R.I.T.I. nº 9 Enero-Abril 2019

Revista Internacional TRANSPARENCIA E INTEGRIDAD

APLICACIÓN DE MÉTODOS ESTADÍSTICOS, ECONÓMICOS Y DE APRENDIZAJE AUTOMÁTICO PARA LA DETECCIÓN DE LA CORRUPCIÓN (*)

José A. Alvarez-Jareño
Elena Badal-Valero
Jose M. Pavía
Departamento de Economía Aplicada de la Universitat de Valencia

RESUMEN

La actividad de la detección del fraude y la corrupción es tan antigua como los delitos que pretende destapar. En los últimos años, sin embargo, debido al avance de las nuevas tecnologías, los delitos económicos y de corrupción se han sofisticado. Consecuentemente, las técnicas y las metodologías para su detección y análisis han de evolucionar para adaptarse a esta nueva situación. Después de una primera definición de fraude y corrupción, se enumeran las principales características de estas actividades y se exponen las principales técnicas de detección y análisis. Se diferenciará entre las técnicas clásicas y las técnicas más modernas de aprendizaje automático, que se están implementando actualmente. Antes de cerrar con las conclusiones, se enuncia la Ley de Benford como herramienta para la obtención de evidencias de un posible delito, y se muestra con un ejemplo práctico su utilización.

1. INTRODUCCIÓN

La revolución que estamos viviendo de la mano de las tecnologías de la información y comunicación está propiciando que determinados términos como Big Data, Aprendizaje Automático o Ciencia de Datos estén cada vez estén más presentes y estén poniéndose de moda. Esto puede llevar a pensar, erróneamente, que la utilización de procedimientos y metodologías estadísticas para la detección del fraude y la corrupción sea también una actividad novedosa. Nada más lejos de la realidad. De igual modo que la corrupción no es un invento moderno, las técnicas de prevención y de detección tampoco lo son. Algunas de estas técnicas han llevado incluso, a lo largo de la historia, a enunciar principios físicos. Podemos recordar a Arquímedes y su grito triunfal "¡Eureka, Eureka!" tras descubrir cómo podía esclarecer si el orfebre del rey de Siracusa había utilizado todo el oro que le había sido entregado para la fabricación de una corona.

El oro y la plata, como elementos de valor, se transformaron pronto en monedas. Monedas que como primer bien de fabricación en serie debían ya, desde un primer momento, cumplir con unas especificaciones muy concretas. Para poder actuar como un objeto estándar de intercambio, las monedas debían (y deben) tener todas un mismo peso, composición de metal, grosor y diámetro (si con circulares). De ahí que, desde muy antiguo, surgiera la necesidad de comprobar que las nuevas monedas fabricadas (y las que estaban en circulación) cumplían con las especificaciones preestablecidas. Con este objetivo se estableció ya en el año 1279, en Inglaterra, la que se conoce como el "Trial of the Pyx", que ha perdurado hasta la actualidad.

Curran-Everett (2009) afirma que el "Trial of the Pyx" es el contraste de hipótesis de dos colas más antiguo que se conoce, siendo instaurado antes incluso de enunciarse los conceptos de probabilidad o muestreo. Su finalidad era detectar si la Casa de la Moneda utilizaba correctamente el oro y la plata que el Rey entregaba para acuñar moneda. Lo que esta prueba pretendía era detectar la posible existencia de corrupción entre los responsables de fabricar moneda en Inglaterra.

Uno de los pocos Masters de la Casa de la Moneda que tuvo problemas con esta prueba fue Sir Isaac Newton, quien ejerció el cargo desde 1699 hasta su fallecimiento. El jurado de orfebres rechazó las monedas de Newton por considerarlas inferiores al estándar, lo que forzó a Newton a demostrar que los estándares utilizados hasta ese momento en la prueba eran demasiado puros para una comparación justa. Alvarez-Jareño (2012) afirma que no se tienen evidencias de que Newton se beneficiase de esta situación, ya que todos sus esfuerzos estaban encaminados a reducir la variabilidad de las características de las monedas, aportando mejoras en el proceso de acuñación debido a que la metalurgia fue, junto con la alquimia, de gran interés para él.

Nuevos tiempos han traído nuevos delincuentes y formas de delinquir, con técnicas más sofisticadas, aunque también han conducido a nuevos métodos de detección del fraude, más acordes con los conocimientos y los nuevos instrumentos disponibles. Así, por ejemplo, el colombiano Eduardo Salcedo se dedica a combatir la actual delincuencia con inteligencia artificial, algoritmos o herramientas de visualización, siendo conocido como el moderno Sherlock Holmes[1]. Salcedo aboga por emplear intensamente la tecnología en la lucha contra el fraude, la corrupción y los delitos económicos, señalando las limitaciones que tenemos los humanos para procesar grandes cantidades de datos. En este sentido, Salcedo recuerda que el cerebro humano tiene barreras que son insalvables y cita a Robin Dunbar, quién fijó en 150 personas el límite de la estructura social que un ser humano puede abarcar. Por este motivo se hace imprescindible incluir las nuevas técnicas de inteligencia artificial en la prevención y detección del fraude y la corrupción.

