Evasión en IVA: Análisis de redes

Centro de Ciencias de la Complejidad (C3)
Instituto de Física (IF)

Universidad Nacional Autónoma de México (UNAM)

en colaboración con:

Department of Network and Data Science (DNDS)

Central European University (CEU)

Objetivo

El objetivo general de este proyecto consiste en identificar lazos entre actores que comercializan facturas electrónicas que simulan operaciones tales como canales de transacciones (compras, ventas u otras formas de enajenación) mediante algoritmos y técnicas de análisis estadístico de teoría de redes. Esto con el fin de reconocer y agrupar un conjunto de actores con presunta participación en actos de defraudación fiscal, estimar el monto que defraudan del Impuesto al Valor Agregado y diseñar estrategias de combate a este fenómeno basadas en métodos desarrollados en el estudio científico de los sistemas complejos.

Resumen ejecutivo

Usando una versión anonimizada y agregada por mes de todos los comprobantes digitales emitidos en México entre enero del 2015 y diciembre del 2018, *en este estudio analizamos una gran cantidad de datos para detectar posibles evasores fiscales.* Usando técnicas innovadoras de ciencia de redes e inteligencia artificial, exploramos la posibilidad de detectar posibles sospechosas de ser Empresas que Facturan Operaciones Simuladas y posteriormente estimar cuántos recursos han evadido. Aunque las respuestas en este documento son sólo estimaciones, pueden ser útiles para determinar acciones que inhiban la evasión fiscal y establecer bases para estudios futuros más específicos. Analizamos el comportamiento de las emisiones de comprobantes fiscales entre integrantes del Registro Federal de Contribuyentes (RFC), de los cuales un conjunto ha sido ya identificado y publicado previamente por el SAT como EFOS definitivas y presuntas, con el propósito de identificar patrones en la actividad que resulten útiles para la detección de otras posibles EFOS.

Por lo tanto, para cumplir con este objetivo consideramos tres elementos fundamentales: 1) el fundamento legal que permite la trazabilidad de los sujetos obligados; 2) la disponibilidad de contribuyentes categorizados como evasores fiscales; y 3) los métodos necesarios para la detección de contribuyentes con comportamiento similar a los evasores fiscales.

La trazabilidad de los sujetos obligados (personas físicas y morales) al entero y pago del Impuesto al Valor Agregado (IVA) se deriva principalmente: 1) de la obligación enunciada en el artículo 32 fracción III de la Ley del IVA sobre expedir y entregar comprobantes fiscales y 2) de acuerdo a la fracción II del artículo 5, la disposición al contribuyente de utilizar el comprobante fiscal como un medio para trasladar y acreditar el impuesto en términos de la propia Ley, es decir, el contribuyente debe sustentar ante el SAT el origen de su saldo a favor o por pagar por medio de comprobantes fiscales.

Ambos ordenamientos vinculados al artículo 29 del Código Fiscal de la Federación, el cual establece la obligación de expedir Comprobantes Fiscales Digitales por Internet (CFDI) por los actos o actividades que realicen los contribuyentes, proveen al SAT de la información necesaria para conocer los vínculos de compra-venta que cada contribuyente realiza, incluso permite seguir la cadena de las operaciones comerciales entre los sujetos obligados.

Con base en esta información, los métodos utilizados en nuestro estudio son: 1) *ciencia de redes*, 2) *redes neuronales artificiales*, y 3) *bosques aleatorios*. El análisis basado en ciencia de redes es útil para realizar una caracterización de los mecanismos de operación y asociación de EFOS, mientras que los métodos basados en técnicas de aprendizaje de máquinas (redes neuronales artificiales y bosques aleatorios) son utilizados para clasificar una población de contribuyentes como sospechosos de presentar comportamientos similares a las EFOS que ya han sido identificados

por el SAT. Cada uno de estos métodos provee una lista de posibles EFOS, resultado del análisis independiente de distintos aspectos de la actividad fiscal de los contribuyentes. Consideramos que los contribuyentes que aparecen en ambas listas tienen una mayor probabilidad de tener un comportamiento sospechoso similar al de las EFOS, y por lo tanto, después de realizar un proceso adicional de validación basado en métricas obtenidas del análisis de ciencia de redes, los incluimos en el cálculo de un estimado de la evasión de IVA asociado a este tipo de comportamiento ilícito.

El método implementado para la caracterización de patrones de emisión de comprobantes y mecansimos de organización de EFOS está basado en la *ciencia de redes*. Este enfoque ha sido utilizado anteriormente para el modelado y el análisis de sistemas de diversos tipos como: redes genéticas o neuronales, redes de transporte y comunicación, interacciones sociales, colaboraciones científicas y, recientemente, redes de corrupción, criminales y de evasión de impuestos, etc. [1–5]. El que la ciencia de redes sea un enfoque adecuado para la descripción de este tipo de sistemas se debe principalmente a que todos ellos se componen de un conjunto de elementos (nodos) entre los cuales se pueden definir interacciones o relaciones (enlaces), por ejemplo, la activación o inhibición de una neurona o un gen sobre otro, las co-autorías en artículos científicos, grupos sociales compartiendo información, etc. En nuestro caso, la emisión de CFDI nos permite definir una red de interacción en la que los nodos se asocian a contribuyentes y los enlaces a la emisión o recepción de CFDI correspondientes a transacciones comerciales entre ellos. De esta forma nos es posible construir redes de interacción mensuales y anuales considerando criterios basados en la regularidad y los montos asociados a CFDI emitidos y recibidos por EFOS.

El primero de nuestros métodos de aprendizaje automatizado corresponde a una *red neuronal artificial* (RNA). Ésta consiste en un conjunto de unidades, llamadas neuronas artificiales, conectadas entre sí para transmitirse señales. Cada neurona artificial realiza una función, es decir: recibe variables de entrada, las multiplica por un peso modificable en el entrenamiento de la red, aplica una función matemática a sus entradas, y genera un resultado, el cual puede ser usado después por otra neurona. En otras palabras, la información de entrada atraviesa la red neuronal (donde se somete a diversas funciones y operaciones) produciendo valores de salida. Las RNA son apropiadas para aplicaciones en las que no se dispone *a priori* de un modelo identificable que pueda ser programado, pero se dispone de un conjunto básico de ejemplos de entrada. También son altamente robustas tanto al ruido como a la disfunción de elementos concretos y son fácilmente paralelizables. En nuestra implementación, diseñamos una RNA que recibe como entrada datos de las facturas asociadas a un RFC y como salida devuelve un valor entre 0 y 1 que indica la probabilidad de que el RFC sea parte de las posibles EFOS.

El segundo de nuestros métodos de aprendizaje de máquinas es un *bosque aleatorio* (BA), formado por varios árboles de decisión. Un algoritmo de árbol de decisión consiste en realizar cortes en los valores de cada una de la variables o características consideradas en los datos. Tales cortes forman reglas de decisión, y una secuencia de decisiones asociadas constituye un "camino" que muestra cuáles son los aspectos que debe tener un elemento del sistema (un contribuyente) para considerarse perteneciente a la clase deseada (una de las EFOS). Un conjunto aleatorio de varios árboles de decisión da origen al bosque aleatorio, el cual da robustez al algoritmo. El resultado de este algoritmo es similar al de una RNA: un número entre 0 y 1 interpretado como la probabilidad de que un contribuyente sea una EFOS.

Analizando la distribución de montos asociados a las emisiones de CFDI realizadas por EFOS (definitivas y presuntas), hemos identificado que estos realizan emisiones diferenciadas según el tipo de receptor de las transacciones. Los montos asociados a operaciones entre EFOS son mayores (entre cientos de miles y millones de pesos) que las operaciones que realizan hacia contribuyentes no identificados como EFOS (alrededor de decenas de miles de pesos o menores). Este comportamiento diferenciado nos permite definir un *nivel de actividad de EFOS*, el cual utilizamos para filtrar los enlaces en las redes de interacción mensuales.

El análisis de la estructura de las redes de interacción nos ha permitido identificar subredes de operación de contribuyentes alrededor de EFOS publicadas. Estas subredes están asociadas a un flujo circular de emisiones de comprobantes en las que también están involucrados contribuyentes no etiquetados como EFOS por el SAT. Estos contribuyentes generan sospechas debido a su estrecha interacción con EFOS y por ser parte del flujo de emisiones de comprobantes potencialmente asociados a operaciones simuladas. El análisis de redes nos ha permitido definir una métrica para cuantificar el nivel de colusión de contribuyentes no clasificados como EFOS dentro de las subredes de operaciones sospechosas. Dicho índice nos permite realizar una validación adicional de los contribuyentes clasificados como sospechosos de ser EFOS, y así realizar un estimado de los montos evadidos anuales.

En más detalle, obtenemos listas de contribuyentes sospechosos (al aplicar nuestros métodos de clasificación a la base de datos de CFDI proporcionada por el SAT)y el nivel de colusión de sospechosos dentro de redes de operación de EFOS (a través de un índice de cercanía), obtenemos un estimado conservador de evasión del IVA adicional al ya identificado por el SAT. El monto de evasión estimado presenta una tendencia creciente que va de 40,097.2 millones de pesos (MDP) en 2015 a 77,318.6 MDP para 2018. En el periodo en general, se estima un promedio anual de 60,605 MDP y 7,677 RFC sospechosos. Es importante recalcar que la identificación de contribuyentes sospechosos es sólo el resultado de nuestros métodos de clasificación, y complementa los esfuerzos e investigaciones exhaustivas realizadas por el SAT de acuerdo a lo que la ley establece.

Los análisis realizados en este estudio nos permiten identificar algunas características de la forma de asociación de las EFOS y sus métodos de operación. No obstante, es necesario realizar estudios más específicos que se enfoquen en caracterizar detalles de las operaciones asociadas a EFOS y su comportamiento temporal, los cuales conlleven a herramientas y métodos de identificación más robustos. Con base en nuestros resultados, emitimos las siguientes recomendaciones a consideración futura del SAT: a) Complementar los sistemas automáticos del SAT en el monitoreo y detección de EFOS sospechosas con técnicas basadas en los métodos de este estudio o similares; b) mejorar la prevención de auto-facturas y flujos circulares de activos en redes de emisiones y recepciones; y c) disminuir el uso de RFC genéricos y así aumentar la eficacia de métodos de caracterización y clasificación de EFOS. Al final de este documento incluimos otros mecanismos que podrían ser útiles para el SAT, así como líneas de investigación en las que nuestro grupo de investigación podría contribuir en un futuro cercano.

1. Introducción Fiscal

Las contribuciones tributarias dentro del Estado mexicano juegan un papel fundamental, ya que mediante su recaudación se puede invertir en programas y obras públicas (a corto, mediano, y largo plazo), mantenimiento de infraestructura y otras acciones que promueven el desarrollo de distintos sectores de la población[1].

En el mismo sentido, las leyes fiscales establecen las contribuciones que deberán aportar los ciudadanos al Estado. En las leyes fiscales se define el sujeto, la base, la tasa o tarifa que cada impuesto contiene, así mismo, incluye la periodicidad, forma de pago y demás apreciaciones que la autoridad fiscal determine para poder alcanzar los objetivos planteados en la Ley de Ingresos de la Federación que año con año, plantea los recursos con los que dispondrá el gobierno para poder hacer frente a todos los compromisos contraídos con la población.

El Servicio de Administración Tributaria (SAT), órgano desconcentrado de la Secretaría de Hacienda y Crédito Público (SHCP), es quien tiene la responsabilidad de facilitar e incentivar el cumplimiento voluntario de las obligaciones tributarias de las personas físicas (ciudadanos) y morales (compañías). A pesar de la estructura y metas establecidas por el Estado mexicano por medio del SAT, existen contribuyentes que buscan eludir una obligación que la Constitución impone de participar en los gastos que requiere la nación, mismos que le son necesarios para su existencia y desarrollo. De esta forma se genera el fenómeno de *evasión fiscal* [6], definida de acuerdo al SAT como "toda acción, u omisión, parcial o total, tendiente a reducir o retardar el cumplimiento de la obligación tributaria" [7].

La evasión fiscal incluye la omisión de ingresos percibidos, el incremento no justificado de deducciones (aplicación de gastos no deducibles), el pago de un monto menor de impuestos, entre otras actividades ilícitas. Algunas de sus posibles causas de la evasión, son el costo-beneficio de quien evade, una escasa conciencia o cultura tributaria, el comercio informal, la corrupción, las lagunas legales y la simulación de operaciones.

La evasión fiscal disminuye la equidad horizontal y vertical [8], pues los evasores pagan menos impuestos que contribuyentes con igual capacidad de pago, y porque una tasa impositiva elevada aumenta el estímulo para no pagar impuestos. Por lo tanto, la evasión fiscal conlleva a una gama de problemas como la reducción de ingresos tributarios, la desigualdad de la carga tributaria, una competencia injusta entre contribuyentes y evasores y una percepción de ineficiencia por parte de la autoridad fiscal.

Los evasores fiscales tienen algunos comportamientos que pueden llegar a ser desde muy simples hasta muy complejos. De alto interés es el de aquellos contribuyentes que simulan operaciones sin haber realizado alguna actividad económica que los ampare. A una empresa que emite comprobantes sin la prestación de un servicio o la comercialización de un bien se le conoce como Empresa que Factura Operaciones Simuladas (EFOS). Las EFOS se caracterizan frecuentemente por no tener personal activo o registrado en el Instituto Mexicano del Seguro Social (IMSS), así como por no contar con la infraestructura necesaria para mantener operaciones que generen los ingresos esperados de acuerdo a su giro comercial o de servicios. Además, suelen indicar un domicilio fiscal falso o lo cambian constantemente, volviéndose no localizables.

Las EFOS, de acuerdo al procedimiento implementado por el SAT para atender lo establecido en el artículo 69-B del Codigo Fiscal de la Federación (CFF), se catalogan como [9]:

[1] La Constitución Política de los Estados Unidos Mexicanos (CPEUM), en su artículo 31 fracción IV, dispone la obligación que todos los mexicanos tienen que contribuir para los gastos públicos de la Federación, del Distrito Federal (ahora Ciudad de México) o del Estado y Municipio en el que residan, de la manera proporcional y equitativa que dispongan las leyes. Y del mismo modo la CPEUM faculta al Congreso de la Unión para imponer las contribuciones necesarias mediante la legislación que corresponda.

Presunta: Estatus inicial del SAT para la notificación de contribuyente con operaciones sospechosas que asimilan a EFOS. Este mismo hace referencia a las empresas que por su operación variable, domicilio no válido y falta de activos ya sea de carácter material o humano pudieran estar generando operaciones simuladas [2].

Definitiva: Son las empresas que no presentaron un proceso para desvirtuar o su proceso no fue satisfactorio para poder demostrar que su operación es real, por lo cual quedan observadas como EFOS y no podrán efectuar transacciones de facturación con terceros ya que sus certificados y RFC quedarán inválidos para la generación de las mismas. Así mismo este tipo de contribuyente podrá tener un proceso penal en el cual puede llevar hasta los 6 años de cárcel.

Es importante mencionar que una empresa que tiene relación ya sea como cliente o proveedor de algún contribuyente que se encuentre en este estatus, podrá ser llamada para procesos fiscales por estar relacionada con EFOS.

Las empresas que reciben comprobantes fiscales de las EFOS se denominan Empresas que Deducen Operaciones Simuladas (EDOS). Aunque también realizan actos fiscales ilícitos, las EDOS suelen tener una estabilidad y una formalidad comprobable en la plantilla de su nómina, en sus activos fijos y en el pago de sus contribuciones. Las EDOS se pueden describir como contribuyentes regulares; al adquirir un comprobante fiscal derivado de una operación simulada, no obstante, las EDOS buscan reducir su base de impuestos, y así acreditar el Impuesto al Valor Agregado (IVA) para anular o disminuir el pago de este impuesto y eventualmente generar beneficios fiscales que en el extremo podrían ser devoluciones o compensaciones.

El CFF[3] en el artículo 69-B[4] prevee un procedimiento para mitigar este tipo de esquemas de evasión fiscal y con ello reducir el impacto que esto genera a la recaudación de los impuestos en la forma y términos que las distintas leyes fiscales señalan. Para el caso específico por el cual se elabora este trabajo, se necesita tener sumamente claro que tanto el CFF y la Ley del IVA se complementan una a la otra a fin de poder determinar precisamente cuáles son todas las obligaciones que involucran todos los aspectos en la emisión de comprobantes fiscales y en su caso, en el momento en que estos son emitidos de manera indebida a través de ciertos canales de transacción (compra y venta de comprobantes fiscales).

Además se deben tomar en cuenta los criterios de la Resolución Miscelánea Fiscal (RMF)[5] la cual de acuerdo al IMCP [11], pretende precisar la regulación establecida en las leyes y reglamentos fiscales, con el fin de lograr su eficaz aplicación y facilitar el cumplimiento de la ley, respetando en todo tiempo la seguridad jurídica de los contribuyentes en cuanto a los principios de reserva y primacía de ley.[6].

Al desarrollar tecnología para la emisión de comprobantes fiscales, es posible conocer cuales son los comportamientos que se generan durante las operaciones y transacciones comerciales o de servicios[7] Mediante el análisis de redes se pueden usar diversos métodos computacionales y estadísticos para clasificar atributos e identificar los enlaces que existen dentro de una red de emisores y receptores de comprobantes fiscales. Estos métodos también funcionan como

[2]Una vez notificado como presunta existe un procedimiento en el SAT para poder desvirtuar cualquier observación. Las empresas que son observadas como presuntas ya NO pueden salir de la lista del SAT.

[3]Para Reyes Caballero [10], el CFF es un compendio de diversos aspectos fiscales, cuyo objetivo es determinar las contribuciones y las diversas obligaciones que se deben cumplir en relación con los impuestos federales.

[4]Este articulo fue adicionado a finales de 2013 como parte de las diversas reformas planteadas en ese entonces.

[5]El Servicio de Administración Tributaria tiene la obligación, de acuerdo a la fracción I del artículo 33 del CFF, de publicar anualmente las resoluciones dictadas por la autoridad que establezcan disposiciones de carácter general.

[6]Dentro de la estructura que compone a la RMF se encuentran diversos Títulos, los cuales están agrupados de acuerdo a la disposición en especifico que busca precisar. También contiene una serie de anexos, los cuales buscan profundizar de manera aún mas especifica acerca de un elemento fiscal que necesite ser detallado para su debido cumplimiento.

[7]Dadas las características del comprobante fiscal o CFDI, el cual contiene diversos nodos o campos, se pueden determinar redes de interacción entre emisores y receptores que permiten visualizar los flujos de operaciones y determinar como es el comportamiento entre los diversos actores que componen una red.

una forma de validación a la clasificación de los atributos anteriormente mencionada. Tal red puede ser lícita, ilícita o una mezcla de ambas.

Dos factores que permiten estudiar y conocer la evasión del IVA a través del uso de análisis de redes son: 1) la posible existencia de redes ilícitas donde están involucrados contribuyentes que emiten y reciben comprobantes fiscales con el único fin de erosionar la base gravable, ocasionando la evasión del IVA, y 2) la inmensa cantidad de datos estructurados que se generan con el uso de CFDI en la vida diaria.

2. Revisión del marco legal

En este. apartado se analizará el IVA, sus principales características, los sujetos obligados para su cumplimiento y sus obligaciones a cumplir. Asimismo, relacionado con la emisión del CFDI se encuentra el Código Fiscal de la Federación y el Anexo 20 de la Resolución Miscelánea Fiscal que establecen los requisitos para su emisión y tipos de CFDI, respectivamente.

2.1. El Impuesto al Valor Agregado (IVA)

De acuerdo al Centro de Estudios de Finanzas Públicas [12], el IVA es un impuesto indirecto que grava el consumo de los contribuyentes y no repercute directamente sobre los ingresos, sino que recae sobre los costos de producción y venta de las empresas y se traslada a los consumidores mediante los precios. Se dice que es un impuesto indirecto, pues el agente económico que lo recauda no es quien termina soportando la carga fiscal, además de no ser recaudado directamente por el ente fiscalizador, sino que es cobrado y enterado por el vendedor de un bien o servicio gravado al momento de la transacción comercial. Para GPM Contadores y Auditores S.C. [13] se considera un impuesto real debido a que esta directamente relacionado con el consumos de bienes y servicios independientemente de las circunstancias personales del contribuyente y por otra parte es un impuesto interno porque grava únicamente las operaciones llevadas a cabo dentro del territorio nacional (aunque cuando un producto se importa, dependiendo de su naturaleza, se grava conforme en términos de la ley).

