PAPELES DE TRABAJO

8/2019

¿PSM o imputación?: herramientas para la fusión estadística de las encuestas ECV y EPF

NURIA BADENES PLÁ

BORJA GAMBAU-SUELVES

Instituto de Estudios Fiscales

Avda. Cardenal Herrera Oria, 378, 28035-Madrid

I.S.S.N.: 1578-0252

N.I.P.O.: 188-19-032-6

ÍNDICE

Resumen

En este trabajo se realiza una fusión estadística entre las encuestas ECV y EPF para el año 2016. La motivación reside en la obtención de una base de microdatos que contenga simultáneamente información sobre consumo y renta de las familias y su posterior explotación para el análisis de la fiscalidad directa e indirecta. La fusión se realiza asignando los valores de renta a la EPF mediante tres metodologías: Regresión Simple, Imputación Múltiple, y *Propensity Score Matching* (PSM). Una vez obtenidos los resultados, se comparan atendiendo a los valores medios, a los ajustes de los histogramas, y a las diferencias en las distribuciones utilizando herramientas tradicionales del análisis de la desigualdad. Si la bondad del ajuste tiene en cuenta exclusivamente la distribución de la propia renta, PSM se revela como el mejor método de fusión. Al analizar las distribuciones por subgrupos poblacionales, la dominancia de este método no se mantiene en todo caso.

Palabras clave: fusión estadística, *Propensity Score Matching* (PSM), imputación múltiple, análisis de consumo y renta.

Clasificación JEL: C16, C53, D12.

1. INTRODUCCIÓN

La motivación de este trabajo se origina en la inexistencia de una base de microdatos que contenga simultáneamente información sobre el consumo y sobre la renta de las familias. El principal interés de los autores de contar con esta información unida reside –al margen de otras explotaciones– en la posibilidad que ello brinda para el análisis conjunto de la fiscalidad sobre la renta y sobre el consumo, que suponen las principales cargas contributivas de las familias, así como el efecto de las transferencias monetarias que las familias reciben.

Con el objetivo de cubrir la motivación descrita, este documento de trabajo trata de desarrollar una metodología concreta para la unión de dos fuentes estadísticas fiables que nos suministren tal información: la Encuesta de Condiciones de Vida (ECV) y la Encuesta de Presupuestos Familiares (EPF). Si bien es verdad que la EPF dispone de las dos fuentes de información de utilidad (renta y consumo), la EPF tiende a infraestimar los valores de la renta respecto a los que reporta la ECV, ampliamente aceptados por la comunidad investigadora. Así, el deseo y la disponibilidad –vía EPF– de tener una base de datos que contenga datos de renta y consumo para el análisis de los efectos de la imposición directa e indirecta, no son suficientes ya que previamente se debe resolver la cuestión metodológica que plantea la obtención de una base de datos común para el análisis.

La fusión estadística, o *statistical matching* en inglés, es un área de investigación relativamente nueva que ha recibido mucha atención en los últimos años en respuesta al flujo de datos disponible en la actualidad. Sus orígenes se remontan a mediados de los sesenta cuando Okner (1972) consiguió una base de datos condensada de variables sociodemográficas e impositivas de los hogares fusionando o *matcheando* el *Tax File* de 1966 y la *Survey of Economic Opportunities* de 1967 para el caso de Estados Unidos. A partir de allí el interés sobre la materia empezó a crecer continuadamente hasta el día de hoy. Tanto es así que EUROSTAT (2013) publicó dos documentos de trabajo con estudios de caso y cuestiones metodológicas para proceder correctamente a la hora de fusionar estadísticamente fuentes de información diversas.

Como exponen en su trabajo Taylor *et al*. (2001), Decoster y Van Camp (2002) y Decoster *et al.* (2007, 2013), los métodos y aproximaciones empíricas para proceder a la fusión estadística difieren radicalmente según el objetivo del estudio y las características intrínsecas de las bases de datos originales. Sin embargo estas técnicas se pueden clasificar en dos grandes grupos: los métodos explícitos, basados en modelos de regresión paramétricos (modelos econométricos) o no paramétricos (funciones de densidad), y los métodos implícitos, en los que cada observación de una fuente de datos es *matcheada* con la información de la variable de interés de la otra fuente. Cada uno de los métodos plausibles dentro de estas categorías tiene características deseables desde el punto de vista matemático-estadístico pero generan resultados muy diversos en términos de validación del proceso de fusión condicionando, así mismo, la veracidad tanto interna como externa de la posterior explotación de la fuente de datos fusionada.

Algún ejemplo interesante de la modelización del impacto de impuestos directos e indirectos como principal explotación de la fusión estadística entre la ECV y la EPF, así como las implicaciones del uso de los diferentes métodos usados en la unión, se pueden encontrar en Savage y Ca-

Ilan (2015) para el caso de Reino Unido, en Decoster *et al.* (2013) para el caso de Alemania, o Decoster *et al.* (2014) para el caso de Bélgica.

En el caso de España, la fusión estadística entre la ECV y la EPF así como su utilidad para la investigación de los efectos de la tributación directa e indirecta sobre los hogares españoles ha sido ampliamente tratada por López Laborda *et al*. (2016, 2017). Estos trabajos, basados en métodos explícitos paramétricos y en la aplicación de las curvas de Engel, imputan las variables relacionadas con el gasto, provenientes de la EPF, en la fuente de datos original de renta, la ECV, y con ello logran estimar los impuestos pagados por los hogares españoles en 2013.

A diferencia de los trabajos anteriores, este documento trata de mostrar los resultados de la aplicación de diferentes herramientas de fusión estadística así como una propuesta de análisis para la validación de los resultados obtenidos en la fusión. En concreto, y como nueva aportación al debate metodológico, se utiliza la técnica del *Propensity Score Matching* (PSM) como una técnica mixta para la fusión de encuestas, al usar de manera combinada los métodos explícitos basados en modelos de regresión y los métodos implícitos de asignación de valores en función de un criterio de decisión. La ventaja del PSM radica en que queda fuera del análisis el criterio de decisión subjetivo del investigador ya que el método, basado en la estimación de una probabilidad a través de modelos *logit* o *probit*, resume toda la información común de las fuentes de información en un solo valor; la probabilidad de pertenecer a una encuesta o a otra en función de las características observables del conjunto muestral. Esta probabilidad o *score,* estimada para todos los hogares de las dos encuestas, es el criterio objetivo que se utiliza para emparejar, de tal manera que si dos unidades tienen una puntuación *(score)* muy similar, perteneciendo a dos encuestas diferentes, las características no observables, en este caso la renta para la EPF o el consumo para la ECV, también deberían de serlo pudiéndose asignar el valor de una en otra o viceversa. Al ser la utilización de esta metodología una novedad en el análisis empírico español, la comprobación de los resultados incluirá siempre PSM en comparación con otras metodologías más extendidas en la fusión estadística como la regresión simple o la imputación múltiple.

Por otro lado, en este trabajo, se realiza el emparejamiento considerando la EPF como encuesta receptora, a la que se le asigna la información faltante relativa a la renta.

Finalmente, y como novedad, además de comprobar la bondad del ajuste a través de los valores medios de la renta asignada por subgrupos y el análisis de los histogramas, se añade como herramienta de validación la utilización del análisis distributivo tradicional mediante la comprobación de las diferencias entre las distribuciones de renta y el error cometido mediante las distintas asignaciones con respecto de la renta real. Para ello, se comprueban las diferencias en las Curvas de Lorenz y se obtienen los índices de Reynolds-Smolensky cuyo valor puede descomponerse en el efecto de la diferencia de medias y diferencia por los errores de asignación.

El texto se organiza como sigue: tras esta introducción, el apartado dos realiza una descripción estadística de las bases de datos que se van a unir. En el apartado tres, se explican las herramientas utilizadas para la fusión estadística. El cuarto apartado, se dedica a la comparación de los resultados considerando tanto la distribución global como por subgrupos. En el quinto apartado, se propone la utilización del índice de Reynolds-Smolensky como elemento de comproba-

ción adicional de la bondad de ajuste. El sexto apartado concluye. El apéndice uno se destina a la comprobación de las diferencias de medias subgrupales de las distintas metodologías y el dos, a mostrar un ejemplo aclaratorio del apartado cinco.

2. CONOCIENDO EL CONTEXTO DE PARTIDA: INICIO DEL PROCESO Y DESCRIPCIÓN ESTADÍSTICA DE LAS BASES DE DATOS ECV Y EPF

La unión de los datos que explicamos en el presente documento toma como referencia la información más reciente disponible en el momento actual tanto de Encuesta de Presupuestos Familiares (EPF) y la Encuesta de Condiciones de Vida (ECV). Ambas son del año 2016, si bien la ECV contiene información del año previo a la publicación de la encuesta.

Existen varios escollos que es necesario superar para realizar una buena unión de los datos. La ECV cuenta con muy buena información sobre la renta de las familias, pero no contiene datos sobre el consumo. Por su parte, la EPF detalla mucho el consumo de las familias en diferentes bienes y servicios, pero la información relativa a la renta es muy incompleta, poco detallada, y faltante para muchas familias. Para unir las dos encuestas, se puede optar por dos alternativas:

a) Tomar como referencia la EPF y a esta encuesta asignarle la renta de la ECV.

b) Tomar como referencia la ECV y a esta encuesta asignarle el consumo de la EPF.

Se ha optado por la primera alternativa por una razón práctica: la información de la cesta de consumo incluye muchas variables, mientras que tomar como referencia la EPF implica trasladar una variable, la renta, o unas pocas si se quiere unir además de la renta el IRPF pagado por las familias y sus cotizaciones, lo cual facilita el trabajo. También podría optarse por liquidar el IVA y los impuestos especiales y trasladar únicamente estas dos variables a la ECV desde la EPF. Dependiendo del uso que se quiera dar a la base de datos unida, se puede optar por una u otra alternativa, pero en principio explicamos los pasos seguidos para "rellenar" los datos de la renta en la EPF.

Para poder emparejar ambas encuestas de manera adecuada, sea cual sea el mecanismo utilizado, es menester contar con información común en ambas encuestas que nos permitan identificar las observaciones que más se parecen. La búsqueda de variables comunes y con poder explicativo es por tanto el primer paso, y para poder utilizar la información común, es preciso codificarla de manera uniforme en ambas encuestas.

Las variables comunes halladas para el análisis en la ECV (13.791 observaciones) y la EPF (21.280 observaciones) y la distribución porcentual por categorías en cada una de las encuestas son los siguientes:

— Sexo del sustentador principal de la familia.

Sexo SP	ECV	EPF
Hombre	62,2	67,5
Mujer	37,8	32,5

La interpretación de las tablas se produce siempre de la misma manera. Los valores de la columna bajo la rúbrica ECV representan qué porcentaje de observaciones se sitúa en cada categoría. Así, el 62,2% de los sustentadores principales de la ECV son hombres y el 37,8% son mujeres, sumando 100% en vertical. En la EPF, el 67,5% de los sustentadores principales son hombres y el 32,5% son mujeres.

— Estado civil del sustentador principal.

Estado civil SP	ECV	EPF
Soltero	21,8	18,2
Casado	55,4	60,5
Separado	2,7	2,9
Viudo	13,8	11,9
Divorciado	6,3	6,6

— País de nacimiento del sustentador principal.

País nacimiento SP	ECV	EPF
España	91,7	91,5
Extranjero UE	2,3	2,5
Extranjero no UE	6,0	6,0

— Estudios del sustentador principal.

Estudios SP	ECV	EPF
Primaria o menos	29,0	20,6
1.ª etapa secundaria	23,3	30,8
2.ª etapa secundaria	19,1	18,0
Superior	28,6	30,6

— Situación laboral del sustentador principal.

