, donde ocho de cada diez han perdido habitantes en la última década. Si nos fijamos en los municipios de menos de 1.000 habitantes, la proporción de casos en despoblación asciende al 86%. Cuando se amplía el análisis a municipios no urbanos de hasta 20.000 habitantes en los que todos los núcleos de población son menores de 5.000 habitantes, el número de "municipios en reto demográfico" se eleva hasta 6.974, lo que representa alrededor del 86% de los municipios y 74% del territorio, donde vive un 14% de la población española. Estas cifras muestran que el fenómeno no es anecdótico, sino que afecta a una parte sustancial del territorio (IDAE, s. f.).

En segundo lugar, esas comarcas rurales que no dejan de desangrarse demográficamente son precisamente las menos densamente pobladas de la geografía del país. En el mapa de la siguiente figura, se pueden apreciar grandes extensiones prácticamente inhabitadas, especialmente en el interior, en las mesetas castellanas, que se encuentran en situación

de riesgo (<12,5 hab/km^2) y riesgo extremo (<8 hab/km^2) según criterios de la UE de riesgo demográfico.

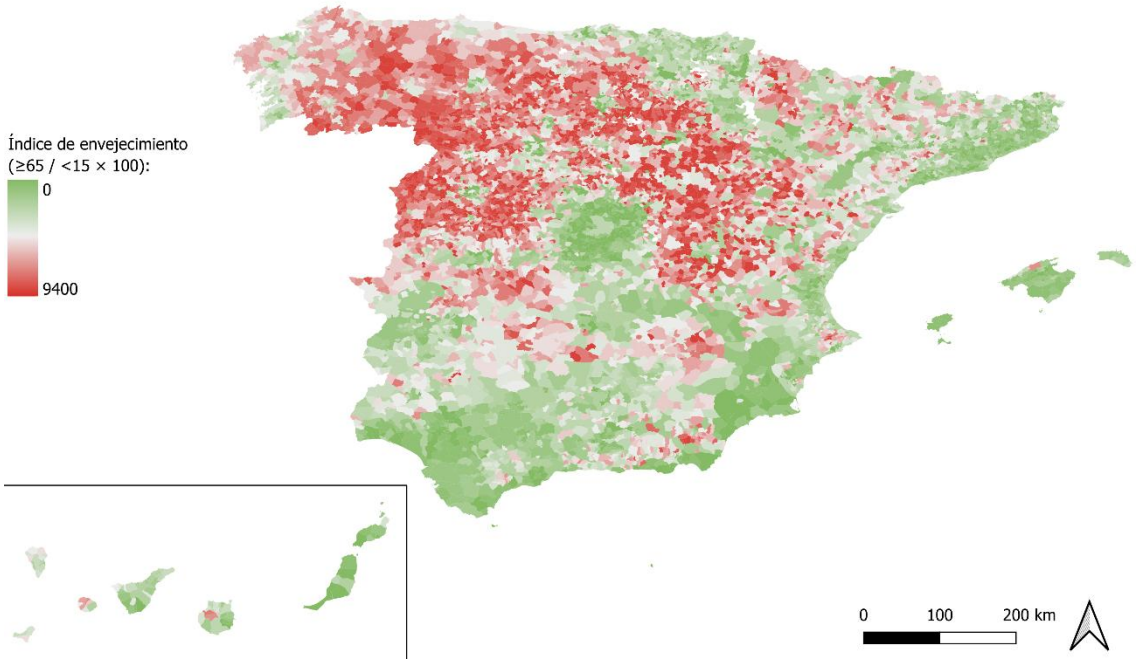
Figura 4. *Densidad de población por municipio en España en 2024*



Fuente: Instituto Nacional de Estadística e Instituto Geográfico Nacional (Padrón municipal y Nomenclátor Geográfico de Municipios). Elaboración propia.

En tercer lugar, la fatal y no casual coincidencia de dichas comarcas más despobladas siendo también las más envejecidas, como se puede observar en el siguiente mapa.

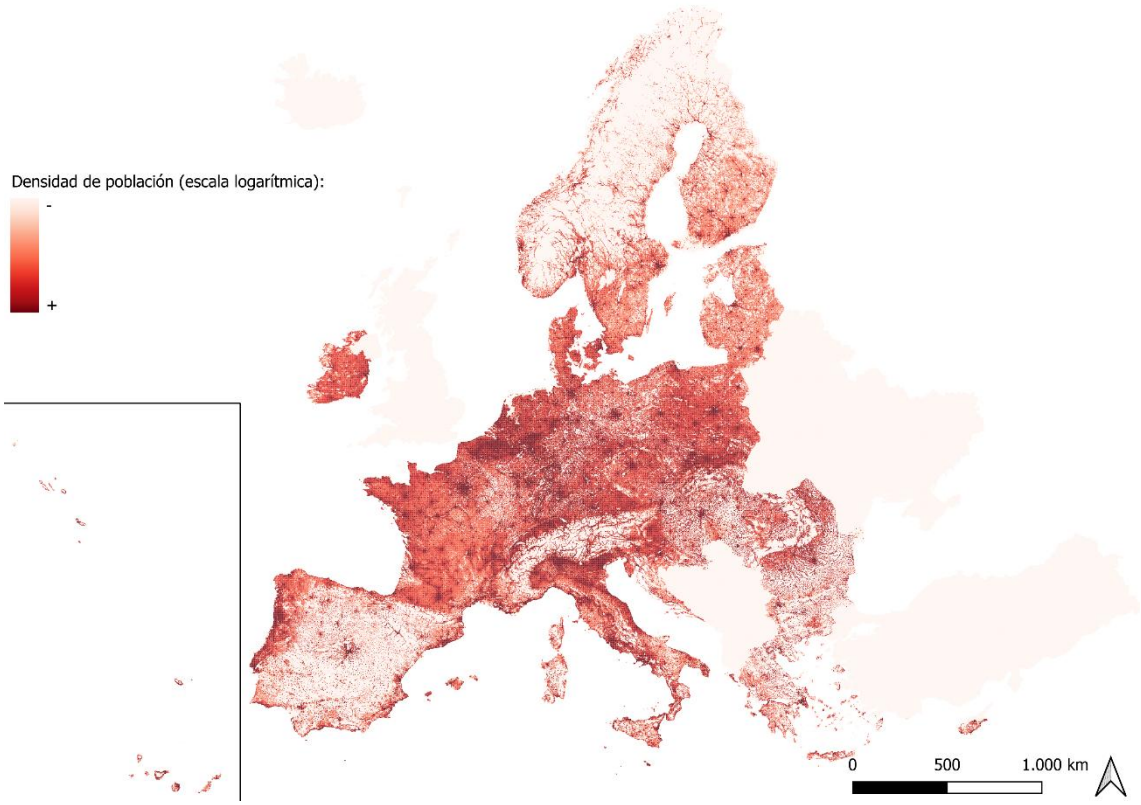
Figura 5. *Índice de envejecimiento por municipio en España en 2024*



Fuente: Instituto Nacional de Estadística (Nomenclátor. Población del Padrón). Elaboración propia.

Por último, puede ser que alguien se preguntase si este fenómeno fuese una tendencia general de las naciones desarrolladas y, aunque es cierto que estos patrones se dan también en otros países, es en España donde presenta un cariz alarmante. Si levantamos la mirada a nuestros vecinos europeos, en el siguiente mapa de densidad de población, observamos que la Península Ibérica presenta algunas de las zonas más desiertas. De hecho, existe en Guadalajara un señorío, Molina de Aragón, coloquialmente conocida como 'la Siberia española', no sólo por ser una de las zonas más frías de la Península, sino por ser igual de inhóspita que dicha región rusa o la Laponia, con una densidad de población de 2'63 habitantes/km^2 (Ávila, 2019). Entre los territorios de Europa, son los alrededores de ese municipio, la Serranía Celtibérica, considerado el mayor desierto demográfico por despoblación de la Unión Europea e incluida en las *Áreas Escasamente Pobladas del Sur de Europa* (Southern Sparsely Populated Areas, SSPA), caracterizadas por densidades inferiores a 12,5 habitantes por kilómetro cuadrado y por una pérdida demográfica persistente, situando a España como el país más afectado del sur de Europa en términos de despoblación territorial (Red SSPA, 2024).

Figura 6. *Densidad de población en Europa (escala logarítmica)*

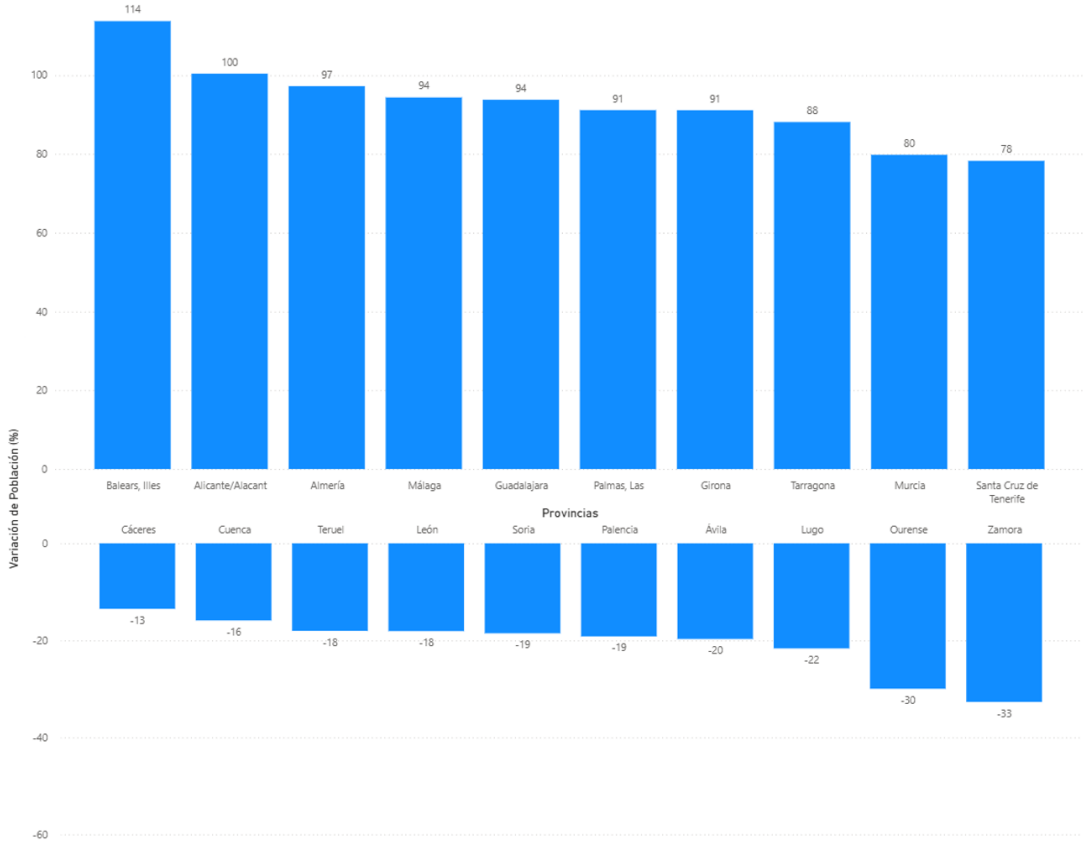


Fuente: Eurostat (GISCO, Population grid data). Elaboración propia.

Por todos estos agravantes, consideramos especialmente crítica esta situación en la que las zonas rurales, que son las que más población han perdido en los últimos lustros, tienden a ser las más desiertas y las más envejecidas con menor relevo generacional, dirigiéndose con mayor velocidad al precipicio y la extinción.

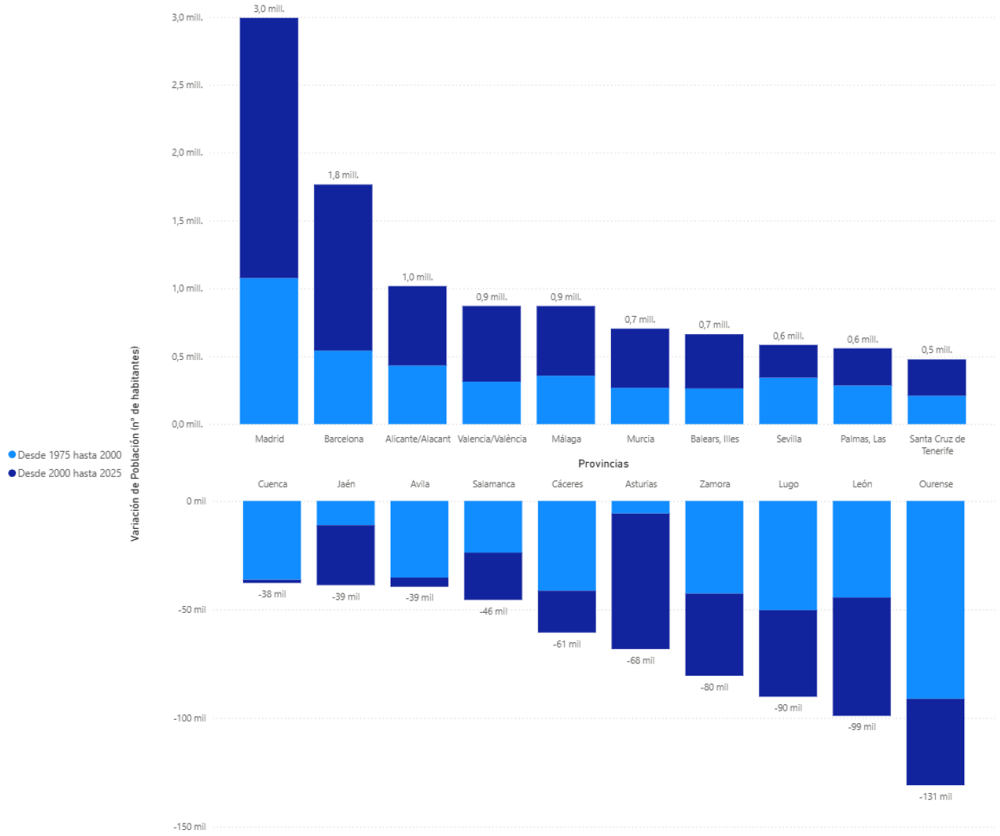
Si entramos más en detalle, en las siguientes figuras se puede observar lo siguiente. Por un lado, las provincias que más perjudicadas se han visto por la pérdida de población han sido Zamora y Orense, que ha perdido casi un tercio de su población desde 1975 hasta la actualidad. Además, resulta interesante ver cómo a lo largo de este periodo hay provincias, como Orense o Lugo, que han perdido más población en la primera mitad, mientras otras como Asturias mayoritariamente en la segunda mitad. Por otro lado, las que más han visto aumentar los números de sus censos han sido, en términos absolutos, las provincias con las ciudades más grandes del país y, en términos relativos, destacan zonas del Levante, como Baleares y Alicante, que han doblado la población solamente en los últimos 50 años.

Figura 7. *Provincias con mayor ganancia y pérdida de población en España desde 1975 hasta 2025 (variación relativa)*



Fuente: Instituto Nacional de Estadística (Estadística continua de población). Elaboración propia.

Figura 8. *Provincias con mayor ganancia y pérdida de población en España desde 1975 hasta 2025 (variación absoluta)*



Fuente: Instituto Nacional de Estadística (Estadística continua de población). Elaboración propia.

Todo esto tiene implicaciones profundas. La pérdida de población y el envejecimiento en las áreas rurales se traducen en cierres de escuelas, dificultades para mantener consultorios médicos, reducción del transporte público, menor inversión en infraestructuras y servicios, así como en un debilitamiento del tejido productivo local. A ello se suma la erosión de tradiciones, formas de vida y patrimonio cultural asociados históricamente a estos territorios.

En este escenario que venimos dibujando a lo largo de la introducción surge el término “España vaciada”, utilizado no solo como simple eslogan mediático, sino como concepto utilizado en la literatura especializada para hacer referencia al conjunto de procesos demográficos, económicos y sociales que han conducido al progresivo vaciamiento de amplias áreas del territorio nacional (Del Molino, 2016). Esta categoría engloba cuestiones estructurales como la pérdida de población, envejecimiento, declive de servicios, reducciones de oportunidades y desigualdades territoriales, que ponen en cuestión la cohesión del país y suponen uno de los mayores retos a los que se enfrenta la España contemporánea. ¿Y qué municipios se consideran parte de la “España vaciada”? En la revisión bibliográfica posterior se abordarán con detalle los estudios realizados sobre este fenómeno hasta el momento y los requisitos que se consideran necesarios para formar parte de este conjunto.

En este contexto, surge la necesidad de ir más allá de la mera descripción del problema. Para comprender de forma rigurosa qué municipios están en mayor riesgo, cuáles son los factores que mejor explican la pérdida de población y qué tendencias pueden esperarse en el futuro, es necesario disponer de herramientas analíticas basadas en datos. Este trabajo buscará aportar desde esta perspectiva: a partir de la integración de información estadística a nivel municipal y del uso de técnicas de ciencia de datos, se plantea analizar la variación poblacional municipal e identificar los factores que explican la despoblación rural en España, ofreciendo instrumentos que puedan contribuir a una toma de decisiones más informada en materia de políticas públicas.

Relevancia del tema

Este fenómeno, como se ha explicado en la sección anterior, constituye un reto de primera magnitud para España, lo que se evidencia en las múltiples reacciones que han tenido lugar desde distintos ámbitos de la sociedad como respuesta a esta problemática: desde organismos institucionales del Estado español y de la Unión Europea hasta los medios de comunicación, la comunidad académica y la sociedad civil.

En primer lugar, a nivel institucional se ha producido una respuesta en diferentes categorías de gobierno. A escala autonómica y local, comunidades autónomas y ayuntamientos han empezado a promulgar normativa y planes específicos para afrontar la despoblación. Por un lado, una sistematización de este marco normativo puede encontrarse en el repositorio elaborado por la Universidad de Zaragoza, que recopila las principales leyes estatales y autonómicas vinculadas al Reto Demográfico y al Desarrollo Territorial en España, organizadas por nivel administrativo y comunidad autónoma (Universidad de Zaragoza, s. f.). Por otro lado, un ejemplo representativo de los numerosos ayuntamientos que ha comenzado a diseñar e implementar planes estratégicos propios para hacer frente a los procesos de pérdida y envejecimiento de la población ha sido el de San Esteban del Valle (Ávila), con su *Plan Estratégico contra la Despoblación 2019-2022* (Ayuntamiento de San Esteban del Valle, 2019). A escala estatal, el Gobierno de España creó el Ministerio para la Transición Ecológica y el Reto Demográfico (MITECO) mediante el Real Decreto 2/2020, el 12 de enero de 2020, que concedió al ministerio "la elaboración y el desarrollo de la política del Gobierno frente al reto demográfico y el despoblamiento territorial" (*art.* 14.1 del Real Decreto). De hecho, este organismo implantó en 2021 el *Plan de Recuperación: 130 medidas frente al Reto Demográfico*, e inició a principios de 2025 un proceso de consulta pública para la nueva *Estrategia Nacional para la Equidad Territorial y el Reto Demográfico 2030*, entre otras iniciativas que han asignado a la lucha contra el abandono rural como una prioridad política. A escala supranacional, desde la Unión Europea se ha reconocido la importancia de esta cuestión mediante el European Agricultural Fund for Rural Development (EAFRD) 2021-2027, que incluye partidas concretas de €8.100 millones procedentes de los fondos NextGenerationEU para afrontar la crisis de la COVID-19 y promover el desarrollo rural resiliente (Comisión Europea, 2025).

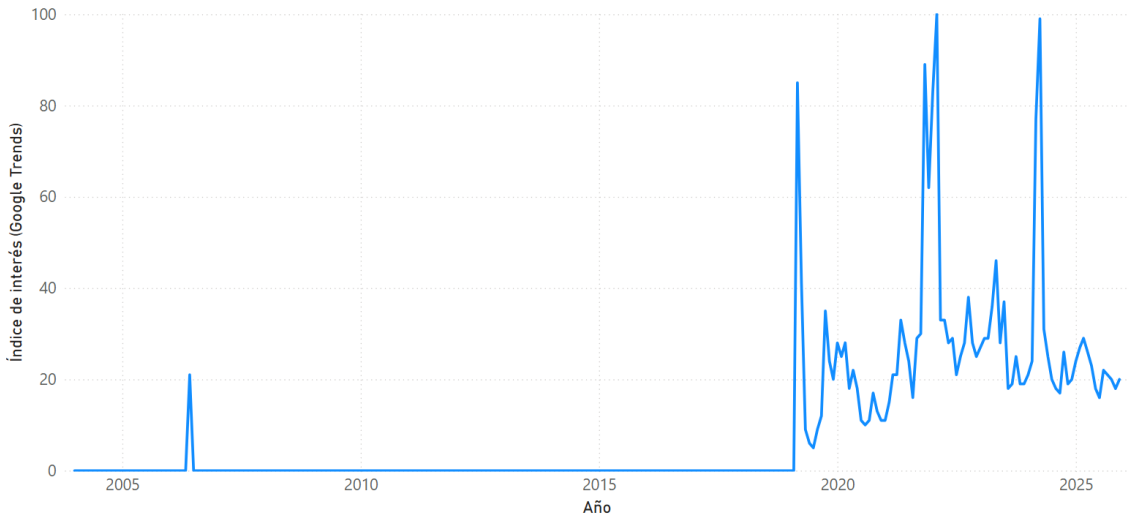
En segundo lugar, también ha habido una respuesta social notoria desde las zonas rurales. El 31 de marzo de 2019 tuvo lugar en Madrid la manifestación denominada *La Revuelta de la España Vaciada*, en la que decenas de miles de personas procedentes de más de veinte provincias se concentraron para exigir un pacto de Estado contra la despoblación (Sosa Troya, 2019). Más recientemente, el 1 de abril de 2023, diferentes asociaciones que luchan

por la pervivencia del medio rural despoblado convocaron protestas en numerosas capitales de provincia bajo el lema "No queremos ser territorio de sacrificio" (Agencia EFE, 2023).