Por otra parte, para los efectos del IVA en México [14], con base en el artículo 1 de la Ley del IVA, están considerados como obligados al pago del impuesto todas las personas físicas y morales que, en territorio nacional, realicen: la enajenación de bienes, presten servicios independientes, otorguen el uso o goce temporal de bienes (arrendamiento) o importen algún bien o servicio. Dentro del mismo artículo, se señala que el cálculo del impuesto resultará de aplicar a los valores que señala la Ley, la tasa del 16 % [8]. Asimismo se especifica que el impuesto al valor agregado determinado en ningún caso formará parte de dichos valores, es decir, no formará parte del valor del bien o servicio que sirvió como base para determinar el impuesto.

El Impuesto al Valor Agregado señalado en el párrafo anterior deberá trasladarse, en forma expresa y por separado, a quienes adquieran o arrenden bienes, o reciban los servicios que fueron pactados como parte de una operación o actividad económica. Para tener un poco mas en contexto a que se refiere la Ley con el traslado, éste se define como el cobro o cargo que el contribuyente debe hacer a quien entregó el bien o el servicio por un monto equivalente al impuesto establecido en la Ley, inclusive cuando se retenga el impuesto en términos de la misma.

En el mismo sentido, la Ley contempla un impuesto acreditable, el cual se debe de entender como aquel que fue

[8] La Ley del IVA contempla una tasa 0 % y exenciones, lo cual para propósitos de este estudio no es analizado

trasladado al contribuyente así como el propio impuesto que hubiese pagado con motivo de la importación de los bienes y servicios. Con lo anterior, aparece la figura del acreditamiento, el cual consiste en restar el impuesto acreditable, de la cantidad que resulte de aplicar a los valores señalados en esta Ley la tasa que corresponda.

De la diferencia de aplicar el impuesto trasladado al impuesto acreditable, se determina el impuesto a cargo que el contribuyente tiene que pagar ante las oficinas autorizadas, para esto también se resta el impuesto que se le hubiere retenido a dicho contribuyente durante el período del que se trate. De igual forma, como resultado de ésta diferencia, puede resultar un saldo a favor del contribuyente, esto quiere decir que en un período determinado el impuesto acreditable fue mayor al impuesto trasladado. Dicho saldo a favor puede solicitarse que se acredite contra un saldo de un impuesto a pagar a futuro en meses subsecuentes o se solicite la devolución total del saldo a favor determinado.

Ahora bien, para considerar que el impuesto sea trasladado o acreditable se deben cumplir ciertas disposiciones establecidas en la Ley objeto del análisis. En el caso del impuesto trasladado, para que sea considerado efectivamente trasladado, se prevé que la contraprestación pactada por los actos o actividades gravadas para efectos del IVA estén efectivamente cobradas y que el impuesto se encuentre trasladado en forma expresa y por separado.

En el caso del impuesto acreditable para que el impuesto sea considerado para su acreditamiento deben cumplirse los siguientes requisitos:

- Que el impuesto al valor agregado corresponda a actividades estrictamente indispensables por las que deba pagarse el impuesto establecido en la Ley. En este caso se entiende como estrictamente indispensables aquellas erogaciones efectuadas que sean deducibles para los fines del Impuesto Sobre la Renta (ISR), aún y cuando no se esté obligado al pago de dicho impuesto.
- Que el impuesto al valor agregado se encuentre trasladado expresamente en los comprobantes fiscales mencionados en el artículo 32 fracción III de la Ley.
- Que el impuesto trasladado al contribuyente haya sido efectivamente pagado en el mes de que se trate.
- Se enteren, en su caso las retenciones del impuesto al valor agregado trasladado en los términos y plazos establecidos en la Ley.

Dicho lo anterior se puede observar que al estar involucradas dos partes dentro de un mismo acto que causa el traslado y el acreditamiento del impuesto, ambos derivados de llevar a cabo alguna de las actividades mencionadas al inicio de este apartado, el IVA ocasiona una trazabilidad que permite visualizar el resultado final vinculado a las referencias especificas generadas por las actividades sujetas al IVA y que están expresadas en los comprobantes fiscales que para tal efecto contempla la Ley.

Asimismo, la Ley del IVA contempla ciertas obligaciones en específico que tienen que ser cumplidas por los sujetos señalados para acatar las disposiciones establecidas [14]. Las obligaciones más relevantes para efectos de este estudio son:

- Expedir y entregar comprobantes fiscales.
- Expedir comprobantes fiscales por las retenciones del impuesto y proporcionar mensualmente a través de los medios electrónicos que señale el Servicio de Administración Tributaria, la información sobre a quienes se les retuvo el impuesto establecido en la Ley.

- Proporcionar mensualmente a través de los medios electrónicos que para tal efecto el SAT señale, la información sobre el pago, retención acreditamiento y traslado del impuesto al valor agregado en las operaciones con sus proveedores, dentro de la cual, se desglosa el valor de los actos o actividades por las que el contribuyente esta obligado al pago del impuesto. Esta información se presentará mas tardar el día 17 del mes inmediato posterior al que corresponda la información.

Para esta última obligación mencionada, el SAT a efecto de que los contribuyentes puedan cumplir con dicha disposición, tiene dentro de su portal web la forma o formato electrónico A-29, el cual denomina "Declaración Informativa de Operaciones con Terceros" (DIOT)[9]

Esta declaración busca dar un panorama más profundo y detallado de las declaraciones mensuales de IVA que presentan tanto personas físicas como morales. Es por esto que dentro de la DIOT es necesario señalar el RFC, los montos de las actividades por las cuales se pago el IVA al 16 por ciento, los montos de las actividades pagadas al IVA del 0 por ciento o los montos de las operaciones por los cuales estuvo exento al pago del Impuesto. Lo anterior conlleva a que el SAT tenga prácticamente en tiempo real no solamente los montos de IVA declarados, sino que también puede contar y determinar la veracidad de los montos declarados para este impuesto.

2.2. Emisión de comprobantes fiscales

De acuerdo al SAT en el Anexo 20 de la RMF [15], los comprobantes fiscales deben emitirse por los actos o actividades que se realicen, por los ingresos que perciban o por las retenciones de contribuciones que efectúen los contribuyentes ya sean personas físicas o morales. Asimismo, expedir CFDI, es una obligación de los contribuyentes personas físicas y morales de conformidad con el artículo 29, párrafos primero y segundo, fracción IV y penúltimo párrafo del CFF y 39 del Reglamento del CFF, en relación con la regla 2.7.5.4., y el Capítulo 2.7 De los Comprobantes Fiscales Digitales por Internet o Factura Electrónica de la Resolución Miscelánea Fiscal vigente.

Para el caso de la Ley del IVA se tiene la obligación de expedir un comprobante fiscal en los siguientes casos particulares:

- En el caso de factoraje financiero, se tiene que expedir el estado de cuenta de acuerdo a lo establecido en el artículo 29-A del CFF.
- Para el caso de los contribuyentes del Régimen de Incorporación Fiscal se emiten los comprobantes fiscales de acuerdo a lo establecido a las fracciones II y IV del artículo 112 de la Ley del ISR.

[9]Desde su implementación la DIOT ha sufrido algunos cambios respecto de la obligación en su presentación, los cuales al ser de carácter correctivo en el formato, en la carga o en la forma del envío, no han sido considerablemente importantes para su cumplimiento. La Regla 2.8.4.3. de la Resolución Miscelánea Fiscal publicada en el DOF el 30 de diciembre de 2015 estableció de manera indirecta que la presentación de la DIOT sustituía a la presentación de la Declaración Anual Informativa de Clientes y Proveedores, ya que, para quien había cumplido con presentar la DIOT por cada uno de los meses anteriores al año 2015, esta Regla en la Resolución Miscelánea exentaba de la presentación de la Declaración Anual Informativa de Clientes y Proveedores.

El SAT puso a disposición a través de su página web en la sección "Mis cuentas", el clasificador del gasto, en donde de manera automática, aparecían todos los CFDI emitidos y recibidos durante el mes para presentar la declaración correspondiente[10]. Al hacer uso de este "formulario prellenado" que contenía los CFDI que amparaban los ingresos y deducciones autorizadas, se tenía la opción de no presentar la DIOT, lo anterior con el fin de facilitar e incentivar a que los contribuyentes usaran la información generada de manera automática por el SAT. Está opción fue eliminada, derivado de entre otras cosas, que se incumplía con el principio de la auto determinación de las contribuciones de acuerdo a lo establecido en el artículo 6 del CFF.

- En el caso del artículo 32 fracción III como parte de las obligaciones de los contribuyentes para el cumplimiento de la Ley del IVA se deben expedir y entregar comprobantes fiscales.
- En el caso del articulo 32 fracción V, por las retenciones que se efectúen en los casos previstos en el artículo 1-A de la Ley del IVA.
- Cuando se venda un bien o se preste un servicio de forma accidental.

Hasta este punto solamente se ha estado mencionando el término "comprobante fiscal" el cual, está así considerado dentro de la Ley del IVA. El término que se le ha dado al comprobante fiscal al paso del tiempo es muy diverso y no es materia de este estudio profundizar en ello, sin embargo la definición de comprobante fiscal se establece en el artículo 29 del CFF, el cual menciona: Cuando las leyes fiscales establezcan la obligación de expedir comprobantes fiscales por los actos o actividades que realicen, por los ingresos que se perciban o por las retenciones de contribuciones que efectúen, los contribuyentes deberán emitirlos mediante documentos digitales a través de la página de Internet del SAT. Las personas que adquieran bienes, disfruten de su uso o goce temporal, reciban servicios o aquéllas a las que les hubieren retenido contribuciones deberán solicitar el comprobante fiscal digital por Internet respectivo.

De lo anterior podemos concluir que la emisión de comprobantes fiscales que se refiere la Ley del IVA, se deberán realizar mediante documentos digitales a través de la página de internet del SAT y que quien reciba dicho documento obtendrá el respectivo Comprobante Fiscal Digital por Internet (CFDI).

2.3. Comprobante Fiscal Digital por Internet (CFDI)

El CFDI es un documento XML[11] que contiene ciertos requisitos y particularidades que el SAT publica anualmente mediante el Anexo 20 de la RMF, "Guía de llenado para Comprobantes Fiscales". Este esquema de facturación electrónica se caracteriza por tener un sello de certificación o "Timbre" que únicamente puede ser emitidos por los Proveedores Autorizados de Certificación (PAC) avalados por el SAT. El comprobante describe el bien o servicio adquirido, la fecha de la transacción, su costo, y desglosa los impuestos correspondientes al pago, así como las retenciones que en su caso proceden a efectuarse [16].

El CFDI ofrece ventajas directas como: identificar a los participantes en distintas transacciones comerciales, comprobar las transacciones que pagan impuestos, ayuda a investigaciones de lavado de dinero, recuperar información de transacciones y como consecuencia de su uso se puede evitar la evasión fiscal y determinar redes de interacción entre quien emite y recibe el CFDI.

La emisión del CFDI legalmente esta regulado principalmente por el Código Fiscal de la Federación, en los artículos 27, 29, 29-A y 69-B, en relación con lo establecido con el capítulo 2.7 de la Resolución Miscelánea Vigente. En dichos artículos se hace la precisión de qué es un CFDI y en que casos se tiene que expedir (artículo 29), que requisitos debe cumplir su emisión (artículo 29-A), en que momento se considera un comprobante que ampara una operación simulada o inexistente (artículo 69-B), de entre otras especificaciones.

A su vez, para efectos de la Ley del Impuesto al Valor Agregado, de acuerdo a la fracción II del artículo 5, el CFDI es utilizado como un medio para trasladar y acreditar el impuesto en términos de la propia Ley. Asimismo,

[11] Este comprobante fiscal es el mas reciente y el mas moderno. Actualmente se utiliza la versión 3.3 la cual fue publicada desde el año 2017. La diferencia con el CFD, su antecesor que dejo ser usado a finales de 2013, consiste en que una vez que se emite el documento se envía a un proveedor de certificación, quien le asigna un folio fiscal, verifica que cumpla con los requisitos del comprobante, lo sella digitalmente y lo regresa al emisor.

en la fracción III del artículo 32 se establece la obligación de expedir y entregar CFDI con el fin de cumplir con las obligaciones correspondientes al IVA. Aquí, con una relación estrecha con el Código Fiscal de la Federación, el CFDI busca satisfacer una necesidad tributaria mediante la cual, el contribuyente puede cumplir sus obligaciones. Estas obligaciones en términos del IVA, prácticamente son aquellas de trasladar y acreditar el Impuesto y dependiendo de su actividad económica, el contribuyente puede expedir y recibir diversos tipos de CFDI.

De acuerdo al Anexo 20 "Guía de llenado de los comprobantes fiscales por internet", publicado por el SAT, existen cinco tipos distintos de CFDI, los cuales se mencionan a continuación:

1. *Comprobante de ingreso*: Se emiten por los ingresos que obtienen los contribuyentes, ejemplo: prestación de servicios, arrendamiento, honorarios, donativos recibidos, enajenación de bienes y mercancías, incluyendo la enajenación que se realiza en operaciones de comercio exterior, etc.

2. *Comprobante de egreso*: Amparan devoluciones, descuentos y bonificaciones para efectos de deducibilidad y también puede utilizarse para corregir o restar un comprobante de ingresos en cuanto a los montos que documenta, como la aplicación de anticipos. Este comprobante es conocido como nota de crédito.

3. *Comprobante de traslado*: Sirve para acreditar la tenencia o posesión legal de las mercancías objeto del transporte durante su trayecto. De este tipo de CFDI se puede expedir de dos maneras: Emisión de CFDI de traslado por el propietario de las mercancías cuando las transporte el mismo o Emisión de CFDI por el transportista, siempre que el propietario de las mercancías contrate los servicios de transportación.

4. *Comprobante de Recepción de pagos*: Es un CFDI que incorpora un complemento para recepción de pagos, el cual debe emitirse en los casos de operaciones con pago en parcialidades o cuando al momento de expedir el CFDI no reciban el pago de la contraprestación y facilita la conciliación de las facturas contra pagos.

5. *Comprobante de Nómina*: Es un CFDI al que se incorpora el complemento recibo de pago de nómina, el cual debe emitirse por los pagos realizados por concepto de remuneraciones de sueldos, salarios y asimilados a estos, es una especie de una factura de egresos.

La emisión de cualquier tipo de CFDI satisface una necesidad tributaria para los contribuyentes, ya que el CFDI es el medio de comprobación con el que fiscalmente podrán dentro de sus obligaciones fiscales calcular, presentar, cumplir y en su caso pagar o determinar un impuesto a favor. Actualmente los contribuyentes (de acuerdo a la magnitud de su actividad económica) establecen para cumplir en el ramo fiscal estrategias integrales que les permita mantener un equilibrio para asegurar la manutención económica de su actividad y a la par, seguir contribuyendo de acuerdo a las disposiciones fiscales a las que estén sujetos. Este tipo de estrategias se han mantenido a lo largo del tiempo en una dualidad que las divide de una interpretación correcta e incorrecta de los lineamientos fiscales. Dentro de estas estrategias se encuentran aquellas donde se emiten CFDI sin ningún tipo de respaldo económico, técnico, de infraestructura, personal a cargo o sin otro tipo de justificación y debido a esto, el CFF prevé en el artículo 69-B en que momento y bajo que procedimientos, las autoridades fiscales podrán determinar y sancionar a este tipo de contribuyentes[12].

Este artículo, ha sido un medio de control como consecuencia de los esquemas agresivos de evasión fiscal que implementan algunos contribuyentes, a través de la simulación de comprobantes fiscales [17]. En el mismo sentido,

[12]A finales del 2013 se dieron importantes modificaciones a distintos ordenamientos fiscales (Ley del Impuesto sobre la Renta, Ley del Impuesto al Valor Agregado, Código Fiscal de la Federación, entre otros), las cuales fueron impulsadas con motivo de las reformas planteadas por el titular del Ejecutivo Federal a cargo en ese entonces, entre ellas se encontró la adición del artículo 69-B al CFF.

este artículo estipula el procedimiento encaminado a detectar y sancionar tanto a los contribuyentes que expiden comprobantes derivados de operaciones inexistentes, así como a quienes reciben estos comprobantes para generar un efecto fiscal a su favor[13]. Dicho procedimiento principalmente se refiere a:

- El momento en que la autoridad fiscal podrá determinar la inexistencia de las operaciones amparadas con este tipo de comprobantes.
- Mediante el buzón tributario, notificación personal y una publicación en el DOF, se notificará a los contribuyentes que se encuentran dentro del supuesto que considera este artículo.
- Los contribuyentes que fueron notificados por estar dentro del supuesto de emitir comprobantes simulados, podrán aclarar su situación para ser desvirtuados de este supuesto, tienen 15 días para tal efecto y tienen la posibilidad de solicitar prórroga por 5 días adicionales.
- La consideración de quienes hayan recibido comprobantes simulados y les hayan dado un efecto fiscal, podrán corregir su situación fiscal mediante la declaración o declaraciones complementarias que correspondan.
- La determinación de los créditos fiscales que las autoridades pueden calcular de acuerdo a sus facultades de comprobación y la consideración como actos o contratos simulados la emisión de comprobantes fiscales de este tipo para efecto de los delitos previstos en el CFF.

De acuerdo a lo señalado en los puntos anteriores, estos lineamientos son los que actualmente se han estado aplicando con el fin de poder determinar la materialidad de las operaciones y para generar la detección de quienes estén operando bajo estos esquemas de emisión de comprobantes. Antes de concluir este estudio, el Congreso de la Unión aprobó algunas otras medidas que endurecen y amplían el combate a este tipo de esquemas, de estos cambios aprobados al artículo 69-B del CFF[14] destacan los siguientes [18]:

- La cancelación de certificados de sello digital cuando: en el ejercicio de sus facultades, la autoridad fiscal detecte que el contribuyente no puede ser localizado en su domicilio fiscal, desaparezca durante el procedimiento, desocupe su domicilio fiscal sin presentar el aviso de cambio correspondiente en el registro federal de contribuyentes, se ignore su domicilio, o bien, dentro de dicho ejercicio de facultades se tenga conocimiento de que los comprobantes fiscales emitidos se utilizaron para amparar operaciones inexistentes, simuladas o ilícitas.
- La autoridad detecte que el contribuyente emisor de comprobantes fiscales no desvirtuó la presunción de la inexistencia de las operaciones amparadas en tales comprobantes y, por tanto, se encuentra definitivamente en dicha situación, en términos del artículo 69-B, cuarto párrafo del CFF.
- La autoridad detecte que se trata de contribuyentes que dieron efectos fiscales a comprobantes expedidos por un contribuyente que aparece en el listado definitivo de contribuyentes que realizan operaciones simuladas, y que en un plazo de 30 días contados a partir de la publicación del listado no acreditaron la efectiva adquisición de los bienes o recepción de los servicios, ni corrigieron su situación fiscal.
- La denominación y creación del Tercero Colaborador Fiscal, el cual, es aquella persona que proporcione a la autoridad fiscal información y documentación necesaria para substanciar el procedimiento establecido en el artículo 69-B del CFF (operaciones inexistentes), así como para motivar las resoluciones del mismo.

[13] Este artículo ha sufrido diversas modificaciones desde 2013, sin embargo para efectos del presente estudio debido a la importancia y al momento de su modificación, se centra en la modificación publicada en el DOF el 25 de junio de 2018 del artículo 69-B.

[14] Los cambios fueron aprobados el día 31 de octubre de 2019 y entraran en vigor a partir del 1 de enero del 2020.

3. Experiencias internacionales de investigación

La mayoría de los métodos estadísticos que se han aplicado para la detección de fraude fiscal pueden clasificarse en dos categorías. Primero se encuentran amplias aplicaciones de la ley de Newcomb-Benford, la cual permite detectar anomalías en conjuntos grandes de números que siguen la ley. En Polonia, se aplicó a datos entre 2009-2015 de empresas de venta al por mayor [19], encontrando una posible manipulación de datos por algunas empresas. También se ha aplicado esta ley sobre datos agregados de reportes de impuestos entre 2007-2011 para todas las regiones de Italia [20], hallando anomalías en ciertas regiones. Otra aplicación interesante es sobre datos del US Bank Holding Company antes y durante la crisis financiera del 2000 [21], donde se concluye que hubo una tendencia a manipular reportes del tamaño y ganancias de bancos con dificultades financieras. También se ha aplicado para estudiar donativos de campañas electorales en USA [22], Puerto Rico y Venezuela [23].

La segunda categoría de métodos para detección de fraude fiscal se basa en otros análisis estadísticos, particularmente en hallar anomalías estadísticas comparando los datos observados con valores esperados [24]. Esta categoría incluye muchos métodos para clasificación estadística, entre los que destacan las redes neuronales. Éstas se han empleado repetidamente para detectar fraude en transacciones de tarjetas de crédito [25–27], así como para detectar firmas que reportan estados financieros fraudulentos en Grecia [28].