2. FRAUDE Y CORRUPCIÓN

La primera cuestión a determinar es, sin embargo, que se entiende por fraude y por corrupción, y en que ámbito se va a aplicar. Lo más socorrido en estos casos es empezar por buscar cómo se definen en los diccionarios. El Diccionario de la Real Academia de la Lengua indica que la palabra fraude proviene del latín fraus, fraudis, y lo define como "acción contraria a la verdad y a la rectitud, que perjudica a la persona contra quien se comete". El fraude por lo tanto consta de dos partes, la primera una acción no legítima, la segunda es que debe generar un perjuicio. Se puede decir, que sin perjuicio no hay fraude.

El Diccionario de uso del español María Moliner, por su parte, lo define como "engaño hecho con malicia, con el cual alguien perjudica a otro y se beneficia a sí mismo". En este caso, además añade que el fraude busca proporcionar un beneficio en la persona que lo comete.

Respecto al vocablo corrupción, y atendiendo a la definición que la ONG Transparencia Internacional (2009) hace del mismo: "abuso del poder para beneficio propio", encontramos que la corrupción sería una forma de fraude. De hecho así es como se establece en los Manuales de ACFE (Association of Certified Fraud Examiners). La clasificación que establece Transparencia Internacional diferencia tres tipos, según la cantidad de fondos perdidos y el sector en el que se produzca: corrupción a gran escala, menor y política. Aunque se ha de señalar que también existe la corrupción en la empresa privada y entre empresas privadas.

[1] El Mundo, José Fajardo, 9 de enero de 2018
http://www.elmundo.es/papel/historias/2018/01/09/5a53b8cc22601dda7a8b463c.html

A pesar de lo anterior, sin embargo, tal como indica Queralt (2016), en el "derecho español y en la mayoría de ordenamientos europeos, no existen tipos penales específicos de corrupción". La corrupción se manifiesta a través de otros delitos que conforman una larga lista de comportamientos ilícitos. Entre los comportamientos que cabría destacar como corrupción se encontrarían el cohecho activo y pasivo, el soborno, la colusión, la malversación de caudales públicos, el tráfico de influencias, el delito fiscal, la prevaricación común y la prevaricación urbanística. Todos ellos delitos que como se puede comprobar atendiendo a sus definiciones pueden ser catalogados de fraudes. En todos los casos, se perjudica a alguien (muchas veces a los contribuyentes) y se beneficia a los que llevan a cabo estas acciones.

Bolton y Hand (2002) distinguen entre prevención y detección del fraude. La prevención del fraude englobaría el conjunto de medidas que se toman para detener o frenar el fraude. Y comprende desde las marcas de agua de los billetes bancarios hasta el código PIN del teléfono móvil. Su objetivo es evitar, prevenir el fraude. La detección del fraude, por su parte, consiste en identificar un hecho fraudulento y en tratar de hacerlo, una vez ha ocurrido, tan rápido como sea posible. Los procesos de detección entran en juego cuando la prevención ha sido burlada. Las tarjetas de crédito disponen de claves o números secretos para su utilización, sin embargo, al mismo tiempo las compañías de estas tarjetas evalúan todas y cada una de las operaciones en busca de posibles utilizaciones fraudulentas.

3. CARACTERÍSTICAS DEL FRAUDE Y LA CORRUPCIÓN

Evidentemente, los individuos que cometen fraude o corrupción persiguen, además de obtener un provecho personal, un objetivo secundario, no expuesto hasta este momento: pasar desapercibidos. Es decir, no ser detectados; que nadie se entere. Salvo en aquellos casos en que el fraude es muy burdo, siempre tienen la esperanza de no ser descubiertos. Para ello realizan una labor de encubrimiento, tratan de dotar de legitimidad a las acciones que realizan. Ello provoca que el fraude y la corrupción no sean, en general, fáciles de detectar.

En el caso de la corrupción existen incluso individuos que piensan que no serán denunciados e investigados debido a su status social o político, y ocurre que a veces no se esconden ni tratan de ocultar sus actividades ilícitas. De hecho, en algunos países la corrupción es consentida[2] por la clase dirigente y se ha convertido en una forma de vida. El fraude y la corrupción es un coste para todo negocio o país, y si el fraude supera un cierto umbral, puede convertir una actividad rentable en ruinosa. Por este motivo, el fraude no tiende a ser muy grande en un negocio, aunque, en otros niveles puede ser mucho mayor. En algunos países, el porcentaje de economía sumergida o de utilización indebida de servicios públicos puede llegar a tener un peso elevado.

