Estimación de la tasa de pobreza en Cantabria mediante matching estadístico

Autores: Pablo Lobete López
Francisco J. Parra Rodríguez

DOC. Nº 2/2018
ISSN 2444 - 1627
Santander, Cantabria

Instituto Cántabro de ESTADÍSTICA

DOCUMENTOS TÉCNICOS

ÍNDICE

RESUMEN

El objetivo del presente estudio es establecer una metodología para el cálculo de la tasa de riesgo pobreza en Cantabria a partir de los datos de la Encuesta Social de Cantabria (ESOC) 2015, realizando un matching estadístico con la Encuesta de Presupuestos Familiares (EPF) 2016.

La ESOC es una encuesta realizada por el ICANE tiene como finalidad conocer las aptitudes sociales y condiciones de vida de la población de Cantabria. La muestra objetivo de la ESOC son 1.800 hogares.

La EPF tiene como objetivo suministrar información anual sobre la naturaleza y destino de los gastos de consumo, así como sobre diversas características relativas a las condiciones de vida de los hogares, en base a una muestra de aproximadamente 24.000 hogares para el conjunto del Estado. Para Cantabria, el tamaño de la muestra en 2016 fue de 762 hogares.

La tasa de pobreza monetaria en Cantabria puede calcularse tanto con los microdatos de la Encuesta de Condiciones de Vida (ECV) como con las de la EPF, si bien en el Documento Técnico 1/2017 del ICANE, se concluye que ambas encuestas se realizan con una muestra insuficiente para obtener medidas de pobreza.

Mediante un matching estadístico entre la ESOC y la EPF se pretende realizar un cálculo la tasa de riesgo de pobreza de Cantabria más consistente en el tiempo. En concreto se trata de estimar los ingresos mensuales de los hogares en la ESOC, a partir de una serie de variables comunes a la EPF y de la respuesta a la pregunta sobre el intervalo en que se sitúa la renta disponible de l hogar ingresos en la primera. Señalar que el matching entre la ESOC y la ECV se descartó por la menor muestra de la ECV (347 hogares).

Finalmente, se realiza una estimación de la tasa de riesgo de pobreza en Cantabria a partir de los ingresos de los hogares de la ESOC.

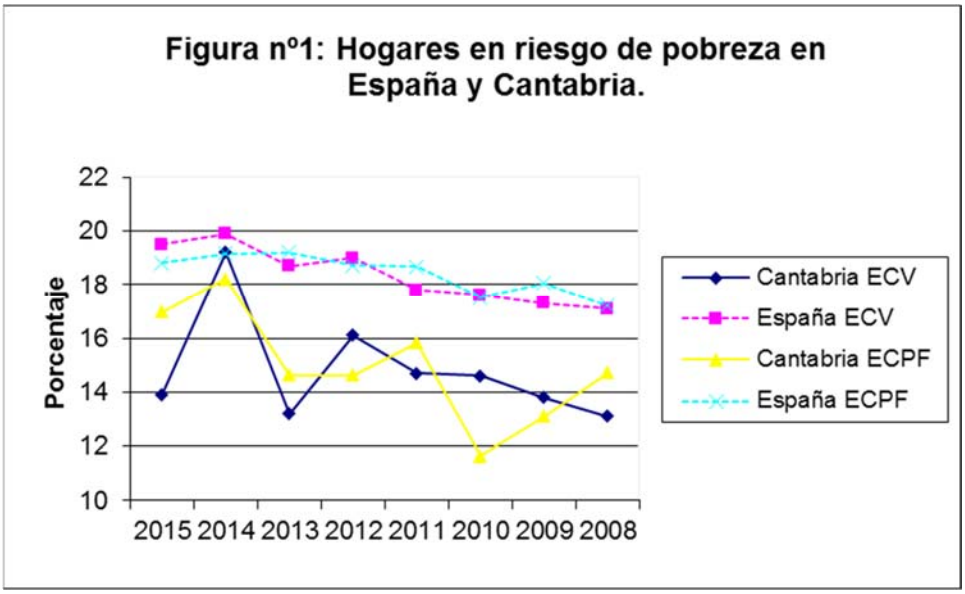
Palabras clave: Tasa de riesgo de pobreza, matching estadístico, Cantabria.

I. INTRODUCCIÓN

El contexto de la crisis económica ha supuesto un aumento de la incidencia de la pobreza y la exclusión social y, al mismo tiempo, ha despertado un mayor interés por estas cuestiones desde un punto de vista social, académico y de políticas públicas. La lucha contra la pobreza y la exclusión social se determinó como uno de los cinco grandes objetivos de la Estrategia Europa 2020, lo que requiere de metodologías fiables para su medición, que permitan evaluar su impacto desde un punto de vista estático y dinámico.

La utilización de técnicas cuantitativas que nos permitan obtener una información precisa es un requisito indispensable para la evaluación y planificación de políticas públicas. En ese sentido, las herramientas para la medición de la tasa de riesgo de pobreza en Cantabria presentan importantes limitaciones, derivadas fundamentalmente de un reducido tamaño muestral en la Comunidad Autónoma de la ECV y la EPF.

En la figura nº1, se aprecia la variabilidad temporal que tienen los cálculos de las tasas de pobreza monetaria que se obtienen en Cantabria con la ECV y EPF. En estos cálculos la línea de pobreza es el 60% de la mediana de los datos regionales.



Fuente; Parra (2017)

La ECV es una encuesta cuya muestra nacional es de alrededor de 13.000 hogares y 35.000 personas, en tanto que la EPF tiene una muestra nacional de aproximadamente 24.000 hogares al año. Si bien para obtener datos nacionales el tamaño muestral de la ECV es considerado suficiente, para detalles regionales de la pobreza la muestra con la que se trabaja puede constituir una seria limitación para interpretar de modo adecuado sus resultados. En Cantabria, por poner un ejemplo, las respuestas válidas de la ECV de 2015 fueron de 347 hogares. La EPF de 2016 investigó 762 hogares en Cantabria.

En Parra (2017), se calcula un intervalo de confianza para la tasa de pobreza en la ECV de 2015 y un 95% de significación estadística del: 12% -18%, en el caso de la tasa de

pobreza de ingresos de la EPF el intervalo para el mismo nivel de significación estadística estaría entre el 14% -20%.

Obtener un cálculo menos variable de la tasa de pobreza en Cantabria requeriría realizar una nueva operación estadística con una metodología similar a la ECV, pero con una mayor muestra de hogares, lo que supondría un gasto adicional para el sistema estadístico, y con el inconveniente de que podría ser visto como una duplicación de estadística, si esta la realizara el centro regional de estadística oficial, el ICANE. No obstante, existen referencias bibliográficas que permite abordar este problema bajo otro enfoque metodológico: el matching estadístico.

El matching estadístico ha sido empleado en la estimación de la privación material y el gasto de los hogares en el estudio Statistical matching of European Union statistics on income and living conditions (EU-SILC) and the household budget survey (Serafino and Tonkin, 2013), donde realiza un matching estadístico entre la Encuesta de Presupuestos Familiares (HBS) y la Encuesta Europea de Ingresos y Condiciones de Vida (EU-SILC), con el objetivo de estimar el gasto y la privación material a partir de un número limitado de variables comunes (Región, grupo de edad, estado de tenencia del hogar, el tipo de vivienda, el tamaño del hogar, la tenencia o no de automovil y el nivel de ingresos disponibles en el hogar). En Eursotat existe también una referencia sobre este tipo de matching: Statistical matching: a model based approach for data integration (Leulescu y Agafitei, 2013), donde se presentan los resultados de dos estudios piloto y se plantean una serie de recomendaciones para la aplicación de esta técnica, así como para optimizar el diseño futuro de las encuestas sociales. En España destacar la fusión de datos entre EPA y ECV para la estimación de la tasa de pobreza de Cataluña que realiza Montoya (2015).

La Encuesta Social de Cantabria, en su diseño, no está planteada como una encuesta sobre distribución de ingresos y gastos, su finalidad es conocer las aptitudes sociales y condiciones de vida de la población de Cantabria, por lo que la información disponible en ésta acerca de los ingresos del hogar, variable necesaria para medir el riesgo de pobreza, solo se pregunta agregada por intervalos de renta bastante amplios, si bien en la encuesta se realizan otra serie de preguntas sobre las características de los hogares y personas entrevistadas que permite un matching estadístico a partir de las variables comunes de la ESOC y la ECV o la EPF. El menor número de hogares que responden a la ECV en Cantabria, unido al hecho de que las estimaciones de la tasa de pobreza en ambas encuestas son muy semejantes, aconseja la preferencia de la EPF para este propósito

El documento se estructura en un apartado descriptivo de la metodología del matching estadístico, un segundo relativo a las fuentes estadísticas utilizadas, un tercero a describir la armonización de variables comunes entre la ESOC y la EPF, un cuarto a la selección de variables y los modelos estadísticos sobre los que se va a realizar la fusión, en un quinto se evalúan los resultados obtenidos, para después, en un sexto aparatado, proceder a calcular la tasa de pobreza de Cantabria. En un apartado final se expondrán las conclusiones.

2. EL MATCHING ESTADÍSTICO: METODOLOGÍA PARA EL CÁLCULO DE LA TASA DE POBREZA EN CANTABRIA

La integración de datos provenientes de diferentes fuentes tiene su desarrollo inicial en los Estados Unidos y en la década de 1960. En Montoya (2015) se ofrece una ajusta revisión bibliográfica sobre el desarrollo de estas técnicas y su desarrollo en el Sistema Estadístico Europeo.

La integración de datos se puede realizar por medio de tres metodologías diferentes: data merging, record linkage y statistical matching (D'Orazio et al., 2001). El matching estadístico es un enfoque basado en modelos para proporcionar información estadística conjunta con variables e indicadores recopilados a través de dos o más fuentes, cuyas unidades de análisis provienen de una misma población y poseen variables en común pero no se superponen. Básicamente consiste en imputar a unos individuos (receptores) información para algunas variables a partir de la información proveniente de otros individuos (donantes), a los que se les han observado algunas características comunes y que se relacionan con la información que se quiere estimar (Leulescu y Agafitei, 2013).

Las diferentes bases de datos empleadas se ponen en relación a través de una serie de variables comunes, que deben cumplir algunos requisitos básicos en cuanto a su distribución y deben ser similares en su estructura. La fusión de ficheros se inicia, por tanto, con el estudio de las bases de datos, localizando esas variables comunes, que va acompañado de una armonización de las mismas, recodificándolas para que coincidan completamente en una y otra base de datos. Solo podrán emplearse para el matching estadístico aquellas variables que puedan armonizarse para que sean similares en su definición, teniendo que descartar aquellas donde no sea posible. En cuanto a su distribución, al tratarse de variables representativas de una misma población, se requiere que las variables comunes, una vez armonizadas, compartan una distribución similar.

En un segundo paso, se genera un modelo a partir de esas variables comunes para tratar de estimar el valor de la variable objetivo Z que solo existe en una de las bases de datos, pero no en la otra. Los procedimientos de correspondencia se pueden considerar como un problema de imputación de las variables objetivo de una encuesta de donante a receptor, tomando una serie de campos comunes entre ambas que están correlacionadas con Z, por lo que el conjunto de datos del donante será explorado y utilizado para imputar a las unidades del otro conjunto de datos: el destinatario (Leulescu y Agafitei, 2013).

Para la realización del matching estadístico entre la ESOC y la EPF, siguiendo las recomendaciones metodológicas se han establecido los siguientes pasos:

Estudio de las bases de datos e identificación de variables comunes a ambos conjuntos de datos.

