

---

## INTRODUCCIÓN

---

Las bases de microdatos procedentes de registros administrativos desempeñan un papel cada vez más relevante en el análisis económico aplicado y especialmente en el ámbito de la economía pública, al permitir el acceso a fuentes de información hasta hace relativamente poco tiempo no utilizables y cuyo potencial de explotación es enorme. La captación sistemática de la información a nivel individual mediante procedimientos administrativos implantados en la gestión de los programas de ingresos y gastos de las administraciones pública permite disponer de los datos con regularidad y precisión, y sin tener que incurrir en los importantes costes que acarrea el inevitable trabajo de campo necesario para construir cualquier base de microdatos basada en encuestas. Asimismo, la posibilidad de contar con los valores poblacionales de referencia para las variables incluidas en las bases de microdatos permite realizar diseños muestrales donde los márgenes de error estadístico al inferir resultados poblacionales son muy reducidos. Otra de las grandes ventajas que aporta la utilización de estas fuentes de información es la posibilidad de contar con bases de datos de un tamaño muestral muy grande a un coste de producción estadística bastante reducido. Por último, el desarrollo y tratamiento informático de estos registros administrativos permite incluso la conexión entre distintas bases de datos a partir de la identificación en las mismas de los individuos u hogares a los que se refiere la información registrada.

En el ámbito de los ingresos públicos, en los países anglosajones y nórdicos existe una importante tradición en la utilización de información procedente de fuentes tributarias para la construcción de bases de microdatos. Un antecedente de obligada referencia es el caso de Estados Unidos, donde su agencia federal de administración tributaria, el *Internal Revenue Service (IRS)*, elabora muestras anuales con información individualizada de las declaraciones del impuesto sobre la renta personal desde el año 1916<sup>1</sup>. Inicialmente, la construcción de estas bases de datos estaba motivada por la necesidad de desarrollar análisis de la política tributaria federal desde la propia administración pública, si bien sus propiedades estadísticas las convirtieron más tarde en una fuente de información fundamental en el campo de la investigación socioeconómica<sup>2</sup>.

Esta preocupación por disponer de bases de datos con información individualizada procedente de los registros tributarios se vio reforzada por la gran revolución

---

<sup>1</sup> Esta base de datos pionera fue el *Individual Complete Report File*.

<sup>2</sup> Véase WEBER y BRYANT (2005).

de la informática en la década de los años setenta del pasado siglo. La posibilidad de procesamiento computacional de cantidades enormes de observaciones en tiempos muy reducidos, que brindaba la tecnología de los ordenadores personales y sus programas de análisis estadístico, impulsó decididamente el diseño de este tipo de bases de microdatos. Así, a finales de los setenta, la *Statistics of Income Division (SOI)* del IRS elaboró los primeros paneles de microdatos, aunque limitados en cuanto a las variables incluidas<sup>3</sup>. El primer panel de microdatos completo procedente del registro de las declaraciones del impuesto sobre la renta personal (*Federal Individual Income Tax*) fue elaborado por la SOI para el ejercicio fiscal de 1987 y su horizonte temporal se extendió hasta el ejercicio de 1996. Se trataba de una base de microdatos con estructura de panel expandido, en la que los distintos cortes transversales anuales, extraídos mediante muestreo estratificado, eran representativos de la población declarante del impuesto sobre la renta personal federal norteamericano. Posteriormente, en 2001 vio la luz un nuevo panel que incorpora los ejercicios impositivos desde 1999 hasta 2007, el último disponible.

En Europa, los antecedentes de la creación de bases de microdatos procedentes de registros administrativos tributarios los encontramos en países como Alemania y Suecia. En Alemania, fue una Ley Federal de Estadística de 1981 la que reconoció la necesidad de construir este tipo de bases estadísticas, cuyo ejemplo más representativo es el «Income tax scientific-use-file 1995» que elabora el *Research Data Center (RDC)* de la *Federal Statistical Office (Destatis)* con la colaboración de las oficinas estadísticas de los Laenders desde 1995 con una periodicidad trienal<sup>4</sup>. Las muestras incluyen un 10% de las declaraciones del impuesto federal sobre la renta personal alemán (unos 3 millones de observaciones muestrales)<sup>5</sup>. Los países nórdicos también cuentan con una amplia tradición en la construcción de bases de microdatos con información de origen tributario, destacando los casos de Suecia y Dinamarca, donde en algunos casos estas muestras son enlazadas con otras bases de microdatos con información sobre prestaciones sociales o contratos laborales.