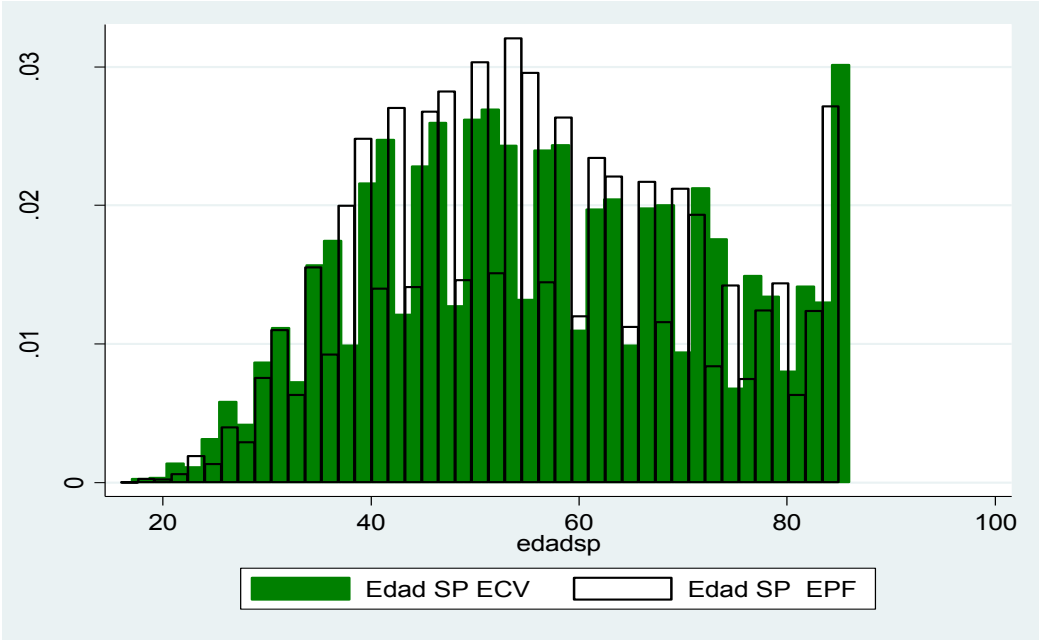
Situación laboral SP	ECV	EPF
Trabaja	53,4	54,3
Inactivo	1,2	2,8
Parado	7,6	6,7
Jubilado	28,7	29,9
Estudiante	0,3	0,1
Amo de casa	5,8	4,8
Incapacidad	3,1	1,4

— Edad del sustentador principal.

Edad SP	Obs.	Media	Desv.	Mín.	Máx.
ECV	13.791	56,6	16,2	17	86
EPF	21.180	55,9	15,1	16	85

Al tratarse la edad de una variable continua y no categórica, aportamos los valores de la edad media del sustentador principal (56,6 años en la ECV y 55,9 en la EPF), así como los valores de las personas más jóvenes y de más edad (17 y 86 en la ECV y 16 y 85 en la EPF). También se muestra el número de observaciones y la desviación típica.

Los histogramas de las edades del sustentador principal superpuestos, permiten comprobar el grado de coincidencia en las edades en ambas encuestas:



— Tipo de edificación en la que residen los miembros del hogar.

Tipo de vivienda	ECV	EPF
Unifamiliar independiente	55,4	52,0
Unifamiliar adosada	25,1	28,6
Edificio <10 viviendas	11,5	12,8
Edificio >=10 viviendas	2,3	1,0
Otros	5,7	5,5

— Régimen de tenencia de la vivienda del hogar.

Régimen tenencia vivienda	ECV	EPF
Propiedad sin hipoteca	55,4	52,0
Propiedad con hipoteca	25,1	28,6
Alquiler mercado	11,5	12,9
Alquiler inferior mercado	2,3	1,0
Cesión	5,7	5,5

— Número de habitaciones de la vivienda.

N.º habitaciones vivienda	ECV	EPF
Una	0,4	0,5
Dos	2,0	1,3
Tres	9,3	5,2
Cuatro	25,8	17,6
Cinco	37,0	43,6
Seis o más	25,5	31,9

— Número de miembros del hogar.

N.º de miembros hogar	ECV	EPF
1	23,0	18,7
2	32,0	32,1
3	21,9	22,9
4	17,6	20,2
5	4,1	4,5
6	1,0	1,0
7	0,3	0,4
8	0,1	0,1
9 o más	0,1	0,1

— Tipo de hogar.

Tipo de hogar	ECV	EPF
Adulto solo >65 años	11,8	8,8
Persona sola 30 a 64 años	10,5	9,4
Persona sola <30 años	0,6	0,5
Adulto solo y algún <16 años	3,4	9,2
Pareja sin hijos, alguno >=65	16,8	12,8
Pareja sin hijos ambos <65	13,2	11,3
Pareja 1 hijo <16	10,8	8,0
Pareja 2 hijos <16	11,3	8,9
Pareja 3 o más hijos <16	2,0	1,5
Otros	19,6	29,5

— Comunidad Autónoma.

CCAA	ECV	EPF
Andalucía	10,3	11,0
Aragón	3,6	4,4
Asturias	4,1	4,1
Baleares	2,8	3,6
Canarias	3,1	4,2
Cantabria	2,7	3,5
Castilla y León	6,2	6,8

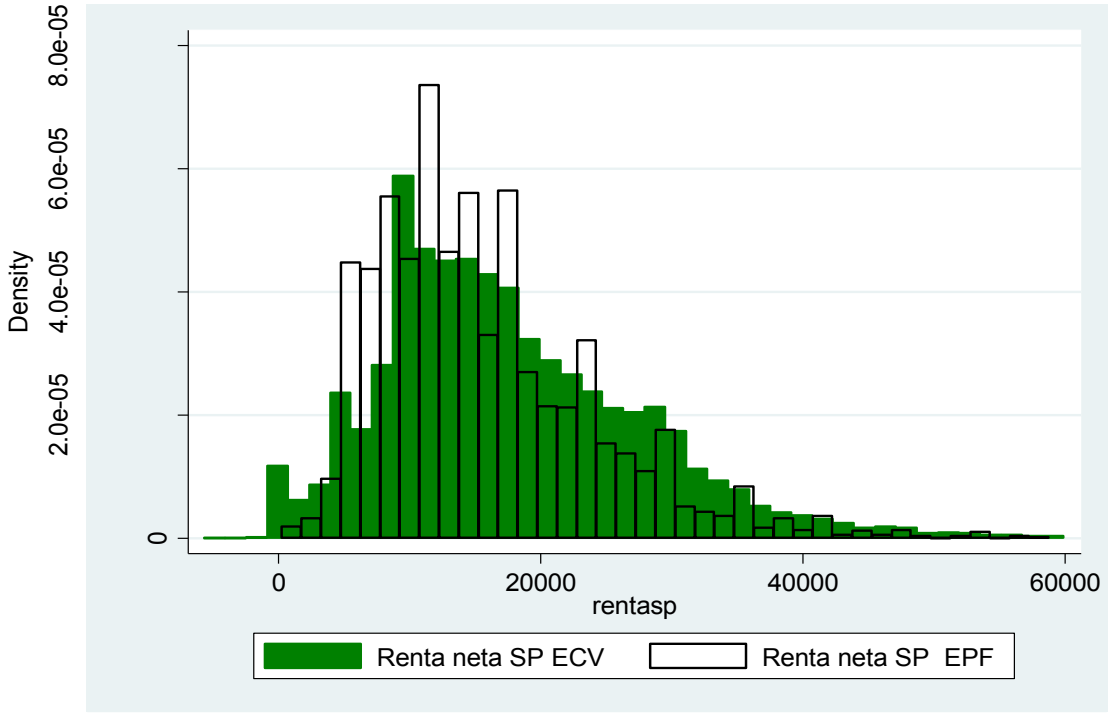
CCAA	ECV	EPF
Castilla-La Mancha	4,1	5,5
Cataluña	20,9	9,3
Comunidad Valenciana	6,6	7,8
Extremadura	3,9	4,5
Galicia	6,0	6,4
Madrid	9,9	7,3
Murcia	3,7	4,1
Navarra	3,1	3,5
País Vasco	4,9	9,7
Rioja	2,6	3,3
Ceuta	0,6	0,6
Melilla	0,8	0,6

— Tramo de renta neta mensual del sustentador principal.

Tramo mensual de renta (€) del SP	ECV	EPF
Menos de 500	9,0	6,2
Entre 500 y 1.000	23,4	28,2
Entre 1.000 y 1.500	26,4	30,8
Entre 1.500 y 2.000	17,2	17,1
Entre 2.000 y 2.500	12,7	10,2
Entre 2.500 y 3.000	6,5	4,0
Más de 3.000	4,7	3,5

— Renta neta anual del sustentador principal:

Mostramos el histograma de las rentas netas anuales del sustentador principal en la ECV y en la EPF, considerando hasta 60.000 €.



Nótese que el número de hogares en los que se cuenta con información de la renta del sustentador principal en la EPF es solamente de 9.197, frente a las más de 21.000 observaciones con las que se cuenta en relación a las demás variables comunes con la ECV.

— Por percentiles de renta, los valores son los siguientes:

Renta anual SP	ECV	EPF
Percentiles		
1%	0	3.600
5%	4.370	5.112
10%	6.620	6.960
25%	10.379	9.492
50%	15.899	14.040
75%	23.523	20.400
90%	30.736	27.600
95%	35.585	32.004
99%	48.304	44.796
N	13.791	9.042
Media	17.654	15.878
Desv	10.117	9.197

3. LA FUSIÓN ESTADÍSTICA: HERRAMIENTAS PARA LA UNIÓN DE LOS DATOS

Existen múltiples formas alternativas de unir la información referente a consumo y renta en las familias.

Se han utilizado tres mecanismos para realizar la unión con el fin de sopesar a partir de los resultados finales el método más conveniente en función de la facilidad de cálculo en la construcción de la base unida y su explotación posterior y el máximo acercamiento en la renta emparejada en comparación con la que se conoce en la ECV. Estos métodos son:

1. Imputación puntual.
2. Imputación múltiple.
3. *Propensity Score Matching* (PSM).

La estrategia seguida es aplicar distintos métodos para predecir la renta que le corresponde al hogar en la ECV en función de una serie de características (presentes tanto en la ECV como en la EPF). Como la renta en la ECV es conocida, se puede comprobar cuál de los métodos conduciría a un mejor ajuste, y éste será el utilizado para asignar la renta en la EPF, donde esta variable es desconocida y no permite comprobación.

3.1. Imputación puntual

Mediante este mecanismo, la variable desconocida en la EPF (rentatotal) y conocida en la ECV es utilizada como variable dependiente en una regresión en la que se utilizan como variables inde-

pendientes las comunes a la ECV y EPF con poder explicativo. Una vez construido un modelo que ajuste la explicación de esta renta en la ECV, se utilizan los parámetros estimados para predecir el valor de la renta en la EPF.

Todas las variables explicativas son categóricas excepto la renta del sustentador principal y su edad (única variable no significativa). Se construye una variable "rentaspi" (renta imputada del sustentador principal) para poder utilizarla en la predicción de renta total del hogar. Se ha constatado que la renta del sustentador principal (también el tramo de renta al que pertenece) tiene un gran poder explicativo sobre la renta del hogar, pero precisamente esta información no aparece en muchas de las observaciones de la EPF, por ello se imputa en los casos que es desconocida para no prescindir de su capacidad explicativa.

Source	SS	df	MS
Model	4.4450e+12	11	4.0409e+11
Residual	1.1782e+12	13665	86216765.2
Total	5.6232e+12	13676	411172220

Number of obs=13677; F(11, 13665)=4686.96; Prob>F=0.0000; R-squared=0.7905;
Adj R-squared=0.7903; Root MSE=9285.3

rentatotal	Coef.	Std. Err.	t	P>|t|	[95% Conf. Interval]	
rentaspi	1.498247	.0091962	162.92	0.000	1.480221	1.516273
sexosp	2806.277	177.8951	15.77	0.000	2457.578	3154.976
ecivilsp	-407.566	80.01113	-5.09	0.000	-564.3988	-250.7332
paisnacsp	-1289.271	169.9842	-7.58	0.000	-1622.463	-956.0785
estudsp	1763.292	85.74825	20.56	0.000	1595.213	1931.37
edadsp	.5467021	6.631366	0.08	0.934	-12.45169	13.54509
tipoedif	-212.334	77.62701	-2.74	0.006	-364.4937	-60.1744
regten	-534.6256	79.505	-6.72	0.000	-690.4663	-378.7848
nhabit	549.2385	84.26616	6.52	0.000	384.0653	714.4118
hogar	1795.542	31.15467	57.63	0.000	1734.474	1856.609
ccaa	87.62893	17.48134	5.01	0.000	53.36309	121.8948
cons	-13994.78	812.0399	17.23	0.000	-15586.49	-12403.07

Las 13.791 observaciones de la ECV se reducen a 13.677 después de realizar las imputaciones de la renta del sustentador principal.