En cualquier caso, y salvo casos extremos, de acuerdo con Jensen (1997), las dificultades para detectar el fraude y la corrupción residen en que:

- El porcentaje de fraude es pequeño (distribución de la clase sesgada). Desde el punto de vista estadístico, el número de personas corruptas, o el número de operaciones fraudulentas, es un porcentaje relativamente pequeño frente al número de personas honradas o el número de operaciones legítimas. Este problema es conocido como de datos desequilibrados (imbalanced data set). A veces, la sensación puede ser diferente, especialmente si solo se presta atención, si solo se focaliza el interés en los casos de corrupción, desconociéndose los casos en los que se ha actuado de forma correcta.

[2] La corrupción consentida, Otto Granados https://www.nexos.com.mx/?p=24569

- Es necesario disponer de bases de datos correctamente etiquetadas, es decir, tener identificados los individuos corruptos y honrados o las operaciones fraudulentas y legítimas. Este proceso de etiquetado lleva tiempo y es caro, y aun así, no se tendrá la seguridad de que sea completamente correcto. Normalmente será un experto el que realice esta labor para una posterior investigación. El resultado del análisis facilitará una clasificación futura de individuos u operaciones. No obstante, hay muchas operaciones que no son etiquetadas como fraudulentas, y pudiera ser que algunas de ellas sí que lo fueran. Nunca se puede tener la seguridad que no hay operaciones fraudulentas entre aquellas etiquetadas como legítimas.
- Los defraudadores y los corruptos son innovadores y tienen una gran creatividad y capacidad de adaptación. Algunos casos[3] tienen una preparación y una profesionalidad que los hace difícilmente detectables a menos que se realice una investigación amplia y coordinada que puede afectar a varias empresas o instituciones. Un caso aislado puede no ser detectado como fraude, precisando la colaboración conjunta de diferentes agentes para identificar como fraude un conjunto de casos.
- Los fraudes o la corrupción se tienen que detectar por un experto o investigador. El análisis de los datos puede facilitar la labor de identificación de posibles operaciones fraudulentas o individuos corruptos, pero será un experto o un investigador el que deberá reunir las pruebas para determinar si una operación es fraudulenta o si un individuo es corrupto. La investigación que se debe llevar a cabo tendrá un coste para la administración, la empresa o la sociedad, que cuentan con unos recursos limitados.

Como bien conocen los responsables de la justicia en los diferentes países, el fraude y la corrupción no son fáciles de detectar y de probar, y no sólo por la escasez de medios que hay en muchos países para luchar contra esta lacra social sino porque el fraude es un fenómeno dinámico. Consecuentemente, los sistemas de prevención y detección del fraude y la corrupción deberán estar constantemente monitorizados, actualizados y mejorados para que sean efectivos.

4. TÉCNICAS PARA LA DETECCIÓN DEL FRAUDE Y LA CORRUPCIÓN

El fraude y la corrupción han cambiado a lo largo de la historia debido a la mejora de la tecnología. Algunos crímenes que pasaron desapercibidos en otras épocas, y que podrían ser considerados como crímenes perfectos, hoy no son utilizados por los delincuentes porque con la tecnología actual serían detectados con bastante facilidad. No obstante, como los beneficios de las actividades ilícitas pueden ser importantes, existe siempre una motivación importante para la innovación. Los defraudadores y los corruptos tienden a refinar sus actuaciones en la medida que los investigadores disponemos de mejores herramientas para su detección y mayor conocimiento de sus estrategias.

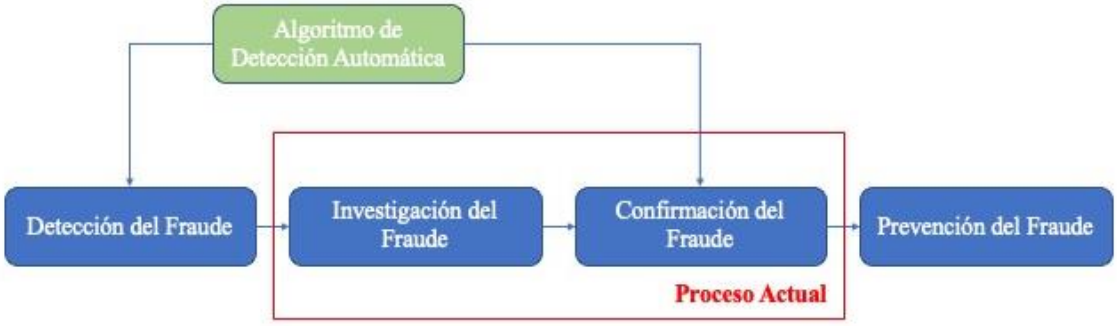
La detección del fraude mediante el empleo de un proceso automático permite clasificar de forma masiva operaciones o sujetos e identificar casos de alto riesgo. Las técnicas son muchas y variadas por lo que inicialmente las identificaremos como Algoritmo de Detección Automática, sin especificar ninguna metodología. El proceso de detección del fraude o ciclo del fraude implica, de acuerdo con Van Vlasselaer et al. (2017), ver Figura 1, principalmente cuatro actividades esenciales:

1. Detección del fraude: Aplicar los modelos de detección disponibles a observaciones nuevas y asignar un riesgo de fraude a cada una de ellas.

