BIG DATA Y GESTIÓN DEL DATO EN LA ADMINISTRACIÓN TRIBUTARIA: PASADO, PRESENTE Y FUTURO

Mª LUZ GÓMEZ LÓPEZ

Servicio de Estudios Tributarios y Estadísticas AEAT

Cuando recibí el encargo o posibilidad de escribir este artículo me planteé varias preguntas; la primera, ¿era yo la persona más adecuada para hablar de Big data e Inteligencia Artificial? Sobre esta cuestión y con sinceridad la respuesta es que «no». Sin embargo, me hice otra pregunta, ¿se puede considerar que hacemos tratamientos de grandes volúmenes de datos, con procesamientos de los mismos en tiempos muy reducidos, y obteniendo valor añadido de los datos en virtud de esos tratamientos estadísticos, presentando los resultados al usuario de una manera muy intuitiva y accesible?, y a esta pregunta la respuesta era afirmativa. Y esta es la razón fundamental por la que me encuentro escribiendo este artículo.

Antes de ponerme a la tarea, he hecho un cierto acercamiento a las definiciones más conocidas y citadas de Big Data atribuidas a Gartner y O'Reilly, porque respecto a este concepto existen múltiples interpretaciones, y algunos usos del término claramente erróneos.

El primer autor define Big Data como información de mucho volumen procesada a gran velocidad y muy variada que requiere sistemas de información innovadores y efectivos para poder facilitar la obtención de conocimiento y la toma de decisiones (Gartner; 2012).

El segundo de los autores va más allá, porque define Big Data como los datos que sobrepasan la capacidad de procesamiento de las bases de datos tradicionales. La particularidad de los datos y la necesidad de procesamientos novedosos distingue para este autor el concepto. Respecto a los datos hace referencia a dos características: los datos se mueven demasiado rápido o éstos no cuadran en la arquitectura tradicional de bases de datos (datos no estructurados). En consecuencia, para obtener valor de estos datos debe buscarse un modo alternativo para procesarlos (O'Reilly; 2012).

En ambas definiciones, si bien hay notables diferencias, hay algunos puntos comunes: Volumen de datos, Velocidad de procesamiento del dato (en ocasiones en tiempo real) y Variedad, que debe entenderse como la capacidad de tratar fuentes de datos diversas e incluso en distintos formatos, incluyendo tanto datos estructurados, como no estructurados.

De esas definiciones Gartner pone el énfasis en el volumen de datos y su capacidad de procesamiento; mientras que O'Reilly enfatiza más en la forma de procesarlos, y en la extracción de valor a datos no estructurados. En cualquier caso, las organizaciones

tanto empresas como administraciones han realizado fuertes inversiones en infraestructura (almacenamiento, transmisión y transformación de datos) y en el uso de lenguajes de programación innovadores para realizar explotaciones o analítica de esos grandes volúmenes de datos almacenados de una forma más eficiente, Phyton, Hadoop y R.

Tras este somero análisis se llega a la conclusión que la Agencia Estatal de la Administración Tributaria posiblemente sea la administración pública en la que encajen mejor ambas definiciones de esos autores, y las características mencionadas en los párrafos anteriores.

Sin embargo, como estadístico y sin minusvalorar los aspectos anteriores, me parece imprescindible resaltar que el dato es lo más importante, o dicho de otro modo, disponer de datos en gran cantidad y calidad, y por esta razón voy a empezar el artículo analizando la importancia del «dato» desde una perspectiva global, dado que la misma resulta perfectamente trasladable al dato obtenido de un registro administrativo. En la sociedad actual el dato se ha convertido en una unidad de negocio alrededor del cual se mueven multitud de nuevas empresas e instituciones públicas y, a la par, se ha convertido al científico de datos y a los informáticos en los puestos más demandados en las ofertas de empleo. Y es que hoy en día la inferencia, entendiendo ésta como la información que se puede extraer a partir del tratamiento de los datos, proporciona utilidades muy valoradas en el mundo empresarial. En muchas ocasiones las empresas dedicadas a estas actividades son buscadoras de datos, los recopilan de distintas fuentes, tanto públicos como privados y los tratan técnicamente para proporcionar resultados que son servicios demandados por otras empresas especialmente dedicadas a realizar estudios, la comunidad académica en su vertiente investigadora y la sociedad en su conjunto. El beneficio obtenido de la analítica de los datos es evidente, y tanto las operadoras de telefonía, como las entidades financieras y muchas grandes empresas de otras actividades crean unidades de negocio para el tratamiento de los datos de sus clientes, obtener sus preferencias, clasificarlos, y de esa manera ir conociendo más a sus clientes o potenciales clientes para ofrecerles productos más a la medida de los perfiles diseñados. En todo este proceso evolutivo de nuestra economía, la tecnología juega un papel fundamental, y no solo ha producido grandes avances en distintos sectores de la actividad, sino que ha venido para quedarse y revolucionar la ciencia de datos. Las propias unidades estadísticas se plantean por qué conformarse con hacer inferencia o simulación a partir de una muestra de datos, si se puede hacer sobre la totalidad de los datos disponibles y así, evitar o minorar los errores de muestreo. Y justamente ese es el paradigma en el que estamos ahora y en el que se mueve la estadística oficial, la tributaria y en general toda la Administración pública en su conjunto.