Recodificación y armonización de las variables comunes cuando ha sido necesario.

Evaluación individual de dichas variables, comprobando que su distribución es similar.

Formulación de diferentes modelos explicativos o funciones de enlace de la variable objetivo Z en la base de datos de la EPF (2016).

Evaluación de la capacidad explicativa de cada modelo y elección de los más adecuados.

Estimación de la variable objetivo en la Encuesta Social de Cantabria (2015) a partir del modelo seleccionados.

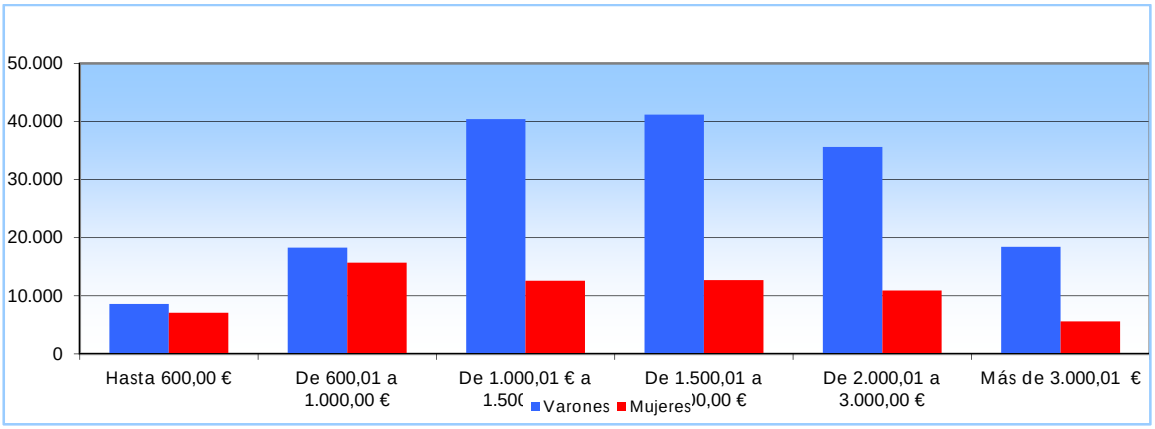
3. FUENTES UTILIZADAS

La EPF de 2016 realizada por el Instituto Nacional de Estadística (INE) tiene un total de 761 casos válidos para la Comunidad Autónoma de Cantabria, y recoge el importe exacto del nivel de ingresos del hogar como una variable continua, con la siguiente definición: conjunto de los percibidos regularmente por el hogar y sus miembros perceptores de ingresos individuales, cualquiera que sea su origen, una vez descontadas las cotizaciones a la Seguridad Social y otros pagos asimilados (Entidad de Previsión Social, Mutualidades Obligatorias y Derechos Pasivos), así como las cantidades satisfechas en concepto de impuestos.

Los ingresos se investigan son los regulares mensuales del hogar tanto en valor puntual como en intervalo, así como los tipos de fuentes de ingresos para cada hogar, también se recoge el número de miembros del hogar que son perceptores de ingresos .Los ingresos incluyen las rentas procedentes del trabajo por cuenta ajena, por cuenta propia, subsidios, prestaciones, pensiones, rentas del capital y de la propiedad, así como otros ingresos regulares (donaciones de instituciones, transferencias de otros hogares, remesas de emigrantes y otros ingresos regulares distintos a las prestaciones sociales). Si bien, hay que señalar que la EPF realiza una imputación de los registros en el caso que el hogar no proporcione el valor puntual, a partir del intervalo el que se encontrarían sus ingresos, esta imputación se lleva a cabo ajustando la información disponible a un modelo de regresión[1]

La Encuesta Social de Cantabria (2015), realizada por el Instituto Cántabro de Estadística (ICANE), tiene una muestra de un total de 1606 casos válidos. La tabulación de los resultados a la pregunta sobre el intervalo de ingresos en el hogar se presenta en la figura nº 2

Figura n°2. Hogares por tramos de Renta y sexo del sustentador principal



Fuente: ESOC.2013 y 2015

Por su parte en la tabla nº1 se puede apreciar la respuesta a la pregunta sobre fuente de renta mensual en la ESOC:

[1] Para ello se ha utilizado el módulo de imputación del programa IVE (Imputation and Variance Estimator), desarrollado por el Institute of Social Research de la Universidad de Michigan.

Tabla nº1. Ingresos por fuentes de Renta

Fuentes de renta	Número	% total hogares
Ingresos por trabajo	170,367	71.96%
Prestaciones contributivas (pensiones por jubilación,...)	86,428	36.51%
Prestaciones no contributivas	12,135	5.13%
Prestaciones de subsidio por desempleo	20,134	8.50%
Ayudas públicas por familia / hijos (1)	11,713	4.95%
Ayuda para vivienda (subvenciones para compra o alquiler) (1)	2,254	0.95%
Ingresos por rentas	13,171	5.56%
Transferencias periódicas monetarias percibidas entre hogares (1)	1,839	0.78%
Otros ingresos	12,244	5.17%

Nota (1): El número de observaciones muestrales es menor que 20, por lo que la cifra hay que interpretarla con cautela.

Fuente ESOC

Para realizar el matching sobre ambas encuestas se han seleccionado las siguientes las variables comunes (covariables), que entendemos pueden ser utilizadas para la fusión de ambos ficheros con el fin de obtener la tasa de pobreza de los hogares en Cantabria:

- Régimen de tenencia de la vivienda
- Disposición o no de otras viviendas
- Tamaño del municipio
- Sexo del/la sustentador/a principal
- Número de ocupados en el hogar
- Ocupación del sustentador principal
- Número de habitaciones de la vivienda
- Número de mayores de 65 en el hogar

4. TRATAMIENTO DE LOS DATOS

Para la realización del matching estadístico, partimos de las variables comunes a los ficheros de hogares de la ESOC y la EPF. El primer paso es la armonización de las definiciones de dichas variables en ambas encuestas. Los criterios sobre los que se ha basado la armonización se recogen en la Tabla nº2.

Tabla 2. Criterios para la armonización de las variables comunes (EPF-ESOC)

Variables en EPF 2016 y ESOC 2015	Estructura
Régimen de tenencia de la vivienda	Recodificada para armonizarla con la ESOC 2015. Finalmente contiene tres niveles: 1 Propiedad 2 Alquiler 3 Cesión
Disposición o no de otras viviendas	1 Sí 2 No
Tamaño del municipio	Recodificada en EPF, pasado de 6 estratos 5 (por no tener el 6º valores para Cantabria). A su vez, se ha generado la variable "Tamaño del municipio" en la ESOC 2015, a partir de código de municipio del hogar. Finalmente, cuenta con 5 niveles: 1 Menos de 10.000 hab 2 De 10.000 a 20.000 hab 3 De 20.000 a 50.000 hab 4 De 50.000 a 100.000 hab 5 100.000 hab o más
Sexo del/la sustentador/a principal	1 Hombre 0 Mujer
Número de ocupados en el hogar	
Ocupación del sustentador principal	1 Ocupado 2 No ocupado
Número de habitaciones de la vivienda	Finalmente excluida por contener un alto número de casos perdidos en la ESOC 2015
Número de mayores de 65 en el hogar	Esta variable se crea en la EPF a partir de las edades individuales de los miembros del hogar

En el Anexo I se puede consultar en código R el proceso de armonización de las variables utilizadas en el matching de ambas encuestas.

Destacar que uno de los propósitos iniciales era generar una variable que se denominó *hacinamiento*, poniendo en relación el número de miembros del hogar con el número de habitaciones de la vivienda. También se creó otra definición alternativa de la variable *hacinamiento* que únicamente tenía en cuenta el número de personas adultas del hogar.

Sin embargo, dicha variables, hubo que excluirla finalmente del análisis, por presentar un elevado número de casos perdidos en la ESOC.

El proceso de recodificación de las variables implicó varias transformaciones en las variables de la ESOC 2015, así como la generación de varias variables nuevas. Por una parte, se han crearon variables tipo *dummy* para cada intervalo manifestado de ingresos y, por otra se generaron las variables "Número de mayores de 65", "Número de ocupados" y "Número de menores de 14 años" a partir de la suma del resultado de las variables individuales de cada miembro del hogar.

Por otro lado, en la ESOC hubo de obtenerse la variable "Unidades de consumo", que se emplea para el cálculo del umbral de pobreza y de la tasa de riesgo de pobreza.

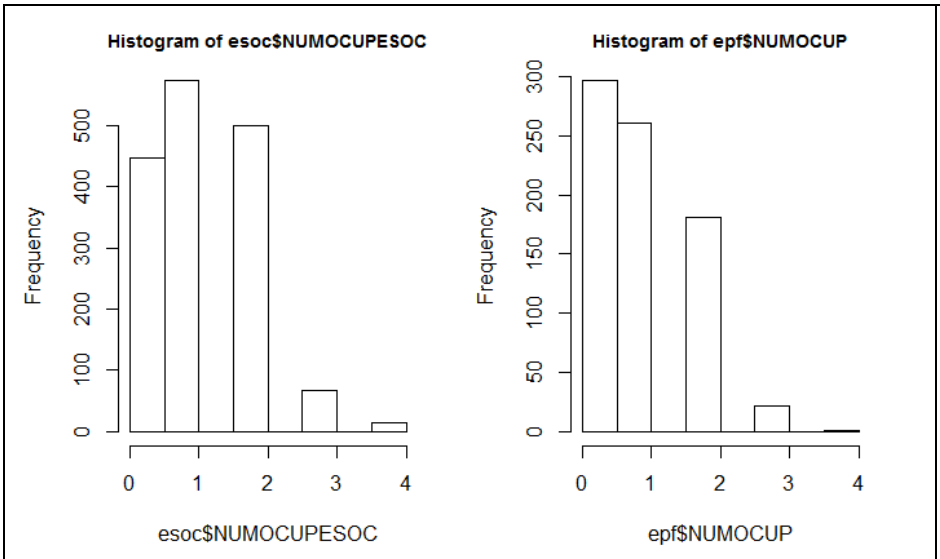
Un requisito para realizar el matching es que las variables seleccionadas deben distribuirse de manera similar en ambos conjuntos. Se evaluó la similitud de la distribución de las variables en cada uno a través de un análisis gráfico, y con la medida de la Distancia de Hellinger (HD), que considera adecuado un resultado inferior al 5% (Leulescu & Agafitei ,2013).

$$HD(V,V') = \sqrt{\frac{1}{2}\cdot\sum_{i=1}^{K}\left(\sqrt{p(V=i)}-\sqrt{p(V'=i)}\right)^2} = \sqrt{\frac{1}{2}\cdot\sum_{i=1}^{K}\left(\sqrt{\frac{n_{Di}}{N_D}}-\sqrt{\frac{n_{Ri}}{N_R}}\right)^2}$$

Donde K es el número total de celdas de la tabla de contingencia, n_{Di} es la frecuencia de la celda "i" de los datos del donante D, n_{Ri} es la frecuencia de la celda "i" de los datos del recipiente R y N el tamaño total de la tabla de contingencia específica.