En tercer lugar, esta movilización social ha sido seguida de una respuesta política, con la aparición de agrupaciones que han llevado la lucha contra la despoblación rural por bandera. Por ejemplo, Teruel Existe, que obtuvo un escaño en el Congreso tras las elecciones generales del 10 de noviembre de 2019, o Soria ¡Ya!, que concurrió a las elecciones generales del 23 de julio de 2023 con la intención de representar a la provincia de Soria. Estas candidaturas ilustran la dimensión política del fenómeno y su inserción en la agenda pública.

Por último, a nivel mediático, el término *"España vaciada"* ha ganado difusión y visibilidad en los últimos años. Su creciente presencia en medios de comunicación, redes sociales e internet, como se puede ver en la figura de debajo, evidencia que la sociedad presta cada vez mayor atención al desafío demográfico que enfrentan los territorios rurales. De hecho, incluso desde el ámbito académico ha aumentado el interés por analizar esta problemática, como se podrá ver en mayor profundidad más adelante, en la sección de estado del arte, por la cantidad de estudios realizados recientemente.

Figura 9. *Evolución del interés de búsqueda en Google por el término "España vaciada" (2004-2025)*



Fuente: Google Trends (interés de búsqueda normalizado, 0–100). Elaboración propia.

Motivación

Por un lado, este trabajo nace de una motivación personal. La despoblación rural no es solo un fenómeno demográfico o económico, sino una transformación profunda de la identidad de un país. España es, en gran parte, el resultado de sus pueblos, de sus costumbres locales y de una forma de vida que durante siglos ha ido tejiendo el carácter de su gente. El éxodo que está sufriendo el mundo rural para abarrotar las grandes urbes cada vez más despersonalizadas y globalizadas supone la pérdida de una parte esencial de nuestra cultura y nuestras tradiciones. Este proyecto pretende, desde la analítica de datos, aportar

conocimiento y conciencia sobre esta realidad, ayudando a comprender las causas del éxodo rural con el fin de preservar la vitalidad de las zonas rurales y su legado.

Por otro lado, por parte de la universidad se insiste en apoyarse en alguno de los Objetivos de Desarrollo Sostenible de la Agenda 2030 como propósito general. En el caso de esta investigación, se alinea con el ODS 11 (Ciudades y comunidades sostenibles), que promueve el desarrollo de asentamientos humanos inclusivos, seguros, resilientes y sostenibles. En este contexto, potenciar el medio rural y frenar la despoblación constituye una vía complementaria para reducir la presión demográfica sobre las grandes ciudades y avanzar hacia un desarrollo territorial más equilibrado y sostenible. También se relaciona con el ODS 8 (Trabajo decente y crecimiento económico) y el ODS 9 (Industria, innovación e infraestructuras), al abordar la necesidad de generar actividad económica, oportunidades laborales y conectividad para favorecer un crecimiento sostenible e inclusivo con el medio rural (Naciones Unidas, 2015).

Estructura del TFG

Este Trabajo de Fin de Grado se estructura en ocho capítulos, que reflejan las distintas fases del proceso de investigación y análisis desarrollado.

En el primer capítulo, se presenta la introducción y contextualización del fenómeno, en la que se expone el problema de la despoblación rural en España, su relevancia político-social, así como la motivación y estructura del estudio.

El segundo capítulo define los objetivos generales y específicos, junto con el alcance del trabajo donde se precisan las metas analíticas que se persiguen y los límites del estudio en términos espaciales, temporales y metodológicos.

En el tercer capítulo, se desarrolla el marco teórico y la revisión bibliográfica, que recoge los principales antecedentes académicos e institucionales sobre la despoblación rural y la "España vaciada". En esta sección se analizan las aportaciones más relevantes del estado del arte y se identifican los enfoques utilizados en estudios previos, sirviendo de base para la formulación del enfoque propio del presente trabajo.

El cuarto capítulo aborda el diseño metodológico y la arquitectura de datos, describiendo las fuentes empleadas, la estructura de los datos municipales utilizados, los procesos de recopilación y limpieza de la información, y las herramientas de análisis aplicadas.

En el quinto capítulo, se presenta el análisis exploratorio y la modelización, donde se estudian las relaciones entre las distintas variables, se identifican patrones territoriales y se desarrollan modelos predictivos orientados a explicar la variación poblacional histórica y construir escenarios de proyección.

El sexto capítulo recoge los resultados, interpretando los hallazgos obtenidos y contrastándolos con la literatura existente, para evaluar la coherencia y el valor añadido de las conclusiones alcanzadas. Además, aborda las conclusiones y líneas futuras, en las que se sintetizan los principales aportes del trabajo, se reconocen las limitaciones del análisis y se plantean posibles vías de investigación y mejora.

Finalmente, el trabajo incluye un apartado de bibliografía, con todas las fuentes académicas, institucionales y técnicas consultadas, así como un conjunto de anexos con información complementaria que apoya el desarrollo del proyecto.

Objetivos y alcance del estudio

Objetivo general

El propósito principal del proyecto es analizar la despoblación rural en España mediante técnicas de ciencia de datos, identificar los factores que explican la pérdida o recuperación de población, y construir escenarios de proyección futura, identificando municipios en situación de mayor vulnerabilidad demográfica. A través del diseño de un modelo explicativo y la elaboración de visualizaciones, se pretende ofrecer una herramienta analítica que permita comprender en profundidad el trasfondo de este fenómeno, detectar municipios en riesgo y contribuir al diseño de estrategias basadas en datos para mitigar el proceso de vaciamiento demográfico.

Este enfoque parte de la consideración de que la despoblación no es únicamente un problema demográfico, económico o social, sino también un problema de datos, derivado de la necesidad de integrar, estructurar y analizar grandes volúmenes de información territorial de forma coherente. Bajo esta perspectiva, el proyecto busca aportar una visión cuantitativa y objetiva, complementando los enfoques descriptivos predominantes en los estudios existentes. Se pretende demostrar cómo la analítica de datos puede convertirse en una herramienta que revitalice la formulación de políticas públicas más informadas y efectivas, al traducir la complejidad del fenómeno en indicadores, métricas y números medibles y cuantificables.

Objetivos específicos

Recopilar y estructurar un conjunto de datos a nivel municipal procedentes de diferentes fuentes oficiales, como el Instituto Nacional de Estadística (INE), el SEPE, el Catastro o el Ministerio para la Transición Ecológica y el Reto Demográfico (MITECO), incorporando tanto archivos estructurados (CSV, Excel) como datos obtenidos mediante APIs cuando sea posible. El objetivo es construir un dataset unificado y limpio que incluya dimensiones demográficas, económicas, laborales, de infraestructura y conectividad.

Diseñar una arquitectura de datos y un modelo de almacenamiento que garanticen la coherencia, trazabilidad y accesibilidad de la información recopilada, empleando para ello modelos relacionales o no relacionales, según la naturaleza de los datos.

Realizar un análisis exploratorio de los datos (EDA) para identificar patrones, correlaciones y relaciones significativas entre variables, empleando librerías de Python como pandas, numpy, matplotlib o seaborn. Además, se buscará detectar al menos cinco variables especialmente influyentes en los procesos de despoblación o mantenimiento de población de los municipios.

Desarrollar un modelo analítico o predictivo basado en técnicas estadísticas o de machine learning, como por ejemplo una regresión lineal o random forest, que permita explicar la despoblación, utilizando librerías como scikit-learn. El modelo tendrá como propósito estimar la pérdida de población de los municipios españoles y analizar el peso relativo de cada predictor. Su rendimiento será evaluado mediante métricas apropiadas, como R^2, entre otros, con el fin de asegurar su validez y utilidad práctica.

Elaborar visualizaciones interactivas y representaciones geográficas mediante el uso de herramientas como Power BI que faciliten la comprensión e interpretación de los resultados obtenidos, permitiendo comunicar las conclusiones de manera clara y visual. Entre las visualizaciones previstas destacan mapas coropléticos por municipio, gráficos de dispersión que relacionen variables clave, mapas de calor regionales y dashboards que permitan explorar las variables más relevantes que explican la despoblación.

Alcance del estudio

El estudio se centrará en los municipios españoles incluidos en el Registro de Entidades Locales (Ministerio de Política Territorial, s. f.), durante el periodo de 1998-2024, para variables con un enfoque cuantitativo como demográficas, económicas, laborales, de infraestructura y conectividad.

Se excluyen variables con un enfoque cualitativo, como factores políticos, históricos o legislativos que, aunque pudieran ser relevantes para comprender mejor el fenómeno de la despoblación rural, exceden el propósito analítico de este trabajo.

Por último, destacar las limitaciones de este trabajo particular. En primer lugar, la principal sería la disponibilidad y accesibilidad insuficiente de datos. Para las diferentes y numerosas variables se necesita un nivel de desagregación municipal que no siempre presentan, y puede haber brechas o inconsistencias temporales, pues muchos indicadores no existen para largos periodos y no todos los organismos publican información con la misma frecuencia. En segundo lugar, problemas de calidad de los datos, por valores perdidos, variables incompletas o cambios metodológicos en la recopilación de datos a lo largo del tiempo. En tercer lugar, restricciones computacionales y de procesamiento, por la capacidad limitada de hardware personal para ejecutar modelos y manejar grandes volúmenes de datos. En caso necesario, se podrían estudiar alternativas, como el uso de servicios en la nube para aumentar la capacidad de procesamiento, condicionado a un presupuesto limitado. Por último, la limitación temporal de meses de la que se dispone para entregar este trabajo en tiempo y forma que limitará la profundidad del análisis y su exhaustividad.

Revisión bibliográfica y contextualización del fenómeno

Marco conceptual del fenómeno

Dado que este trabajo versa sobre la "España vaciada", es importante comenzar aclarando varios aspectos sobre este término para partir de una base sólida y clara para el posterior análisis. Comprender qué es, por qué ocurre y sus consecuencias será el fin de esta sección.

En primer lugar, como se explicó en la introducción, este término se utiliza para hacer referencia al conjunto de procesos demográficos, económicos y sociales que han conducido al progresivo vaciamiento de amplias áreas del territorio nacional. Este fenómeno no solo se centra en la despoblación rural, sino que engloba otras cuestiones estructurales como el envejecimiento, declive de servicios, reducciones de oportunidades y desigualdades territoriales, que ponen en cuestión la cohesión del país.

Siguiendo la literatura especializada, la despoblación rural puede definirse como un proceso prolongado de pérdida de población en los territorios de baja densidad, generalmente asociado a un saldo migratorio negativo, un envejecimiento acusado y tasas de natalidad reducidas. Collantes y Pinilla (2019), en su obra *¿Lugares que no importan? La despoblación de la España rural desde 1900 hasta el presente*, muestran cómo buena parte del medio rural español lleva más de un siglo encadenando fases de estancamiento y retroceso demográfico, en contraste con el dinamismo de las áreas urbanas e industriales. Estos autores subrayan que la despoblación no es únicamente una cuestión de pérdida de habitantes, sino también de cambio estructural: transformación del modelo productivo, concentración de empleo y servicios en las ciudades, abandono de actividades agrarias tradicionales y creciente dependencia de centros urbanos para el acceso a oportunidades laborales, educativas y sanitarias. En términos funcionales, la despoblación rural implica, por tanto, una reducción sostenida de la capacidad de estos territorios para retener y atraer población, especialmente joven y en edad activa.

El término "España vacía" fue popularizado por Sergio del Molino en 2016 con la publicación de su libro, *La España Vacía, viaje por un país que nunca fue*, para describir el conjunto de territorios del interior peninsular caracterizados por una baja densidad de población, un fuerte envejecimiento y una sensación de abandono histórico respecto a los grandes centros de decisión política y económica. Poco después, diversos movimientos sociales y plataformas procedentes de estas zonas empezaron a hablar de "España vaciada", enfatizando que el vaciamiento no es un fenómeno natural o espontáneo, sino el resultado de decisiones políticas, económicas y de planificación territorial que, de forma acumulada, han favorecido la concentración de recursos y oportunidades en determinados espacios en detrimento de otros (Verón Lassa y Hernández Ruiz, 2021; Cabello, 2024).

La literatura reciente sobre el reto demográfico y la cohesión territorial en España recoge este concepto para referirse a aquellos territorios donde la combinación de baja densidad, envejecimiento y declive socioeconómico plantea problemas de sostenibilidad a medio y largo plazo. De hecho, en el ámbito de las políticas públicas, el Ministerio para la Transición Ecológica y el Reto Demográfico (MITECO) ha introducido la noción de "municipios de reto

demográfico" para delimitar, según unos criterios muy generales, los espacios más afectados que podrían beneficiarse de determinadas subvenciones estatales. En distintas convocatorias y programas, como el PREE 5000, estos hacen referencia a aquellos de menos de 5.000 habitantes (6.827 municipios), y municipios no urbanos de hasta 20.000 habitantes en los que todas sus entidades singulares tienen menos de 5.000 habitantes (148 municipios) (Real Decreto 691/2021). LA UE usa en ocasiones otros criterios de riesgo de despoblación, calificando de riesgo y riesgo severo de despoblación a municipios con densidades de población por km^2 de menos de 12,5 y 8 respectivamente.

En segundo lugar, para comprender el surgimiento de este fenómeno es necesario atender las causas estructurales que explican la despoblación rural, como se esbozó brevemente en la introducción. La literatura académica coincide en señalar que el fenómeno no puede atribuirse a un único factor, sino a la interacción de procesos económicos, demográficos y territoriales de largo recorrido. Collantes y Pinilla (2019) explican que desde mediados del siglo XX la mecanización agrícola redujo drásticamente la necesidad de mano de obra en el campo, acelerando la salida de población joven hacia destinos urbanos e industriales. Asimismo, desde la entrada de España en la Comunidad Europea en 1986 ha habido un drástico cambio de la estructura económica del país. En paralelo a una desindustrialización del sector secundario que conllevó la decadencia de zonas industriales, como la industria minera de las cuencas de Asturias y León, la industria siderúrgica de los Altos Hornos de Vizcaya o los astilleros de Cádiz y Ferrol, entre otros ejemplos, el país experimentó una progresiva terciarización de la economía, especializándose en sector servicios y turismo en litorales y grandes urbes, concentrando oportunidades en estas zonas (FUNCAS, 2015). Sumado a esto, un deterioro en las infraestructuras y servicios de los entornos rurales agrandó la brecha entre las ciudades y los pueblos, como analiza el estudio *Thinking in rural gap: mobility and social inequalities* de Camarero y Oliva (2019). No debemos olvidar que las zonas rurales han sufrido, como parte de la sociedad española, el mismo fenómeno de caídas de las tasas de natalidad que se ha dado en todos los países desarrollados, como señala Macarrón (2017). Finalmente, factores socioculturales como la búsqueda de mejores oportunidades educativas, laborales y vitales en las ciudades han reforzado durante décadas estas dinámicas migratorias. El estudio *La representación del éxodo rural en el cine español (1900–2020)* muestra el reflejo en el séptimo arte, el cine, de esta idea en la cultura popular que ha reforzado, durante más de un siglo, una concepción de lo urbano y las ciudades como desarrollo, modernidad y calidad de vida donde las ciudades representan oportunidades y progreso, mientras que el medio rural aparece ligado al atraso, la dureza de las condiciones laborales o la falta de perspectivas.

En tercer lugar, las consecuencias de este proceso trascienden la mera pérdida de habitantes en las zonas rurales. La despoblación genera desequilibrios territoriales profundos, que se traducen en una creciente vulnerabilidad económica y social, y comprometen la viabilidad misma de numerosos municipios a corto y medio plazo. La reducción sostenida de población implica menor demanda y, por tanto, cierre progresivo de servicios esenciales como escuelas, centros de salud, transporte público, pequeños comercios o actividades culturales. Este deterioro impacta directamente en la calidad de vida de los habitantes rurales y acelera un círculo vicioso de abandono, dificultando aún más el arraigo. A ello se suma la potencial pérdida de una herencia cultural y formas de vida tradicionales que caracteriza al medio rural, poniendo en riesgo la continuidad de saberes y modos de vida que forman parte de la identidad histórica de la nación.

La gravedad de estas dinámicas se refleja en indicadores demográficos contundentes: desde 2015 España registra más defunciones que nacimientos a nivel nacional, según datos del Movimiento Natural de la Población del INE, una tendencia que se intensifica en los municipios de menor tamaño (INE, s.f.). Este crítico hecho motivó que el Pleno del Senado aprobara en 2015 la creación de la Ponencia de estudio para la adopción de medidas en relación con la despoblación rural (Senado de España, 2015), y que, posteriormente, la VI Conferencia de Presidentes Autonómicos (2017) acordara impulsar de manera coordinada una Estrategia Nacional frente al Reto Demográfico, reconociendo explícitamente que la despoblación constituye uno de los principales desafíos territoriales de la España contemporánea (Gobierno de España, 2017).

En conjunto, estos elementos permiten entender la "España vaciada", más que como un simple problema de "poca gente en muchos kilómetros cuadrados", como el resultado de la interacción entre despoblación rural y desequilibrio territorial que podrían conducir a la extinción de cientos de pueblos. Sobre este marco conceptual se apoyará el análisis cuantitativo que desarrolla este trabajo en los siguientes capítulos.

Estado del arte

Si se investiga sobre los trabajos ya realizados hasta la fecha sobre este fenómeno, es notorio el creciente número de estudios que han ido apareciendo durante los últimos años. Los enfoques son variados y diversos, aunque la mayoría se centra en los aspectos sociológicos, políticos, históricos y económicos. Entre los estudios más conocidos y citados sobre el tema, que constituyen la base de la literatura académica sobre esta cuestión, destacan: los trabajos de Collantes y Pinilla, que abarcan un análisis histórico de la despoblación rural española durante más de un siglo, comparándolo con los casos de otros países europeos; los informes de Funcas, que examinan la evolución demográfica del interior peninsular y los cambios recientes en la distribución territorial de la población; y los documentos elaborados por el Ministerio para la Transición Ecológica y el Reto Demográfico, donde se proponen las principales estrategias y políticas públicas diseñadas para afrontar el reto demográfico.

Sin embargo, son pocos los trabajos que tienen un enfoque principalmente cuantitativo a la hora de abordar este problema. Entre ellos, los más resaltables por su naturaleza analítica son los siguientes.

Por un lado, *Growth and decline in rural Spain: an exploratory analysis* de Gómez Valenzuela & Holl (2023) es uno de los más completos y metodológicamente avanzados trabajos que usa datos municipales, análisis espacial y variables multidimensionales para estudiar el crecimiento y declive de las áreas rurales españolas empleando datos altamente desagregados. Proponen una nueva tipología de áreas rurales en declive y muestran que factores como la accesibilidad a zonas urbanas, la conectividad y la estructura económica son claves para explicar las diferencias internas dentro del medio rural, sentando precedente de análisis territorial avanzado.

Por otro lado, López-Penabad, Iglesias-Casal y Rey-Ares (2022) desarrollan un índice compuesto de desarrollo sostenible rural (RSDI) para municipios gallegos, integrando dimensiones económicas, demográficas, sociales y ambientales. Su enfoque permite

identificar municipios vulnerables y analizar los factores que explican su situación, proporcionando una herramienta cuantitativa útil para priorizar intervenciones públicas

Por tanto, se pueden encontrar estudios que buscan comprender la realidad de manera numérica sirviéndose de diferentes fuentes de datos para llevar a cabo análisis descriptivos. No obstante, existe un margen para incorporar análisis predictivos mediante un enfoque que combine de manera integrada múltiples variables demográficas, económicas, territoriales y de conectividad a escala municipal. La comunidad académica ya ha cubierto el análisis descriptivo para diagnosticar la situación actual, saber qué y por qué ocurrió, pero falta dar el siguiente paso para entender qué ocurrirá en el futuro con el fin de poder prescribir las mejores medidas para remediar esta tendencia demográfica.