EL DATO EN LA ADMINISTRACIÓN PÚBLICA Y ADMINISTRACIÓN ELECTRÓNICA

Este planteamiento global de la importancia del dato válido para las empresas es igualmente válido para las administraciones y, de hecho, ya hay una corriente general favorable a utilizar los registros administrativos de manera masiva en las principales y más costosas estadísticas oficiales, así como en investigaciones académicas y científicas. Sin embargo, en la administración la revolución tecnológica y el cambio del paradigma en el uso de los datos ha llegado con distinta velocidad de implantación, si bien la normativa nacional y la comunitaria han impulsado de forma decidida esta revolución digital de las distintas administraciones. En esa velocidad de implantación tecnológica y de apuesta decidida a relacionarse con los agentes económicos, sujetos pasivos de los impuestos, colaboradores o entidades financieras, por vía telemática, ha sido pionera la Agencia Estatal de la Administración Tributaria, aunque será más adelante cuando se presentará de una forma más pormenorizada la evolución natural y las apuestas tecnológicas llevadas a cabo por esta Administración.

La tradición estadística, hasta hace pocos años, consistía en estudiar la población para obtener la mínima muestra posible con la estratificación necesaria para obtener el mejor estimador con el menor error de muestreo. El estudio de la población, el canal de recogida en el trabajo de campo, eran decisivos también para seleccionar un tipo de muestreo u otro: muestreo aleatorio simple, estratificado, bietápico, por conglomerados, etc.

Pero la estadística oficial ha probado que en muchos casos se pueden reducir los tamaños de las muestras con los enormes ahorros de costes que supone y sin pérdida de calidad recurriendo a los registros administrativos. Registros administrativos que han llegado a tener unas propiedades excepcionales: son exhaustivos o casi censales, sus datos están estructurados, los datos suelen tener una calidad relevante, tanto más cuanto más ligados estén al procedimiento que ha generado la necesidad del registro administrativo, y además gozan de una puntualidad en su recepción derivada del hecho de que los plazos en los procedimientos administrativos están perfectamente delimitados.

El impulso de la Administración electrónica vino de la aprobación de la Ley 11/2007 que promovía la modernización de la Administración española al consagrar el derecho de los ciudadanos a relacionarse electrónicamente con la Administración. Si bien era un derecho para el ciudadano, esta norma constituía una obligación correlativa para la Administración, que resultaba obligada a introducir los canales electrónicos en las distintas fases de los procedimientos administrativos que implicarán una interacción con el ciudadano, y que puede considerarse el punto de partida para disponer de forma creciente de datos de los ciudadanos de forma estructurada, y a su vez, el inicio del destierro paulatino del formato papel en los formularios de los distintos procedimientos administrativos.

No obstante, hay que destacar que la AEAT (Agencia Estatal de la Administración Tributaria) era una administración pionera porque se había anticipado a esta norma, y había iniciado la senda de la digitalización muchos años antes. Aunque la presentación telemática y obligatoria de las declaraciones estaba limitada solo a algunos procedimientos y a la relación con pocos contribuyentes (grandes empresas), sin embargo, el proceso de digitalización no se quedaba en la entrada de datos, porque toda la información se convertía en datos estructurados y eran grabados en su totalidad porque siempre estuvo claro que había que optimizar la utilización de los datos. También es cierto, que en esos años el uso estadístico de esa inagotable fuente de datos era residual, porque el objetivo de los mismos no era otro que su gestión, elaboración de filtros de inconsistencias y selección en aras de actuaciones gestoras e inspectoras, y por supuesto, ir construyendo un repositorio de información de todos los contribuyentes con toda la información disponible, tanto la directa aportada por el propio contribuyente, como tan importante como la primera, la imputada por terceros y en la que se basa el contraste de información fiscal.

En años más recientes, la Directiva 2019/1024/CE pone en valor los datos administrativos. Esta directiva supranacional impulsa la necesidad y obligación de las administraciones de proporcionar «datos abiertos y reutilizables» procedentes de la información recabada por el sector público y define los «conjuntos de datos de alto valor», entre los que relaciona los siguientes: geoespaciales, meteorológicos, observación del territorio y el medioambiente, movilidad, estadísticos y empresariales. Estos datos tienen que ser gratuitos (con algunas excepciones), estar en formato digital vía API y permitir su descarga masiva. En el conjunto de datos de alto valor, sin ánimo de exhaustividad, figuran: códigos postales, unidades administrativas territoriales, direcciones, datos catastrales, datos de agricultura, consumo de energía, imágenes de satélite, meteorológicos, sensores, indicadores demográficos y económicos (datos estadísticos de la AEAT, INE y Seguridad Social), registros de empresas, tráfico, vías de comunicación, etc.…

Respecto a la simple digitalización y almacenamiento del dato, las administraciones han implementado procedimientos estadísticos que han irrumpido de forma relevante en la gestión y en el tratamiento de los datos y las técnicas estadísticas se han introducido en las propias herramientas de selección, al incorporar utilidades de presentación y cálculo de estadísticos básicos y, al menos en el ámbito de la administración tributaria, también se ha incorporado un curso de estadística descriptiva en el Plan de formación continua con el fin de que el uso de esos estadísticos sea adecuadamente interpretados por los usuarios internos de la administración tributaria.

