Universidad Autónoma de Aguascalientes
Centro de Ciencias Básicas
Departamento de Sistemas de Información

TESIS

Nombre del proyecto:

Análisis comparativo de algoritmos de búsqueda de coincidencia de nombres por variación fonética contra la lista de OFAC para determinar su eficacia contra una lista de nombres en español

Presenta:

José de Jesús Martínez Cámara

PARA OBTENER EL GRADO DE MAESTRO EN INFORMÁTICA Y TECNOLOGÍAS COMPUTACIONALES

TUTOR:

Dr. Juan Muñoz López

COMITÉ TUTORAL:

Dra. Eunice Esther Ponce de León Sentí

c. Dr. Lizeth Itziguery Solano Romo

Aguascalientes, Ags., a 01 de Junio de 2016

UNIVERSIDAD AUTONOMA
DE AGUASCALIENTES

JOSÉ DE JESÚS MARTÍNEZ CÁMARA
MAESTRÍA EN INFORMÁTICA Y TECNOLOGÍAS COMPUTACIONALES
P R E S E N T E.

Estimado alumno:

Por medio de este conducto me permito comunicar a Usted que habiendo recibido los votos aprobatorios de los revisores de su trabajo de tesis y/o caso práctico titulado: **"Análisis comparativo de algoritmos de búsqueda de coincidencia de nombres por variación fonética contra la lista de OFAC para determinar su eficacia contra la lista de nombres en español",** hago de su conocimiento que puede imprimir dicho documento y continuar con los trámites para la presentación de su examen de grado.

Sin otro particular me permito saludarle muy afectuosamente.

A T E N T A M E N T E
Aguascalientes, Ags., a 01 de junio de 2016
"Se lumen proferre"
EL DECANO

M. en C. JOSE DE JESUS RUIZ GALLEGOS

c.c.p.- Archivo.

UNIVERSIDAD AUTONOMA DE AGUASCALIENTES

M. EN C. JOSÉ DE JESÚS RUÍZ GALLEGOS
DECANO (A) DEL CENTRO DE CIENCIAS

P R E S E N T E

Por medio del presente como Tutor designado del estudiante **JOSÉ DE JESÚS MARTÍNEZ CÁMARA** con ID 159788 quien realizó la tesis titulada: **ANÁLISIS COMPARATIVO DE ALGORITMOS DE BÚSQUEDA DE COINCIDENCIA DE NOMBRES POR VARIACIÓN FONÉTICA CONTRA LA LISTA DE OFAC PARA DETERMINAR SU EFICACIA CONTRA UNA LISTA DE NOMBRES EN ESPAÑOL**, y con fundamento en el Artículo 175, Apartado II del Reglamento General de Docencia, me permito emitir el **VOTO APROBATORIO**, para que él pueda proceder a imprimirla, y así como continuar con el procedimiento administrativo para la obtención del grado.

Pongo lo anterior a su digna consideración y sin otro particular por el momento, me permito enviarle un cordial saludo.

A T E N T A M E N T E
"Se Lumen Proferre"
Aguascalientes, Ags., a 27 de Mayo de 2016.

Dr. Juan Muñoz López
Tutor de Tesis.

Dra. Eunice Esther Ponce de León Sentí
Asesor de Tesis

c. Dr. Lizeth Itziguery Solano Romo
Asesor de Tesis

c.c.p.- Interesado
c.c.p.- Secretaría de Investigación y Posgrado
c.c.p.- Consejero Académico
c.c.p.- Minuta Secretario Técnico

