UIMP
Universidad Internacional
Menéndez Pelayo

ANÁLISIS DE LOS PATRONES DE MOVILIDAD RESIDENCIAL EN ESPAÑA MEDIANTE ÁRBOLES DE CLASIFICACIÓN

ANALYSIS OF RESIDENTIAL MOBILITY PATTERNS IN SPAIN USING CLASSIFICATION TREES

Trabajo de Fin de Máster
para acceder al

MÁSTER EN CIENCA DE DATOS
MASTER IN DATA SCIENCE

Autor: Matin Khakpour

Directores: Francisco Matorras Weinig

Olga de Cos Guerra

Julio de 2021

Resumen:

Este trabajo estudia una fuente demográfica principal en España, las Estadísticas de Variaciones Residenciales (EVR); un informe anual del Instituto Nacional de Estadística (INE) que abarca todos los cambios residenciales comunicados por los individuos cuando hay modificación en su municipio de residencia. Usando las herramientas de Ciencia de Datos se realiza un análisis descriptivo de esta fuente. Después se procede a enlazarlo con más de 70 variables censales, padronales y territoriales. Aplicando una serie de técnicas de Machine Learning, en concreto los árboles de decisión, como un método de aprendizaje supervisado, se intenta detectar los patrones más influyentes a los flujos migratorios entre los municipios españoles. Se intenta aumentar el poder predictivo mediante diferentes modelos, muestreos y categorizaciones. Al final se detectan un par de variables con relativa importancia sobre el patrón de cambios hacia municipios grandes y pequeños.

Palabras clave:

Fuentes demográficas; Migraciones interiores; Aprendizaje automático; Árboles de clasificación

Abstract:

This paper studies a principal demographic resource in Spain, the Residential Variation Statistics (EVR in Spanish); an annual report compiled by the National Institute of Statistics (INE) which covers all residential changes reported by individuals when there is a change in their municipality of residence. Using Data Science tools, a descriptive analysis of this source is carried out. Then proceeds to its aggregation with more than 70 demographic and territorial variables gathered from census and municipal registers. Applying a series of Machine Learning techniques, specifically Decision Trees, as a supervised learning method, an attempt is made to detect the most influential patterns of migration flows between Spanish municipalities. We try to increase the predictive power through different models, sampling, and categorizations. Finally, a couple of variables with relative importance on the pattern of changes toward big and small municipalities are introduced.

Key words:

Demographic sources; Internal migration; Machine Learning; Classification Trees

Índice

1. Introducción

1.1. El contexto de EVR

Estudiar los cambios demográficos y territoriales de una sociedad es una tarea multidisciplinaria que requiere un abanico de conocimientos geográficos, estadísticos, sociológicos y económicos.

Estos estudios resultan de interés, no solo en el ámbito académico de Ciencias Sociales, sino de cara a la administración pública y establecer políticas territoriales a nivel nacional, autonómico o municipal, con el fin de combatir la vulnerabilidad geográfica, la despoblación, visibilizar los efectos de envejecimiento poblacional y analizar la migración, entre otros objetivos.

El rasgo cambiante de la población en el territorio nacional es el enfoque de estudios que intentan detectar los procesos y patrones existentes para crear una base sobre la cual se pueda dar pautas y resultados estratégicos para la planificación de equipamientos, servicios, etc.

Entre los enfoques posibles, estudiar los flujos de migración, tanto interior como exterior, es un terreno de especial interés.

En la dinámica demográfica española, desde los años 80, se está observando una tendencia importante de descentralización (especialmente residencia periférica que predomina la inmigración intra-provisional) como resultado de la expansión metropolitana de grandes núcleos urbanos, sobre todo, Madrid y Barcelona (De Cos, 2007).

La migración de los españoles hacia otros países europeos como la consecuencia de la crisis económica produjo más de 1,334,000 bajas hacia extranjero durante los años 2008-2011 (Domingo y Sabater, 2013). La llegada de inmigrantes extranjeros a Europa durante las últimas décadas también ha sido el enfoque de muchos debates sociales y políticos.

Por otro lado, el surgimiento de conceptos como *España vacía* hace hincapié en la distribución no regular de flujos interiores y la despoblación de muchas parcelas del territorio nacional.

Se entiende por migración el cambio de residencia habitual de un individuo que implica atravesar algún tipo de división geográfica (Susino, 2012). Esta definición implica su doble dimensión temporal y espacial, por lo cual cada estudio sobre migración tiene que delimitar sus fuentes en base a sus dimensiones disponibles.

En España existen varias fuentes comunes de datos demográficos para estudiar la migración que principalmente provienen de los censos, los padrones municipales, la Encuesta de Migración (EM) y otros informes no periódicos sino puntuales.

La Estadística de Variaciones Residenciales (EVR) es un informe anual elaborado por el Instituto Nacional de Estadística (INE) a partir de la explotación de la información relativa a las altas y bajas por cambios de residencia registradas en los padrones municipales (Metodología EVR).

En cuanto a la dimensión espacial, una *variación residencial* es un cambio de residencia de un individuo de un municipio a otro, al extranjero o desde el extranjero. Sobre su dimensión temporal, tomando un año de referencia, las fechas reflejadas en cada publicación de EVR se extienden hasta el mes de marzo (inclusive) del año siguiente al de estudio.

Además de variaciones residenciales como tal, se incluyen en EVR registros de otro tipo como altas por omisión, bajas por inclusión indebida, bajas por caducidad, etc. pero tienen una proporción escasa.

Las variables que se registran sobre cada variación son el sexo, la fecha, el lugar de nacimiento (país si es en el extranjero) y la nacionalidad del ciudadano, además de la procedencia y el destino del movimiento. En el caso de los municipios españoles (de procedencia, destino o nacimiento) también hay una variable categórica indicando el tamaño poblacional del municipio.

Es importante señalar que los registros de EVR no hacen referencia al número de personas que llevan a cabo una variación residencial, sino al número de variaciones efectuadas, ya que un ciudadano puede cambiar su residencia de un municipio a otro más de una vez en un año.

1.2. Los objetivos de estudio

A pesar de las discrepancias entre las fuentes y sus inconsistencias, hay una imagen bastante precisa de los procesos migratorios que han afectado a España durante la segunda mitad del siglo XX y el principio del siglo XXI. (Susino, 2012). La Estadística de Variaciones Residenciales, además de ser la más utilizada (Susino, 2011) se erige como la fuente estadística más adecuada para estudiar las migraciones en España (Martí y Ródenas, 2006) tanto de la población española como extranjera (Domingo y Sabater, 2013).

Gracias a usar las altas padronales en EVR, el registro de la inmigración internacional e interna en España respecto a otros países es bastante eficiente. Pero, en comparación con la inmigración, la emigración tanto de nacionales como de extranjeros se caracteriza por su borrosidad estadística, siendo el cómputo de las bajas de limitada cobertura, incompleto y sesgado (Domingo y Sabater, 2013). Por esta razón se decidió enfocar este estudio en flujos interiores.

No es escasa la literatura que, sobre migraciones interiores, se ha producido en las últimas décadas. Sin embargo, son escasos los estudios que se basen en el análisis de flujos entre municipios, ya sea a base de los censos o bien las EVR (Susino, 2011).

A partir de esta premisa, la aspiración de este trabajo consiste en usar EVR como la fuente principal y enlazarlo con otras fuentes de datos demográficos para estudiar el patrón de la migración interior, es decir, variaciones residenciales cuyos ambos destino y procedencia son municipios del territorio español.

Para este objetivo se implementan las herramientas y softwares habituales de Ciencia de Datos para demostrar la potencialidad que pueden tener en aplicar a áreas como Ciencias Sociales. Además, como una aportación analítica, se aprovecha de los métodos de aprendizaje automático, concretamente los árboles de clasificación, para intentar detectar los patrones de movimiento.

No se pueden utilizar las EVR para analizar la evolución de las migraciones a largo plazo debido a sus cambios metodológicos en diferentes etapas (Susino, 2011). Por consiguiente, este estudio se delimita a la última publicación disponible de EVR a la hora de realizar este análisis, la del año 2019.

2. Análisis descriptivo de EVR

2.1. Estructura de la fuente

Los datos de EVR se publican en el portal del INE (www.ine.es), bajo la sección de «Demografía y población», la subsección de «Padrón. Población por municipios». Para obtener todos los registros sin categorización previa, en el apartado de «Microdatos» se puede descargar una carpeta zip con el conjunto de ficheros del año seleccionado. En esta carpeta existe un diccionario de variables donde explica los metadatos usados en EVR. Los datos están preparados en diferentes formatos como `txt` de ancho fijo, `csv` y otros formatos utilizables en softwares como SAS, SPSS y STATA.[1]

Para este trabajo, se optó por usar el lenguaje de programación R como una herramienta versátil de análisis estadístico. El trabajo se ha desarrollado en un entorno de Jupyter Notebook con kernel de R y el código es accesible en repositorios abiertos.

De la carpeta de microdatos del año 2019, se ha usado el fichero `csv` (de hecho, es un `tsv` porque tiene como separador la tabulación). El fichero tiene 2,868,942 filas y 16 columnas. La estructura de las variables se ve en la tabla 1 (se puede consultar sus definiciones en el diccionario del anexo):

```
$ SEXO    : num [1:2868942] 1 6 1 1 6 6 1 1 6 6 ...
$ PROVNAC : chr [1:2868942] "01" "01" "01" "01" ...
$ MUNINAC : chr [1:2868942] NA NA NA NA ...
$ EDAD    : num [1:2868942] 4 1 36 26 37 75 14 11 28 9 ...
$ MESNAC  : num [1:2868942] 4 4 6 3 6 1 7 11 10 1 ...
$ ANONAC  : num [1:2868942] 2015 2018 1982 1993 1982 ...
$ CNAC    : num [1:2868942] 108 108 108 108 108 108 108 108 108 108 ...
$ PROVALTA: chr [1:2868942] "01" "01" "01" "01" ...
$ MUNIALTA: chr [1:2868942] NA NA NA "059" ...
$ MESVAR  : num [1:2868942] 7 7 3 5 6 1 9 1 12 1 ...
$ ANOVAR  : num [1:2868942] 2019 2019 2019 2019 2019 2019 ...
$ PROVBAJA: chr [1:2868942] "01" "01" "01" "01" ...
$ MUNIBAJA: chr [1:2868942] NA NA NA NA ...
$ TAMUALTA: num [1:2868942] 1 1 1 6 6 6 6 6 6 6 ...
$ TAMUBAJA: num [1:2868942] 1 1 1 1 1 1 1 1 1 1 ...
$ TAMUNACI: num [1:2868942] 1 1 1 1 1 1 1 1 1 1 ...
```

Tabla 1. Estructura inicial de variables del dataset EVR_2019 (en entorno R)

Los códigos de provincias y municipios de nacimiento, alta y baja son del tipo carácter (string) y el resto son numéricos (enteros).

Algunas variables tienen valores distintivos para España/Extranjero:

PROVNAC, PROVALTA, PROVBAJA: "66" = extranjero

CNAC: "108" = nacionalidad española

[1] Se debe mencionar que el INE no se responsabiliza de los resultados que los receptores de los datos obtengan a partir de estos ficheros basados en sus propios cálculos.

Para los valores correspondientes a otros países hay que recurrir al diccionario de EVR.

Las variables de tamaño[2] del municipio de nacimiento, alta o baja cogen este rango de valores:

1	Municipio no capital hasta 10.000 habitantes
2	Municipio no capital de 10.001 a 20.000
3	Municipio no capital de 20.001 a 50.000
4	Municipio no capital de 50.001 a 100.000
5	Municipio no capital de más de 100.000
6	Municipio capital de provincia

Tabla 2. Categorías poblacionales (variables TAMUALTA y TAMUBAJA)

2.2. Tratamiento inicial de los valores nulos

Se hace un contaje de los campos nulos de cada variable:

SEXO	0
PROVNAC	0
MUNINAC	224362
EDAD	0
MESNAC	0
ANONAC	0
CNAC	0
PROVALTA	0
MUNIALTA	505738
MESVAR	0
ANOVAR	0
PROVBAJA	0
MUNIBAJA	424934
TAMUALTA	345749
TAMUBAJA	873842
TAMUNACI	1614660

Tabla 3. El número de registros nulos por cada variable en el dataset EVR_2019.

La alta cantidad de nulos se debe a que, por cuestiones de anonimato de datos personales, si alguien ha nacido o se ha dado de alta o baja en un municipio de categoría 1 (población inferior a 10.000 habitantes), se ha blanqueado el código del municipio correspondiente.

En el caso de tamaño de municipio de nacimiento, además la variable tiene valor nulo para todos los nacidos en extranjero, porque la población de un país no es comparable con los intervalos poblacionales de un municipio.

Se convierten las variables oportunas al tipo categórico, pero se incluyen los nulos coma una clase más, porque luego se necesita calcular su proporción al respecto de los demás valores. El

[2] El tamaño del municipio se calcula a fecha 1 de enero del año al que hace referencia la EVR correspondiente.

comando `addNA()` permite hacer ambas cosas a la vez, sin cambiar las etiquetas. Este procedimiento se aplica a MUNINAC, MUNIALTA, MUNIBAJA, TAMUALTA, TAMUBAJA, TAMUNACI.

Los códigos provinciales (PROVNAC, PROVALTA, PROVBAJA) se mantienen en formato carácter, porque más adelante se necesita hacer alguna transformación textual con ellos.

El resto de las variables numéricas en esta fase se usan sin factorizar (SEXO, EDAD, MESNAC, ANONAC, CNAC, MESVAR, ANOVAR).

2.3. Visualización de variables

En primer paso, se visualizan los valores de cada variable, de manera más simple posible, para obtener una idea de partida sobre su distribución y encontrar las posibles anomalías.

Tal como son las codificaciones, esto no tiene sentido para algunas variables como MUNINAC porque, para los nacidos en España se codifica el municipio de nacimiento y para los nacidos en el extranjero se codifica el país de nacimiento. Además, en el caso de España, este código municipal no es el código estándar de 5 dígitos, entonces habrá que juntarlo con su correspondiente código de provincia (lo mismo para MUNIALTA y MUNIBAJA[3]).

También se descarta considerar las variables ANONAC (porque hay otra para la edad) y ANOVAR (porque es igual para todos los registros de un fichero anual).

Para el resto, en el caso de variables continuas se usa histograma y para las variables discretas se usa barplot o pie chart. Las gráficas son las siguientes:

SEXO:

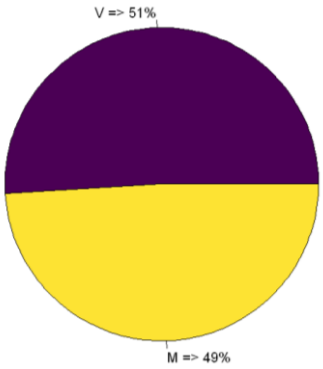

Figura 1. Proporción de varones y mujeres[4]

[3] Se puede hacer el mismo análisis sobre subconjuntos correspondientes. Domingo y Sabater (2013) han desarrollado este tipo de estudio sobre los principales países de origen y de destino en la inmigración española.

[4] Todas las figuras usadas en este trabajo son de elaboración propia y creadas con lenguaje R.

Provincia de nacimiento (PROVNAC):

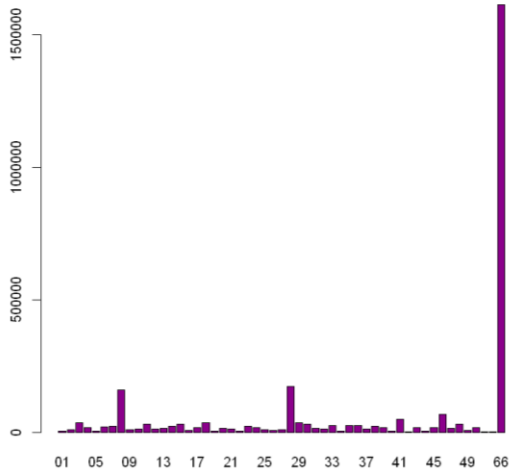

Figura 2. Frecuencia de cada código provincial de lugar de nacimiento

Se observa en la figura 2 que un alto porcentaje de todas las variaciones son de personas nacidas en extranjero (PROVNAC = 66). La proporción exacta es 56.3%.