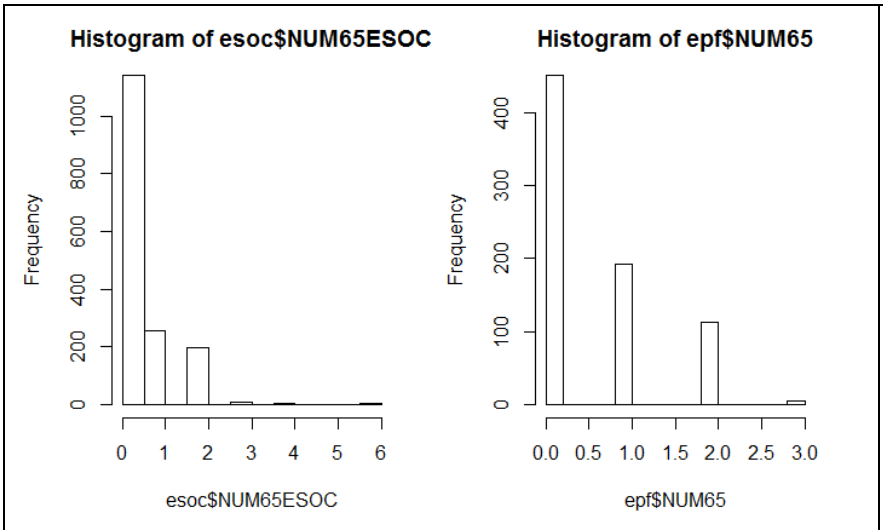
Como podemos observar en las figuras 3-8, la distribución por lo general es bastante similar en ambas estadísticas, obteniendo Distancias de Hellinger inferiores a 0,05 para todos casos.

Figura 3. Histogramas de la variable armonizada *número de ocupados* en las bases de datos de EPF y ESOC y Distancia de Hellinger entre ambas distribuciones.



Distribución de la variable número de ocupados en ESOC (izquierda) y EPF (derecha)

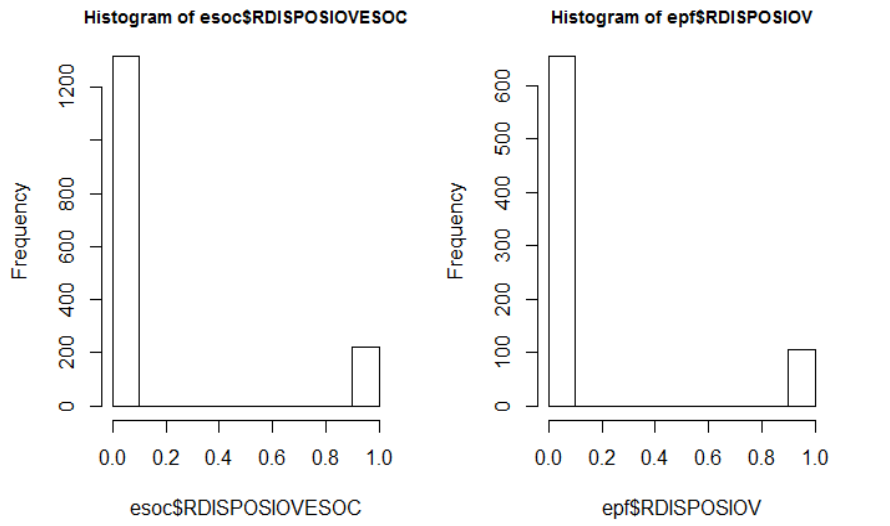
Figura 4. Histogramas de la variable armonizada *número de mayores de 65 años* en las bases de datos de EPF y ESOC y Distancia de Hellinger entre ambas distribuciones.



Hellinger Distance: 0.01554986

Distribución de la variable número de mayores de 65 años en ESOC (izquierda) y EPF (derecha)

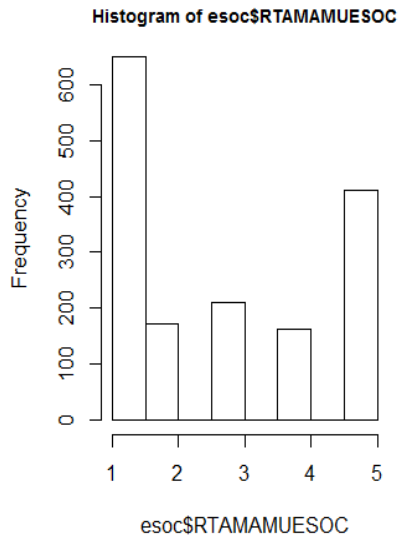
Figura 5. Histogramas de la variable armonizada *disposición de 2ª vivienda* en las bases de datos de EPF y ESOC y Distancia de Hellinger entre ambas distribuciones.

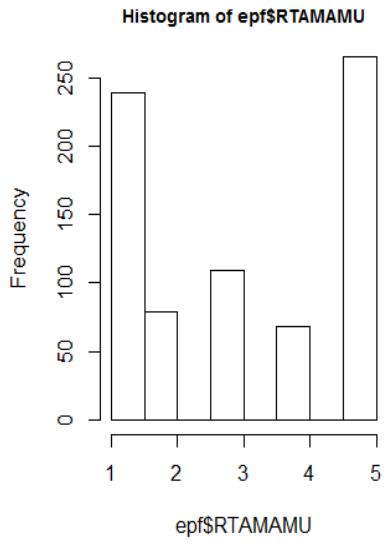


Hellinger Distance: 0.003663777

Distribución de la variable Disposición de segunda vivienda en ESOC (izquierda) y EPF (derecha)

Figura 6. Histogramas de la variable armonizada *tamaño del municipio* en las bases de datos de EPF y ESOC y Distancia de Hellinger entre ambas distribuciones.

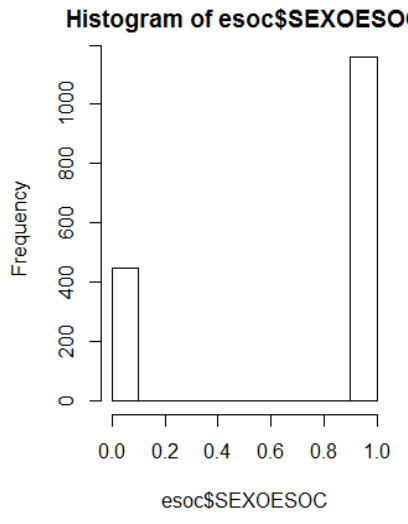


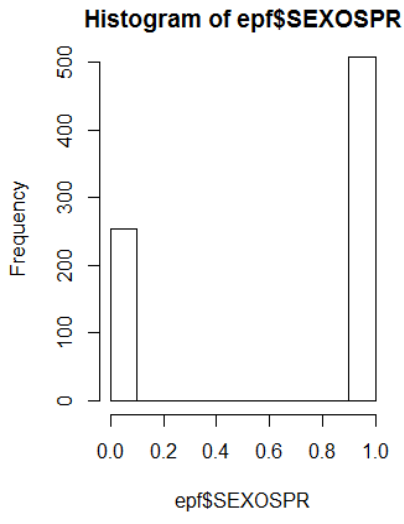


Hellinger Distance: 0.008964882

Distribución de la variable Tamaño del municipio en ESOC (izquierda) y EPF (derecha)

Figura 7. Histogramas de la variable armonizada sexo del sustentador principal en las bases de datos de EPF y ESOC y Distancia de Hellinger entre ambas distribuciones.

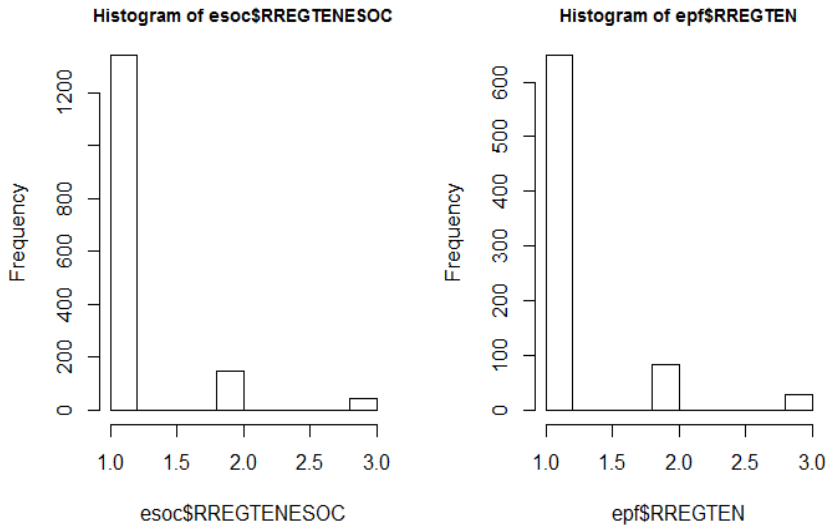




Hellinger Distance: 0.01178758

Distribución de la variable Sexo en ESOC (izquierda) y EPF (derecha)

Figura 8. Histogramas de la variable armonizada *régimen de tenencia de la vivienda principal* en las bases de datos de EPF y ESOC y Distancia de Hellinger entre ambas distribuciones.



Hellinger Distance: 0.01453584

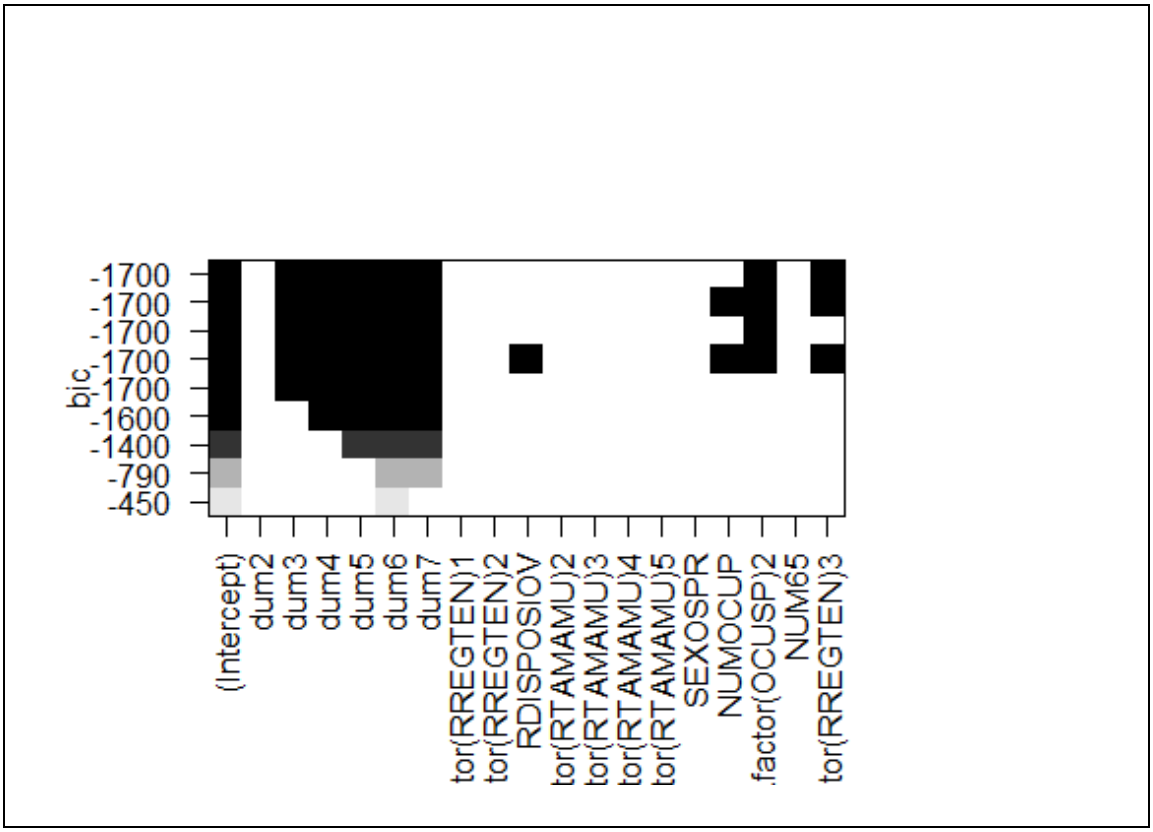
Distribución de la variable Sexo en ESOC (izquierda) y EPF (derecha)

5. SELECCIÓN DE VARIABLES Y APLICACIÓN DE LAS TÉCNICAS DE FUSIÓN

Para reducir el conjunto de variables explicativas se ha realizado preselección de variables utilizando un procedimiento automático: la librería "leaps" de R, realizando el ejercicio de selección automática con el método "forward"[2]. En el Anexo II y Figura 9 se presentan el código R utilizado y los resultados obtenidos.