El modelo para realizar las imputaciones de la renta del sustentador principal es:

Glm rentasp sexosp ecivilsp paisnacsp estudsp situacsp tipoedif edadsp nhabit nmiemb hogar ccaa tramo

Iteration 0: log likelihood = -213384.35;

Generalized linear models No. of obs = 22726

Optimization : ML Residual df = 22713

Scale parameter = 8381769

Deviance = 1.90375e+11 (1/df) Deviance = 8381769
Pearson = 1.90375e+11 (1/df) Pearson = 8381769
Variancefunction: V(u) = 1 [Gaussian]
Link function : g(u) = u [Identity]
AIC = 18.78002
Log likelihood = -213384.3504 BIC = 1.90e+11

rentasp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
sexosp	-200.7591	43.87078	-4.58	0.000	-286.7442	-114.7739
ecivilsp	44.6583	19.40283	2.30	0.021	6.629459	82.68715
paisnacsp	-54.22954	40.28828	-1.35	0.178	-133.1931	24.73404
estudsp	136.0119	21.34758	6.37	0.000	94.17142	177.8524
situacsp	12.66894	15.23096	0.83	0.406	-17.18319	42.52107
tipoedif	-43.12901	18.97754	2.27	0.023	-80.32431	-5.933713
edadsp	8.189516	1.824466	4.49	0.000	4.613629	11.7654
nhabit	19.24182	20.71181	0.93	0.353	-21.35259	59.83623
nmiemb	32.54633	29.56697	1.10	0.271	-25.40388	90.49654
hogar	-14.62508	12.26387	1.19	0.233	-38.66182	9.411659
ccaa	3.380156	3.989255	0.85	0.397	-4.43864	11.19895
tramo	6128.571	15.14486	406.66	0.000	6098.887	6158.254
cons	-3872.912	186.9977	-20.71	0.000	-4239.42	-3506.403

3.2. Imputación múltiple

El método de imputación múltiple (MI) presenta mejoras respecto a la imputación simple. En el caso de imputación simple, no se refleja la incertidumbre de los valores desconocidos en las predicciones, por lo que las varianzas de los parámetros estimados están sesgadas hacia cero. En lugar de rellenar los datos *missing* con un único valor, el procedimiento MI (Rubin 1987) reemplaza cada valor *missing* por un conjunto de valores plausibles que representan la incertidumbre sobre el valor más adecuado para ser imputado. Los conjuntos de datos múltiples se analizan utilizando procedimientos estándar para completar datos y se combinan los resultados. El proceso para combinar los distintos conjuntos es esencialmente el mismo, independientemente de qué sistema para completar datos se utilice. El método MI no trata de estimar los valores faltantes con valores simulados sino conseguir una muestra aleatoria de los valores *missing*. Este procedimiento conduce a inferencias estadísticas válidas que reflejan fielmente la incertidumbre debida a los valores *missing* ofreciendo intervalos de confianza para los parámetros.

El método de Rubin para MI se desarrolla en varios pasos:

1) El primer paso implica la utilización de un método adecuado de imputación de los valores *missing* que incorpore variación aleatoria.

2) El segundo paso consiste en repetir el procedimiento anterior *m* veces.

3) En el tercer paso se lleva a cabo un análisis sobre cada uno de los m conjuntos de datos utilizando métodos estándar.

4) En cuarto lugar se calcula la media de los parámetros estimados a lo largo de los conjuntos de valores *missing* para obtener una estimación puntual.

5) Finalmente, se calculan los errores estándar como media de los errores estándar cuadráticos de las estimaciones de los valores *missing*. Después es posible calcular la varianza de los parámetros de los valores *missing* a lo largo de las muestras.

Este método presenta varias características deseables. Por un lado es posible obtener estimaciones insesgadas de todos los parámetros a partir de los errores estándar, lo cual no puede lograrse con imputación determinista. Es posible además obtener buenas estimaciones de los errores estándar. Sin embargo, hay ciertas condiciones que deben mantenerse antes de utilizar este procedimiento, como el hecho de que los valores *missing* deben aparecer de forma aleatoria, y no ligados de manera sistemática a una característica, y que el método de imputación utilizado sea el apropiado.

El mecanismo de imputación múltiple se basa entonces en la misma idea que la imputación puntual, pero la estimación de los parámetros se realiza de forma repetida mediante un mecanismo de *bootstrap*, lo que permite contar con distintos conjuntos de datos sobre los que realizar la estimación. En nuestro trabajo además se ha utilizado un proceso doble de imputación. En primer lugar para obtener una predicción de la renta del sustentador principal, que cuenta con gran poder explicativo sobre la renta de las familias y se encuentra *misssing* en muchas observaciones, especialmente en la EPF. Una vez imputado este valor, se procede a una segunda imputación múltiple de la renta final, que es la variable final de interés.

El modelo utilizado para la imputación múltiple es:

mi impute pmm rentatotal2 yhatsp sexosp ecivilsp paisnacsp edadsp estudsp situacsp tipoedif regten nhabit nmiemb hogar ccaa tramo, forceknn(2) add(20)

note: tramo omittedbecause of collinearity

Univariate imputation Imputations = 20; Predictive mean matching added = 20; Imputed: m=1 through m=20 updated = 0; Nearest neighbors = 2

Observations per m				
Variable	**Complete**	**Incomplete**	**Imputed**	**Total**
Rentatotal2	13791	22011	21196	35802

(complete + incomplete = total; impute disthe minimum across m of the number of filled-in observations.)

glm rentatotal2 yhat spsexosp ecivilsp paisnacsp edadsp estudsps ituacsp tipoedif regten nhabit nmiemb hogar ccaa Multiple-imputation estimates Imputations = 20

Generalized linear models Number of obs = 34873

Average RVI = 1.4755

Largest FMI = 0.7727

DF adjustment: Large sample; DF: min = 32.95; avg = 63.96; max = 110.62

Model F test: Equal FMI F(13, 657.1) = 3421.04Within VCE type: OIM
Prob> F = 0.0000

rentatotal2	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
yhatsp	1.577203	.0109485	144.06	0.000	1.555201	1.599205
sexosp	3111.632	176.0792	17.67	0.000	2759.611	3463.654
ecivilsp	-349.7447	83.32439	4.20	0.000	-516.9574	-182.532
paisnacsp	-1084.672	176.7266	-6.14	0.000	-1439.869	-729.4746
edadsp	33.31438	6.885386	4.84	0.000	19.64702	46.98174
estudsp	1463.603	85.69225	17.08	0.000	1292.312	1634.894
situacsp	-516.6326	59.956	-8.62	0.000	-636.1751	-397.0901
tipoedif	329.307	96.03448	-3.43	0.002	-524.1258	-134.4882
regten	-344.0319	67.83995	-5.07	0.000	-478.4664	-209.5974
nhabit	434.1877	83.91433	5.17	0.000	266.5121	601.8634
nmiemb	152.0886	137.583	1.11	0.276	-126.0035	430.1807
hogar	1616.337	64.25615	25.15	0.000	1485.6	1747.074
ccaa	73.81149	15.44782	4.78	0.000	42.97725	104.6457
cons	-14578.17	704.1266	-20.70	0.000	-15974.25	-13182.09

Multiple-imputation estimates Imputations = 20

Generalized linear models

Variance information						
	Within	Between	Total	RVI	FMI	efficiency
yhatsp	.000045	.000071	.00012	1.65039	.637208	.969123
sexosp	13786.3	16397.	31003.9	1.2489	.569102	.972332
ecivilsp	2740.84	4002.01	6942.95	1.53314	.619623	.96995
paisnacsp	11727.7	18575.8	31232.3	1.66313	.639023	.969038
edadsp	26.3188	20.0854	47.4085	.801317	.456064	.977705
estudsp	3281.14	3868.59	7343.16	1.23799	.566899	.972436
situacsp	1738.68	1767.74	3594.81	1.06755	.529361	.974214
tipoedif	2496.92	6405.43	9222.62	2.6936	.743244	.964169
regten	2694.9	1816.53	4602.26	.707766	.424747	.979204
nhabit	3182.66	3675.19	7041.61	1.21249	.561663	.972684
nmiemb	5863.23	12443.7	18929.1	2.22844	.7047	.965964
hogar	993.769	2985.79	4128.85	3.15474	.7727	.962802
ccaa	111.526	121.056	238.635	1.13973	.54601	.973425
cons	285357	200417	495794	.737455	.435058	.97871

Los métodos de imputación simple, al contrario que los *listwisedeletion* (que eliminan del análisis cada registro en el que aparece un valor *missing* en cualquiera de las variables) no descartan los valores *missing*. Ello evita el sesgo en la estimación de parámetros, pero a cambio infraestima la varianza y ofrece valores con precisión sobreestimada y test muy optimistas. La imputación múltiple rectifica este problema al crear imputaciones múltiples y tener en cuenta la variabilidad muestral debida a los datos *missing*. El proceso se lleva a cabo en dos fases separadas, una de imputación que rellena los datos *missing* y otra de análisis, que proporciona inferencia sobre los resultados imputados de forma múltiple. Debido a la facilidad computacional, se recomiendan 20 imputaciones para reducir el error muestral debido a las imputaciones (StataCorp, 2013), y es la opción escogida en este análisis.

3.3. *Propensity score matching* (PSM)

La técnica PSM se utiliza normalmente en el ámbito de la evaluación de impacto como herramienta de construcción de contrafactuales válidos cuando no existe la posibilidad de comparar grupos de tratamiento y control en el contexto de diseños experimentales. La idea subyacente es la de encontrar entre el grupo de no tratados aquellos elementos que sean lo más parecidos posible a los tratados, para que, por diferencia entre los resultados en uno y otro grupo, pueda aislarse el efecto de un tratamiento, sin incluir ninguna otra característica o circunstancia que no sea achacable al mismo. La construcción del grupo de no tratados comparable (control-contrafactual) utiliza la información de las características relevantes que determinan la probabilidad de ser tratado para escoger las unidades más parecidas. La misma mecánica de construcción de contrafactuales es la que se utiliza para emparejar observaciones de dos encuestas diferentes. A partir de las variables comunes en la ECV y la EPF se determina la probabilidad de pertenecer a la ECV mediante un modelo probit. Una vez estimado el modelo se calcula el score, que es la predicción de la probabilidad de dicha pertenencia para las observaciones de la EPF. La información de todos los regresores queda subsumida en el score, de manera que emparejando observaciones de distintas encuestas con el score más parecido, logramos un método de emparejamiento. La correspondencia no es biunívoca, de manera que hay observaciones de la ECV que pueden ser pareja de distintas observaciones de la EPF, y alguna observación de la ECV puede no ser usada en el emparejamiento. Las parejas se buscan solamente en la región de soporte común, y el criterio de emparejamiento utilizado es el de distancia mínima entre los scores. Una vez identificada la observación de la ECV más parecida para cada EPF, se asigna la renta correspondiente, pudiéndose construir una base que incorpora la renta, a otro conjunto de variables demográficas y de consumo.

En el modelo que se ha utilizado para calcular el score, además de utilizar la información de las variables continuas (edad y renta del sustentador principal), se categoriza toda la información común en la ECV y la EPF en forma de variables dicotómicas. Se introducen entonces todos los cruces posibles como variables explicativas en el modelo probit.

Se prueba también un modelo probit en el que se introducen como variables explicativas todas las comunes en la ECV y la EPF, pero la comparación de los resultados nos hace decantarnos por el primero descrito.

Model	probit	probit	diff
N	34987	34987	0
Log-Lik Intercept Only	-23461.548	-23461.548	0.000
Log-Lik Full Model	-20147.354	-20925.018	777.663
D	40294.709(34899)	41850.035(34770)	-1555.326(129)
LR	6628.388(78)	5073.061(215)	1555.326(-137)
Prob> LR	0.000	0.000	0.000
McFadden's R2	0.141	0.108	0.033
McFadden'sAdj R2	0.138	0.099	0.039
Maximum Likelihood R2	0.234	0.183	0.051
Cragg & Uhler's R2	0.173	0.135	0.038
McKelvey and Zavoina's R2	0.310	0.222	0.087
Efron's R2	0.179	0.138	0.042
Variance of y*	1.449	1.286	0.163
Variance of error	1.000	1.000	0.000
Count R2	0.701	0.676	0.025
Adj Count R2	0.240	0.177	0.063
AIC	1.157	1.209	-0.052
AIC*n	40470.709	42284.035	-1813.326
BIC	-324844.170	-321939.151	-2905.019
BIC'	-5812.294	-2823.574	-2988.721

Difference of 2988.721 in BIC' provides very strong support for current model.

Además de estos tres métodos que servirán para realizar una primera unión y comparar los resultados obtenidos, durante el desarrollo del trabajo se ha planteado la posibilidad de una extensión futura que no explicamos aquí, y que consistiría en realizar un procedimiento de imputación doble. Con este método se imputaría la renta en la EPF (tal y como se ha descrito hasta ahora) y el consumo en la ECV, lo que permitiría contar con una base de datos más extensa. Ello requeriría una reponderación de los factores para que la representatividad global de la base nueva se mantuviese al realizar ejercicios que requieran elevación a total poblacional.