La incorporación de una unidad de estadística en la AEAT se remonta al año 1992, en el mismo momento de la creación de la Agencia Tributaria, pero en su inicio no tenía la función de realizar estadísticas, porque esa tarea había quedado en el Instituto de Estudios Fiscales (IEF) en esos inicios. Sólo años más tarde esa competencia fue adquirida por la AEAT. En ese primer momento, el cometido de la unidad era la creación de simuladores de las distintas figuras tributarias para evaluar el coste de las distintas políticas fiscales, sus efectos redistributivos, progresividad, y delimitar el conjunto de beneficiados y perjudicados de cada medida concreta, así como el coste aparejado a cada iniciativa. El embrión del simulador de Renta también había surgido en el IEF, pero se consolidó y mejoró en la AEAT. Adicionalmente se construyeron simuladores de Sociedades, IVA, Patrimonio y Retenciones. Aunque tradicionalmente la estadística descriptiva antecede a la estadística predictiva, en la AEAT se siguió el sentido inverso. Sin embargo, la publicación estadística jugo un papel relevante en la transparencia exigida a una administración y desde luego, fue mucho más divulgada que el de la predicción e impacto de reformas fiscales que quedo siempre opacada en un conocimiento interno en la Administración Fiscal. En esos primeros años se firmó el primer Convenio con el INE para la utilización de los datos fiscales para los marcos poblacionales de empresas, que más tarde sería conocido como el DIRCE (Directorio Central de Empresas).

En mi opinión, la creación de unidades estadísticas en los organismos que disponen de registros administrativos es imprescindible porque la utilización en primera instancia de los datos en estas unidades sirve para la mejora continua de la información tratada, introduciendo una perspectiva de calidad y una funcionalidad estadística necesaria para reportar los datos a terceros, es decir, a organismos o particulares que pueden no tener conocimiento previo de las particularidades de la información contenida en los registros que pretende utilizar.

Esto es muy relevante y en el entorno de la Estadística tributaria quiero detenerme en un dato poco o nada conocido y es que la producción estadística inventariada de la AEAT alcanza la cifra de 27 operaciones estadísticas, la mayoría de ellas incluidas en el Plan Estadístico Nacional y además se colabora en 44 operaciones estadísticas de las 168 inventariadas por el INE.

Para hacernos una idea de lo que supone la explotación estadística de las principales fuentes de datos más conocidas. Del IRPF se producen 5 estadísticas anuales a partir de más de 20 millones de declaraciones por año, de la información de rentas de trabajo, pensiones y prestaciones por desempleo se obtiene una estadística a partir de más de 35 millones de registros por año. Las estadísticas que parten de las declaraciones de Sociedades tratan 1,5 millones de declaraciones y más de 4000 partidas por registro /año. Las declaraciones del Impuesto sobre el valor añadido tratan 3,2 millones de declaraciones, 60.000 de ellas generadas informáticamente a partir del tratamiento de 2.500 millones de registros por año (facturas emitidas y recibidas en el Sistema Inmediato de Información). Aparte destacar que también se explotan otros modelos tributarios de menor dimensión para producir, por

ejemplo, estadísticas de patrimonio, de matriculación de vehículos, y otras diversas que quedan en el ámbito interno de la Agencia porque parte de información que se considera más sensible.

Para tener una medida aproximada de la magnitud de los suministros anuales de información al Instituto Nacional de Estadística para la elaboración de sus operaciones estadísticas inventariadas, destacar de mayor a menor la colaboración con los Censos: el de población con más de 46 millones de personas físicas y las relaciones familiares de los integrantes del hogar, Directorio Central de empresas con 4,3 millones de registros en el último año, Censo Agrario con más de 1 millón de registros con actividad agrícola, Proyecto Atlas (ERGEO) estadística experimental con información completa de todos los modelos informativos de renta y la declaración de IRPF de todas las personas físicas de territorio de régimen fiscal común publicada a nivel de sección Censal, datos de rentas salariales para las Encuestas de Población Activa y de Costes Salariales obtenidas de las declaraciones de los empresarios pagadores de retribuciones salariales, las encuestas Industrial y de Servicios que utilizan información combinada procedente de las actividades empresariales declaradas por las personas físicas y las sociedades en sus declaraciones respectivas, completadas con la información procedente del Impuesto sobre el Valor Añadido y del modelo de atribución por fuentes de renta.

Por supuesto a lo anterior, hay que añadir la colaboración para la realización del módulo fiscal de la Muestra Continua de Vidas Laborales de la Seguridad Social, las Muestras de IRPF para investigadores que se solicita en el portal del Instituto de Estudios Fiscales y las Muestras de Patrimonio y el Anuario Estadístico accesible en la Web de la AEAT.