UNIVERSIDAD AUTÓNOMA DE AGUASCALIENTES

POSGRADOS UAA

DICTAMEN DE REVISIÓN DE LA TESIS / TRABAJO PRÁCTICO

DATOS DEL ESTUDIANTE	
NOMBRE: JOSÉ DE JESÚS MARTÍNEZ CÁMARA	ID (No. de Registro): 159788
PROGRAMA: MAESTRÍA EN INFORMÁTICA Y TECNOLOGIAS COMPUTACIONALES.	ÁREA: LA INGENIERÍA / ENFOQUE DE SISTEMAS PARA EL MEJORAMIENTO DE LOS PROCESOS ORGANIZACIONALES USANDO SI/TI
TUTOR/TURORES: DR. JUAN MUÑOZ LÓPEZ	
TESIS (X) TRABAJO PRÁCTICO ( )	
OBJETIVO: REALIZAR UN ANÁLISIS COMPARATIVO DE ALGORITMOS DE BÚSQUEDA DE COINCIDENCIA DE NOMBRES POR VARIACIÓN FONÉTICA CONTRA LA LISTA DE OFAC PARA DETERMINAR SU EFICACIA CONTRA UNA LISTA DE NOMBRES EN ESPAÑOL.	
DICTAMEN	
CUMPLE CON CRÉDITOS ACADÉMICOS:	(SI)
CONGRUENCIAS CON LAS LGAC DEL PROGRAMA:	(SI)
CONGRUENCIA CON LOS CUERPOS ACADÉMICOS:	(SI)
CUMPLE CON LAS NORMAS OPERATIVAS:	(SI)
CONINCIDENCIA DEL OBJETIVO CON EL REGISTRO:	(SI)

Aguascalientes, Ags. a 1 de JUNIO de 2016

FIRMAS

DRA. MARÍA DOLORES TORRES SOTO.
CONSEJERO ACADÉMICO DEL ÁREA

DR. JOSÉ MANUEL MORA TAVAREZ.
SECRETARIO TÉCNICO DEL POSGRADO

DR. JUAN JAUREGUI RINCÓN.
SECRETARIO DE INVESTIGACIÓN Y POSGRADO

Código: FO-040200-23
Revisión: 00
Emisión: 21/02/11

AGRADECIMIENTOS

Quisiera aprovechar este espacio para agradecer a todas las personas que siempre estuvieron apoyándome y que me tendieron su mano más de una vez en diferentes ocasiones durante el curso de mi maestría, y que sin ellas este trabajo no hubiera sido posible.

Le agradezco a mis profesores dentro de la maestría, por dedicar ese tiempo a prepararme y a transmitirme su conocimiento siempre con el fin de hacerme mejor persona y más preparada. Al Dr. Juan Muñoz, mi tutor de tesis, quien desde el día 1, me acompaño y me guío en la elaboración de esta tesis. A mis asesores de tesis, la Dra. Eunice y la c. Dr. Lizeth, que aun dentro de sus múltiples ocupaciones, tenían siempre un momento para mí, y para mis dudas.

A la Universidad Autónoma de Aguascalientes, porque al ser esa institución con tanto prestigio y aceptarme dentro de su curso de maestría, elevó mi nivel de estudio y me dio la oportunidad de sobresalir. A CONACYT, por el apoyo financiero, que me ayudo para poder cursar mis estudios de maestría.

A mis padres por su amor, apoyo y comprensión cuando a veces me extravío. Sé que por ellos en mi nació el gusto por aprender y por crecer cada día más.

A mis hermanos, que son mi ejemplo y mis amigos de toda la vida.

A mi familia y amigos, que siempre han creído en mí, y que constantemente me demuestran lo mucho que significo para ellos.

Y dejo a lo último, porque para mí es lo más importante, a mi esposa. Nada de esto, hubiese sido posible sin su apoyo, sin su ayuda, su guía y sus regaños, es simplemente la raíz de este y más documentos que he hecho, hice y hare, y no solo eso, el motor que me impulsa a ser mejor cada día y a superarme como persona. Te amo IMA.