Otros métodos que cabe mencionar incluyen los basados en reglas, los cuales producen clasificadores que emplean reglas de inferencia o condiciones para filtrar los datos, por ejemplo los clasificadores Bayesianos, o los árboles de decisión [29]. También se ha aplicado análisis de enlaces en el ámbito de telecomunicaciones [30] para generar y estudiar comunidades de interés alrededor de individuos con comportamiento fraudulento.

En cuanto a métodos basados en *inteligencia computacional*, a pesar de que no hay estudios definitivos en la literatura, hay algunos casos particulares, incluyendo el uso de co-evolución en Estados Unidos [31], redes neuronales artificiales en Malasia [32], un modelo híbrido en Irán [33] y minería de datos en Brasil [34]. En cuanto a la ciencia de redes, se ha usado para modelar el fenómeno de la corrupción a nivel gobierno [35], pero a nuestro saber este es el primer estudio donde se usa para detectar evasión fiscal.

El presente estudio se distingue de los anteriores por diversas razones. Las principales son la gran cantidad de datos y la colaboración directa con la autoridad recaudadora, lo cual permite que los resultados tengan un impacto directo en la fiscalización. Los resultados presentados a continuación podrían tener un impacto internacional, ya que los métodos explorados podrían servir a otros países.

4. Descripción de los datos suministrados

Los datos suministrados por el SAT son:

- Un catálogo de RFC anonimizados (RFCA). Es decir, un conjunto de RFC que son encriptados para proteger la identidad de las personas físicas y morales sujetas a este estudio.
- Un conjunto de CFDI agregados por mes, correspondientes a cada par emisor-receptor que tuvieron alguna transacción en el periodo estudiado.
- Una lista de RFCA identificados como EFOS o presuntos EFOS.

En este estudio utlizamos un catálogo de 81,511,015 RFCA con la siguiente información: tipo, situación y estado del contribuyente, entidad federativa, municipio, fecha de inicio de operaciones, sector y actividad. Tenemos los datos correspondientes a los CFDI de enero 2015 a diciembre 2018, agregados por mes para cada par emisor-receptor. Los datos tienen 6,823,415,757 registros con los siguientes campos: RFCA del emisor, RFCA del receptor, ejercicio, periodo, tipo, número de facturas activas, número de facturas canceladas, monto total activo, monto total cancelado, monto subtotal activo, monto subtotal cancelado, monto descuento activo, monto descuento cancelado, monto IVA trasladado activo, monto IVA trasladado cancelado, monto IEPS trasladado activo, monto IEPS trasladado cancelado, monto total trasladado activo, monto total trasladado cancelado, monto IVA retenido activo, monto IVA retenido cancelado, monto ISR retenido activo, monto ISR retenido cancelado, monto total retenido activo, monto total retenido cancelado, monto total parcial activo, y monto total parcial cancelado.

Contamos con una lista de 8,570 RFCA identificados anteriormente por el SAT como EFOS definitivas y 1,488 RFCA que presuntamente son EFOS [15]. En los 48 meses proporcionados de actividad se encuentran 7,571,093 RFCA con al menos una factura, por lo que las EFOS definitivas representan el 0.0046 % del total, las EFOS presuntas el 0.0028 %, y el resto componen el 99.94 % de los contribuyentes considerados. Las cifras anteriores indican que los datos están desbalanceados: las proporciones entre la clase identificada (EFOS) y la no identificada (desconocida) son distintas. Esto tiene un impacto en el diseño de la solución del presente caso de estudio, lo cual hemos implementado de forma satisfactoria.

En lo sucesivo cuando se mencione a una EFOS, ya sea definitiva o presunta, se hará referencia a las que ya han sido identificados por el SAT y que fueron suministradas para la realización de este estudio. Cuando hagamos mención a un RFCA desconocido, nos referimos a todos aquellos RFCA que no han sido clasificados como EFOS (presuntas o definitivas) por el SAT. Finalmente, también se suministraron los datos correspondientes a las declaraciones DIOT, las declaraciones de IVA y los datos de los saldos a favor generados desde el año 2015 al año 2018. [16]

5. Metodología de investigación

El estudio de los datos se realizó inicialmente con tres diferentes métodos que posteriormente fueron integrados para proveer una lista de RFCA sospechosos de ser EFOS. Primero, construimos redes de interacción entre contribuyentes conectados de acuerdo a las emisiones y recepciones de CFDI que realizan, a partir de EFOS. Esto nos permitió reconocer comportamientos típicos de emisión de CFDI y mecanismos de asociación de las EFOS ya identificadas y también encontrar RFCA con patrones similares dentro de las redes. Posteriormente, se implementaron dos métodos diferentes e independientes de aprendizaje automatizado basados en metodologías distintas al análisis de redes, que permiten detectar patrones en los registros de emisiones de CFDI y realizar una clasificación de RFCA desconocidos entre sospechosos y no sospechosos de presentar un comportamiento similar al de las EFOS ya identificadas. Estos métodos fueron integrados mediante un índice de cercanía a las EFOS para cada RFCA, que representa su nivel de colusión dentro de las emisiones de las EFOS. Finalmente, estimamos montos de evasión de IVA con base en los resultados anteriores.

[15] Las EFOS proporcionadas corresponden a las identificadas por el SAT hasta Octubre de 2019

[16] Estos datos fueron explorados pero no se consideraron para el calculo de los montos estimados de evasión que se reportan en el estudio.

5.1. Ciencia de redes

En esta sección se describe la forma en la que se construyen redes de interacción entre EFOS y RFCA desconocidos basadas en las emisiones y recepciones de CFDI entre ellos. Así mismo, se describen los análisis realizados sobre la estructura de los enlaces y el papel de las EFOS y el resto de los RFCA en las redes de interacción. Dicho análisis nos permite definir medidas que posteriormente serán útiles para el cálculo de un estimado de evasión de IVA.

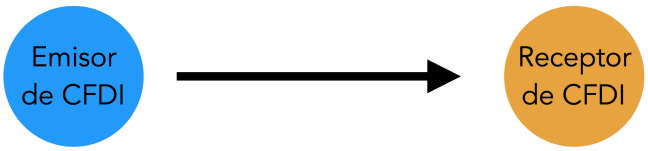


Figura 1: Un enlace dirigido en la red de interacción corresponde a un comprobante fiscal emitido entre contribuyentes. Dichos comprobantes pueden ser de tipo ingreso, egreso o traslado.

5.1.1. Definición de las redes de interacción entre contribuyentes

El registro de la actividad económica de los contribuyentes por medio de la emisión de comprobantes fiscales nos permite definir redes de interacción, las cuales están compuestas por nodos y enlaces. Cada nodo en la red corresponde a un contribuyente (identificado por su RFCA) el cual se etiqueta en una de tres categorías: EFOS definitivos (aquellos ya publicados por el SAT), EFOS presuntos (actualmente bajo sospecha de ser EFOS, pero aun sin certeza) y RFCA desconocidos (el resto). Los enlaces en la red[17] corresponden a emisiones dirigidas de CFDI entre contribuyentes, véase la figura 1.

Una vez hemos definidos nodos y enlaces, tenemos una red. La estructura de dicha red describe, entre otras cosas, las relaciones entre *grupos de contribuyentes*. Asumimos que el estudio de esta estructura permite realizar una caracterización de algunos patrones de asociación en los que han operado las EFOS, y los contribuyentes a su alrededor, de forma histórica. La detección de estos patrones nos permitió identificar contribuyentes con actividad sospechosa.

Partiendo de los datos proporcionados construimos redes de interacción anuales y mensuales. Por un lado, a escala anual consideramos las emisiones y recepciones desde y hacia EFOS, lo cual nos permite identificar los RFCA desconocidos con los que interactúan de forma regular y conjuntos de RFCA desconocidos y EFOS que conforman subredes organizadas de emisión y recepción de operaciones potencialmente simuladas. Por otro lado, hemos identificado que a escala mensual, los montos totales asociados a los CFDI emitidos por EFOS (definitivas y presuntas) ocurren con mayor frecuencia dentro de un intervalo que hemos denominado como el *nivel de operaciones de EFOS*. Considerando los CFDI (enlaces) dentro de este nivel de operaciones construimos las redes de interacción mensual. De manera análoga a las redes de interacción anual, identificamos subredes y cuantificamos la importancia de los nodos dentro de la red por medio del cálculo de *medidas de centralidad*, con el fin de evaluar su utilidad para la descripción del comportamiento de EFOS.

En las siguientes secciones se describen detalladamente los distintos análisis realizados de las redes de interacción entre contribuyentes, así como los resultados obtenidos.

[17]Los enlaces en la red, al igual que los comprobantes fiscales emitidos, pueden ser de tipo ingreso, egreso o traslado. En este estudio nos restringimos a los enlaces asociados a emisiones y recepciones de CFDI de tipo ingreso.

5.1.2. Redes de interacción anual

Como primer ejercicio, consideramos la red de interacción anual inducida por el agregado de CFDI emitidos y recibidos por todos los nodos asociados a EFOS, tomando solamente en consideración los enlaces de tipo ingreso que tienen asociados al menos 10 facturas durante un año y montos positivos[18]. Esta restricción selecciona los enlaces entre EFOS y otros RFCA que interactúan con mayor frecuencia durante el año los cuales, de acuerdo al principio de homofilia [36–38] en redes sociales, estarían asociados a nodos que son más parecidos entre ellos.

En la red de interacción anual identificamos componentes fuertemente conectados que se relacionan con *subredes de operación* en las se organizan contribuyentes con actividad fiscal anómala. Pare definir un componente fuertemente conectado, necesitamos introducir la noción de camino. Entre dos nodos existe un camino, si es posible ir de un nodo al otro mediante los enlaces dirigidos en la red. Un componente fuertemente conectado es aquel en el que existen caminos, en ambas direcciones, entre cualquier pareja de nodos. En la figura 2 se muestran los componentes fuertemente conectados más grandes en los años de 2015 y 2016. Recordando que los enlaces en la red están asociados a comprobantes de intercambios potenciales de bienes o servicios, la presencia de este tipo de estructuras implica un flujo circular dentro de un conjunto de contribuyentes. Al estar construidas estas redes alrededor de nodos asociados a EFOS, es posible que estén asociadas al intercambio de comprobantes de operaciones simuladas.

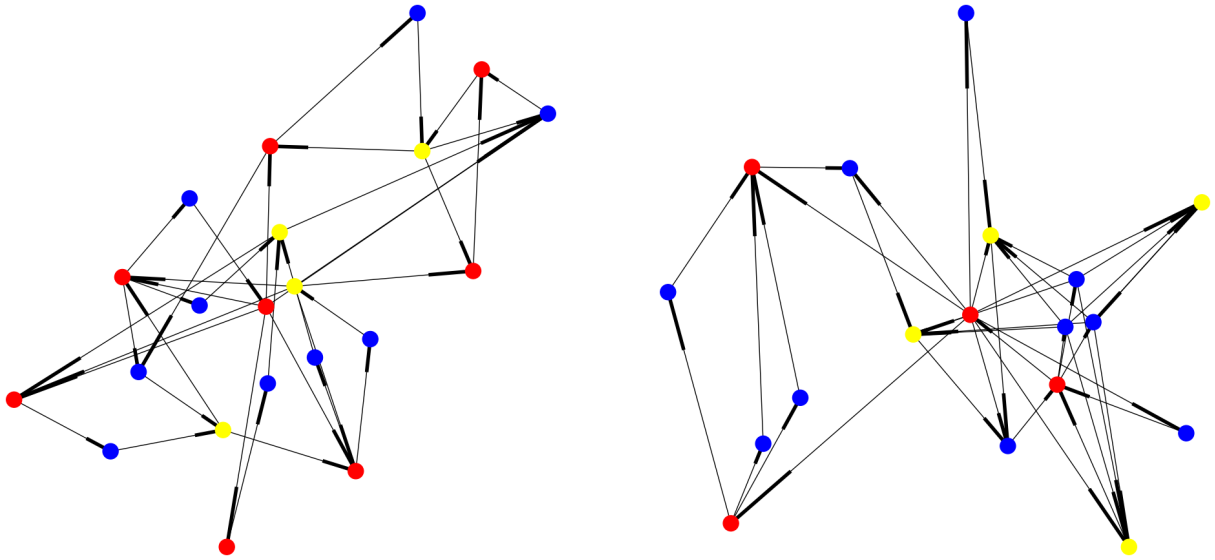


Figura 2: Ejemplos de subredes de operación en las redes de interacción agregadas anuales en 2015 (izquierda) y 2016 (derecha). Los nodos rojos corresponden a EFOS definitivas, amarillos a presuntas y los azules a RFCA desconocidos.

Una característica notable de las subredes de operación identificadas, es la proporción de los tipos de nodos en ellas. Como se puede observar en la figura 3, la mayor parte de los nodos en las subredes detectadas corresponden a RFCA desconocidos, lo cual sugiere que dichos contribuyentes al ser partícipes de la emisión o recepción de CFDI relacionados con EFOS, podrían también tener una actividad fiscal anómala.

Este método en el que se toman como *semillas* a los nodos asociados a EFOS y sus emisiones o recepciones de

[18]En los datos proporcionados también se incluyen facturas canceladas que no fueron tomadas en cuenta para la construcción de las redes.

CFDI para construir redes de interacción anuales, nos permitió identificar comunidades de operación asociadas a transacciones potencialmente simuladas alrededor de EFOS. Usando esta información se podría tener una idea de los mecanismos de organización de este tipo de contribuyentes.

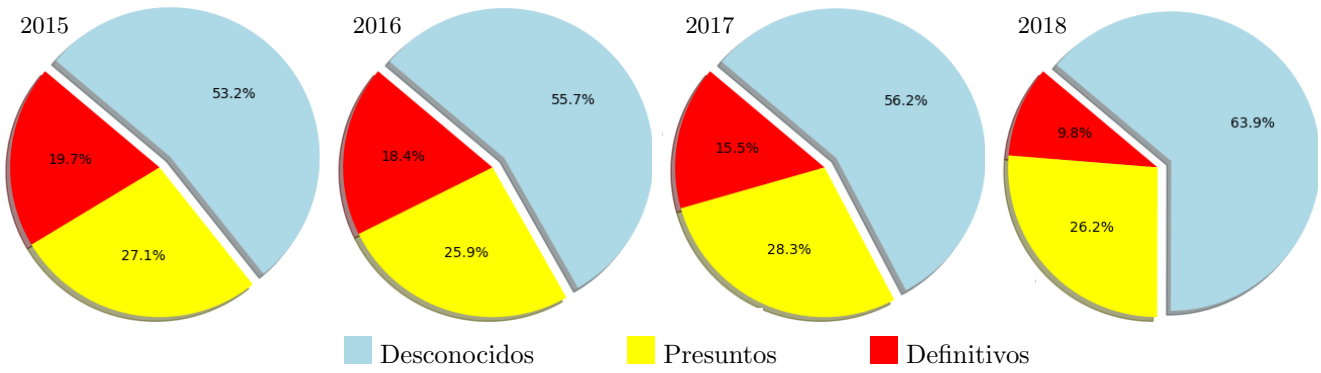


Figura 3: Proporción de tipos de nodos presentes en las subredes de operación en las redes de interacción agregadas anuales de 2015 a 2018. Cabe hacer notar que la mayoría de los nodos presentes en las subrededes corresponden a RFCA desconocidos, lo que nos muestra que partiendo de una población pequeña de EFOS nos es posible identificar posibles RFCA sospechosos.

5.1.3. Redes de interacción mensual

En esta sección estudiamos las redes inducidas por actividad mensual. A diferencia de las redes agregadas anuales, consideramos ahora enlaces asociados a emisiones y recepciones entre todos los tipos de nodos (EFOS definitivas, presuntas y RFCA desconocidos). Sin embargo, dado que el conjunto total de emisiones de CFDI en un mes es muy grande, es necesario definir un criterio para reducir la red a un tamaño manejable.

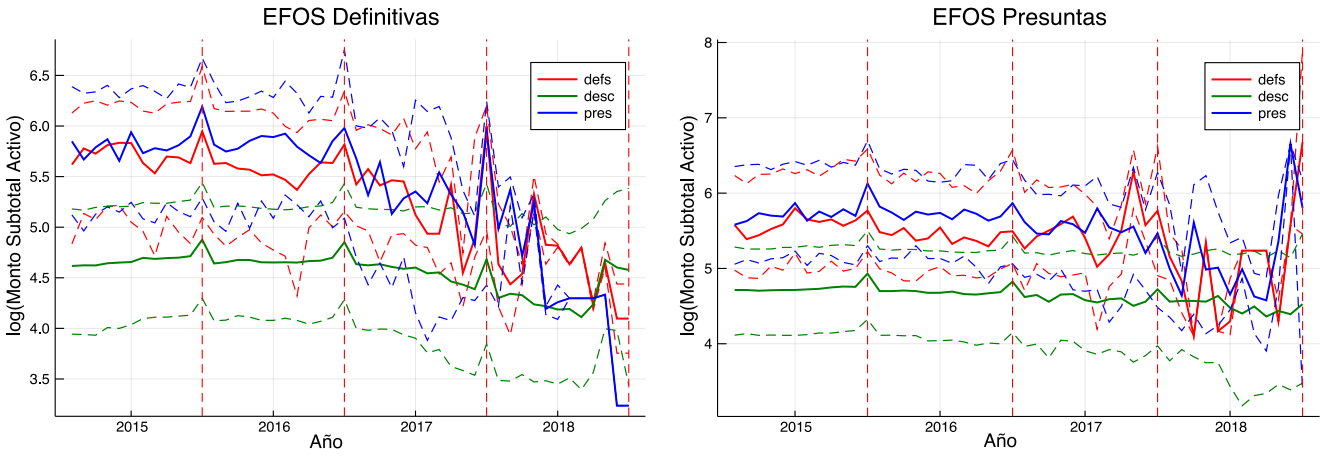


Figura 4: Comportamiento temporal del logaritmo del Monto Subtotal Activo asociado a emisiones desde EFOS definitivas y presuntas hacia los tipos restantes de nodos. Las líneas sólidas muestran la media y las líneas punteadas al rango intercuartil de la distribución. Se observa que comprobantes emitidos por las EFOS, ya sean definitivas o presuntas, corresponden típicamente a montos en el rango entre diez mil y un millón de pesos. Definimos a este rango como el régimen de actividad de las EFOS. Las líneas punteadas verticales corresponden al mes de Diciembre de cada año.

Con el fin de seleccionar los enlaces más relevantes calculamos la distribución de los Montos Totales de los comprobantes emitidos por EFOS (definitivas o presuntas) hacia los demás tipos de nodos. Como se muestra en la figura 4, la media de la distribución cambia en el tiempo, mostrando un aumento a fin de año. Cabe hacer notar que las transacciones que las EFOS realizan entre ellas corresponden a montos mayores que los que emiten hacia contribuyentes no etiquetados, i.e. las EFOS realizan *emisiones diferenciadas* según si los receptores son otras EFOS (ya sean presuntas o definitivas) o RFCA desconocidos. Definimos como *nivel de actividad* de EFOS al intervalo de montos definido por los rangos intercuartiles de las distribuciones asociadas a las emisiones desde EFOS, el cual utilizamos para seleccionar los enlaces que conforman las redes de interacción mensual, tomando en cuenta solamente los enlaces que corresponden a operaciones con montos dentro del nivel de actividad.

Con los enlaces seleccionados, construimos redes mensuales y calculamos el componente fuertemente conectado más grande. Por ejemplo, para el mes de diciembre de 2015 consta de transacciones entre 653,588 contribuyentes; obtenemos valores similares para los meses restantes de todos los años. Cabe recordar que debido a la definición de un componente fuertemente conectado estos 600 mil contribuyentes forman parte de un flujo circular de emisiones de CFDI en el que está involucrado un conjunto de EFOS. Sin embargo, con la información que tenemos no es posible identificar cuáles de los enlaces en la red están asociados a operaciones simuladas[19].

Definiremos el *alcance promedio* de un conjunto de nodos, para continuar caracterizando la estructura de la red alrededor de las EFOS. La distancia entre nodos se define como mínimo de pasos que se deben de dar siguiendo los enlaces de la red para llegar de un nodo a otro. El alcance, $R_i(d)$, para el nodo i, es el número de nodos a una distancia d (o menor) de dicho nodo, mientras que el alcance promedio $R(d)$ es simplemente el promedio de los alcances $R_i(d)$, de los nodos de algún conjunto seleccionado; por ejemplo, podemos hablar del alcance de las EFOS.