[2] En la selección de variables por este método se comienza por un modelo que no contiene ninguna variable explicativa y se añade como primera de ellas a la que presente un mayor coeficiente de correlación -en valor absoluto- con la variable dependiente. En los pasos sucesivos se va incorporando al modelo aquella variable que presenta un mayor coeficiente de correlación parcial con la variable dependiente dadas las independientes ya incluidas en el modelo. El procedimiento se detiene cuando el incremento en el coeficiente de determinación debido a la inclusión de una nueva variable explicativa en el modelo ya no es importante.

Figura 9. Selección de variables explicativas por "leap"



Analizados los resultados se decidió utilizar como covariables, en la función de enlace las siguientes:

- Las variables dummies que recogen los estratos correspondientes a los niveles de renta que se utilizan en la ESOC.
- Régimen de tenencia de la vivienda
- Sexo del/la sustentador/a principal
- Número de ocupados en el hogar
- Número de mayores de 65 en el hogar

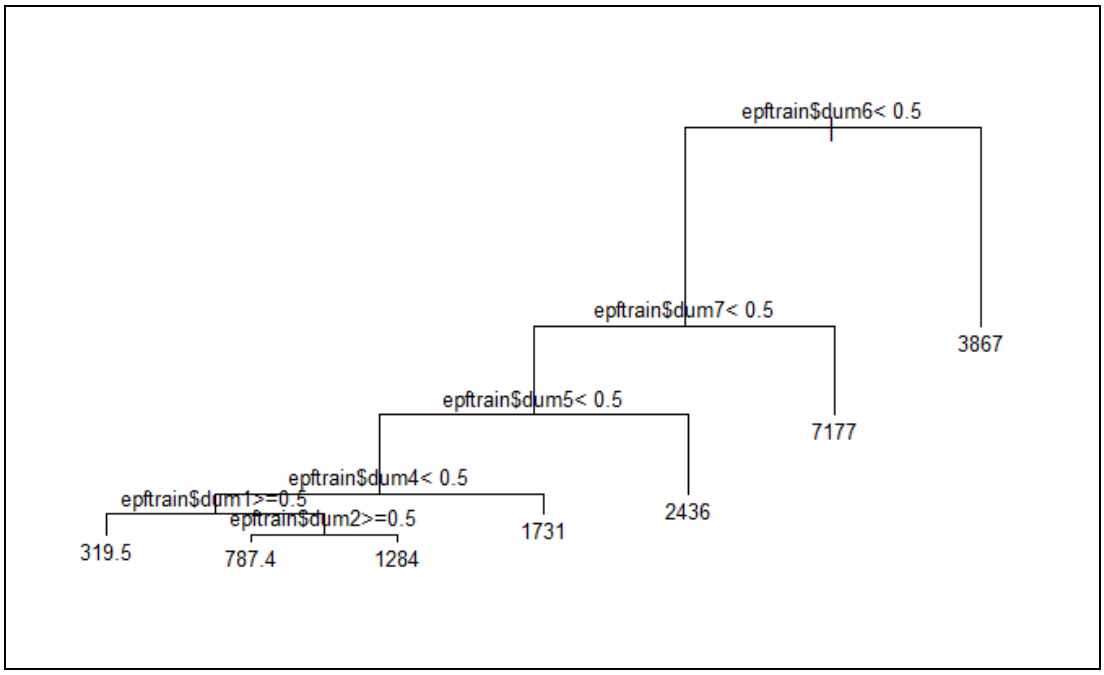
Para especificar y cuantificar la función de enlace con la que realizar el matching, exploramos métodos tanto no paramétricos como paramétricos.

En cuanto a los métodos no paramétricos, se ha realizado:

- **Un árbol de regresión**. Los árboles de decisión o de clasificación son un modelo surgido en el ámbito del aprendizaje automático (Machine Learning), que partiendo de una base de datos, crea diagramas de construcciones lógicas que nos ayudan a resolver problemas. A esta técnica también se la denomina segmentación jerárquica. Es una técnica explicativa y descomposicional que utiliza un proceso de división secuencial, iterativo y descendente, que parte de una variable dependiente, forma grupos homogéneos definidos específicamente mediante

combinaciones de variables independientes en las que se incluyen la totalidad de los casos recogidos en la muestra. Los árboles resuelven tanto problemas de clasificación como de regresión. Aquí se ha utilizado el package "rpart", *Recursive Partitioning and Regression Trees*. El resultado obtenido se presenta en la figura nº9

Figura 10. Árbol de regresión

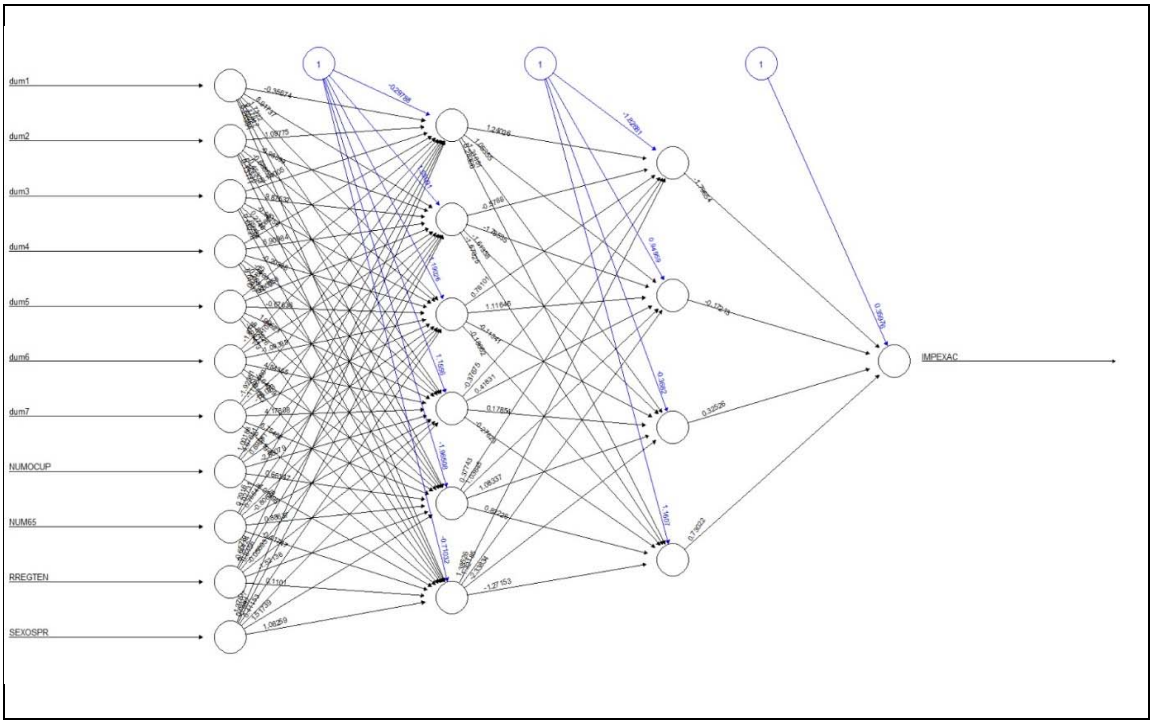


En la representación del árbol de regresión se aprecia que las particiones secuenciales se realizan exclusivamente con las variables *dummys*, dado que , el árbol, normalmente usa la media o la moda de las observaciones de entrenamiento en la región a la que pertenece (James, G et al, 2013) para realizar las predicciones, el resultado devolverá la estructura de factores que más se acercan a la explicativa y no una variable continua con los ingresos estimados para cada hogar. Estos resultados descartan la utilidad del árbol para la finalidad que se persigue.

- **Una red neuronal**. Es un modelo compuesto por nodos, donde se sigue un patrón adaptativo y de "aprendizaje" en los datos, de modo que a medida que se recopilan nuevos casos y conjuntos de datos, los modelos se ajustan y permanecen generalizables. Son eficaces cuando se dan casos de no linealidad en los datos, por lo que optamos por explorar esta opción por si no fuera posible predecir los valores de la variable objetivo a través de modelos paramétricos. (DeTienne et al., 2003).

Para la aplicación y entrenamiento de la red neuronal, utilizamos el package de R "neuralnet". En primer lugar, es necesario generar una base de datos con las variables normalizadas. A continuación, se ha segmentado la muestra en dos submuestras, una de test y otra de entrenamiento. La muestra de entrenamiento recoge el 70% de los casos, mientras que la muestra de test recoge el 30% restante. El objetivo es obtener la definición de red con mejor capacidad explicativa. En el Anexo III, se presenta el código R utilizado para definir la red neuronal. El resultado se representa en la figura 11.

Figura 11. Red Neuronal



En cuanto a los métodos paramétricos, realizamos varios modelos:

- Regresión lineal
- Regresión gamma
- Regresión logarítmica

La especificación general de estos modelos es:

$$I = D\alpha + \beta X + u$$

Donde I es el vector del ingreso estimado para cada hogar i, D, es la matriz de las variables *dummys* que establecen los intervalos de ingresos en donde se sitúa cada hogar, X es la matriz del resto de las covariables para cada hogar y u un vector de errores aleatorios.

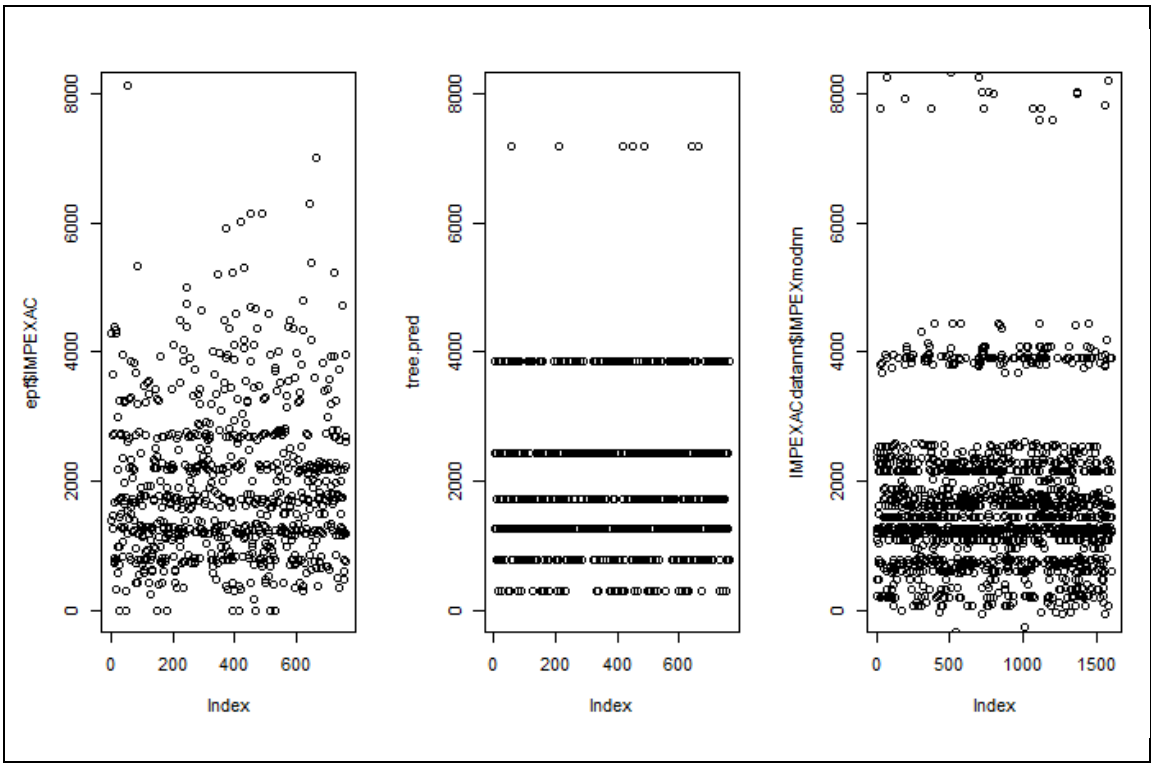
En estos modelos, los α, asociados a las variables *dummys* que recogen los intervalos de ingresos del hogar, determinan un valor medio para cada intervalo de ingreso, siendo el resto de variables explicativas las que establecen el mayor o menor valor con respecto a ese α. A efectos de evitar posibles problemas de multicolinealidad, la ecuación se especifican sin término constante.