EDAD:

Se ha añadido al histograma la media de la población española en 2019 según los datos del INE. También se ha marcado una curva de distribución normal con la misma media y desviación estándar que la muestra de EDAD:

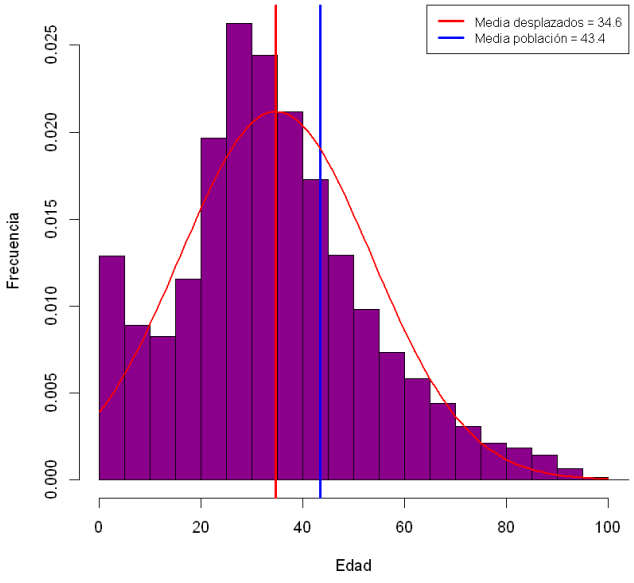

Figura 3. Histograma de edad

Parece llamativa la diferencia de casi 9 años entre la edad media de los desplazados (34.6) y la población general (43.4). La estructura de la población española es envejecida, entonces se infiere que las edades más avanzadas forman menos parte de los cambios residenciales.

Mediante un test de hipótesis, se comprobó si la diferencia observada es estadísticamente significativa. En principio se contempló usar el test de t-student (aunque es para variables continuas, teniendo una muestra bastante grande y basándose en el teorema central del límite, se podría proceder a usarlo) pero al ver la curva normal pareció que la muestra no cumple la condición de normalidad. Mediante la figura 4 y test de Shapiro se comprobó que efectivamente la muestra no tiene distribución normal:

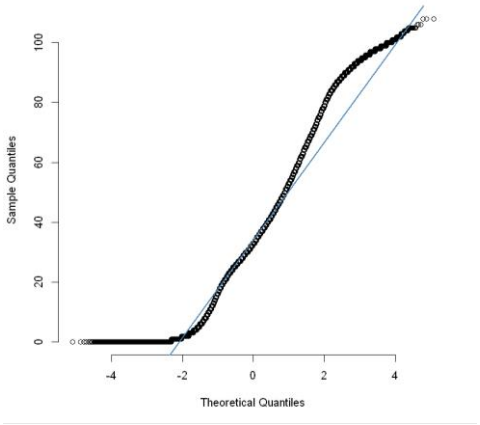

Figura 4. Q-Q plot de la muestra para comparar con la distribución normal

```
        Shapiro-Wilk normality test

data:  sample(db$EDAD, 5000)
W = 0.977, p-value < 2.2e-16
```

(En el Shapiro, la hipótesis nula es la normalidad de la distribución. Entonces, obteniendo un p-valor bastante menor a 0.05, se puede rechazarla.)

Entonces, habría que usar otro test no-paramétrico apto para muestras no-normales. Teniendo en cuenta que solo existe una muestra y el único parámetro poblacional conocido es la media, se usó el test de One-Sample Wilcoxon Signed Rank:

$$H_0\colon \mu_d = \mu_p = 43.4$$
$$H_1\colon \mu_d < \mu_p$$

```
        Wilcoxon signed rank test with continuity correction

data:  db$EDAD
V = 1.0147e+12, p-value < 2.2e-16
alternative hypothesis: true location is less than 43.4
```

Se debe aceptar la hipótesis alternativa y afirmar que la media de edad de la población migrante es significativamente menor que la de la población general.

Este hecho puede indicar que se está analizando una muestra de población en edad de trabajar y posiblemente muchos de los cambios se efectúan por motivos laborales. También puede ser que existen muchos jóvenes y adulto-jóvenes con estructuras de hogares nucleares con hijos/as entre

la muestra, porque hay una subida en edades más bajas de la pirámide (los primeros bins de la derecha que justo distorsionan el histograma de la curva normal).

Código de nacionalidad (CNAC):

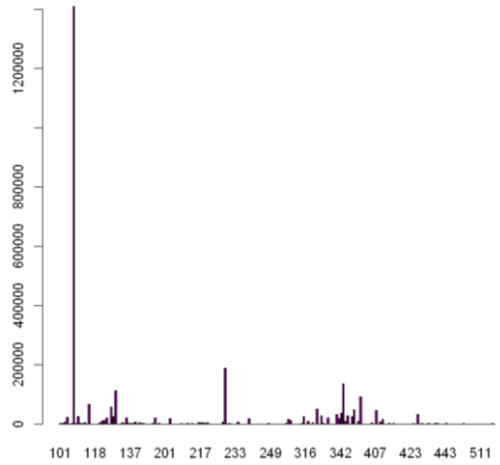

Figura 5. Frecuencia de cada código de nacionalidad

El bin más alto es el del código 108 (España). La proporción es 49.1% (nacionalidad española) versus 50.9% (nacionalidad extranjera). Suponiendo que en España unos 12% de la población general es extranjera, es una evidencia de que, en el caso de las variaciones residenciales, los extranjeros tienen una proporción bastante más alta que su representación demográfica. Ródenas y Martí (2006) afirman una tendencia similar (sobrepresentación de extranjeros en los cambios residenciales) en su estudio de los EVR desde el año 1996 hasta 2004.

Provincia de alta (PROVALTA):

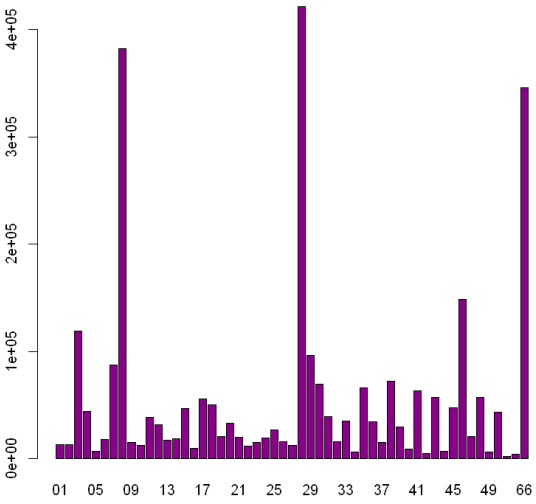

Figura 6. Frecuencia de cada código provincial de lugar de alta

El último bin de la derecha corresponde a variaciones con destino extranjero; 12.1% del total.

Provincia de baja (PROVBAJA):

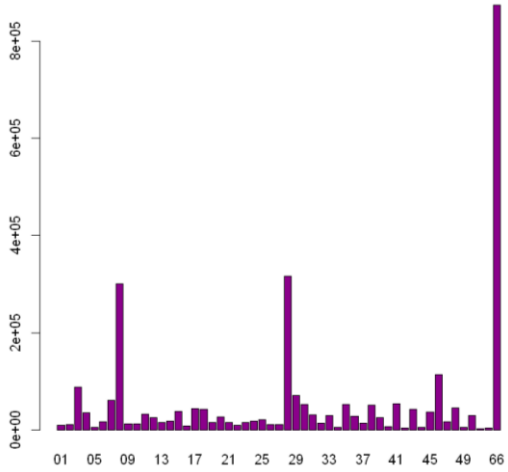

Figura 7. Frecuencia de cada código provincial de lugar de baja

En el caso de las variaciones con origen en extranjero, la proporción es 30.5% del total.

Mes de variación (MESVAR):

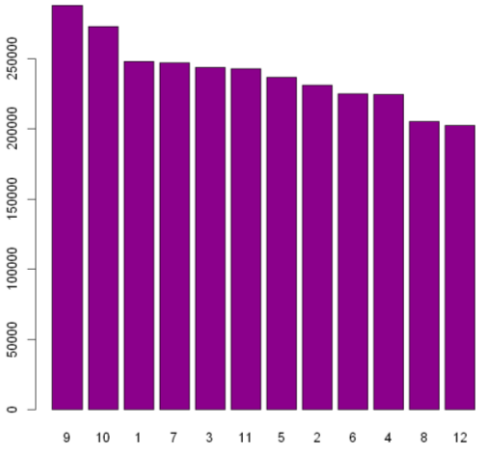

Figura 8. Frecuencia de variaciones por mes

Aunque la mayor representación tiene desarrollo no excesivo, la mayoría de las variaciones se realizan durante los meses de septiembre y octubre. Una posible motivación puede ser el comienzo del curso escolar en el caso de los estudiantes y núcleos familiares con hijos. También coincide con los ciclos laborales tras el verano; nuevas incorporaciones y fines de contratos estivales.

Tamaño de municipio de alta (TAMUALTA):

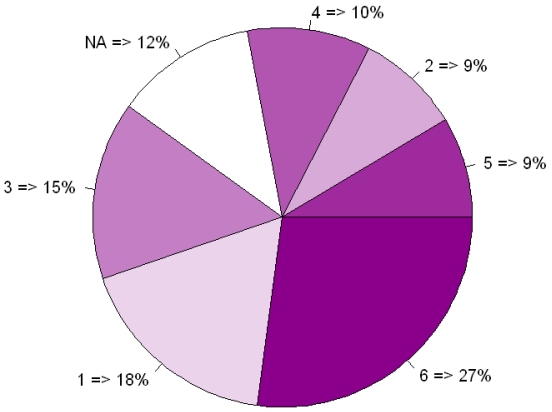

Figura 9. Proporción de categorías poblaciones de municipios de alta
(La intensidad de color representa mayor población)

La definición de intervalos de la figura 9 se dio en la tabla 2. Los nulos (NA) corresponden a destinos en extranjero.

De la figura 9 se puede deducir una tendencia importante. Era de esperar que los municipios capitales de provincia sean el primer destino de las variaciones. Pero resulta que los municipios de categoría 1 (menos de 10,000 habitantes) son el segundo destino más frecuentado en las variaciones. Es decir, la tendencia se halla en los extremos de categorías poblacionales. Esto se puede deber a que buena parte de esos municipios estarán dentro de algún sistema metropolitano. Entonces, se mezcla la capacidad de atracción de los grandes núcleos urbanos con el proceso de periurbanización centro-periferia.

Tamaño de municipio de baja (TAMUBAJA):

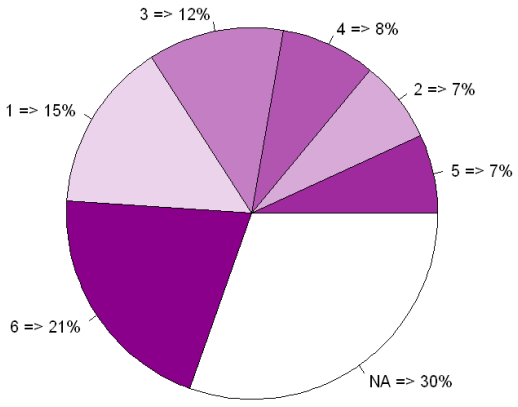

Figura 10. Proporción de categorías poblacionales de municipios de alta
(La intensidad de color representa mayor población)

Hay una considerable proporción de variaciones con origen extranjero (casi un tercio del total). También es importante observar que, dentro de España, la tendencia poblacional entre los municipios de baja es la misma que los municipios de alta; por orden descendente:

6 -> 1 -> 3 -> 4 -> 2 -> 5

Esto indica el mencionado proceso de descentralización de grandes núcleos urbanos, hasta el punto que en algunos casos como Madrid y Barcelona desafía la delimitación del espacio urbano-metropolitano y requiere otras perspectivas de categorización para reflejar la realidad territorial (Reques y De Cos, 2013).

Tamaño de municipio de nacimiento (TAMUNACI):

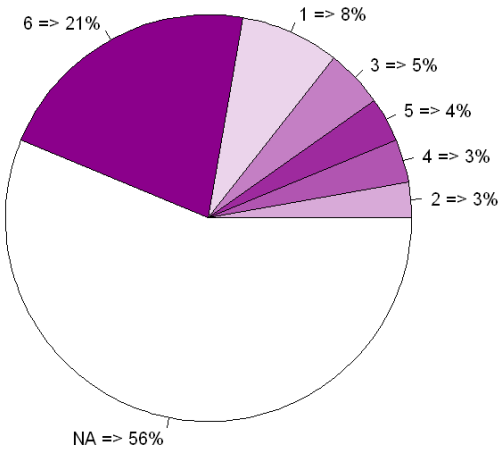

Figura 11. Proporción de categorías poblacionales de municipios de nacimiento
(La intensidad de color representa mayor población)

Se comentó en la figura 2 que 56.3% de las personas migrantes son nacidos en extranjero y entonces en la figura 11 se desconoce la información poblacional de su lugar de nacimiento.

2.4. Análisis de subconjuntos de interés

En el segundo paso, se producen otras visualizaciones mediante subconjuntos y agrupaciones de variables, de manera que puedan aportar alguna información añadida o perspicacia nueva.

Se pretende ver si hay alguna relación significativa entre cada par de origen y destino a nivel provincial (tanto dentro de España como con el extranjero). Para esto se crea una tabla de frecuencia cruzada entre cada par de provincias como origen y como destino mediante el comando `xtabs` y a base de esa tabla se crea un `heatmap` de la figura 12:

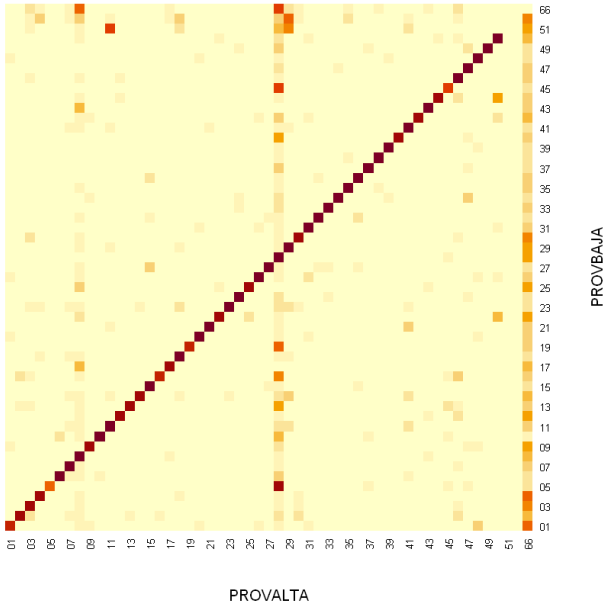

Figura 12. Frecuencia de variaciones entre cada par de provincias
(Rojo intenso = frecuencia alta. Amarillo = frecuencia baja)

A simple vista, se entiende de la diagonal de la figura 12 que muchas de las variaciones se hacen dentro de una misma provincia. 34.7% del total de variaciones son INTRA-provinciales. Se intenta ver ¿en qué provincias son más frecuentes las variaciones intra-provinciales?

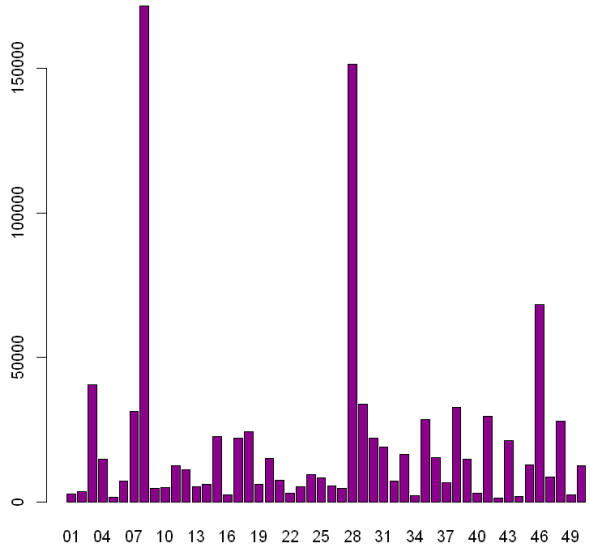

Figura 13. Frecuencia de variaciones intra-provinciales por cada código de provincia

Las principales provincias con flujo interior son 08 (Barcelona), 28 (Madrid) y 46 (Valencia).