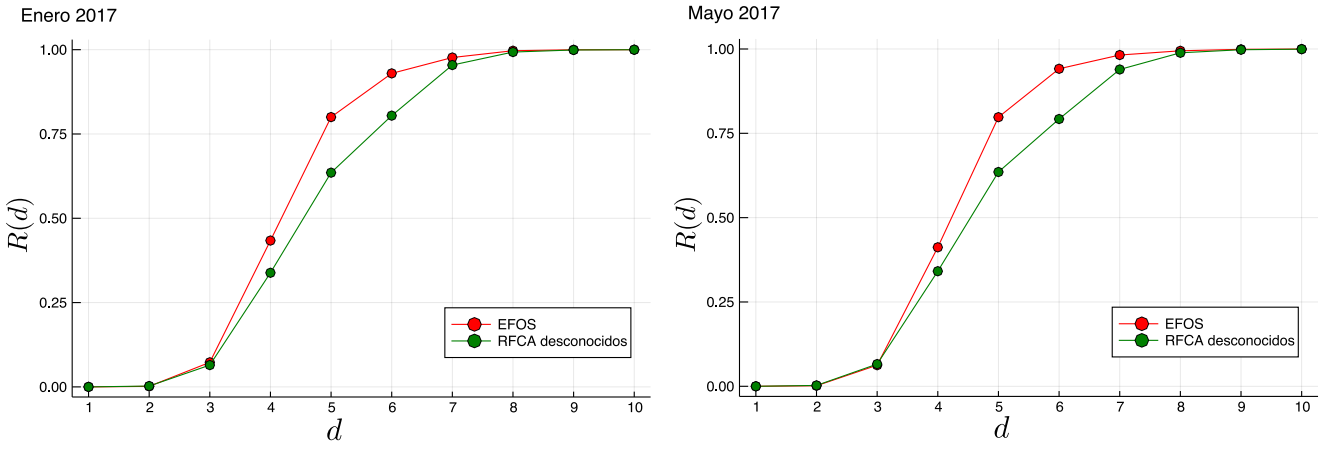


Figura 5: Alcance en la red en función de la distancia d para nodos asociados a EFOS y RFCA no etiquetados. Se observa que el alcance de las EFOS es mayor que el de los RFCA desconocidos para distancias intermedias y de más del 75 % para $d \geq 5$. La curva asociada a los RFCA desconocidos corresponde al promedio sobre 10 muestras aleatorias del mismo número que EFOS en la red. Los datos mostrados corresponden a enero (izquierda) y mayo (derecha) de 2017.

La estructura de la red de interacción mensual es tal que, como se muestra en la figura 5, partiendo de la mayoría de nodos asociados a una EFOS es posible llegar a más del 75 % de los nodos en la red realizando cinco pasos. También

[19]Consideramos que dicha identificación es difícil y requiere de información más detallada de los comprobantes emitidos para reconocer operaciones/RFCA fraudulentos. De igual forma, no es posible asegurar que todos los nodos asociados a RFCA presentes en la red están involucrados en actividades de evasión de impuestos, por lo que utilizamos dos métodos distintos para identificar a los RFCA que con mayor probabilidad podrían ser EFOS.

se puede observar que la curva asociada a EFOS se encuentra arriba de la asociada a RFCA desconocidos, que se puede interpretar como una mayor eficiencia de las EFOS para distribuir sus operaciones en la red. Esta observación sugiere un mecanismo de operación de las EFOS con el posible objetivo de limitar la trazabilidad de sus operaciones. Por otro lado, el número de EFOS cercanas a un RFCA arbitrario dentro de la red, es un indicador del nivel de colusión de un RFCA dentro de la red de operaciones de las EFOS. Esto se puede determinar de forma mensual o considerar el total de EFOS cercanas a un RFCA durante un año. Como se muestra en la figura 6, hay casos en los que las EFOS cercanas a un RFCA son más de 20 e incluso pueden llegar a 100 en un mes.

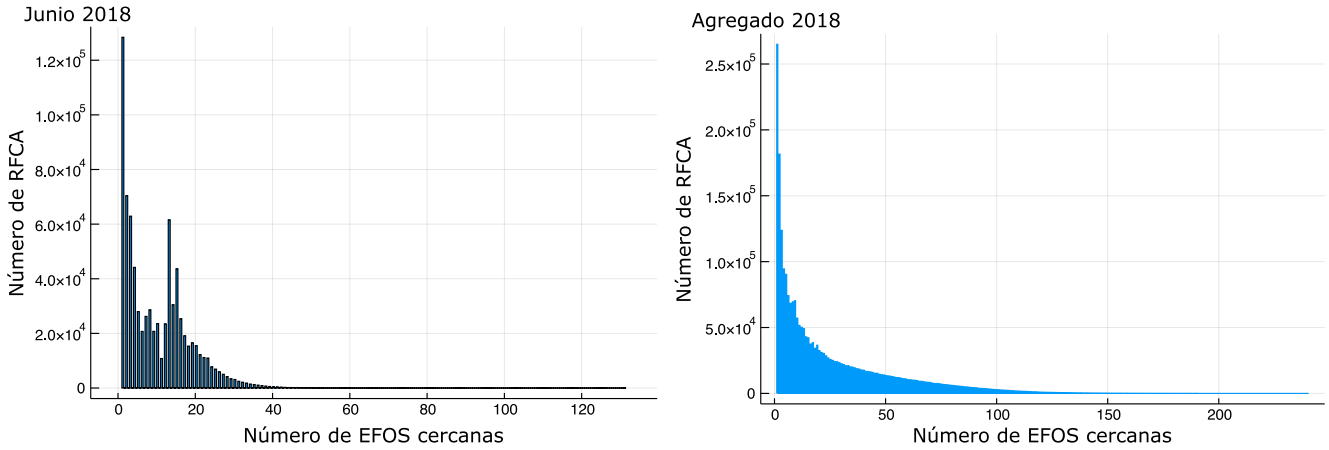


Figura 6: Distribución de EFOS cercanas a RFCA dentro de las redes de interacción mensual (izquierda) y el agregado anual (derecha). El número de EFOS cercanas es un indicador del nivel de colusión de un RFCA dentro de la red de operaciones de las EFOS. Se observan casos en los que RFCA se encuentran en la cercanía de un gran número de EFOS tanto en el caso mensual (izquierda) como en el agregado anual (derecha).

También estudiamos otras medidas, como el *betweeness*, *closeness*, *stress* y *page rank*, entre otros, sin embargo, no se observaron comportamientos atípicos de las EFOS que permitiera identificarlas.

La descripción que hemos realizado de las redes de interacción entre contribuyentes, tanto anuales como mensuales, nos ha permitido identificar características de los métodos de asociación de EFOS y la estructura local de la red a su alrededor, tales como: la organización en subredes de operaciones, asociadas a emisiones circulares de CFDI entre ellos y compuestas en su mayoría por RFCA desconocidos en las que las EFOS publicadas fungen como semillas (véase figura 3) y emisiones diferenciadas por parte de las EFOS, en las que hemos identificados que las emisiones entre las EFOS corresponden a montos mayores que las que realizan hacia RFCA desconocidos, lo cual nos sugiere que las EFOS operan entre ellas dentro de un nivel de actividad definido por los montos de sus operaciones (véase figura 4). También hemos podido cuantificar, por medio del alcance de los nodos en la red, el nivel de colusión de otros contribuyentes dentro de la actividad de las EFOS publicadas, como se puede observar en la figura 6, existen RFCA desconocidos que son cercanos a un gran número de EFOS, tanto en un mes como a lo largo de un año, los cuales consideramos más coludidos con las EFOS. Estos resultados sugieren que el análisis de redes aplicado a la descripción de los mecanismos de asociación y patrones de emisión de contribuyentes, es una herramienta útil y con un amplio potencial para la caracterización e identificación de prácticas nómalas.

5.2. Redes neuronales

Como primer método de clasificación de RFCA desconocidos como sospechosos de presentar un comportamiento similar al de las EFOS publicadas se implementó una *red neuronal artificial* (RNA). Las RNA son un modelo de aprendizaje automático inspirado en la arquitectura de un cerebro. Consisten de una colección de unidades interconectadas, de manera semejante a como se conectan las neuronas en un cerebro, por lo que comúnmente se le conocen a estas unidades como *neuronas*. Cada neurona alterna su estado entre activa e inactiva de acuerdo a la información que recibe de las neuronas con las que está conectada. Alterando el peso de la interacción entre neuronas, se cambia la manera en que se procesa la información. Es justamente mediante la modificación de dichos pesos como una red neuronal aprende a identificar patrones; a este proceso se le denomina *entrenamiento*.

Las neuronas en una red neuronal artificial a menudo se dividen en diferentes capas: una capa de entrada, la cual recibe los datos a clasificar; capas ocultas, que realizan el proceso de clasificación de los datos de entrada mediante la modificación de los pesos entre neuronas y el ajuste de las ponderaciones de los datos de entrada hasta que la clasificación que realiza la red sea óptima; y una capa de salida, de la cual se obtiene el resultado final del proceso de clasificación realizado por la red sobre los datos de entrada. La salida de la red se compara con la salida deseada mediante una función de pérdida, lo que da como resultado un cuantificador para el error. Durante el entrenamiento, estos errores se propagan a través de la red para actualizar los pesos y minimizar la función de pérdida.

Las RNA se han utilizado en una variedad de tareas, incluyendo visión por computadora [39], reconocimiento de voz [40], traducción automática [41], juegos de mesa y videojuegos [42–44] y diagnósticos médicos [45]. También se han utilizado en una variedad de aplicaciones en servicios financieros, desde pronósticos y estudios de mercado [46–48] hasta detección de fraudes [49] y evaluación de riesgos [50, 51]. Una red neuronal puede evaluar los datos de precios y descubrir oportunidades para tomar decisiones comerciales basadas en el análisis de datos. Las redes pueden distinguir interdependencias sutiles no lineales y patrones que otros métodos de análisis técnico no pueden.

5.2.1. Preparación de datos

En esta implementación, se diseñó una RNA que recibe como entrada los datos de todos los CFDI asociados a un RFCA emisor. Mediante una técnica llamada *re-sampling* (sobre-muestreo) [52], se forma una muestra balanceada de RFCA desconocidos y EFOS definitivos. El método de sobre-muestreo considerado en esta implementación consiste en volver a muestrear la clase pequeña (CFDI emitidos por EFOS definitivos) al azar hasta que contenga tantos ejemplos como la otra clase, para al final tener un gran conjunto de datos con la misma cantidad de CFDI emitidos por RFCA desconocidos y EFOS definitivos.

El modelo asocia a cada RFCA un valor entre 0 y 1 relacionado con la probabilidad de que éste sea EFOS. A continuación, describiremos el procedimiento que se realizó para diseñar, entrenar y evaluar la RNA. Posteriormente, presentaremos algunos resultados y conclusiones.

5.2.2. Modelado

Diseño de la RNA.– Una *red neuronal dinámica recurrente* (RNDR) es un tipo particular de redes neuronal que permite introducir un número arbitrario de renglones de datos (variables de entrada) a la vez, lo que resulta útil en este contexto, ya que los RFCA tienen cantidades varias de CFDI emitidos. Las redes neuronales recurrentes son arquitecturas en las cuales la salida de cada paso de la ejecución se provee como entrada al paso siguiente; esto

les permite conservar información aprendida a lo largo del tiempo. *Long short term memory* (LSTM) [20] describe el diseño de las neuronas artificiales, y son las que le otorgan memoria a la RNA. Estas neuronas tienen el mejor desempeño conocido en la actualidad y son particularmente efectivas para conjuntos de datos provenientes de series de tiempo [55–57] En particular, de varias arquitecturas que se probaron, se obtuvo el mejor desempeño con una RNDR con tres capas de celdas LSTM con 256 neuronas cada una, utilizando una función tangencial hiperbólica para calcular estados internos[21]. Cabe añadir que las conexiones de una RNDR no son únicamente entre diferentes capas, sino que también están conectadas de una neurona a sí misma a través del tiempo. Esto significa que la propagación del error para el ajuste de pesos se produce no sólo entre nodos diferentes, sino también entre el mismo nodo en diferentes pasos de tiempo, como se muestra en la figura 7.

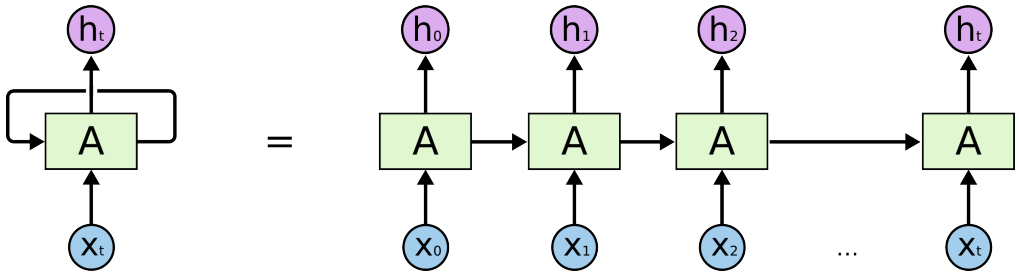


Figura 7: Una parte de una red neuronal A, observa una entrada x_t y calcula un valor h_t. El ciclo permite que la información fluya de un paso de la red al siguiente. Si desenrollamos el ciclo, una red neuronal recurrente puede considerarse como múltiples copias de la misma red, cada una de las cuales pasa un mensaje a un sucesor.

Una celda LSTM está controlada por tres compuertas: la compuerta de olvido, la compuerta de entrada y la compuerta de salida. Cada compuerta dentro de la celda es una red neuronal diferente que decide qué información se permite en el estado de la celda, el cual funciona como memoria de la red. Las compuertas pueden aprender qué información es relevante guardar u olvidar durante el entrenamiento. La compuerta de olvido controla la cantidad de información que se guardará en la memoria y elimina información que no es relevante. La compuerta de entrada controla la cantidad de nueva entrada que se almacenará en la memoria, en otras palabras, determina qué tan importante es la nueva información. Por último, la compuerta de salida determina las características de la información analizada para obtener una salida que permita clasificar correctamente.

La arquitectura de la red neuronal utilizada para clasificar RFCA como posibles EFOS cuenta con tres capas ocultas de celdas LSTM con 256 neuronas cada una, conectando cada neurona en una capa a cada neurona en la siguiente capa. La red se desenrolla a través del tiempo para analizar todas las facturas emitidas por un RFCA y, a partir de lo analizado, lo clasifica como EFOS o no EFOS.

Entrenamiento de la RNA.– El entrenamiento de la RNA se realiza a partir de los siguientes pasos. Se dividen todos los RFCA previamente identificados como EFOS definitivas en dos conjuntos, uno con 2,981 de los RFCA, llamado conjunto de entrenamiento, y otro con el 745, llamado conjunto de prueba. Al conjunto de prueba se agrega la misma cantidad de RFCA desconocidos. Al conjunto de entrenamiento se agregan 1,000,000 de RFCA desconocidos y luego

[20] Las celdas LSTMs son una topología de red neuronal presentada por primera vez por Hochreiter y Schmidhuber [53] con el propósito de eliminar el problema del desvanecimiento del gradiente [54] mediante la introducción de un mecanismo de memoria. Un gradiente mide cuánto cambia la salida de una función si cambian un poco las entradas. El problema es que, para redes muy profundas, el gradiente de los errores se disipa rápidamente en el tiempo, termina siendo muy pequeño y esto evita que los pesos cambien su valor. Las redes con este problema son capaces de aprender dependencias a corto plazo, pero a menudo tienen dificultades para aprender las dependencias a largo plazo.

[21] Estas tres capas corresponden a las capas ocultas que realizan el proceso de clasificación de los datos de entrada, además de las capas ocultas, la red cuenta con una capa de entrada y una de salida.

se copian los 2,981 EFOS definitivas hasta tener la misma cantidad que de RFCA desconocidos, terminando con un conjunto de 2,000,000 de RFCA. Así, ambos conjuntos estarán formados por 50 % datos de EFOS definitivas, correspondiente a los registros de los CFDI de tipo ingreso de EFOS definitivas seleccionadas al azar, y 50 % de RFCA desconocidos, que se compone de los registros de los CFDI de tipo ingreso de un conjunto de RFCA seleccionados al azar de la población total. Para cada RFCA se obtienen datos de los CFDI asociados. Estos datos son los que se proporcionan a la RNA y son sobre los que se entrena a la RNA ajustando parámetros internos. Después del proceso de entrenamiento se presenta a la RNA el conjunto de datos de prueba, el cual nunca antes ha visto, para evaluar su desempeño.

Variables adicionales consideradas.– Además de incorporar las variables cuantitativas mencionadas en la sección 4, se probó incorporar datos categóricos como el tipo y situación del contribuyente, la descripción de situación, el estado del contribuyente, la fecha de inicio de operaciones, el sector, y la entidad Federativa. También consideramos incorporar datos referentes a las redes de interacciones (véase la sección 5.1) como el grado de salida, grado de entrada, *betweenness*, *closeness*, *stress*, *radiality*, y *page rank*. Sin embargo, todas las RNA entrenadas con éstas variables tuvieron un desempeño igual o peor que la RNA que usa solamente datos de CFDI.

5.2.3. Evaluación de desempeño

Utilizamos el F1-score [58] como medida para evaluar la competencia del modelo entrenado. El F1-score se obtiene calculando la media armónica de la precisión y la recuperación. La precisión es la proporción de las instancias relevantes clasificadas correctamente entre todas las instancias que el modelo cree que son relevantes. Si VP son los verdaderos positivos y FP los falsos positivos, la precisión estaría dada por VP/(VP + FP), véase la tabla 1. La precisión contesta la pregunta *¿cuántos de los RFCA seleccionados realmente son EFOS?* La recuperación es la proporción de las instancias relevantes clasificadas correctamente entre todas las instancias realmente relevantes, VP/(VP+FN) con FN los falsos negativos, contesta la pregunta *de todos los RFCA que realmente son EFOS ¿cuántos fueron clasificados correctamente?* La media armónica se define como el valor obtenido cuando el número de valores en el conjunto de datos se divide por la suma de sus recíprocos. Es un tipo de promedio generalmente utilizado para números que representan una tasa o proporción (como la precisión y la recuperación) porque iguala los pesos de cada punto de datos. Un F1-score alcanza su mejor valor en 1 (precisión y recuperación perfecta) y el peor en 0. En la tabla 1 se muestra una forma de separar las clasificaciones que hace la red neuronal para poder evaluarla.

		Clase predicha	
		P	N
Clase real	P	Verdaderos Positivos (VP)	Falsos Negativos (FN)
	N	Falsos Positivos (FP)	Verdaderos Negativos (VN)

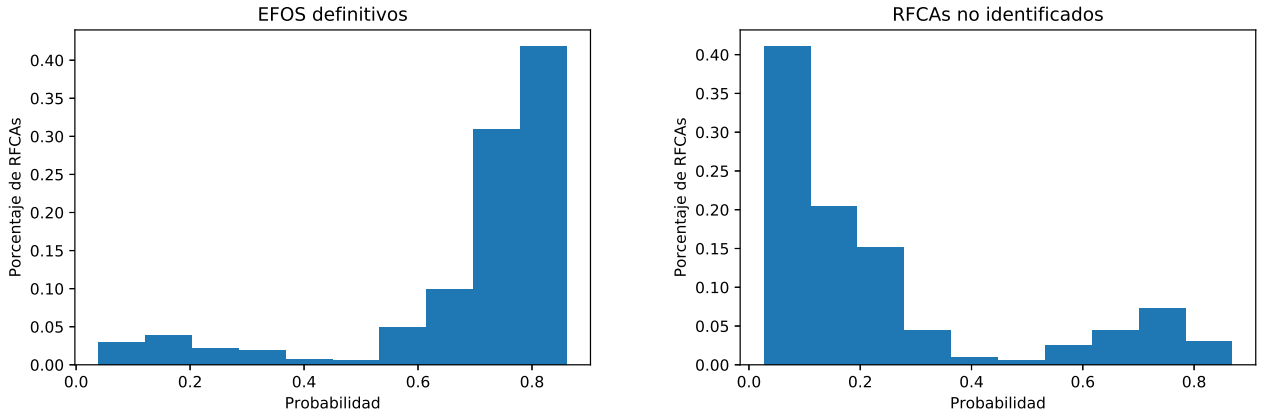
Tabla 1: Matriz de confusión para clasificación binaria. Los verdaderos positivos (VP) son los ejemplos que el modelo clasificó correctamente como EFOS. Los falsos negativos (FN) son los ejemplos que el modelo clasificó como No EFOS, pero que son en realidad EFOS. Los verdaderos negativos (VN) son los ejemplos que el modelo clasificó como No EFOS y no se han detectado como EFOS anteriormente. Los falsos positivos (FP) son los ejemplos que el modelo clasificó como EFOS, que no se han detectado como tales anteriormente.