Contribución del presente trabajo

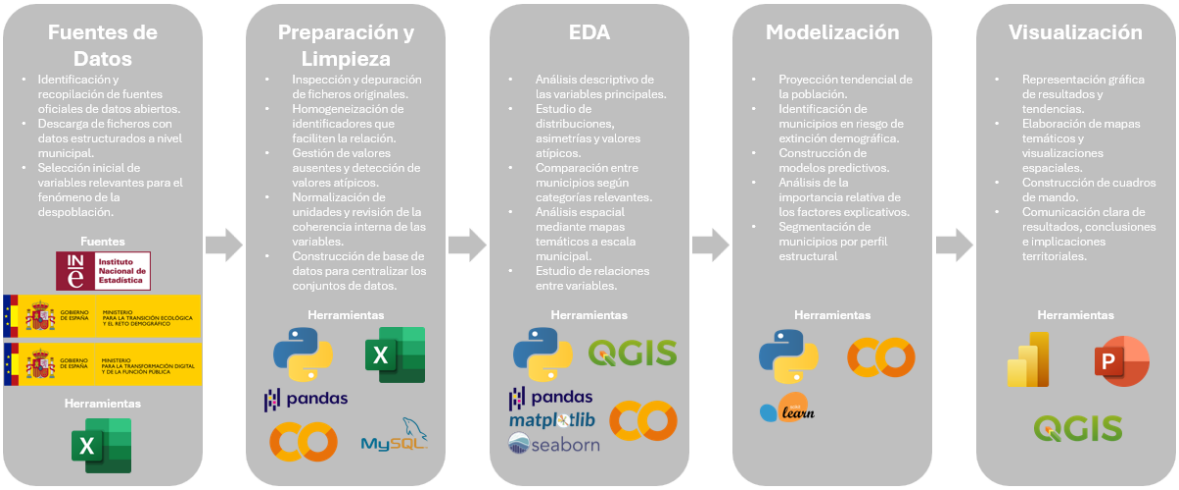
El presente trabajo aborda el fenómeno de la despoblación rural en España desde un enfoque cuantitativo basado en datos, de carácter analítico descriptivo y predictivo, aplicando las herramientas y metodologías propias del ámbito de ciencia de datos. A diferencia de los estudios tradicionales ya realizados de carácter político o sociológico, este análisis se centrará en la integración, tratamiento y modelización de información cuantitativa procedente de diversas fuentes oficiales, como el Instituto Nacional de Estadística (INE), el Servicio Público de Empleo Estatal (SEPE) o el Ministerio para la Transición Ecológica y el Reto Demográfico (MITECO), con el fin de identificar patrones y relaciones entre un rango diverso de variables demográficas, económicas, territoriales y de conectividad, entre otras.

Se emplearán técnicas estadísticas y de aprendizaje automático que permitan detectar los factores que explican la variación poblacional a nivel municipal, así como visualizar estos resultados mediante herramientas como Power BI, que faciliten una interpretación intuitiva de los datos.

El propósito de este trabajo no es formular propuestas políticas, sino aportar evidencia empírica que ayude a comprender mejor el fenómeno de la “España vaciada” y sirva como apoyo a la toma de decisiones basadas en datos. De este modo, el proyecto combina el rigor cuantitativo contribuyendo al conocimiento académico sobre el reto demográfico en España.

Metodología y diseño del análisis

Figura 10. *Diagrama a alto nivel del pipeline metodológico*



Elaboración propia.

Fuentes de datos y variables

Como trabajo de ciencia de datos con un fuerte componente cuantitativo, será especialmente relevante aclarar las fuentes de dónde se obtendrán los datos para su posterior uso. Al tratarse de un Trabajo de Fin de Grado se ha pretendido emplear fuentes oficiales para asegurar la fiabilidad de las conclusiones que alcance el análisis. Para ello, única y exclusivamente se han usado datos de administraciones públicas del Estado que ofrecen de manera abierta por internet. A continuación, se explicará cada fuente en concreto y la información que ofrece.

En primer lugar, el Instituto Nacional de Estadística (INE), al tratarse del organismo público español encargado de la coordinación general de los servicios estadísticos y ofrecer una gran variedad de datos de diferentes ámbitos. Entre ellos, para este trabajo es especialmente útil toda la documentación demográfica y poblacional, como las cifras oficiales de datos del padrón a nivel municipal, la relación de municipios y sus códigos de provincia, y la evolución de los censos de población a diferentes niveles de granularidad territorial según edad y sexo. Se han podido descargar varios ficheros en formato Excel (.xlsx) con datos de población a nivel provincial desde 1975 hasta 2025, datos de población a nivel municipal desde 1996 hasta 2024, y la evolución del código de cada municipio desde 1996 hasta la actualidad.

En segundo lugar, el Ministerio para la Transformación Ecológica y el Reto Demográfico (MITECO), al ser el órgano del Gobierno de España responsable del reto demográfico. Ha sido oportuno no solo por su clasificación de municipios según criterios de riesgo de despoblación de 2021, sino también porque ofrece un fichero de datos municipales con numerosas variables de distintas temáticas llamado Sistema Integrado de Datos Municipales (SIDAMUN), ambos en formato tipo Excel (.xlsx). Esta última en principio es una herramienta desarrollada por la Secretaría General para el Reto Demográfico que sólo

permite la visualización de dichos datos. No obstante, tras contactar con ellos y aclarar la finalidad académica de este trabajo, facilitaron rápidamente el fichero para su explotación. El archivo cuenta con 250 variables recientes de demografía, economía, geografía, servicios, vivienda y medioambiente que ha centralizado de diferentes organismos oficiales como el INE, Eurostat, Instituto Geográfico Nacional, Tesorería General de la Seguridad Social, Servicio Público de Empleo Estatal, Ministerio de Sanidad, Ministerio de Educación, Ministerio de Cultura, Banco de España, Dirección General de Tráfico, Correos, Ministerio para la Transformación Digital, entre otros (Secretaría General para el Reto Demográfico, s. f.).

En tercer lugar, el Ministerio para la Transformación Digital, al ser el órgano del Gobierno de España responsable de las telecomunicaciones del país. Este publica datos de conectividad a nivel municipal, con indicadores de red y cobertura, desde 2021 hasta 2024 en formato Excel también (.xlsx).

De todas las fuentes anteriores se han podido obtener numerosas variables a partir de los ficheros con datos estructurados mencionados. Dado el elevado número de variables disponibles, se ha realizado una selección previa atendiendo a criterios de relevancia teórica en la literatura, disponibilidad homogénea a nivel municipal, interpretabilidad y potencial capacidad explicativa, así como a la necesidad de evitar redundancias y problemas de multicolinealidad. Las variables consideradas se presentan a continuación:

Tabla 1. *Descripción de las variables seleccionadas*

Nombre	Tipo	Fuente	Año	Unidades	Descripción
Población municipal	Cuantitativa	INE	1996–2024	Habitantes	Población empadronada por municipio y año
Código INE	Categórica	INE (Nomenclátor)	1996–2024	—	Identificador único municipal
Población provincial	Cuantitativa	INE	1975–2025	Habitantes	Población registrada por provincia y año
Cobertura de acceso inalámbrico fijo	Cuantitativa	Ministerio para la Transformación Digital	2024	% de viviendas	Porcentaje de viviendas con acceso inalámbrico fijo, cobertura FTTH, cobertura HFC, cobertura ≥ 30 Mbps, cobertura ≥ 100 Mbps, cobertura ≥ 1 Gbps, cobertura 4G, cobertura 5G, cobertura 5G banda 3,5 GHz
Grado de urbanización (DEGURBA)	Categórica ordinal	Eurostat	2023	Categorías: Rural, Semiurbana, Urbana	Clasificación del municipio según grado de urbanización
Población total y por sexo	Cuantitativa	INE (SIDAMUN)	2024	Habitantes	Población residente total y desagregada por sexo
Edad media de la población (total y por sexo)	Cuantitativa	INE	2024	Años	Edad media de la población residente
Estructura por grupos de edad	Cuantitativa	INE	2024	% sobre total	Distribución de población (<16, 16–64, ≥65)
Ratios demográficos	Cuantitativa	INE	2024	%	Tasa de dependencia, índice de envejecimiento, ratio de masculinidad, mujeres en edad fértil
Pirámide de población por edad y sexo	Cuantitativa	INE	2024	% sobre total	Distribución por tramos quinquenales de edad y sexo
Variación absoluta y relativa de población (2003–2024)	Cuantitativa	INE	2003–2024	Habitantes y %	Cambio neto y porcentual de población en el periodo
Movimientos naturales de población	Cuantitativa	INE	2023	‰	Nacimientos, defunciones, crecimiento vegetativo y matrimonios por 1.000 habitantes
Nacionalidad de la población	Cuantitativa	INE	2024	%	Proporción de población española y extranjera
Lugar de nacimiento	Cuantitativa	INE	2024	%	Distribución según lugar de nacimiento (municipio, provincia, CCAA, extranjero)
Superficie, altitud y perímetro	Cuantitativa	IGN	2024	km^2 / m	Características físicas del municipio

Densidad de población	Cuantitativa	IGN, INE	2024	hab./km^2	Intensidad de ocupación del territorio
Entidades singulares	Cuantitativa	INE	2023	nº / %	Número de entidades y peso de la principal
Área costera	Binaria	Eurostat	2023	Sí / No	Municipio con salida al mar
Usos del suelo	Cuantitativa	IGN	2017	% superficie	Superficie agrícola, artificial, natural, humedales, agua y otros
Afiliación a la Seguridad Social	Cuantitativa	TGSS	Diciembre 2024	% / x1.000	Afiliados por 1.000 hab. en edad activa y por régimen (Régimen General, R.G. sector agrario, R.G.S.E. hogar, R.E. Mar, R.E.T. Autónomos, R.E. Minería del Carbón)
Desempleo	Cuantitativa	SEPE	Diciembre 2024	% / x1.000	Paro total, por sexo, edad y sector
Contratación laboral	Cuantitativa	SEPE	Diciembre 2024	x1.000	Contratos totales, por sexo, duración y sector
Rentas	Cuantitativa	INE	2022	€	Renta neta y bruta media por persona según fuente de ingreso (salario, pensiones, prestaciones por desempleo, otras prestaciones, otros ingresos) y hogar
Desigualdad de renta	Cuantitativa	INE	2022	Índices	Índice de Gini y ratio P80/P20
Pensiones contributivas	Cuantitativa	Seguridad Social	Julio 2025	nº / €	Número y cuantía media de pensiones
Empresas por sector	Cuantitativa	INE	2024	%	Distribución sectorial de empresas
Infraestructura sanitaria	Binaria / discreta	Ministerio de Sanidad	2025	Sí / nº	Consultorios, centros de salud y hospitales
Establecimientos sanitarios	Binaria / discreta	Ministerio de Sanidad	2025	Sí / nº	Farmacias, clínicas dentales y centros ópticos
Transporte y movilidad	Cuantitativa	DGT	2024	x1.000	Parque de vehículos por 1.000 habitantes
Servicios financieros y postales	Binaria / discreta	Banco de España / Correos	2024/2022	Sí / nº	Oficinas bancarias y de correos
Infraestructura educativa	Binaria / discreta	Ministerio de Educación	2023	Sí / nº	Centros educativos por nivel y CRA
Equipamientos culturales	Binaria	Ministerio de Cultura	2025	Sí / No	Disponibilidad de biblioteca
Accesibilidad a servicios	Cuantitativa	IGN	2024	Minutos	Tiempo a núcleos urbanos, hospital y vías principales
Parque de viviendas	Cuantitativa	INE	2021	nº	Número total de viviendas
Tipología de viviendas	Cuantitativa	INE	2021	%	Viviendas principales y no principales
Tamaño medio del hogar	Cuantitativa	INE	2021	Personas	Tamaño medio de los hogares
Hogares unipersonales	Cuantitativa	INE	2021	%	Proporción de hogares unipersonales
Vivienda turística	Cuantitativa	INE	2024	nº / x100	Establecimientos y plazas turísticas

Elaboración propia.

Preparación y limpieza de datos

Para pasar de todas esas fuentes dispersas y heterogéneas a una base consolidada con las variables listas para el análisis, se deberán acometer los siguientes pasos de ingeniería de datos:

Primero, después de descargar todos los ficheros habrá que relacionarlos entre sí para conseguir centralizar los datos en una misma base de datos. En el caso de este trabajo, a partir de todos estos se pueden conseguir variables de diferente índole. Por un lado, a partir del padrón municipal, información de tipo panel con datos de población para municipios observadas a lo largo de un rango temporal. Por otro lado, a partir del SIDAMUN y del MTDFP (Ministerio para la Transformación Digital y Función Pública), de tipo de corte transversal con numerosos datos diferentes para cada municipio en un momento temporal concreto. Para conseguirlo, usando Python y Excel, se deberán eliminar encabezados innecesarios y variables no relevantes.

Segundo, la unidad de observación es el municipio y será crucial encontrar una clave de identificación homogénea que nos permita relacionar las diferentes variables de cada municipio entre las diferentes tablas con las que se cuenta en todos los ficheros descargados. El nombre del municipio no será una buena opción, ya que no solo puede haber diferencias entre tablas por signos de exclamación, sino que también pueden estar escritos en diferentes idiomas, como aquellos que pertenezcan a regiones de España que cuentan con una lengua cooficial. Por ejemplo, el municipio de Sangenjo se podría llamar Sanxenxo en gallego, y dificultar la unión de información entre diferentes fuentes. Por ello, el código de 5 dígitos establecido por el INE para cada municipio del *Registro de Entidades Locales (REL)* será una mejor opción, al tratarse de un identificador distintivo de cada municipio donde los 2 primeros dígitos indican la provincia y los 3 últimos el municipio concreto. No obstante, se debe tener en cuenta que durante el periodo temporal estudiado ha habido variaciones debido a secesiones y uniones de municipios. Cada modificación se deberá tener en cuenta para homogeneizar la información y que en la serie temporal no haya interrupciones porque un municipio hubiese desaparecido por segregarse en dos diferentes o fusionarse a otro. Para ello, en caso de que desapareciera por segregarse en dos municipios diferentes, se sumarán las poblaciones de estos dos para el municipio original suponiendo que no se hubiesen separado. En caso de desaparecer por unirse entre dos municipios diferentes, se supondrá que se hubiesen unido al principio del periodo y se sumará la población en los años en que fueron diferentes. Se deberá estudiar también la relevancia de este fenómeno según el número de municipios del total afectado por estas modificaciones.

Tercero, el tratamiento de variables para dejarlas listas para el análisis posterior. Para ello, habrá que ver: qué hacer con valores ausentes (NaNs) y atípicos (outliers), homogeneizar unidades y la revisión de la coherencia interna de las variables.

Una vez hecho todo eso, se podrán consolidar los ficheros en MySQL. A continuación se puede observar una figura que refleja esquemáticamente mediante un diagrama entidad-relación la estructura final de la base de datos con campos representativos y la vinculación entre tablas a través del código INE.

Figura 11. *Diagrama simbólico de la base de datos*

Elaboración propia.

Análisis exploratorio de datos (EDA)

Una vez preparada y consolidada la base de datos, y con carácter previo a la estimación de cualquier modelo, se lleva a cabo un análisis exploratorio de los datos (Exploratory Data Analysis, EDA) para comprender la distribución de las variables, detectar patrones preliminares y evaluar posibles problemas que puedan afectar al análisis posterior. Se emplearán librerías específicas de Python, como matplotlib o seaborn, y QGIS, herramienta potente para visualizaciones geográficas.

En primer lugar, mapas temáticos para visualizar la geografía española a nivel municipal según diferentes variables como la densidad de población, el envejecimiento, la pérdida o crecimiento de población, y otros indicadores socioeconómicos, lo que permite contextualizar los fenómenos observados desde una perspectiva espacial.

En segundo lugar, se estudiará la distribución de las variables clave mediante el uso de histogramas y diagramas de caja, según su naturaleza, comparándolas por categorías como el grado de urbanización o el tamaño poblacional. Además, se calcularán estadísticos descriptivos básicos, como la media, mediana y varianza, entre otros. Todo esto servirá para comprender en profundidad cada variable y se podrán identificar valores atípicos.

Por último, se analiza la relación entre las variables numéricas mediante el cálculo de una matriz de correlaciones, representada a través de un mapa de calor. Este análisis permite identificar posibles relaciones lineales entre variables y detectar problemas de colinealidad que puedan condicionar la selección de variables y la especificación de los modelos analíticos posteriores.

Modelo predictivo o analítico

Una vez se cuente con una profunda comprensión preliminar de los datos, se podrán usar técnicas estadísticas y de aprendizaje automático que nos ayuden a proyectar la evolución futura de la población, explicar la despoblación y detectar patrones estructurales entre municipios. Se programará con Python, concretamente con librerías habituales para Machine Learning como scikit-learn, para construir los modelos tanto de aprendizaje supervisado como no supervisado. Además, se usará Google Colab, un servicio gratuito en la nube que te permite ejecutar código en máquinas virtuales de Google, mejorando la capacidad de procesamiento de mi ordenador personal y acelerando resultados.

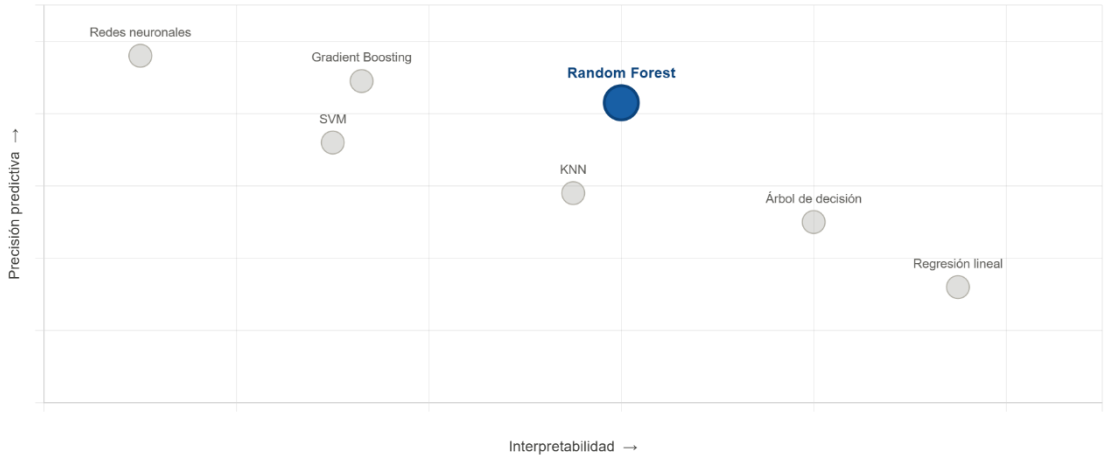
El primer modelo servirá para estimar la proyección de población durante las próximas décadas, y se alimentará del registro histórica de población desde 1996 hasta 2024 y de la estructura demográfica actual (2024). Se podrán utilizar sus resultados para identificar el año en el que cada municipio caerá por debajo de un determinado umbral. Así, se sabrá el año en el que se extinguirán los municipios si continúan el curso reciente.

Para ello, se comenzará ajustando un modelo de regresión lineal por municipio para capturar su trayectoria propia. Después, como segunda derivada de esta aproximación, se descontará el efecto de la estructura demográfica actual añadiendo un ajuste adicional para que los municipios más envejecidos tengan una mayor pendiente de despoblación. No se emplean series temporales ya que no se busca recoger ninguna estacionalidad, simplemente la tendencia a largo plazo.

El segundo modelo tiene como objetivo explicar la variación porcentual de población de cada municipio e identificar los factores con mayor capacidad explicativa del fenómeno de la despoblación. Para ello se construye un Random Forest Regressor, algoritmo que combina múltiples árboles de decisión y promedia sus predicciones, lo que le permite capturar relaciones no lineales entre variables con una precisión superior a los modelos lineales, manteniendo al mismo tiempo la capacidad de extraer medidas de importancia de variables interpretables. La Figura 12 ilustra el equilibrio que ofrece este algoritmo entre precisión predictiva e interpretabilidad en comparación con otras alternativas de aprendizaje automático.

El modelo se entrena sobre el 80% de los municipios disponibles, reservando el 20% restante como conjunto de test para evaluar su capacidad de generalización a datos no vistos durante el entrenamiento. El rendimiento se evalúa mediante métricas estándar de regresión: el coeficiente de determinación R^2, que indica la proporción de varianza explicada por el modelo, y el error cuadrático medio (RMSE), que cuantifica el error de predicción en las mismas unidades que la variable objetivo. Una vez entrenado y validado, el modelo se emplea para realizar simulaciones que permiten estimar el impacto que habría tenido sobre la variación poblacional predicha de un municipio concreto la modificación de alguna de sus variables accionables, como la apertura de un consultorio médico o la mejora de la cobertura de banda ancha. Esta capacidad convierte el modelo en una herramienta de apoyo a la toma de decisiones, útil para que los organismos administrativos puedan priorizar intervenciones en función de su impacto estimado sobre la dinámica demográfica de los territorios más vulnerables.

Figura 12. *Mapa de modelos de aprendizaje automático: equilibrio entre interpretabilidad y precisión predictiva*



Elaboración propia.

El tercer modelo se empleará para reconocer patrones estructurales entre municipios que comparten similitudes, ayudando a comprender con mayor profundidad el fenómeno de la "España vaciada". Se diseñará un K-means y un clustering jerárquico que permita agrupar municipios según características socioeconómicas y demográficas, y ver su comportamiento. Por ejemplo, municipios pequeños pero turísticos, municipios dinámicos semiurbanos del extrarradio de grandes ciudades, etc.

Visualizaciones y herramientas interactivas

Para construir las visualizaciones que ayuden a comprender mejor el fenómeno estudiado en este trabajo, se realizarán desde simples gráficos hasta mapas geográficos, cualquier ilustración que sirva para comunicar con efectividad y claridad no solo los resultados y conclusiones alcanzados durante el desarrollo técnico, sino también las causas, consecuencias y contexto actual del problema.

Se usarán herramientas como librerías de visualización de Python, como matplotlib y seaborn, Power BI, al ser una herramienta gratuita muy completa que ofrece la posibilidad de diseñar una gran variedad de gráficos, y PowerPoint, como soporte para presentación visual. En caso de necesitar una herramienta más potente para crear mapas, QGIS es una excelente opción para crear ilustraciones espaciales.