En la actualidad se facilitan datos individuales anonimizados a muchos peticionarios, y se proporciona información en muestras para el uso de investigadores; y esta tendencia sigue aumentando y consolidando la producción de estos servicios desde hace muchos años. Entre los suministros de datos individuales anonimizados al alcance de los investigadores, cabe citar las Muestras de IRPF (modelo 100), las de Retenciones a cuenta de IRPF (modelo 190), las muestras de Patrimonio (modelo 714), la Muestra continua de vidas laborales con el módulo fiscal y la muestra de microdatos de la Encuesta de Población Activa con datos fiscales.

Hay que destacar también, que todos los suministros de información individualizada identificada están amparados por Reglamentos Comunitarios que posibilitan la utilización plena de los registros administrativos para algunos fines, por ejemplo, censales u operaciones estadísticas oficiales de obligado reporte a Eurostat. Igualmente se encuentran amparados en el artículo 95 de la Ley General Tributaria algunas cesiones individuales para gestión de procedimientos de terceros validados por una autorización del peticionario para la consulta de sus datos fiscales, ejemplo de este tipo de suministros son datos para becas, para solicitar el Ingreso Mínimo Vital, y para acceder a algunos servicios públicos.

Respecto a producción de estadísticas propia, desde el primer momento se abordó exclusivamente en formato digital por la actual Subdirección Gral. de Estadísticas, entonces ubicada en el Departamento de Informática Tributaria, pero hay que destacar que hasta el 2008 no se creó el Servicio de Estudios Tributarios y Estadísticas que incrementó las competencias y que impulsó aún más la producción estadística propia y la colaboración mediante el suministro de datos a otras oficinas estadísticas de la Administración. No obstante, no hubiera sido posible generar esta cadena de valor estadístico sin la decidida implicación de la Dirección por la digitalización de la información fiscal y la implementación de métodos estadísticos y de contraste para mejorar sustancialmente la calidad de los datos; en resumen, para llevar a cabo una estrategia de Tecnologías, Información y Comunicación, a la que me refiero en el siguiente apartado.

LOS PRINCIPIOS RECTORES DE LA ESTRATEGIA TIC DEL DATO EN LA ADMINISTRACIÓN TRIBUTARIA

Vivimos en una sociedad altamente datificada, nuestros datos van quedando en la red, dejando huella de nuestros gustos, aficiones, ubicación y búsquedas en la red, datos que en sí mismos son de alto valor para las empresas. Sin embargo, los datos que las administraciones disponen de los ciudadanos son muchos más y seguramente más valiosos, aunque con la fuerte restricción de que éstos no permiten su comercialización, aunque sí su uso, para proporcionar información estructurada en forma de estadísticas a disposición del público, y también, como no puede ser de otro modo, para colaborar con otras oficinas de estadística en su producción propia.

La evolución en el número de datos disponibles anualmente y la mejora en la calidad del dato hace de los registros administrativos una mina para la ciencia de datos, y por supuesto, tiene una utilidad creciente como campo de pruebas de nuevas herramientas tecnológicas que optimicen el tratamiento de ese ingente volumen de datos. El grado de madurez que el tratamiento de los datos ha experimentado en la AEAT desde su creación no ha sido casual. Durante ese horizonte temporal la AEAT ha apostado su estrategia en seis ejes:

Digitalización: la AEAT apostó por el canal digital como preferente para establecer la relación con los ciudadanos y con las empresas, ampliando cada vez en mayor medida el canal de presentación telemática de las declaraciones en detrimento de las declaraciones en papel. Este cambio por sí solo permitió implantar una gestión inteligente de la información y de los datos integrados en su conjunto, así como una reducción muy significativa de los procedimientos gestores, convirtiendo la tecnología en el recurso básico para

que los empleados públicos desempeñen *sus labores con* eficiencia y eficacia, promoviendo recursos para la innovación y la ampliación de los servicios.

Servicios: en los que se persigue la mejora continua en la calidad del desarrollo y en la entrega de los servicios a los usuarios, facilitando el cumplimiento voluntario y desarrollando herramientas cada vez más sofisticadas para proporcionar pre-borradores o datos fiscales, o ambas utilidades, y garantizando una disponibilidad permanente a los servicios y al acceso a la información requerida para el procedimiento gestor.

También se han firmado diversidad de Convenios con colaboradores y otros organismos de la administración para gestionar prestaciones sociales (becas, Ingreso Mínimo Vital, prestaciones de Entidades Locales y Comunidades Autónomas), y se ha adquirido un rol relevante en la función de proporcionar datos económicos y demográficos a la estadística oficial (INE e Institutos de Estadística de las Comunidades Autónomas)

- *Tecnología*: la evolución tecnológica en los últimos años ha generado tratamientos y tiempos de respuesta impensables hace tan solo una década y la AEAT nunca ha estado al margen de esa evolución, utilizando diversidad de tecnologías adecuadas a cada servicio concreto y teniendo en cuenta las necesidades para su implementación y capacidad para cumplir los objetivos concretos marcados

- *Personal*: en este terreno, y dadas las necesidades de formación requeridas para atender los servicios y las tecnologías empleadas, la AEAT es una institución destacada en la integración de personal con perfiles diversos, de forma tal que, junto a las áreas funcionales de la organización representada por los cuerpos de Hacienda, es también relevante la presencia de otros cuerpos especializados, informáticos, cuerpos de Estadística, Administradores Cíviles, Abogados del Estado, etc. Desde el año 2008 se dispone de un Servicio de Estudios y Estadísticas que constituye una unidad muy pequeña pero que representa claramente esta mezcla de perfiles profesionales con distintas habilidades y competencias para el desarrollo de sus funciones.