4. LA BONDAD DE AJUSTE: UNA COMPARACIÓN ENTRE LOS TRES MÉTODOS DE FUSIÓN

La comprobación de la bondad del ajuste realizado, pasa en cualquier caso por comparar la realidad conocida con la que se asume después de aplicar la metodología PSM o imputación. Es decir, si se conoce la renta verdadera declarada en la ECV y se compara con la que se la asigna con cada metodología, se puede comprobar si la diferencia es aceptable o no. Si se considera que lo es, el mecanismo puede darse por válido para asignar las rentas en la ECV, donde no hay forma de comprobar la bondad, ya que la renta verdadera no se conoce.

Por esta razón, los métodos utilizados han asignado una nueva renta también a las observaciones de la ECV: aunque el dato verdadero era conocido, se necesita comprobar cuánto se acerca a la realidad la renta que se asigna.

La comprobación de la bondad de la nueva distribución de renta creada no puede centrarse exclusivamente en la propia renta, y sin condicionar a la pertenencia a determinadas categorías. Es necesario que se compruebe que la distribución de renta creada es similar a la verdadera cuando se dan diferentes características, en particular las reflejadas en las variables utilizadas en la unión, que son además comunes a las dos encuestas que se unen.

Esta aclaración es necesaria porque el fin último no es la creación de una distribución de renta que replique las características estadísticas de la distribución de renta conocida, sino la utilización de la base para distintos análisis posteriores, como cuál es la carga fiscal que se soporta por vía de la imposición directa e indirecta simultáneamente. Si las rentas de los hogares no son coherentes con las características de los mismos, el análisis que utilice la base de datos construida no tendría validez.

4.1. Comparación de los ajustes atendiendo exclusivamente a la renta

Si solamente se consideran la renta global, sin atender a ningún grupo de características, los resultados de los distintos métodos arrojan los siguientes valores:

	Obs.	Media	Std. Dev.	Mín.	Máx.
Renta ECV	13.791	30.707	20.239	0	99.996
Predicción lineal	13.677	30.749	18.028	-10.611	139.881
PSM	13.791	30.654	20.104	0	99.996
Imputación múltiple	13.677	30.590	17.926	-7.467.691	78.859

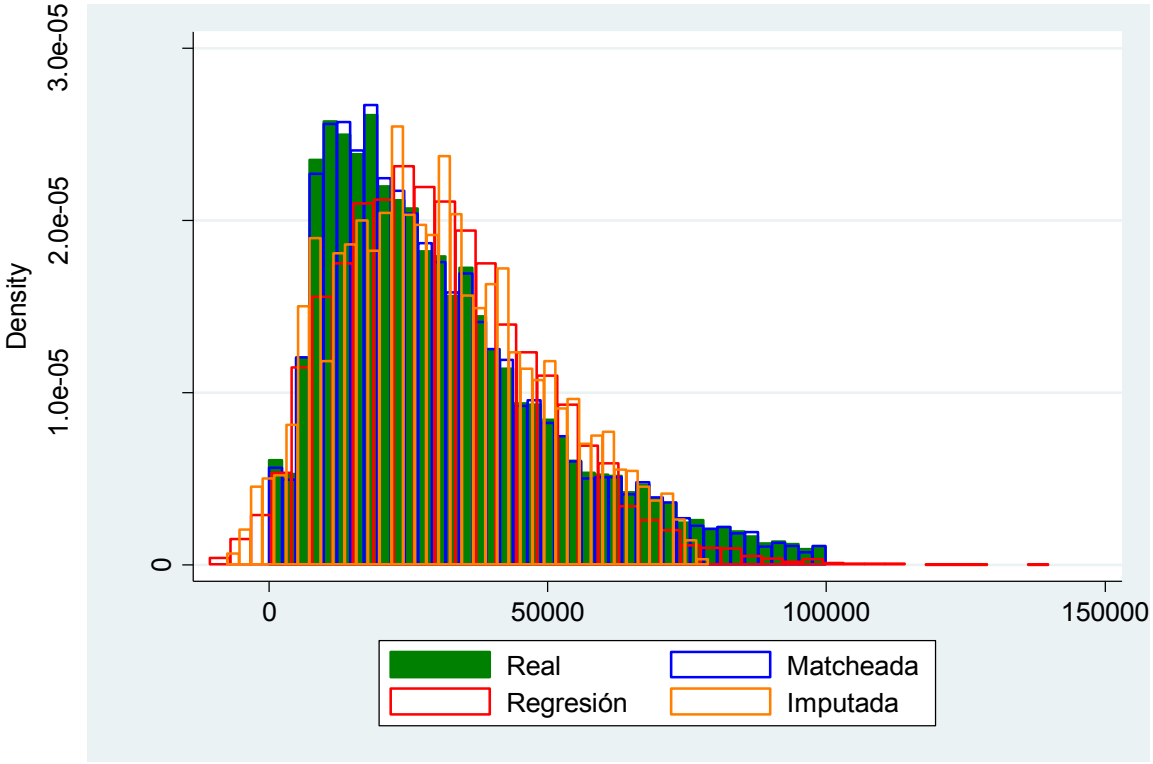
Y con valores detallados para distintos percentiles:

	Renta ECV	Predicción lineal	PSM	Imputación múltiple
1%	1.345	-2.136	1.613	-2.234
5%	6.585	5.544	6.759	4.779
10%	9.164	8.929	9.293	7.970
25%	15.060	17.302	15.093	16.837
50%	26.045	28.824	26.007	28.914
75%	41.478	41.814	41.389	42.585
90%	60.595	54.792	60.300	56.363
95%	71.810	62.290	71.653	63.314
99%	90.906	81.378	90.579	72.451

Las diferencias absolutas y porcentuales ente el valor de la renta en la ECV y las distintas rentas calculadas son las siguientes:

Percentil	Absolutas			Porcentuales		
	Predicción lineal	PSM	Imputación múltiple	Predicción lineal	PSM	Imputación múltiple
1%	3.478	-269	3.578	259%	-20%	266%
5%	6.031	-174	1.806	92%	-3%	27%
10%	-8.920	-129	1.193	3%	-1%	13%
25%	-2.242	-33	-1.776	-15%	0%	-12%
50%	-2.779	38	-2.869	-11%	0%	-11%
75%	-336	88	-1.107	-1%	0%	-3%
90%	5.803	294	4.231	10%	0%	7%
95%	9.520	157	8.495	13%	0%	12%
99%	9.528	327	18.456	10%	0%	20%

Histograma para los valores de renta en la ECV, mediante predicción lineal (Regresión), utilizando PSM (Matcheada) y con Imputación múltiple (Imputada).



Cuando se analiza la bondad del ajuste exclusivamente considerando la renta, el método PSM parece que permite los mejores resultados, tanto cuando se comparan los valores de la renta que se asignaría con la verdadera renta de la ECV para distintos percentiles de renta, y superponiendo los histogramas.

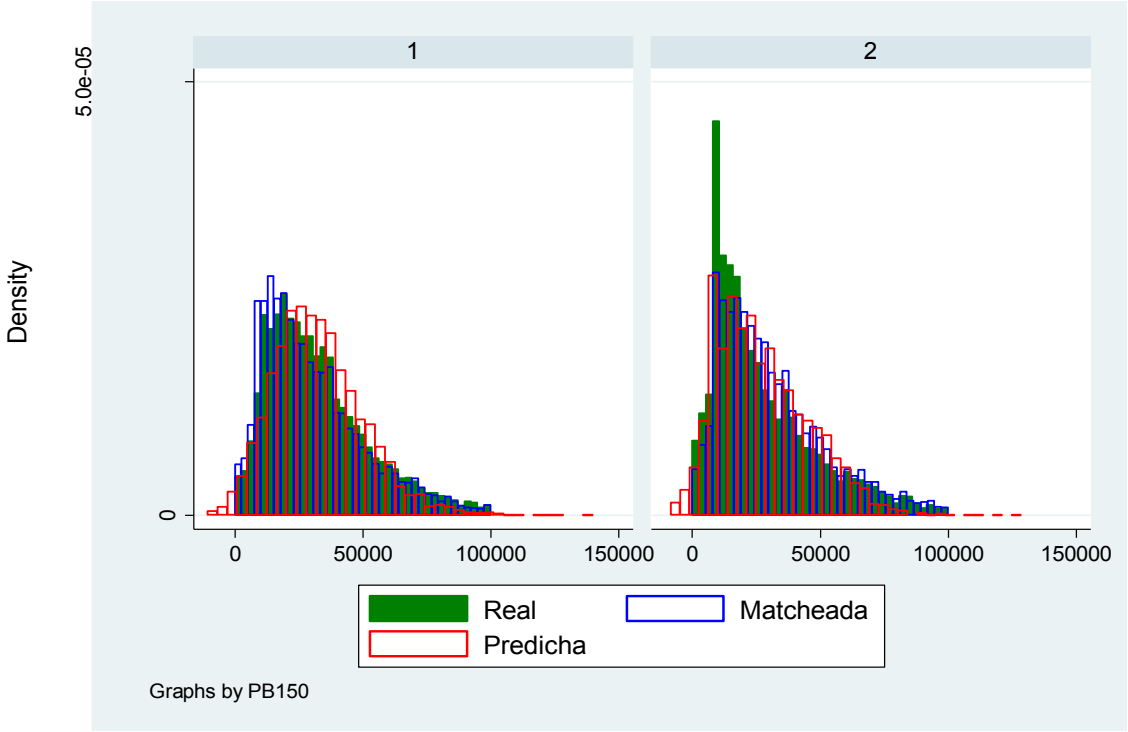
Un resultado adicional que ofrecemos como alternativa para llevar a cabo la comprobación de la bondad del ajuste, es comparar las diferencias en los índices de Gini de la verdadera distribución de renta de la ECV y la asignada según distintos métodos. Se trata de resumir la información de las distribuciones en un solo valor y comparar cuál de las posibles asignaciones de renta se asemeja más a la real.

4.2. Comparación de los ajustes atendiendo a otras características de las encuestas

El objetivo último del trabajo reside en construir una base de datos que reúna de forma simultánea información sobre la renta y sobre el consumo de las familias, con la finalidad de poder utilizar esa base de datos fusionada para realizar análisis de los efectos de la fiscalidad sobre el gasto y sobre la renta de forma simultánea. Los impuestos sobre la renta se establecen atendiendo a dimensiones adicionales a la propia renta, especialmente circunstancias familiares, para tratar de aproximar la carga impositiva a la verdadera capacidad de pago. Por esta razón, es menester que la base de datos reproduzca de forma fidedigna no solamente la distribución global de la renta, sino la distribución atendiendo a distintas características. Por esta razón, el análisis de la bondad del ajuste se ha realizado atendiendo a distribuciones subgrupales.

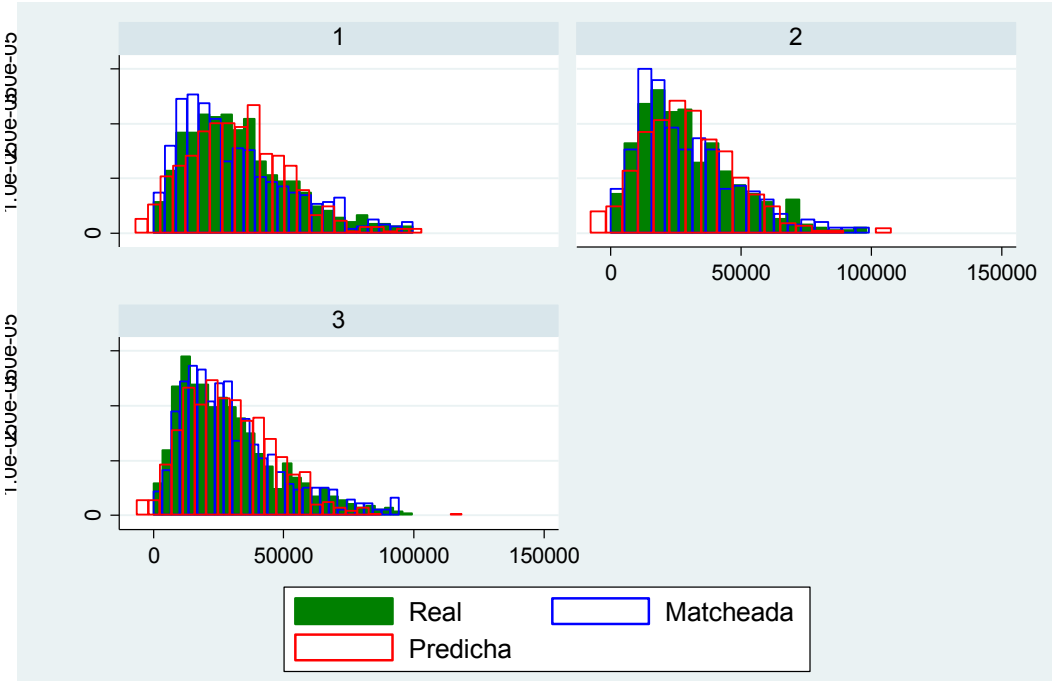
En las siguientes líneas exponemos superpuestos, los histogramas obtenidos para la distribución real de la renta (Real), la que se obtendría utilizando PSM (Matcheada), y la que se obtendría prediciendo mediante un método de imputación simple (Predicha). Estas comprobaciones se realizan a modo de ejemplo por sexo del sustentador principal, y por CCAA.