En el Anexo IV, se presenta el código R utilizado para la estimación de los modelos paramétricos y los resultados obtenidos:

6. EVALUACIÓN DE LA CALIDAD DE LOS RESULTADOS DE LA FUSIÓN

En las figuras 12 y 13, se puede analizar de forma gráfica la distribución de las estimaciones de ingresos obtenidas a partir de los modelos utilizados anteriormente. En la primera imagen de la figura 9 se presenta la distribución de los ingresos en la EPF (variable Z).

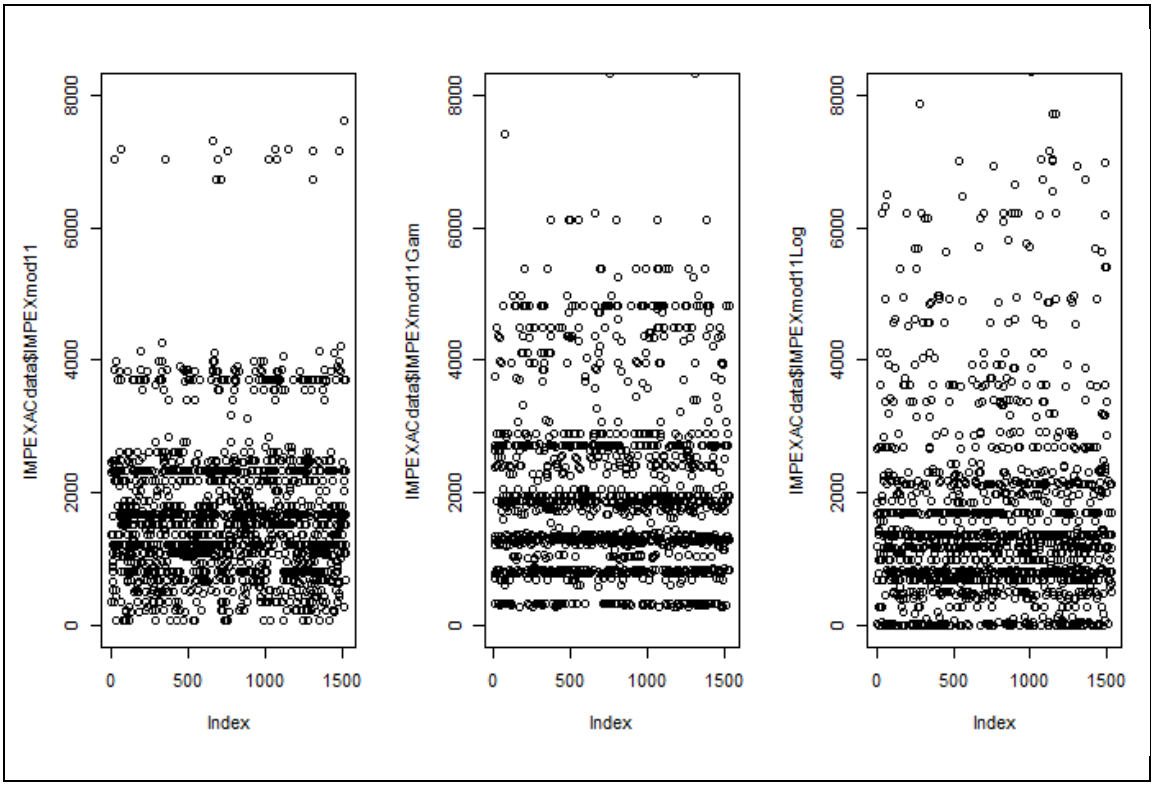
Figura 12. Puntuaciones de la variable ingresos en la EPF 2016, estimación de ingresos de la red neuronal y estimación de ingresos del árbol de regresión



En la figura 12, donde se representan los resultados de los métodos no paramétricos junto a la distribución de la variable objetivo, se observa, como era de esperar, que el árbol de regresión concentra todos los de casos en a los coeficientes asociados a las dummys que recogen los intervalos de ingresos. Esta falta de continuidad en la estimación, invalida el resultado para el objetivo que pretendemos. Las estimaciones de la red neuronal tienen una mayor variabilidad en los estratos más bajos, apartándose del objetivo en los estratos de rentas más altas.

Como vemos, los modelos paramétricos de la figura 12 presentan una distribución más semejante a la distribución de la variable ingresos en la EPF, las estimaciones para los niveles de renta más altos son más adecuadas en los modelos de la Regresión Logarítmica y de la Regresión Gamma.

Figura 13. Estimaciones del modelo lineal, modelo gamma y modelo logarítmico.



En la Tabla 3, recogemos una serie de estadísticos sobre los errores para evaluar el resultado de las estimaciones realizadas. Se trata de los estadísticos AIC[3], RMS (del inglés root mean square) y la tasa o porcentaje de aciertos. El AIC y el RMS se calculan lógicamente con los datos utilizados para estimar el modelo, es decir los de la EPF, los otros dos estadísticos No así la tasa de aciertos, esta tasa requiere estimar el modelo con los datos de la ESOC, obtener los intervalos de renta de acuerdo a como se preguntan en la ESOC y comparar si los nuevos intervalos coinciden con las respuestas dadas por los hogares (en el Anexo V se puede consultar la programación en R de dicho cálculo).

[3] Criterio de Información de Akaike

Tabla 3. AIC, suma del error cuadrático medio y % de acierto de cada modelo

Modelo	AIC	Suma del error cuadrático medio	% de acierto[1]
Regresión lineal	11325.44	164867.67	87,54%
Regresión gamma	12384.02	494310.80	90,47%
Regresión logarítmica	2300.67[2]	7399601	85,29%
Red neuronal	-	209242.85	79,04%
Árbol de regression	-	114144.31	94,14%

[1] El % de acierto o tasa de acierto se calcula a partir de la matriz de confusión entre la estimación realizada por el modelo y el intervalo de ingresos de la ESOC para cada hogar. El cociente de los hogares clasificados correctamente (verdaderos positivos), entre todos los elementos clasificados en esa clase (verdaderos positivos + falsos positivos)

[2] No es posible comparar el valor de AIC de la distribución logarítmica con el AIC de las otras distribuciones.

Los resultados de los estadísticos de regresión apuntan al modelo de Regresión Lineal como el más acertado de los modelos estimados. La tasa de aciertos del 87,54% estarían ligeramente más baja que la Regresión Gamma. Destacar por último que es el Árbol de Regresión el que ofrece el menor RMS y la mayor tasa de aciertos, pero por las razones anteriormente expuestas queda descartado para la finalidad del análisis. En definitiva, a partir del análisis gráfico y de los estadísticos que aparecen en la tabla, se ha elegido como el más apropiado el modelo de regresión lineal.

7. ESTIMACIÓN DE LA TASA DE RIESGO DE POBREZA

Una vez realizado el matching entre la ESOC y la EPF, con las estimaciones que nos proporciona el modelo lineal, disponemos de una estimación nominal de la renta disponible de cada hogar, que obtenida por unidad de consumo nos permite calcular de la tasa de riesgo de pobreza para la ESOC 2015.

Para calcular el número de unidades de consumo del hogar se emplea la escala de la OCDE modificada, concediendo un peso de 1 al primer adulto, un peso de 0,5 a los demás adultos y un peso de 0,3 a los menores de 14 años. Una vez calculado el ingreso por unidad de consumo del hogar se adjudica éste a cada uno de sus miembros.

Tomamos como línea de pobreza el umbral fijado en el 60% de la mediana de los ingresos por unidad de consumo de los hogares de Cantabria.

La tasa de riesgo de pobreza monetaria en la ESOC de 2015 se situaría en el 19,28%, resultando ser más alta que las estimaciones de dicha tasa en la ECV y ESP, los intervalos de confianza calculados mediante un bootstrapping con 1000 replicaciones, sitúan los intervalos de confianza para dicha tasa entre el 17,69% y el 20,95% para un nivel de significación del 90% y de entre un 17,38% y un 21,26% para un nivel del 95%. La amplitud de estos intervalos es menor que el cálculo realizado para la ESOC y la ESP (Parra, 2017).

8. CONCLUSIONES

La estadística oficial que recoge información sobre la tasa de pobreza monetaria presenta dificultades muestrales en Cantabria.

Se ha empleado la técnica de matching estadístico, obteniendo una estimación nominal del nivel de ingresos del hogar, lo que permite el cálculo de la tasa de riesgo de pobreza para Cantabria en el año 2015 con datos de la ESOC del ICANE.

A nivel metodológico se han encontrado con algunas dificultades para la aplicación del matching. En primer lugar, a la hora de identificar variables comunes se ha de tener en cuenta que la existencia de pequeñas diferencias en el diseño de las preguntas, puede traducirse en importantes discrepancias en los datos que imposibilitan la utilización de esas variables en el proceso. La armonización de las definiciones de las variables comunes ha permitido una mejor integración entre conjuntos de datos. Por otra parte, la existencia de variables con un número relativamente elevado de casos perdidos supuso una dificultad añadida.

La función de enlace utiliza como regresores los intervalos o tramos de ingresos del hogar manifestados por la persona entrevistada, teniendo el resto de variables del modelo (nº de ocupados, nº de mayores de 65 años, régimen de tenencia de la vivienda, sexo del sustentador principal del hogar) un peso menor en el modelo, actúan como graduadores del nivel de ingresos entre los intervalos. Los modelos no paramétricos son los que ofrecen los peores resultados para nuestros propósitos.

Con respecto a la tasa de riesgo de pobreza estimada, obtenemos un resultado que resulta coherente en relación a las tasas de pobreza conocidas mediante otras fuentes estadísticas hasta la fecha, pero con menor variabilidad estadística.

Bibliografía

D'Orazio, M., Di Zio, M., & Scanu, M. (2001). Statistical Matching: a tool for integrating data in National Statistical Institutes. Second International Seminar of Exchange of Technology and Know-How/Fourth New Techniques and Technologies for Statistics Seminar. Creta, junio de 2001.

DeTienne, K. B.; DeTienne, D. H.; & Johsi, S. A. (2003). *Neural networks as a statistical tool for business researchers.* Organizational Research Methods, 6, 236-265.