- *Calidad del dato*: la búsqueda de la calidad ha venido de la mano de la prestación de ayuda a la cumplimentación de las declaraciones. El primer hito y quizás el más conocido es el borrador de renta y el suministro de datos fiscales; la mejora de la calidad de los modelos informativos proporcionados por terceros, con el control de la identificación y de la calidad de los datos proporcionados en estos, lo ha hecho posible. Pero también esta organización ha introducido un mecanismo para disponer de las facturaciones de las grandes empresas en el denominado Suministro Inmediato de Información y en sus antecesores, los formularios 340 y 347 de Ventas y Compras declaradas e imputadas por terceros.

En resumen, los avances llevados a cabo han propiciado un incremento espectacular del volumen de datos, una mejora de calidad basada en cruces y algoritmos de cálculo y una integración entre los diferentes procesos y sistemas de información que ha revolucionado los mecanismos de selección y de evaluación de inconsistencias en la propia organización y en el tratamiento de errores de cumplimentación de las declaraciones. La implicación de la Dirección en el proceso y el desarrollo de sistemas de consulta de fácil utilización por los usuarios ha proporcionado sofisticados procedimientos de selección y análisis a los usuarios internos, incluyendo el aprendizaje de las herramientas en la formación continua de la organización.

Otro reto relevante es la integración en el sistema de información de la AEAT de información externa, ejemplos de ello, son la información catastral, determinadas operaciones societarias proporcionada por los notarios, intercambios de información fiscal con las Haciendas Forales, intercambio de información con otras administraciones tributarias (Country by Country), y con entidades financieras extranjeras respecto a los contribuyentes de nuestro país (CRS), etc.

En este sentido, la tecnología ha facilitado la utilización de consultas sofisticadas y su uso se ha puesto a disposición de usuarios internos y externos, mejorando a su vez los procesos de control; la integración automática de información del sujeto con fuentes internas y externas que permite obtener información ampliada de los contribuyentes, sofisticando y depurando aún más si cabe, el tratamiento estadístico de la información, multiplicando sus finalidades, sin abandonar la necesidad de datos más exactos y relevantes para la selección y el control.

En esta evolución tampoco se ha abandonado la parametrización de estados de cumplimiento de las obligaciones tributarias de los contribuyentes y se ha mejorado sustancialmente la clasificación sectorial de las empresas y su calificación censal, lo que ha posibilitado la elaboración de indicadores por actividad y la incorporación de análisis clúster, inferencia estadística y demás procesos que permiten introducir criterios de control adicionales.

La mejora continua, la digitalización de toda la información y la puesta a disposición de los usuarios de la mayoría de los procedimientos en sede electrónica, así como la optimización de los procedimientos internos, ha hecho posible la integración de casi el 100% de la plantilla en situación de teletrabajo por el Covid-19 y esta nueva realidad, no exenta de problemas en su etapa inicial, ha permitido mantener el estándar de servicios que ofrece la AEAT.

En aras de la transparencia y la colaboración con otras administraciones, la producción estadística basada en distintas fuentes fiscales está muy consolidada, al igual que el suministro de datos para más de 40 operaciones estadísticas del Plan Estadístico Nacional producidas por otras instituciones. La digitalización completa del dato constituye un esfuerzo de la Administración

Tributaria para garantizar la transparencia y el uso eficiente de los datos y constituye por sí mismo un objetivo reconocido en la transformación de la sociedad que se ha acostumbrado al cumplimiento de sus obligaciones fiscales mediante presentaciones telemáticas con programas de ayuda o pre-borradores en su cumplimentación. En dicha transformación la mejora continua y la tecnología han jugado un papel crucial en la interacción de la Administración con los contribuyentes.

PRESENTE Y FUTURO

En la actualidad, una vez superada la etapa anterior, la AEAT avanza en la aplicación de inteligencia artificial (IA), y particularmente en el aprendizaje de la máquina (machine learning) para resolver problemas complejos buscando la solución óptima, programas de entrenamiento para la realización de inferencia estadística y otras utilidades que están en fase experimental, pero que a buen seguro en el futuro proporcionarán vías alternativas para mejorar las estadísticas. Por ejemplo, incorporando volumen de ingresos y gastos a los empresarios en Estimación objetiva no agraria (EONA), y en otros usos, como vía para proporcionar más elementos de evaluación útiles para los actuarios que intervienen en la selección y control de contribuyentes, mejorando la gestión de riesgo de fraude y la priorización de las actuaciones.