[3] El pueblo que fingía accidentes de tráfico: engañar al seguro se profesionaliza. http://www.elconfidencial.com/espana/2016-09-25/fraude-banda-organizada-pueblo-accidente-coche_1264170/ Cortarse la mano y otras ocurrencias para cobrar el seguro. http://www.elconfidencial.com/espana/2015-12-20/cortarse-la-mano-y-otras-ocurrencias-para-defraudar-al-seguro_1123969/

2. Investigación del fraude: A menudo es necesario que un experto humano investigue los casos sospechosos dada la complejidad y la sutileza de los mismos. Generalmente, este proceso de investigación es el que consume más tiempo y hace que el proceso global sea ineficiente.
3. Confirmación del fraude: Determinar si un fraude es verdadero, posiblemente con la ayuda de un equipo de investigación. Ellos serán los que finalmente determinen si el modelo acierta o se equivoca.
4. Prevención del fraude: Prevenir el fraude que se efectuará en el futuro.

Figura 1. Ciclo del Fraude.



(Fuente: Van Vlasselaer et al. (2017)).

Actualmente, las herramientas estadísticas para la detección del fraude se podrían dividir en dos grandes grupos:

- Técnicas clásicas o tradicionales.
- Técnicas de aprendizaje automático (machine learning).

4.1. Técnicas Clásicas

De igual forma que el fraude y la corrupción tienen una larga historia igualmente la tienen los métodos establecidos para la prevención. Incluso en algunos casos estos métodos han precedido a los conceptos matemáticos y estadísticos que los sustentan, como fue el "Trail of the Pyx" en la Inglaterra del siglo XIII. Aunque no todos los métodos clásicos son igualmente sofisticados, se pueden diferenciar cuatro metodologías catalogadas como clásicas:

- Listados: Listados de control de usuarios.
  Se trata de listados donde se dispone de información de presuntos estafadores. Por ejemplo, en el sector financiero español se dispone de ficheros de morosos, como el famoso RAI (Registro de Aceptaciones Impagadas) o muchos otros como ASNEF EQUIFAX, BADEXCUG, CIRBE (Banco de España), etc.; también se dispone de un Fichero de Inquilinos Morosos (FIM). Su finalidad es identificar clientes que cometieron algún tipo de fraude en el pasado y alertar a las entidades de que el comportamiento de estos clientes puede no ser honesto. Muchas de las empresas dedicadas a confeccionar estas listas han evolucionado en su negocio, pasando de recopilar información sobre determinados comportamientos, a realizar sus propios modelos de predicción y vender estos resultados como una media del riesgo de fraude.
- Reglas: Sistemas de reglas definidas por expertos.
  Los investigadores del fraude determinan una serie de reglas por las que un individuo puede estar cometiendo un fraude o una operación puede ser fraudulenta y se genera una alerta. Si en un evento se cumplen una serie de reglas identificadas por los expertos, se generará un procedimiento de investigación.

Estos sistemas basados en reglas son en ocasiones costosos de crear, ya que precisan de conocimiento experto, y normalmente, difíciles de mantener y ejecutar. Todos los eventos tienen que procesarse según las reglas, y los que cumplan con los requisitos serán marcados como sospechosos, pasando a ser investigados por expertos en la materia.

- Detección de comportamientos anómalos: Métodos estadísticos de clasificación para detectar comportamientos anómalos, como el análisis de regresión.
  El objetivo es detectar cambios en el comportamiento de los individuos para identificar un posible fraude. Se describe el comportamiento de un colectivo en base a un modelo estadístico o econométrico. Los individuos que no cumplen con el comportamiento esperado son marcados como fraude potencial.
- Análisis de relaciones: Análisis de relaciones que permitirá encontrar relaciones entre diferentes elementos de información. Redes de individuos que se dedican al fraude y que son detectadas por elementos que son comunes, y que individualmente no se podrían detectar.

Lo métodos de clasificación basados en la estadística tradicional (Hand (1981) y McLachlan (1992)), como el análisis discriminante lineal y la regresión logística, han sido herramientas efectivas para algunas aplicaciones en la detección del fraude y la corrupción, tal como exponen Bolton y Hand (2002).