Fritsch, S. and Guenther, F. (2016). *neuralnet: Training of Neural Networks.* R package version 1.33. **https://CRAN.R-project.org/package=neuralnet**

Intrator O. and Intrator N. (1993) *Using Neural Nets for Interpretation of Nonlinear Models.* Proceedings of the Statistical Computing Section, 244-249 San Francisco: American Statistical Society (eds).

James, G.; Witten, D.; Hastie, T.; Tibshirani. R. (2013). *An Introduction to Statistical Learning (with Applications in R)*: **http://www-bcf.usc.edu/~gareth/ISL/**

Leulescu, A. and Agafitei, M. (2013) *Statistical matching: a model based approach for data integration*. Eurostat methodologies and working paper, Eurostat.

Montoya, J (2015): *Fusión de datos entre EPA y ECV para la estimación de indicadores estadísticos regionals*. Facultat de Matemàtiques i Estadística. Universitat Politècnica de Catalunya.Trabajo de fin de Máster

Parra, F (2017). *Indicadores de pobreza infantil en Cantabria*. Documento Técnico 1/2017. Instituto Cántabro de Estadística (ICANE). **http://www.icane.es/c/document_library/get_file?uuid=27c26368-7e14-43e4-b1d9-b724eb84121a&groupId=10138**

Parra, F. and Campo, L. (2015) *Estimación de la tasa de pobreza en Cantabria en Área Pequeña.* Documento Técnico 2/2015. Instituto Cántabro de Estadística (ICANE). **http://www.icane.es/c/document_library/get_file?uuid=353fa74f-a325-4b30-a78e-df9db6ee1e24&groupId=10138**

Serafino, P. and Tonkin, R. (2013) *Statistical matching of European Union statistics on income and living conditions (EU-SILC) and the household budget survey.* Statistical Working Papers. Eurostat.

Therneau, T and Atkinson, B (2018). *rpart: Recursive Partitioning and Regression Trees.* R package version 4.1-13. **https://CRAN.R-project.org/package=rpart**

Anexo I.

```
#Recodificación EPF

library(readxl)
epf <- read_excel("F:/…/EPF_2016-2.xls")

attach(epf)

#Régimen de tenencia RECODIFICADA
epf$RREGTEN[epf$REGTEN == "1"] <- 1
epf$RREGTEN[epf$REGTEN == "2"] <- 1
epf$RREGTEN[epf$REGTEN == "3"] <- 2
epf$RREGTEN[epf$REGTEN == "4"] <- 2
epf$RREGTEN[epf$REGTEN == "5"] <- 3
epf$RREGTEN[epf$REGTEN == "6"] <- 3

#Disposición de otras viviendas RECODIFICADA
epf$RDISPOSIOV[epf$DISPOSIOV == "6"] <- 0
epf$RDISPOSIOV[epf$DISPOSIOV == "1"] <- 1

#Creación de la variable "mayores de 65" NUM65
epf$NUM65 <- epf$NMIEM13+epf$NMIEM12

#Creación de la variable "Hacinamiento" HACIN
epf$HACIN <- epf$NMIEMB/epf$NHABIT

#Creación de la variable "Hacinamiento adultos" HACIN18
epf$HACINAD <- epf$NMIEM6/epf$NHABIT

#Recodificación de TAMAMU
epf$RTAMAMU[epf$TAMAMU == "1"] <- 1
epf$RTAMAMU[epf$TAMAMU == "2"] <- 2
epf$RTAMAMU[epf$TAMAMU == "3"] <- 3
epf$RTAMAMU[epf$TAMAMU == "4"] <- 4
epf$RTAMAMU[epf$TAMAMU == "5"] <- 5
epf$RTAMAMU[epf$TAMAMU == "6"] <- 5

#Nueva variable +20000 habitantes (2) o menos (1)
epf$RTAMAMU2[epf$RTAMAMU == "1"] <- 1
epf$RTAMAMU2[epf$RTAMAMU == "2"] <- 1
epf$RTAMAMU2[epf$RTAMAMU == "3"] <- 2
epf$RTAMAMU2[epf$RTAMAMU == "4"] <- 2
epf$RTAMAMU2[epf$RTAMAMU == "5"] <- 2

#Variable SEXO
epf$SEXOSPR[epf$SEXOSP == "1"] <- 1
```

```r
epf$SEXOSPR[epf$SEXOSP == "6"] <- 0
detach(epf)

#Creación de una variable DUMMY para cada intervalo
epf$dum1 <- ifelse (epf$IMPEXAC < 600, 1,0)
epf$dum2 <- ifelse (epf$IMPEXAC < 1000 & epf$IMPEXAC > 600, 1,0)
epf$dum3 <- ifelse (epf$IMPEXAC < 1500 & epf$IMPEXAC > 1000, 1,0)
epf$dum4 <- ifelse (epf$IMPEXAC < 2000 & epf$IMPEXAC > 1500, 1,0)
epf$dum5 <- ifelse (epf$IMPEXAC < 3000 & epf$IMPEXAC > 2000, 1,0)
epf$dum6 <- ifelse (epf$IMPEXAC < 6000 & epf$IMPEXAC > 3000, 1,0)
epf$dum7 <- ifelse (epf$IMPEXAC > 6000, 1,0)

#Recodificación ESOC

attach(esoc)

#Variable NUMOCUP en la ESOC
#Obtención del número de ocupados por hogar
esoc$P4_RR[P4_ocup1SP == "Ocupado/a tiempo completo"] <- 1
esoc$P4_RR[P4_ocup1SP == "Ocupado/a tiempo parcial"] <- 1
esoc$P4_RR2[P4_ocup2 == "Ocupado/a tiempo completo"] <- 1
esoc$P4_RR2[P4_ocup2 == "Ocupado/a tiempo parcial"] <- 1
esoc$P4_RR3[P4_ocup3 == "Ocupado/a tiempo completo"] <- 1
esoc$P4_RR3[P4_ocup3 == "Ocupado/a tiempo parcial"] <- 1
esoc$P4_RR4[P4_ocup4 == "Ocupado/a tiempo completo"] <- 1
esoc$P4_RR4[P4_ocup4 == "Ocupado/a tiempo parcial"] <- 1
esoc$P4_RR5[P4_ocup5 == "Ocupado/a tiempo completo"] <- 1
esoc$P4_RR5[P4_ocup5 == "Ocupado/a tiempo parcial"] <- 1
esoc$P4_RR6[P4_ocup6 == "Ocupado/a tiempo completo"] <- 1
esoc$P4_RR6[P4_ocup6 == "Ocupado/a tiempo parcial"] <- 1
esoc$P4_RR7[P4_ocup7 == "Ocupado/a tiempo completo"] <- 1
esoc$P4_RR7[P4_ocup7 == "Ocupado/a tiempo parcial"] <- 1
esoc$P4_RR8[P4_ocup8 == "Ocupado/a tiempo completo"] <- 1
esoc$P4_RR8[P4_ocup8 == "Ocupado/a tiempo parcial"] <- 1
esoc$P4_RR9[P4_ocup9 == "Ocupado/a tiempo completo"] <- 1
esoc$P4_RR9[P4_ocup9 == "Ocupado/a tiempo parcial"] <- 1
esoc$P4_RR10[P4_ocup10 == "Ocupado/a tiempo completo"] <- 1
esoc$P4_RR10[P4_ocup10 == "Ocupado/a tiempo parcial"] <- 1
esoc$P4_RR3[is.na(esoc$P4_RR3)] <- 0
esoc$P4_RR4[is.na(esoc$P4_RR4)] <- 0

#Suma de las nuevas variables
esoc$NUMOCUPESOC = rowSums (esoc[,53:62])
detach(esoc)

#Variable NUM65 en la ESOC
attach(esoc)
#Obtención del número de mayores de 65 por hogar
esoc$P4_R[P4_edad1SP > 65] <- 1
esoc$P4_R2[P4_edad2 > 65] <- 1
esoc$P4_R3[P4_edad3 > 65] <- 1
```

```r
esoc$P4_R4[P4_edad4 > 65] <- 1
esoc$P4_R5[P4_edad5 > 65] <- 1
esoc$P4_R6[P4_edad6 > 65] <- 1
esoc$P4_R7[P4_edad7 > 65] <- 1
esoc$P4_R8[P4_edad8 > 65] <- 1
esoc$P4_R9[P4_edad9 > 65] <- 1
esoc$P4_R10[P4_edad10 > 65] <- 1
#Suma de las nuevas variables
esoc$NUM65ESOC = rowSums (esoc[,64:73])
detach(esoc)

#Variable HACIN en la ESOC
attach(esoc)
esoc$HACINESOC <- esoc$P2_Npers/esoc$P11_NumHab
detach(esoc)

#Variable RREGTEN en la ESOC
attach(esoc)
esoc$RREGTENESOC[P9_RegTene == "Propiedad"] <- 1
esoc$RREGTENESOC[P9_RegTene == "Alquiler"] <- 2
esoc$RREGTENESOC[P9_RegTene == "Otra forma"] <- NA
esoc$RREGTENESOC[P9_RegTene == "Ns/nc"] <- NA
esoc$RREGTENESOC[P9_RegTene == "Cedida gratis o a bajo precio"] <- 3
detach(esoc)

#Variable RDISPOSIOV en la ESOC
attach(esoc)
esoc$RDISPOSIOVESOC[P12_SegViv == "No"] <- 0
esoc$RDISPOSIOVESOC[P12_SegViv == "Sá"] <- 1
detach(esoc)

#Variable TAMAMU en la ESOC
attach(esoc)
esoc$TamanoESOC[Estrato == "1"] <- "Menos de 10.000 hab"
esoc$TamanoESOC[Estrato == "2"] <- "Menos de 10.000 hab"
esoc$TamanoESOC[Estrato == "3"] <- "Menos de 10.000 hab"
esoc$TamanoESOC[Estrato == "4"] <- "Menos de 10.000 hab"
esoc$TamanoESOC[Estrato == "5"] <- "10.000-20.000 hab"
esoc$TamanoESOC[Estrato == "6"] <- "20.000-50.000 hab"
esoc$TamanoESOC[Estrato == "7"] <- "50.000-100.000 hab"
esoc$TamanoESOC[Estrato == "8"] <- "100.000 hab o mas"
detach(esoc)

#Recodificación de TAMAMU
attach(esoc)
esoc$RTAMAMUESOC[TamanoESOC == "Menos de 10.000 hab"] <- 1
esoc$RTAMAMUESOC[TamanoESOC == "10.000-20.000 hab"] <- 2
esoc$RTAMAMUESOC[TamanoESOC == "20.000-50.000 hab"] <- 3
esoc$RTAMAMUESOC[TamanoESOC == "50.000-100.000 hab"] <- 4
```

```
esoc$RTAMAMUESOC[TamanoESOC == "100.000 hab o mas"] <- 5
detach(esoc)

#Recodificación de SEXO
esoc$SEXOESOC[esoc$P4_sexo1SP == "Mujer"] <- 1
esoc$SEXOESOC[esoc$P4_sexo1SP == "Hombre"] <- 0

#Creación de una variable DUMMY para cada intervalo.
esoc$dum1ESOC <- ifelse (esoc$P5_rentahog == "Hasta 250 euros" | esoc$P5_
rentahog == "De 250,01 a 600 euros", 1,0)
esoc$dum2ESOC <- ifelse (esoc$P5_rentahog == "De 600,01 a 1.000 euros", 1
,0)
esoc$dum3ESOC <- ifelse (esoc$P5_rentahog == "De 1.000,01 a 1.500 euros",
 1,0)
esoc$dum4ESOC <- ifelse (esoc$P5_rentahog == "De 1.500,01 a 2.000 euros",
 1,0)
esoc$dum5ESOC <- ifelse (esoc$P5_rentahog == "De 2.000,01 a 3.000 euros",
 1,0)
esoc$dum6ESOC <- ifelse (esoc$P5_rentahog == "De 3.000,01 a 6.000 euros",
 1,0)
esoc$dum7ESOC <- ifelse (esoc$P5_rentahog == "De 6.000,01 a 12.000 euros"
 | esoc$P5_rentahog == "Más de 12.000 euros", 1,0)

#Creación de número menores de 14 años hogar
esoc$pts2 <- ifelse(esoc$P4_edad2<14,1,0)
esoc$pts3 <- ifelse(esoc$P4_edad3<14,1,0)
esoc$pts4 <- ifelse(esoc$P4_edad4<14,1,0)
esoc$pts5 <- ifelse(esoc$P4_edad5<14,1,0)
esoc$pts6 <- ifelse(esoc$P4_edad6<14,1,0)
esoc$pts7 <- ifelse(esoc$P4_edad7<14,1,0)
esoc$pts8 <- ifelse(esoc$P4_edad8<14,1,0)
esoc$pts9 <- ifelse(esoc$P4_edad9<14,1,0)
esoc$pts10 <- ifelse(esoc$P4_edad10<14,1,0)
#Nulos en ceros
esoc$pts2[is.na(esoc$pts2)] <- 0
esoc$pts3[is.na(esoc$pts3)] <- 0
esoc$pts4[is.na(esoc$pts4)] <- 0
esoc$pts5[is.na(esoc$pts5)] <- 0
esoc$pts6[is.na(esoc$pts6)] <- 0
esoc$pts7[is.na(esoc$pts7)] <- 0
esoc$pts8[is.na(esoc$pts8)] <- 0
esoc$pts9[is.na(esoc$pts9)] <- 0
esoc$pts10[is.na(esoc$pts10)] <- 0
esoc$nninos <- (esoc$pts1+esoc$pts2+esoc$pts3+esoc$pts4+esoc$pts5+esoc$pt
s6+esoc$pts7+esoc$pts8+esoc$pts9+esoc$pts10)


#Creamos unidades de consumo por hogar
esoc$udconsum=1+(0.5*(esoc$P2_Npers-esoc$nninos-1))+(0.3*esoc$nninos)
```