Por ejemplo, si tomamos 500 EFOS definitivas y 500 desconocidas, y las alimentamos a nuestra red, encontramos que VP = 448, FN = 52, VN = 416 y FP = 84, por lo que la precisión fue de 0.845, la recuperación 0.896 y se obtuvo un 0.87 de F1-score. Si realizamos el cálculo con 1000 EFOS presuntas, obtenemos VP = 881, FN = 119 (VN = FP = 0

por definición), por lo que la precisión es de 1, mientras que la recuperación es de 0.881. Con esto, se obtiene un F1-score de 0.94.

Los RFCA en el conjunto de "presuntos" muestran el mismo comportamiento que el modelo identificó al entrenar con el conjunto de "definitivos", y termina identificando como EFOS al 88 %.

Calculamos la distribución de la probabilidad que obtiene el modelo para EFOS definitivas y RFCA que hasta ahora no han sido identificados como EFOS. En este último grupo existen RFCA que realmente no son EFOS y RFCA que son EFOS pero no han sido detectados.



(a) (b)

Figura 8: Histogramas de las probabilidades asignadas por la red neuronal a diferentes conjuntos de RFCA. Al la izquierda, se consideran EFOS definitivas. Observamos que la red correctamente asigna a la mayoría de ellos una probabilidad alta de ser EFOS. A la derecha, consideramos RFCA no identificadas. Observamos una distribución bimodal en la que hay un porcentaje considerable RFCA a los que se les asigna una probabilidad alta de ser EFOS.

En la figura 8 podemos observar que el modelo está seguro de su decisión la mayoría de las veces (termina con muy alta o muy baja probabilidad). Además en la distribución de probabilidad de los RFCA no identificados, existe un porcentaje que el modelo está clasificando con alta probabilidad (el modelo esta seguro que es EFOS) pero no ha sido identificado anteriormente como EFOS.

Uno de los mayores desafíos en redes neuronales es interpretar lo que la red está aprendiendo de los datos. No sólo es importante desarrollar una solución sólida con un gran poder de predicción; también es interesante entender cómo funciona el modelo desarrollado: qué variables son las más relevantes, la presencia de correlaciones, las posibles relaciones de causalidad, etcétera. Para profundizar en el entendimiento de los resultados, realizamos dos técnicas para conocer las variables más relevantes que detallamos a continuación.

La primer técnica se basa en el análisis hipotético o de simulación, y se usa para medir la importancia relativa de las variables de entrada en los resultados de un modelo. En particular, para medir la importancia de las variables, tomamos una muestra de nuestros datos X y calculamos las predicciones del modelo ya entrenado Y. Luego, para cada variable x_i perturbaremos esa variable (y solo esa variable) mediante una distribución normal aleatoria centrada en 0 con escala del 20 % del promedio de la variable y calcularemos una predicción Y_i. Mediremos el efecto que tiene esta perturbación calculando la diferencia de raíz cuadrática media entre la salida original Y y la perturbada Y_i. Una diferencia de raíz cuadrática media mayor significa que la variable es "más importante". conoce como análisis hipotético o de simulación. es un modelo que determina cómo se ven afectadas función de cambios en las variables

de entrada. Este como análisis hipotético o de simulación. En la tabla 2 (izquierda) se reportan las cinco variables con mayor importancia para la red neuronal.

La segunda técnica consiste en el análisis de componentes principales, la cual es una técnica estadística para convertir datos de alta dimensión en datos de baja dimensión seleccionando las características más importantes que capturan la mayoría de la información sobre el conjunto de datos. Las características se seleccionan en función de la variación que causan en la salida. Podemos extraer las características más importantes del conjunto de datos que son responsables de la máxima variación en la salida. La característica que causa la mayor varianza es el primer componente principal. La característica responsable de la segunda varianza más alta se considera el segundo componente principal, y así sucesivamente. Es importante mencionar que los componentes principales no tienen ninguna correlación entre sí. La importancia de cada variable se refleja en la magnitud de los valores correspondientes en los vectores característicos de una transformación lineal (mayor magnitud - mayor importancia). En la tabla 2 (derecha) se reportan las cinco variables que mejor caracterizan el conjunto de datos de acuerdo al primer componente principal, el cual representa el 99 % del total de la varianza. Las magnitudes de las variables están normalizadas para que la suma de los cuadrados sea igual a 1.

Variable	Efecto de perturbación
Monto Sub Activo	0.2099
Monto Total Activo	0.1813
Monto Total Tras Activo	0.1419
Monto Iva Tras Activo	0.1083
Monto Total Cancelado	0.0748

Variable	Magnitud
Monto Total Activo	0.74925125
Monto Sub Activo	0.64598303
Monto Total Tras Activo	0.10326791
Monto Iva Tras Activo	0.1032678
Monto Total Cancelado	0.0000125

Tabla 2: (izquierda) Efecto de la perturbación en la probabilidad asignada por la red neuronal. (derecha) Importancia de las variables de acuerdo al valor absoluto de su magnitud en el primer componente principal usado para caracterizar el conjunto de datos.

5.2.4. Resultados del modelo

La RNA clasifica de forma eficiente las EFOS identificados que se le han presentado, y utilizando el modelo entrenado clasificamos los RFCA desconocidos somo sospechosos a los que la RNA les asigna una mayor probabilidad de presentar un comportamiento similar al presentado por las EFOS publicadas. La RNA clasificó a 149,921 RFCA desconocidos, correspondientes al 1.98 % del total, como sospechosos con alta probabilidad (> 0.8).

5.3. Bosque aleatorio

Como segundo método de clasificación se usó la técnica de aprendizaje automático denominada *random forest*, o bosque aleatorio (BA). Las técnicas de clasificación automática (incluyendo BA) detectan grupos de elementos con patrones estadísticos similares en una base de datos disponible y, a partir del conocimiento adquirido, toman decisiones sobre la pertenencia a estos grupos de elementos nuevos. En nuestro caso, consideramos las características de EFOS publicadas por el SAT, y las comparamos con RFCA desconocidas.

Un BA se construye combinando aleatoriamente distintos *árboles de decisión*, a fin de obtener resultados robustos a fuentes de ruido inherentes al algoritmo. Un árbol de decisión es un algoritmo matemático formado por un conjunto de preguntas ordenadas y conectadas entre sí a través de sus respuestas (es decir, la formulación de una pregunta

depende de la respuesta a la pregunta anterior). Estas preguntas involucran las variables o características de los datos utilizados. Al construir un árbol de decisión, cada nodo representa una de las preguntas y cada bifurcación depende de su respuesta. Así, al terminar de construir un árbol de decisión, podemos seguir un camino determinado de preguntas y respuestas y contestar la pregunta principal: ¿qué probabilidad tiene este RFCA de ser parte de las EFOS?

En modelos estadísticos como BA es necesario mantener un equilibrio entre medidas como la *varianza* (la variabilidad en la predicción de los modelos para distintos elementos) y el *sesgo* (el grado de diferencia entre el valor real y el predicho). Para lograr dicho equilibrio, una técnica eficaz es la combinación de varios modelos (como la combinación de árboles de decisión para formar un BA) . Así, cada árbol de decisión emite una clasificación (i.e. una probabilidad de sospecha de ser EFOS asociada a los RFCA) y el resultado final del BA es la clasificación más probable entre todos los árboles construidos. Una de las tareas a resolver al momento de construir un BA es encontrar el número óptimo de árboles de decisión utilizados para determinar la combinación que genere el resultado final.

En nuestro caso, la técnica de BA se considera adecuada dado que ofrece las siguientes ventajas:

- La preparación de los datos es mínima. Únicamente se necesita contar con un set de datos donde cada elemento a clasificar, en este caso cada RFCA, sea único y tenga un número determinado de características asociadas a cada una de las clases involucradas, en este caso definitivo o desconocido.
- Tiene un buen manejo para números grandes de variables sin discriminar ninguna.
- Está demostrado que es uno de los métodos con precisión más alta entre los algoritmos de clasificación [59].
- Tiene buen desempeño en bases de datos de gran volumen (lo cual aplica al presente caso de estudio).

El resultado del BA es un número entre 0 y 1 para cada RFCA evaluado, el cuál será interpretado como la probabilidad de cada RFCA desconocido de ser una posible EFOS.

5.3.1. Preparación de datos

Para la implementación del algoritmo de BA, inicialmente se realiza la agrupación de información por emisor, dado que el presente análisis se enfoca en clasificar a los RFCA emisores. Como resultado se obtiene un registro único por cada RFCA emisor para cada uno de los 48 meses contemplados.

Posteriormente, mediante una técnica llamada *undersampling* (submuestreo) [60], se forma una muestra balanceada de RFCA desconocidos y EFOS definitivos. Esta técnica busca el número óptimo de RFCA que permita tener una muestra de los datos que además de balanceada (que tenga la misma cantidad de desconocidos y definitivos) sea representativa (que con el número de RFCA elegido se logre captar las características de toda la población). Como resultado del proceso anterior se llega a una muestra con 1561 EFOS definitivos y 1561 RFCA desconocidos. La muestra obtenida hasta este momento es el conjunto de datos base utilizado para la implementación del algoritmo de BA.

Como parte de la fase de preparación de datos, se aplican dos tratamientos independientes a la muestra anteriormente generada:

1. Se aplicó un análisis para determinar qué tipo de transformación de datos es viable para cada una de las

variables en la muestra. Se usó la familia de transformaciones *box cox* para mejorar la normalidad e igualar la varianza de los datos con el objetivo de mejorar el desempeño del algoritmo [61].

2. Se utilizó el método de *componentes principales*. Este consiste en reducir la dimensionalidad unificando variables existentes para crear nuevas. Este procedimiento se recomienda para mejorar el desempeño de los algoritmos en cuestión [62].

5.3.2. Construcción del modelo

Utilizando el algoritmo de BA se construyeron tres modelos que corresponden a los siguientes escenarios y que utilizan la muestra generada en la sección anterior:

1. Primer escenario: Implementación del algoritmo de BA sin ninguna transformación.
2. Segundo escenario: Implementación del algoritmo de BA utilizando la muestra de datos a la cual se aplicó la técnica de componentes principales.
3. Tercer escenario: Implementación del algoritmo de BA utilizando la muestra de datos en la cual se aplicaron las transformaciones *box cox*.

Para cada uno de los escenarios anteriores, al entrenar el algoritmo de BA se busca el número óptimo de árboles de decisión que lo conformarán. Esto se logra realizando iteraciones del algoritmo, modificando el número de árboles utilizado y observando en qué momento el error producido se estabiliza en un mínimo. Se llegó a la conclusión de que el número óptimo de árboles de decisión es 100.

5.3.3. Evaluación de desempeño

Para evaluar los escenarios anteriores se utilizaron las siguientes medidas.

- Curva ROC: Es una medida de desempeño con valores entre 0 y 1; mientras más grande el valor, dicho desempeño se considera mejor. Una curva ROC se construye utilizando la información de dos puntos: la sensibilidad (posibilidad de clasificar bien a un individuo positivo, en este caso a un EFOS definitivo) y la especificidad (posibilidad de clasificar bien a un individuo negativo, en este caso a un RFCA desconocido que en la realidad no es un RFCA definitivo) [63].
- Error: Es una medida de penalización. Mientras más cercano a 0, se considera mejor. El error cuantifica la parte del modelo que se esta equivocando al clasificar a los RFCA, y en el caso del BA se obtiene mediante una combinación del error producido por cada uno de los árboles individuales, así como la correlación que existen entre estos [59].

Como se puede observar en la tabla 3, a pesar de que hay una mejora en el desempeño para el primer escenario, se privilegia la disminución del error, por lo que el modelo elegido fue el que incluye la transformación de variables *box cox*. Este es el modelo que se usó en los siguientes pasos.

Considerando el modelo elegido, se realizó una validación más, la cual consiste en clasificar los EFOS definitivos utilizando el modelo (los cuales ya sabemos *a priori* que tendrían que tener una probabilidad alta) y observar qué

Escenario	ROC	Error
Bosque aleatorio	0.912	0.164
Bosque aleatorio más componentes principales	0.886	0.161
Bosque aleatorio más transformación de variables	0.893	0.157

Tabla 3: Comparación de medidas de desempeño para las diferentes maneras en que se transformaron los datos de entrada.

	Años con actividad			
Años clasificados como EFOS	1 año	2 años	3 años	4 años
0	17 % (133)	5 % (56)	3 % (11)	6 % (8)
1	83 % (631)	13 % (143)	6 % (24)	4 % (6)
2		82 % (893)	17 % (71)	11 % (16)
3			74 %(307)	26 % (37)
4				53 % (77)

Tabla 4: Estudiamos el desempeño del algoritmo de BA año a año. Consideramos los EFOS definitivos, separados por el número de años que tienen actividad (columnas). En las diferentes filas, consideramos el número de años en los que el algoritmo clasifica el RFCA como EFOS; así, un EFOS definitivo debería ser detectado por el algoritmo en al menos uno de los años de actividad. Por ejemplo, de las EFOS con actividad reportada durante 3 años, BA clasifico erróneamente el 3 % del total de EFOS definitivos con actividad reportada por 3 años, correspondiente a 11 EFOS definitivos.

resultados se obtienen. Se estableció un punto de corte de 0.8; es decir, si el índice de riesgo obtenido es 0.8 o mayor se considera al RFCA clasificado como EFOS, de lo contrario, no). Además, se consideraron los años de actividad de cada EFOS definitivo para el diagnóstico final. Es decir, si tuvo actividad, por ejemplo, dos años, se consideran las dos calificaciones y así sucesivamente. Desarrollando lo anterior se obtuvieron los resultados de la tabla 4, donde se puede observar que cerca del 92 % de los EFOS definitivos están siendo clasificados correctamente por el algoritmo, y el error es únicamente del 8 %.

Calificación	Frecuencia	Porcentaje
EFOS	1,908	79 %
No EFOS	505	21 %

Tabla 5: Calificación de los diferentes tipos de contribuyentes.

Unificando los resultados anteriores se consideraron clasificados como posibles EFOS aquellos RFCA que en todos los años de actividad fueron detectados por el modelo y como No EFOS el caso contrario. La tabla 5 muestra que de todos los EFOS definitivos, únicamente 505 fueron clasificados como No EFOS, lo que significa que son los únicos donde el algoritmo se está equivocando completamente. Dicho comportamiento se considera normal debido a la posibilidad de que sólo en algunos años las EFOS pudieron haber tenido actividades ilícitas.

5.3.4. Resultados

Utilizando el modelo construido y validado en las secciones anteriores (tercer escenario), se toman cuatro grupos de RFCA desconocidos (uno por cada año de estudio) y se obtienen el índice de riesgo. Nótese que si el RFCA tiene

más de un año de actividad, éste tendrá un índice diferente por cada año.

Tomando en cuenta los resultados anteriores, se definieron los siguientes grupos para el total de los RFCA desconocidos:

- Sospechosos: Son todos aquellos RFCA desconocidos que en cada uno de los años que tienen actividad poseen un índice de riesgo mayor o igual a 0.8.
- No sospechosos: Son todos aquellos RFCA desconocidos que en al menos uno de los años que tienen actividad posee un índice de riesgo menor a 0.8.

Con estas definiciones, el algoritmo clasificó a 7,438,448 RFCA como no sospechosos (98.3 %) y a 128,227 RFCA como sospechosos (1.7 %) de ser EFOS.

5.4. Integración de los distintos métodos de clasificación

Resultado de la aplicación de cada uno de estos métodos sobre la base de datos de emisiones de CFDI que nos fue proporcionada, obtuvimos una lista de RFCA sospechosos por cada uno de los métodos, de los cuales, considerando su nivel de colusión dentro de las redes de operación de EFOS por medio de la definición de un Índice de Cercanía, se realizó una estimación de la evasión de IVA. Cabe notar que los RFCA identificados como sospechosos por los distintos métodos parten de supuestos y definiciones realizadas al interior del SAT. Por tal motivo, los datos inducen un sesgo sobre los mecanismos de clasificación de las técnicas. Este sesgo es inevitable dado que es el punto de partida del entrenamiento de los métodos y es necesario considerar a futuro otros métodos que permitan realizar una caracterización más completa e imparcial de otros mecanismos de evasión además de los que ya se han identificado.

Una vez que fueron entrenados y evaluados ambos métodos de clasificación y se procedió a presentarles los datos de los RFCA desconocidos, obtuvimos dos poblaciones de sospechosos por cada uno de los dos métodos, el tamaño de las cuales se reporta en la tabla 6. Los RFCA sospechosos obtenidos de la RNA corresponden a los que les fue asignada una probabilidad > 0.8 de pertenecer a la clase de EFOS y de forma análoga, la lista obtenida del BA corresponde a los RFCA para los cuales se obtuvo un índice de probabilidad > 0.8. Con el objetivo de realizar un refinamiento de estas dos poblaciones de sospechosos consideramos la intersección de las dos listas, obteniendo un total de 43,650 RFCA, los cuales consideramos pueden tener una probabilidad más alta de ser posibles EFOS, debido a que fueron identificados por ambos métodos de forma independiente.

Método	Sospechosos	No Sospechosos
Red Neuronal	149,921	7,416,754
Bosque Aleatorio	128,227	7,438,448

Tabla 6: Número de RFCA clasificados como sospechosos por cada método de clasificación empleado.

5.4.1. Comparación de comportamiento temporal

Con el fin de comparar el comportamiento de los distintos valores de los CFDI entre las poblaciones de EFOS definitivas, presuntas y las definidas por los RFCA identificados como sospechosos de cada uno de los métodos considerados, calculamos las distribuciones de los valores asociados a distintos campos de los CFDI emitidos, en

particular del número de Facturas Activas, el número de Facturas Canceladas, el Monto Total Cancelado y el Monto Subtotal Activo. Representamos dichas distribuciones por medio de *diagramas de caja* (boxplots), en las cuales, como se muestra en la figura 9, se representa la mediana, la región intercuartil y los valores atípicos (outliers) de las distribuciones de cada uno de los valores para las distintas poblaciones de RFCA.

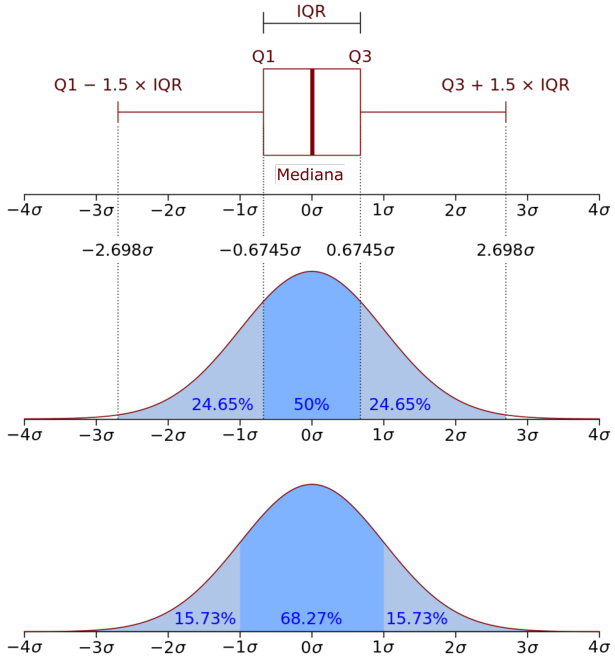


Figura 9: Representación en diagrama de caja (boxplot) de una distribución de valores. La caja central representa el rango intercuartil mientras que la línea representa la mediana y σ es la desviación estándar de la distribución. Todos los puntos que quedan fuera de los rangos $Q1 - 1.5 \times IQR$ y $Q3 + 1.5 \times IQR$ se consideran atípicos (outliers).

Como se muestra en la figura 10, las variables que resultan ser más características de las EFOS y los RFCA sospechosos son: el Monto Total Cancelado y el Monto Subtotal Activo. Las variables restantes asociadas a los CFDI no mostraron una diferencia significativa entre poblaciones. La figura 11 muestra el comportamiento temporal de estas dos variables. Es posible observar que la diferencia en el comportamiento del Monto Total Cancelado y Subtotal Activo de los CFDI asociados a EFOS y sospechosos con respecto a los RFCA que no fueron catalogados como sospechosos se mantiene durante los 48 meses que fueron analizados, separando a las EFOS y sospechosos del comportamiento de la población general.

5.4.2. Número de EFOS cercanas a RFCA sospechosos

Como se introdujo en la sección 5.1, el alcance de las EFOS en las redes de interacción mensuales permite identificar el número de EFOS cercanas a los RFCA desconocidos dentro de la red (a una distancia $d \leq 3$) y así identificar a los que se encuentran más inmersos o coludidos dentro de las operaciones de EFOS. Si consideramos a los RFCA sospechosos pertenecientes a la intersección de las dos listas obtenidas por los métodos de clasificación (RNA y BA, 43 mil RFCA) y calculamos el agregado anual de EFOS cercanas a cada una de ellas, observamos que son cercanas a un número alto de EFOS a lo largo del año (figura 12), lo cual nos indica que los RFCA clasificados como sospechosos por ambos métodos corresponden a distintos RFCA con un nivel alto de colusión con las EFOS publicadas por el SAT, lo cual nos da confianza sobre los métodos de clasificación que fueron implementados.