ÍNDICE GENERAL

ÍNDICE DE TABLAS

ÍNDICE DE GRÁFICAS

ACRÓNIMOS

Acrónimo	Significado
OFAC	Oficina de Control de Activos Extranjeros del Departamento de Tesoro de Estados Unidos
GAFI	Grupo de Acción Financiera sobre el Blanqueo de Capitales y Financiamiento al Terrorismo
ONU	Organización de las Naciones Unidas
CFATF	Grupo de Acción Financiera del Caribe
PC-R-EV	Comité de Expertos para la Evaluación de Medidas Contra el Lavado de Dinero del Consejo de Europa
ESAAMLG	Grupo Contra el Lavado de Dinero del Este y Sur de África
APGML	Grupo de Asia Pacífico Contra el Lavado de Dinero
GAFISUD	Grupo de Acción Financiera de América del Sur
LFPIORPI	Ley Federal para la Prevención e Identificación de Operaciones con Recursos de Procedencia Ilícita
SHCP	Secretaría de Hacienda y Crédito Público
KYC/CDD	Conociendo a tu Cliente/Debida Diligencia de Clientes
SAM	Monitoreo de Actividades Sospechozas
WLF	Revisión de Listas Negras
CMS	Servicio de Manejo de Casos
PwC	Price Waterhouse Cooper
SCT	Secretaría de Comunicaciones y Transportes
BSA	Ley del Secreto Bancario
PIB	Producto Interno Bruto
ONG	Organización No Gubernamental
INEGI	Instituto Nacional de Estadística y Geografía
PGR	Procuraduría General de la República
SMVDF	Salario Mínimo Vigente en el Distrito Federal
FBI	Buro Federal de Investigación
SDN	Nacionales Especialmente Designados
SSN	Número de Seguridad Social
CURP	Clave Única de Registro Poblacional
TP	True Positive o Positivo Verdadero por sus siglas en inglés
TN	True Negative o Negativo Verdadero por sus siglas en inglés
FP	False Positive o Falso Positvio por sus siglas en inglés
FN	False Negative o Falso Negativo por sus siglas en inglés
F-Measure	Armonía entre la precisión y la exactitud

RESUMEN

Este trabajo de tesis fue elaborado con el fin de poder analizar 3 algoritmos fonéticos, para su utilización en la búsqueda de nombres en español dentro de la lista negra de OFAC.

Todo esto con el objetivo de poder elegir el algoritmo que mejores resultados mostrase en relación a su índice de falsos positivos y efectividad en la búsqueda, para el combate al lavado de dinero y el financiamiento al terrorismo.

Se generó una lista de nombres hispanos con 10,000 entradas distribuidas aleatoriamente para hacer las pruebas de los algoritmos, tales que fueron obtenidos mediante el uso de la paquetería de algoritmos fonéticos provista por apache y que están programados en el lenguaje java.

Posteriormente se elaboró un software capaz de utilizar los algoritmos fonéticos, y que recibiera de entrada un conjunto de datos divididos en 15 muestras de 20 nombres y así poder ver el índice de coincidencias positivas de estos algoritmos, cuando se comparaban los conjuntos de datos contra la lista negra de OFAC.

Una vez obtenidas las coincidencias se procedió a analizar los datos, mediante un estudio estadístico de ANOVA de una sola vía, esto para un total de 6 variables distintas que al final nos ayudarían a determinar que algoritmo tiene un mejor comportamiento y eficacia al ser utilizado con una lista de nombres en español.

Al final se analizaron los resultados y se arrojaron las conclusiones pertinentes, seleccionando el algoritmo que mejor cumplía con las premisas de los objetivos planteados en este documento

Por último, se platicaron las recomendaciones para un estudio posterior que parta de la investigación hecha en este documento y que se beneficie de los resultados arrojados al final del análisis de esta tesis.

ABSTRACT

The main purpose of this work was to analyse the characteristics of 3 phonetic algorithms in order to discern which of them is the most suitable to be used in the search of names written in Spanish at the OFAC blacklist.

The chosen algorithm was the one with the lowest false positive rate while continues being effective in the search of any given name. The aim is to use the result of this study to help the combat of money laundry and terrorism financing.

The phonetic algorithms were obtained from the java encode package provided by apache, and it was generated a list with 10,000 names in spanish with a random generation pattern to be used with the phonetic algorithms and the OFAC blacklist.

A special software was developed in order to use the phonetic algorithm package and the lists of names and blacklist, this software accepted as an entry value a list of files that was matched with the phonetic algorithms and the OFAC blacklist.

The sample was created forming 15 groups of 20 entries and put them in files that were processed by the software created, all this in order to create some metrics that were used to analyze the algorithms and the different variables that helped proving the objetives.

Finally the information was compared using the one way ANOVA test in order to determine which algorithm showed the best results with the different variables evaluated.