Desarrollo técnico y modelización

Este capítulo recoge el proceso técnico completo que va desde la carga y preparación de los datos hasta la construcción e interpretación de los modelos analíticos. Cada etapa sigue el pipeline metodológico descrito en el capítulo anterior, y el código completo se encuentra disponible en el repositorio de GitLab referenciado en los anexos.

Carga inicial de datos

El punto de partida del análisis técnico es la carga de los tres conjuntos de datos descritos en el apartado de fuentes: el Sistema Integrado de Datos Municipales (SIDAMUN) del MITECO, el padrón municipal histórico del INE y el fichero de conectividad del Ministerio para la Transformación Digital, todos ellos en formato Excel (.xlsx). Para dejarlos en condiciones de ser analizados fue necesario aplicar algunos retoques previos, dado que los tres presentaban encabezados irregulares con metadatos y cabeceras multinivel en las primeras filas. En cada caso se identificó manualmente la fila con los nombres reales de las variables, se reasignó como encabezado y se normalizaron los nombres de columna eliminando saltos de línea y espacios redundantes. Sobre cada fichero se realizó la selección de variables detallada en la sección de fuentes de datos y variables.

En el caso del padrón municipal, que contiene la evolución de la población de cada municipio desde 1996 hasta 2024 en formato ancho con un año por columna, fue necesario realizar una armonización adicional para garantizar la coherencia de la serie temporal. Durante el período analizado se han producido en España fusiones y segregaciones de municipios que, de no tratarse, generarían discontinuidades artificiales en las series. En concreto, este fenómeno ha afectado a 42/8132 municipios, menos de un 0,6% del conjunto. El criterio adoptado, descrito en detalle en el apartado de preparación y limpieza de datos de la metodología, consiste en tratar los municipios fusionados como si lo hubieran estado desde el inicio del período, y los segregados como si nunca se hubieran separado, de modo que la unidad de observación permanezca estable a lo largo de toda la serie.

Preparación y limpieza de datos

Con los tres ficheros cargados, se acomete la fase de preparación y limpieza con el objetivo de construir una base de datos unificada, coherente y lista para el análisis. Esta fase abarca la identificación y transformación de tipos de variables, el tratamiento de valores ausentes y la integración de los conjuntos de datos mediante el código INE como clave relacional.

Tipos de variables y transformaciones

El primer paso consiste en revisar el tipo asignado por defecto a cada variable tras la carga. En el caso del SIDAMUN, un número significativo de variables numéricas son interpretadas como texto debido a que los valores ausentes se codifican con símbolos no estándar, en este caso, con asterisco (*), en lugar de celdas vacías. Antes de proceder a la conversión de tipos, estos símbolos se sustituyen de forma sistemática por valores nulos para que el tratamiento posterior sea homogéneo. Para las variables de naturaleza cualitativa se aplican las siguientes transformaciones: las variables binarias codificadas como SI/NO (área costera, disponibilidad de biblioteca y pertenencia a un Centro Rural Agrupado) se

recodifican como variables dicotómicas (1/0), y la variable de grado de urbanización DEGURBA, de naturaleza ordinal, se transforma en una escala numérica ordenada (1 = urbano, 2 = semiurbano, 3 = rural), preservando la jerarquía territorial entre categorías.

Valores ausentes

Tras la normalización de tipos, se realiza un análisis sistemático de los valores ausentes por variable para determinar el tratamiento más adecuado en cada caso. La siguiente tabla ilustra la distribución del porcentaje de nulos por variable que había en las fuentes originales, ordenada de mayor a menor, con un porcentaje mayor al 5%.

Tabla 2. *Distribución del porcentaje de nulos por variable*

Variable	Nº municipios	% municipios
Nacidos en España en la misma CCAA (% hab. s/ total)	8132	100 %
Nacidos en España en la misma CCAA y misma provincia (% hab. s/ total)	8132	100 %
Empresas Sector Industria (% s/ Total)	5153	63,35 %
Empresas Sector Construcción (% s/ Total)	5132	63,10 %
Empresas Resto Sector Servicios (% s/ Total)	5008	61,57 %
Empresas Comercio, transporte y hostelería (% s/ Total)	4989	61,34 %
Renta bruta media por persona. Fuente de ingreso: Otras prestaciones (% s/ Total)	1807	22,21 %
Renta bruta media por persona. Fuente de ingreso: Prestaciones por desempleo (% s/ Total)	1807	22,21 %
Renta bruta media por persona. Fuente de ingreso: Pensiones (% s/ Total)	1807	22,21 %
Renta bruta media por persona. Fuente de ingreso: Salario (% s/ Total)	1556	19,13 %
Renta bruta media por persona. Fuente de ingreso: Otros ingresos (% s/ Total)	1556	19,13 %
Índice de Gini (%)	1361	16,73 %
Distribución de la renta P80/P20	1361	16,73 %
Total Empresas	1243	15,28 %

Elaboración propia.

Se establece el criterio de corte en el 5%, umbral habitual en estadística para distinguir entre ausencias que comprometen la fiabilidad de una imputación y aquellas que pueden ser tratadas sin distorsión significativa. Las variables con más de un 5% de valores ausentes se eliminan del análisis, al considerarse que imputar una proporción tan elevada introduciría un sesgo inaceptable en las distribuciones. Para las variables con una proporción residual de nulos igual o inferior al 5% se aplica imputación por la media, dado que el reducido volumen de observaciones afectadas garantiza un impacto mínimo sobre los estadísticos del conjunto.

Con carácter previo a la aplicación de este criterio general, se estudia caso por caso la causa de los valores ausentes, lo que permite identificar situaciones en las que el tratamiento estándar no es el más adecuado. El índice de envejecimiento presenta valores ausentes en municipios sin población menor de 16 años, ya que el indicador se define como el cociente entre la población mayor de 65 años y la menor de 16, lo que implica una división por cero en estos casos. Lejos de ser una ausencia de información, estos nulos reflejan una situación de envejecimiento extremo, por lo que se sustituyen por el valor máximo observado del índice en la muestra. Las variables de tiempo de acceso al municipio más cercano de más de 20.000 y 50.000 habitantes, y el tiempo a la autopista o autovía más cercana, presentan valores ausentes en municipios de determinadas islas donde no existen ni municipios de ese tamaño ni autopistas. En estos casos la ausencia refleja la inexistencia del servicio, por lo que se imputa igualmente el valor máximo observado en la muestra, indicando que el acceso a dicho servicio es máximamente difícil o inexistente. Por último, durante el análisis se detecta un municipio con valores ausentes

en la práctica totalidad de sus variables, sin que exista una causa técnica o geográfica que lo justifique. Dado que la imputación de un municipio del que se desconoce casi toda su información introduciría observaciones, se opta por eliminarlo del dataset. Se trata de un caso marginal y su exclusión no compromete la representatividad del análisis.

En el caso del padrón municipal histórico, se detecta que el año 1997 carece completamente de observaciones. Aunque se podría haber optado por una interpolación entre 1996 y 1998, se decide eliminar los dos primeros años de la serie al encontrarse en el extremo inicial de la serie, iniciando el análisis en 1998 y manteniendo un rango temporal de 26 años suficientemente amplio para capturar las tendencias estructurales de largo plazo. Se identifican y eliminan asimismo los municipios sin ningún registro de población en toda la serie, que corresponden a un par de casos residuales sin representación en el padrón durante el período analizado.

Integración de los datasets y relativización de variables

La integración de los tres conjuntos de datos se realiza mediante el código INE de cinco dígitos como clave relacional, garantizando la unicidad del cruce y evitando los problemas de concordancia que surgirían al usar el nombre del municipio, especialmente en regiones con lengua cooficial. Antes del análisis exploratorio y la modelización, las variables de conteo absoluto (número de centros sanitarios, educativos, bancarios, etc.) se relativizan expresándolas como tasas por cada 1.000 habitantes, lo que hace comparables entre sí municipios de tamaños muy diferentes y elimina diferencias triviales atribuibles únicamente al tamaño poblacional y no al nivel real de dotación de servicios.

Análisis exploratorio de datos (EDA)

Con la base de datos limpia e integrada, se realiza un análisis exploratorio para comprender la distribución de las variables, detectar patrones preliminares y evaluar posibles problemas que puedan condicionar la modelización posterior. El análisis completo cubre las variables de los tres datasets; en este capítulo se presentan los resultados más relevantes, mientras que el código con la exploración exhaustiva está disponible en el repositorio de GitLab que se encuentra en los anexos.

Estadísticos descriptivos

Se calculan los principales estadísticos descriptivos (media, mediana, desviación típica, mínimo, máximo y percentiles 25 y 75) para todas las variables numéricas. La siguiente tabla recoge los estadísticos de las variables más relevantes para el fenómeno estudiado.

Tabla 3. *Tabla de estadísticos descriptivos de las variables clave*

Dimensión	Variable	Media	Mediana	Desv. típica	Mín.	P25	P75	Máx.
Demografía	Índice envejecimiento (%)	941,28	276,92	2.061,7	0	153,39	625	9.400
Origen	Nacidos en el extranjero (% hab. s/ total)	10,58	8,53	8,53	0,00	4,6	14,52	82,28

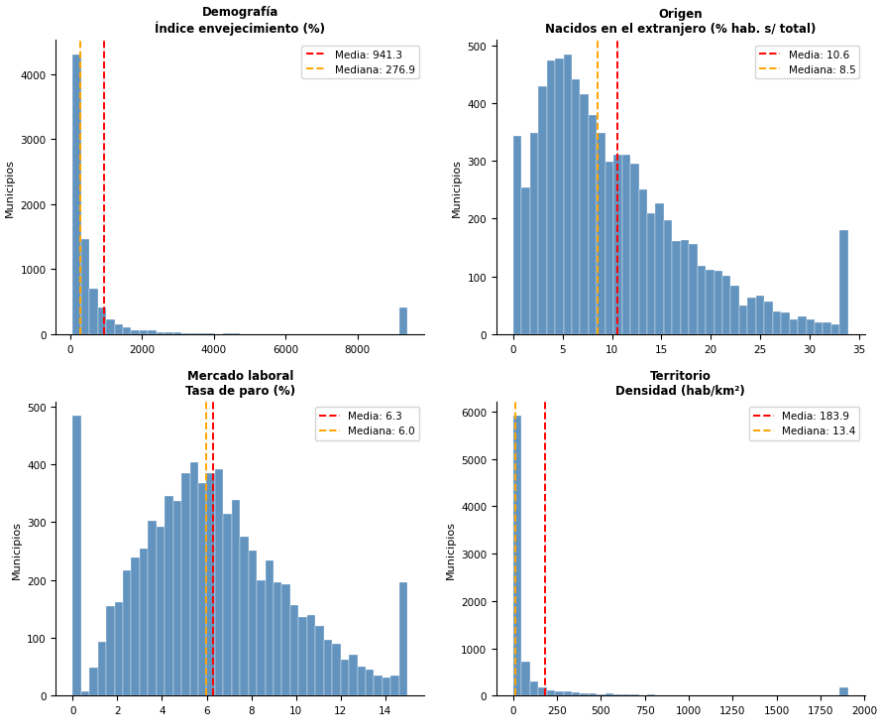
Dinámica	Crecimiento vegetativo x c/ 1.000 hab.	−10,15	−6,84	13,79	−150	−15,41	−1,01	83,33
Economía	Renta neta media por persona (€)	13.489,54	13.382	2.510,27	6.274	11.675	15.084	29.258
Mercado laboral	Tasa de paro (%)	6,29	5,96	3,8	0	3,84	8,33	100
Territorio	Densidad (hab/km^2)	183,9	13,35	938,31	0,22	4,58	57,07	26.877,13
Servicios	Tiempo hospital más cercano (minutos)	30,46	27,82	17,34	0	17,4	40,63	108,58
Conectividad	FTTH (% de viviendas)	0,74	0,87	0,31	0	0,68	0,94	1

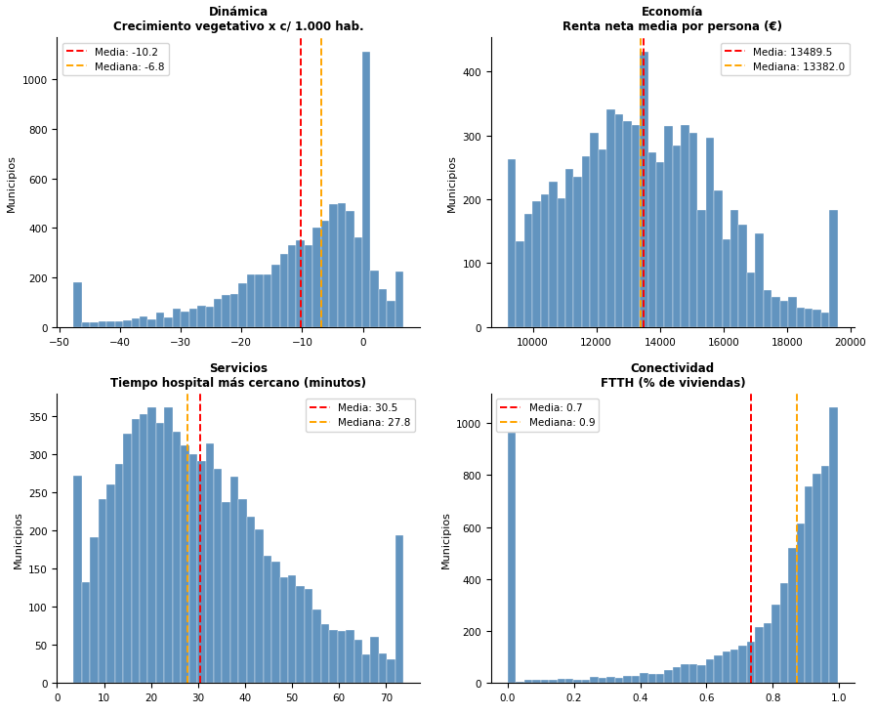
Elaboración propia.

Distribución de las variables

Se analizan las distribuciones de las variables clave mediante histogramas para comprender su forma, identificar asimetrías y detectar la presencia de valores extremos. La Figura 13 recoge, a modo de ejemplo representativo, la distribución de ocho variables relevantes de distintas dimensiones del análisis donde se puede apreciar asimetría y heterogeneidad entre municipios por la diferencia entre media y mediana.

Figura 13. *Distribución de variables clave por municipio (P2-P98, outliers extremos no mostrados)*

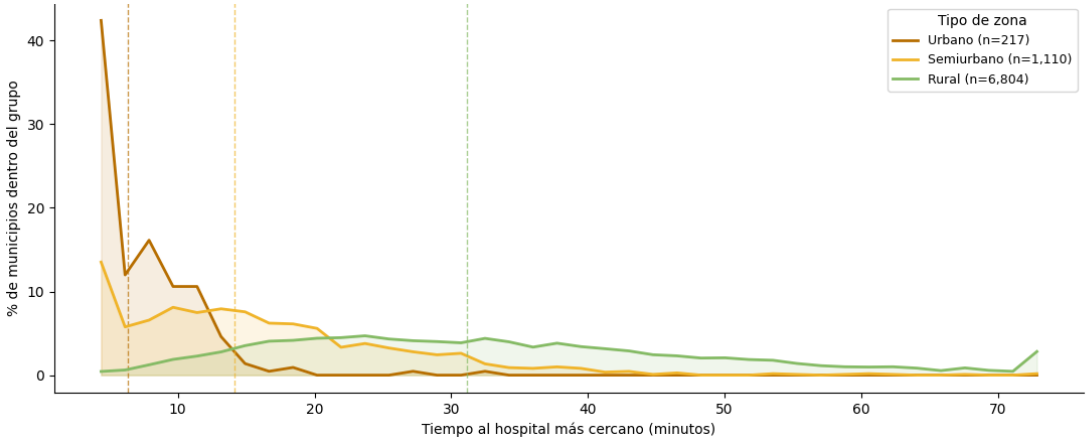




Elaboración propia.

Para profundizar en las diferencias estructurales entre tipos de municipio, se compara la distribución de las variables clave según el grado de urbanización (DEGURBA). La Figura 14 ilustra este análisis para la variable de tiempo hasta el hospital más cercano, una de las variables que refleja mayor diferencia entre tipo de municipio, ya que los municipios rurales son los más perjudicados en comparación con los urbanos.

Figura 14. *Distribución del tiempo al hospital más cercano por DEGURBA (P2-P98, outliers extremos no mostrados)*

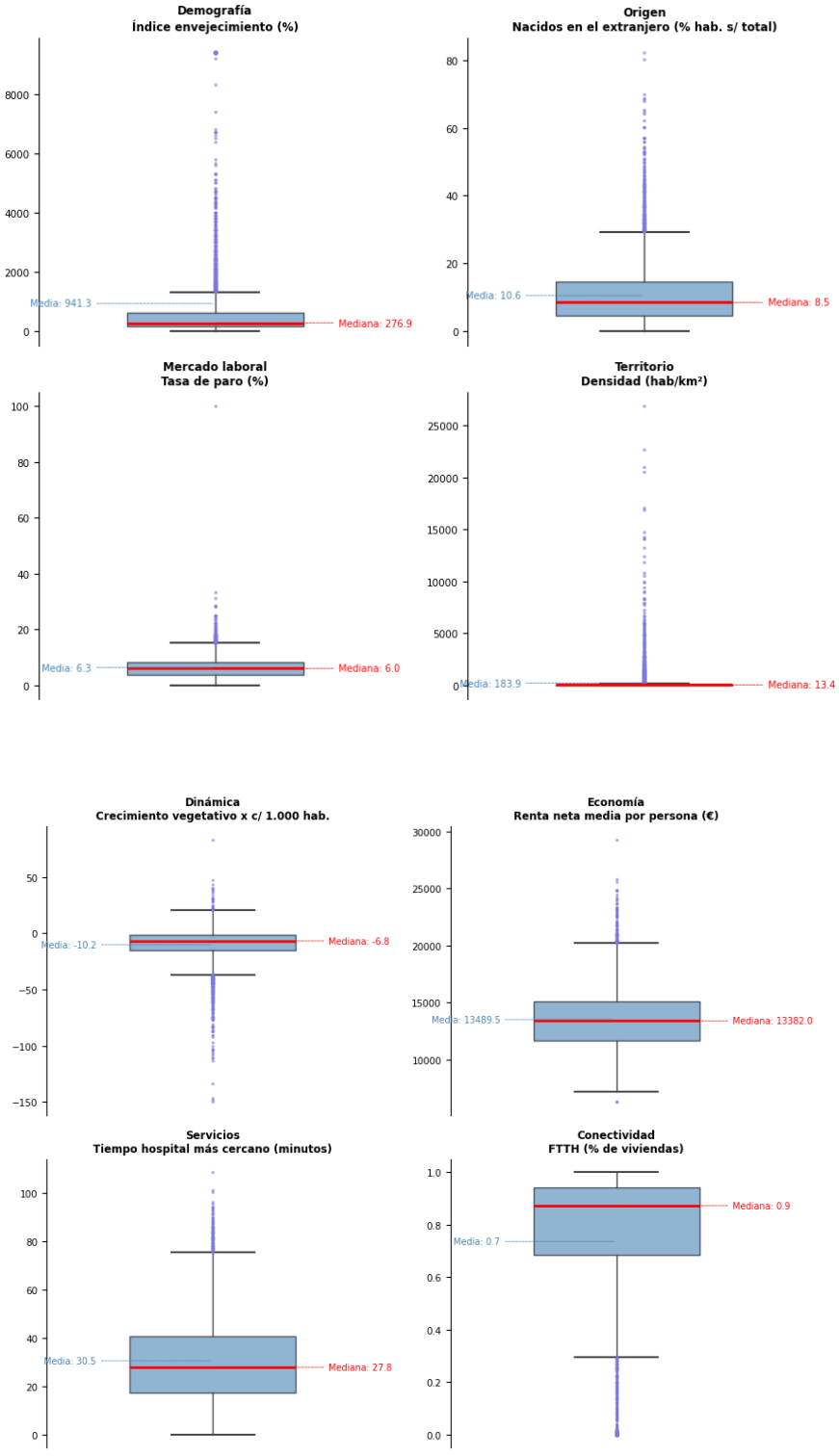


Elaboración propia.

Detección de valores atípicos

Se generan diagramas de caja para todas las variables con el objetivo de identificar valores atípicos. La mayor parte de los indicadores presentan outliers significativos en los extremos, como se observa en la Figura 15, lo que es esperable dada la heterogeneidad del territorio español: el abanico entre los municipios más pequeños y más aislados del interior y los grandes núcleos urbanos y de las costas es enorme.