— Comprobando los ajustes por sexo del sustentador principal (1: hombre, 2: mujer).

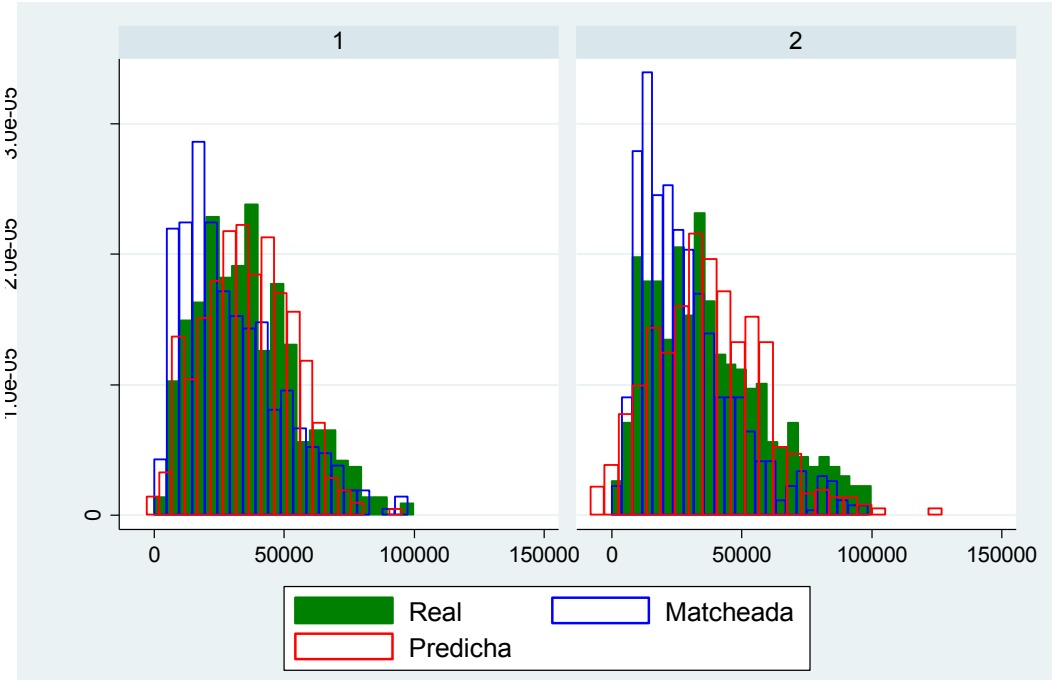


— Por CCAA.

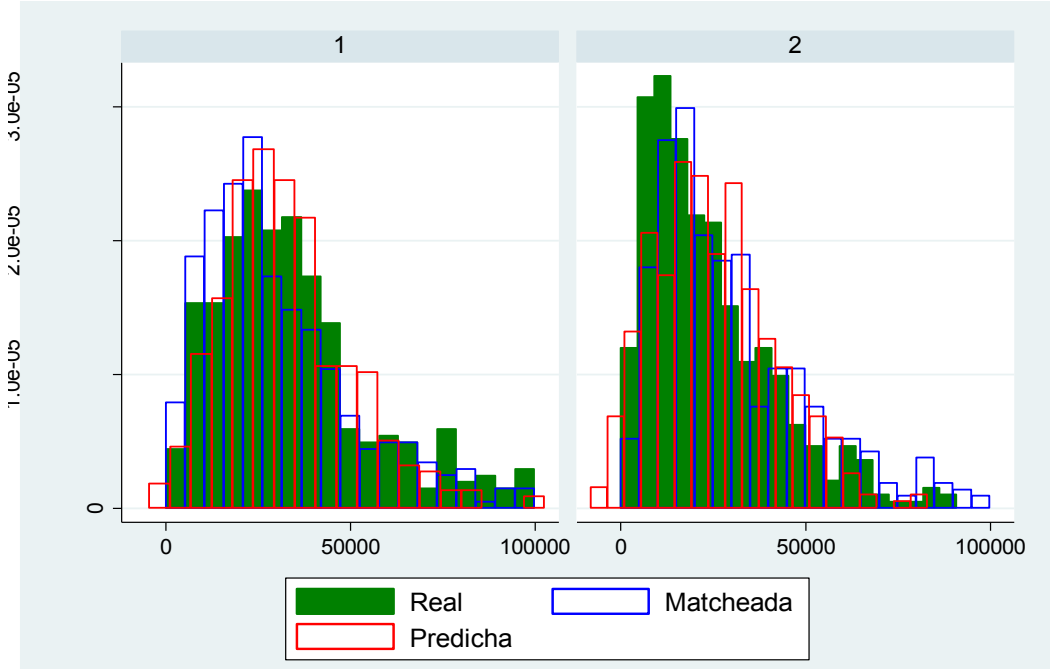
Norte. 1: Asturias, 2: Cantabria, 3: Galicia.



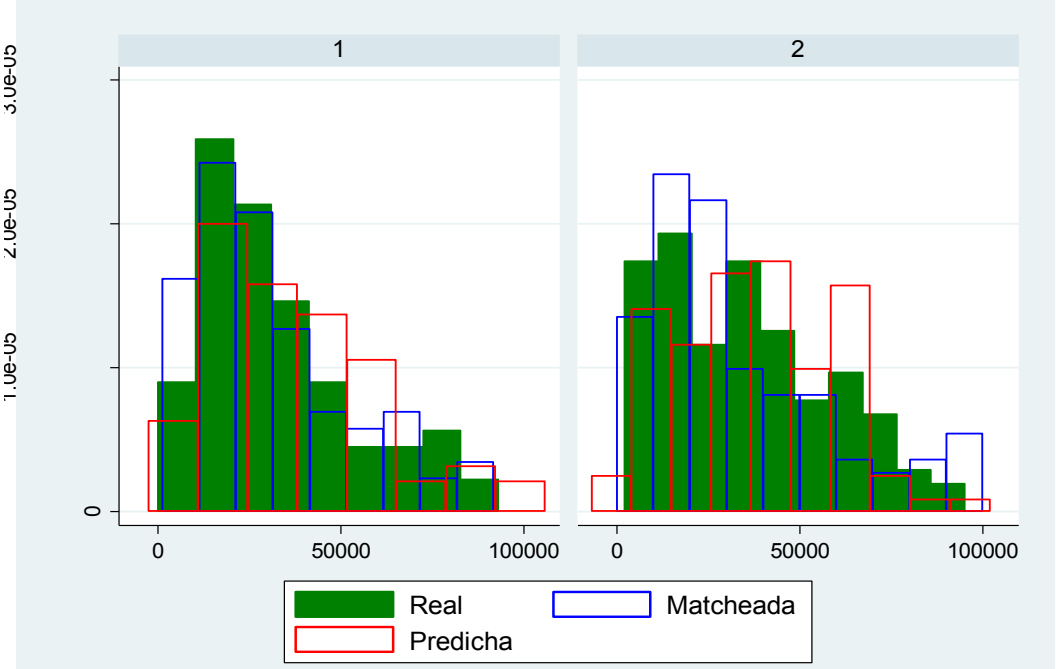
Forales: 1: Navarra, 2: País Vasco



Islas: 1: Baleares, 2: Canarias



África: 1: Ceuta, 2: Melilla



La comprobación visual de los histogramas no puede ser suficiente, por ello se presentan en el Apéndice I los resultados de las medias obtenidas utilizando PSM y predicción mediante regresión simple.

Al comparar los valores de las medias se puede comprobar que una predicción mediante regresión lineal es capaz de ajustar mucho mejor la renta media por subgrupos que la distribución obtenida mediante PSM, pero el propio método está sesgado a minimizar los errores en positivo y negativo con respecto a la media, por lo que no podríamos decidir cuál de los métodos es mejor atendiendo a esta sola dimensión.

Esta es la razón por la que en el quinto apartado mostramos un método alternativo de comprobación de la bondad del ajuste, rescatando herramientas propias del análisis distributivo.

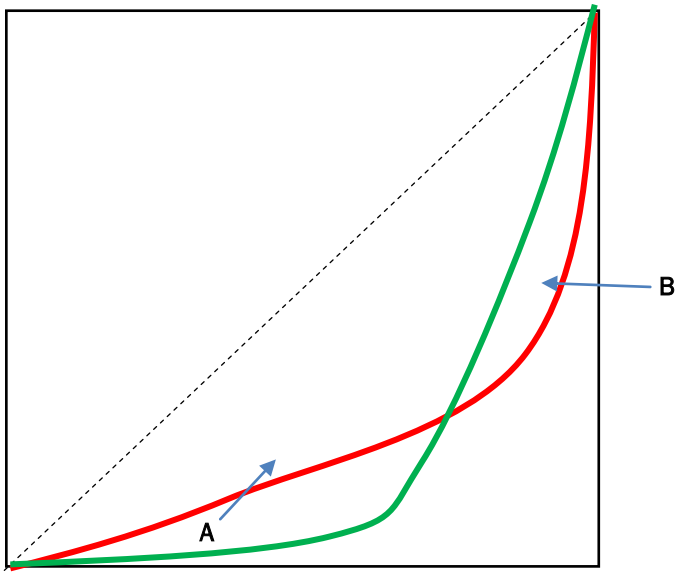
5. EL ANÁLISIS REDISTRIBUTIVO: LA UTILIZACIÓN DEL ÍNDICE DE REYNOLDS-SMOLENSKY PARA COMPRENDER LA BONDAD DE AJUSTE

Un resultado adicional que ofrecemos como alternativa para llevar a cabo la comprobación de la bondad del ajuste, es comparar las diferencias en los índices de Gini de la verdadera distribución de renta de la ECV y la asignada según distintos métodos, y descomponerla de manera que se aísle el efecto reordenación. El índice de Reynolds-Smolensky (RS) captura diferencias entre los índices de desigualdad de dos distribuciones, y el índice de desigualdad de cada distribución es una dimensión que resume la distribución en un solo valor. Hay que tener en cuenta que dos distribuciones diferentes pueden arrojar el mismo valor de la desigualdad, por lo que además de comprobar las diferencias en desigualdad, comprobaremos las diferencias en las distribuciones a través de las curvas de Lorenz.

Esta forma de comparar de la bondad del ajuste debería realizarse después de haber comprobado que las medidas de tendencia central, así como las de dispersión de la distribución son correctas. De esta forma se cuenta con más criterios para la elección, ya que la mera comprobación de la media puede cegar lo que ocurre a lo largo de toda la distribución. Por ejemplo, dado que las curvas de Lorenz se construyen sobre proporciones, si la distribución B es igual que la A pero duplicada, las curvas de Lorenz serían coincidentes aunque sus medias fueran una el doble que la otra.

Además de conocer cuánto se separan las distribuciones en términos proporcionales, se puede utilizar la descomposición del efecto total del RS en función del tipo medio efectivo, la reordenación y el índice de progresividad de Kakwani, reinterpretando su concepción tradicional para aportar más información acerca de la posible separación entre la renta verdadera y la asignada.

El principio, sería deseable para comparar la similitud entre dos distribuciones que el índice de RS sea mínimo, porque ello indicará que los índices de Gini de la distribución real y asignada serían muy cercanos. Pero esta condición es necesaria pero no suficiente para concluir bondad en el ajuste de las distribuciones proporcionales. Al ser Gini un índice construido sobre áreas, puede ocurrir que se compensen diferencias debidas a cruces en las curvas de Lorenz, por ejemplo si A y B fuesen iguales.