INTRODUCCIÓN

El combate al lavado de dinero y al financiamiento al terrorismo es hoy en dia en México, y en el mundo un punto importante de seguridad, tanto así que en cada informe de gobierno en México es agregado un rubro donde se presentan los resultados en estos dos temas.

Es de recalcar el esfuerzo que distintos organismos alrededor del mundo invierten para reformular las leyes que regirán en la mayoría de los países y que ayudaran a tener una homologación de esfuerzos para poder así ser más efectivos combatiendo todo tipo de actividades delictivas relacionadas al lavado de dinero.

Hoy en día, donde el número de transacciones y operaciones bancarias que se hacen por medios electrónicos va en crecimiento, así también la seguridad debe de ir a la par y por este motivo es necesario contar con software que ayude a discernir entre los distintos tipos de delitos para poder encontrar patrones o detalles que apoyen en la lucha contra todo tipo de fraude electrónico y bancario, así como, la prevención del blanqueo de capitales y permitiendo así poder detener una transferencia o una operación bancaria a aquellas personas que no tengan permitido hacer este tipo de operaciones, dado que se encuentran dadas de altas en alguna lista con nombres de personas con las que no se tiene permitido hacer negocios (lista negra).

En el mundo existen más de 400 listas negras, alimentadas por gobiernos de distintos países alrededor del mundo u organizaciones de la iniciativa privada, Una de las listas más conocidas y usadas alrededor del mundo, es la lista de OFAC (Oficina de Control de Activos Extranjeros del Departamento de Tesoro de Estados Unidos, por sus siglas en inglés) que concentra información de muchas de las listas existentes como la del FBI, o países como Reino Unido, Nigeria, entre otros. Esta lista es tomada como base para ser utilizada en otros países, uno de ellos, México (Videgaray Caso, 2014).

Siendo conscientes que un banco puede llegar a manejar miles de transacciones diarias, entre cientos o miles de personas, hacer la revisión de nombres de los clientes contra la lista de OFAC tiene que ser un proceso que permita poder detener o dejar pasar una transacción en cuestión de segundos, es aquí donde la informática juega un papel importante, ya que la implementación y uso de algoritmos que permitan hacer estas búsquedas es imperativa.

Los algoritmos de búsqueda aproximada son de distintos tipos y algunos pueden ayudar a otros a precisar los resultados, por ejemplo, los algoritmos que mejores resultados dan y en menor tiempo, son los algoritmos de lógica difusa, pero se puede mejorar este tipo de búsquedas y mejorar los tiempos al hacer uso de algoritmos de coincidencia de nombres por variación fonética, esto es, según como se pronuncie un nombre en determinado país, puede haber similitud con el nombre escrito de una forma distinta, ejemplo, Steve y Stephen, los cuales para algunos algoritmos de tipo fonético, tendrían similitudes y podrían ser considerados como el mismo nombre.

Los algoritmos fonéticos generan cadenas de caracteres llamados tokens que al ser ingresados en el proceso de algoritmo de lógica difusa, podrían ayudar en la mejora de los resultados (Ramachandran, 2014).

En esta investigación se revisaron 3 de los más usados algoritmos de coincidencia de nombres por variación fonética: soundex, NYSIIS, metaphone (Lait & Randell, 1995), adaptados a distintos lenguajes y se buscará determinar que algoritmo tendría mejores resultados para la búsqueda de nombres contrastada contra la lista de OFAC.

Para tal fin, la siguiente investigación consta de 5 capítulos principales:

1. **Planteamiento del problema**
    - En este capítulo se revisan los antecedentes del tema, su relevancia, justificación y objetivos y se establece un precedente para abordar la investigación contenida en este documento.

2. **Marco Teórico**
    - En este apartado se trata con más detalle el sustento teórico de los temas utilizados en la elaboración de esta investigación. Se hace una revisión a los algoritmos utilizados para la búsqueda de nombres así como se profundizará en temas de prevención de crimen financiero, en este caso, lavado de dinero y financiamiento al terrorismo.
3. **Diseño Metodológico**
    - En este capítulo se describe la metodología que dará soporte a las diferentes pruebas, esto con el fin de apoyar a la generación de conclusiones y de resultados significativos.
4. **Análisis de Resultados**
    - Este tema se revisan los resultados obtenidos, para cimentar las conclusiones finales, y en base a estas conclusiones derivar en la selección de un algoritmo que apoye a tener mejores resultados en la búsqueda de nombres contra la lista de OFAC.
5. **Conclusiones**

    Para finalizar, en este capítulo, se platica de la enseñanza que nos dejó esta investigación, posibles aplicaciones y futuros estudios que puedan derivar de los datos obtenidos.