Estadísticamente tenemos las herramientas, disponemos de datos y también contamos con la tecnología adecuada para efectuar múltiples análisis y complementar la información escasa o deficiente. No obstante, hay que ser realistas, en muchas circunstancias y desde una perspectiva puramente estadística existen distintos algoritmos de aproximación y ello puede conducir a soluciones distintas y estas prácticas, que en el plano estadístico tienen múltiples aplicaciones y están plenamente aceptadas y consolidadas, generan ciertas incertidumbres en el entorno fiscal. Para exponer con más claridad la afirmación anterior, por ejemplo, la imputación de un dato a efectos estadísticos cuando se aprecia un error flagrante es habitual y no tiene mayor trascendencia con el objetivo de mejorar el dato agregado y sin repercusión directa para el contribuyente. Sin embargo, el planteamiento es distinto cuando esa corrección automática se realiza sobre una declaración y tiene repercusión fiscal, claramente la Administración no puede practicar liquidaciones tributarias con distintos resultados, por lo que es preciso realizar correcciones pero con un dato cierto y real, y si eso no es posible, la Administración tiene que dotarse de los procedimientos o algoritmos de imputación reglamentados que permitan la sustitución del dato en la declaración con la repercusión tributaria que ello conlleve y este camino si bien se ha iniciado, todavía requerirá de un tiempo para normalizarlo en la sociedad.

Es muy reciente el envío de cartas informativas a los empresarios informando del alejamiento de los valores de los indicadores de su actividad en relación a las medias estadísticas obtenidas por la Administración, si bien, estas comunicaciones no conllevan de momento ninguna actuación de control. Por otro lado, la selección de magnitudes a comunicar tampoco es completa, sino que es una selección de los ratios más conocidos.

Las restricciones impuestas a la utilización de los datos tributarios previstas en la Ley General Tributaria no ha permitido extender la colaboración público –privada para aplicar beneficios sociales (ejemplos bono_social electricidad, teléfono, etc.) sin que el beneficiario tenga que solicitarlo; tampoco se han habilitado usos como poner en conocimiento de los usuarios los posibles beneficios sociales que les resultarían de aplicación en función de sus rentas y circunstancias personales y familiares conocidas. De hecho, en materia de colaboración social, la información tributaria tendría mucho más recorrido.

Sin embargo, desde una finalidad puramente estadística se están realizando grandes avances, utilizando intensivamente muestras de empresarios clasificados en una determinada actividad en estimación directa simplificada, introduciendo en nuestra selección los criterios objetivos de los empresarios que tributan en renta en Estimación Objetiva por coeficientes No Agrarios (EONA). Esa selección de empresarios personas físicas en estimación directa constituye nuestra selección de entrenamiento con el fin de que la herramienta sea capaz de imputar por similitud ingresos y gastos consistentes con el rendimiento del grupo de empresarios (EONA) del que no disponemos de esos elementos. Hay que ser conscientes de que estos mecanismos difícilmente darán una solución única, pero sí que se constata que la delimitación de la semejanza de las unidades de entrenamiento permite iterar soluciones que proporcionan aproximaciones estadísticas muy eficientes. Posiblemente estamos lejos de que estos mecanismos de inferencia puedan ser utilizados jurídicamente para imputar deuda tributaria, pero todo ello puede reportar avances significativos, y cambios en los usos y elementos de apoyo para levantar actuaciones inspectoras; y por supuesto, en el corto plazo redundara en el aprovechamiento estadístico de estas fuentes de datos derivados de esas inferencias automáticas.

Otro elemento novedoso en el que la AEAT se ha implicado en estas nuevas tecnologías es la puesta a disposición del público de los Asistentes Virtuales, basados en árboles de decisión y diálogos. Ejemplos de ellos ya de uso público serían: el Asistente virtual IVA, el Asistente de IVA Comercio Exterior y el Asistente virtual Censal. Estas herramientas de atención automática resuelven un porcentaje muy elevado de dudas que anteriormente tenían que ser solventadas por el personal de gestión y son capaces de responder a un gran número de cuestiones, al menos las que más frecuentemente han sido consultadas por los contribuyentes, y adicionalmente, las dudas no resueltas por el Asistente Virtual se reportan a través de un formulario por correo electrónico que será objeto de análisis por un experto y quedará en el repositorio de soluciones para futuras interacciones, lo que produce una continua mejora y actualización de la herramienta.

Otra herramienta que es utilizada cotidianamente en la AEAT por un volumen creciente de su personal se remonta a principios del año 2000, y se trata de una aplicación basada en minería de datos, la filosofía de la herramienta es maximizar la utilización de todos los datos del sistema de información tributaria (de todos los modelos y varios ejercicios), permitiendo cruces entre distintas bases de datos, combinando información, filtrando la misma, generando expresiones nuevas, que a su vez permiten su utilización como filtros y generando repositorios de consultas que pueden ser utilizadas de forma privada, pública, con los usuarios con permisos a esa fuente de información, o compartirla con un grupo de usuarios. Incorpora sobre las variables seleccionadas estadísticos básicos descriptivos y tramificaciones en decilas y centilas. La herramienta permite el acceso a datos identificados (con auditoria) o a datos anonimizados a efectos analíticos, y realmente es la herramienta estrella de la AEAT en los años más recientes.