Luego se dividen las variaciones INTER-provinciales (65.3%) en otro subconjunto:

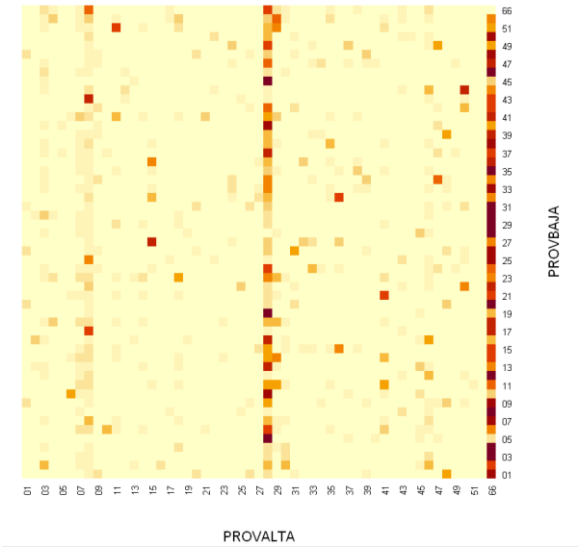

Figura 14. Frecuencia de variaciones inter-provinciales
(Rojo intenso = frecuencia alta. Amarillo = frecuencia baja)

Algunas inferencias:

- La primera fila horizontal muestra que las variaciones con origen extranjero mayormente se afincan en provincias 28 (Madrid) y 08 (Barcelona).
- La columna del medio muestra que Madrid es -con diferencia- el destino más frecuentado y que sus llegadas provienen de todas provincias, pero siendo algunas más destacadas como 05 (Ávila), 19 (Guadalajara) y 45 (Toledo). Es decir, sus provincias vecinas. Esto confirma que el proceso de metropolización de Madrid excede su límite autonómico, configurando una gran región metropolitana.[5]
- La primera columna de la derecha, muestra que las variaciones con destino extranjero están extendidas entre muchas provincias de origen. Se visualizan en concreto para averiguar que las personas que se van al extranjero ¿desde qué provincias se van?

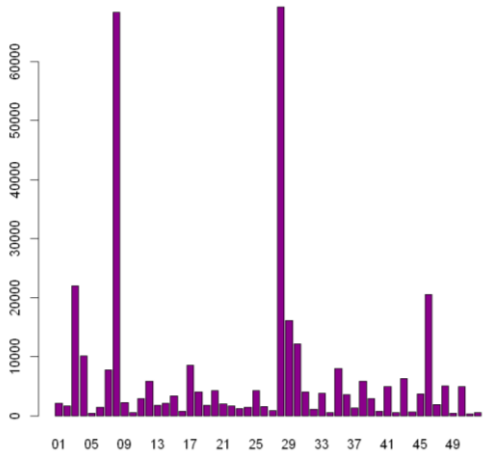

Figura 15. Frecuencia de cada código provincial como origen de variaciones con destino extranjero

[5] García y Pozo (2010) han estudiado este efecto definiendo 6 "coronas" metropolitanas y periurbanas para la comunidad, analizando además el peso de emigración versus inmigración en este fenómeno.

Después se hace la inversa; las personas que vienen del extranjero ¿en qué provincias se afincan? Pero se combinan estos datos con los de la figura 15 para notar además la diferencia entre frecuencias de cada caso:

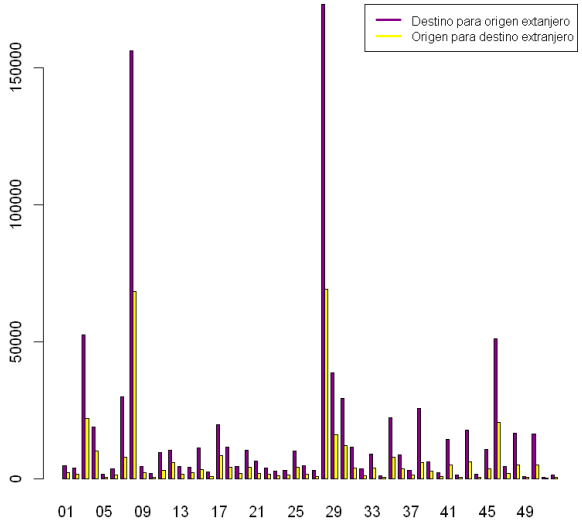

Figura 16. Frecuencia de cada código provisional para variaciones con origen o destino extranjero (Morado = destino para origen extranjero. Amarillo = origen para destino extranjero)

Es notable que la tendencia es muy parecida en ambos casos; los mayoritarios son 28 (Madrid), 08 (Barcelona), 46 (Valencia) y 03 (Alicante). Pero, la proporción de origen/destino es diferente; hay bastantes más variaciones con origen extranjero que con destino extranjero.

En la figura 3 se contempló la distribución de edad de la población migrante, pero también se puede comprobar la edad según la categoría poblacional del municipio de destino. Es decir, en cada franja de edad, ¿a qué tipo de municipios se mudan las personas?

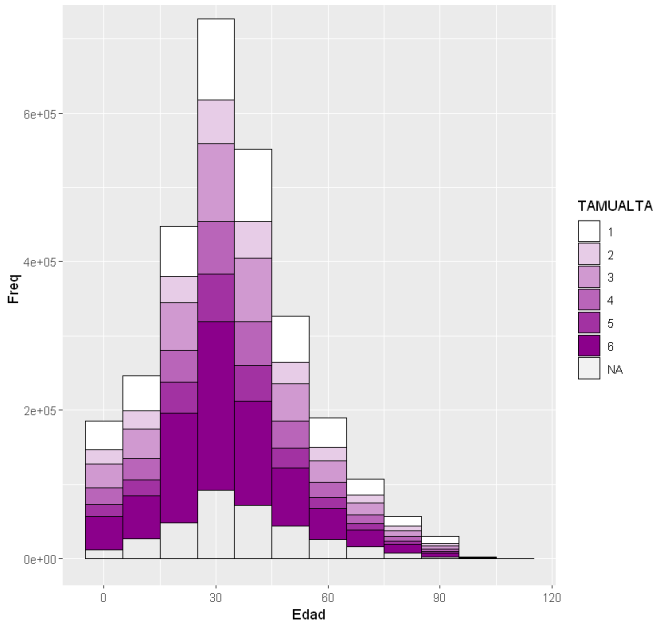

Figura 17. Histograma de edad de desplazados segmentado por tamaño de municipio de destino (La intensidad de color representa mayor población)

Es notable en la figura 17 que las categorías poblacionales se mantienen casi en todas franjas de edad. Es decir, todas las personas, incluso los más jóvenes, tienden a mudarse primero hacia capitales de provincia y luego hacia municipios muy pequeños.

La pregunta inversa es, en cada franja de edad, ¿desde qué tipo de municipio se mudan?

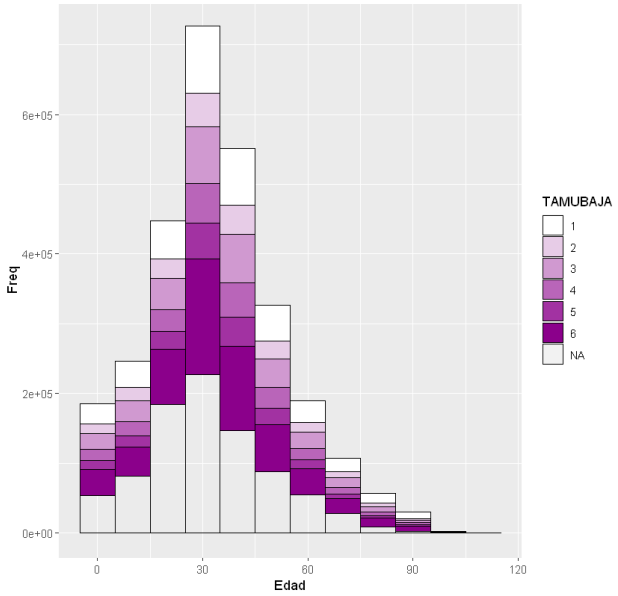

Figura 18. Histograma de edad de desplazados segmentado por tamaño de municipio de origen
(La intensidad de color representa mayor población)

En la figura 18 se ve el efecto de los nulos por la alta proporción de bajas en el extranjero.

Se debe mencionar que todo lo relacionado con «extranjero» se puede definir, por lo menos, según estos tres aspectos:

- Lugar de alta o baja
- Lugar de nacimiento
- Nacionalidad

La gráfica de origen/destino provincial en variaciones con origen/destino extranjero se presentó en la figura 16. Sobre el lugar de nacimiento y la nacionalidad se puede pensar en 4 subconjuntos, de las cuales mayormente se refiere al tipo 4:

1) Nacidos en extranjero (independientemente de su nacionalidad)
2) Extranjeros nacidos en España (hijos de residentes extranjeros)
3) Españoles nacidos en extranjero (extranjeros nacionalizados o hijos de españoles residentes en extranjero)
4) Extranjeros nacidos en extranjero (estudiantes, trabajadores, etc.)

Se compruebe sus proporciones (no del total, porque tienen solapo, sino comparativamente):

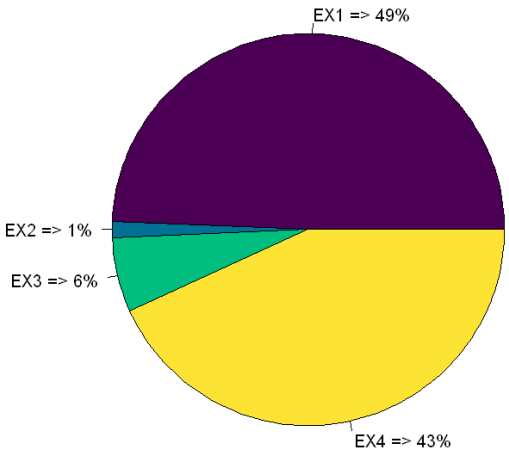

Figura 19. Proporción de categorías relacionadas con extranjero

Se puede calcular cualquier métrica de las anteriores u otras posibles sobre estos subconjuntos dependiendo de lo que sea relevante para otros estudios.

Siguiendo con la potencialidad de subconjuntos, se puede crear categorías sobre una provincia o un municipio. Por ejemplo, se puede ver la distribución de edad de los cántabros (ciudadanos con nacionalidad española y nacidos en Cantabria) que se han ido de Cantabria en comparación con toda la población migrante:

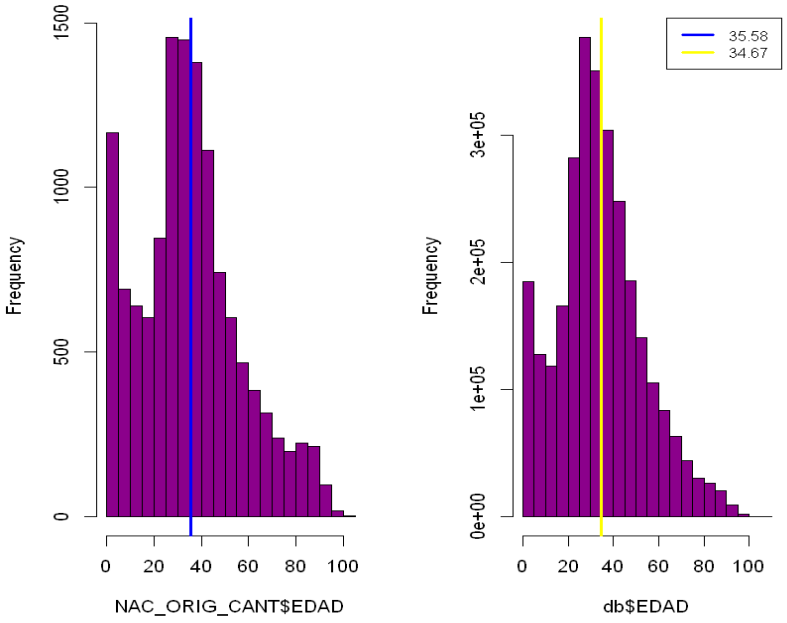

Figura 20. Histogramas de edad de los cántabros migrantes (izq.) y toda la muestra (dcha.)

Mediante un test de Kolmogórov-Smirnov (KS) se averigua la semejanza de distribuciones entre estas dos muestras:

```
        Two-sample Kolmogorov-Smirnov test

data:  NAC_ORIG_CANT$EDAD and db$EDAD
D = 0.044155, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Se rechaza la hipótesis nula (la igualdad de distribuciones). También se puede comprobarlo con la función de distribución acumulada de cada muestra:

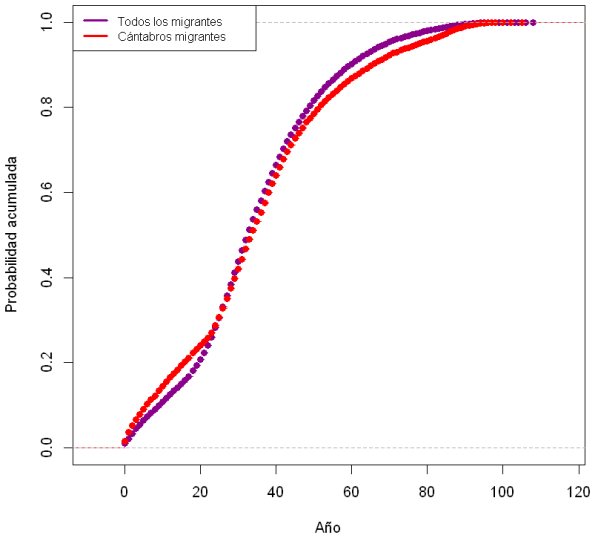

Figura 21. Función de distribución acumulada empírica para la variable EDAD
(Morado = todos los migrantes. Rojo = cántabros migrantes)

En la figura 21 las curvas se ven cercas, porque la distancia resultante de la prueba KS era pequeña (0.044).

Además, se puede comparar las medias:

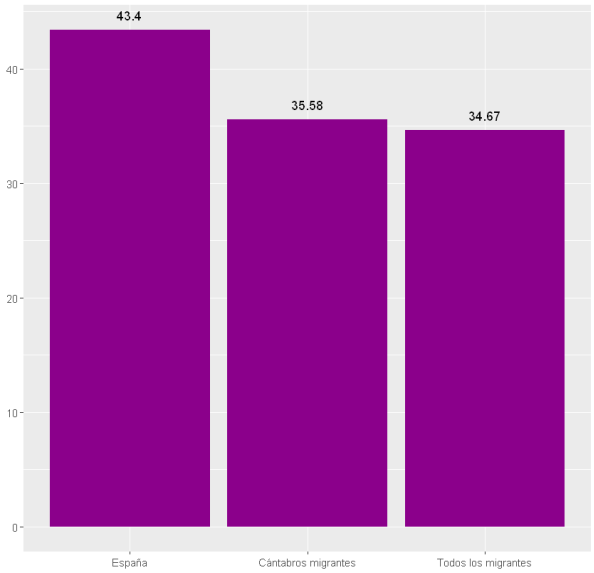

Figura 22. Comparación de edad media de la población española (izq.), los cántabros desplazados (cent.) y toda la población migrante (dcha.)

Se evalúa si la edad media de los cántabros migrantes (35.58) es significativamente mayor a la media de total de migrantes (34.67):

$$H_0: \mu_c = \mu_m$$
$$H_1: \mu_c > \mu_m$$

(Probado con Shapiro que la distribución no es normal, se usa el Wicoxon:)