Capítulo 1: Planteamiento del Problema

1.2 Antecedentes

1.2.1 Lavado de Dinero y Combate al Terrorismo

1.2.1.1¿Qué es el Lavado de Dinero, o Lavado de Activos?

El lavado de activos puede referirse de varias maneras. La mayoría de los países aceptan la definición aprobada por la Convención de las Naciones Unidas contra el Tráfico Ilícito de Estupefacientes y Sustancias Psicotrópicas (ONU, 1988) y la Convención de las Naciones Unidas contra la Delincuencia Organizada Transnacional (ONU, 2000):

Una definición apropiada para el lavado de activos la da Paul Schott (Schott, 2007) y dice: "Es *la conversión o la transferencia de bienes, a sabiendas de que tales bienes provienen de alguno o algunos de los delitos [de narcotráfico], o de un acto de participación en tal delito o delitos, con objeto de ocular o encubrir el origen ilícito de los bienes o de ayudar a cualquier persona que participe en la comisión de tal delito o delitos a eludir las consecuencias jurídicas o de sus acciones.*

La ocultación o el encubrimiento de la naturaleza, el origen, la ubicación, el destino, el movimiento o la propiedad de bienes, o de derechos relativos a tales bienes, a sabiendas de que proceden de un delito o delitos, o de un acto de participación en tal delito o delitos

La adquisición, posesión o utilización de bienes, a sabiendas, en el momento de recibirlos, de que tales bienes proceden de un delito o delitos, o de un acto de participación en tal delito o delitos."

Por su parte, el órgano intergubernamental denominado Grupo de Acción Financiera sobre el Blanqueo de Capitales y Financiamiento al Terrorismo (GAFI, por

sus siglas en inglés), define el lavado de dinero como: "*procesamiento de ganancias derivadas de la actividad criminal para disfrazar su procedencia ilícita, permitiendo a los criminales gozar de ellas sin arriesgar su fuente*" (Gamboa, 2013).

Basado en esta definición, podemos ejemplificar el ciclo de lavado de dinero o de activos en el siguiente diagrama.

Ilustración 1- Ciclo de vida del lavado de dinero

Fuente: Secretaría de Hacienda y Crédito Público, (SHCP, 2015)

1.2.1.2 Antecedentes del Lavado de Dinero

Uno de los efectos que ha sido cuestionado a la globalización económica, es que de forma simultánea a las actividades ilícitas vinculadas a los procesos productivos internacionales, se propician las condiciones para el crecimiento de la consolidación de organizaciones criminales dedicadas a actividades como el tráfico de drogas, el comercio ilegal de armas, la inmigración clandestina, la pornografía infantil y los fraudes financieros, entre otro ilícitos (González Rodríguez, 2009).

El inicio del fenómeno del lavado de dinero, generalmente se ubica en los años sesentas, a la par del desarrollo e incremento lucrativo de los mercados masivos de droga; sin embargo, autores como Córdova y Palencia (Córdoba Gutiérrez & Palencia Escalante, 2001) señalan lo siguiente:

"… los primeros capitales blanqueados se dieron en Estados Unidos, en la época de los gánsteres y de la llamada Ley Seca. Para ese entonces, se dice que Chicago, en la década de 1920, un grupo de delincuentes con negocios en el alcohol, el juego, la prostitución y otras actividades ilícitas, compraron una cadena de lavanderías. Al final de cada día, juntaban las ganancias ilícitas provenientes de los otros negocios, quedando en conjunto justificadas como obtenidas de actividades legales"

Así el origen del término "lavado de dinero", que es relativamente reciente, se remonta a la época del mafioso norteamericano Meyer Lanski, bien conocido en el tiempo de la prohibición. Este delincuente, por aquel entonces creó en Nueva York una cadena de "lavanderos" que servían para blanquear los fondos provenientes de la explotación de casinos ilegales. Bastaba con poner cantidades importantes en efectivo, que recogía de los casinos, dentro de las cajas de su cadena de lavanderías; para que esos fondos ingresaran al círculo bancario.