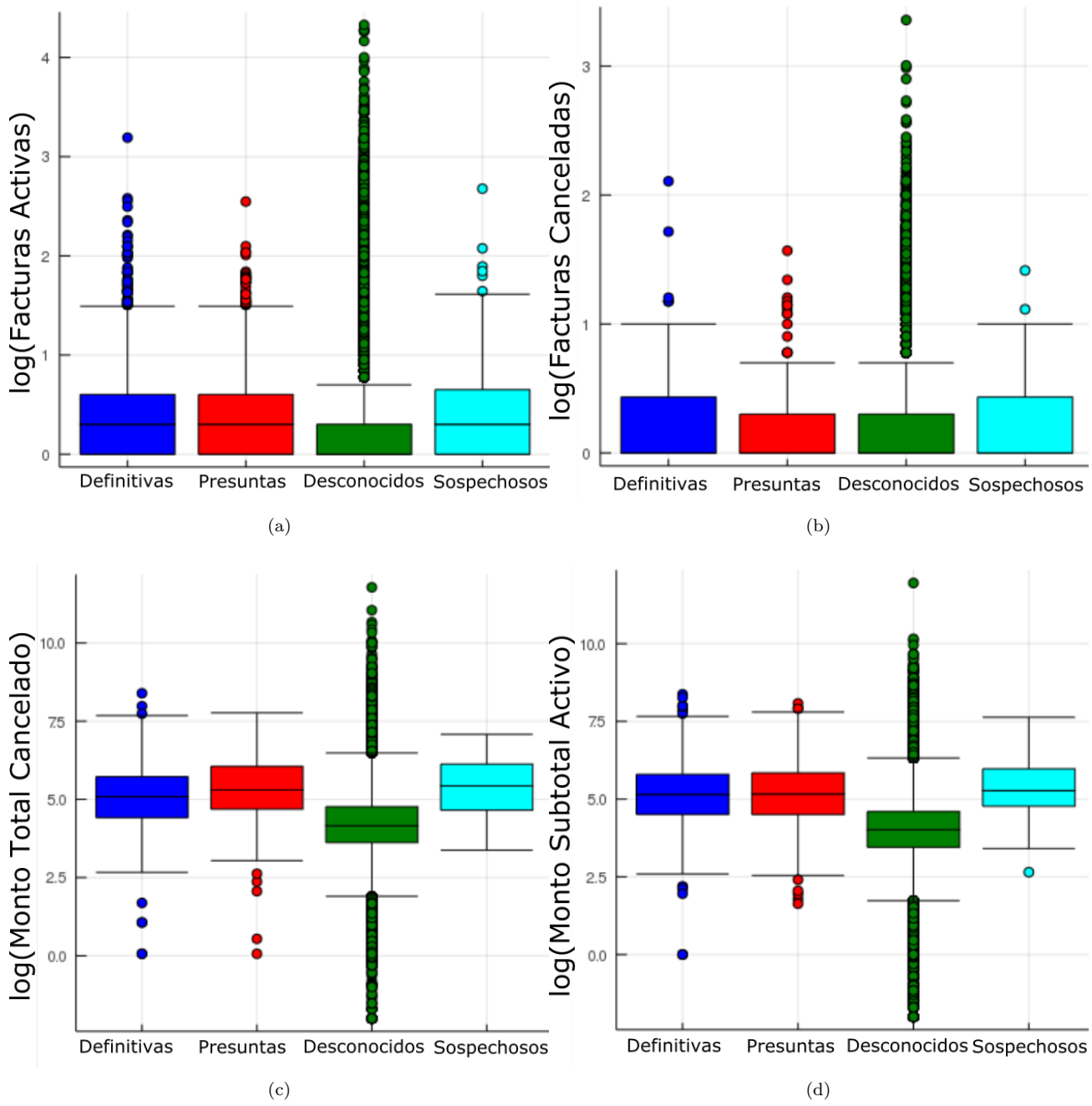
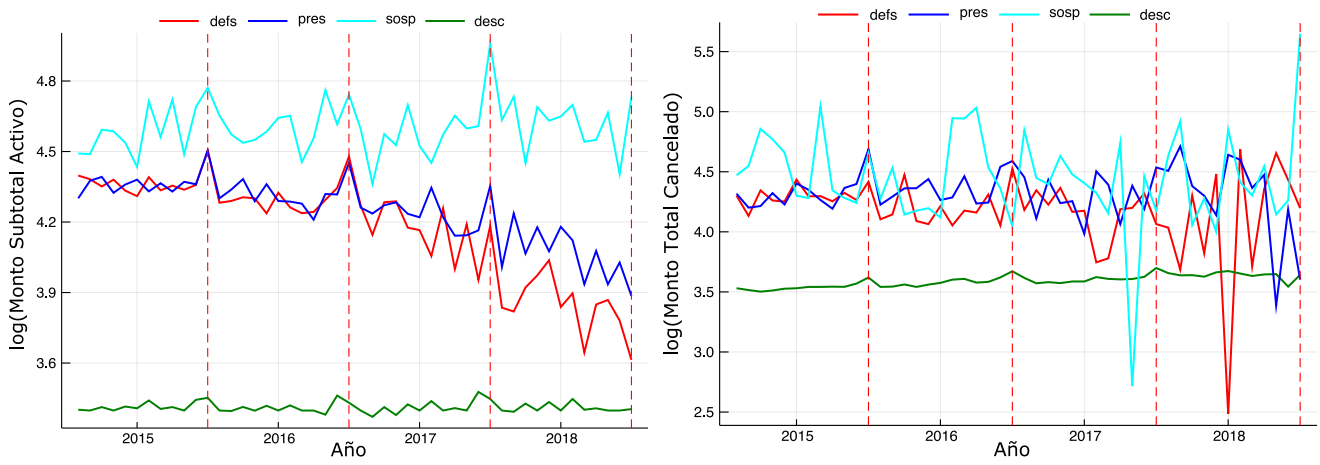


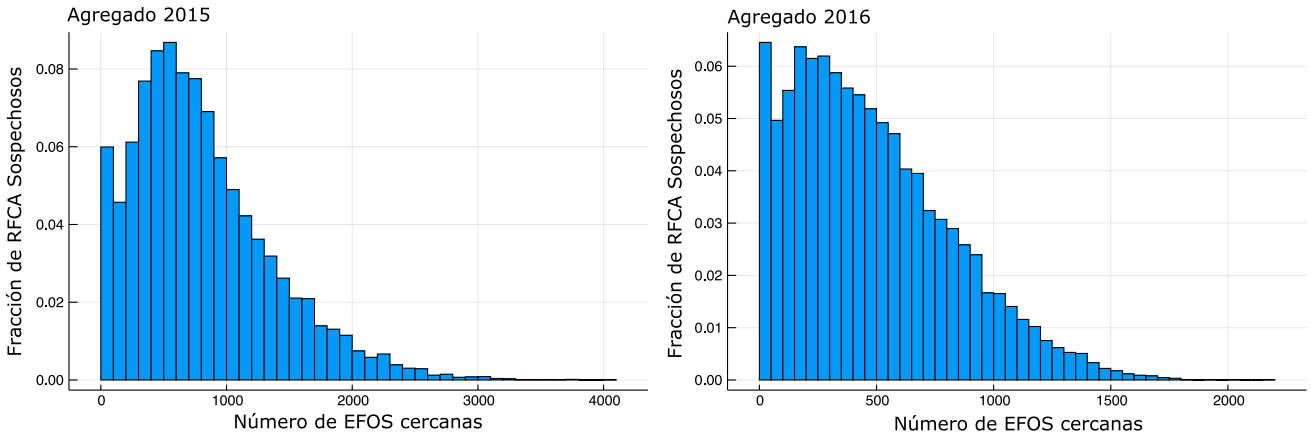
Figura 10: Diagramas de caja para distintos valores asociados a los CFDI para cada una de las poblaciones de EFOS definitivas, presuntoa, RFCA desconocidos y los presentes en la intersección de los RFCA clasificados como sospechosos por los métodos de Redes Neuronales y Bosque Aleatorio. Se presentan en escala logarítmica: s (a) Facturas Activas. (b) Facturas Canceladas. (c) Montos Rotales Cancelados. (d). Monto Subtotal Activo. Se puede observar que para el caso de los Montos Totales Cancelados y el Monto Subtotal Activo, las distribuciones entre EFOS y sospechosos son muy parecidas y corresponden a montos mayores que la distribución de los RFCA desconocidos.

Cabe hacer notar que la cercanía a EFOS no fue parte de las variables utilizadas por los métodos de clasificación para la identificación de RFCA sospechosos, ya que estos solamente se basaron en los datos de CFDI, sino que se calcula después de la clasificación y se compara con los resultados obtenidos de la caracterización basada en las redes de interacción entre contribuyentes.



(a) (b)

Figura 11: Comportamiento temporal de (a) Monto Subtotal Activo y (b) Monto Total Cancelado para cada una de las poblaciones consideradas: EFOS definitivos (azul), EFOS presuntos (rojo), RFCA desconocidos (verde), RFCA sospechosos (cian). Las líneas corresponden a las medias de la distribución de cada población y las líneas punteadas corresponden al mes de Diciembre de cada año. Se puede observar cómo la tendencia del comportamiento de EFOS y sospechosos se separa del de la población de RFCA descononcidos.



(a) (b)

Figura 12: Total anual de EFOS cercanas a los RFCA identificados como sospechosos por los distintos métodos de clasificación (a) 2015 y (b) 2016. Se puede observar que un alto porcentaje de RFCA sospechosos son cercanos a varias EFOS, lo cual nos indica que se encuentran inmersos en sus grupos de operaciones.

5.4.3. Descripción de otros aspectos categóricos

El conjunto de datos que nos fue proporcionado para la realización del estudio incluye variables categóricas de los RFCA, las cuales incluyen información como: el tipo de persona, fecha de inicio de operaciones, entidad federativa y actividad económica, entre otras. En esta sección realizamos una descripción de las variables categóricas asociadas a los RFCA que nuestros métodos han clasificado como sospechosos con el objetivo de complementar con información categórica la caracterización que se ha realizado en las secciones anteriores del comportamiento y mecanismos de asociación de RFCA sospechosos de realizar emisiones anómalas.

Tipo de persona	Porcentaje de la población
Moral	81.52 %
Física	10.22 %
Sin Información	8.3 %

Situación Fiscal	Porcentaje de la población
Activo	91.15 %
Cancelado	0.13 %
Suspendido	0.46 %
Sin información	8.3 %

Tabla 7: Tipo de persona de los RFCA sospechosos identificados por los métodos de clasificación y su situación fiscal. La mayor parte de los RFCA sospechosos son personas morales y el 91.5 % del total se encuentran activos, lo cual los hace susceptibles de ser investigados.

Como se muestra en la tabla 7, el 81.52 % de los RFCA sospechosos corresponde a personas morales, lo cual nos indica que una gran parte del intercambio de CFDI emitidos asociados a operaciones potencialmente simuladas se hace entre empresas. Esto nos lleva a suponer que se elige este tipo de figura jurídica dado que permite, dependiendo de la naturaleza de su constitución, que la responsabilidad legal de los actos o hechos potencialmente ilícitos efectuados, caigan en la persona moral y no en una persona física. En la misma tabla 7 se reporta que el 91.15 % de los RFCA sospechosos se encuentran activos y solamente una fracción menor al 1 % se reportan como cancelados o suspendidos, lo cual muestra que la mayor parte de los RFCA sospechosos participan de la actividad económica cotidiana y son susceptibles de ser investigados.

Los RFCA sospechosos se distribuyen en todo el territorio nacional (véase la figura 13), sin embargo estos se concentran en la Ciudad de México, Nuevo León, Estado de México y Jalisco principalmente. También se observa que la mayor parte de los RFCA sospechosos se constituyeron e iniciaron operaciones en los últimos 10 años (véase la figura 14). Cabe mencionar que hay casos en los que los años de constitución e inicio de operaciones reportados corresponden a hace más de 40 o 50 años, lo cual puede estar asociado a errores o abusos.

Asi mismo, al analizar las DIOT de las EFOS definitivas publicadas por el SAT, observamos que la mayoría de ellas no presentan declaraciones de forma regular en el año, siendo los primeros y los últimos meses cuando se observa el mayor número de ellas. Por otro lado, también se observa que usualmente presentan varias declaraciones para el mismo periodo y en algunos casos se declara el mismo monto en periodos distintos. Por este motivo, es que decidimos no basarnos en los datos incluidos en las DIOT para la realización del cálculo de evasión.

La constancia y el número de DIOT presentadas por este tipo de contribuyentes no es coherente con lo establecido en el artículo 32 fracción VIII de la Ley del Impuesto al Valor Agregado, en donde se señala que se tendrá hasta el último día del mes posterior al que se está declarando para presentar todas las operaciones relacionadas con el acreditamiento y retención del IVA.

De forma similar a los registros de DIOT que nos fueron proporcionados, en el caso de las declaraciones de saldos a favor y acreditados del IVA, solo el 6 % de las EFOS definitivas del padrón proporcionado presentaron declaraciones de saldos a favor en el periodo 2015-2018. Cabe mencionar que las declaraciones presentadas en este periodo hacen referencia a ejercicios fiscales entre 2002 y 2018, lo cual indica una irregularidad en el comportamiento de las declaraciones presentadas por EFOS, tanto en su frecuencia como la diferencia de tiempo con respecto a la presentación de la declaración y el ejercicio y periodo a las que hacen referencia. Dado que los datos de CFDI con los que contamos corresponden al periodo del 2015 al 2018, solamente se consideran las declaraciones correspondientes al mismo periodo.

Otra característica particular de las declaraciones de saldos a favor del IVA por parte de EFOS, es el hecho de que en distintas declaraciones correspondientes al mismo periodo y ejercicio se reportan montos a favor distintos, lo cual

hace que sea difícil determinar cuál es el monto real. Cabe recordar que, al ser estas declaraciones realizadas por medio de un formulario por los mismos contribuyentes, son susceptibles de ser manipuladas y no son una fuente confiable de información en la que se pueda basar el cálculo de los estimados de evasión.

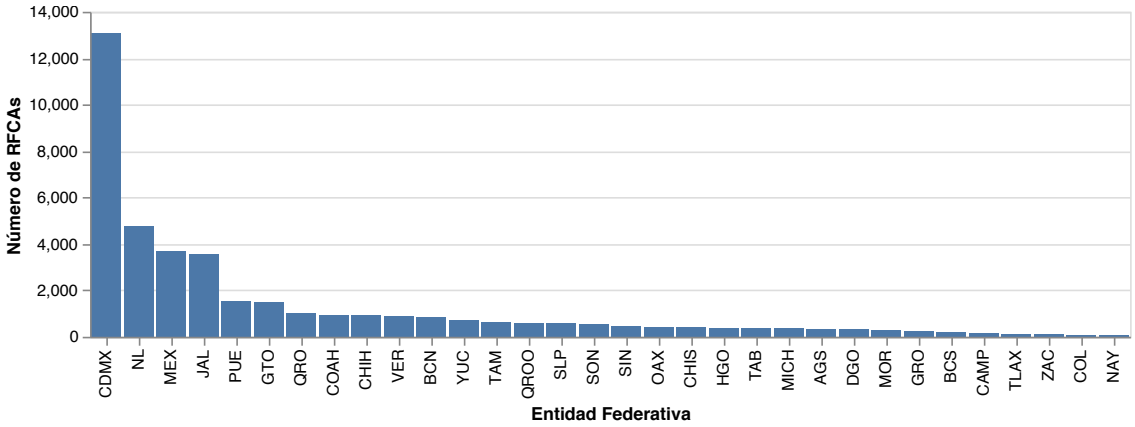


Figura 13: RFCA sospechosos por entidad federativa. Los 43,650 RFCA sospechosos mencionados al inicio de esta sección, se distribuyen en todo el territorio nacional, acumulándose en la Ciudad de México, Nuevo León, Estado de México y Jalisco.

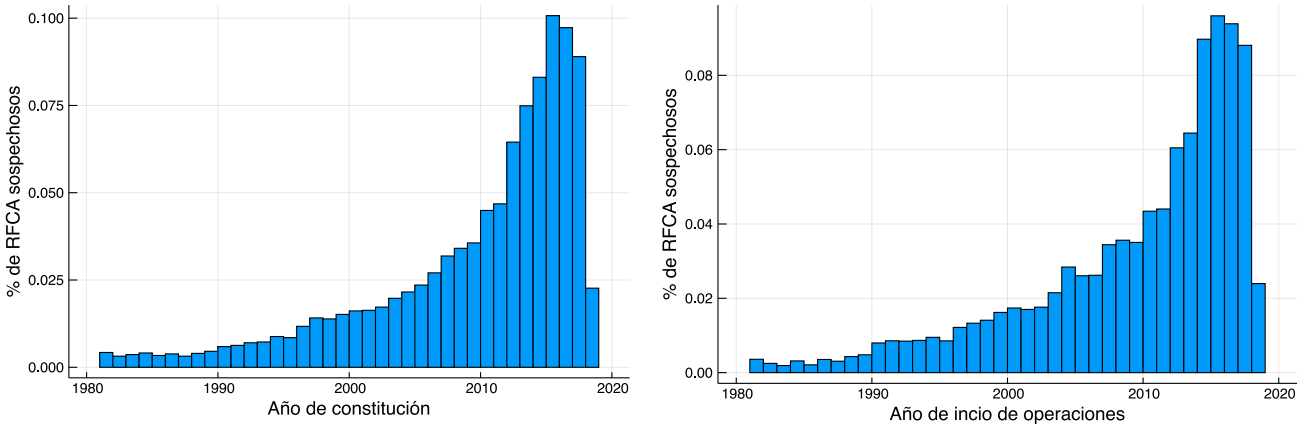


Figura 14: Año de constitución y de inicio de operaciones de los RFCA sospechosos. Se puede observar que la mayoría son de reciente creación.

5.5. Cálculo de la evasión del IVA

Dentro de los ingresos tributarios en México, de acuerdo a los datos presentados por la Secretaría de Hacienda y Crédito Público, destaca por su importancia el IVA, que es el segundo más importante al aportar el 29.44 % de la recaudación total tributaria entre el año 2015 y el año 2018 [64]. Comparado con el promedio recaudado por el mismo impuesto en América Latina para el año 2017, el cual fue del 27.9 % [65], uno podría considerar que las acciones llevadas a cabo para impulsar la recaudación están siendo efectivas. Sin embargo, existen varias causas por las cuales este impuesto aún no alcanza su máximo nivel de recaudación. Algunas de estas causas derivan de las facilidades administrativas, estímulos fiscales o tasas especiales, la exención del impuesto a ciertas actividades o por la evasión fiscal determinada por un bajo nivel de cumplimiento y por la emisión de comprobantes fiscales derivados

de operaciones simuladas.

El impacto de esto último, así como sus emisores, se estima año con año. El SAT publica en su página web y en el DOF, el listado de aquellos contribuyentes que de acuerdo al artículo 69-B del CFF, están dentro del supuesto de la emisión de facturación de operaciones simuladas[22] [66]. De acuerdo con datos proporcionados por el SAT, las pérdidas generadas por este tipo de contribuyentes alcanzan los 354 mil millones de pesos, equivalentes al 1.4 % del Producto Interno Bruto (PIB)[23]. Dicho lo anterior, tener datos relevantes y confiables que puedan determinar la evasión fiscal del IVA derivada de la emisión de CFDI por operaciones simuladas, es vital para que, se pueda conocer los daños económicos que generan y, por otra parte, se puedan tener métodos prácticos que ayuden al combate de este tipo de operaciones.

Cabe resaltar, que el cálculo de la estimación que representa la evasión del IVA, considerando directamente el análisis de redes, como tal, no se había realizado anteriormente. Es por ello que el método y los montos presentados en este estudio son totalmente perfectibles. Además, hay que tener en cuenta que de acuerdo a lo que dispone el artículo 6 párrafo tercero del CFF, los contribuyentes son quienes determinan las contribuciones a su cargo[24]. Así mismo, se precisa que, dado que se tomó como punto de partida el comportamiento de EFOS ya publicadas por el SAT, las estimaciones de evasión del IVA que se realizan en este estudio están asociadas a la presunta emisión de comprobantes de operaciones simuladas. Es probable que existan otros mecanismos de evasión del IVA que, debido al sesgo inherente en los datos trabajados, no son considerados en este estudio.