4.2. Técnicas de Aprendizaje Automático

Se dispone también de una extensa variedad de herramientas en el Aprendizaje Automático (la gran mayoría de los cuales son y tienen su origen en los métodos estadísticos) que permitirán la clasificación de los sujetos o los eventos en base a su comportamiento en relación con la variable objetivo. Dentro del Aprendizaje Automático se disponen de los métodos supervisados y los no supervisados. Para los métodos supervisados es necesario disponer de una base de datos de casos conocidos, algunos de los cuales serán legítimos y otros serán fraudulentos, con los que se construirá un modelo para asignar una probabilidad de fraude a los nuevos casos que se analicen.

Los métodos no supervisados se utilizarán cuando los ejemplos del conjunto de datos no tengan asignada ninguna variable que distinga las operaciones legales de las ilegales. Las técnicas utilizadas con esta metodología son básicamente una combinación de métodos de segmentación y detección de valores extremos.

En este sentido, se podrían clasificar tres tipos de modelos para la detección del fraude:

- Detección de comportamiento inusual.
  Se realizaría en primer lugar un análisis de segmentación (clustering) para determinar grupos con comportamientos similares y posteriormente un análisis de valores extremos (outliers) para observar aquellos que tienen un comportamiento poco habitual para su grupo. Los ejemplos que *socialmente* pertenecen a un grupo suelen tener un comportamiento muy similar, por lo que cuando se separan de ese comportamiento puede indicar una anomalía a tener en cuenta. La ley de Benford sería un ejemplo de esta metodología.
- Detección de relaciones inexplicables.
  Este análisis se basaría en las técnicas de agrupamiento (clustering) y las reglas de asociación. También se pueden utilizar las técnicas de grafos que muestran grupos y relaciones para una determinada actividad. Algunas relaciones entre individuos/operaciones son difíciles de explicar, siempre y cuando no tengan intereses comunes. La relación entre un juez y un mafioso no debería ser muy habitual, ya que

los intereses de ambos están enfrentados. Si se detectan relaciones poco frecuentes deberán ser analizadas para ver qué implicaciones tienen.

- Detección de las características generales del fraude.
  Se realizará un análisis multivariante para identificar las características de los fraudes detectados hasta el momento y poder realizar predicciones. En este caso, las herramientas disponibles para efectuar el análisis son muy amplias, y se dispone desde las clásicas técnicas de regresión o los árboles de decisión hasta las redes neuronales o los support vector machine (SVM). Con esta metodología se pretende aislar las características del fraude y poder encontrar casos similares en otros eventos futuros.

Ngai et al. (2011) realizaron un análisis sobre 49 artículos científicos relacionados con el fraude financiero y llegaron a la conclusión que las técnicas más utilizadas en la minería de datos para su detección son las siguientes:

- Regresión Logística, que se utiliza principalmente para resolver problemas de fraude corporativo y en el seguro de automóviles.
- Redes Neuronales, cuya aplicación se realiza para la detección del fraude en tarjetas de crédito y en el seguro de automóviles.
- Redes Bayesianas, que se pueden utilizar para todo tipo de problemas de detección del fraude.
- Árboles de Decisión, que también suelen ser utilizados para una amplia resolución de problemas.

A las técnicas anteriores se deberían añadir las modernas técnicas de Random Forest y Extrem Gradient Boosting, que se han implementado recientemente en diversos trabajos para la detección del blanqueo de capitales (e.g., Badal-Valero, et al. (2018).

Jensen (1997) detectó algunas debilidades en la utilización del data mining para la detección del fraude, entre las que destacan:

- Utilización inapropiada de la representación de los datos. Considera que no se utiliza adecuadamente las técnicas de visualización de datos en la detección del fraude.
- Demanda de un conjunto inicial etiquetado. Aunque no siempre es necesario, porque se dispone de métodos no supervisados, la necesidad de un conjunto de datos correctamente etiquetado es un problema en la solución de este problema.
- Solo son adecuados para conjuntos muy grandes con mucha información y muchas variables. La cantidad de datos siempre es una limitación, aunque la captura y recopilación de datos es cada vez más sencilla.

4.3. Estabilidad de los modelos

El fraude y la corrupción son fenómenos dinámicos. Los corruptos prueban los sistemas de prevención y detección del fraude para comprender su funcionamiento y descubrir las debilidades, para a continuación adaptar sus métodos y estrategias de funcionamiento. Si el fraude es dinámico, los algoritmos para su prevención y detección también deberán serlo, y por ese motivo se deberá re-entrenar o actualizar los modelos con cierta asiduidad. Los factores que determinarán cada cuanto tiempo se deben actualizar los modelos son los siguientes:

- La volatilidad del comportamiento de los criminales. Cuanto mayor sea la volatilidad menos tiempo tendrá validez el modelo y con mayor frecuencia precisará ser actualizado.
- El poder de detección del modelo, que está relacionada con la volatilidad del comportamiento de los criminales. Si el modelo fuera muy preciso, descubriría mucho

fraude, e inmediatamente, los corruptos cambiarían su forma de actuar para evitar ser descubiertos. Será importante no dar difusión de los modelos utilizados, o bien utilizarlos selectivamente para evitar que los criminales conozcan su utilización.