Claudia Gamboa (Gamboa, 2013) señala, que se han practicado ciertas formas de lavado de dinero desde que surgió la necesidad de ocultar la índole o la existencia de ciertas transferencias financieras por razones ya sean políticas, comerciales o jurídicas.

El lavado de dinero es uno de los delitos más graves de la criminalidad organizada contemporánea, su evolución en el derecho internacional y en los marcos legales de los estados, demuestra con suficiencia que se trata de una práctica que ha marcado sus propias tendencias en la sociedad actual según lo comenta José González (González Rodríguez, 2009).

Las actividades ligadas al lavado de dinero representan riesgos en diversos aspectos, además de las obvias lesiones al tejido social, esta actividad puede

afectar el sistema económico en la medida en que debilita la integridad de los mercados financieros, pudiendo generar el riesgo de disminuir el control de la política económica, contribuyendo a introducir distorsiones e inestabilidad en los mercados, propiciando la pérdida de ingresos fiscales y representando un riesgo para las instituciones financieras y la economía en su conjunto.

1.2.1.3 Antecedentes Legales

La separación de actividades que tipifican que recursos pueden ser considerados de procedencia ilícita, tiene su origen en la Ley del Secreto Bancario (The Bank Secrecy Act) de 1970, que impuso a las instituciones financieras de Estados Unidos obligaciones de mantener constancia de determinadas operaciones y de reportarlas a las autoridades.

Desde sus inicios, ese sistema de reportes financieros implementado por la Ley del Secreto Bancario resultó ser un instrumento ineficaz para luchar de forma efectiva contra el lavado de dinero ya que esa ley únicamente estableció la obligación de reportar las posibles operaciones ilícitas, de forma que los posibles lavadores de dinero podían seguir ejerciendo sus actividades sin posibilidad de una sanción. Derivado de este contexto, el Congreso de los Estados Unidos, expidió la denominada "Ley de Control de Lavado de Dinero" en 1986, que tipifico el delito de lavado de dinero, sancionándolo con una pena de hasta 20 años. Esta ley, que al mismo tiempo federalizo tales actividades, autorizó la confiscación de ganancias obtenidas por los lavadores y proporcionó a las autoridades federales herramientas adicionales para investigar el lavado de dinero (Córdoba Gutiérrez & Palencia Escalante, 2001).

Asumiendo que el primer antecedente normativo sobre el tema se encuentra en la Ley de Secreto Bancario de 1970 y la Ley de Control de Lavado de Dinero de 1986, es posible señalar que la internacionalización de este ilícito ha sido rápida. Así, la comunidad internacional ha reaccionado con eficacia por lo menos en cuanto hace a la regulación legal (González Rodríguez, 2009).

1.2.1.4 Instrumentos Internacionales

A partir de finales de la década de 1980, la comunidad internacional a través de diversas instituciones, ha venido desarrollando un marco normativo orientado a prevenir la utilización del sistema financiero para el blanqueo del dinero proveniente de las actividades ilícitas del crimen organizado, dando origen a los siguientes instrumentos multilaterales:

- La Declaración de Basilea, (Basilea, 1988).
- La Convención de las Naciones Unidas contra el Tráfico Ilícito de Estupefacientes y Sustancias Sicotrópicas, -Convención de Viena- (ONU, 1988).
- El Informe del Grupo de Acción Financiera (GAFI, 1989).
- La Convención en Lavado, Registro, Embargo y Confiscación de los Productos del Crimen (CICAD, 1990).
- El Tratado de la Comunidad Económica Europea que provee las bases del Consejo Directivo de Prevención del Uso del Sistema Financiero con propósitos de Lavado de Dinero (Gibson García, 2009).
- El Plan de Acción de Buenos Aires (UNDP, 1995).
- La Declaración Política y el Plan de Acción contra el Lavado de Dinero de la Sesión Especial de la Asamblea General de las Naciones Unidas sobre el Problema Mundial de Drogas y los Principios de Wolfsberg (Schott, 2007).