Además de comprobar la diferencia entre las curvas de Lorenz, para no analizar exclusivamente la diferencia ente los índices de Gini se puede acudir a la descomposición del mismo a partir del efecto del tipo medio (t), la progresividad (K) y la reordenación (D)

La descomposición RS se puede escribir como:

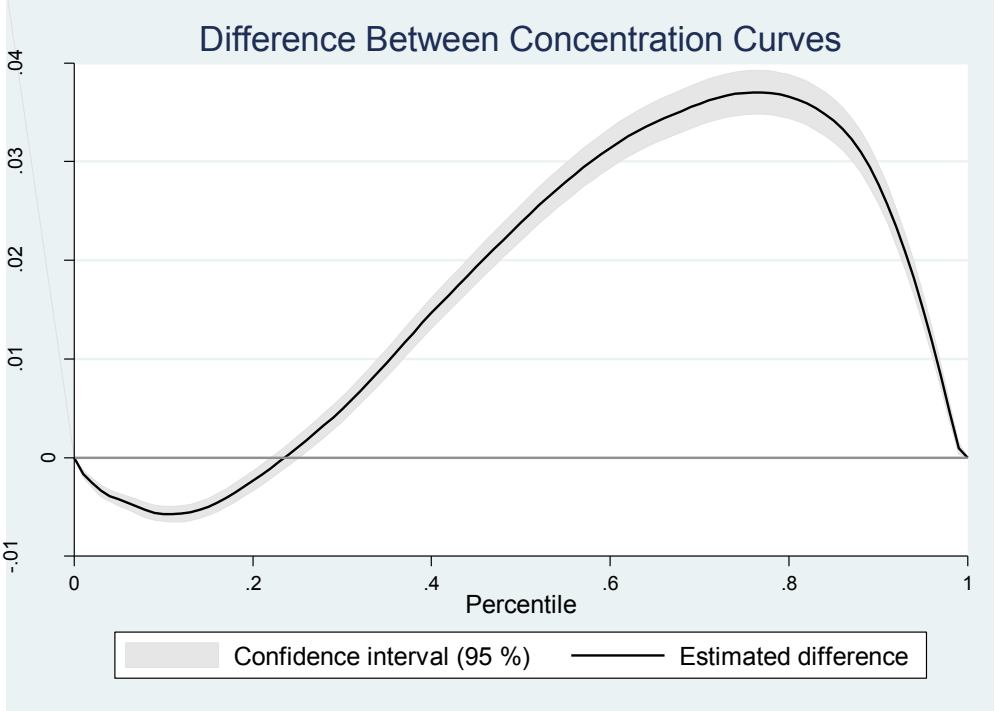
$$RS = \frac{t}{1-t} \cdot K - D \qquad (1)$$

Donde RS indica la diferencia entre el índice de Gini de la distribución real y el de la asignada a partir del tipo medio, t, que determina la diferencia entre la media de la distribución inicial y la asignada; K, que captura de qué forma se relaciona esa diferencia con el nivel de renta, y el efecto reordenación D, que indica cómo cambian de posición las distintas unidades por adoptar en la distribución asignada el orden de la verdadera renta, y no estar ordenadas de menor a mayor. En el apéndice 2 se presenta un ejemplo con 25 observaciones para ilustrar el uso de las herramientas distributivas al análisis de la bondad del ajuste de la fusión estadística que puede aclarar estas cuestiones.

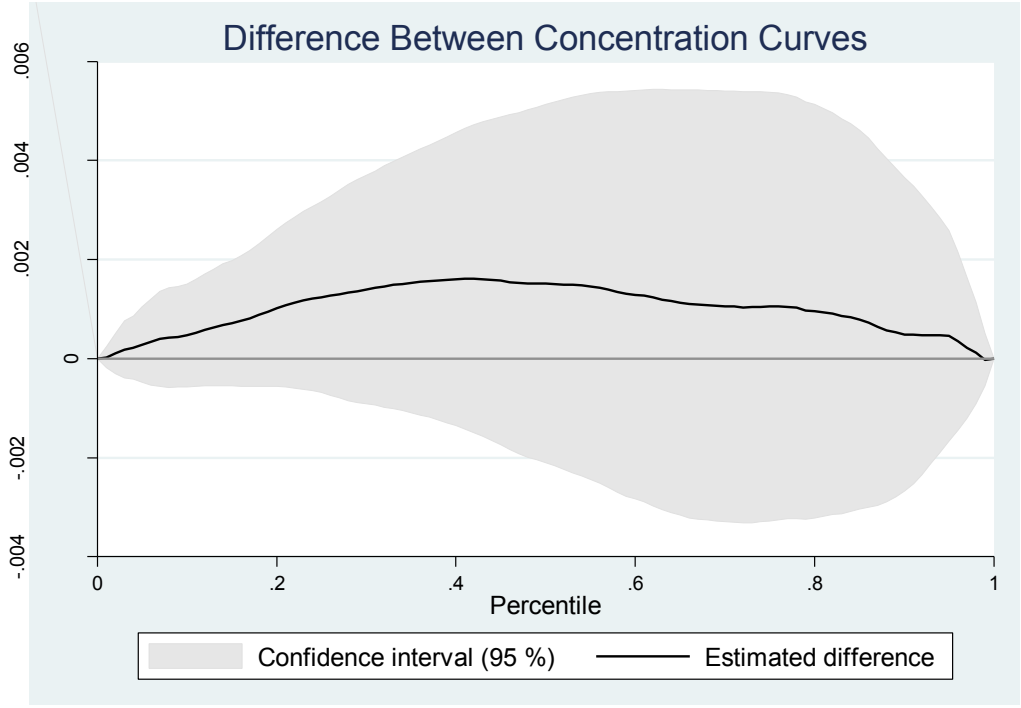
Para lograr una mayor similitud entre la distribución real y la asignada, sería deseable entonces que el tipo medio fuese cercano a cero, porque ello indicaría que la media de la distribución asignada se diferencia poco de la verdadera. Pero esta no es la única cuestión que hay que tener en cuenta, como ya se expuso al comparar los resultados del Apéndice 1. El índice de Kakwani, K, se construye a partir de las diferencias entre la verdadera renta y la asignada acumuladas una a una de todas observaciones, y no presentará los valores habituales obtenidos en el estudio de la progresividad, ya que el "impuesto pagado" por cada renta es equiparable ahora a la diferencia entre la renta real y la asignada, y no tiene porqué ser mayor cuanto mayor es la renta como ocurre con la fiscalidad progresiva. Se obtendrán ahora valores absolutos mucho más elevados de K que en análisis tradicional de la progresividad y con valores tanto positivos como negativos.

Por último, D captura en términos ordinales, la diferencia entre la distribución de renta asignada ordenada de menor a mayor nivel de renta asignada, y la renta asignada ordenada por la distribución real.

En primer lugar se exponen las diferencias entre las curvas de Lorenz de la distribución real y la renta asignada con una predicción obtenida con regresión lineal:

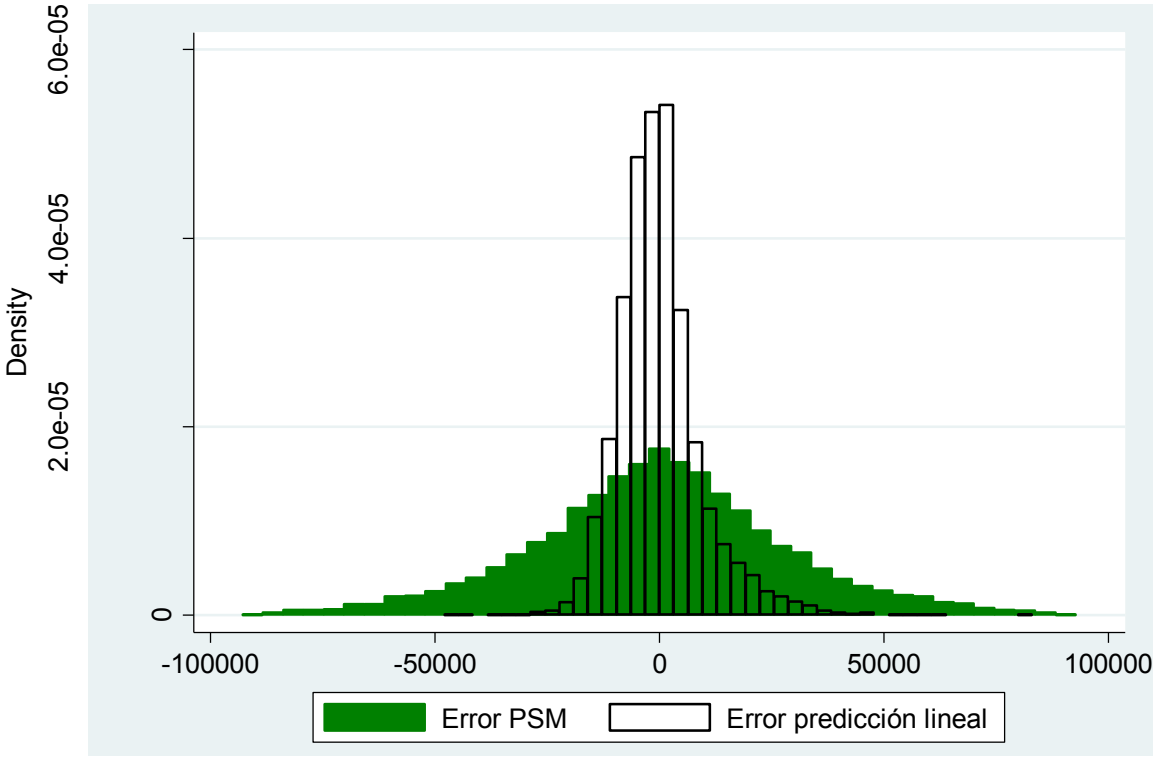


A continuación, se grafica la diferencia entre la curva de Lorenz de la distribución real y la asignada mediante PSM:



Es importante tener en cuenta la escala del error, para concluir que las distribuciones según curva de Lorenz son más parecidas al comparar la renta real y la asignada con PSM que al asignarla con predicción del modelo de regresión lineal.

Otra forma de comprobar las diferencias entre la bondad de cada aproximación es graficar la distribución de los errores en cada sistema, entendiendo por "Error PSM" la diferencia entre el verdadero valor de la distribución y el que se le asigna a cada observación mediante la técnica PSM, y por "Error predicción lineal" la diferencia entre la verdadera renta de la ECV y la asignada mediante un sistema de predicción lineal.



Para comprobar cómo estos resultados se muestran en concordancia con el índice RS y su descomposición, mostramos los valores obtenidos.

	PSM	Predicción lineal
Gini renta real	0,3599215	0,3599215
Gini renta asignada	0,3579225	0,326693
RS	**0,001999**	-0,0332285

Los valores anteriores ponen de manifiesto que la diferencia en la desigualdad entre la distribuciones real y la asignada mediante PSM es menor (RS=0,001999) que al asignar con un método de predicción lineal (RS=-0,0332285).

También se han calculado las diferencias en desigualdad (RS) por distintas características, y se muestran a modo de ejemplo dos categorías, según el sexo del sustentador principal.

RS por sustentandor principal	PSM	Predicción lineal
Hombre	**-0,0218313**	0,0336642
Mujer	0,0362442	**0,0316266**

En este caso, la mayor igualdad (al menos entre las desigualdades de las distribuciones) se obtiene mediante PSM para el grupo de sustentadores principales hombres y con predicción lineal si el grupo considerado es el de sustentadoras principales mujeres.

Calculando los RS por cuartiles de renta se comprueba que la mayor coincidencia entre distribuciones (medida exclusivamente por la desigualdad) se logra mediante PSM en el primer cuartil, y mediante predicción lineal en los restantes.

Rs por cuartil de renta	PSM	Predicción lineal
Cuartil 1	**-0,037592**	0,0674653
Cuartil 2	-0,1332659	**0,0449593**
Cuartil 3	-0,1579136	**0,078338**
Cuartil 4	-0,1603623	**0,0574822**

6. CONCLUSIONES

En este trabajo se ha explorado una forma novedosa de emparejamiento entre la ECV y la EPF. Se trata de la aplicación de la metodología PSM, propia de la evaluación de impacto. Este método tiene la ventaja de resumir toda la información de variables en un score, que se utiliza como elemento para el emparejamiento de los datos de la EPF con los de la ECV.

Antes de proceder a realizar el emparejamiento, es necesario comprobar que el ajuste entre la renta real y la que se está asignando es suficientemente buena: para ello se comprueba cuán parecida es la distribución obtenida con PSM a la verdadera distribución de renta de la ECV. Al mismo tiempo se compara el emparejamiento realizado con PSM con métodos alternativos, para comprobar si PSM es superior a los demás.

El fin práctico del trabajo que se ha expuesto es construir una base de datos que permita analizar de forma simultánea la imposición sobre la renta y el consumo. Si el fin último residiese exclusivamente en replicar una distribución de renta, sin condicionar por ninguna otra característica, la técnica del PSM nos permitiría un ajuste casi perfecto. Pero piénsese que con esta técnica se busca el *score* más cercano para el emparejamiento, y una vez emparejadas observaciones con datos de la ECV se asigna la renta de esa pareja. Ello logra muy buen ajuste de la renta, pero al condicionar la representación de los histogramas de la renta por otras características, el ajuste no es tan bueno como para el total. Es decir, si se representan los histogramas de la renta verdadera y la asignada con PSM, las distribuciones de la renta se superponen prácticamente, pero si se representa la distribución por comunidades autónomas, por tipo de hogar, o por percentiles de renta, por poner algunos ejemplos, las coincidencias ya no se dan con tanta exactitud.