En el campo de la construcción de bases de microdatos de origen tributario puede decirse que España no ha permanecido ajena a esta corriente liderada por Estados Unidos y los países del norte y centro de Europa. En el ámbito de la imposición sobre la renta de las personas físicas, el Instituto de Estudios Fiscales, organismo dependiente del Ministerio de Economía y Hacienda abordó a finales de la década de los años ochenta el primer Panel de Declarantes de IRPF, que inicialmente abarcó los ejercicios impositivos comprendidos entre 1982 y 1994, siendo posteriormente prolongado, en 2006, hasta el ejercicio de 1998<sup>6</sup>. Se trataba de un panel expandido,

---

<sup>3</sup> Inicialmente estas bases de microdatos incluían información enlazada de la *Continuous Work History Sample (CWHS)* y de las *Individual Income Tax Return Cross Sectional Sample* anuales. Ver WEBER (2004).

<sup>4</sup> Véase IZA (2009).

<sup>5</sup> Véase MERZ *et al.* (2005).

<sup>6</sup> Véase AYALA *et al.* (2004) y ONRUBIA *et al.* (2007).

representativo de la población declarante por el IRPF en cada año, y construido mediante extracción aleatoria proporcional (2% de las declaraciones presentadas por los contribuyentes de este impuesto ante la administración tributaria). En la primera mitad de los años noventa, fruto de la colaboración entre el Instituto de Estudios Fiscales y la Agencia Estatal de Administración Tributaria también se construyó un panel con microdatos correspondientes a las declaraciones presentadas por el Impuesto de Sociedades, que comprendía los ejercicios de 1991 a 1994<sup>7</sup>.

La reconocida utilidad de este tipo de bases de datos para el análisis económico aplicado, bien en su explotación directa, bien a través de su fusión estadística con otras bases de microdatos procedentes de encuestas o de otros registros administrativos, impulsó el interés del Ministerio de Economía y Hacienda por continuar, incluso ampliar esta oferta<sup>8</sup>. Así, en la segunda mitad de esta década y nuevamente fruto de la colaboración entre el Instituto de Estudios Fiscales y la Agencia Estatal de Administración Tributaria (AEAT), ha visto la luz una nueva serie de muestras anuales de microdatos de IRPF, iniciada con el ejercicio de 2002 y que hasta el momento incluye los ejercicios comprendidos hasta 2006<sup>9</sup>.

La experiencia obtenida por investigadores y analistas en la utilización del Panel de Declarantes de IRPF 1982-1998, junto con el auge creciente de los trabajos internacionales que utilizan bases de microdatos fiscales con estructura de panel, ponían de manifiesto la necesidad de dar continuidad al proyecto iniciado en el IEF hace veinte años. La potencialidad de las muestras anuales de declarantes del IRPF es muy grande, pero su carácter de muestras transversales (*cross-section samples*) limitan la realización de determinados análisis longitudinales que requieren la disposición de muestras de panel puro. Esta realidad ha llevado al Instituto de Estudios Fiscales, con la inestimable ayuda del Servicio de Estudios Tributarios y Estadísticas de la AEAT, a diseñar y construir un nuevo Panel de Declarantes de IRPF que cubra el periodo que se inicia en 1999 —año de entrada en vigor de la Ley 40/1998, que reformó con profundidad el IRPF español— y que abarque hasta el último ejercicio para el que se dispone de microdatos, 2007.

Los casi veinte años transcurridos desde el diseño del anterior Panel de IRPF y el momento en que se comenzó el proceso de construcción del nuevo Panel han permitido incorporar mejoras en el nuevo diseño estadístico. Asimismo, la generalización de la confección automática de las declaraciones de IRPF a partir de la implantación del programa de ayuda PADRE, junto con los avances informáticos en el registro de las declaraciones por parte de la AEAT, han permitido también contar con una fuente de información poblacional mucho más fiable y consistente en cuanto a la existencia de errores de cumplimentación que la utilizada en los ejercicios del an-

---

<sup>7</sup> Véase ROMERO (2001).

<sup>8</sup> Sobre la fusión de muestras de microdatos tributarios de IRPF con bases de microdatos procedentes de encuestas puede verse PICOS (2006).