```
        Wilcoxon signed rank test with continuity correction

data:  NAC_ORIG_CANT$EDAD
V = 40962412, p-value = 0.808
alternative hypothesis: true location is greater than 34.66642
```

Como p-valor es mayor que 0.05, no se puede afirmar que la edad media de los cántabros migrantes sea significativamente más alta que la edad media de todos los migrantes y puede que estemos observando básicamente el efecto de muestreo.

Como se hizo dos subconjuntos para variaciones INTER e INTRA provinciales, se puede también combinarlos en el análisis. Por ejemplo, para ver si hay alguna diferencia en cuanto al mes de variaciones con destino Santander, según la persona viene desde otra parte de Cantabria o bien desde otras provincias:

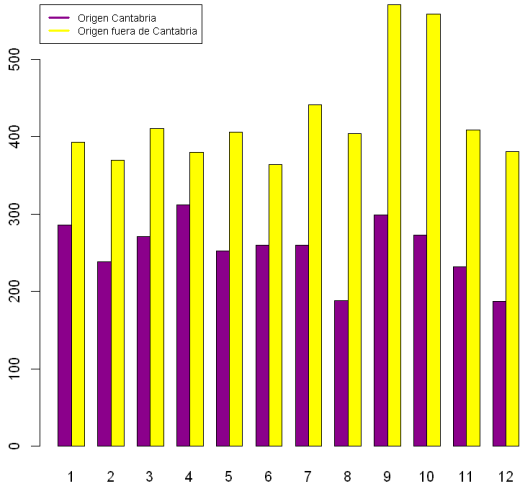

Figura 23. Comparación de frecuencia de variaciones con destino Santander
(Morado = con origen en Cantabria. Amarillo = con origen fuera de Cantabria)

Otro subconjunto se puede hacer para comparar el perfil de las variaciones hacia capitales versus cualquier otro tipo de municipio. Para esto se descartan los valores nulos (altas en extranjero) y luego se compara la distribución de alguna variable. En la figura 24, se ven los histogramas de edad para estos dos grupos:

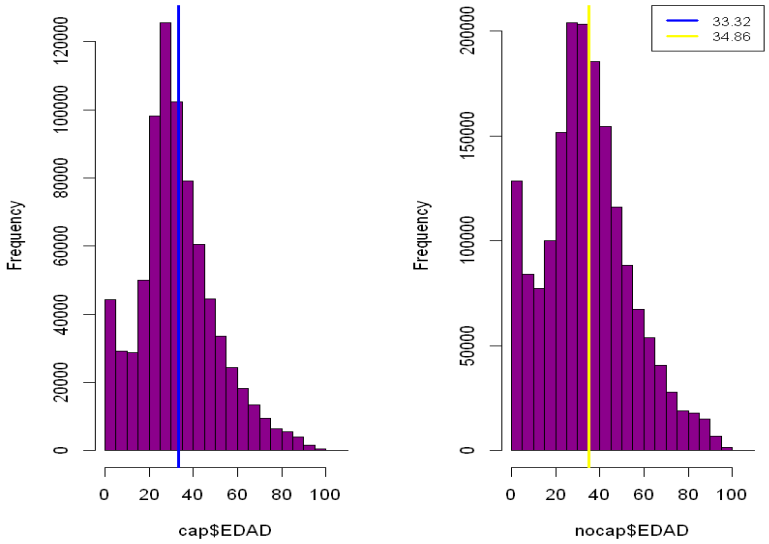

Figura 24. Histograma de edad de migrantes a capitales de provincia (izq.) frente a migrantes a municipios no capitales (dcha.)

En la prueba de Kolmogorov-Smirnov se rechaza la semejanza de distribuciones. Las curvas de FDA teniendo una *distancia* de separación pequeña (0.07) se ven en la figura 25:

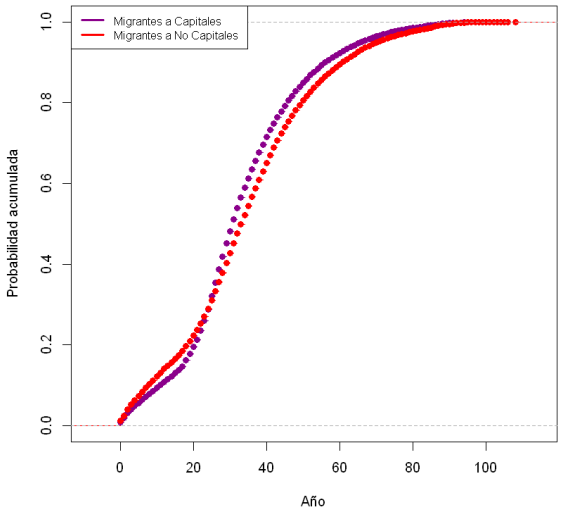

Figura 25. Función de distribución acumulada empírica para la variable EDAD (Morado = migrantes a capitales. Rojo = migrantes a no capitales)

También parecía que la media de edad de los migrantes a capitales (33.3) es menor a la media de migrantes a municipios no capitales (34.9). Aunque mediante los test se afirmó que la diferencia es real, la cantidad de diferencia (1.6 años) no parece muy significativa.

Esta tendencia es diferente entre los españoles y los extranjeros. Se comprobó en la figura 26 que los extranjeros están más orientados hacia las capitales de provincia:

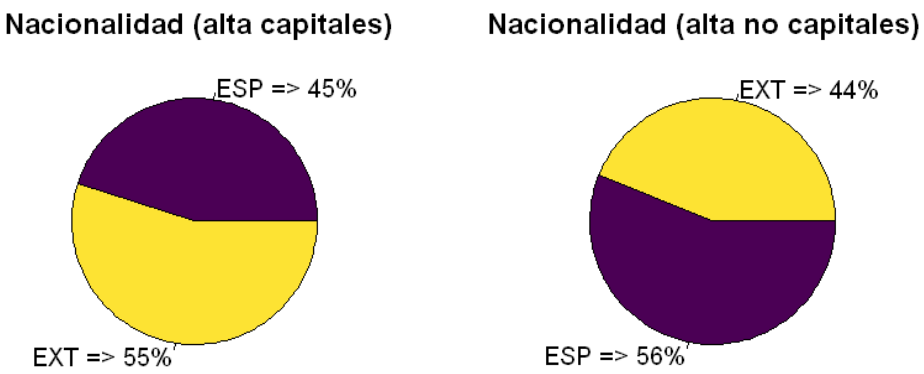

Figura 26. Proporción de altas en municipios capitales (izq.) y no capitales (dcha.) según la nacionalidad.

La misma comparación para otra variable como mes de variación:

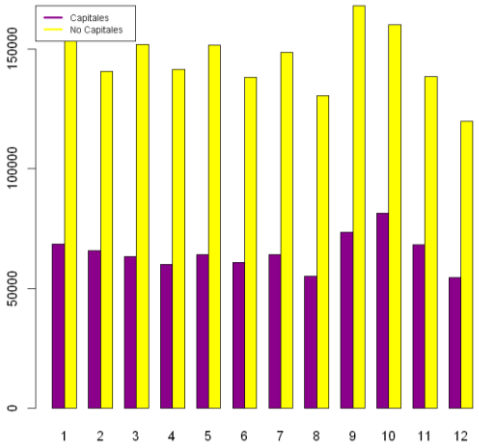

Figura 27. Frecuencia de variaciones por mes según tipo de municipio (Morado = capital de provincia. Amarillo = no capital)

Otra vez, parece que la mayoría de las variaciones de ambos subconjuntos se realizan durante los meses de septiembre y octubre.

Usando el mismo esquema de código, se puede extender y enfocar el análisis con otros subconjuntos y crear visualizaciones de interés.

3. Agregación de bases de datos

3.1. Descarga y preparación de datasets adicionales

Después de haber hecho una aproximación analítica sobre el dataset EVR 2019 en la fase anterior, se procede a vincularlo con otros datasets y alimentar el análisis con nuevas variables que aportan estas fuentes.

Los datasets que inicialmente se incorporaron en el estudio son de *Atlas de distribución de renta de los hogares* enmarcado en los proyectos de Estadística Experimental del INE, dentro del marco establecido por el Sistema Estadístico Europeo. Concretamente estos 3:

- Indicadores demográficos
- Indicadores de renta media y mediana
- Distribución por fuente de ingresos

La última serie de datos disponibles son de los años 2015, 2016 y 2017.

Los ficheros vienen en formato `xlsx` por lo cual se hizo la parte de preparación en Excel (siguiendo las instrucciones dadas en el Datalab de Ciencias Sociales). A modo de resumen, han sido los siguientes pasos de curación para todos los ficheros:

- Eliminar las líneas extras de cabecera
- Renombrar las variables
- Preparar un diccionario de campos
- Crear dos columnas distintas para ID y nombre de los municipios, sacando éstos de la columna ETIQUETA
- Sustituir los valores "." con vacío

Después se hace el contaje de los nulos. Por ejemplo, para las variables demográficas:

IDMUN	0
NOMBREMUN	0
ETIQUETA	0
EDADMED17	1346
EDADMED16	1314
EDADMED15	1281
PRMEN17	1346
PRMEN16	1314
PRMEN15	1281
PRMAY17	1346
PRMAY16	1314
PRMAY15	1281
TAMHOG17	1346
TAMHOG16	1314
TAMHOG15	1281
PRHOGUNI17	1346
PRHOGUNI16	1314
PRHOGUNI15	1281
POB17	1346
POB16	1314
POB15	1281

Tabla 4. El número de registros nulos por cada variable en el dataset de datos demográficos.

Se ve que hay muchos campos con valores nulos. Lo mismo ocurre para los demás datasets. Esto llega a unos 17% de todos los municipios. En algunos casos los nulos son dispersos y en otros casos se concentran en 4 provincias (01-Álava, 20-Gipuzkoa, 48-Bizkaia y 31-Navarra) porque, debido a la estructura tributaria de las provincias forales, algunos de sus datos no están disponibles en las tablas del INE.

Primero, se intentó acceder a estos datos a través de los institutos de estadística de las comunidades autónomas correspondientes (Nastat y Eustat). Al parecer, no tenían los datos para todas las variables que faltan con el mismo formato, metodología o periodo.

Se probó con renta media por persona y por hogar. Se observó una discrepancia entre estas fuentes con los datos del INE. Es decir, para los que no son nulos, los valores no son iguales para el mismo municipio en el mismo año. En otros casos que parecen similares, les faltan otras variables del *Atlas*. Se entiende que esto se debe a tener tablas y estructuras diferentes en las fuentes nacionales y regionales, pero esta alteración al final puede afectar a la fiabilidad del modelo y las conclusiones derivadas.

Teniendo en cuenta que encontrar las equivalencias exactas entre estas fuentes, si es que sea posible, puede derivarse a un terreno ajeno a la competencia de este trabajo, al final se decidió no mezclar las fuentes del INE con las fuentes regionales y se descartó el uso del *Atlas de distribución de renta*.

Además, se ha trabajado en enlazar EVR con otra serie de ficheros de datos censales, demográficos y económicos sacados de otras tablas del INE durante el Datalab de Ciencias Sociales. Cada uno ha sido el resultado de un propio proceso de descarga y curación. En principio se trabajó con 4 ficheros con datos a nivel municipal:

- Evolución de la población
- Estructura demográfica de la población y su dinámica natural
- Datos censales de edificios e inmuebles, hogares y núcleos familiares
- Datos censales de la actividad laboral de personas

Además, se creó un fichero de variables territoriales que no es del INE, sino el resultado de un proyecto cartográfico usando el software QGIS.

Para darles un uso libre en este trabajo se tuvo que adaptarlos. En concreto, se realizaron estos pasos de curación en Excel:

- Eliminar caracteres especiales (*%) que aparecían en algunas filas, posiblemente debido al proceso de anonimato.
- Rectificar algunos valores manualmente (por ejemplo, cuando el denominador de una fracción era cero).
- En el caso de las variables que conceptualmente formaban la partición de un conjunto, se sustituyeron los nulos con cero, porque tiene sentido (por ejemplo, ramas de actividad laboral que son concluyentes teniendo la opción NO APLICABLE)

- Borrar las filas de los municipios para los cuales no había ningún campo con valor relleno (posiblemente debido a los cambios en divisiones territoriales). Han sido escasos.
- Eliminar los registros "condominio de varios municipios" en el dataset de variables territoriales, porque son zonas no habitadas y, por tanto, sin correspondencia en las fuentes estadísticas, tan solo en las cartográficas.

Cada uno de los datasets tiene alrededor de 8100 filas y entre 10 y 40 variables. El número total de municipios españoles en 2019 ha sido 8131. Sin embargo, hay algunos municipios que su división territorial es reciente y resultado de su segregación de otros municipios, por lo cual lógicamente no tienen datos para los años pasados. Estos cambios de unidades administrativas de base responden al conocido problema en Ciencias Sociales para estudios evolutivos denominado Unidad Espacial Modificable (UEM) (De Cos, 2004). Por esta razón, de cada serie se usaron las variables disponibles para el último año, normalmente 2020.

Por ejemplo, para la tasa de crecimiento entre 2016 y 2020, se chequeó manualmente los 8 municipios con valores nulos. Efectivamente se tratan de municipios nuevos, inexistentes en el año 2016. Se les asignó un valor 0, teniendo en cuenta que el efecto de segregación será reflejado también en disminución poblacional de sus municipios padres.

Los nulos de las algunas variables de ramas de actividad laboral son errores derivados de división con denominador cero (÷TOTAL RAMAS = 0, porque solo tienen registros en NO APLICABLE). Entonces se sustituyeron por cero.

Al final se hizo el contaje de los nulos restantes en los 5 datasets. Del dataset de evolución de la población se escogieron las variables del año 2020 y del resto, todas las variables, salvo 5 variables del dataset de edificios e inmuebles que todavía tenían nulos. Esto equivale a 90 variables en total.

3.2. Preparación y delimitación de EVR

Como este estudio se centra en los patrones de movilidad residencial a nivel nacional y debido a que no están disponibles los datos agregados para países extranjeros, se creó un subconjunto de EVR de todos los cambios residenciales cuyos ambos municipios de origen y de destino son del territorio español. El número de registros baja de 2,868,942 a 1,649,351.

Entre las 16 variables de EVR, se incluyen sus variables significativas a nivel personal y las demás variables a nivel municipal tanto de EVR como de los otros 5 datasets nuevos.

Se descartaron algunas variables que no son relevantes:

- Mes de nacimiento
- Año de nacimiento (ya que existe la edad)
- Año de variación (es igual para todos los registros de un fichero anual)

En el caso de código de nacionalidad (CNAC), se convirtió en variable binaria (0 para española y 1 para extranjera):

0	1
1240566	408785

Tabla 5. Contaje de valores de la variable CNAC después de binarización

Se hizo algo similar sobre el lugar de nacimiento (0 en España y 1 en extranjero), pero añadiéndola como una variable nueva (BPROVNAC) porque luego se necesita PROVNAC para otro procedimiento.

0	1
1146640	502711

Tabla 6. Contaje de valores de la nueva variable BPROVNAC

Chequear los nulos restantes:

SEXO	0
PROVNAC	0
MUNINAC	212266
EDAD	0
CNAC	0
PROVALTA	0
MUNIALTA	397718
MESVAR	0
PROVBAJA	0
MUNIBAJA	383643
TAMUALTA	0
TAMUBAJA	0
TAMUNACI	502711
BPROVNAC	0

Tabla 7. El número de nulos restantes por cada variable en el dataset EVR_2019

El TAMUNACI es nulo para todos los nacidos en extranjero. En el modelo previsto de momento no se incluirá, pero, si en algún momento se hace otro subconjunto solo con los nacidos en España se puede incluir.

Hay que generar nuevos campos ID para municipios de ALTA y BAJA, concatenando sus correspondientes códigos de provincia y de municipio para llegar a tener el código municipal estándar de 5 dígitos. El problema es que para los municipios con menos de 10,000 habitantes el código de municipio es desconocido (MUNIALTA y MUNIBAJA. Unos 400,000 registros). En estos casos se generó un código ejemplar "999". Después se puede concatenar todos los códigos provinciales y municipales, creando así un estándar IDMUNIALTA e IDMUNIBAJA.

Según los enfoques del estudio, se definieron dos escalas de arraigo/distancia para medir la cercanía entre la provincia de nacimiento y la provincia de alta o baja:

ARRALTA:

1 -> PROVNAC = PROVALTA (nacido en España)

2 -> PROVNAC ≠ PROVALTA (nacido en España)

3 -> ninguno de los anteriores (nacido en extranjero)

ARRBAJA:

1 -> PROVNAC = PROVBAJA (nacido en España)

2 -> PROVNAC ≠ PROVBAJA (nacido en España)

3 -> ninguno de los anteriores (nacido en extranjero)

Posteriormente, a los 5 ficheros nuevos que se enlazarán con EVR hay que agregar una fila por cada provincia con el ID 999 que representa el municipio poco-poblado ejemplar. Asignar un valor apropiado de cada variable para este municipio imaginario no es trivial. Se optó por asignarle, por cada variable, *el promedio de valores de los municipios con menos de 10,000 habitantes en la misma provincia*. En el caso de métricas cardinales como la población, hay que convertir la media a un número entero.

Para estos *workarounds* se apartaron los municipios 51001 y 52001 (Ceuta y Melilla), porque solo tienen un municipio que a su vez tiene una población mayor a 10,000, entonces no necesitan municipio ejemplar.

Para este procedimiento apareció la necesidad de añadir dos columnas CPRO y POB2020 a todos los demás datasets que carecen de estas variables, para poder filtrar los municipios pequeños de cada provincia y hacer el subconjunto y calcular las métricas.

Resultó que el número total de municipios en estos datasets es ligeramente diferente (por las razones anteriormente explicadas). Entonces se tuvo que tomar como referencia el dataset de EVOLUTIVO y corregir los registros no-coincidentes (unos 30 casos). Para algunos se buscó la población directamente en las tablas de INE o los portales de la Comunidad Autónoma correspondiente. En otros casos que son municipios desaparecidos ya desde hace años debido a las nuevas divisiones y segregaciones territoriales, se borró el registro, ya que obviamente tampoco estará presente entre los registros de EVR de 2019. Después se puede aplicar el procedimiento anterior a otros datasets.

3.3. Enlazamiento de fuentes

Para usar el IDMUN como PRIMARY KEY, se debe tener en cuenta que en cada fila de EVR hay dos ID para municipios de alta y de baja. Entonces, se crearon dos copias de cada dataset

y se les dio los nombres adecuados a sus columnas (añadiendo la palabra ALTA o BAJA al nombre de cada variable en su correspondiente copia del dataset).

Al final, hay que hacer dos LEFT JOIN por cada dataset original, uno con su versión que tiene IDMUNIALTA y otro con la versión IDMUNIBAJA, tomando como PRIMARY KEY de cada registro su correspondiente IDMUNIALTA y IDMUNIBAJA de EVR. Ejemplo:

```
db_joined <- dplyr::left_join(db, evolutivo_alta, by = "IDMUNIALTA")
db_joined <- dplyr::left_join(db_joined, evolutivo_baja, by = "IDMUNIBAJA")
```

La versión de EVR editada tenía 18 variables. En esta fase se añadieron 72 variables provenientes de los datasets nuevos:

- 5 de evolución de la población
- 35 de estructura demográfica
- 8 de edificios e inmuebles
- 17 de actividad laboral
- 7 de territoriales

Cada variable está duplicada en cada fila con valores para el municipio de ALTA y de BAJA.

Al final, el dataset resultante tiene una dimensión de 1,649,351 filas por 162 columnas y pesa más de 2 GB.

4. Aplicación de aprendizaje automático

Al dataset agregado que se creó en la fase anterior se aplicarán técnicas de Machine Learning que permitan entender el problema, es decir cuáles son los factores que afectan a los flujos migratorios interiores. Para demostrar la viabilidad de estas técnicas, se concentrará en los movimientos hacia capitales de provincia. Se puede formularlo en la siguiente pregunta:

> Si alguien se muda a una capital de provincia[1], ¿cuáles son los factores más influyentes/predictivos que se puede deducir tanto de sus características personales como de las características municipales de su origen?

4.1. Introducción sobre los árboles de decisión

Para el lector interesado, se repasa brevemente algunos aspectos generales de estas técnicas.

Los modelos de aprendizaje automático (Machine Learning) son unas técnicas de computación, subconjunto de inteligencia artificial, que se dividen en dos ramas mayoritarias: aprendizaje supervisado y aprendizaje no-supervisado (aunque existen otros tipos híbridos como aprendizaje por refuerzo, semisupervisado, etc.)

El aprendizaje supervisado se aplica cuando hay una serie de variables predictoras para predecir una variable objetivo. En otras palabras, existen unos ejemplos a base de los cuales se entrena un modelo para predecir un resultado.

El aprendizaje no-supervisado recibe de entrada un conjunto de observaciones e intenta detectar patrones o clústeres entre ellos según las características que tengan los datos.

En este estudio se intenta crear un modelo predictivo, para después estudiar la función predictora y ver qué variables y cómo juegan un papel importante en determinar el destino de un cambio residencial según si es hacia una capital o no. Entonces, se trata de aprendizaje supervisado.

Entre la panoplia de métodos de aprendizaje supervisado, se optó por usar los CART (Classification And Regression Trees) de la familia de árboles de decisión, sobre todo por la facilidad de interpretación y presentar la jerarquía y los umbrales de cada variable clave.

Un árbol de decisión es un diagrama para tomar decisiones a base de los valores de variables predictoras y llegar a un valor probabilístico para la variable objetivo. Está compuesto por:

- Nodos: cada nodo es una decisión a base del valor de una única variable predictora.
- Raíz (nodo principal): es el primer nodo de arriba.
- Rama: cada rama corresponde a un valor de la variable de su nodo padre.
- Hoja (nodo terminal): un conjunto de observaciones que su mayoría determina un valor probabilístico asignado a la variable objetivo.

[1] En principio se creó la binarización según capitalidad del municipio de alta, pero, como se explicará más adelante, se probó también otras categorizaciones a base de umbrales poblacionales.

- Etiquetas: indican el número o porcentaje de observaciones que caen en cada nodo.

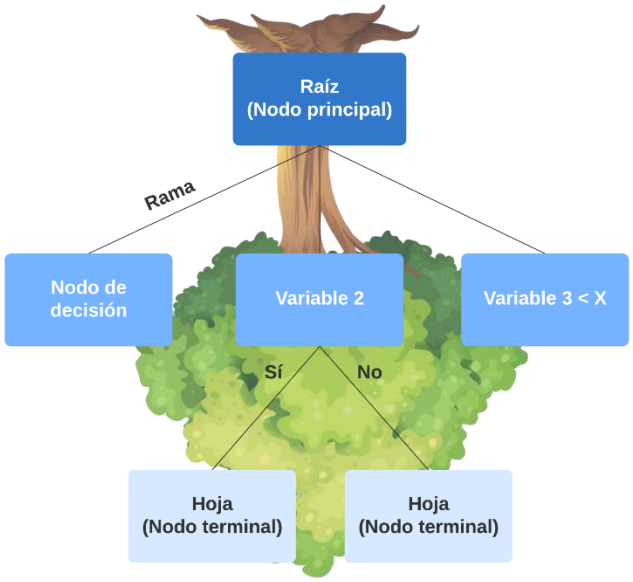

Figura 28. Esquema de un árbol de decisión

Hay dos tipos de árboles de decisión: árboles de regresión (donde la variable objetivo es continua) y árboles de clasificación (variable objetivo discreta). En ambos casos, las predictoras pueden ser continuas o categóricas.

Las ventajas de árboles de clasificación para este problema son:

- Manejan bien la mezcla de variables categóricas y continuas sin necesitar mucho preprocesamiento.
- La estructura del árbol expresa explícitamente la jerarquía de las variables influyentes y los intervalos importantes.
- Son fácilmente interpretables/legibles, incluso para expertos de otras áreas que no tengan conocimientos de Machine Learning. De hecho, se puede usar el resultado como una función o instrucción sin necesitar una implementación programática. Estas características no existen en otros métodos como regresión, redes neuronales o máquinas de vectores de soporte.

Para crear un árbol se inicia con calcular el *poder separativo* de cada variable predictora sobre la variable predictando. Existen diferentes métricas como error de clasificación, Information Gain, Índice de Gini, Cross-entropía, etc. Los paquetes que se implementará en este estudio usan el índice de Gini que es una medida para calcular la impureza de cada nodo terminal.

$$Gini = 1 - \sum_{i=1}^{n} p_i^2$$

Donde p_i es la probabilidad de caer x número de observaciones de cada clase de la variable objetivo en un nodo terminal.

Se escoge una variable predictora y se coloca en la raíz. Las ramas que salen desde la raíz consisten en valores correspondientes a esa variable. Por ejemplo, si es binaria, una rama para SÍ y otra para NO, y si es continua, dos intervalos <= y >. Se crean las hojas resultantes y se calcula el Gini para cada hoja. Se hace una suma ponderada del Gini de todas las hojas (ponderada por el número total de observaciones que caen en cada hoja). Esta suma será el índice de impureza de esa variante del árbol con la predictora escogida como la raíz. Haciendo lo mismo para todas las predictoras, al final se elige la que tiene el índice de impureza menor y se fija como la raíz. Crear las siguientes divisiones es una iteración del bucle anterior.

Este procedimiento continúa hasta que en una hoja no haya observaciones que se puedan dividir en dos grupos (el árbol completo) o que se alcance un determinado mínimo de muestras en cada hoja que indica la profundidad óptima del árbol.

La salvedad es que crecer el árbol normalmente resulta en sobreajuste, es decir el modelo funciona muy bien cuando se aplica a la muestra aprendida, pero tendrá mucho error al aplicarse a una muestra nueva. La solución es podar el árbol completo. Para esto se puede tunear diferentes hiperparámetros, por ejemplo, establecer una profundidad máxima o un mínimo de observaciones en cada hoja. Una vía habitual es aplicar un coste para la complejidad del árbol:

$$coste\ de\ complejidad = R(t) + \alpha\ .T$$

Donde R(t) es el error de clasificación, α es el parámetro de penalización y T es el número total de nodos terminales. Por cada valor de α una variante del árbol con un determinado número de hojas saldrá como el mejor resultado. Normalmente se hace una validación cruzada para ver cómo cambia el error general del modelo según se altera el valor de α y al final se elige la α que minimice el error.

Hay dos aproximaciones complementarias al usar los árboles:

En el método Random Forest, en vez de usar un árbol único como el modelo, se crea un conjunto de árboles. Para crear cada árbol se escoge una muestra aleatoria con un número determinado de variables predictoras. La predicción de variable objetivo será la clase que indiquen la mayoría de los árboles del bosque para cada observación. La idea es disminuir la variabilidad del modelo individual.

En el método Boosting, se sacan los resultados finales de un árbol individual y sobre esta muestra se entrena un nuevo árbol y así sucesivamente durante un determinado número de iteraciones. La idea es bajar el sesgo y mejorar la predicción del modelo total mediante concatenación de unos modelos menos robustos.

Para detallar más sobre la metodología de árboles se puede recurrir a una referencia conocida (también usada para esta síntesis) de aprendizaje estadístico como James et al. (2013).

4.2. Preparación del dataset para aplicar el modelo

Primero, hay que preparar un dataset de entrada para el modelo. Del dataset agregado, se extraen como predictoras todas las variables personales y las variables del municipio de BAJA. Del municipio de ALTA simplemente se extrae TAMUALTA, que se utilizará para definir el objetivo.

Entre las predictoras no se pueden incluir PROVNAC, PROVBAJA y IDMUNIBAJA porque son variables categóricas con muchos niveles (cada código de provincia se considera una categoría) y el límite de predictoras categóricas en R es de 32 categorías.

De esta manera, se quedan 1,649,351 observaciones de 79 predictoras y 1 variable objetivo. Como la muestra es innecesariamente grande, se trunca el número de filas a 10,000 con selección aleatoria. Luego se hace una división de train (80%) y test (20%).

4.3. Análisis preliminar de variables

Antes de crear los árboles, se realizan unos análisis estadísticos para conocer mejor las variables y sus relaciones (para aplicar algunas fórmulas hay que separar las variables categóricas de numéricas). El mapa de correlación entre las predictoras numéricas se ve en la figura 29:

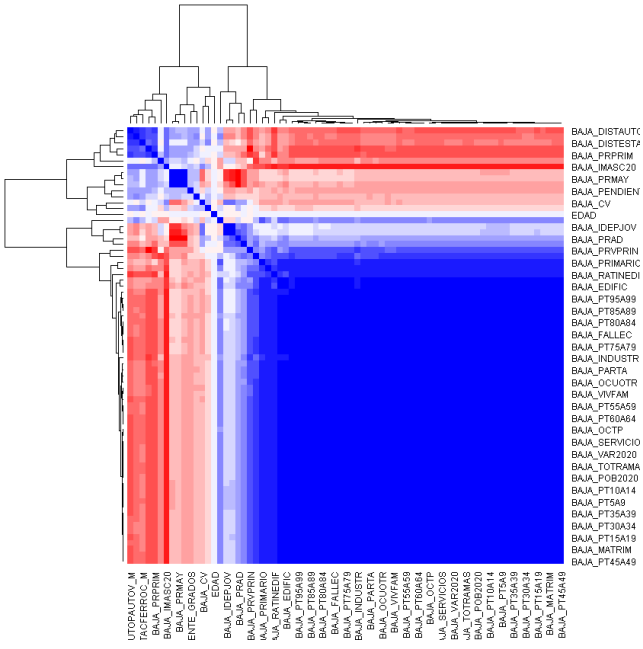

Figura 29. Correlación entre predictoras numéricas (azul = +1 y rojo = -1)

A simple vista, se entiende que hay mucha colinealidad entre varias predictoras. Esto es entendible, teniendo en cuenta que, por ejemplo, varias de estas variables se tratan del número de habitantes en cada intervalo de edad, entonces están altamente correlacionadas.

Si la idea fuera aplicar otro método de Machine Learning, se tendría que bajar la dimensionalidad del dataset aplicando métodos como regularización o Análisis de Componentes Principales (PCA). De hecho, el resultado de PCA se presenta en las figuras 30 y 31:

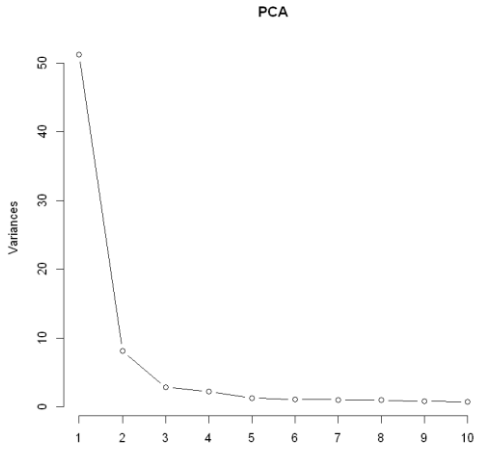

Figura 30. Varianza explicada por cada componente principal

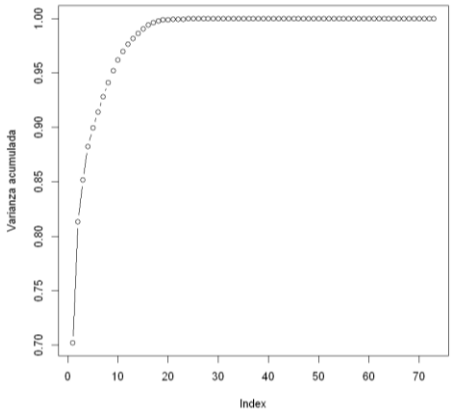

Figura 31. Varianza acumulada por componentes principales

Donde pocos conjuntos de variables explican la mayor parte de la varianza. Pero, en este estudio, no resultará aprovechable aplicar estos métodos, porque:

- Se pierde el significado interpretable de las variables originales.
- No se puede aplicar a predictoras categóricas (salvo usar otros métodos como Multiple Factor Analysis).
- Tal como funcionan los algoritmos de árboles, no es un preprocesamiento necesario porque, aunque existan muchas variables correlacionadas, el árbol escogerá la que tiene un índice de Gini ligeramente mejor y genera menos impureza en los resultados (mayor poder separativo). Si resulta que dos variables tienen un índice exactamente igual, cogerá la primera según entra en el algoritmo.

También se puede comprobar la correlación entre las predictoras y la variable objetivo:

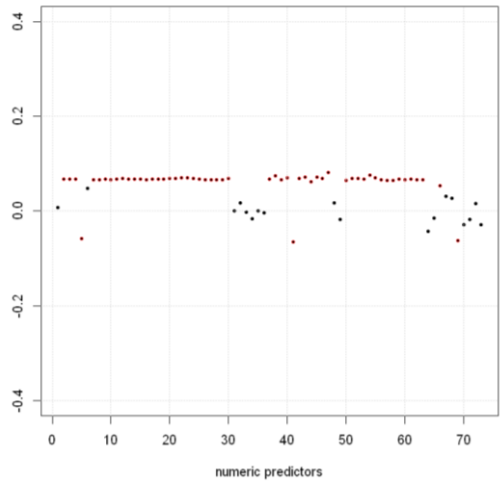

Figura 32. Correlación entre ALTA_CAPITAL y las predictoras numéricas (rojo = abs(corr) > 0.05). La línea horizontal que se ha formado arriba son las mismas variables que forman un bloque azul en la figura 29.

Se observa que el grado de correlación es muy débil. Entonces, no se encontrará una sola variable con alto poder predictivo. Puede que la acumulación de unas variables aporte más información. Usando la técnica de árboles de clasificación se probará si se puede obtener mejores resultados para detectar las predictoras significativas.

4.4. Probar diferentes modelos

Se abordará el problema con distintos enfoques que se detallan a continuación. Se intenta que los árboles clasifiquen entre cambios residenciales a diferentes tipos de destino. Si se consigue una buena clasificación, las ramificaciones indicarán las variables más relevantes, así como sus interrelaciones e intervalos.

4.4.1. Modelo con ALTA_CAPITAL binaria (CAP vs. NoCAP)

En esta primera prueba se intenta distinguir entre migraciones con destino en capitales de provincia y migraciones que como destino tienen cualquier otro municipio que no es capital. Para reflejar la situación de capitalidad, la variable objetivo se convierte en binaria:

- CAP si el municipio de alta es capital de provincia (TAMUALTA = 6).
- NoCAP si el municipio de alta no es capital de provincia (TAMUALTA ≠ 6).

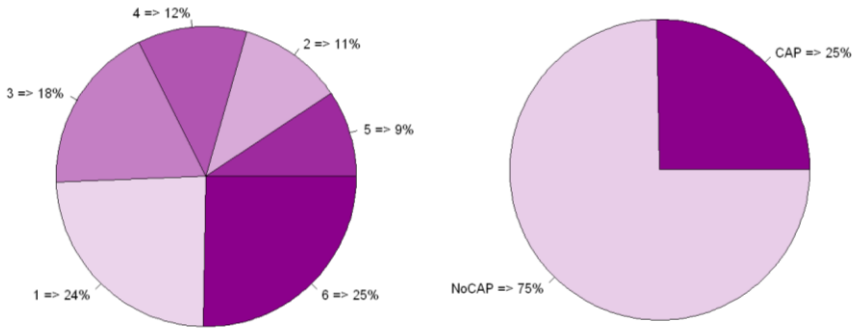

Figura 33. Proporción de categorías poblacionales de municipio de alta en la muestra, antes y después de binarización.

(Hay que mencionar que, aunque las categorías municipales siguen la estructura mencionada en la tabla 2, cuando un municipio es capital de provincia, se le asigna la categoría 6, independientemente de su población.)

No se incluyen las variables del municipio de ALTA entre las predictoras. La razón es que, si a la vez se introducen los datos de ambos municipios de ALTA y BAJA, como la variable objetivo es del municipio de ALTA, puede que el modelo solo esté buscando relaciones entre las variables X de ALTA y relacionarlo con la Y de ALTA, absteniendo así las de BAJA. Por ejemplo, el modelo cogerá la población del municipio de ALTA para determinar si es capital o no, una cifra que no tiene ninguna relación con la variación residencial en cuestión. De hecho, se hizo una prueba de este modelo que se verá más adelante.

Como un procedimiento general, en cada caso se empieza con crear un árbol completo como se ve en la figura 34. Se han probado ambos paquetes `tree` y `rpart` y el segundo normalmente muestra más estabilidad y mejor funcionamiento.

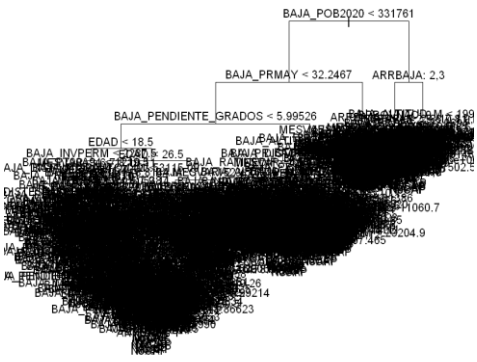

Figura 34. Esquema de un árbol completo (paquete `tree`).
La alta complejidad y densidad del árbol no resulta útil y legible, aunque sí se ven las variables más discriminantes en la raíz y las primeras ramificaciones desde arriba.

Luego se intentó podar el árbol mediante una validación cruzada (5 folds) aplicando como la función de coste el error de clasificación. Con la curva resultante de `tree` no se consiguió una estabilización. Se probó lo mismo con `rpart` (la función `printcp` y `plotcp`) y tampoco se consiguió una curva significativa para tunear el parámetro de penalización de complejidad (cp). Es decir, la división se queda en una ruta sola (*singlenode tree*). Cuando se fuerza un *cp* concreto, el árbol crece hasta cierto nivel, pero la división final de las observaciones que caen en cada nodo terminal es muy desproporcionada:

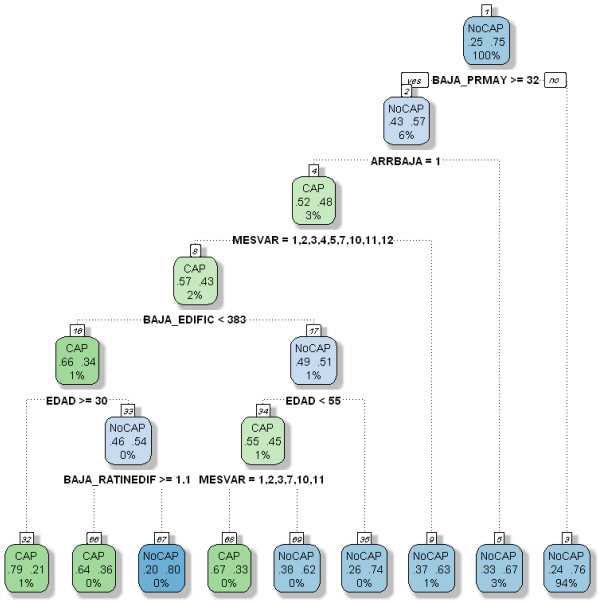

Figura 35. Un árbol con división desproporcional.
La mayoría de observaciones caen en una sola hoja y bajo la misma categoría (NoCAP).

4.4.2. Modelo con ALTA_CAPITAL multiclase (6 categorías)

Barajando la hipótesis de que el problema puede provenir de la proporción desigual de clases de variable objetivo, se preparó otra versión del dataset para probar un modelo multiclase. Es decir, la variable objetivo en vez de 2 clases tenga 6 (las mismas categorías poblacionales municipales de la figura 33 izquierda) y entonces las distintas clases tienen tamaños menos desequilibrados.

En este caso, la curva de *pruning* muestra una tendencia un poco más estable y el árbol resultante se ve a continuación:

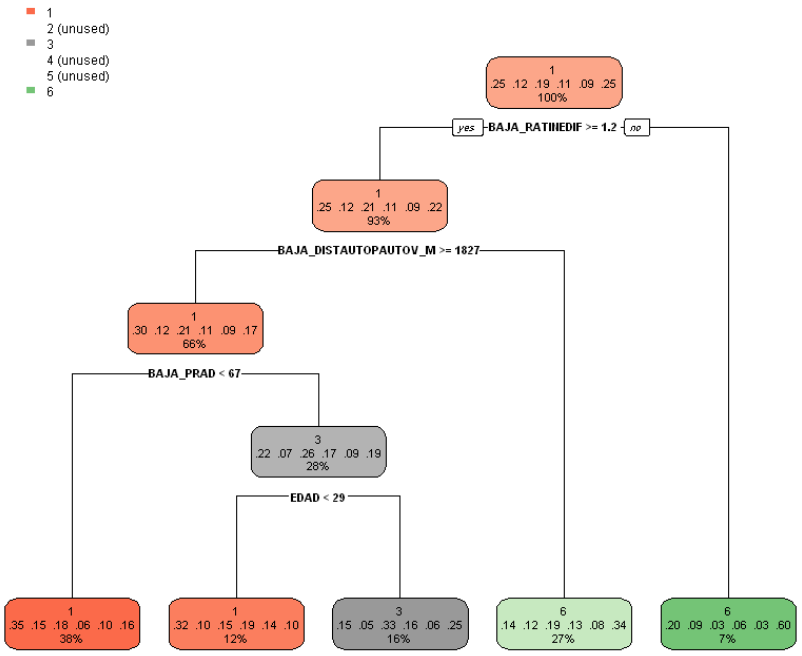

Figura 36. Árbol con variable objetivo multiclase

Observando la figura 36, tres de las clases (2, 4 y 5) tienen subpresentación y la proporción de otras clases en los nodos terminales tiene mucha impureza. La precisión del árbol podado ronda unos 40% que aun así es mayor a la media de proporción de clases.

Se debe mencionar que, la precisión de todos los modelos se tiene que comparar, en cada caso, con el tamaño de clases en la muestra correspondiente. Por ejemplo, si la variable objetivo de un modelo tiene dos clases con 50% de representación cada una, un modelo con 60% de accuracy está mejorando la precisión en unos 10% respecto a una predicción totalmente aleatoria. En el caso de muestras desbalanceadas existen también otras métricas como la precisión balanceada, etc. (Hall, 2020).

Visto estos resultados y para entender por qué el árbol no es capaz de clasificar bien entre las clases, primero se buscó la distinción entre las distribuciones de variables predictoras en subconjuntos CAP y NoCAP.

Se arrancó desde una comprobación visual mediante histogramas de todas las variables numéricas. La diferencia no resultó detectable a ojo. Después, se realizó un test de Kolmogorov-Smirnov sobre semejanza de distribuciones. Filtrando los resultados con un p-valor menor a 0.05 se destacan estas variables:

	names	pvalues	distance
5	BAJA_IMASC20 & BAJA_IMASC20	0.0366310527	0.20
14	BAJA_PT35A39 & BAJA_PT35A39	0.0366310527	0.20
15	BAJA_PT40A44 & BAJA_PT40A44	0.0366310527	0.20
20	BAJA_PT65A69 & BAJA_PT65A69	0.0243103130	0.21
21	BAJA_PT70A74 & BAJA_PT70A74	0.0004452597	0.29
38	BAJA_MUFET & BAJA_MUFET	0.0013646561	0.27
49	BAJA_PRVNPRIN & BAJA_PRVNPRIN	0.0243103130	0.21
50	BAJA_OCTC & BAJA_OCTC	0.0158141003	0.22
56	BAJA_ESTUD & BAJA_ESTUD	0.0158141003	0.22
69	BAJA_ALTITUD_M & BAJA_ALTITUD_M	0.0023184583	0.26

Tabla 8. Variables con una distribución significativamente diferente entre subconjuntos de CAP y NoCAP según el resultado de test de Kolmogorov-Smirnov (p-valor < 0.05)

Como se observa, la métrica de *distancia* es baja (el rango original oscila entre 0 y 1, con 0 representado el mayor grado de similitud y 1 el mayor grado de diferencia). Esto quiere decir que no se observan diferencias muy significativas entre la distribución de las predictoras en estas dos muestras y por lo tanto es esperable que los árboles no consigan demasiada discriminación entre estos subconjuntos.

4.4.3. Modelo con ALTA_CAPITAL binaria balanceada

La derivada aproximación consiste en balancear la muestra de entrenamiento. El efecto negativo de muestras desequilibradas es un tema conocido y estudiado en Machine Learning. Parece que los árboles en concreto muestran un funcionamiento sensible y débil ante desequilibrio. Para compensar este efecto existen varias estrategias como sobremuestreo, submuestreo, error de clasificación ponderado, etc. (Jordan, 2018). De todos modos, al compensar el desequilibrio existente entre diferentes clases, se fuerza al modelo para que aprenda mejor las características específicas y distintivas de cada clase y evitar que la predicción o categorización se realice meramente a base de la proporción desigual de clases (cosa que ocurrió en los primeros intentos cuando el árbol se quedó en un *single node*).

Se creó una muestra balanceada cogiendo de la clase mayoritaria el mismo número de observaciones que tiene la clase minoritaria (submuestreo) y luego barajando el conjunto.

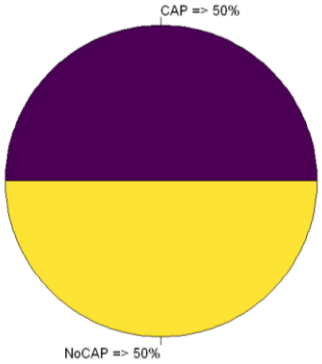

Figura 37. Proporción de clases en la muestra binaria balanceada

Se entrenó un árbol completo sobre esta muestra y luego se podó con el *cp* que minimiza el error. El árbol podado tiene un accuracy de test de 59%. En este paso también se probó crear un árbol más simple solo a base de 3 predictoras que resultaban importantes en el árbol completo (BAJA_POB2020, MESVAR y EDAD), pero la precisión resultante fue más baja (53%).

4.4.4. Modelo multiclase triple

Dado los resultados anteriores, se desarrolló la idea de que quizás la distinción capital vs. no-capital no resulta la más discriminativa, porque habrá bastante similitud, por ejemplo, entre municipios no capitales con más de 100,000 habitantes (categoría 5) y municipios capitales de provincia (categoría 6) con una población ligeramente más grande. Entonces, se planteó la siguiente categorización:

- Los municipios de categoría 1 y 2: Pequeño
- Los municipios de categoría 3 y 4: Mediano
- Los municipios de categoría 5 y 6: Grande

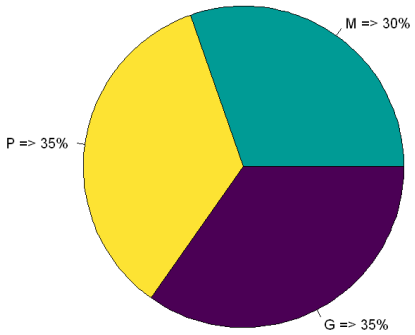

Figura 38. Proporción de clases en la muestra con categoría triple.

Como las proporciones resultan bastante similares, no se balancea nuevamente. Se entrena un árbol completo y se poda basándose en su curva de *cp*:

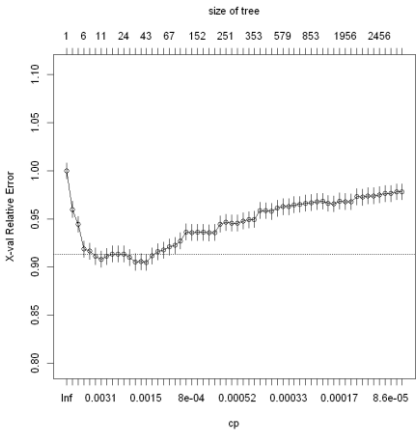

Figura 39. Curva de pruning.
Relación entre el factor de penalización de complejidad, el tamaño del árbol y el error resultante.

El árbol podado tiene casi 42% de precisión en test. Como el número de nodos terminales fue muy grande, se entrenó un árbol más pequeño a base de las 4 predictoras más importantes (BAJA_PRAD, BAJA_TDEP, BAJA_IDEPMAY y BAJA_PRMAY). El modelo escogió 3 de ellas, resultando en 40% de precisión con la siguiente estructura:

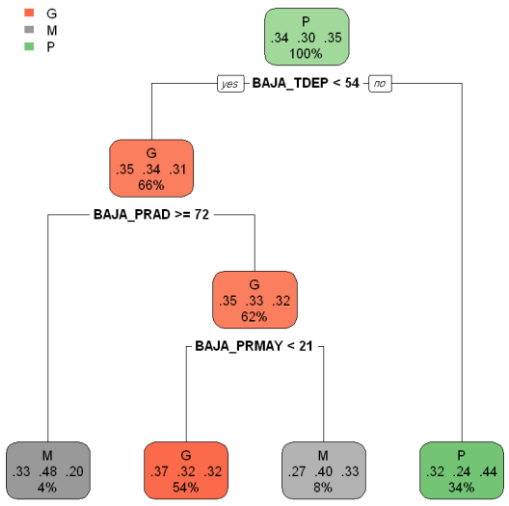

Figura 40. Árbol multiclase triple con las 3 predictoras más importantes (tasa de dependencia, porcentaje de población adulto y porcentaje de población mayor en el municipio de origen).

Del árbol de la figura 40 se deduce que, por ejemplo, si la tasa de dependencia (población de menores de 15 y mayores de 65 años entre la población general) es igual o mayor a 54% en un municipio (perfil poblacional con muchos ancianos o menores), las personas que se mudan desde allí en 44% de los casos se mudan a un municipio pequeño, en 32% a un municipio grande y en 24% a un municipio mediano (según intervalos poblacionales anteriormente mencionados). Otro ejemplo es cuando la tasa de dependencia es menor a 54% y el porcentaje de adultos mayor a 72% (alta presentación de ancianos) en cuyo caso casi la mitad de los migrantes se van hacia municipios de tamaño poblacional mediano.

4.4.5. Penalización de misclasificación en modelo multiclase triple