A continuación se describe la forma, datos y consideraciones que se tomaron en cuenta para llevar a cabo una estimación del monto evadido del IVA, principalmente por la emisión de CFDI generados por los RFCA sospechosos para los años 2015, 2016, 2017 y 2018. Principalmente se busca determinar la recaudación potencial. Esta representa la recaudación que se obtendría si todos los contribuyentes legalmente obligados pagasen sus impuestos y para este caso especifico, sería aquella recaudación del IVA que se obtendría si los CFDI simulados no existieran y, con ello, el IVA Trasladado expresado en los CFDI, no fuera efectivamente acreditado por el receptor de los mismos. Esto porque el propósito de la emisión de este tipo de comprobantes es reducir el monto del IVA que en realidad se tiene la obligación de pagar.

Partiendo de la información reportada en los CFDI, y solamente considerando los campos asociados a montos activos, definimos la recaudación potencial asociada a un RFCA arbitrario, $\text{rec}_{\text{IVA}}\phi_i$, como la diferencia entre el agregado anual del IVA Trasladado asociado a cada uno de sus CFDI de ingreso emitidos, el cual denotamos como $\text{IVA}_{\text{TA}\,i}$, y el IVA Neto pagado por el RFCA en el mismo periodo, $\text{IVA}_{\text{Neto}\,i}$, *i.e.*

$$\text{rec}_{\text{IVA}}\phi_i = \sum \text{IVA}_{\text{TA}\,i} - \text{IVA}_{\text{Neto}\,i}. \tag{1}$$

Posteriormente, definimos la recaudación potencial total de una población de RFCA $\text{REC}_{\text{IVA}}\phi$ como:

$$\text{REC}_{\text{IVA}}\phi = \sum_i \text{rec}_{\text{IVA}}\phi_i, \tag{2}$$

[22] Del 2014 al 2018, el SAT ha publicado 7,200 contribuyentes definitivos y 1,374 contribuyentes que presuntamente son EFOS. El listado publicado por el SAT es modificado constantemente de acuerdo al proceso que se lleva a cabo para desvirtuar a los contribuyentes como EFOS definitivos y presuntos. Es posible que al consultar el listado, las cifras hayan cambiado.

[23] Cifras presentadas mediante comunicado de prensa emitido el 25 de junio de 2019.

[24] Son los contribuyentes quienes deciden declarar totalmente, parcialmente o en su caso no considerar los CFDI que les corresponda para el cálculo en este caso del IVA.

i.e., corresponde a la suma de las recaudaciones potenciales individuales de cada uno de los miembros de la población. Por ejemplo, una población pueden ser los RFCA sospechosos miembros de la intersección de las listas de los métodos de clasificación.

Cabe aclarar que solamente se tomó en consideración el total del agregado del IVA Trasladado, debido a que suponemos que el impuesto, al provenir de CFDI emitidos por RFCA sospechosos, fue efectivamente acreditado por el receptor del comprobante. Por otra parte, lo que no tuvimos a nuestro alcance para determinar el cálculo, fue la interacción detallada entre el emisor y el receptor de los CFDI, a fin de conocer si los montos expresados en los mismos tuvieron algún efecto fiscal que pudo haber determinado algún otro comportamiento o dato a considerar para la realización del cálculo.

5.5.1. Datos para el cálculo

Los datos que utilizamos para estimar la evasión del IVA son los CFDI de tipo ingreso emitidos por los RFCA sospechosos entre 2015 y 2018 y declaraciones de IVA realizadas por los contribuyentes en las que se incluye el IVA neto pagado durante el mismo periodo. Dentro de la información proporcionada por el SAT respecto al CFDI de tipo ingreso, se incluyen aquellos que se encuentran Activos y Cancelados. Se hace esta precisión ya que los CFDI de tipo ingreso y que tienen un estatus de Cancelado, causan un interés especial, ya que dentro del proceso para el cálculo de la evasión, se percibió que el monto subtotal así como el monto del IVA Trasladado generado en estos comprobantes, en algunos años incluso superó los montos para los CFDI de tipo Ingreso considerados como Activos y tomados en cuenta para la realización del cálculo.

Consideramos que la información de las operaciones entre contribuyentes registradas en los CFDI es la que nos permite describir de mejor manera la actividad y mecanismos de evasión, ya que los montos reportados en las declaraciones, tanto de DIOT como del IVA son susceptibles de ser manipulados y pueden no corresponder a los ingresos y montos reales plasmados en los CFDI. Una diferencia significativa entre los montos expresados en el CFDI y los de las declaraciones presentadas, puede ser un indicador de prácticas ilícitas en la emisión de comprobantes.

En específico, como se describe en la ecuación 2, utilizamos para el cálculo de los estimados de evasión del IVA los agregados anuales de los *Montos del IVA Trasladados* reportados en los CFDI de ingreso emitidos por los RFCA sospechosos, y los montos de *IVA Neto efectivamente pagado* obtenidos de la base contable con información de las declaraciones del IVA presentadas por los RFCA sospechosos proporcionada por el SAT.

5.6. Estimados de montos evadidos anuales

Como se discutió en la sección 5.1, el número de EFOS cercanas a un RFCA dentro de la red de interacción es un indicador de su nivel de colusión dentro de las subredes de operaciones asociadas a EFOS, de tal forma que se puede formular la hipótesis de que un RFCA cercano a un gran número de EFOS publicadas, es mucho más susceptible de incurrir en el mismo tipo de prácticas a diferencia de uno que es cercano a un número bajo, el cual se puede considerar como menos susceptible de realizar emisiones asociadas a operaciones simuladas o anómalas. Con esto en mente, consideramos para el cálculo de los estimados de evasión en cada uno de los años considerados (2015 a 2018) solamente a los RFCA sospechosos más cercanos a las EFOS publicadas por el SAT (a una distancia $d \leq 3$, véase la sección 5.1), los cuales, como se muestra en la figura 16, corresponden a entre el 28 % y el 38 % del total de los RFCA sospechosos en cada año.

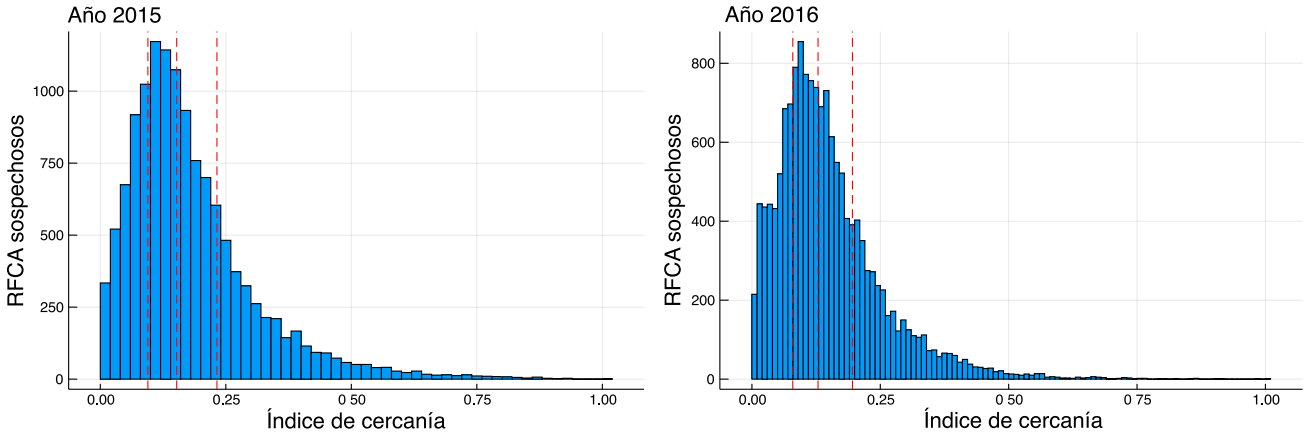
Con el objetivo de refinar el conjunto de RFCA sospechosos cercanos considerando sus características en las redes de interacción, definimos el *índice de cercanía*, $\sigma_i(y)$, asociado a un RFCA sospechoso arbitrario i en el año y, como el cociente entre el número total de EFOS cercanas a un RFCA durante un año, entre el número de meses en que dichas EFOS fueron cercanas al RFCA, *i.e.*:

$$\sigma_i(y) = \frac{\text{EFOS cercanas en } y}{\text{Meses en que fueron cercanas}}. \tag{3}$$

Cabe hacer notar que el número de meses en que las EFOS fueron cercanas al RFCA sospechoso no son necesariamente 12, ya que se pude dar el caso en que haya meses en los que el RFCA no fue cercano a ninguna EFOS en la red. Dado que el número de EFOS cercanas a los RFCA sospechosos cambia año con año (véase figura 12), para definir un criterio que pueda ser aplicado a cualquiera de los años tomados en cuenta, normalizamos el índice de cercanía de los RFCA, el cual denotamos por $\hat{\sigma}_i$, con respecto al valor máximo observado cada año, *i.e.*

$$\hat{\sigma}_i(y) = \frac{\sigma_i(y)}{\text{máx}(\sigma_i(y))}, \tag{4}$$

donde $\hat{\sigma}_i(y)$ tiene valores en el intervalo $[0, 1]$ y nos permite definir de forma cuantitativa un umbral para cada periodo, $\theta_\sigma(y)$, el cual junto con la condición $\hat{\sigma}_i(y) \geq \theta_\sigma(y)$ permite filtrar a los RFCA sospechosos con respecto a su nivel de colusión. Cabe hacer notar que $\theta_\sigma(y) = 0$ selecciona a todos los RFCA sospechosos cercanos a EFOS en ese año (la fracción de RFCA mostrada en la figura 16) y un valor $\theta_\sigma \approx 1$ selecciona a los más coludidos en la red. Dado que el valor del umbral es arbitrario damos estimados de los montos de evasión para los valores $\theta_\sigma = 0$ y a los primeros tres cuartiles de la distribución del índice de cercanía para cada año, lo que corresponde a todos los RFCA sospechosos cercanos a EFOS, y al 75, 50 y 25 % de los RFCA más coludidos (véase la figura 15).



(a) (b)

Figura 15: Índice de cercanía para los RFCA sospechosos en la intersección de las listas de los dos métodos de clasificación (RNA y BA) cercanos a EFOS ya identificadas. El índice de cercanía se utiliza como un método adicional de validación o refinamiento de la lista de sospechosos considerando propiedades observadas en las redes de interacción. Se muestran los resultados obtenidos para (a) 2015 y (b) 2016. Las líneas punteadas corresponden a los cuartiles asociados al 25, 50 y 75 % de los datos en la distribución

En la figura 16 se muestra el número de facturas emitidas y los estimados anuales de la evasión del IVA asociado a la emisión de CFDI de operaciones probablemente simuladas en millones de pesos (MDP) para los años 2015 a

2018 considerando un umbral de colusión $\theta_\sigma = 0$, *i.e.*, el total de los RFCA sospechosos presentes en las redes de interacción. En ambos casos se observa un comportamiento creciente que, junto con la observación de que el 91.15 % de los RFCA sospechosos se encuentran activos, podemos suponer que estos evasores potenciales al no haber sido identificados siguen realizando, e incluso incrementando, la emisión de CFDI potencialmente asociados a operaciones simuladas. Así mismo, se muestra el comportamiento de los montos evadidos en función del umbral de colusión definido como los tres primeros cuartiles de la distribución del índice de cercanía para cada año. Se puede observar que aun considerando solamente al 25 % de los RFCA sospechosos más coludidos con las EFOS en cada año[25], se obtiene un estimado entre 40,097.27 y 77,318.59 MDP entre 2015 y 2018.

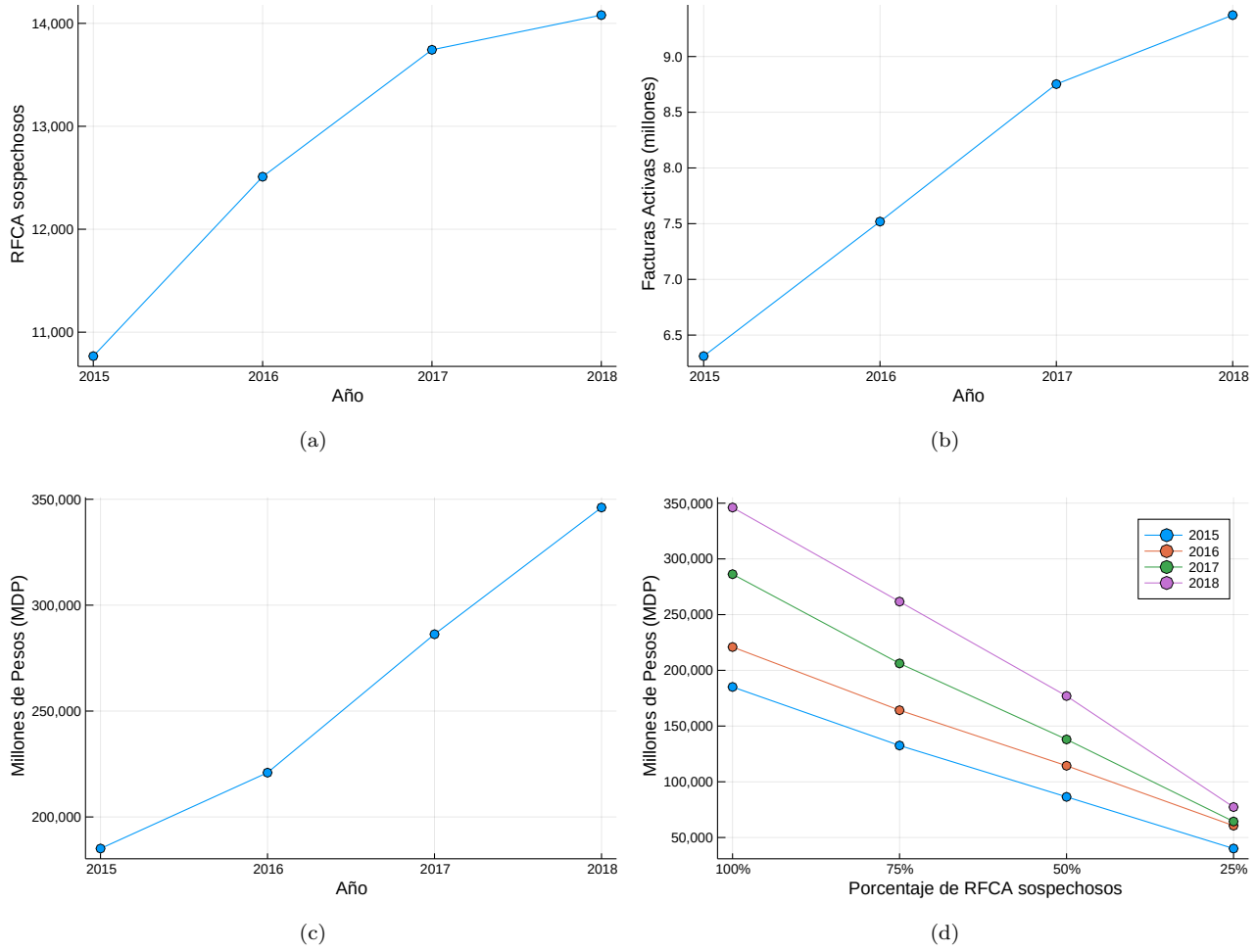
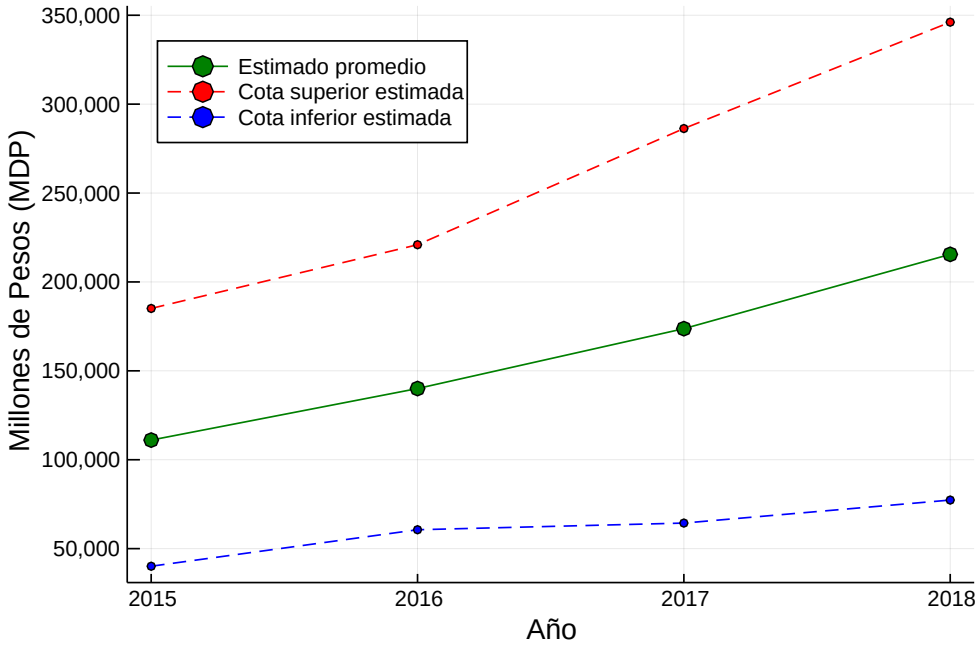


Figura 16: (a) RFCA sospechosos a distancia $d \leq 3$ en las redes de interacción. (b) Número de facturas activas emitidas por todos los RFCA sospechosos en las redes de interacción. (c) Estimados anuales de los montos de evasión del IVA en MDP asociado a operaciones potencialmente simuladas considerando todos los RFCA sospechosos que participan en las redes de interacción ($\theta_\sigma = 0$). (d) Estimados de los montos evadidos del IVA en MDP en función del porcentaje de RFCA sospechosos elegidos de acuerdo a su nivel de colusión dado por su índice de cercanía $\hat{\sigma}_i(y)$.

Los resultados obtenidos en función del umbral de colusión nos permiten establecer cotas superiores e inferiores para los estimados de los montos de evasión del IVA. Las cotas superiores que se reportan corresponden al estimado de

[25]Este conjunto de RFCA tiene los valores más altos del índice de cercanía EFOS en cada año y, como ya se ha mencionado en varias ocasiones, son los que se encuentran más coludidos con los EFOS publicados y por lo tanto más sospechosos de presentar un comportamiento similar



Estimados de Evasión MDP			
Año	Cota Inferior	Promedio	Cota Superior
2015	40,097.27	111,048.36	185,087.23
2016	60,626.86	140,041.13	220,922.03
2017	64,377.11	173,717.06	286,273.35
2018	77,318.59	215,518.71	346,106.32
Promedio anual	60,604.96	135,081.31	259,597.23

RFCA únicos		
Año	Cota Inferior	Cota Superior
2015	2,686	10,767
2016	3,132	12,510
2017	3,461	13,743
2018	3,541	14,080
Total	7,677	17,769

Figura 17: Cotas para los montos estimados de evasión de IVA en Millones de Pesos (MDP) asociado a la emisión de CFDI de operaciones potencialmente simuladas realizadas por los RFCA sospechosos para los años en el periodo 2015-2018. Se reporta también el número de RFCA únicos con los que se realiza el cálculo de la evasión en cada año. El número total de RFCA únicos que realizaron operaciones simuladas durante los cuatro años estudiados se estima entre 7,677 (cota mínima) y 17,769 (cota máxima).

evasión considerando a todos los RFCA sospechosos cercanos a EFOS en cada año y las cotas inferiores corresponden al estimado de evasión considerando al 25 % de los RFCA sospechosos que presentan los valores más altos del índice de cercanía, que corresponden a un total de 7,677 RFCA únicos con operaciones entre 2015 y 2018 asociado a un estimado promedio de 60,604.96 MDP anuales. Finalmente, el estimado promedio que reportamos corresponde al promedio de los montos evadidos obtenidos para cada corte del umbral de colusión. Las cotas reportadas en la figura 17 no deben ser consideradas como valores definitivos, ya que pueden existir otros mecanismos de evasión del IVA diferentes a la simulación de operaciones que no se consideran en este estudio. Más aún, cabe mencionar que un factor de incertidumbre adicional es que no conocemos que porcentaje de las operaciones asociadas a un RFCA sospechoso es simulado y asimismo no conocemos de manera precisa si este IVA trasladado fue realmente acreditado por el contribuyente receptor del CFDI. Dada la ausencia de un dato de referencia, consideramos arriesgado determinar un porcentaje por nuestra cuenta y hemos optado por considerar el 100 % de los comprobantes emitidos por los RFCA sospechosos como operaciones simuladas y que el IVA trasladado en estos comprobantes fue acreditado en su totalidad por otro contribuyente. Por este motivo, de manera moderada consideramos más realista usar la cota inferior. Un

estudio posterior enfocado en la trazabilidad o seguimiento de CFDI podría ser útil para ayudar a determinar dicho porcentaje de simulación de operaciones de manera cuantitativa y hacer un cálculo más preciso.