<sup>9</sup> Véase PEREZ y DIAZ DE SARRALDE (2009).

tigo Panel, en cuyos primeros años, la informatización de la administración tributaria estaba en su fase inicial. La utilidad de contar con una base de datos como el nuevo Panel de IRPF ha sido reconocida por el Consejo Superior de Estadística, lo que ha supuesto la inclusión del mismo como una Operación Estadística del Plan Estadístico Nacional. Este hecho supone dotar al proyecto de una continuidad en el tiempo, además de la obligación de realizar la difusión de acuerdo a los compromisos adquiridos.

El interés del Instituto de Estudios Fiscales por divulgar con la máxima amplitud el nuevo Panel de IRPF, especialmente entre los analistas socioeconómicos y los investigadores, está en el origen de la elaboración de este libro. Los principios de transparencia y generalización en el acceso a la información estadística han presidido desde los momentos iniciales la realización del proyecto. Por ello, se ha sido consciente de la necesidad de contar con una publicación que incorporase de forma detallada las especificaciones técnicas del diseño estadístico realizado, así como una descripción lo más completa posible de su estructura y contenido informativo. También es indudable que la amplia utilización de esta base de microdatos, menos conocida que las procedentes de las encuestas principales elaboradas por el Instituto Nacional de Estadística (INE), depende en gran medida de la difusión de sus potencialidades y de la clarificación de cómo llevar a cabo su explotación, al menos, en un nivel básico del análisis. Por ello, este libro incluye también una guía de utilización del Panel que cubre distintos aspectos, como la estimación poblacional de la población de referencia, el cálculo de magnitudes agregadas a partir de los microdatos, los cálculos poblacionales de las medidas estadísticas fundamentales y el cálculo de medidas de distribución habituales en la utilización de este tipo de bases de microdatos. La posibilidad de vincular las declaraciones de IRPF de los cónyuges que, de acuerdo con la legislación del IRPF español, optan por tributar separadamente dentro del matrimonio, abre una interesante aproximación de estudio, cambiando la unidad observacional de análisis del declarante al hogar fiscal. Esta guía de utilización aborda esta cuestión, lo mismo que la forma de trabajar longitudinalmente con las muestras, empleando la posibilidad de seguir ejercicio a ejercicio tanto a los declarantes del impuesto como a los hogares fiscales en los que éstos se encuadran.

En cuanto a su contenido, el libro se organiza de la siguiente manera. Tras esta introducción, el primer capítulo está dedicado a presentar el diseño estadístico del Panel de IRPF, justificando la elección realizada en cuanto al tipo de base de datos y la estructura de sus ficheros. En este capítulo también se describe el proceso de extracción de las muestras tanto en el año base, el ejercicio 2003, como en el resto de años incluidos, tanto retrospectivamente (de 1999 a 2002) como prospectivamente (de 2004 hasta 2007). El segundo capítulo de la obra recoge las características básicas del Panel, describiendo el contenido de los ficheros y sus variables, tanto monetarias como no monetarias. También se incluye en el mismo la definición del hogar fiscal como unidad observacional alternativa a la declaración, que fue la unidad de muestreo utilizada en la construcción del Panel. El interés por construir una base

de microdatos lo más rica posible en cuanto a su potencial de explotación ha llevado a los autores del proyecto a incorporar una serie de variables adicionales construidas a partir de la información original aportada por la AEAT en sus declaraciones. El tercer capítulo describe la creación de estas variables, así como el procedimiento diseñado para facilitar el seguimiento longitudinal, con «entradas» y «salidas» tanto de declarantes como de hogares fiscales. El capítulo cuarto del libro tiene por objeto ofrecer un análisis de la representatividad de las muestras del Panel. Para ello se analiza su capacidad de inferencia poblacional para las principales variables económicas del IRPF. Este capítulo incorpora además un estudio sobre el «desgaste» (*attrition*) del Panel, así como un análisis detallado de las anomalías detectadas en la identificación de algunas observaciones originales, especialmente referidas a las relaciones entre declarantes y en la conformación de los hogares. Junto con la guía de utilización del Panel antes reseñada, el capítulo quinto incluye también la presentación de la estructura final de los ficheros y de las variables incluidas en los mismos, con los valores que pueden tomar éstas.