Anexo 2.

```
#Selección automática de variables explicativas con leaps

library(leaps)
regfit= regsubsets(IMPEXAC ~ 0 + dum1 + dum2 + dum3 + dum4 + dum5 + dum6
+ dum7  + as.factor(RREGTEN) + RDISPOSIOV + as.factor(RTAMAMU) + SEXOSPR
+ NUMOCUP  +  as.factor(OCUSP) + NUM65, data = epf,method="forward")

## Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax,
## force.in = force.in, : 1 linear dependencies found

## Reordering variables and trying again:

summary(regfit)

## Subset selection object
## Call: regsubsets.formula(IMPEXAC ~ 0 + dum1 + dum2 + dum3 + dum4 +
##     dum5 + dum6 + dum7 + as.factor(RREGTEN) + RDISPOSIOV + as.factor(R
TAMAMU) +
##     SEXOSPR + NUMOCUP + as.factor(OCUSP) + NUM65, data = epf,
##     method = "forward")
## 18 Variables  (and intercept)
##                       Forced in Forced out
## dum2                      FALSE      FALSE
## dum3                      FALSE      FALSE
## dum4                      FALSE      FALSE
## dum5                      FALSE      FALSE
## dum6                      FALSE      FALSE
## dum7                      FALSE      FALSE
## as.factor(RREGTEN)1       FALSE      FALSE
## as.factor(RREGTEN)2       FALSE      FALSE
## RDISPOSIOV                FALSE      FALSE
## as.factor(RTAMAMU)2       FALSE      FALSE
## as.factor(RTAMAMU)3       FALSE      FALSE
## as.factor(RTAMAMU)4       FALSE      FALSE
## as.factor(RTAMAMU)5       FALSE      FALSE
## SEXOSPR                   FALSE      FALSE
## NUMOCUP                   FALSE      FALSE
## as.factor(OCUSP)2         FALSE      FALSE
## NUM65                     FALSE      FALSE
## as.factor(RREGTEN)3       FALSE      FALSE
## 1 subsets of each size up to 9
## Selection Algorithm: forward
##          dum2 dum3 dum4 dum5 dum6 dum7 as.factor(RREGTEN)1
## 1  ( 1 ) " "  " "  " "  " "  "*"  " "  " "
## 2  ( 1 ) " "  " "  " "  " "  "*"  "*"  " "
## 3  ( 1 ) " "  " "  " "  "*"  "*"  "*"  " "
## 4  ( 1 ) " "  " "  "*"  "*"  "*"  "*"  " "
## 5  ( 1 ) " "  "*"  "*"  "*"  "*"  "*"  " "
```

```
## 6  ( 1 ) " "  "*"  "*"  "*"  "*"  "*"  " "
## 7  ( 1 ) " "  "*"  "*"  "*"  "*"  "*"  " "
## 8  ( 1 ) " "  "*"  "*"  "*"  "*"  "*"  " "
## 9  ( 1 ) " "  "*"  "*"  "*"  "*"  "*"  " "
##          as.factor(RREGTEN)2 as.factor(RREGTEN)3 RDISPOSIOV
## 1  ( 1 ) " "                 " "                 " "
## 2  ( 1 ) " "                 " "                 " "
## 3  ( 1 ) " "                 " "                 " "
## 4  ( 1 ) " "                 " "                 " "
## 5  ( 1 ) " "                 " "                 " "
## 6  ( 1 ) " "                 " "                 " "
## 7  ( 1 ) " "                 " "                 " "
## 8  ( 1 ) " "                 " "                 " "
## 9  ( 1 ) " "                 "*"                 " "
##          as.factor(RTAMAMU)2 as.factor(RTAMAMU)3 as.factor(RTAMAMU)4
## 1  ( 1 ) " "                 " "                 " "
## 2  ( 1 ) " "                 " "                 " "
## 3  ( 1 ) " "                 " "                 " "
## 4  ( 1 ) " "                 " "                 " "
## 5  ( 1 ) " "                 " "                 " "
## 6  ( 1 ) " "                 " "                 " "
## 7  ( 1 ) " "                 " "                 " "
## 8  ( 1 ) " "                 " "                 " "
## 9  ( 1 ) " "                 " "                 " "
##          as.factor(RTAMAMU)5 SEXOSPR NUMOCUP as.factor(OCUSP)2 NUM65
## 1  ( 1 ) " "                 " "     " "     " "                " "
## 2  ( 1 ) " "                 " "     " "     " "                " "
## 3  ( 1 ) " "                 " "     " "     " "                " "
## 4  ( 1 ) " "                 " "     " "     " "                " "
## 5  ( 1 ) " "                 " "     " "     " "                " "
## 6  ( 1 ) " "                 " "     "*"     " "                " "
## 7  ( 1 ) " "                 " "     "*"     " "                "*"
## 8  ( 1 ) " "                 "*"     "*"     " "                "*"
## 9  ( 1 ) " "                 "*"     "*"     " "                "*"

plot(regfit,scale="bic")
```

Anexo 3.

```
#Red Neural con neuralnet
library(MASS)
library(neuralnet)
library(ggplot2)
set.seed(75) #Semilla para aleatoriedad
datos <- epf
n <- nrow(datos)
muestra <- sample(n,n*.70)
train <- datos[muestra,]
```

```r
test <- datos[-muestra,]
#Normalización de variables
datos <- datos[, sapply(datos,is.numeric)]
maxs <- apply(datos,2,max)
mins <- apply(datos,2,min)
datos_nrm <- scale(datos,center=mins, scale=maxs - mins)
datos_nrm <- as.data.frame(datos_nrm)
#Dataframe con las variables que incluiremos en el modelo
datos_nrm2 <- subset(datos_nrm,select=c(IMPEXAC, dum1, dum2, dum3, dum4,
dum5, dum6, dum7, NUMOCUP, NUM65, RREGTEN, SEXOSPR))
train_nrm <- datos_nrm2[muestra,]
test_nrm <- datos_nrm2[-muestra,]
#Fórmula
nms <- names(train_nrm)
frml <- as.formula(paste("IMPEXAC ~", paste(nms[!nms %in% "IMPEXAC"], col
lapse = "+")))
#Modelo
modelo.nn <- neuralnet(IMPEXAC ~ dum1 + dum2 + dum3 + dum4 + dum5 + dum6
+ dum7 + NUMOCUP + NUM65 + RREGTEN + SEXOSPR, data=train_nrm, hidden = c(
6,4), threshold =0.05, algorithm = "rprop+")
#Predicción
pr.nn <- compute(modelo.nn,within(test_nrm,rm(IMPEXAC)))
IMPEXAC.predict <- pr.nn$net.result*(max(datos$IMPEXAC)-min(datos$IMPEXAC
))+min(datos$IMPEXAC)
IMPEXAC.real <- (test_nrm$IMPEXAC)*(max(datos$IMPEXAC)-min(datos$IMPEXAC)
)+min(datos$IMPEXAC)
#Suma de error cuadrático
(se.nn <- sum((IMPEXAC.real-IMPEXAC.predict)^2)/nrow(test_nrm))
```

```
## [1] 209242.8531
```