Además de la superposición de los histogramas, se han analizado formas alternativas para tratar de elegir cuál es la mejor forma de asignar las rentas. La comprobación de los valores medios de las rentas para distintos niveles de la misma es necesario para juzgar los mecanismos de emparejamiento, pero no es suficiente.

En aras de contar con mecanismos adicionales de comprobación, que sean al mismo tiempo sencillos de aplicar, se han calculado las mediciones propias del análisis distributivo, para sintetizar la información de las distribuciones y comprobar las diferencias. En particular, se han calculado las diferencias entre las desigualdades de la distribución real y las construidas para el emparejamiento mediante el índice de Reynolds-Smolensky. Al aplicar estas mediciones, de nuevo se obtiene mejor ajuste con PSM que con otros métodos al tratar de replicar la distribución global. Cuando las distribuciones se establecen para subgrupos, PSM no aparece siempre como la mejor alternativa.

No se puede establecer una conclusión fuerte en cuento a la bondad de PSM frente a otros métodos de emparejamiento: lo es sin duda para replicar la distribución de renta global, pero no lo es siempre para distribuciones por subgrupos.

Puesto que se trata de un trabajo en curso, que lleva tiempo desarrollándose y que se prevé seguir desarrollando en el medio plazo, es conveniente plantear posibles extensiones del mismo: por un lado, de cara a explotar más las herramientas distributivas para comprobar la bondad de los ajustes entre distribuciones, y por otro, para buscar mecanismos de emparejamiento alternativos.

Bibliografía

CALIENDO, M. y KOPEINIG, S. (2005): "Some Practical Guidance for the Implementation of Propensity Score Matching", *DIW Discussion Papers*, No. 485, http://ftp.iza.org/dp1588.pdf.

CONTIA P. L.; MARELLAB, D., y SCANUC, M. (2016): "StatisticalMatchingAnalysisforComplexSurveyDataWithApplications", *Journal of the American Statistical Association*, vol. 111, n.º516.

DECOSTER A. y VAN CAMP, G. (2002): De constructie van één samengesteld bestand op basis van twee bestanden: koppeling van de budgetenquete 1997-98 en het fiscaal bestand 1999 (inkomstens 1998) [i.e. Match the expenditure survey of 1997-98 to the income survey of 1999].

DECOSTER A.; DE SWERDT K., y VAN CAMP, G. (2004): "Matching of income and expenditure data by means of nonparametric estimation of Engel Curves", *Report of the D.W.T.C.* project AG/01/079.

DECOSTER A., ROCK, B. D.; SWERDT, K. D.; LOUGHREY, J.; O'DONOGHUE, C., y VERWERFT, D. (2007): "Techniques to impute expenditures into an income data set EUROMOD AIMAP deliverable 3.4", Institute for Social and Economic Research, https://www.iser.essex.ac.uk/files/msu/emod/aimap/deliverables/AIM-AP1.1b.pdf.

DECOSTER, A.; OCHMANN, R., y SPIRITUS, K. (2013): "Integrating VAT into EUROMOD. Documentation and results for Germany", *EUROMOD Working Paper Series*, EM20/13, https://www.iser.essex.ac.uk/research/publications/working-papers/euromod/em20-13.pdf.

DECOSTER, A.; OCHMANN, R., y SPIRITUS, K. (2014): "Integrating VAT into EUROMOD. Documentation and results for Belgium", *EUROMOD Working Paper Series*, EM12/14, https://www.iser.essex.ac.uk/research/publications/working-papers/euromod/em12-14.pdf.

D'ORAZZIO, M.; DI ZIO, M., y SCANU, M. (2013): "Old and new approaches in statistical matching when samples are drawn with complex survey designs", https://www.istat.it/it/files/2013/12/Old-and-new-approaches-in-statistical-matching.pdf.

EUROPEAN COMISSION (2013): "Statistical matching: a model based approach for data integration" *EUROSTAT Methodologies and Working Papers*, https://ec.europa.eu/eurostat/web/products-statistical-working-papers/-/KS-RA-13-020?inheritRedirect=true.

EUROPEAN COMISSION (2013): "Statistical matching of EU-SILC and the Household Budget Survey to compare poverty estimates using income, expenditures and material deprivation" *EUROSTAT Methodologies and Working Papers*, https://ec.europa.eu/eurostat/documents/3888793/5857145/KS-RA-13-007-EN.PDF/37d4ffcc-e9fc-42bc-8d4f-fc89c65ff6b1.

LÓPEZ LABORDA, J.; MARÍN GONZÁLEZ, C., y ONRUBIA, J. (2016): "Estimación de los impuestos pagados por los hogares españoles en 2013 a partir de la Encuesta de Presupuestos Familiares y la Encuesta de Condiciones de Vida. Metodología", *FEDEA. Estudios sobre la Economía Española-2016/20,* http://documentos.fedea.net/pubs/eee/eee2016-20.pdf.

LÓPEZ LABORDA, J.; MARÍN GONZÁLEZ, C., y ONRUBIA, J. (2016): "Observatorio sobre el reparto de los impuestos entre los hogares españoles", *FEDEA. Estudios sobre la Economía Española-2016/21,* http://documentos.fedea.net/pubs/eee/eee2016-21.pdf.

LÓPEZ LABORDA, J.; MARÍN GONZÁLEZ, C., y ONRUBIA, J. (2017): "Estimating Engel curves: A new way to improve the SILC-HBS matching process", Metodología. *FEDEA. Documentos de Trabajo-2017/15,* http://documentos.fedea.net/pubs/dt/2017/dt2017-15.pdf.

OKNER, B. (1972a): *Constructing a New Data Base from Existing Microdata Sets: the 1966 Merge File. Annals of Economic and Social Measurement*, 1(3), 325-342.

OKNER, B. (1972b): REPLY AND COMMENTS. ANNALS OF ECONOMIC AND SOCIAL MEASUREMENT, 1(3), 359-362.

OKNER, B. (1974): *Data Matching and Merging: An Overview. Annals of Economic and Social Measurement,* 3(2), 347-352

RUBIN, D. B. (1987): *Multiple imputation for non-responses in surveys*, New York, NY: John Wiley&Sons.

RUBIN D. B. (1996): "Multiple Imputation After 18+ Years", *Journal of the American Statistical Association,* Vol. 91, No. 434 (Jun., 1996), pp. 473-489, http://www.jstor.org/stable/2291635.

STATACORP (2013): *Stata Multiple Imputation reference manual. Release 13*, Statistical Software, College Station, TX: Stata Corp LP, https://www.stata.com/manuals13/mi.pdf.

TAYLOR, R.; SUTHERLAND, H., y GOMULKA, J. (2001): Using POLIMOD to evaluate alternative methods of expenditure imputation, Microsimulation unit research note MU/RN38.

APÉNDICE 1. RENTA MEDIA POR SUBGRUPOS EN LA DISTRIBUCIÓN REAL, IMPUTADA Y PREDICHA

Renta neta media por CCAA.

Andalucía	Obs.	Media	Desv. típica	Mínimo	Máximo
Renta ECV	1415	25095.45	18032.24	0	99505.3
Renta PSM	1415	30236.56	19922.68	0	99383.1
Renta regresión	1402	26487.56	16697.38	-6801.308	97197.55

Aragón	Obs.	Media	Desv. típica	Mínimo	Máximo
Renta ECV	498	32545.46	19868.7	39.4	98275.91
Renta PSM	498	30761.74	19844.72	0	99962.9
Renta regresión	497	31329.71	17064.07	-7116.374	98577.29

Asturias	Obs.	Media	Desv. típica	Mínimo	Máximo
Renta ECV	565	33078.31	20057.95	59.1	99651.7
Renta PSM	565	31300.95	21659.82	0	99715.18
Renta regresión	564	32778.77	18879.37	-6784.655	102967.4

Baleares	Obs.	Media	Desv. típica	Mínimo	Máximo
Renta ECV	385	34217.74	21111.3	49.7	99733.3
Renta PSM	385	30294.83	20355.1	49.1	99715.18
Renta regresión	384	32385.02	17343.88	-4408.108	102532.8

Canarias	Obs.	Media	Desv. típica	Mínimo	Máximo
Renta ECV	422	24134.37	17594.05	0	90906
Renta PSM	422	30225.12	20352.87	0	99819.6
Renta regresión	415	25550.86	16401.74	-8176.478	82971.3

Cantabria	Obs.	Media	Desv. típica	Mínimo	Máximo
Renta ECV	377	30028.19	18627.98	134.6	98368.9
Renta PSM	377	29509.7	19523.17	0	99188.2
Renta regresión	375	29938.27	17776	-7798.578	107640.7

Castilla y León	Obs.	Media	Desv. típica	Mínimo	Máximo
Renta ECV	859	29615.62	18975.75	0	99819.6
Renta PSM	859	30459.59	19626.37	0	99819.6
Renta regresión	857	30222.9	17670.64	-6716.389	96624.63

Castilla-La Mancha	Obs.	Media	Desv. típica	Mínimo	Máximo
Renta ECV	574	26138.08	18587.8	0	94000
Renta PSM	574	29469.17	19671.71	0	99505.3
Renta regresión	569	27514.85	16108.61	-8804.654	95840.55

Cataluña	Obs.	Media	Desv. típica	Mínimo	Máximo
Renta ECV	2886	33286.87	21435.9	0	99996
Renta PSM	2886	32306.33	20661.3	0	99733.3
Renta regresión	2847	31421.86	18427.9	-10293.7	128452.6

Com. Valenciana	Obs.	Media	Desv. típica	Mínimo	Máximo
Renta ECV	916	27901.93	18363.11	0	99452.2
Renta PSM	916	30192.06	20086.16	0	99996
Renta regresión	905	29538.79	17631.76	-8500.999	123485.9

Extremadura	Obs.	Media	Desv. típica	Mínimo	Máximo
Renta ECV	531	24924.89	17238.01	0	94860.1
Renta PSM	531	29804.76	18946.73	0	94961.3
Renta regresión	528	26896.15	16127.39	-4953.845	82774.19

Galicia	Obs.	Media	Desv. típica	Mínimo	Máximo
Renta ECV	826	29246.89	19223.1	0	99383.1
Renta PSM	826	30418.82	19628.72	0	94451.2
Renta regresión	820	29545.35	17076.42	-6342.813	118484.5

Madrid	Obs.	Media	Desv. típica	Mínimo	Máximo
Renta ECV	1370	34125.89	21753.32	0	98534.98
Renta PSM	1370	30987.47	20595.14	0	98852.9
Renta regresión	1368	33668.02	18390.41	-8653.847	98762.17

Murcia	Obs.	Media	Desv. típica	Mínimo	Máximo
Renta ECV	512	25203.81	16806.19	0	92454.2
Renta PSM	512	28382.18	19282.15	201	97564.7
Renta regresión	510	27128.16	15698.32	-10610.57	96317.63

Navarra	Obs.	Media	Desv. típica	Mínimo	Máximo
Renta ECV	430	36034.43	18800.12	53.7	99753.8
Renta PSM	430	29903.53	19193.86	0	97516.6
Renta regresión	430	35772.97	16716.72	-2954.3	95399.26

País Vasco	Obs.	Media	Desv. típica	Mínimo	Máximo
Renta ECV	672	37493.71	22461.93	0	99715.18
Renta PSM	672	28947.49	18768.64	56.4	98721.4
Renta regresión	669	37405.26	20935.95	-8139.178	127056.6

Rioja	Obs.	Media	Desv. típica	Mínimo	Máximo
Renta ECV	356	33296.96	20037.46	0	97628
Renta PSM	356	30486.45	18596.41	1431.6	99996
Renta regresión	356	33349.15	17755.87	-7045.47	139880.8

Ceuta	Obs.	Media	Desv. típica	Mínimo	Máximo
Renta ECV	86	32363.37	21380.99	0	93060.4
Renta PSM	86	30849.78	21574.57	1206.3	91786
Renta regresión	70	37355.13	23819.62	-2638.49	105839.7

Melilla	Obs.	Media	Desv. típica	Mínimo	Máximo
Renta ECV	111	35630.98	22437.34	2013	95239.1
Renta PSM	111	33251.07	25342.43	0	99962.9
Renta regresión	111	37276.87	20823.81	-6869.371	101950.1

Los valores de las rentas medias según el sexo del sustentador principal del hogar:

Hombre	Obs.	Media	Desv. típica	Mínimo	Máximo
Renta ECV	8574	32632.97	19911.6	0	99996
Renta PSM	8574	30446.29	19882.2	0	99962.9
Renta regresión	8497	32661.22	17740.32	-10610.57	139880.8