- El número de casos parecidos o similares confirmados disponibles en la base de datos. Los modelos serán más estables cuanto mayor sean el número de casos en la base de datos que están correctamente clasificados. Es decir, cantidad y calidad de los datos.
- La velocidad a la que se confirman los nuevos casos. El tiempo necesario para identificar, investigar y confirmar un nuevo fraude será importante a la hora de actualizar los modelos, ya que los datos que se incluyan en el conjunto de entrenamiento deberán tener la suficiente calidad para no introducir un sesgo en el algoritmo. Si se incluyen datos clasificados como lícitos, y después de algún tiempo pasan a ser fraude, durante ese período de tiempo el modelo estaría dejando de identificar algún comportamiento que se está utilizando en ese momento.
- El esfuerzo requerido para re-entrenar el modelo. El tiempo de computación de los modelos es cada vez menos importante, ya que en la actualidad se pueden trasladar el proceso a sistemas distribuidos que reducen sustancialmente los tiempos de ejecución de los algoritmos.

Los modelos tienen una vida limitada y será necesario actualizarlos constantemente. Una vez que los criminales conocen que se está aplicando un método para detectar un tipo de fraude, modifican su comportamiento para evitar ser detectados. Si de acuerdo con la normativa se debe monitorizar todo ingreso bancario de más de 3.000 euros, inmediatamente las personas que no deseen ser identificadas o monitorizadas pasarán a realizar ingresos de menos de esta cantidad.

5. LA LEY DE BENFORD

Simon Newcomb (1835-1909), astrónomo y matemático americano, descubrió por primera vez el "fenómeno del primer dígito" en 1881. Newcomb observó que los libros de tablas logarítmicas de la biblioteca estaban considerablemente más usados en las primeras páginas y que, progresivamente, las marcas iban disminuyendo a medida que se inspeccionaba en las páginas finales. Como las tablas logarítmicas se utilizaban para conocer el resultado de esta operación, no podía ser casual que los diferentes libros de tablas no tuvieran las mismas evidencias de desgaste. Si estaban más utilizados en las primeras páginas, que son las que incluyen los logaritmos que empiezan por 1 y 2, debería ser porque es más probable que en la *naturaleza* se encuentren números que empiezan por el dígito 1 ó 2 que por los dígitos 8 ó 9. Newcomb fue capaz de deducir una fórmula que en teoría daba la probabilidad de que un número al azar empezara por un dígito concreto. Este descubrimiento fue expuesto en el American Journal of Mathematics en 1881, aunque pasó totalmente desapercibido. La publicación constaba únicamente de dos páginas.

Tuvieron que pasar 57 años, hasta que el físico Frank Benford de la General Electric redescubriera, de forma independiente, la misma ley y testara la validez de la misma en diferentes conjuntos de datos. Sus resultados se ajustaban tan correctamente a la fórmula propuesta que, actualmente, esta ley probabilística es conocida como Ley de Benford. Una de las características principales de la Ley de Benford es que no depende de cambios de unidades de medida. En términos económicos, no depende de la moneda utilizada.

Todos los conjuntos de datos no obedecen a la Ley de Benford, existen múltiples ejemplos como los números de las guías telefónicas o la tabla de raíces cuadradas que no la siguen. Sin embargo, existen muchos otros conjuntos, que surgen de manera natural, que sí se ajustan a esta ley. La serie de números de Fibonacci, habitual en problemas de computación o biología, es un ejemplo de cumplimiento de la Ley de Benford. Estos campos no son los únicos. Son muchas las variables de origen económico que se ajustan a la Ley de Benford. Entre ellas los

índices bursátiles, tal como analizan Ley y Varian (1994) para el índice Down-Jones, Ley (1996) para el Down-Jones Industrial Average y el Standard & Poor's 500, o Tödter (2007) para los valores bursátiles de la bolsa alemana.

El profesor Nigrini (1992) fue pionero en utilizar la Ley de Benford para detectar irregularidades contables y evasión de impuestos, y desde entonces es usada para detectar fraude y datos 'manufacturados' en documentos financieros. En este sentido, Rauch et al. (2011) inspeccionaron las cifras macroeconómicas reportadas a Eurostat entre 1999 y 2009 por los países miembros de la UE y encontraron que Grecia es el país cuyos datos más se alejan de cumplir con la ley de Benford. Recientemente, Das, Mishra y Rajib (2017) han propuesto utilizar la ley de Benford para la auditoría contable.

Para ejemplificar la ley de Benford utilizaremos un conjunto de datos conocido: la serie de números de Fibonacci. Se utilizarán los primeros 1000 números para comprobar si esta serie se ajusta a la ley de Benford (ver Figura 2). La comprobación se realizará con la librería benford.analysis desarrollada por Cinelli (2017) para R. Con sus funciones se efectúan diferentes análisis, entre ellos un contraste de bondad de ajuste.