Figura 15. *Diagrama de bigotes de variables clave por municipio*



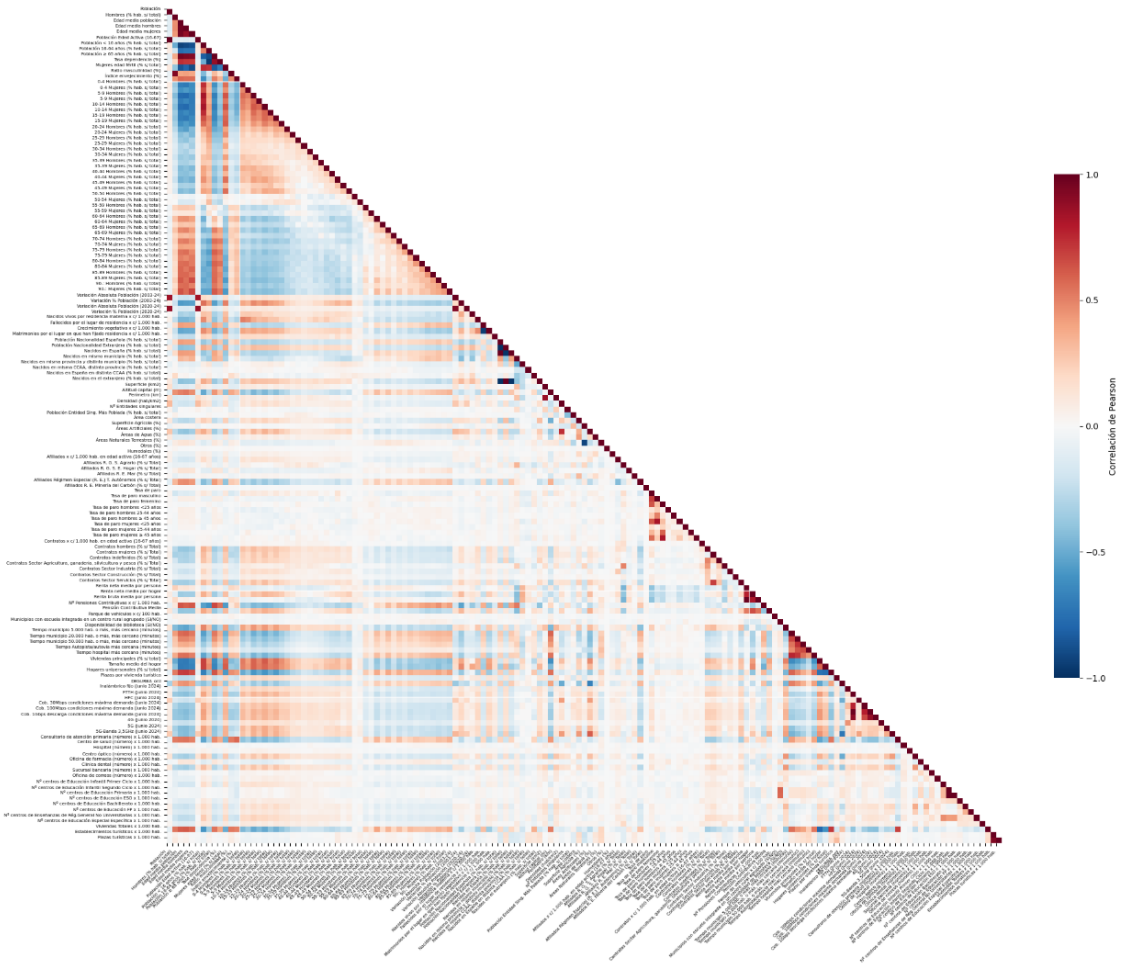
Elaboración propia.

A pesar de la presencia de valores extremos, se toma la decisión de mantener todas las observaciones en el análisis. La razón principal es que estos valores no son errores de medición sino realidades geográficas y demográficas legítimas: un municipio con un índice de envejecimiento de 9.400 existe y su situación es precisamente la que el análisis busca capturar. Eliminar los outliers equivaldría a excluir del estudio los casos más extremos de despoblación o de ciudades urbanas, que son los más relevantes para el objetivo del trabajo.

Análisis de correlaciones

Se calcula la matriz de correlaciones de Pearson entre todas las variables numéricas del SIDAMUN para identificar relaciones lineales entre ellas y detectar posibles problemas de multicolinealidad antes de la modelización. La Figura 16 muestra el mapa de calor de la matriz completa.

Figura 16. *Heatmap de la matriz de correlaciones*



Elaboración propia.

Del análisis se identifican los pares de variables con correlación superior a 0,90, que son candidatos a la eliminación en la fase de modelización para evitar que la importancia se distribuya artificialmente entre variables redundantes. Este umbral conservador permite eliminar las redundancias más evidentes preservando al mismo tiempo la riqueza informativa del conjunto.

Modelización

Una vez comprendida la estructura de los datos y las relaciones entre variables, se procede a la construcción de los tres modelos analíticos. El primero proyecta la evolución futura de la población municipal para estimar cuándo podrían extinguirse los municipios en declive. El segundo, un Random Forest Regressor, identifica los factores que mejor explican la variación histórica de población. El tercero agrupa los municipios en perfiles estructurales mediante técnicas de clustering.

Proyección de población municipal (2025-2050)

El primer modelo tiene como objetivo proyectar la evolución futura de la población de cada municipio y estimar el año en el que podría caer por debajo de un umbral crítico, fijado en 0 habitantes como escenario de extinción demográfica, bajo el supuesto de que se mantienen las tendencias recientes observadas (desde 1998 hasta 2024). El propósito no es una predicción precisa del futuro, sino la construcción de escenarios coherentes y comparables que permitan identificar los territorios en situación de mayor vulnerabilidad.

Se construyen dos escenarios complementarios:

Escenario base

Para cada municipio *i* se estima individualmente el siguiente modelo de regresión lineal sobre la serie histórica del padrón municipal (1998-2024):

$$Población_{it} = \beta_{0i} + \beta_{1i} \cdot t + \varepsilon_i$$

donde *t* es el año y β_{1i} es la pendiente, es decir, el cambio medio anual de población en habitantes por año. La proyección futura se obtiene evaluando el modelo en los años 2025-2050, imponiendo una cota inferior de cero habitantes. Se elige regresión lineal porque el objetivo no es capturar estacionalidad sino la tendencia estructural de largo plazo, y porque su interpretabilidad la hace útil para la comunicación de resultados.

Escenario ajustado

El escenario base presenta una limitación importante: asume implícitamente que todos los municipios mantienen la misma capacidad de reproducción demográfica, independientemente de su estructura por edades. Dos municipios con la misma pendiente histórica pueden tener perspectivas muy distintas si uno cuenta con una población relativamente joven y el otro presenta un envejecimiento severo. Para corregir esta limitación se introduce un ajuste demográfico que intensifica la pendiente de declive en los

municipios más envejecidos, utilizando el índice de envejecimiento como indicador del potencial de reemplazo generacional. La pendiente ajustada se calcula como:

$$\beta_{1i\_adj} = \beta_{1i} \cdot (1 + \alpha \cdot (IE\_cap_i / 100 - 1))$$

donde IE_cap_i es el índice de envejecimiento del municipio *i* capado a 300 para evitar distorsiones en municipios muy pequeños y envejecidos, y $\alpha = 0{,}5$ es un parámetro de intensidad deliberadamente conservador que reconoce que el modelo no incorpora explícitamente natalidad ni migraciones. El ajuste solo se aplica a municipios con pendiente histórica negativa. La proyección ajustada se obtiene evaluando:

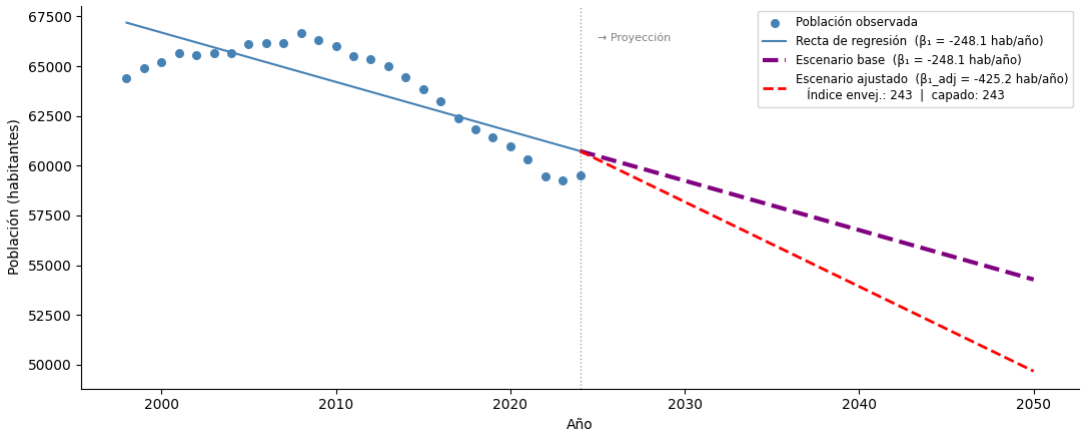
$$Población_{it} = P_{i,2024\_reg} + \beta_{1i\_adj} \cdot (t - 2024)$$

donde $P_{i,2024\_reg}$ es la población predicha por la regresión en 2024, garantizando la continuidad entre la recta histórica y la proyección futura.

Resultados

En la siguiente figura se puede ver en un scatter plot cómo actúa la regresión en el caso de un municipio concreto, Zamora, observando tanto las cifras de población histórica observadas, la regresión, y la proyección de ambos escenarios hasta 2050.

Figura 17. *Scatter plot con proyección de población para Zamora según regresión del escenario base vs ajustado*



Elaboración propia.

Los resultados espaciales y el detalle de los municipios con mayor urgencia demográfica según ambos escenarios se presentan en el capítulo de resultados.

Modelo explicativo: Random Forest Regressor

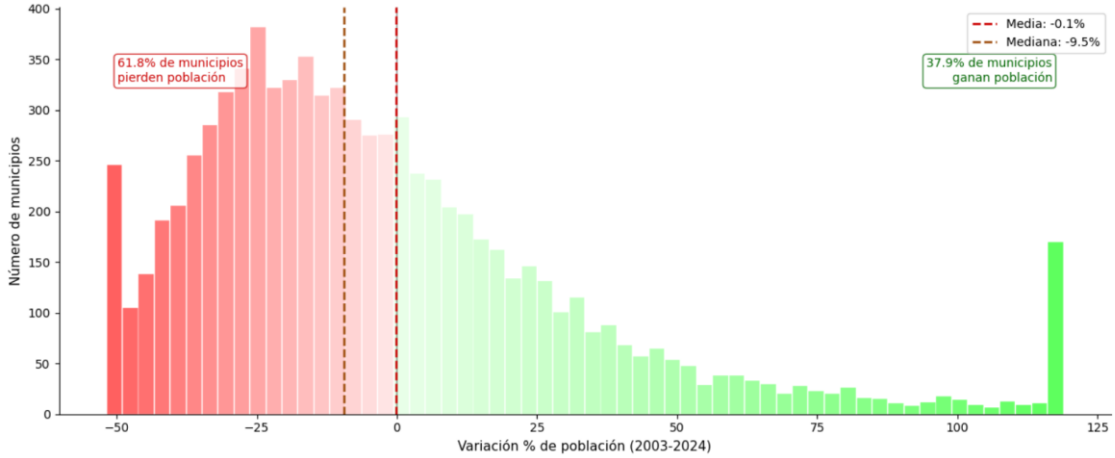
El segundo modelo tiene un doble propósito: explicar la variación porcentual de población de cada municipio en el período 2003-2024 y, sobre todo, identificar los factores con mayor capacidad explicativa del fenómeno de la despoblación. La elección de un Random Forest Regressor responde a tres razones principales: su capacidad para capturar relaciones no

lineales entre variables, su robustez frente a la multicolinealidad, y la posibilidad de extraer medidas de importancia de variables directamente interpretables.

Variable objetivo y variables explicativas

La variable objetivo es la variación porcentual de población entre 2003 y 2024, que captura el cambio estructural de largo plazo de cada municipio. Se elige este período porque 2003 es el primer año con cobertura completa en el SIDAMUN. La Figura 18 muestra su distribución, que presenta una marcada asimetría negativa: la mayoría de los municipios pierden población, con una mediana negativa.

Figura 18. *Distribución de la variable objetivo — variación porcentual de población (2003-2024)*



Elaboración propia.

Las variables explicativas se seleccionan siguiendo dos criterios: relevancia teórica contrastada en la literatura sobre despoblación y disponibilidad homogénea a nivel municipal. El conjunto inicial comprende 76 variables de dimensiones demográficas, de dinámica natural y origen de la población, territoriales, de mercado laboral, económicas, de accesibilidad y servicios, de conectividad digital y de vivienda.

Cabe señalar, no obstante, una limitación relevante en la interpretación del modelo: la mayoría de las variables explicativas corresponden a mediciones recientes, en torno a 2024, mientras que el target captura la variación acumulada entre 2003 y 2024. Esta asimetría temporal implica que el modelo no predice el futuro, sino que explica las asociaciones observadas al final del período entre el estado actual de los municipios y el cambio demográfico acumulado. No se dispone de series históricas completas para todas las variables a nivel municipal, lo que impide construir un modelo con variables y target alineados temporalmente. Por ello el enfoque adoptado es estrictamente explicativo del cambio 2003-2024, no predictivo de la evolución futura.

Sobre esta selección se aplica un filtro de multicolinealidad con umbral 0,90: de cada par con correlación superior a ese valor, se elimina la variable con mayor correlación media con el resto del conjunto. Se identifican 8 pares con correlación superior al umbral, lo que conduce a la eliminación de 8 variables, detalladas en la siguiente tabla. El conjunto final consta de 68 variables explicativas.

Tabla 4. *Variables eliminadas por alta colinealidad (correlación ≥ 0,90). Para cada par se elimina la variable con mayor correlación media con el resto del conjunto*

Variable retenida	Variable eliminada	Correlación
Nacidos en el extranjero (% hab. s/ total)	Población Nacionalidad Extranjera (% hab. s/ total)	1,00
Nº centros de Ed. Infantil Segundo Ciclo x 1.000 hab.	Nº centros de Educación Primaria x 1.000 hab.	0,99
FTTH (junio 2024)	Cob. 100Mbps condiciones máx. demanda (junio 2024)	0,98
Edad media población	Población ≥ 65 años (% hab. s/ total)	0,96
Fallecidos por lugar de residencia x c/ 1.000 hab.	Crecimiento vegetativo x c/ 1.000 hab.	0,95
Ratio masculinidad (%)	Hombres (% hab. s/ total)	0,93
Edad media población	Población < 16 años (% hab. s/ total)	0,92
Superficie Agrícola (%)	Áreas Naturales Terrestres (%)	0,91

Elaboración propia.

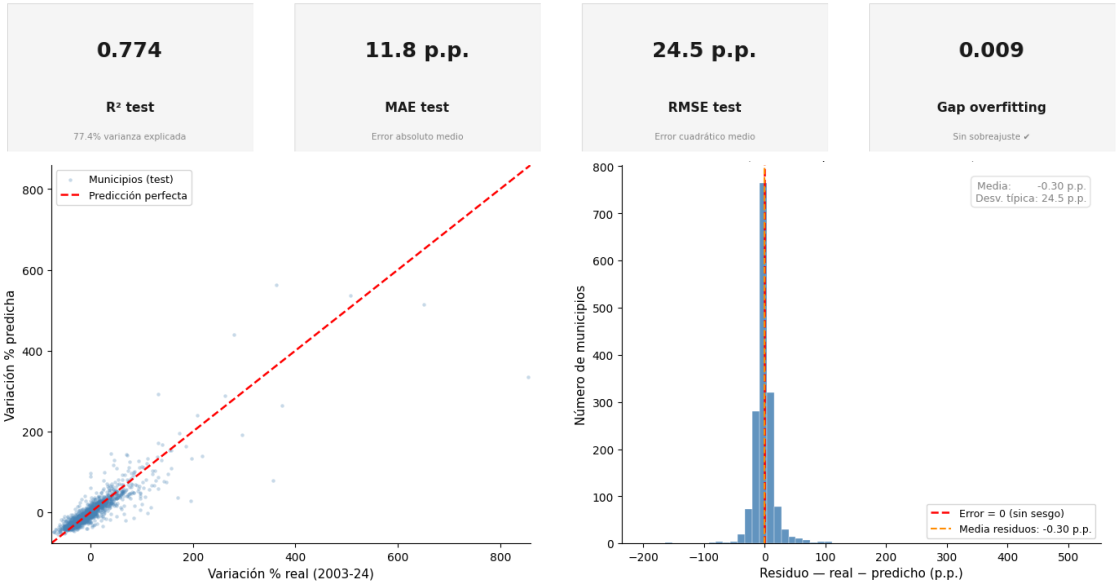
Entrenamiento y búsqueda de hiperparámetros

El dataset se divide en un conjunto de entrenamiento (80%, 6.505 municipios) y un conjunto de test (20%, 1.626 municipios), con semilla fija para garantizar la reproducibilidad. La búsqueda de hiperparámetros se realiza mediante RandomizedSearchCV con 20 iteraciones y validación cruzada de 5 folds, explorado sobre el número de árboles, el número de variables candidatas en cada split, el criterio de impureza y la profundidad máxima de los árboles. Los hiperparámetros óptimos encontrados son: 200 árboles, profundidad máxima de 10, selección aleatoria del 30% de variables en cada split, mínimo de 5 observaciones por hoja y criterio de error cuadrático medio.

Métricas de evaluación

La Figura 19 recoge métricas de evaluación en el conjunto de test después de haber entrenado el modelo. Además, también se puede ver debajo un gráfico de predicción frente a valores reales y la distribución de los residuos.

Figura 19. *Métricas de evaluación del Random Forest en conjunto de test (arriba). Gráfico de Predicción vs Real (abajo izquierda) y distribución de residuos (abajo derecha)*

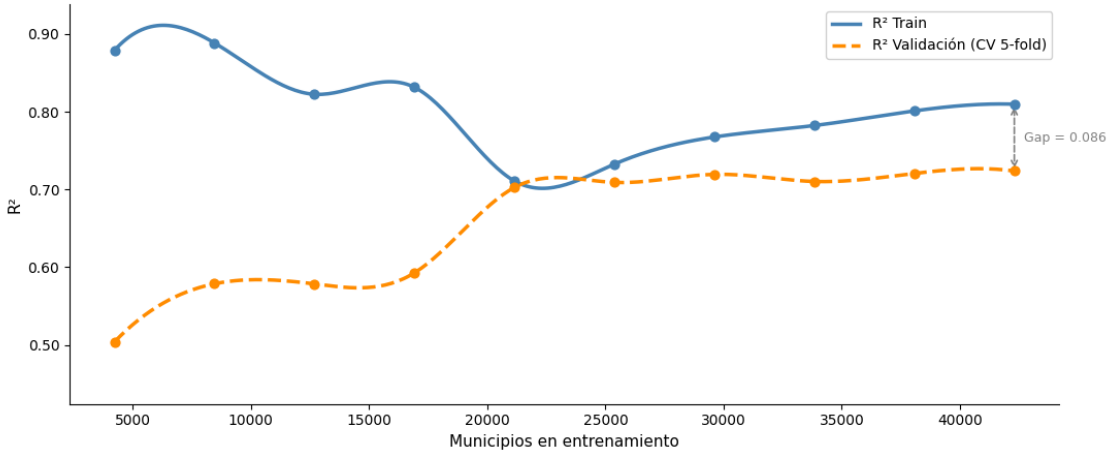


Elaboración propia.

El modelo explica el 77,4% de la varianza de la variación poblacional en el conjunto de test (R^2 = 0,774), con un error absoluto medio de 11,8 puntos porcentuales y un error cuadrático medio de 24,5 puntos porcentuales. El gap entre train y test es de apenas 0,009, lo que indica que el modelo generaliza correctamente sin sobreajuste relevante, resultado que se aprecia visualmente en el gráfico de predicción frente a valores reales, donde los puntos se agrupan en torno a la diagonal de predicción perfecta, y en la distribución de residuos, centrada en torno a cero sin sesgo sistemático apreciable.

Las curvas de aprendizaje de la Figura 20 permiten confirmar visualmente la ausencia de sobreajuste. La curva de entrenamiento se estabiliza en torno a R^2 = 0,78 y la de validación converge hacia ese mismo valor a medida que aumenta el tamaño de la muestra, con una brecha contenida entre ambas. Este patrón es indicativo de un modelo bien ajustado: ni memoriza el conjunto de entrenamiento ni carece de capacidad explicativa.

Figura 20. *Curvas de aprendizaje del Random Forest. La convergencia de ambas curvas confirma la ausencia de sobreajuste significativo*



Elaboración propia.

En la sección de resultados se muestra la importancia de permutación de las veinte variables con mayor capacidad explicativa, evaluada sobre el conjunto de test. Esta métrica mide la caída en R^2 al barajar aleatoriamente los valores de cada variable: si permutar una variable no empeora el modelo, esa variable no aporta información útil. Se prefiere frente a la importancia MDI porque no está sesgada hacia variables de alta cardinalidad y se evalúa sobre datos no vistos durante el entrenamiento. Asimismo, se presenta un gráfico SHAP para ver cómo se relacionan con la variable objetivo.

Simulaciones what-if

Una de las capacidades más valiosas de un modelo de Random Forest entrenado es la posibilidad de realizar simulaciones que permiten estimar el efecto que tendría sobre la variación poblacional predicha la modificación de una o varias variables de entrada, manteniendo el resto de condiciones constantes. Esta funcionalidad convierte el modelo

en una herramienta de apoyo a la toma de decisiones: dado el perfil actual de un municipio, ¿qué intervenciones de política pública habrían tenido mayor impacto sobre su evolución demográfica?