Anexo 4.

```r
#Modelo de Regresión lineal
mod11<- lm(formula = epf$IMPEXAC ~ 0 + dum1 + dum2 + dum3 + dum4 + dum5 +
 dum6 + dum7 + NUMOCUP + NUM65 + RREGTEN + SEXOSPR, data = epf)
summary(mod11)
```

```
##
## Call:
## lm(formula = epf$IMPEXAC ~ 0 + dum1 + dum2 + dum3 + dum4 + dum5 +
##     dum6 + dum7 + NUMOCUP + NUM65 + RREGTEN + SEXOSPR, data = epf)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1147.7  -171.8    13.8   166.5  3264.7
##
## Coefficients:
##         Estimate Std. Error t value Pr(>|t|)
## dum1     -133.05      75.51  -1.762   0.0785 .
```

```
## dum2      319.02      58.09   5.492 5.44e-08 ***
## dum3      726.05      53.71  13.518  < 2e-16 ***
## dum4     1163.43      58.70  19.819  < 2e-16 ***
## dum5     1826.77      58.86  31.036  < 2e-16 ***
## dum6     3193.10      68.89  46.348  < 2e-16 ***
## dum7     6518.09     163.93  39.761  < 2e-16 ***
## NUMOCUP   146.52      25.40   5.769 1.17e-08 ***
## NUM65     138.92      24.12   5.759 1.24e-08 ***
## RREGTEN   215.13      29.39   7.320 6.42e-13 ***
## SEXOSPR   145.39      31.95   4.551 6.23e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 409 on 750 degrees of freedom
## Multiple R-squared:  0.9686, Adjusted R-squared:  0.9681
## F-statistic:  2101 on 11 and 750 DF,  p-value: < 2.2e-16
```

```
# Modelo Regresión Gamma
attach(epf)
library(Gammareg)
modgamm <- epf$IMPEXAC2 ~ 0 + dum1 + dum2 + dum3 + dum4 + dum5 + dum6 + d
um7 + NUMOCUP + NUM65 + RREGTEN + SEXOSPR
estrucgamm = ~ 0 + dum1 + dum2 + dum3 + dum4 + dum5 + dum6 + dum7 + NUMOC
UP + NUM65 + RREGTEN + SEXOSPR
##Gammareg(modgamm, estrucgamm, meanlink="log")
mod11Gam=glm(modgamm,family=Gamma)
summary(mod11Gam)
```

```
##
## Call:
## glm(formula = modgamm, family = Gamma)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.3393  -0.1110  -0.0250   0.0834   1.4873
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## dum1     2.798e-03  1.325e-04  21.114  < 2e-16 ***
## dum2     9.791e-04  3.843e-05  25.476  < 2e-16 ***
## dum3     5.219e-04  2.226e-05  23.444  < 2e-16 ***
## dum4     2.858e-04  1.935e-05  14.771  < 2e-16 ***
## dum5     1.240e-04  1.582e-05   7.840 1.55e-14 ***
## dum6    -3.686e-05  1.586e-05  -2.325 0.020357 *
## dum7    -1.543e-04  1.982e-05  -7.784 2.32e-14 ***
## NUMOCUP  2.157e-05  6.465e-06   3.336 0.000891 ***
## NUM65    2.613e-05  6.777e-06   3.855 0.000126 ***
## RREGTEN  2.007e-04  1.196e-05  16.781  < 2e-16 ***
## SEXOSPR  4.546e-05  9.266e-06   4.906 1.14e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for Gamma family taken to be 0.08137551)
##
##     Null deviance:    NaN  on 761  degrees of freedom
## Residual deviance: 237.15  on 750  degrees of freedom
## AIC: 12384
##
## Number of Fisher Scoring iterations: 6
```

```
# Modelo Regresión Logarítmica
mod11Log <- lm(log(IMPEXAC2)~ 0 +dum1 + dum2 + dum3 + dum4 + dum5 + dum6
+ dum7 + NUMOCUP + NUM65 + RREGTEN + SEXOSPR, data=epf)
summary(mod11Log)
```

```
##
## Call:
## lm(formula = log(IMPEXAC2) ~ 0 + dum1 + dum2 + dum3 + dum4 +
##     dum5 + dum6 + dum7 + NUMOCUP + NUM65 + RREGTEN + SEXOSPR,
##     data = epf)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.8716 -0.2326  0.1039  0.5018  6.4995
##
## Coefficients:
##         Estimate Std. Error t value Pr(>|t|)
## dum1     1.65977    0.29946   5.543 4.13e-08 ***
## dum2     4.63377    0.23037  20.114  < 2e-16 ***
## dum3     4.96098    0.21300  23.291  < 2e-16 ***
## dum4     5.10781    0.23281  21.940  < 2e-16 ***
## dum5     5.32192    0.23343  22.799  < 2e-16 ***
## dum6     5.54337    0.27323  20.289  < 2e-16 ***
## dum7     6.20001    0.65013   9.537  < 2e-16 ***
## NUMOCUP  0.53571    0.10073   5.318 1.39e-07 ***
## NUM65    0.52913    0.09567   5.531 4.41e-08 ***
## RREGTEN  1.04993    0.11656   9.007  < 2e-16 ***
## SEXOSPR  0.45690    0.12670   3.606 0.000331 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.622 on 750 degrees of freedom
## Multiple R-squared:  0.9528, Adjusted R-squared:  0.9521
## F-statistic:  1376 on 11 and 750 DF,  p-value: < 2.2e-16
```

Anexo 5.

```
# Cálculo de las tasas de aciertos

#Vamos a crear un dataframe con las predicciones y los ingresos por inter
valos, con la que evaluar el grado de acierto (de coincidencia dentro del
```

```
 intervalo).
IMPEX <- na.omit(data.frame esoc$P5_rentahog))

IMPEX <- cbind(IMPEX,IMPEXACdata)
IMPEX$IMPEXmod11_IN <- cut(IMPEX$IMPEXmod11, c(0, 250, 600, 1000, 1500, 2
000, 3000, 6000, 12000, 1000000, 2000000), labels=c('Hasta 250 euros', 'D
e 250,01 a 600 euros', 'De 600,01 a 1.000 euros', 'De 1.000,01 a 1.500 eu
ros', 'De 1.500,01 a 2.000 euros', 'De 2.000,01 a 3.000 euros', 'De 3.000
,01 a 6.000 euros', 'De 6.000,01 a 12.000 euros', 'Más de 12.000 euros',
'No sabe/No contesta'))
IMPEX$IMPEXmod11Gam_IN <- cut(IMPEX$IMPEXmod11Gam, c(0, 250, 600, 1000, 1
500, 2000, 3000, 6000, 12000, 1000000, 2000000), labels=c('Hasta 250 euro
s', 'De 250,01 a 600 euros', 'De 600,01 a 1.000 euros', 'De 1.000,01 a 1.
500 euros', 'De 1.500,01 a 2.000 euros', 'De 2.000,01 a 3.000 euros', 'De
 3.000,01 a 6.000 euros', 'De 6.000,01 a 12.000 euros', 'Más de 12.000 eu
ros', 'No sabe/No contesta'))
IMPEX$IMPEXmod11Log_IN <- cut(IMPEX$IMPEXmod11Log, c(0, 250, 600, 1000, 1
500, 2000, 3000, 6000, 12000, 1000000, 2000000), labels=c('Hasta 250 euro
s', 'De 250,01 a 600 euros', 'De 600,01 a 1.000 euros', 'De 1.000,01 a 1.
500 euros', 'De 1.500,01 a 2.000 euros', 'De 2.000,01 a 3.000 euros', 'De
 3.000,01 a 6.000 euros', 'De 6.000,01 a 12.000 euros', 'Más de 12.000 eu
ros', 'No sabe/No contesta'))

varaci <- IMPEX$IMPEXmod11_IN #Modelo lineal
varaci=na.omit(ifelse(varaci==IMPEX$esoc.P5_rentahog,1,0))
y=rep(1,length(varaci))
aciertos=data.frame(varaci,y)
poraciertosLog <- sum(aciertos$varaci)/sum(aciertos$y)
poraciertosLog

## [1] 0.8754254595

varaci <- IMPEX$IMPEXmod11Gam_IN #Modelo Gamma
varaci=na.omit(ifelse(varaci==IMPEX$esoc.P5_rentahog,1,0))
y=rep(1,length(varaci))
aciertos=data.frame(varaci,y)
poraciertosGam <- sum(aciertos$varaci)/sum(aciertos$y)
poraciertosGam

## [1] 0.9046970728

varaci <- IMPEX$IMPEXmod11Log_IN #Modelo logarítmico
varaci=na.omit(ifelse(varaci==IMPEX$esoc.P5_rentahog,1,0))
y=rep(1,length(varaci))
aciertos=data.frame(varaci,y)
poraciertosLog <- sum(aciertos$varaci)/sum(aciertos$y)
poraciertosLog

## [1] 0.8189542484

#Aciertos neuralnet
IMPEXNN <- (data.frame(esoc$P5_rentahog))
```

```
IMPEXNN <- cbind(IMPEXNN,IMPEXACdatann)
IMPEXNN$IMPEXmodnn <- cut(IMPEXNN$IMPEXmodnn, c(0, 250, 600, 1000, 1500,
2000, 3000, 6000, 12000, 1000000, 2000000), labels=c('Hasta 250 euros', '
De 250,01 a 600 euros', 'De 600,01 a 1.000 euros', 'De 1.000,01 a 1.500 e
uros', 'De 1.500,01 a 2.000 euros', 'De 2.000,01 a 3.000 euros', 'De 3.00
0,01 a 6.000 euros', 'De 6.000,01 a 12.000 euros', 'Más de 12.000 euros',
 'No sabe/No contesta'))
varacinn <- IMPEXNN$IMPEXmodnn #Seleccionar modelo a testar % aciertos
varacinn=na.omit(ifelse(varacinn==IMPEXNN$esoc.P5_rentahog,1,0))
ynn=rep(1,length(varacinn))
aciertosnn=data.frame(varacinn,ynn)
poraciertosnn <- sum(aciertosnn$varacinn)/sum(aciertosnn$ynn)
poraciertosnn
```

```
## [1] 0.7904642409
```

```
#Aciertos árbol
IMPEXNN <- (data.frame(esoc$P5_rentahog))
IMPEXNN <- cbind(IMPEXNN,IMPEXACdatann)
IMPEXNN$IMPEXarb <- cut(IMPEXNN$IMPEXarb, c(0, 250, 600, 1000, 1500, 2000
, 3000, 6000, 12000, 1000000, 2000000), labels=c('Hasta 250 euros', 'De 2
50,01 a 600 euros', 'De 600,01 a 1.000 euros', 'De 1.000,01 a 1.500 euros
', 'De 1.500,01 a 2.000 euros', 'De 2.000,01 a 3.000 euros', 'De 3.000,01
 a 6.000 euros', 'De 6.000,01 a 12.000 euros', 'Más de 12.000 euros', 'No
 sabe/No contesta'))
varaciarb <- IMPEXNN$IMPEXarb #Seleccionar modelo a testar % aciertos
varaciarb=na.omit(ifelse(varaciarb==IMPEXNN$esoc.P5_rentahog,1,0))
yarb=rep(1,length(varaciarb))
aciertosarb=data.frame(varaciarb,yarb)
poraciertosarb <- sum(aciertosarb$varaciarb)/sum(aciertosarb$yarb)
poraciertosarb
```

```
## [1] 0.9414694894
```

Anexo 6.

```
#Creación base de datos para el cálculo de la tasa de riesgo de pobreza
viv=rep(1,length(IMPEXmod11))
viv2=rep(1,length(IMPEXmodnn))
IMPEXACdata <- data.frame(IMPEXmod11,IMPEXmod11lin,IMPEXmod11Rob,IMPEXmod
10Rob,IMPEXmod11Gam,IMPEXmod11Log,as.integer(viv))
IMPEXACdatann <- data.frame(IMPEXmodnn,as.integer(viv2))
#Se generan dos bases de datos (regresiones y neuralnet).

#Tasas regresiones
reg <- IMPEXACdata$IMPEXmod11/nuevos$UDCONSUM
tmen <- median(reg)*0.6
tmen
```

```
## [1] 491.0813982
```

```
tp=ifelse(reg<=tmen,1,0)
tu=rep(1,length(tp))
datos=data.frame(tp,tu)
tpob <- sum(datos$tp)/sum(datos$tu)
tpob
```

```
## [1] 0.1928104575
```

```
#Boostrap para calcular IC
library(boot)
x=ifelse(reg<=tmen,1,0)
u=rep(1,length(x))
datos=data.frame(x,u)
porcentaje <- sum(datos$x)/sum(datos$u)
porcentaje
```

```
## [1] 0.1928104575
```

```
porcentaje.fun <- function(data,w) {sum(data$x*w)/sum(data$u*w)}
prcent.boot <- boot(datos,porcentaje.fun,R=1000,stype="w",sim="ordinary")
boot.ci(prcent.boot,conf=c(0.90,0.95))
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
```

```
CALL :
```

```
boot.ci(boot.out = prcent.boot, conf = c(0.9, 0.95))
```

```
Intervals :
```

```
Level      Normal              Basic
90%   ( 0.1769,  0.2095 )   ( 0.1771,  0.2091 )
95%   ( 0.1738,  0.2126 )   ( 0.1739,  0.2118 )

Level     Percentile            BCa
90%   ( 0.1765,  0.2085 )   ( 0.1771,  0.2093 )
95%   ( 0.1739,  0.2118 )   ( 0.1745,  0.2118 )

Calculations and Intervals on Original Scale
```