Los resultados para la serie de Fibonacci son los siguientes:

- p-valor del contraste chi-cuadrado: 1.
- p-valor del contraste del Arco de la Mantisa: 0,9997.
- Desviación Absoluta Media: 0,00082.

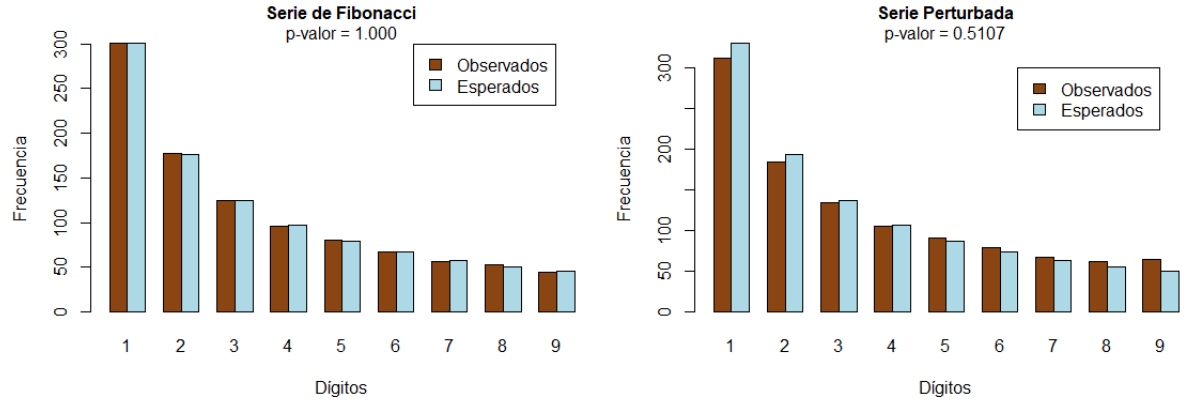
A continuación, introducimos datos fraudulentos (datos 'manufacturados') en el conjunto de datos, generando al azar, con la función sample de R, 100 números aleatorios (lo que representa solo un 10% del número de datos de la serie original). Tras agrupar ambos conjuntos de datos (los primeros dígitos de la serie de Fibonacci y de los números generados aleatoriamente), recalculamos los estadísticos de los contrastes y las medidas, obteniendo:

- P-valor del contraste chi-cuadrado: 0,5107.
- P-valor del contraste del Arco de la Mantisa: 0,1519.
- Desviación Absoluta Media: 0,00664.

Como se observa, aunque los resultados del p-valores no cambian la decisión de los contrastes de hipótesis, para ambos contrastes se reducen sustancialmente los p-valores, y el incremento que se produce en la desviación absoluta media es considerable. La inclusión de solo un 10% de datos generados aleatoriamente con la función sample de R genera un cambio en la distribución de los números que puede ser detectado mediante los contrastes y las medidas facilitadas en la librería benford.analysis. Estos resultados se pueden representar a través de unos diagramas de barras como el que se presenta en la Figura 2, donde en las barras claras corresponden con la distribución teórica de la Ley de Benford y las barras oscuras corresponden a los datos analizados (serie de Fibonacci más datos aleatorios).

Al realizar una visualización de los resultados se puede apreciar, por ejemplo, que el número nueve aparece con mayor asiduidad que la proporción esperada. Las diferencias en el estadístico chi-cuadrado (p-valores representados en los gráficos) son sensiblemente mayores en el número nueve, lo que indicaría que su distribución difiere sustancialmente de la esperada.

Figura 2. Comparación de las distribuciones con Ley de Benford.



(Fuente: Elaboración Propia.)

Estos números han sido generados con una función, pero muchas veces los datos son "generados" por los humanos. La generación de datos por humanos no producirá ni siquiera números pseudo-aleatorios, como lo son la mayoría de las funciones utilizadas por los principales programas informáticos. Los datos que eligen las personas estarán afectados por algún tipo de criterio o preferencia que, analizados bajo criterios estadísticos, podría mostrar una tendencia o sesgo a elegir más unos números que otros. Ello facilitaría a un algoritmo detectar estas modificaciones o cambios en la distribución de los números.

Kahnemann (2015) lo explica de la siguiente forma: "Así, la gente espera que las características esenciales del proceso estarán representadas no solo globalmente en la secuencia entera, sino también localmente en cada una de sus partes. Pero una secuencia localmente representativa se desvía de manera sistemática de la posibilidad esperada: contiene demasiadas alternancias y muy pocas repeticiones."