Es importante señalar, antes de interpretar los resultados, que este análisis es de naturaleza correlacional y no causal: el modelo captura asociaciones estadísticas aprendidas de datos observacionales, no el efecto real de una intervención. Modificar una variable en el perfil de un municipio no equivale a simular una política pública real, ya que no se tienen en cuenta los efectos sistémicos, la endogeneidad ni las respuestas de comportamiento que una intervención real generaría. Los resultados deben leerse, por tanto, como una exploración de la sensibilidad del modelo ante cambios en las variables de entrada, no como una estimación causal del impacto de medidas concretas.

Para ilustrar esta capacidad se toma como caso de estudio el municipio de Ribeira de Piquín (Lugo, código INE 27053), un ejemplo representativo de la España vaciada en su expresión más extrema: 492 habitantes en 2024; una pérdida del 40,5% de su población entre 2003 y 2024; un índice de envejecimiento de 1.167, es decir, casi doce mayores de 65 años por cada menor de 16; y un crecimiento vegetativo profundamente negativo. El municipio carece de consultorio médico propio, tiene una cobertura de fibra óptica (FTTH) del 7,7% de las viviendas y se encuentra a 46 minutos del municipio más cercano con más de 20.000 habitantes.

Se simulan tres intervenciones simultáneas que representan medidas concretas y reales de política pública: la apertura de un consultorio médico de atención primaria, la extensión de la cobertura de fibra óptica al 100% de las viviendas, y la reducción de 15 minutos en el tiempo de desplazamiento al municipio más cercano de más de 20.000 habitantes por una mejora significativa de la red viaria. Los resultados de esta simulación y su interpretación se presentan en el capítulo de resultados.

Clustering

El tercer modelo aplica técnicas de clustering, modelo de aprendizaje no supervisado, para identificar grupos de municipios con perfiles estructurales similares. A diferencia del Random Forest, que predice y explica la variación de población, el clustering tiene como objetivo comprender mejor la heterogeneidad del territorio español y analizar cómo ha afectado la despoblación a cada tipo de municipio.

Selección de variables y preparación

Para el clustering se utiliza un subconjunto de catorce variables seleccionadas para cubrir las dimensiones clave de diferenciación entre municipios (tamaño, densidad, demografía, economía, estructura sectorial, aislamiento, infraestructura, turismo y conectividad) maximizando la independencia entre ellas. La variación de población se excluye deliberadamente del input del clustering, con el objetivo de agrupar los municipios por sus características estructurales y analizar después cómo ha evolucionado la población en cada grupo, no condicionando los grupos a dicha evolución.

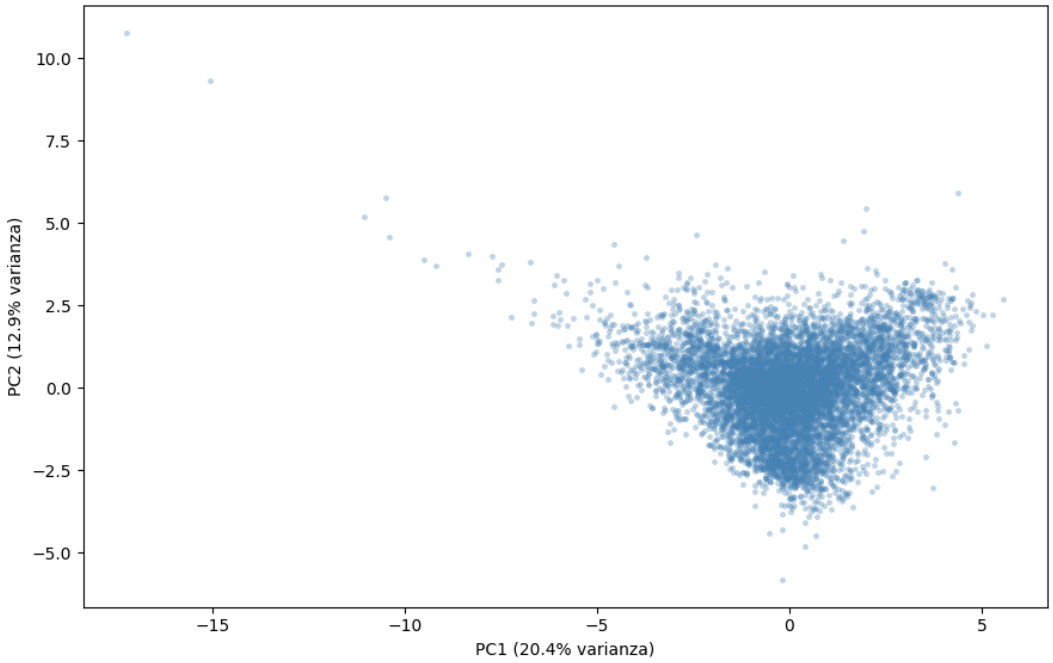
Dado que K-means utiliza distancias euclídeas, las variables se estandarizan previamente (media 0, desviación típica 1) para evitar que variables con escalas muy distintas dominen las distancias. Para el clustering jerárquico, en cambio, se utiliza la distancia de Gower, que

normaliza internamente cada variable al rango [0,1] y aplica la métrica apropiada según el tipo de variable (numérica o binaria), por lo que no requiere estandarización previa. Esta decisión está motivada por la presencia de variables binarias en el conjunto (área costera), cuyo tratamiento con distancia euclídea no sería metodológicamente correcto.

Análisis de componentes principales (PCA)

Con carácter previo al clustering se realiza un Análisis de Componentes Principales con el objetivo de explorar visualmente si existen agrupaciones naturales en los datos. La primera componente principal explica el 20,4% de la varianza total entre municipios, y la segunda el 12,9%, acumulando entre ambas un tercio. La Figura 21 muestra la proyección de todos los municipios en el espacio definido por estas dos componentes.

Figura 21. *Proyección de los municipios españoles en las dos primeras componentes principales*

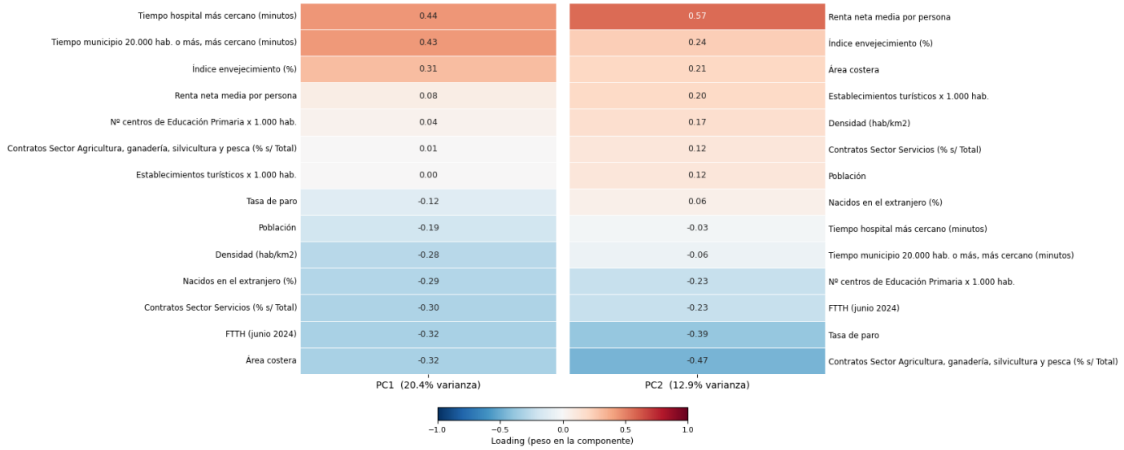


Elaboración propia.

Como se puede apreciar, no emergen agrupaciones visualmente nítidas ni separadas en el espacio bidimensional del PCA. No existen tipos discretos y bien delimitados, sino un espectro de situaciones que los algoritmos de clustering buscarán segmentar de forma aproximada.

Las componentes principales no son más que combinaciones lineales ponderadas de las variables originales, donde los pesos, denominados *loadings*, indican la contribución de cada variable a cada componente. Analizar qué variables tienen mayores *loadings* en las primeras componentes permite identificar qué dimensiones estructurales son las que más diferencian a los municipios entre sí. La Figura 22 muestra los *loadings* de las catorce variables de clustering en PC1 y PC2, ordenados de mayor a menor peso.

Figura 22. *Pesos (loadings) de las variables en PC1 y PC2*



Elaboración propia.

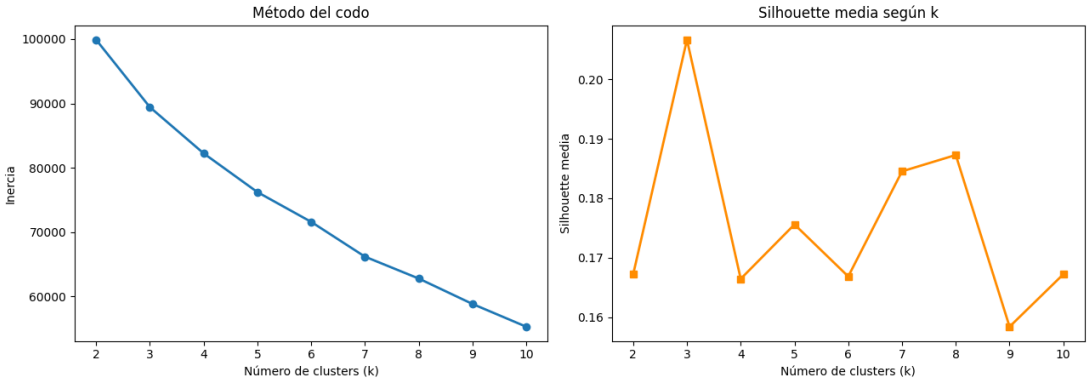
PC1, que explica el 20,4% de la varianza total, está dominada por variables de aislamiento e infraestructura (tiempo al hospital y tiempo al municipio más cercano de más de 20.000 habitantes) con pesos positivos altos y por variables de conectividad, orientación terciaria y si es costero (FTTH, contratos en servicios, área costera) con pesos negativos. Esta componente puede interpretarse como un eje que opone municipios aislados y poco conectados frente a municipios más bien costeros, con infraestructura y economía de servicios. PC2, que explica el 12,9%, está liderada por la renta media con peso positivo y la agricultura y la tasa de paro con pesos negativos, configurando un eje de nivel económico y estructura productiva que separa municipios prósperos y terciarizados de municipios agrarios con mayor desempleo.

Selección del número óptimo de clusters y método

Se prueban dos algoritmos de clustering, K-means y clustering jerárquico con criterio Ward, evaluados para distintos valores de k mediante el coeficiente de silhouette, que mide la calidad de la separación entre grupos en una escala de −1 a 1, donde valores más altos indican clusters más compactos y bien separados. En ambos casos la evaluación de la silhouette se realiza sobre el espacio euclídeo estandarizado para que los resultados sean comparables.

Para el K-means, la Figura 23 muestra el método del codo y la silhouette media para valores de k entre 2 y 10. El método del codo no presenta un punto de inflexión claro. La silhouette media alcanza su valor máximo en k=3 (0,207), siendo este el criterio determinante para la elección del número de clusters.

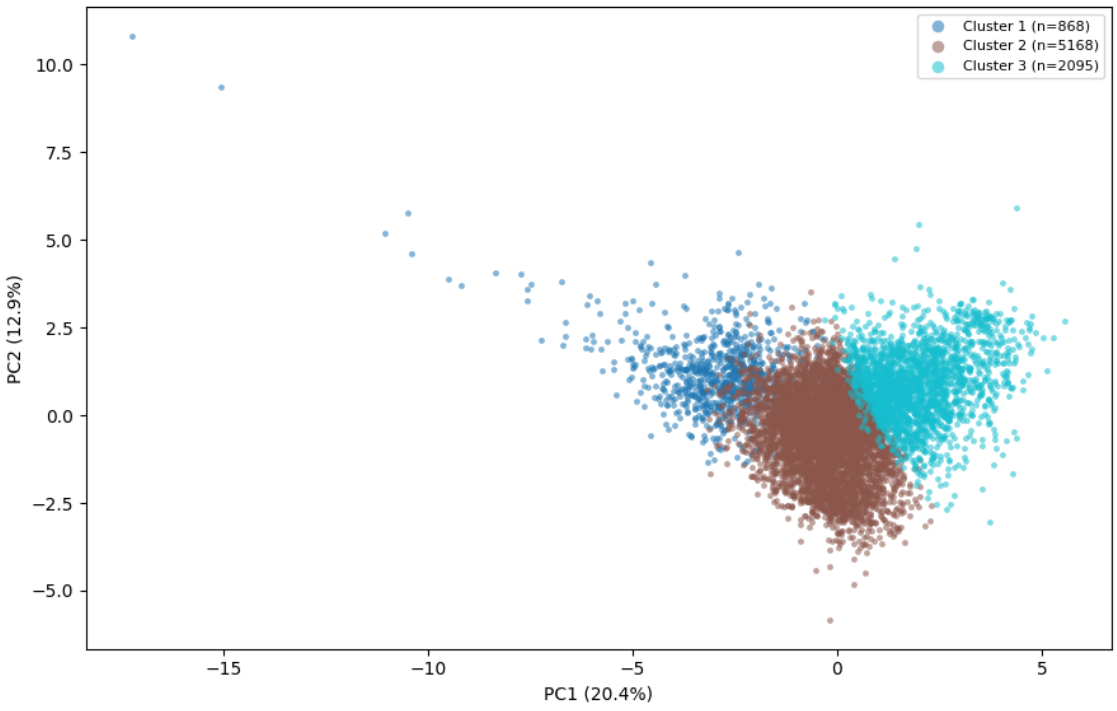
Figura 23. *Método del codo (izquierda) y silhouette media (derecha) para K-means*



Elaboración propia.

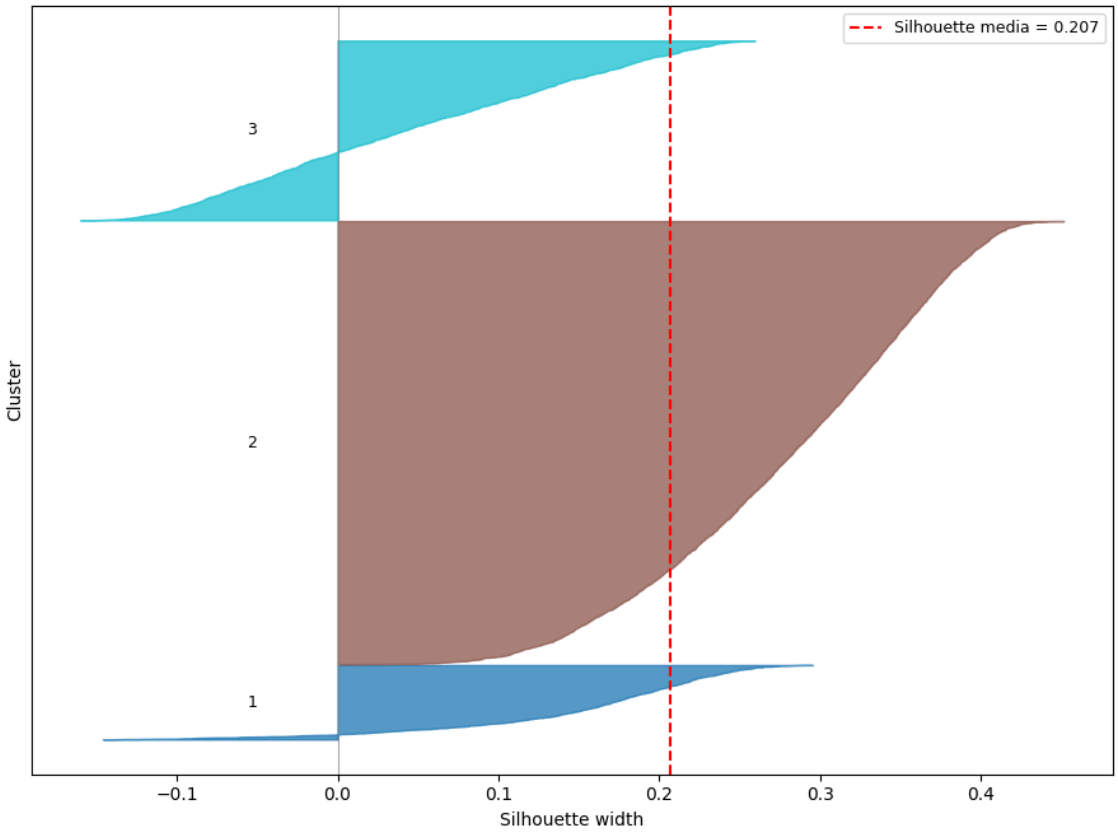
La Figura 24 muestra la proyección en el espacio PCA y la Figura 25 el silhouette plot del K-means para k=3.

Figura 24. *Visualización del K-means (k=3) en el espacio PCA*



Elaboración propia.

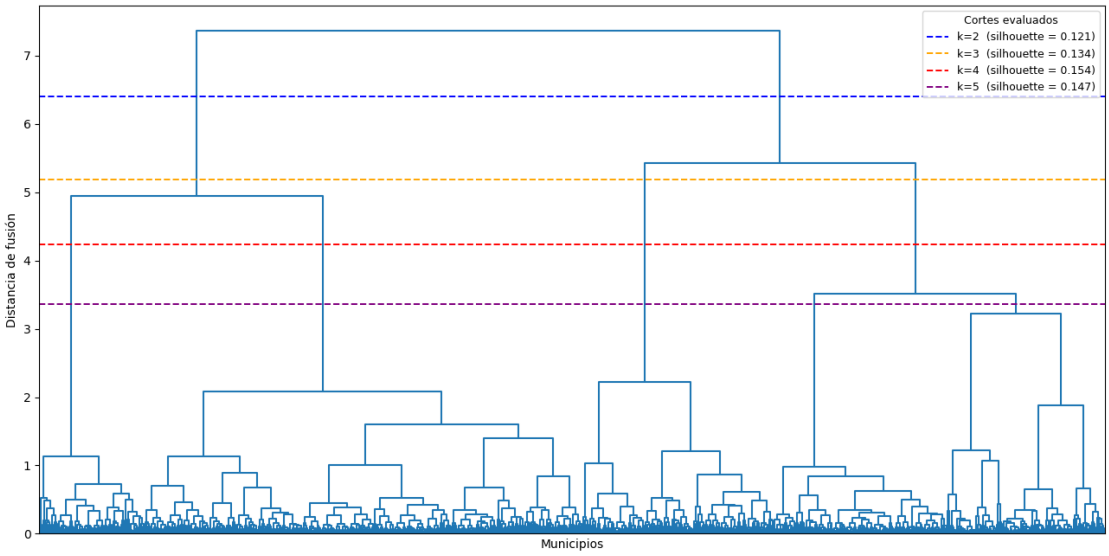
Figura 25. *Gráfico de silueta por observación y cluster del K-means (k=3)*



Elaboración propia.

Para el clustering jerárquico se utiliza la distancia de Gower con criterio de fusión Ward. El dendrograma de la Figura 26 muestra la estructura jerárquica de las fusiones, con las líneas discontinuas indicando los cuatro puntos de corte evaluados y la silhouette media obtenida en cada caso. El valor más alto corresponde a k=4 (silhouette = 0,154), que es el número de clusters seleccionado para este método.

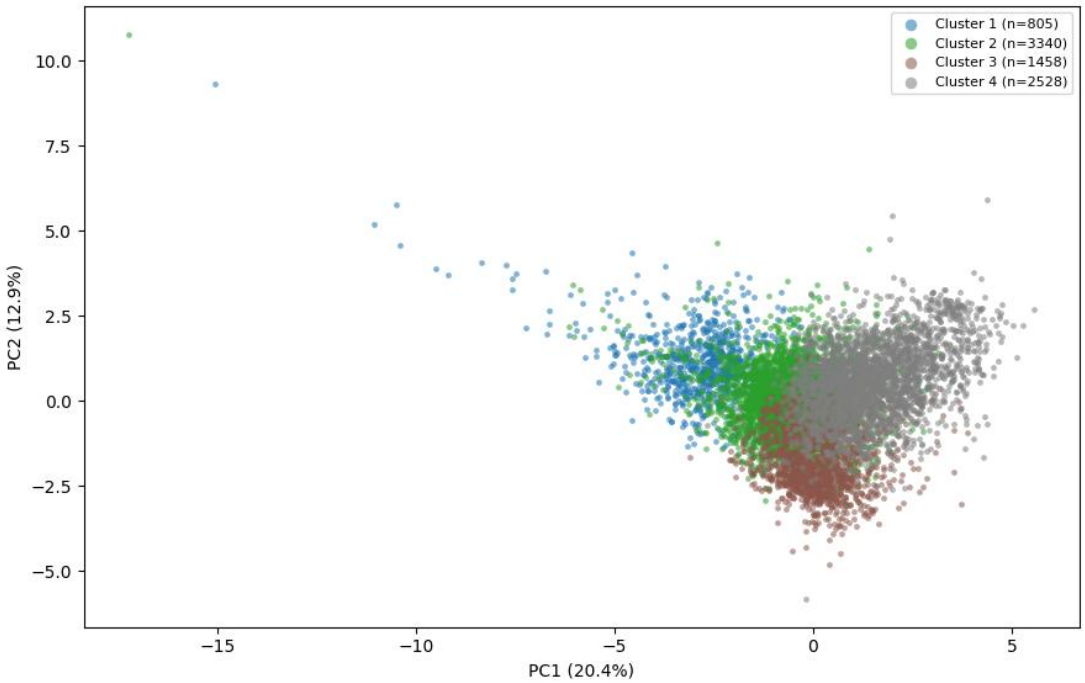
Figura 26. *Dendrograma del clustering jerárquico (distancia Gower, criterio Ward). Las líneas discontinuas indican los posibles puntos de corte para k=2, 3, 4 y 5, con la silhouette media correspondiente a cada uno*



Elaboración propia.