6. Conclusiones y limitaciones

Los métodos usados en este estudio generalizan el comportamiento de EFOS ya detectadas por el SAT a contribuyentes aún no identificados como sospechosos a partir de una comparación cuantitativa de sus actividades tributarias. Por una parte, esto implica que EFOS ocultas con patrones estadísticos fundamentalmente distintos a los ya detectados en principio no pueden ser categorizadas por estos métodos. Por otra parte, es posible que empresas honestas tengan patrones similares a las EFOS detectadas. Por lo tanto, a pesar de haber obtenido resultados alentadores, los métodos propuestos en este estudio no son perfectos, y no reemplazan a humanos (o a una investigación fiscal extensiva) en la decisión de si un contribuyente simula operaciones o no. El objetivo de este estudio no es sustituir los esfuerzos actuales del SAT en la lucha contra la evasión fiscal, sino complementar tales esfuerzos con herramientas cuantitativas en la frontera de la investigación mundial.

Sería exagerado aspirar a eliminar por completo la evasión fiscal. Pero con el desarrollo de herramientas para detectar evasión en conjunto con la actuación de las autoridades correspondientes, se podría inhibir considerablemente una práctica que, aunque no es predominante, tampoco es rara. Y en consecuencia aumentaría la recaudación fiscal de forma considerable.

El número de RFCA sospechosos de ser EFOS y los montos evadidos son estimaciones de valores reales desconocidos. No deberían ser considerados como una estimación final o como iguales a los valores reales. Su utilidad radica en poder estimar de forma rápida y eficiente el orden de magnitud de la evasión del IVA en términos de patrones ya detectados, usando una cantidad de recursos del SAT mucho menor que la asociada a investigaciones tradicionales de evasión fiscal. Los resultados de nuestro estudio pueden servir para identificar nuevas sospechosas de ser EFOS de manera rápida, a fin de que el SAT pueda actuar legalmente antes de que las personas detrás de las EFOS se den de baja, registren otra empresa, o recluten a otra persona para realizar el mismo tipo de operaciones ilícitas.

Por último, los resultados de RFCA sospechosos así como los montos derivados de su análisis, no se pueden tomar o catalogar de ninguna manera como contribuyentes y operaciones de EFOS, esto es en primera instancia porque el equipo de investigadores que realizó el estudio no tiene la facultad legal ni los medios necesarios para hacer tal determinación y por otra parte, de acuerdo al articulo 69-B analizado anteriormente, se tiene que seguir un procedimiento preciso para poder efectuar una determinación de esta magnitud. Por lo tanto, dentro de este procedimiento en su caso el SAT, tendría que notificar mediante buzón tributario, notificación personal o mediante una publicación en el DOF, a los contribuyentes que así lo considere pertinente a partir de los resultados de este estudio.

6.1. Recomendaciones

Con base en los resultados de este estudio, nos permitimos emitir las siguientes recomendaciones:

1. Integrar un sistema automático de monitoreo y detección de EFOS sospechosas, basado en los métodos de este estudio o similares, a las herramientas tecnológicas con las que ya cuenta el SAT. Al analizar la actividad cotidiana de todos los contribuyentes, se podrían identificar presuntos evasores de manera ágil e informar a las

instancias correspondientes. Sistemas similares se podrían desarrollar para detectar y alertar sobre otros tipos de evasión y lavado de dinero.

2. Una posible práctica de los RFCA sospechosos de ser EFOS detectados por este estudio es emitir facturas para que quien las reciba deduzca impuestos y después las cancele. A partir de 2019, los receptores reciben un aviso en su buzón tributario cuando los emisores desean cancelar una factura para que se apruebe la cancelación. Sin embargo, hacen falta mecanismos para asegurar la recuperación de los impuestos evadidos con facturas canceladas. Es recomendable generar un correo electrónico personalizado dirigido a quien fue emisor y receptor del CFDI cancelado con el recordatorio que debe presentar su declaración complementaria en dado caso de haber aplicado el CFDI en cuestión para los efectos fiscales que correspondan.

3. Una práctica similar se da cuando las EFOS se dan de baja después de haber vendido facturas durante algunos meses, antes de que sean detectadas por el SAT. Se recomienda evaluar el proceso actual de suspensión de actividades. Analizando automáticamente el comportamiento previo de las empresas que solicitan darse de baja (montos manejados, fecha de creación, cancelación de facturas, etc.), se podrían generar alertas para tomar las acciones correspondientes, con la intención de reducir la incidencia de esta práctica.

4. En este estudio hemos detectado que muchas EFOS emiten facturas a sí mismos. A partir de 2019 se ha implementado un candado para prevenir auto-facturas. Sin embargo, entre dos o más empresas se pueden obtener efectos similares con técnicas de circularidad. Recomendamos implementar herramientas que detecten automáticamente flujos circulares de activos en redes de emisiones y recepciones.

5. Los RFCs genéricos — usados cuando un receptor no tiene RFC — limitan la posibilidad de rastrear contribuyentes evasores y, por tanto, disminuyen la precisión de nuestros métodos. Recomendamos evaluar medidas para reducir el uso de RFCs genéricos, o bien aumentar la rastreabilidad de los CFDI que los usen.

6. Aprovechar al máximo la capacidad de información que genera el CFDI como base de datos para crear un formulario tanto mensual como anual de cálculo de impuestos que permita el rastreo de los CFDI que se están declarando sin que el contribuyente sienta una imposición por parte del SAT en la información que se tiene que verter en dicho formulario.

7. Crear el comprobante simplificado electrónico que permita vincular las ventas al público en general con la emisión del CFDI que genere el contribuyente para cumplir con su obligación de facturar todas aquellas ventas por las que sus clientes no le pidan un CFDI.

8. Considerar la modificación de la Ley General de Sociedades Mercantiles o las leyes aplicables a efecto de endurecer la responsabilidad de los fedatarios públicos en el acto de constitución de personas morales, ya que, las empresas que se dedican a la comercialización de CFDI simulados, desechan y crean nuevas sociedades a fin de no ser detectados ni localizados a tiempo, siendo estos quienes las constituyen sin corroborar plenamente la capacidad financiera o material de la sociedad.

9. Crear un ID de acreditamiento y traslado que permita la identificación del monto y del CFDI que dio origen al IVA que se pretende acreditar o trasladar en la declaración mensual correspondiente.

6.2. Trabajo a futuro

Los resultados obtenidos en este proyecto abren la puerta a diversas preguntas de investigación que nos gustaría abordar en un futuro cercano:

1. De la lista entregada al SAT de RFCA con alta probabilidad de ser EFOS, sería útil recibir retroalimentación de los resultados de las investigaciones internas del SAT, a fin de mejorar nuestros métodos de clasificación y detección automática de EFOS.

2. Realizar un estudio más específico sobre el comportamiento de EFOS para poder refinar métodos automáticos de identificación.

3. Usar análisis de componentes principales para asignar una ponderación a cada factor de riesgo en un índice global de probabilidad que permita priorizar sospechosos a investigar.

4. Extender nuestro estudio incluyendo un análisis de la evasión de ISR e IEPS.

5. Mejorar la estimación de impuestos evadidos analizando más datos y montos específicos, revisando DIOT más detenidamente y estudiando el flujo de facturas entre los agentes de las redes de evasión.

6. Analizar el destino de los CFDI detectados como simulados, ya que independientemente del uso fiscal que se les dé, estos comprobantes también pueden usarse para lavado de dinero, corrupción, tráfico de mercancías e importaciones y exportaciones ilegales.

7. Analizar los efectos causados por los CFDI emitidos por las EFOS que se encuentran cancelados y verificar el impacto que causa la aplicación de este tipo de comprobantes en la recaudación tributaria.

8. En el contexto de ciencia de redes, realizar un estudio de distribuciones de *motifs* [67] (patrones locales) alrededor de EFOS, para detectar comunidades de evasión. Por ejemplo, detectamos patrones de varios contribuyentes alrededor de EFOS que tanto reciben como emiten facturas de las EFOS.

Créditos

Por orden alfabético.

Directores del proyecto

Dr. Carlos Gershenson, Dr. Gerardo Iñiguez, Dr. Carlos Pineda.

Investigadores

Lic. Rita Guerrero, Lic. Eduardo Islas, Mtro. Omar Pineda, Mtro. Martín Zumaya.

Agradecimientos

Lic. Ana Camila Baltar Rodríguez, Mtro. Romel Calero, Mtro. José Luis Gordillo, Dr. Alejandro Frank Hoeflich, Mtro. Ollin Langle, Juan Antonio López Rivera, Dr. José Luis Mateos Trigos, Ing. Eric Solís Montufar, Dr. Juan Claudio Toledo Roy, Dr. Octavio Zapata Fonseca.

Referencias

[1] Malcolm K. Sparrow. The application of network analysis to criminal intelligence: An assessment of the prospects. *Social Networks*, 13(3):251 – 274, 1991.

[2] Luciano da Fontoura Costa, Osvaldo N. Oliveira Jr., Gonzalo Travieso, Francisco Aparecido Rodrigues, Paulino Ribeiro Villas Boas, Lucas Antiqueira, Matheus Palhares Viana, and Luis Enrique Correa Rocha. Analyzing and modeling real-world phenomena with complex networks: a survey of applications. *Advances in Physics*, 60(3):329–412, 2011.

[3] Amanda L. Andrei, Kevin Comer, and Matthew Koehler. An agent-based model of network effects on tax compliance and evasion. *Journal of Economic Psychology*, 40:119 – 133, 2014. Special Issue on Behavioral Dynamics of Tax Evasion.

[4] Maria R. D'Orsogna and Matjaž Perc. Statistical physics of crime: A review. *Physics of Life Reviews*, 12:1 – 21, 2015.

[5] Luis Natera, Federico Battiston, Gerardo Iñiguez, and Michael Szell. Data-driven strategies for optimal bicycle network growth. *arXiv preprint arXiv:1907.07080*, 2019.

[6] José Tapia Tovar. *La evasión fiscal: Causas, efectos y soluciones*. Porrúa, 2000.

[7] Servicio de Administración Tributaria (SAT). Glosario: Informe tributario y de gestión. `http://www2.sat.gob.mx/sitio_internet/informe_tributario/informe2013t4/glosario.pdf`. Último acceso: Octubre 2019.

[8] Cecilia Licona Vite. *Estudio sobre la evasión y la elusión fiscales en México*. Cámara de Diputados, LXI Legislatura, 2011.

[9] Definición efos y edos. `https://digitalinvoice.com.mx/efos-y-edos/`. Último acceso: Diciembre 2019.

[10] Jorge Alberto Reyes Caballero. La importancia del código fiscal de la federación en la actividad económica. `https://www.soyconta.com/la-importancia-del-codigo-fiscal-de-la-federacion-en-la-actividad-economica/`. Último acceso: Diciembre 2019.

[11] Instituto Mexicano de Contadores Publicos. *Resolucion Miscelanea Fiscal 2017*. Instituto Mexicano de Contadores Publicos, Ciudad de México, 1st edition, 2017.

[12] Centro de Estudios de Finanzas Públicas. Importancia del impuesto al valor agregado. 2017.

[13] GPM Contadores y Auditores S.C. Seminario fiscal. 2016.

[14] Camara de Diputados. Ley del impuesto al valor agregado. `http://www.diputados.gob.mx/LeyesBiblio/pdf/77_091219.pdf`. Último acceso: Diciembre 2019.

[15] SAT. Guía de llenad para los comprobantes fiscales digitales por internet. `http://omawww.sat.gob.mx/tramitesyservicios/Paginas/documentos/GuiaAnexo20.pdf`. Último acceso: Agosto 2019.

[16] Consultoria SAP. Todo sobre cfdi. `https://www.consultoria-sap.com/2018/04/todo-sobre-cfdi.html`. Último acceso: Septiembre 2019.

[17] Colegio de Contadores Públicos de México. Reforma al artículo 69-b del código fiscal de la federación. *Boletín de Investigación de la Comisión Fiscal 3*, (65):1–6, 2018.

[18] PWC. Reforma fiscal 2020. http://explore.pwc.com/c/66-4?x=sTGTPe&utm_source=Website&utm_medium=SiteRF20&utm_content=VerPF. Último acceso: Septiembre 2019.

[19] Domicián Máté, Rabeea Sadaf, Tibor Tarnóczi, and Veronika Fenyves. Fraud detection by testing the conformity to benford's law in the case of wholesale enterprises. *Polish Journal of Management Studies*, 16, 2017.

[20] Marcel Ausloos, Roy Cerqueti, and Tariq A Mir. Data science for assessing possible tax income manipulation: The case of italy. *Chaos, Solitons & Fractals*, 104:238–256, 2017.

[21] Theoharry Grammatikos and Nikolaos Papanikolaou. Applying benford's law to detect fraudulent practices in the banking industry. Working paper, University of Luxembourg, Luxembourg, 2016.

[22] Wendy K Tam Cho and Brian J Gaines. Breaking the (benford) law: Statistical fraud detection in campaign finance. *The american statistician*, 61(3):218–223, 2007.

[23] Luis Pericchi and David Torres. Quick anomaly detection by the newcomb—benford law, with applications to electoral processes data from the usa, puerto rico and venezuela. *Statistical science*, pages 502–516, 2011.

[24] Richard J Bolton and David J Hand. Statistical fraud detection: A review. *Statistical science*, pages 235–249, 2002.

[25] Sushmito Ghosh and Douglas L Reilly. Credit card fraud detection with a neural-network. In *System Sciences, 1994. Proceedings of the Twenty-Seventh Hawaii International Conference on*, volume 3, pages 621–630. IEEE, 1994.

[26] Emin Aleskerov, Bernd Freisleben, and Bharat Rao. Cardwatch: A neural network based database mining system for credit card fraud detection. In *Proceedings of the IEEE/IAFE 1997 computational intelligence for financial engineering (CIFEr)*, pages 220–226. IEEE, 1997.

[27] Raghavendra Patidar, Lokesh Sharma, et al. Credit card fraud detection using neural network. *International Journal of Soft Computing and Engineering (IJSCE)*, 1(32-38), 2011.

[28] Efstathios Kirkos, Charalambos Spathis, and Yannis Manolopoulos. Data mining techniques for the detection of fraudulent financial statements. *Expert systems with applications*, 32(4):995–1003, 2007.

[29] Fan Yu, Zheng Qin, and Xiao-Ling Jia. Data mining application issues in fraudulent tax declaration detection. In *Proceedings of the 2003 International Conference on Machine Learning and Cybernetics (IEEE Cat. No. 03EX693)*, volume 4, pages 2202–2206. IEEE, 2003.

[30] Corinna Cortes, Daryl Pregibon, and Chris Volinsky. Communities of interest. In *International Symposium on Intelligent Data Analysis*, pages 105–114. Springer, 2001.

[31] Erik Hemberg, Jacob Rosen, Geoff Warner, Sanith Wijesinghe, and Una-May O'Reilly. Detecting tax evasion: a co-evolutionary approach. *Artificial Intelligence and Law*, 24(2):149–182, 2016.

[32] Razieh Tabandeh, Mansor Jusoh, Nor Ghani Md Nor, and Mohd Azlan Shah Zaidi. Estimating factors affecting tax evasion in Malaysia: A neural network method analysis. *Persidangan Kebangsaan Ekonomi Malaysia ke VII (PERKEM VII), Transformasi Ekonomi dan Sosial Ke Arah Negara Maju, Ipoh, Perak*, pages 4–6, 2012.

[33] Eghbal Rahimikia, Shapour Mohammadi, Teymur Rahmani, and Mehdi Ghazanfari. Detecting corporate tax evasion using a hybrid intelligent system: A case study of iran. *International Journal of Accounting Information Systems*, 25:1–17, 2017.

[34] Luciano A Digiampietri, Norton Trevisan Roman, Luis AA Meira, Cristiano D Ferreira, Andreia A Kondo, Everton R Constantino, Rodrigo C Rezende, Bruno C Brandao, Helder S Ribeiro, Pietro K Carolino, et al. Uses of artificial intelligence in the Brazilian customs fraud detection system. In *Proceedings of the 2008 international conference on digital government research*, pages 181–187. Digital Government Society of North America, 2008.

[35] Johannes Wachs and János Kertész. A network approach to cartel detection in public auction markets. *Scientific Reports*, 9:10818, 2019.

[36] Luca Maria Aiello, Alain Barrat, Rossano Schifanella, Ciro Cattuto, Benjamin Markines, and Filippo Menczer. Friendship prediction and homophily in social media. *ACM Transactions on the Web (TWEB)*, 6(2):9, 2012.

[37] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.

[38] Sergio Currarini, Jesse Matheson, and Fernando Vega-Redondo. A simple model of homophily in social networks. *European Economic Review*, 90:18–39, 2016.

[39] Lu Hongtao and Zhang Qinchuan. Applications of deep convolutional neural network in computer vision. *J. Data Acquis. Process*, 31(01):1–17, 2016.

[40] Ganesh K Venayagamoorthy, Viresh Moonasar, and Kumbes Sandrasegaran. Voice recognition using neural networks. In *Proceedings of the 1998 South African Symposium on Communications and Signal Processing-COMSIG'98 (Cat. No. 98EX214)*, pages 29–32. IEEE, 1998.

[41] J. Zhang and C. Zong. Deep neural networks in machine translation: An overview. *IEEE Intelligent Systems*, 30(05):16–25, sep 2015.

[42] Gerald Tesauro and Terrence J Sejnowski. A'neural'network that learns to play backgammon. In *Neural Information Processing Systems*, pages 794–803, 1988.

[43] Christopher Clark and Amos Storkey. Training deep convolutional neural networks to play go. In *International conference on machine learning*, pages 1766–1774, 2015.

[44] Sebastian Starke, He Zhang, Taku Komura, and Jun Saito. Neural state machine for character-scene interactions. *ACM Transactions on Graphics*, 38, 11 2019.

[45] Filippo Amato, Alberto López, Eladia María Peña-Méndez, Petr Vaňhara, Aleš Hampl, and Josef Havel. Artificial neural networks in medical diagnosis, 2013.

[46] Takashi Kimoto, Kazuo Asakawa, Morio Yoda, and Masakazu Takeoka. Stock market prediction system with modular neural networks. In *1990 IJCNN international joint conference on neural networks*, pages 1–6. IEEE, 1990.

[47] Hirotaka Mizuno, Michitaka Kosaka, Hiroshi Yajima, and Norihisa Komoda. Application of neural network to technical analysis of stock market prediction. *Studies in Informatic and control*, 7(3):111–120, 1998.

[48] Rick L Wilson and Ramesh Sharda. Bankruptcy prediction using neural networks. *Decision support systems*, 11(5):545–557, 1994.

[49] A. Shen, R. Tong, and Y. Deng. Application of classification models on credit card fraud detection. In *2007 International Conference on Service Systems and Service Management*, pages 1–4, June 2007.

[50] Robert R Trippi and Efraim Turban. *Neural networks in finance and investing: Using artificial intelligence to improve real world performance*. McGraw-Hill, Inc., 1992.

[51] Lean Yu, Shouyang Wang, and Kin Keung Lai. Credit risk assessment with a multistage neural network ensemble learning approach. *Expert systems with applications*, 34(2):1434–1444, 2008.

[52] Nathalie Japkowicz. The class imbalance problem: Significance and strategies. In *Proc. of the Int'l Conf. on Artificial Intelligence*, 2000.

[53] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997.

[54] Sepp Hochreiter. Learning causal models of relational domains. Master's thesis, Institut fur Informatik, Technische Universitat, Munchen, 1991.

[55] Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. Lstm: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 28(10):2222–2232, 2016.

[56] Wenpeng Yin, Katharina Kann, Mo Yu, and Hinrich Schütze. Comparative study of cnn and rnn for natural language processing. *arXiv preprint arXiv:1702.01923*, 2017.

[57] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[58] C. J. Van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd edition, 1979.

[59] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[60] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):539–550, 2008.

[61] Jason W Osborne. Improving your data transformations: Applying the box-cox transformation. *Practical Assessment, Research & Evaluation*, 15(12):1–9, 2010.

[62] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

[63] Edson Zangiacomi Martinez, Francisco Louzada Neto, and Basílio de Bragança Pereira. A curva roc para testes diagnósticos. *Cadernos de Saúde Coletiva*, 11(1):7–31, 2003.

[64] Servicio de Administración Tributaria (SAT). Sitio de estadística. `http://omawww.sat.gob.mx/cifras_sat/Paginas/datos/vinculo.html?page=ListCompleta69B.html`. Último acceso: Noviembre 1 2019.

[65] CEPAL. Estadísticas tributarias para América Latina y el Caribe. Publicación anual, 2019.

[66] Servicio de Administración Tributaria (SAT). Información estadística del sat. `http://omawww.sat.gob.mx/cifras_sat/Paginas/inicio.html`. Último acceso: Noviembre 1 2019.

[67] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: Simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.