Mujer	Obs.	Media	Desv. típica	Mínimo	Máximo
Renta ECV	5217	27542.77	20377.33	0	99651.7
Renta PSM	5217	30995.51	20461.62	0	99996
Renta regresión	5180	27613.55	18059.95	-8430.937	128452.6

El nivel de estudios del sustentador principal se clasifica como sigue:

Primaria o menos	Obs.	Media	Desv. típica	Mínimo	Máximo
Renta ECV	3978	20122.15	12936.91	0	92857.3
Renta PSM	3978	31400.18	20554.14	0	99962.9
Renta regresión	3964	19721.36	12456.76	-10610.57	90272.78

1.ª etapa Secundaria	Obs.	Media	Desv. típica	Mínimo	Máximo
Renta ECV	3195	26289.18	16182.99	0	95799.7
Renta PSM	3195	29538.92	19577.48	0	99642
Renta regresión	3185	26977.03	14213.11	-8139.178	100399.1

2.ª etapa Secundaria	Obs.	Media	Desv. típica	Mínimo	Máximo
Renta ECV	2617	33107.5	19489.56	0	99733.3
Renta PSM	2617	30945.53	20857.86	0	99962.9
Renta regresión	2608	33475.55	16777.73	-6342.813	128452.6

Superior	Obs.	Media	Desv. típica	Mínimo	Máximo
Renta ECV	3926	43471.98	22456.38	0	99996
Renta PSM	3926	30669.8	19638.92	0	99996
Renta regresión	3920	43152.8	18235.3	-7798.578	139880.8

Según el tipo de hogar:

Adulto solo >65 años	Obs.	Media	Desv. típica	Mínimo	Máximo
Renta ECV	1634	16088.85	10559.11	0	94862.9
Renta PSM	1634	30642.06	20649.7	0	99819.6
Renta regresión	1630	14917.51	12326.81	-8430.937	119574

Persona sola 30<64 años	Obs.	Media	Desv. típica	Mínimo	Máximo
Renta ECV	1445	20789.72	15814.62	0	98792.1
Renta PSM	1445	30678.16	19989.73	0	98348
Renta regresión	1439	21558.43	18545.2	-10610.57	107640.7

Persona sola <30 años	Obs.	Media	Desv. típica	Mínimo	Máximo
Renta ECV	88	12848.8	9342.9	0	51382.5
Renta PSM	88	33570.06	23535.86	0	86449.9
Renta regresión	88	15062.09	11673.81	-5278.909	55963.62

Adulto solo y algún <16 años	Obs.	Media	Desv. típica	Mínimo	Máximo
Renta ECV	465	21559.37	17029.67	0	98915.5
Renta PSM	465	27057.71	18143.14	0	98534.98
Renta regresión	465	24017.34	19207.3	-4098.887	113302.8

Pareja sin hijos, alguno >=65	Obs.	Media	Desv. típica	Mínimo	Máximo
Renta ECV	2311	27936.56	16120.3	0	99383.1
Renta PSM	2311	31128.38	20318.68	0	99962.9
Renta regresión	2301	28289.11	14284.38	2307.66	127056.6

Pareja sin hijos, ambos <65	Obs.	Media	Desv. típica	Mínimo	Máximo
Renta ECV	1820	33135.96	20397.3	0	99996
Renta PSM	1820	30478.69	19432.43	0	99819.6
Renta regresión	1794	32224.66	16732.08	-3051.594	118484.5

Pareja 1 hijo <16	Obs.	Media	Desv. típica	Mínimo	Máximo
Renta ECV	1495	35729.04	20360.26	0	99819.6
Renta PSM	1495	32409.13	21141.02	0	99828.7
Renta regresión	1486	35749.55	16479.79	-612.0732	128452.6

Pareja 2 hijos <16	Obs.	Media	Desv. típica	Mínimo	Máximo
Renta ECV	1552	39187.19	21417.04	0	99962.9
Renta PSM	1552	30902.01	20658.04	0	99996
Renta regresión	1545	40586.01	17590.43	542.2281	139880.8

Pareja 3 o más hijos <16	Obs.	Media	Desv. típica	Mínimo	Máximo
Renta ECV	271	34264.09	23960.55	0	99753.8
Renta PSM	271	31435.7	21668.88	0	99996
Renta regresión	267	38107.32	20141.47	1050.401	100399.1

Otros	Obs.	Media	Desv. típica	Mínimo	Máximo
Renta ECV	2707	39700.38	20690.86	0	99733.3
Renta PSM	2707	29702.3	19104.42	0	99733.3
Renta regresión	2664	38989.79	14857.36	1325.839	103746.8

APÉNDICE 2

Veamos un ejemplo sencillo con 25 observaciones en el que comparamos la verdadera distribución de renta (Renta cierta) con la imputada (Imputada ECV). La diferencia entre ambas es lo que denominamos "Efecto imputación" que sería nulo si lográsemos imputar el valor exacto de la renta cierta. La renta imputada no tiene porqué mantener un orden creciente, aunque la comparación entre la distribución real y la ordenada sí que se hará siguiendo un criterio de menor a mayor, por lo que en la última columna se presenta la renta imputada ordenada de menor a mayor. En negrita se señalan los valores que no se reordenan y que ocupan la misma posición que antes de ser ordenada la distribución de menor a mayor.

Orden renta cierta	Renta cierta (X)	Imputada en ECV (X-T)	Efecto imputación (T)	Imputada ordenada X-T ordenada por X
1	1996,9	2023,6	-26,65395273	1999,9
2	2313,7	1999,9	313,78864	2023,6
3	2907,4	3010,8	-103,4143218	**3010,8**
4	2970,7	3165,1	-194,3818954	3051,2
5	3032,3	3051,2	-18,88850084	3165,1
6	3133,6	3209,6	-76,02071215	**3209,6**
7	3689,8	3774,2	-84,44160454	**3774,2**
8	3985,4	4116,2	-130,8433095	3836,9
9	4133,6	3836,9	296,6986493	4116,2
10	4789,7	4449,8	339,9080769	**4449,8**
11	4959,8	5131,9	-172,0994094	**5131,9**
12	5609,6	5422,5	187,0742308	**5422,5**
13	6241,8	6333,5	-91,73902339	**6333,5**
14	7052,9	6979,0	73,85584609	**6979,0**
15	7122,5	7268,2	-145,7289653	**7268,2**
16	7347,7	7444,7	-97,03854312	**7444,7**
17	7784,5	7713,2	71,33133139	**7713,2**
18	7933,8	7724,4	209,389716	**7724,4**
19	8067,5	8300,6	-233,1054062	**8300,6**
20	9091,7	9349,5	-257,8094866	9333,9
21	9368,5	9592,5	-223,9637048	9349,5
22	9564,3	9333,9	230,3746132	9592,5
23	10333,9	10178,5	155,4386985	10105,0
24	10345,3	10105,0	240,2685811	10178,5
25	10923,5	10730,3	193,2163661	**10730,3**

Calculamos los índices de Gini de la renta cierta (X), de Gini y concentración de la renta imputada (X-T), así como el índice de concentración del efecto imputación (T):

Gini o concentración	**0,25625015**	**0,255002148**	**0,679121178**	**0,25539862**
	G_X	C_{X-T}	C_T	G_{X-T}

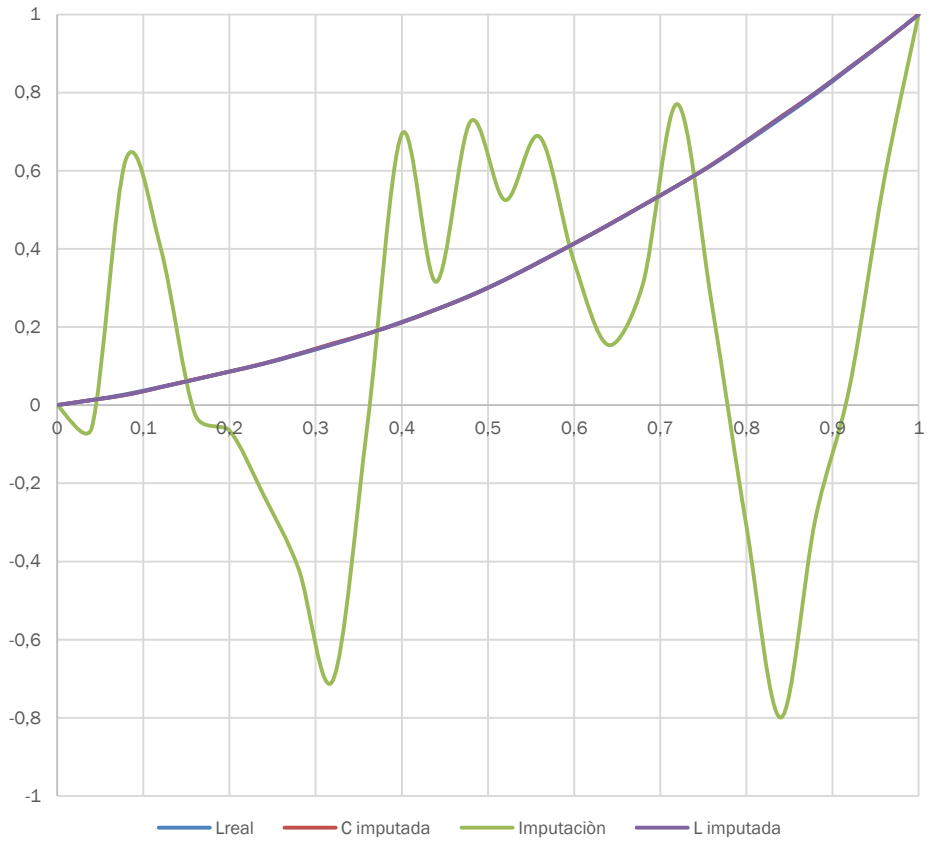
Y a partir de estos obtenemos el índice de Reynolds-Smolensky (RS), el índice de Kakwani (K), el tipo medio (t) y la reordenación (D):

RS	**0,00085152**
K	0,42287103
t	0,00294256
t/(1-t)	0,00295125
D	0,00039648

El valor de RS nos sirve como indicador de lo cercanas que están en términos de desigualdad la distribución de renta real y la imputada. Si fuese cero, las distribuciones exhibirían la misma desigualdad, y solamente en ese caso podrían ser coincidentes. Esta es una condición necesaria para que el ajuste sea bueno en términos de distribución, por lo que un valor grande de RS alerta de que la distribución replicada no es una buena aproximación de la real. En el ejemplo, RS es de magnitud muy pequeña, indicando que el ajuste es muy cercano. La descomposición del RS a partir del efecto tipo medio (t/1-t), progresividad (K) y reordenación (D) ofrece una idea de los efectos internos para lograr el ajuste entre la distribución real y la imputada. El valor de t indica el cuánto difieren las medias de la distribución real y la imputada, ya que le media imputada es 1-t veces la media de la distribución de renta real. El efecto reordenación D, también sería de pequeño tamaño. Lo que muestra el error de imputación a nivel de microdato es el valor de K. En el análisis de distribución tradicional, un valor positivo de K indica que la diferencia entre las dos distribuciones se establece de forma progresiva, es decir, que pagan más impuestos las observaciones que ostentan un nivel de renta más elevado. El patrón de progresividad en términos de bondad de las distribuciones no se puede interpretar de esa misma forma, ya que se estaría hablando de un error de imputación sistemático. De hecho, en el análisis tradicional distributivo, la curva de concentración de T debería ir por debajo de la curva de Lorenz de la renta inicial para que el impuesto fuese progresivo. Pero esta no es la interpretación que se dará ahora a K, que puede resultar positivo o negativo, dependiendo de cómo se distribuya el efecto imputación, que en principio es aleatorio.

Si representamos los efectos anteriores en términos de curvas obtenemos el siguiente resultado:

Interpretación de la desigualdad de distribuciones a partir de RS



Nota: Lreal y Cimputada son prácticamente coincidentes con Limputada, por lo que aparentemente no están representadas en el gráfico.

Como se puede comprobar, las curvas de Lorenz de la renta real imputada (ordenada por renta inicial o no), son prácticamente coincidentes, a pesar de que la curva de concentración de la imputación sigue un patrón aleatorio, y resulta en un valor anormalmente grande en comparación con el análisis distributivo tradicional.

Así, una manera adicional de comparar los ajustes es calcular el índice de RS entre la renta real y la imputada, además de la comprobar medias y varianzas, y coincidencia entre curvas de Lorenz y concentración. Un valor elevado de RS indicará que el ajuste no puede ser bueno, mientras que un valor pequeño nos permitirá no descartar la distribución replicada, aunque requiera comprobaciones adicionales.