Dentro del mismo proyecto se cuenta con una versión de relaciones, que permite establecer relaciones familiares y empresariales entre los distintos sujetos identificados y que es de mucha utilidad para la investigación de tramas y operativas fraudulentas sofisticadas. Esta herramienta no ha sido hasta la fecha explotada estadísticamente.

Un campo que tiene pendiente de abordar la administración tributaria es la georreferenciación entendiendo por tal el análisis, selección, asignación y estudio de los datos tributarios desde una perspectiva gráfica en mapas. Recientemente se han generado mapas Web con QGIS para la representación de los precios medios de alquiler por tipo de inmueble. Pero claramente tenemos que profundizar y conseguir expertos en estos formatos de presentación e incorporarlos a las herramientas de selección y análisis de manera extensiva.

En resumen, aunque ya se ha dicho anteriormente, la disponibilidad de datos, las herramientas tecnológicas y el conocimiento experto permiten su aplicación con diversidad de finalidades. Los anteriores son solo ejemplos, pero constituyen una realidad actual en la Administración Tributaria. No obstante, en la inferencia estadística, en las clasificaciones clúster y en las herramientas de aprendizaje de máquinas somos conscientes de que no hay un algoritmo único, que también éstos diferirán en los métodos existentes, en los datos, y en los posibles pesos o importancia que se atribuya a los mismos. Sin duda constatar esta realidad no debe ser un impedimento para seguir avanzando en estos procedimientos. Las bases necesarias para aplicar estas técnicas las tenemos: gran número de datos, datos de calidad, tamaños de población suficientes para el entrenamiento y contamos con expertos para incorporar la complejidad matemática que se precise en los modelos y, en ocasiones, la mejora en estimación que se obtiene al introducir un mayor número de variables de análisis que nos conducirá en general a un mejor resultado.

Cualquier institución pública o privada precisa disponer de datos e incrementar las funcionalidades y usabilidad de sus sistemas informáticos con la finalidad de simplificar y agilizar los trámites y mejorar la opinión de los usuarios y en ese sentido, la Agencia Estatal de la Administración Tributaria no es una excepción. El objetivo es garantizar la disponibilidad de servicios basados en plataformas en movilidad; incrementar la capacidad para procesar grandes volúmenes de datos; mejorar la integración de distintas fuentes de datos, incluso en soportes distintos; la introducción de operaciones lógicas y matemáticas complejas para manejar el volumen de información y la transformación de esa información en productos útiles, tanto para la toma de decisión política (reformas tributarias), como útiles para la sociedad al poner esa información a su disposición de una forma coherente, adecuadamente clasificada y con el mayor grado de actualidad posible.

En un artículo como éste no se puede relativizar el reto que supuso la introducción en 2017 del Suministro Inmediato de Información (SII), que consiste en trasladar los datos de las facturas recibidas y emitidas por las grandes empresas y los grupos de entidades IVA a la sede electrónica de la AEAT, que ha supuesto el tratamiento de más de dos mil millones de facturas anuales registradas en el SII, acompañada de la complejidad de explotar la información de manera analítica extrapolando esos datos a la generación de declaraciones resumen anual desde una perspectiva estadística y de exploración de nuevas situaciones de riesgos de fraude en IVA con herramientas creadas al efecto.

En la administración tributaria sabemos que una correcta identificación de la actividad económica y de las características de los contribuyentes es importante. La correcta identificación de los mismos puede ser determinante en la gestión de riesgos y tener implicaciones en el cumplimiento tributario, también puede abrir espacios a generar o consolidar beneficios específicos asociados a una serie de características. Y si éstas características y su delimitación exacta constituyen la base para la simulación de impactos de beneficios fiscales ex ante, también resulta muy relevante evaluar el impacto real de esas medidas ex post. De momento, estas técnicas de análisis están orientadas a la evaluación de las políticas públicas. En dicha evaluación es muy interesante analizar los contribuyentes que, cumpliendo los criterios objetivos para haber hecho uso de un incentivo fiscal, no lo han usado. Y esos comportamientos abren también un espacio de estudio basado en teorías del comportamiento; sin embargo, en relación a este objetivo por el momento todavía la predicción del comportamiento no está incluido en los simuladores de política fiscal, aunque no descartamos incorporarlos en un medio plazo.

En resumen, la tecnología nos abre posibilidades enormes, retos y un aprendizaje continuo, y no sólo desde una perspectiva estadística sino desde una operativa de negocio. Pasar de mecanismos tradicionales, que precisan varios empleados-mes para encontrar situaciones de riesgos y realizar contactos con los ciu-

dadanos para dirimir y verificar situaciones, a otras alternativas que permiten solventar las mismas tecnológicamente con un tratamiento exhaustivo de los datos existentes en las bases de datos tributarias, convierte a la administración tributaria en una institución pública a la vanguardia y pionera en la utilización de información de múltiples fuentes para crear un conjunto de datos sintéticos que sirvan para validar o confirmar los resultados. Esos datos pueden ser tanto declaraciones de períodos anteriores, información recibida de terceros (como bancos u operadores de tarjetas de crédito) o información tan detallada y voluminosa como la que viene de las facturas del SII. Estos datos estructurados podrían complementarse con otros no estructurados, como la publicidad que los contribuyentes despliegan en medios de comunicación y redes sociales, anuncios o páginas web donde se realiza comercio electrónico, o plataformas donde se anuncian alquileres turísticos. Todo ello constituye nuevas fuentes de información que se están uniendo a las habituales de los modelos de declaración aprobados por las órdenes ministeriales.