Como se comentó anteriormente, además del balanceo de muestra (*adjusted sampling*), otra aproximación para bregar con el desequilibrio es asignar pesos diferentes a la misclasificación de predicciones. Usando el paquete `C50` se creó una matriz de confusión explícita donde la diagonal es cero y otras entradas son pesos relativos para misclasificación de cada clase en concreto. Aún con este método, no se consiguió una precisión de test más alta de 40%.

4.4.6. Otros modelos binarios (Grande vs. Pequeño)

Al hilo de la sección 4.4.4. se intentó enfrentar los municipios grandes con los pequeños según otras divisiones. Primero se categorizó los municipios de clase 5 y 6 como "grande" y las demás 4 clases como "pequeño". Luego se probó con 4, 5 y 6 como grande y el resto como pequeño.

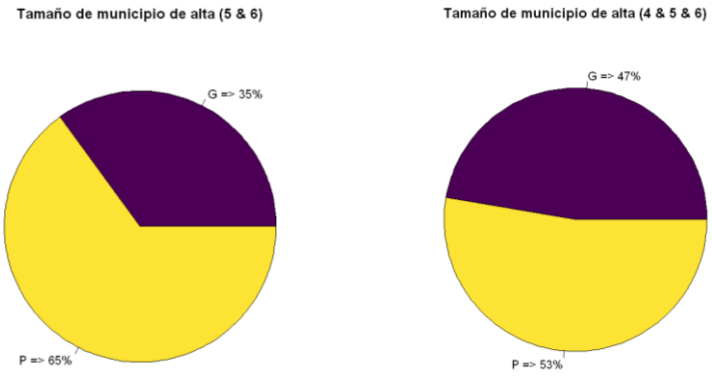

Figura 41. Proporción de clases en dos modelos multiclases binarios (Grande vs. Pequeño)

En el caso del modelo con categorías 5 y 6 como grande, el accuracy de test no superó los 65%. Pero el otro modelo (4, 5, 6 como grande) alcanzó 59% de accuracy con la muestra balanceada y el árbol podado. Una variante de este árbol es la siguiente:

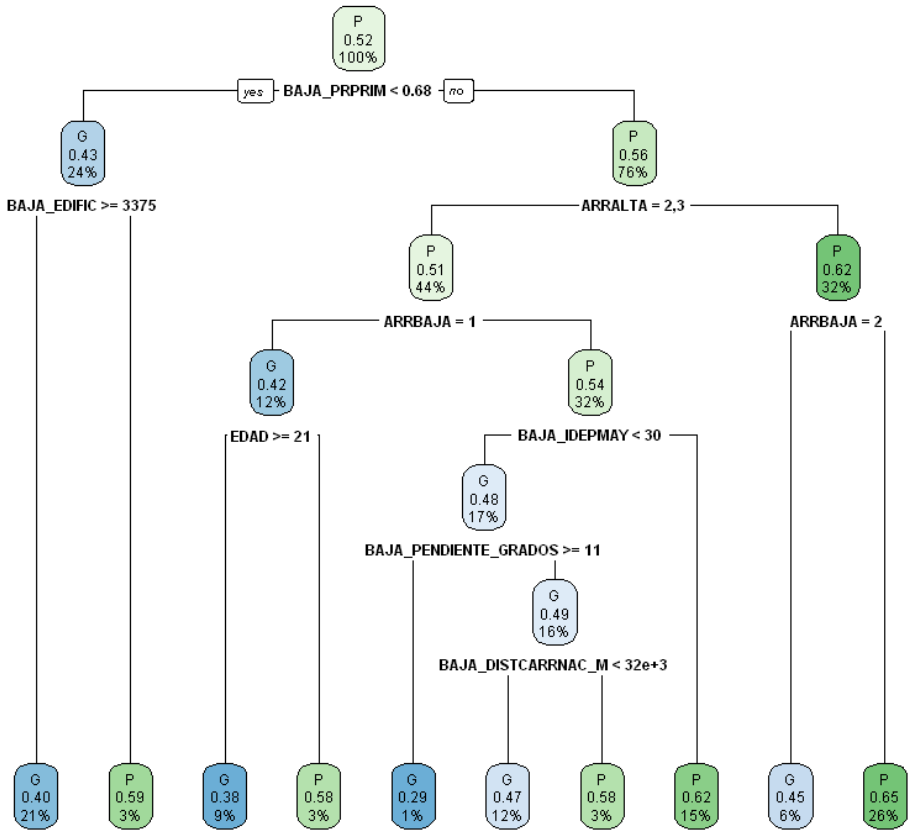

Figura 42. Árbol podado con muestra balanceada y categorías 4,5,6 como Grande

Visto los modelos y variantes anteriores, el árbol de la figura 42 parece un ejemplo informativo, aun padeciendo una predicción débil. Lo interesante sobre este árbol es el hecho de que, entre

sus 8 variables importantes, 1 pertenece a las variables originales de EVR (EDAD), 5 pertenecen a las variables censales de municipio de BAJA añadidas durante la fase de agregación y 2 son las variables definidas en este estudio (ARRALTA y ARRBAJA).

Esto indica que este tipo de migraciones tiene cierta relación con factores como el porcentaje de trabajadores del sector primario y el número de edificios del municipio de origen (que a su vez está relacionado con la población) y la cercanía/lejanía entre el lugar de nacimiento y el origen o el destino. Por ejemplo, solo viendo las 3 ramificaciones desde arriba, se entiende que si el porcentaje de trabajadores del sector primario es mayor a 68% en el municipio de origen (la mayoría se dedican a agricultura y ganadería) y el migrante es un ciudadano español que está moviéndose dentro de la misma provincia de su nacimiento, en 68% de los casos su nuevo destino también es un municipio pequeño.

También se presentó una variante ligera de este árbol mediante valores predeterminados (sin crecer el completo y luego hacer la poda) básicamente para poder explicar su funcionalidad:

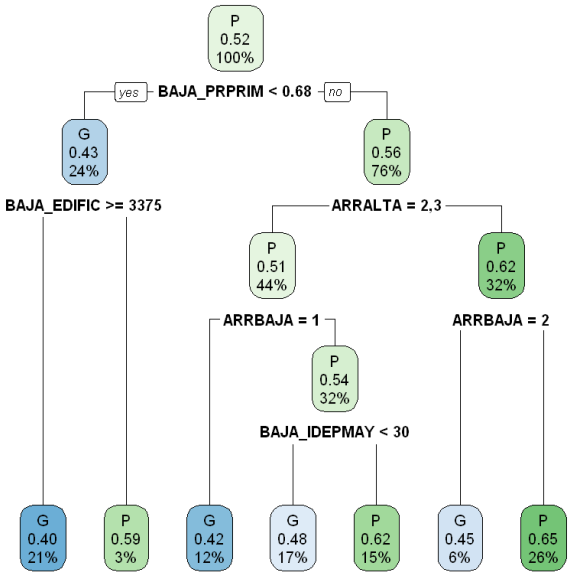

Figura 43. Árbol ligero con muestra balanceada y categorías 4,5,6 como Grande

Comparando con el árbol anterior (figura 42), en el árbol ligero (figura 43) los nodos terminales siguen teniendo impureza, pero la estructura ha ahorrado 3 variables. Entre las variables que se han eliminado, PENDIENTE_GRADOS tampoco generaba una relación explícitamente entendible y en una de sus ramas caían apenas 1% de las observaciones. La desaparición de la variable EDAD también puede ser útil en algún sentido porque, tal como se comentó en la fase de análisis descriptivo, si se supone que entre la muestra hay miembros de familias con hijos, esta distinción entre edades al final no es real, porque esas variaciones son mutuas. Resumiendo, el modelo puede ganar más simplicidad sin perder mucha precisión.

Ante la baja precisión en general, también se debe mencionar que, a la hora de hacer interpretaciones, se puede enfocar más en algunas de las ramas que acaban en una hoja con más pureza. Por ejemplo, en la figura 43, la rama que termina en la primera hoja de la derecha

muestra una relación más probable (65%) que otras ramas, luego la tercera rama de la derecha (62%), etc.

El problema es que sigue existiendo mucha variabilidad en las estructuras de los árboles resultantes según la aleatoriedad de muestra de entrenamiento. Por ejemplo, cogiendo una muestra nueva, la estructura puede resultar bastante diferente a las figuras 42 o 43:

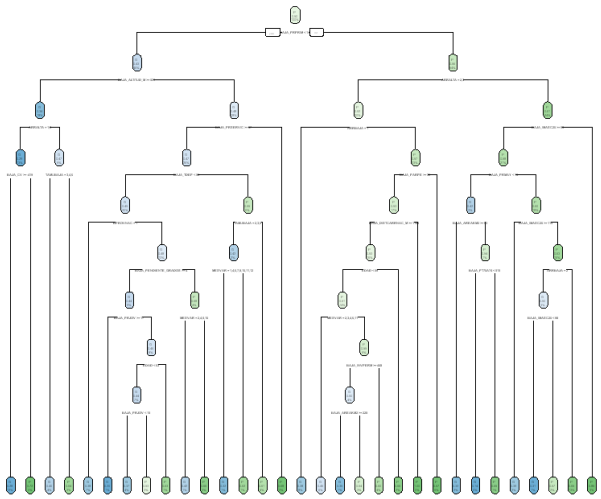

Figura 44. Otra variante del árbol podado con muestra de categorías 4,5,6 como Grande

4.4.7. Bagging

Con la intención de bajar la variabilidad del modelo y conocer las variables verdaderamente influyentes, se crea un `RandomForest`. Hay que realizar varios tuneos sucesivos sobre el número total de árboles del bosque (`ntree`), el número de variables escogidas en cada paso (`mtry`) y el número de observaciones (`sampsize`). Creando un bosque con 387 árboles, 4 predictores y 5141 observaciones, no se observó una mejoría considerable (la precisión sigue girando en torno a 60%). Viendo la importancia de los predictores sobre la precisión general del modelo, los resultados aquí se distinguen de los árboles individuales:

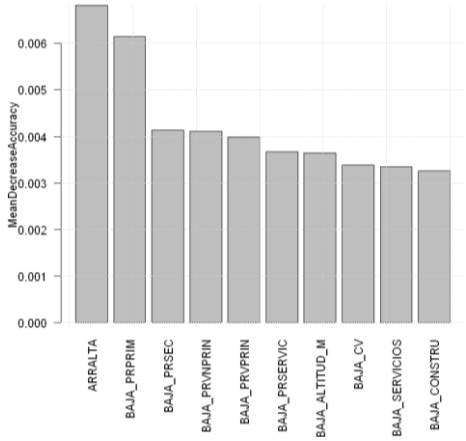

Figura 45. Importancia de las primeras 10 variables (Random Forest)

Donde aparentemente las variables más importantes son las pertenecientes a las ramas de actividad laboral, además de ARRALTA.

4.4.8. Boosting

Con el objetivo de bajar el sesgo de los modelos débiles (árbol individual), usando el paquete `adabag` se crea un conjunto boosted de clasificadores individuales. El número de iteraciones se reduce lentamente, entonces se establece 100 iteraciones.

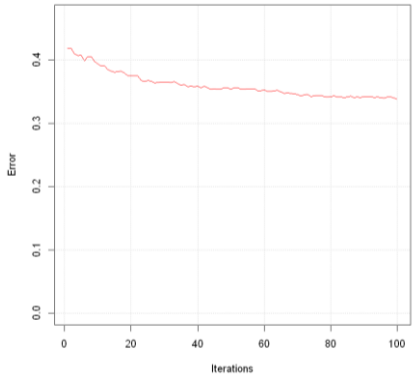

Figura 46. Error de test del modelo boosted en función del número de árboles individuales

La precisión sigue siendo la misma (59 - 60%). Se revisa la importancia relativa de los predictores en el conjunto:

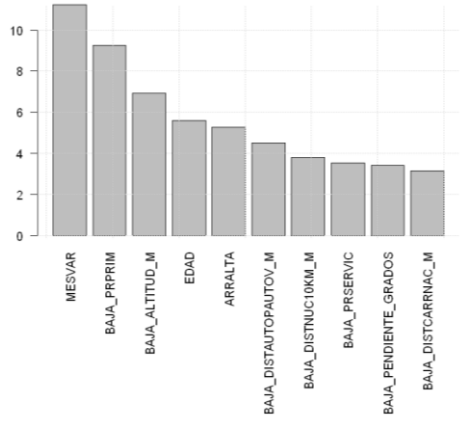

Figura 47. Importancia de las primeras 10 variables (modelo boosted)

El resultado es distinto del caso anterior, aun compartiendo algunas variables como BAJA_PRPRIM, ARRALTA, BAJA_PRSERVIC y BAJA_ALTITUD_M.

4.4.9. Modelo con variables de ALTA

Solo para demostrar el tema mencionado en la sección 4.2. sobre el sesgo derivado de introducir al mismo tiempo las variables del municipio de ALTA, se hizo la prueba. Una vez con la variable

objetivo binaria (CAP vs. NoCAP) y otra vez con la multiclase (P, M, G). Se presentan los árboles resultantes:

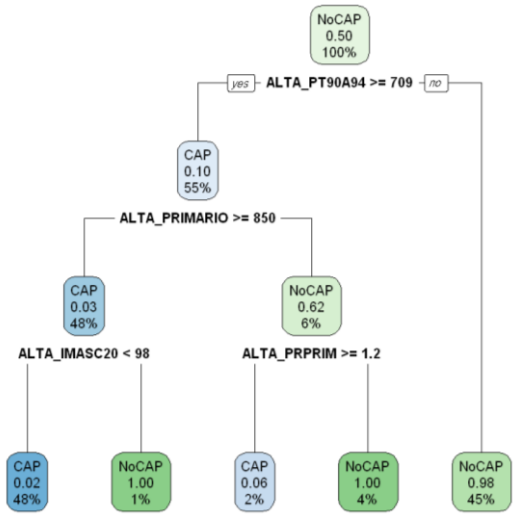

Figura 48. Modelo binario de la muestra con ambas variables de ALTA y BAJA

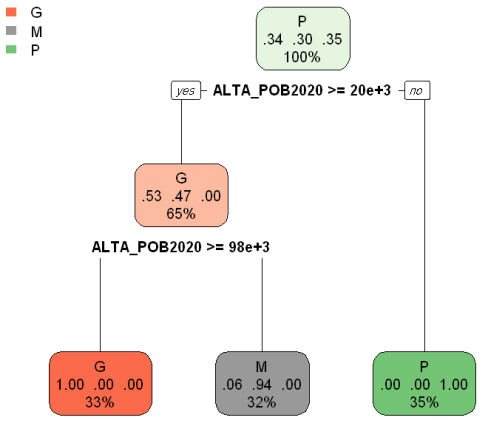

Figura 49. Modelo multiclase de la muestra con ambas variables de ALTA y BAJA

Ambos modelos tienen un accuracy de test de unos 97% y hojas con alta pureza, pero, como está claro en las figuras 48 y 49, todas las variables se han escogido de los datos demográficos del municipio de ALTA, entonces el modelo no está aprendiendo las características de las variaciones residenciales, sino está categorizando las capitales según su población.

4.5. Análisis de los resultados

Se ha comprobado que los árboles muestran mal funcionamiento ante muestras desbalanceadas. También se demostró que cuando las predictoras no tienen mucho poder separativo, los métodos de bagging y boosting tampoco pueden aumentar la precisión de manera considerable, aunque sí aportan alguna información sobre la importancia de variables que no esté alterada por la variabilidad de los árboles individuales.

Ante el problema de baja discriminación, se probó que los modelos funcionan mejor si se agrupan clases similares respecto a la variable objetivo.

A rasgos generales, hay que reconocer que no se puede llegar a una predicción robusta sobre los patrones de movilidad residencial hacia municipios grandes o pequeños usando las variables existentes. Pero, viendo algunas variables en común entre los resultados de los árboles individuales y modelos de bagging y boosting, se puede afirmar que los indicadores que posiblemente tienen cierta relación con la decisión de un ciudadano para mudarse a un municipio grande o pequeño (estableciendo 50,000 habitantes como el umbral) son:

- BAJA_PRPRIM (% de trabajadores del sector primario en el municipio de baja)
- ARRALTA (la cercanía entre la provincia de nacimiento y la provincia de destino)

Se puede afirmar que alto porcentaje de los trabajadores del sector primario (>= 70%) mayormente resulta en elegir municipios pequeños como nuevo destino (en 60% de los casos). La escala de arraigo también muestra una tendencia similar. Si un ciudadano decide mudarse dentro de la misma provincia de su nacimiento, con 62% de probabilidad está mudando a un municipio pequeño.

También el porcentaje de trabajadores en otras ramas de actividad laboral, sobre todo la de servicios, puede tener cierta importancia. Entre las variables territoriales, la altitud parece ser más destacada en este problema.

Básicamente para mostrar la interpretación de un árbol individual, si se elige el árbol de la figura 43 como sugerencia (por su sencillez estructural) y concentrando solo en la rama que acaba en su primera hoja de la derecha, a modo de *storytelling* se puede comentar lo siguiente:

> Entre la población española en el año 2019, si un ciudadano decide cambiar su lugar de residencia dentro del territorio nacional, sin que sepamos su motivación explícita, podemos fijarnos en la tasa de trabajadores del sector primario en su municipio de baja. Si esta tasa es mayor a 68% (es decir, la mayoría de sus habitantes se dedican a agricultura, ganadería, pesca, etc.) preguntamos lo siguiente: ¿se va a desplazar lejos o prefiere estar cerca de su pueblo de origen o donde nació? Si resulta que prefiere estar cerca, normalmente o se trata de un español que antes también vivía en la misma provincia o bien es un extranjero que no tiene esa vinculación. En 65% de los casos, estos perfiles se mudan a un municipio pequeño con menos de 50,000 habitantes.

Usando esta metodología se puede deducir todas las relacionas que expresan los modelos.

El valor añadido es que estos resultados se han obtenido gracias a aplicar una técnica de Machine Learning y no sería posible conseguirlos mediante análisis descriptivo o métodos convencionales de Ciencias Sociales.

5. Conclusiones y trabajo futuro

Este proyecto ha estudiado una fuente demográfica principal en España, el informe de Estadísticas de Variaciones Residenciales (EVR). Usando las herramientas de Ciencia de Datos se hizo un análisis descriptivo de todas las variables del dataset y se presentaron diferentes visualizaciones, aproximaciones y subconjuntos definibles.

Para alimentar el análisis se pasó a enlazar EVR con más de 70 variables demográficas, censales y territoriales provenientes de distintas fuentes oficiales. Se presentó un procedimiento reproducible para la agrupación de fuentes, unificación de identificadores y tratar los valores nulos, acompañado por su implementación programática, que además se puede adaptar para otros estudios sociales y geográficos.