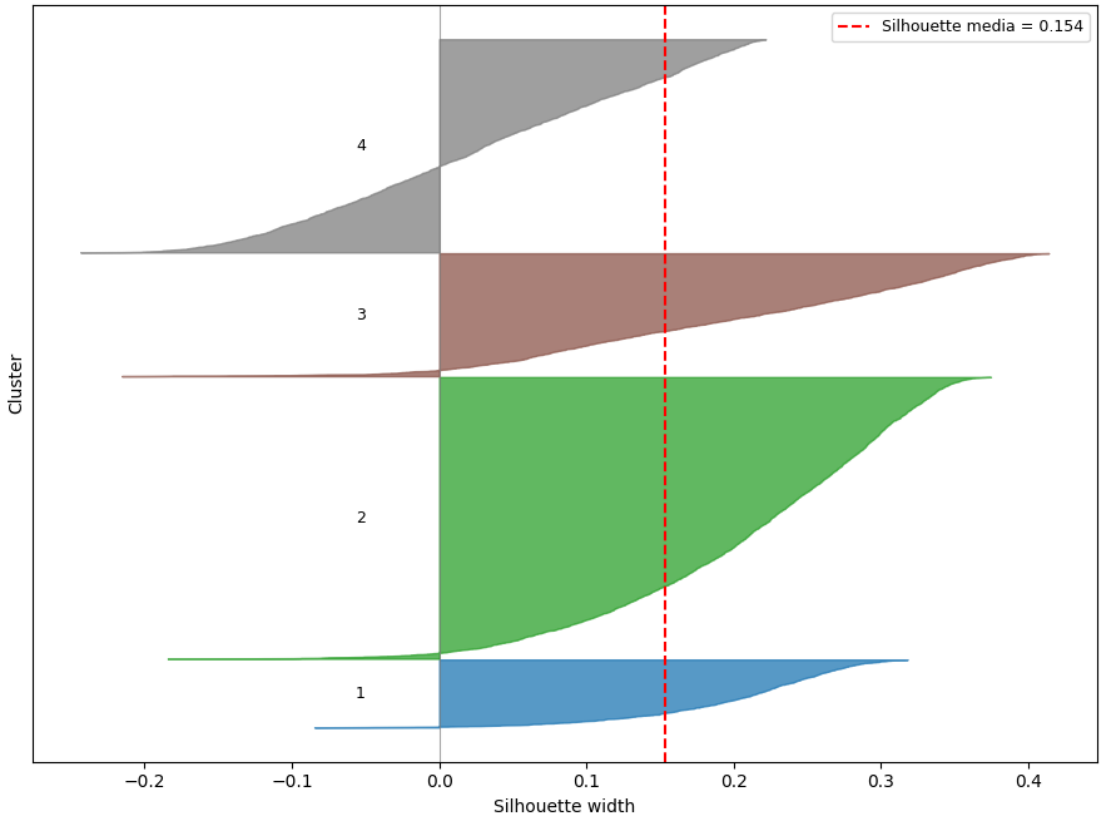
La Figura 27 muestra la proyección en el espacio PCA y la Figura 28 el silhouette plot de la solución jerárquica con k=4.

Figura 27. *Visualización del Jerárquico (k=4) en el espacio PCA*



Elaboración propia.

Figura 28. *Gráfico de silueta por observación y cluster del Jerárquico (k=4)*



Elaboración propia.

Comparando ambos métodos, el K-means con k=3 obtiene una silhouette media de 0,207 frente al 0,154 del clustering jerárquico con k=4, lo que lleva a seleccionar el K-means como método definitivo.

La caracterización detallada de cada perfil de municipio, el análisis del impacto de la despoblación en cada cluster y la distribución geográfica de los grupos identificados se presentan en el capítulo de resultados.

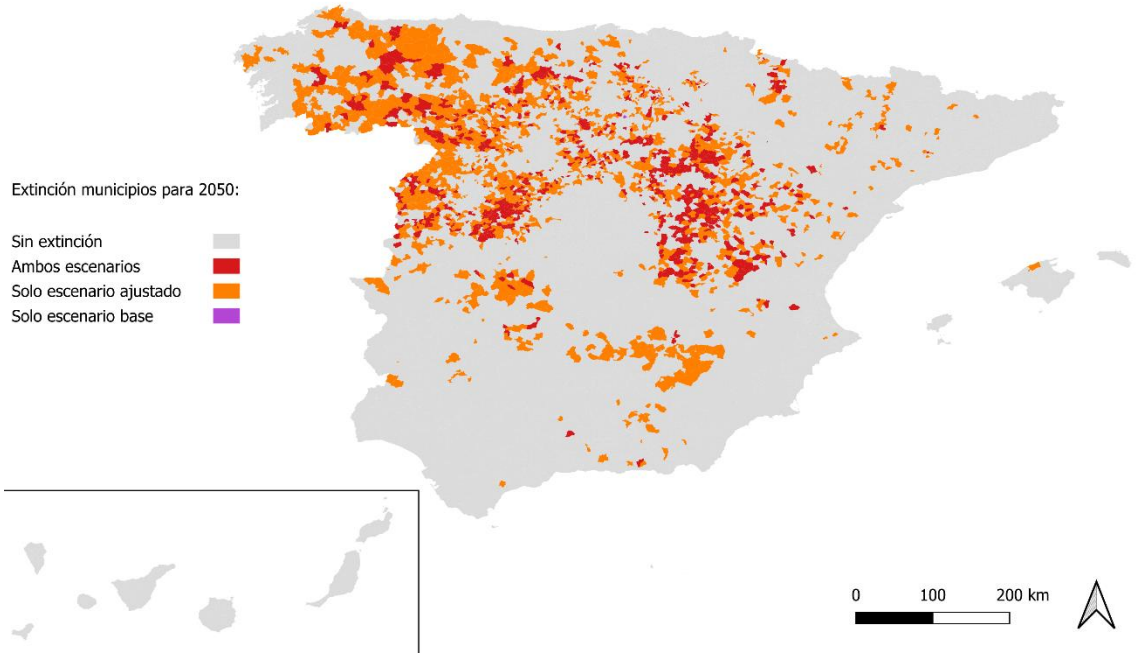
Resultados, conclusiones, limitaciones y futuros pasos

Resultados

Resultado 1: Escenarios de proyección poblacional (2025-2050)

La Figura 29 recoge el mapa de España con los municipios que alcanzarían la extinción demográfica antes de 2050 coloreados según el escenario que los predice.

Figura 29. *Municipios extintos para 2050 según el escenario base y el ajustado*



Elaboración propia.

El patrón espacial es inequívoco: los municipios más vulnerables se concentran en el interior peninsular, especialmente en las dos Castillas, Aragón, Extremadura y zonas del interior de Galicia y Asturias. Las zonas costeras, las grandes áreas metropolitanas y los territorios insulares quedan prácticamente al margen del fenómeno.

La diferencia entre el escenario base y el ajustado revela algo relevante: un subconjunto de municipios aparece únicamente en el escenario ajustado (coloreados en naranja), lo que indica que su tendencia histórica pura no augura extinción antes de 2050 pero su estructura demográfica actual, con índices de envejecimiento muy elevados, intensifica el declive proyectado hasta cruzar el umbral. En sentido contrario, solo dos municipios aparecen solo por el escenario base, pues presentan una tendencia histórica de fuerte caída pero una población relativamente joven que el ajuste demográfico modera.

La siguiente tabla recoge los diez municipios con mayor urgencia según cada escenario, aquellos cuya fecha estimada de extinción es más próxima.

Tabla 5. *Top 10 municipios con fecha de extinción demográfica más próxima según el escenario base (izquierda) y el escenario ajustado (derecha)*

Municipio	Cód. INE	Año extinción
Arandilla del Arroyo	16020	2026
Angón	19031	2026
Cendejas de la Torre	19081	2027
Cigudosa	42062	2027
Rebollosa de Jadraque	19231	2029
Navamorales	37218	2029
Riaguas de San Bartolomé	40168	2029
Arbeteta	19038	2030
Bercimuel	40029	2030
Collado del Mirón	05063	2031

Municipio	Cód. INE	Año extinción
Angón	19031	2027
Cigudosa	42062	2027
Collado del Mirón	05063	2028
Arbeteta	19038	2028
Cendejas de la Torre	19081	2028
Navamorales	37218	2028
Bascuñana	09046	2029
Ibrillos	09178	2029
Arandilla del Arroyo	16020	2029
Valsalobre	16234	2029

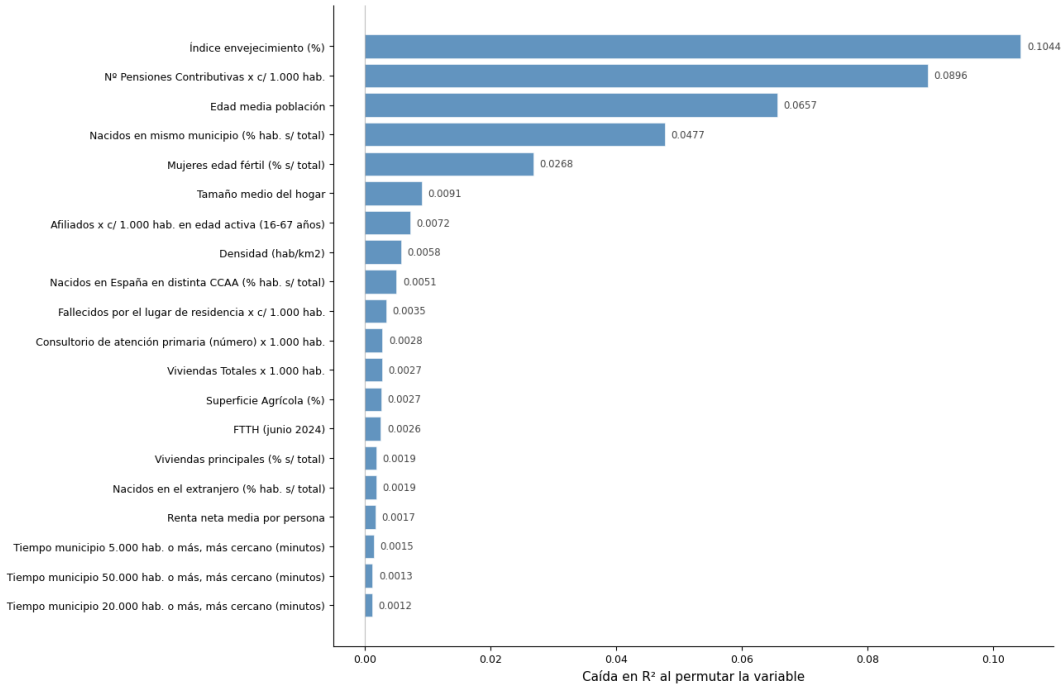
Elaboración propia.

Los municipios identificados son en su mayoría pequeñas localidades con poblaciones inferiores a 200 habitantes y pérdidas acumuladas superiores al 40% en el período analizado. Destaca que varios municipios cambian de posición entre escenarios en función de su índice de envejecimiento, lo que confirma que la estructura demográfica es tan determinante como la tendencia histórica a la hora de estimar la vulnerabilidad futura.

Resultado 2: Factores explicativos del cambio poblacional 2003-2024 (Random Forest) y simulación what-if

La Figura 30 recoge la importancia de permutación de las veinte variables con mayor capacidad explicativa del modelo, evaluada sobre el conjunto de test.

Figura 30. *Importancia de permutación de las veinte variables más relevantes del modelo*

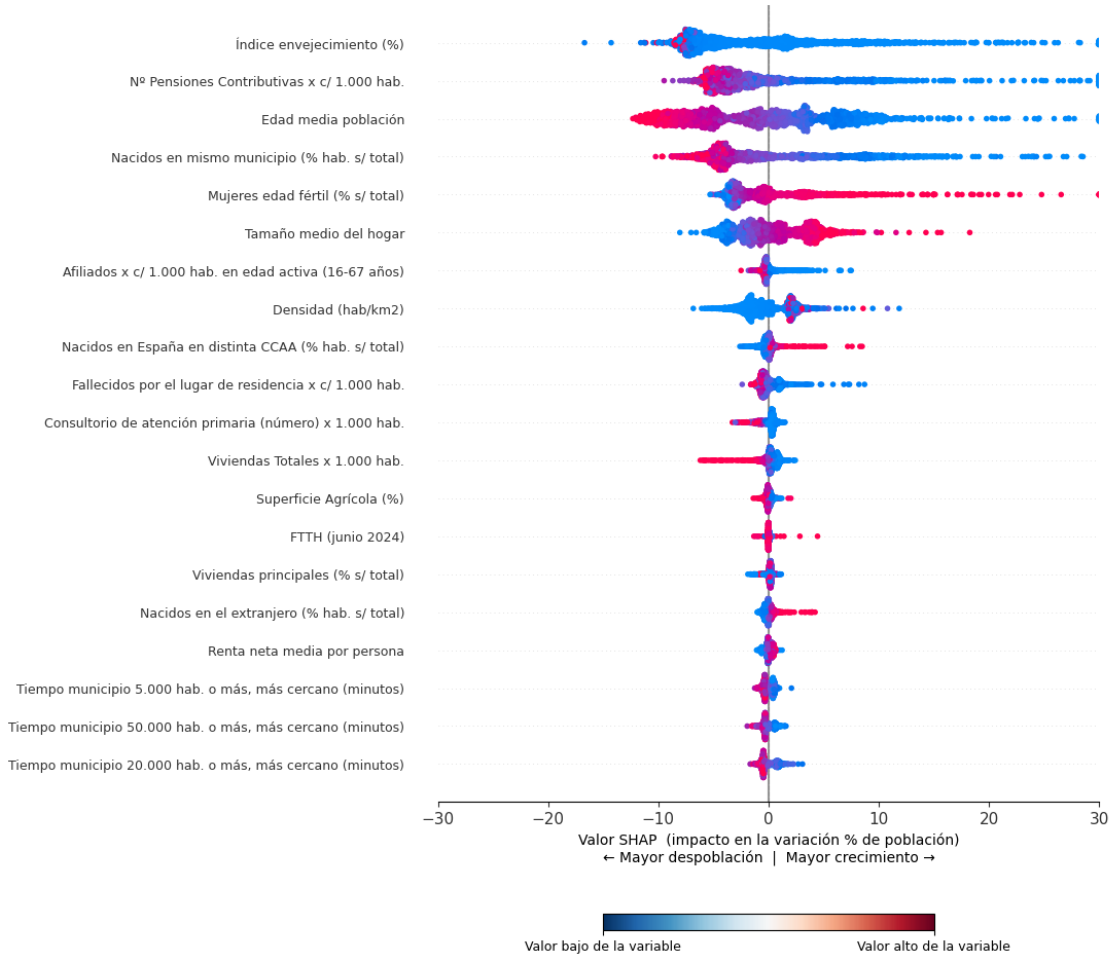


Elaboración propia.

Las cinco variables con mayor importancia de permutación son el índice de envejecimiento (0,104), el número de pensiones contributivas por cada 1.000 habitantes (0,090), la edad media de la población (0,066), el porcentaje de nacidos en el mismo municipio (0,048) y el porcentaje de mujeres en edad fértil (0,027). Los resultados apuntan a una conclusión clara: los factores demográficos estructurales (envejecimiento, arraigo local y capacidad reproductiva) son los que con mayor fuerza explican la pérdida de población a nivel municipal, por encima de los factores económicos o de conectividad. El peso del número de pensiones contributivas refuerza esta lectura: una población con alta dependencia de pensiones es una población envejecida sin relevo generacional.

El siguiente gráfico muestra de esas variables con mayor relevancia, cómo se relacionan con la variable objetivo.

Figura 31. *SHAP importancia y dirección del efecto*



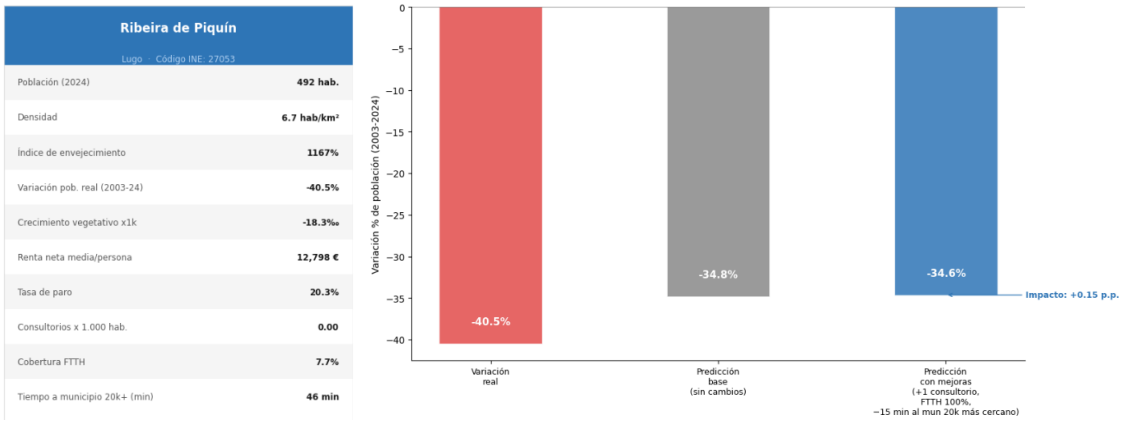
Elaboración propia.

El gráfico SHAP permite visualizar la dirección del efecto de cada variable sobre la predicción del modelo. Cada punto representa un municipio. Su posición horizontal indica el valor de la variable objetivo para ese municipio: los puntos a la izquierda del cero contribuyen a predecir mayor despoblación, y los puntos a la derecha, mayor crecimiento. El color codifica el valor real de la variable en ese municipio: rojo indica un valor alto y azul un valor bajo.

La lectura del gráfico permite extraer conclusiones claras sobre cómo se relaciona cada variable con la evolución poblacional. Por ejemplo, en la edad media de la población se aprecia que los municipios con envejecimiento alto (puntos rojos) se desplazan fuertemente hacia la izquierda, es decir, el modelo predice para ellos mayor despoblación; mientras que los de envejecimiento bajo (puntos azules) se desplazan hacia la derecha, es decir, menor despoblación. Variables como la FTTH, el tiempo de accesibilidad o la renta media aparecen en la parte inferior del gráfico con distribuciones muy concentradas en torno a cero, confirmando que su impacto sobre la variación poblacional es marginal en comparación con los factores demográficos estructurales.

Por otro lado, la Figura 32 muestra los resultados de la simulación para Ribeira de Piquín (Lugo), municipio con 492 habitantes, una pérdida del 40,5% de su población entre 2003 y 2024 y un índice de envejecimiento de 1.167.

Figura 32. *Simulación what-if para Ribeira de Piquín. Perfil estructural del municipio (izquierda) y variación poblacional real, predicción base del modelo y predicción bajo el escenario de mejoras simultáneas en sanidad, conectividad y accesibilidad (derecha)*



Elaboración propia.

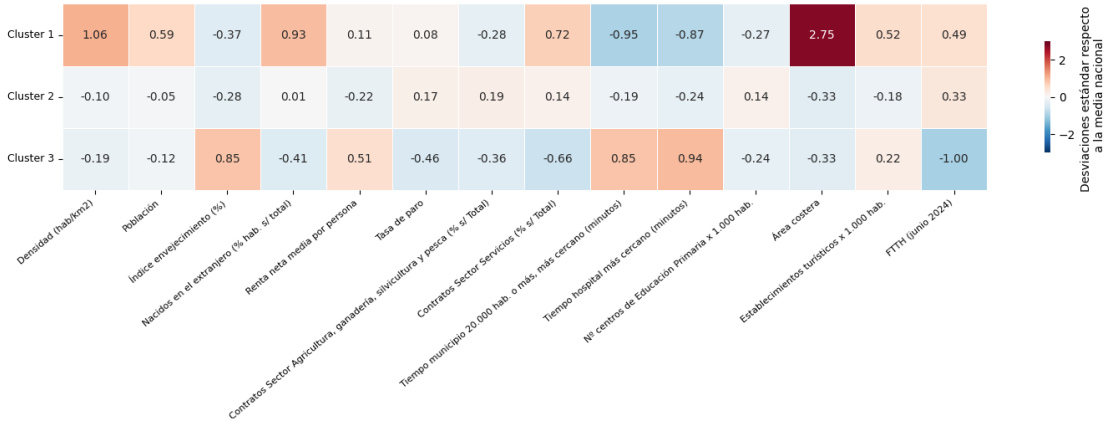
El resultado es revelador: la combinación de las tres intervenciones desplaza la predicción apenas 0,15 puntos porcentuales, de −34,79% a −34,64%. Esta cifra, prácticamente indistinguible de la predicción base, sugiere que en municipios con envejecimiento extremo las variables demográficas estructurales dominan de tal modo sobre las variables accionables que una mejora aislada en servicios o conectividad no se refleja de forma apreciable en el patrón histórico observado. No hay jóvenes que se queden atraídos por la fibra óptica, no hay familias que valoren tener un consultorio cerca porque no hay niños, y la reducción del tiempo de desplazamiento no compensa la ausencia de oportunidades económicas y de relevo generacional.