6. CONCLUSIONES

El análisis del fraude permite varias aproximaciones, diferentes soluciones, que deben ir desarrollándose y adaptándose de forma dinámica, en una lucha infinita, en una carrera sin fin, debido a la creatividad continua del defraudador. Aunque algunos problemas tengan difícil solución, esto no debería representar un obstáculo insalvable. Al contrario, esto debería ser contemplado como como un acicate, un estímulo para los investigadores, ya que el reto, el desafío, la curiosidad forma parte consustancial del ser humano y del investigador.

Detener el fraude en sus fuentes, para pasar de una política de respuesta a una preventiva a través del análisis de los datos es otro de los desafíos. Avanzar en la detección del fraude en el punto de aplicación, utilizando la tecnología que identifica señales de alerta mediante análisis de voz, análisis de texto y análisis del comportamiento puede ser una línea de trabajo futura. Determinar cómo formar al personal para trabajar con las métricas de análisis del fraude es una tarea en la que ya deberíamos estar trabajando.

7. BIBLIOGRAFÍA

Alvarez-Jareño, J.A. (2012): "The Trail of the Pyx", un Control Estadístico de la Calidad con Ocho Siglos de Antigüedad. En *Historia de la Probabilidad y la Estadística VI* (pp. 147-159). Madrid: Ed. UNED.

Badal-Valero, E.; Alvarez-Jareño, J.A. y Pavía, J.M. (2018): Combining Benford's Law and Machine Learning to detect Money Laundering. An actual Spanish court case, *Forensic Science International*, 282, pp. 24-34.

Benford, F. (1938): The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 78, pp. 551-572.

Bolton, R.J. y Hand, D.J. (2002): Statistical Fraud Detection: A Review. *Statistical Science*, 17(13), pp. 235-255.

Cinelli, C. (2017): benford.analysis: Benford Analysis for Data Validation and Forensic Analytics. R package version 0.1.4.1. https://CRAN.R-project.org/package=benford.analysis

Curran-Everett, D. (2009): Explorations in statistics: hypothesis tests and P values. *Advances in Physiology Education*, 33, pp. 81-86.

Das, R.C.; Mishra, C.S. y Rajib, P. (2017): Detection of Anomalies in Accountig Data Using Benford's Law: Evidence from India. *Journal of Social Science Studies*, 4(1), pp. 123-139.

Hand, D.J. (1981): *Discrimiantion and Classification*. Chichester: Ed. Wiley.

Jensen, D. (1997): *Prospective assessment of AI technologies for fraud detection: a case study*. AAAI Workshop on AI Approaches to Fraud Detection and Risk Management, Menlo Park, California, pp. 34-38.

Kahnemann, D. (2015): *Pensar rápido, Pensar despacio*. Madrid: Ed. Debate.

Ley, E. (1996): On the Peculiar Distribution of the U.S. Stock Indexes' Digits. *The American Statistician*, 50(4), 311-313.

Ley, E. y H.R. Varian (1994): Are there psychological barriers in the Down-Jones Index? *Applied Financial Economics*, 4, pp. 217-224.

McLachlan, G.J. (1992): *Discriminant Analysis and Statistical Pattern Recognition*. New York: Ed. Wiley.

Newcomb, S. (1881): Note on the frequency of use of the different digits in natural numbers. *American Journal of Mathematics*, 4, pp. 39-40.

Ngai, E.W.T.; HU, Y.; Wong, Y.H.; Chen, Y. y Sun, X. (2011): The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, 50, pp. 559-569.

Nigrini, M.J. (1992): *The detection of income escape through an analysis of digital distributions*. (PhD Tesis, University of Cincinatti).

Queralt, J.J. (2016): Public Compliance y Corrupción: Análisis Conceptual y Propuestas. *Revista Internacional Transparencia e Integridad*, 2, pp. 1-11.

Rauch, B.; Göttsche, M.; Brähler, G. y Engel, S. (2011): Fact and Fiction in EU-Governmental Economic Data. *German Economic Review*, 12(3), pp. 243-255.

Tödter, K.H. (2007): Das Benford-Gesetz und die Anfangsziffern von Aktienkursen. *Wirtschaftswissen-schaftliches Studium*, 36(2), pp. 93-97.

Transparencia Internacional (2009): *Guía de lenguaje claro sobre la lucha contra la corrupción*. https://transparencia.org.es/wp-content/uploads/2014/10/Gu%C3%ADa-de-lenguaje-claro-sobre-lucha-contra-la-corrupción.pdf [Último acceso: 22 de octubre de 2018]

Van Vlasselaer, V.; Eliassi-Rad, T.; Akoglu, L.; Snoeck, M. y Baesesns, B. (2017): Gotcha! Network-based Fraud Detection for Social Security Fraud. *Management Science*, 63(9), pp. 3090-3110.

(*) La presentación del contenido de este artículo ha recibido financiación a través de la convocatoria pública de la Generalitat Valenciana (DOGV 8064, de 16-6-2017) relativa al Sistema de alertas rápidas en la lucha contra la corrupción.