Usando los algoritmos de Machine Learning se intentó resolver un problema en concreto; detectar los patrones de flujos migratorios entre municipios españoles. Se aprovechó de técnicas de aprendizaje supervisado, intentando examinar el abanico de métodos relacionados con Árboles de Clasificación para llegar a tener una valoración integral sobre su funcionalidad.

Se ha comprobado que las variables existentes no alcanzan una gran separación sobre la variable objetivo (categoría poblacional del municipio de destino), pero sí aportan información sobre ello aumentando en unos 10% la precisión del modelo respecto a una predicción aleatoria.

El balanceo de la muestra de entrenamiento y la división de categorías poblacionales resultaron dos factores determinantes, porque habrá similitud entre perfiles municipales de franjas poblacionales cercanas, aunque algunos sean capitales de provincia y otros no. El mejor resultado se consiguió categorizando los municipios con 50,000 habitantes o más como grande y los de menos de 50,000 habitantes como pequeño, balanceando luego el conjunto.

Viendo los resultados, las variables relacionadas con ramas de actividad laboral (sobre todo, el porcentaje de trabajadores del sector primario y el sector de servicios) resultaron más importantes. Además, la definición propia de la escala de arraigo entre el municipio de nacimiento y el municipio de destino resultó significativa en detectar los patrones de movilidad residencial. Partiendo de estos resultados se puede plantear unos enfoques y variables más concretas para llegar a detectar mejor los patrones.

En cuanto a los aspectos técnicos se puede concluir que:

- La diversidad estructural de fuentes y anomalías naturales en al ámbito de Ciencias Sociales requiere una combinación de métodos programáticos y manuales para trabajar con los datos. Las herramientas habituales de Ciencia de Datos tienen alta potencialidad en abarcar problemas de este tipo.

- Conocer los patrones de movilidad residencial (por lo menos, teniendo como variable objetivo la población del destino) parece ser un problema de alta complejidad estadística

que no se puede resolver con las variables principales demográficas y territoriales. Entonces, la selección de variables de interés y su preanálisis será una tarea esencial en semejantes trabajos.

- Los métodos de árboles en este caso muestran las mismas debilidades generales que se les asigna, es decir baja precisión y alta variabilidad, aunque aportan una interpretación esclarecedora sobre los patrones existentes detrás de los flujos migratorios. Esta interpretabilidad se puede considerar como un punto competitivo en comparación con otras técnicas de Machine Learning.

Este trabajo ha mostrado el interés y la potencialidad de la aplicación de Machine Learning para analizar los patrones complejos de flujos migratorios, pudiéndose desarrollar los futuros estudios en diferentes líneas:

- Usando el mismo dataset agregado y la misma metodología, probar otras variables objetivos para ver cuáles son los patrones más predecibles.

- Separar los núcleos urbanos más grandes del resto de municipios y probar si se llega a una separación más considerable.

- Establecer otra estrategia de selección de variables y crear un dataset nuevo para aplicar el mismo método y comparar los resultados.

- Probar otros métodos de Machine Learning como las redes bayesianas para detectar una relación sobre el mismo objetivo (capitalidad/tamaño del municipio de destino) con las predictoras existentes.

ANEXOS

1. Repositorio:

Los ficheros de los datasets usados en este proyecto y los códigos de R en formato Jupyter Notebook son accesibles en repositorios abiertos:

https://github.com/Analytics-Matin/EVR

https://zenodo.org/record/5105572

ANEXOS

2. Referencias bibliográficas:

De Cos Guerra, Olga. (2004). «Valoración del método de densidades focales (kernel) para la identificación de los patrones espaciales de crecimiento de la población en España». Geofocus, núm. 4, págs. 136-165. http://www.geofocus.org/index.php/geofocus/article/view/46/214

— (2007). «La dinámica metropolitana en España. Análisis estadístico y cartográfico de los municipios a partir de la población y la vivienda». Geographicalia, núm. 51, 2007, págs. 59-80.

Domingo, Andreu y Sabater, Albert (2013). «Crisis económica y emigración: la perspectiva demográfica». Anuario CIDOB de la Inmigración, 2013, págs. 59-88.

Eustat (Instituto Vasco de Estadística): https://www.eustat.eus/indice.html

García Palomares, Juan Carlos y Pozo Rivera, Enrique (2010). «Movimientos migratorios en la Comunidad de Madrid: Unos flujos más intensos y complejos (1991-2006)». Boletín de la Asociación de Geógrafos Españoles, núm. 53, págs. 89-119.

Hale, Jeff (2020). «The 3 Most Important Composite Classification Metrics». https://towardsdatascience.com/the-3-most-important-composite-classification-metrics-b1f2d886dc7b

James G., Witten D., Hastie T., Tibshirani R. (2013). «Tree-Based Methods. In: An Introduction to Statistical Learning». Springer Texts in Statistics, Vol 103. Springer, New York, NY. https://doi.org/10.1007/978-1-4614-7138-7_8 pp. 303-335.

Jordan, Jeremy (2018). «Learning from imbalanced data»: https://www.jeremyjordan.me/imbalanced-data/

Metodología EVR: https://www.ine.es/daco/daco42/migracion/notaevr.htm

Nastat (Instituto de Estadística de Navarra): https://www.navarra.es/home_es/Gobierno+de+Navarra/Organigrama/Los+departamentos/Economia+y+Hacienda/Organigrama/Estructura+Organica/Instituto+Estadistica/

Reques Velasco, Pedro y De Cos Guerra, Olga. (2013). «Los difusos límites del espacio urbano-metropolitano en España». Ciudad y territorio: Estudios territoriales, núm. 176, págs. 267-280.

Ródenas, Carmen y Martí, Mónica (2006). «Reinterpretando el crecimiento de la movilidad de España: La población extranjera y las migraciones repetidas». Cuadernos Aragoneses de Economía, 2ª época, núm. 16 (1), págs. 37-59.

Susino Arbucias, Joaquín (2011). «La evolución de las migraciones interiores en España: una evaluación de las fuentes demográficas disponibles». Papers: revista de sociología, Vol. 96, núm. 3, 2011, págs. 853-881.

— (2012). «Fuentes demográficas para el estudio de la migración en España». REMHU - Revista Interdisciplinar da Mobilidade Humana, Vol. 20, núm. 39, 2012, págs. 51-76.

ANEXOS

3. Diccionario de variables:

Los identificadores "ALTA_" y "BAJA_" al inicio de una variable hacen referencia al municipio de alta o de baja respectivamente.

Variable	Definición
ADULT	Población del grupo de edad adulto
ALTITUD_M	Altitud en m
ANONAC	Año de nacimiento
ANOVAR	Año de variación
AREAKM2	Área en Km2
ARRALTA	Escala de arraigo entre provincias de nacimiento y alta
ARRBAJA	Escala de arraigo entre provincias de nacimiento y baja
BPROVNAC	Binario de provincia de nacimiento (0 España, 1 Extranjero)
CNAC	Código de nacionalidad (codificación según diccionario EVR)
CONSTRU	Personas en el sector de construcción
CPRO	Código de provincia española
CV	Crecimiento vegetativo
DISTAUTOPAUTOV_M	Distancia mínima a autopistas o autovías en m
DISTCARRNAC_M	Distancia mínima a carreteras nacionales en m
DISTESTACFERROC_M	Distancia mínima a vías férreas en m
DISTNUC10KM_M	Distancia mínima a núcleos urbanos en m
EDAD	Edad en el momento de la variación
EDIFIC	Total de edificios
ESTUD	Estudiantes
FALLEC	Fallecidos por el lugar de residencia
IDEPJOV	Índice de dependencia de los jóvenes

IDEPMAY	Índice de dependencia de los mayores
IDMUN	ID de municipio español (formato estándar 5 dígitos)
IMASC2020	Índice de masculinidad en el año 2020
INDUSTR	Personas en otras industrias
INMUEB	Total de inmuebles
INVPERM	Personas con invalidez permanente
JOVEN	Población del grupo de edad joven
JUBPENS	Jubilados, prejubilados, pensionistas o rentistas
MATRIM	Matrimonios por el lugar en que han fijado residencia
MAYOR	Población del grupo de edad mayor
MESNAC	Mes de nacimiento
MESVAR	Mes de variación
MUFET	Muertes fetales tardías por residencia materna
MUJ2020	Población de mujeres en el año 2020
MUNIALTA	Municipio o país de alta (codificación según diccionario EVR)
MUNIBAJA	Municipio o país de baja (codificación según diccionario EVR)
MUNINAC	Municipio o país de nacimiento (codificación según diccionario EVR)
NACIM	Nacidos vivos por residencia materna
NMHIJNUC	Número medio de hijos
NOAPLIC	No aplicable
NOMBREMUN	Nombre de municipio español
NUCFAM	Número de núcleos familiares
OCTC	Ocupados a tiempo completo
OCTP	Ocupados a tiempo parcial
OCUOTR	Otra situación
PARPE	Parados buscando primer empleo
PARTA	Parados que han trabajado antes
PENDIENTE_GRADOS	Pendiente de cambio en grados
POB2020	Población general en el año 2020
PRAD	Porcentaje de población adulto

PRIMARIO	Personas en agricultura, ganadería y pesca
PRJOV	Porcentaje de población joven
PRMAY	Porcentaje de población mayor
PRMONOPAR	Porcentaje de núcleos monoparentales
PRNPRIN	Porcentaje de viviendas no principales
PRNUMEROS	Porcentaje de familias numerosas
PROVALTA	Provincia de alta (codificación según diccionario EVR)
PROVBAJA	Provincia de baja (codificación según diccionario EVR)
PROVNAC	Provincia de nacimiento (codificación según diccionario EVR)
PRPRIM	Porcentaje de trabajadores del sector primario
PRSEC	Porcentaje de trabajadores en construcción e industria
PRSERVIC	Porcentaje de trabajadores en servicios
PRVRPRIN	Porcentaje de viviendas principales
PTXAY	Población por grupos quinquenales de X a Y años
RATINEDIF	Ratio de inmuebles por edificio
SERVICIOS	Personas en el sector de servicios
SEXO	Sexo
TAMEDHOG	Tamaño medio del hogar
TAMUALTA	Tamaño de municipio de alta
TAMUBAJA	Tamaño de municipio de baja
TAMUNACI	Tamaño de municipio de nacimiento
TC16A20	Tasa de crecimiento medio interanual de 2016 a 2020
TDEP	Tasa de dependencia (menores de 15 y mayores de 65 años entre la población general)
TOTRAMAS	Total de ramas de actividad laboral
VAR2020	Población de varones en el año 2020
VIVFAM	Total de viviendas familiares
VIVNOPR	Total de viviendas no principales
VIVPRIN	Total de viviendas principales

4. Listado de herramientas usadas:

Máquina: `Windows 10 x64 (build 19041)`

Entorno: `Anaconda (Versión 2021.05) / jupyter 1.0.0 con kernel de R`

Versión de R: `3.6.1 (2019-07-05)`

Paquetes de data wrangling:

```
readr_1.3.1
readxl_1.3.1
dplyr_0.8.0.1
```

Paquetes de visualización:

```
ggplot2_3.1.1
rattle_5.4.0
lattice_0.20-38
RColorBrewer_1.1-2
rpart.plot_3.0.8
```

Paquetes de Machine Learning:

```
tree_1.0-40
rpart_4.1-15
caret_6.0-83
C50_0.1.3.1
randomForest_4.6-14
adabag_4.2
```