EXPERIENCIAS EN LA AEAT Y APLICACIONES

El mundo interconectado y toda la información que vamos dejando suponen oportunidades de mejorar los análisis y compartir experiencias con las instituciones más avanzadas y digitalizadas, la participación de administraciones, académicos y empresas compartiendo conocimiento, herramientas de código abierto y el uso de inteligencia artificial en un contexto de buenas prácticas, preservando ante todo la confidencialidad de los datos, es una posibilidad cada vez más cercana.

Existen métodos creados por la empresa privada que ya son accesibles a la comunidad científica y que también han sido utilizados experimentalmente en la Administración, entre ellos y quizás el más conocido es el método PageRank, desarrollado por Google, que se trata de un algoritmo de Centralidad Eigenvector o, dicho de otro modo, un algoritmo basado en las medidas de centralidad en la que se ponderan los nodos en función de su respectiva centralidad, cercanía o relación a otros nodos. Utilizando los datos de los nodos de forma iterativa el algoritmo devuelve un valor «Pagerank», que mide la frecuencia o probabilidad de acabar en cada uno de los nodos teniendo en cuenta sus conexiones. El método de «clasificación de páginas» de Google es el ejemplo más conocido de uso de este tipo de algoritmos, utilizando los datos de enlaces, búsquedas y frecuencia de visitas. Este algoritmo, que ha convertido al buscador de Google en el más demandado y utilizado mundialmente, es muy útil en otros contextos, por ejemplo, para localizar tramas de fraude en los que los flujos comerciales y financieros entre distintos operadores económicos se diseminan en distintos países y empresas.

Hace varios años se llevó a cabo un procedimiento experimental de selección de contribuyentes basado en redes neuronales entrenadas a partir de una selección amplia de contribuyentes inspeccionados con fraude fiscal y analizando más de 4000 variables de los mismos. Tras este entrenamiento la red seleccionaba del marco de población global los contribuyentes a inspeccionar. La idea era comparar los resultados en términos de fraude descubierto con esta peculiar selección respecto a los métodos tradicionales basados en contrastes, inconsistencias, imputaciones u otros indicios objetivos obtenidos de la información. Sin embargo, esta técnica que desde una perspectiva informática experimental es muy interesante, en su aplicación es menos efectiva porque no era posible conocer los atributos que habían tenido más peso en esa selección (mecanismo de caja negra), y desde luego, no contaba con los elementos de información de los programas de selección, resultando por tanto más efectivos estos últimos.

También en el IVA se realizó de manera experimental una selección aleatoria de empresas a inspeccionar en el sector de la construcción que se combinó con otra selección obtenida con procedimientos tradicionales, sin que el actuario conociera el método de selección para el control. Esta técnica dio resultados muy aceptables y se utilizó en varios ejercicios.

Lo característico de estas nuevas opciones de detección de fraude es que se trata de técnicas asépticas, que no presuponen un comportamiento razonado o modelizado del fraude. Por el contrario, el conocimiento experto del inspector es preciso para orientar la interpretación correcta de los datos y llevar a cabo una delimitación y estimación económica del fraude fiscal.

Respecto a la evolución tecnológica y el uso de herramientas de Big Data en la AEAT recomiendo acudir a la publicación número 44 de la Revista de Administración Tributaria CIAT/IEF/AEAT escrito por Ignacio González García de la Agencia Tributaria de España (*Analytics y Big Data. La Nueva Frontera*) que analiza las características técnicas básicas de estos nuevos métodos de aproximación basados en técnicas de Big data y ofrece ejemplos concretos de casos de uso desarrollados en la AEAT para la detección de redes familiares extendidas, cálculo de la riqueza societaria de los contribuyentes teniendo en cuenta las participaciones indirectas y cruzadas, detección de tramas de corrupción y blanqueo o de estructuras de fraude.

En el ámbito internacional, el informe de la OCDE de 2016 « *Advanced Analytics for Better Tax Administration: Putting Data to Work*» expone la aplicación del análisis de redes para la prevención del fraude carrusel en el IVA en diversos países (Irlanda, Países Bajos en el entorno de la UE), y apunta a la posibilidad de detectar por este método otros tipos de fraude a través de las conexiones entre distintos operadores económicos.

En la AEAT, la Dirección y los trabajadores estamos comprometidos a evolucionar y enfrentarnos a los desafíos que nos impone el nuevo mundo digital y esperamos estar permanentemente actualizados para desempeñar la función que justifica la existencia de nuestra organización.