Este resultado conecta directamente con los hallazgos del análisis de importancia de variables: las cinco variables con mayor poder explicativo en el modelo son todas de naturaleza demográfica y estructural (índice de envejecimiento, pensiones contributivas, edad media, arraigo local y mujeres en edad fértil) mientras que las variables accionables desde la política pública presentan importancias de permutación marginales. Por tanto, el modelo sugiere que en municipios que han cruzado ciertos umbrales demográficos, la variación de población histórica predicha ya no responde de forma apreciable a variables de servicios o conectividad. Dado que el modelo captura asociaciones estadísticas y no relaciones causales, este resultado no permite afirmar que las intervenciones convencionales sean ineficaces en la práctica, pero sí invita a reflexionar sobre el momento óptimo de intervención, cuestión que se aborda en las conclusiones.

Resultado 3: Tipologías municipales y distribución geográfica de la despoblación

El análisis de clustering identifica tres perfiles estructurales de municipio que capturan la heterogeneidad del territorio español. La Figura 33 muestra el heatmap de centroides estandarizados que permite caracterizar cada grupo.

Figura 33. *Perfil de centroides estandarizados por cluster*

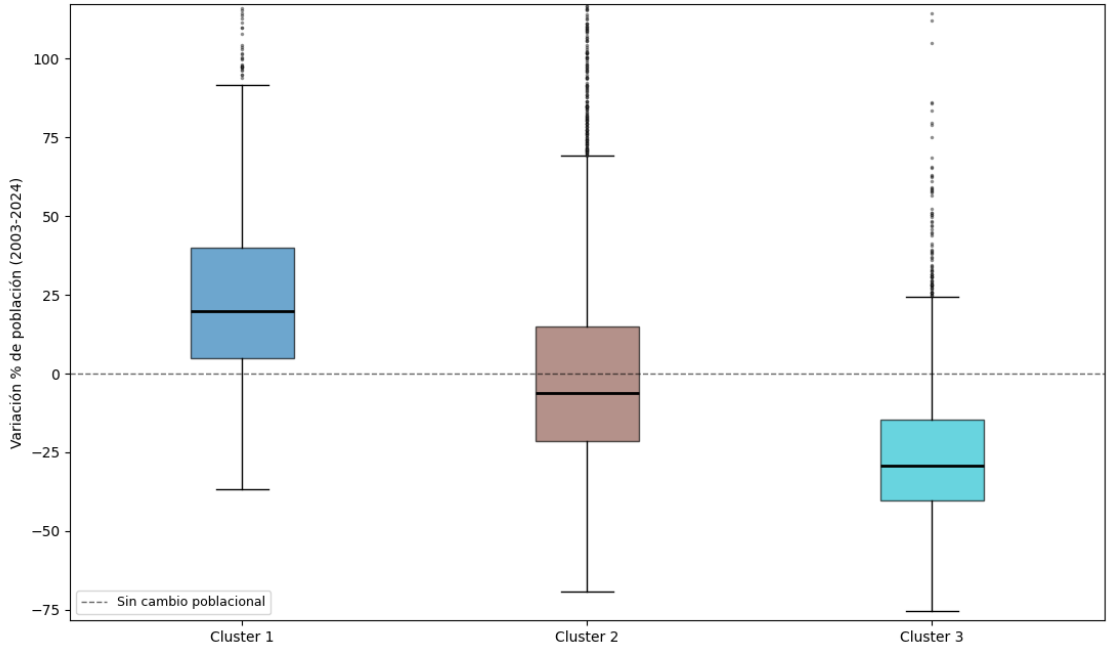


Elaboración propia.

El heatmap de centroides permite caracterizar los tres grupos con claridad. El Cluster 1 se distingue por una densidad y población significativamente superiores a la media nacional, un índice de envejecimiento bajo, una elevada proporción de nacidos en el extranjero, una localización próxima a municipios grandes y hospitales, una fuerte presencia costera y una alta dotación de establecimientos turísticos. Se trata, por tanto, de municipios urbanos, costeros y turísticos, concentrados principalmente en el arco mediterráneo, las islas y las grandes áreas metropolitanas. El Cluster 2 presenta un perfil intermedio, con valores próximos a la media nacional en prácticamente todas las dimensiones y una ligera ventaja en conectividad y servicios respecto al Cluster 3. Agrupa municipios semiurbanos y periurbanos que no destacan por ninguna característica extrema. El Cluster 3, en contraste, muestra un envejecimiento muy por encima de la media, una población y densidad bajas, un fuerte aislamiento geográfico reflejado en los tiempos elevados al municipio más cercano y al hospital, una economía con escasa presencia del sector servicios y una conectividad digital significativamente inferior, con el valor más bajo de cobertura FTTH de los tres grupos. Corresponde al perfil clásico de la España vaciada: municipios rurales aislados, envejecidos y con escasa dotación de servicios e infraestructuras.

Una vez caracterizados los tres grupos, se analiza cómo ha evolucionado la población en cada uno de ellos entre 2003 y 2024, incorporando la variación porcentual de población como variable de validación externa no utilizada en la construcción de los clusters.

Figura 34. *Distribución de la variación porcentual de población (2003-2024) por cluster. Eje recortado al P3-P97*

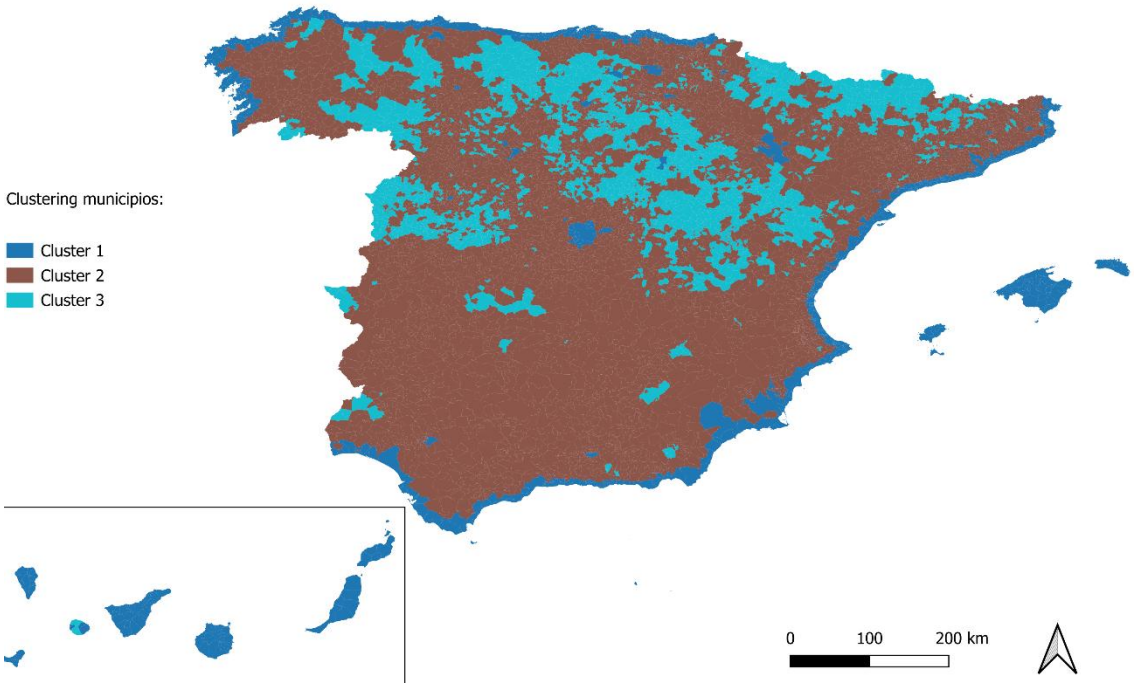


Elaboración propia.

La Figura 34 confirma de forma contundente la relación entre el perfil estructural de cada cluster y su evolución demográfica. El Cluster 1, formado por municipios urbanos y turísticos, registra una variación mediana de población en torno al +20% en el período 2003-2024, con una distribución amplia pero claramente sesgada hacia valores positivos. El Cluster 2, de perfil intermedio, presenta una mediana ligeramente negativa en torno al −10%, con una distribución más centrada en torno a cero que refleja situaciones muy heterogéneas. El Cluster 3, el de los municipios rurales aislados y envejecidos, acumula las pérdidas más severas, con una mediana en torno al −30% y una distribución casi íntegramente por debajo de cero. Este resultado valida la coherencia del clustering: los grupos construidos exclusivamente a partir de características estructurales capturan con precisión las diferencias en la evolución demográfica de los distintos tipos de municipio español, y refuerzan el diagnóstico de que el envejecimiento, el aislamiento y la escasa dotación de servicios son los rasgos que mejor explican la concentración de la despoblación en determinados territorios.

La Figura 35 muestra la distribución geográfica de los clusters, permitiendo visualizar en el mapa de España qué territorios pertenecen a cada perfil.

Figura 35. *Distribución geográfica de los clusters municipales*



Elaboración propia.

La lectura conjunta de este mapa con el de extinción demográfica presentado en el Resultado 1 revela una notable coincidencia espacial: los municipios del Cluster 3 (rurales, aislados y envejecidos) se solapan en gran medida con los territorios que los modelos de proyección identifican como más vulnerables antes de 2050. Ambas aproximaciones, construidas desde ángulos metodológicos distintos, una proyección tendencial y una segmentación estructural no supervisada, convergen en señalar las mismas zonas del interior peninsular como las de mayor urgencia demográfica, lo que refuerza la robustez de los hallazgos y la coherencia del análisis en su conjunto.

Conclusiones

El análisis desarrollado en este trabajo permite extraer un conjunto de conclusiones que abarcan tanto el diagnóstico del fenómeno de la despoblación rural en España como el valor del enfoque metodológico adoptado.

La primera conclusión es de carácter territorial: la despoblación no es un fenómeno homogéneo, sino la otra cara de una concentración demográfica igualmente intensa. Existe una España que crece: las costas, el arco mediterráneo, las islas y las grandes áreas metropolitanas; y una España que se vacía: el interior peninsular, especialmente las dos Castillas, Aragón, Extremadura y zonas del interior de Galicia y Asturias. Esta brecha no ha hecho sino ensancharse en las últimas décadas y los modelos de proyección sugieren que, bajo el supuesto de continuidad de las tendencias recientes, varios municipios del interior desaparecerían demográficamente antes de 2030, con los primeros casos concentrados en provincias de Guadalajara, Soria y Burgos.

La segunda conclusión afecta directamente al diseño de políticas públicas. El modelo explicativo revela que el envejecimiento, las pensiones contributivas, la edad media de la población y el arraigo local explican la variación poblacional municipal de forma abrumadoramente superior a variables accionables como la conectividad digital, la dotación de servicios sanitarios o la accesibilidad viaria.

De esta segunda conclusión se deriva una tercera, quizás la más relevante para los responsables de política territorial. El análisis de sensibilidad muestra que, en municipios con envejecimiento estructural muy elevado, ninguna combinación razonable de mejoras en servicios o conectividad desplaza de forma apreciable la variación poblacional que el modelo predice. Dado que el modelo captura asociaciones estadísticas y no relaciones causales, este resultado no permite afirmar que dichas intervenciones sean ineficaces en la práctica, pero sí es coherente con lo que la literatura especializada lleva tiempo señalando: cuando la estructura demográfica de un territorio se ha deteriorado de forma severa, el margen de acción de las políticas convencionales puede ser muy limitado. En ese sentido, los resultados apuntan a que la prioridad debería ser actuar de forma preventiva en los municipios que todavía conservan suficiente estructura demográfica para beneficiarse de las intervenciones, antes de que entren en la espiral de declive irreversible que caracteriza a casos extremadamente envejecidos como el de Ribeira de Piquín.

Finalmente, este trabajo aporta una conclusión de naturaleza metodológica relevante para el ámbito del Business Analytics. La integración de fuentes de datos heterogéneas procedentes de organismos oficiales (INE, MITECO, SEPE, Ministerio para la Transformación Digital, entre otras) permite construir una visión territorial completa que ninguna fuente individual ofrece por sí sola. La aplicación de técnicas de aprendizaje automático, análisis de importancia de variables y clustering no supervisado al análisis territorial demuestra que estas herramientas no son exclusivas del ámbito empresarial o financiero, sino que tienen un potencial real para enriquecer el diagnóstico de problemas sociales complejos y apoyar la toma de decisiones basada en evidencia en materia de política pública.

Limitaciones y futuros pasos

El presente trabajo presenta una serie de limitaciones que conviene reconocer para situar correctamente el alcance de sus conclusiones.

La principal limitación metodológica deriva de la ausencia de series históricas completas a nivel municipal para la mayoría de las variables explicativas. Al contar únicamente con una fotografía reciente de los indicadores (en torno a 2024) mientras que la variable objetivo captura la variación poblacional entre 2003 y 2024, existe una asimetría temporal que restringe el enfoque a la explicación de asociaciones observadas al final del período. Esta limitación impide extrapolar los resultados como predicciones del comportamiento demográfico futuro: para ello sería necesario disponer de paneles municipales históricos con cobertura homogénea que permitiesen alinear temporalmente variables y target.

En segundo lugar, el modelo no incorpora una serie de variables que podrían resultar muy relevantes para explicar la dinámica demográfica municipal. Entre ellas, destacan: los flujos migratorios internos y el saldo migratorio exterior, componentes fundamentales de la variación de población junto al crecimiento vegetativo; indicadores de calidad de vida y

bienestar subjetivo; variables de capital social y cohesión comunitaria; la presencia de segundas residencias y su efecto sobre la percepción de vitalidad del municipio; o indicadores de actividad emprendedora y tejido empresarial local. La ausencia de estas variables no obedece a una decisión metodológica sino a su falta de disponibilidad con cobertura homogénea a nivel municipal, lo que constituye una limitación de disponibilidad de datos sobre el territorio rural español.

A partir de estas limitaciones se abren varias líneas futuras de investigación que podrían extender el trabajo. La incorporación de nuevas variables enriquecería significativamente la capacidad explicativa del análisis. Asimismo, sería de interés estudiar, por ejemplo, la evolución del porcentaje de población extranjera en los municipios afectados por la despoblación: en varios de ellos este indicador ha crecido de forma muy significativa en las últimas décadas, llegando en algunos casos a superar al de población de nacionalidad española. Este fenómeno añade una dimensión cultural y de cohesión territorial al proceso de vaciamiento que merece un análisis específico y que podría constituir por sí solo un trabajo de investigación independiente.

Finalmente, resultaría especialmente valioso analizar los casos atípicos que el modelo no captura bien: municipios rurales aislados con crecimiento demográfico sostenido, y municipios con condiciones estructurales aparentemente favorables que sin embargo registran pérdidas severas de población. El estudio cualitativo de estas excepciones, casos de éxito y de fracaso que escapan al patrón general, puede aportar claves explicativas que los modelos cuantitativos no son capaces de detectar y orientar de forma más concreta el diseño de políticas públicas diferenciadas por territorio.

Bibliografía

Agencia EFE. (2023, 1 de abril). Cuatro años de la Revuelta de la España Vaciada: Zamora dice "No al territorio de sacrificio". https://efe.com/castilla-y-leon/2023-04-01/revuelta-de-la-espana-vaciada/

Ávila, M. G. (2019, 1 de enero). España, zona cero de la despoblación en la Unión Europea. *El Independiente*. https://www.elindependiente.com/desarrollo-sostenible/2019/01/01/la-zona-cero-de-la-despoblacion-en-la-union-europea/

Ayuntamiento de San Esteban del Valle. (2019). *Plan Estratégico contra la Despoblación 2019-2022*. San Esteban del Valle.

Cabello, S. A. (2024). España vacía, España vaciada: las dimensiones de identidad y simbólicas de las regiones periféricas. Un marco territorial. *Papeles del CEIC*, *2024*(1), 1–7. https://doi.org/10.1387/pceic.25988

Camarero, L., & Oliva, J. (2019). Thinking in rural gap: mobility and social inequalities. *Palgrave Communications*, *5*, 95. https://doi.org/10.1057/s41599-019-0306-x

Collantes, F., & Pinilla, V. (2019). *¿Lugares que no importan? La despoblación de la España rural desde 1900 hasta el presente*. Prensas de la Universidad de Zaragoza.

Comisión Europea. (2025). *Rural development – Common Agricultural Policy*. Unión Europea.

Del Molino, S. (2016). *La España vacía. Viaje por un país que nunca fue*. Turner.

Eurostat. (s. f.). *Population grid data (GISCO)*. https://ec.europa.eu/eurostat/web/gisco/geodata/grids

FUNCAS. (2015). *La desindustrialización de España en el contexto europeo*. Fundación de las Cajas de Ahorros.

Gobierno de España. (2017). *Declaración de la VI Conferencia de Presidentes*.

Gobierno de España. (2020). *Real Decreto 2/2020, de 12 de enero, por el que se reestructuran los departamentos ministeriales*. Boletín Oficial del Estado.

Gobierno de España. (2021). *Real Decreto 691/2021, de 3 de agosto, por el que se regula la concesión directa de ayudas para actuaciones de rehabilitación energética en edificios existentes en municipios de reto demográfico (Programa PREE 5000)*. Boletín Oficial del Estado, núm. 185.

Gómez Valenzuela, V., & Holl, A. (2024). Growth and decline in rural Spain: an exploratory analysis. *European Planning Studies*, *32*(2), 430–453. https://doi.org/10.1080/09654313.2023.2179390

Instituto Nacional de Estadística. (s. f.). *Estadística continua de población*. https://ine.es/dyngs/INEbase/operacion.htm?c=Estadistica_C&cid=1254736177095&menu=resultados&idp=1254735572981

Instituto para la Diversificación y Ahorro de la Energía. (s. f.). *Municipios de reto demográfico: Programa PREE 5000*. https://www.idae.es/ayudas-y-financiacion/para-la-

rehabilitacion-de-edificios/programa-pree-5000-rehabilitacion/municipios-de-reto-demografico

López-Penabad, M., Iglesias-Casal, A., & Rey-Ares, L. (2022). Measuring rural sustainable development at the municipal level: A composite index approach for Galicia (Spain). *Sustainability*, *14*(6), 1–21. https://doi.org/10.3390/su14063561

Macarrón, A. (2017). *El suicidio demográfico de España*. Homo Legens.

Martínez-Puche, A., Martínez Puche, S., García Delgado, F. J., & Amat Montesinos, X. (2022). La representación del éxodo rural en el cine español (1900-2020): evolución, causas y consecuencias territoriales. *Investigaciones Geográficas*, (77), 79–101. https://doi.org/10.14198/INGEO.19337

Ministerio de Política Territorial y Memoria Democrática. (s. f.). *Registro de Entidades Locales*. https://registroentidadeslocales.mpt.es/REL/frontend/inicio/index

Ministerio para la Transición Ecológica y el Reto Demográfico. (2021). *Plan de Recuperación: 130 medidas frente al Reto Demográfico*. Gobierno de España.

Ministerio para la Transición Ecológica y el Reto Demográfico. (2025). *Informe de la consulta pública previa para la nueva Estrategia Nacional para la Equidad Territorial y el Reto Demográfico 2030*. Gobierno de España.

Naciones Unidas. (2015). *Transformar nuestro mundo: la Agenda 2030 para el Desarrollo Sostenible*. Asamblea General de las Naciones Unidas. https://www.un.org/sustainabledevelopment/es/sustainable-development-goals/

Red de Áreas Escasamente Pobladas del Sur de Europa. (2024). *Documento de posición para Europa*. https://sspa-network.eu/wp-content/uploads/Documento-de-Posicion-SSPA2024EU.pdf

Secretaría General para el Reto Demográfico. (s. f.). *Sistema Integrado de Datos Municipales (SIDAMUN): Metodología y manual de usuario*. Gobierno de España. https://public.tableau.com/views/SistemaIntegradodeDatosMunicipales2023/Portada

Senado de España. (2015). *Ponencia de estudio para la adopción de medidas en relación con la despoblación rural*. Comisión General de las Comunidades Autónomas.

Sosa Troya, M. (2019, 31 de marzo). La 'España vaciada' clama por una gran alianza contra la despoblación. *El País*. https://elpais.com/sociedad/2019/03/31/actualidad/1554022545_649884.html

Universidad de Zaragoza. (s. f.). *Legislación sobre el Reto Demográfico y el Desarrollo Territorial en España*. https://educacionterritorio.unizar.es/legislacion-reto-demografico-desarrollo-territorial/

Verón Lassa, J. J., & Hernández Ruiz, J. (2021). Los términos "España vacía, vaciada y despoblada": significado y presencia en la conversación mediática. ESIC University / Universidad San Jorge.

Anexos

Anexo I. Repositorio de código

El código completo del análisis desarrollado en este trabajo, incluyendo la carga y preparación de datos, el análisis exploratorio, los modelos de proyección de población, el Random Forest Regressor y el clustering de municipios, está disponible públicamente en el siguiente repositorio de GitLab:

https://gitlab.com/tfg-ba/tfg-ba-espana-vaciada

El repositorio contiene un único cuaderno de Jupyter (TFG_BA.ipynb) con todo el desarrollo técnico comentado y estructurado por capítulos, siguiendo el orden del presente trabajo. Los datos originales no se incluyen por razones de tamaño, pero las fuentes de acceso público están detalladas en la sección de fuentes de datos y variables.