Como lo menciona José González (González Rodríguez, 2009), se han constituido diversos grupos que tienen como finalidad establecer mecanismos de cooperación enfocados al combate del lavado de dinero, entre tales instancias se encuentran: el Grupo de Acción Financiera (GAFI); el Grupo Egmont; el Grupo de Acción Financiera del Caribe (CFATF, por sus siglas en inglés); el Comité de Expertos para la Evaluación de Medidas Contra el Lavado de Dinero del Consejo de Europa (PC-R-EV Commite, por sus siglas en inglés); el Grupo Contra el Lavado de Dinero del Este y Sur de África (ESAAMLG, por sus siglas en inglés); el Grupo de Asia Pacífico Contra el Lavado de Dinero (APGML, por sus siglas en inglés) y el Grupo de Acción

Financiera de América del Sur (GAFISUD), entre otras instancias internacionales que persiguen el mismo fin.

1.2.1.5 Actividades Vulnerables en México

Las actividades vulnerables en México, son todas aquellas que por su tipo o índole están tipificadas dentro de la ley LFPIORPI (Gobierno de México, 2012) como actividades a ser evaluadas y monitoreadas para prevenir el lavado de dinero y el financiamiento al terrorismo, estas se encuentran listadas en la siguiente ilustración (Ilustración 2).

Ilustración 2- Actividades vulnerables tipificadas en la ley LFPIORPI

Fuente: Secretaría de Hacienda y Crédito Público (SHCP, 2015)

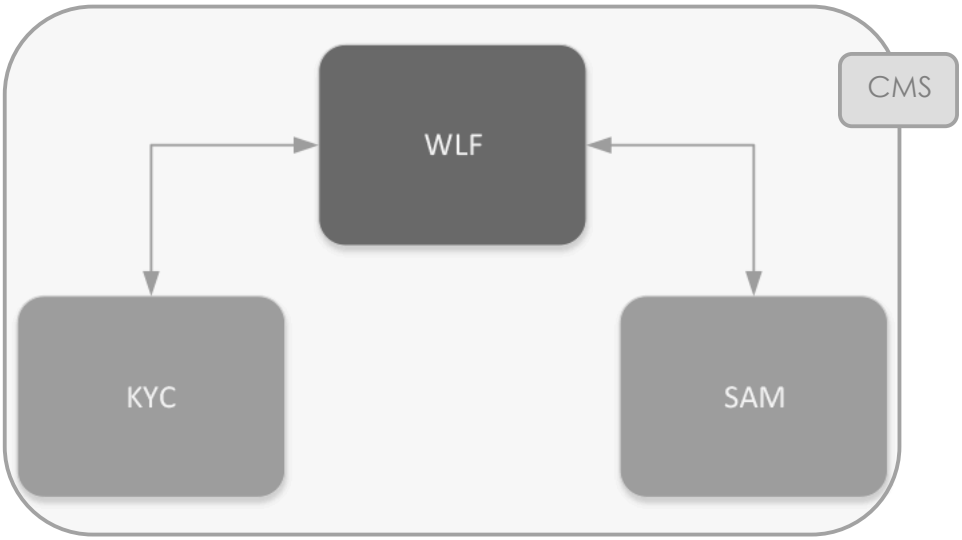
1.2.1.6 Segmentación de Actividades para el Combate al Lavado de Dinero

Una vez revisadas las medidas preventivas para el combate al lavado de dinero, podemos identificar los diferentes rubros a los que la detección y monitoreo de actividades sospechosas se refiere, listadas a continuación:

- Identificación del cliente y debida diligencia (KYC/CDD, Know Your Customer and Customer Due Diligence, por sus siglas en inglés)

- Monitoreo de actividades sospechosas (SAM, Suspicious Activity Monitoring, por sus siglas en inglés)
- Revisión de listas negras (WLF, Watchlist Filtering, por sus siglas en inglés)
- Servicio de Manejo de Casos (CMS, Case Management System, por sus siglas en inglés)

Ilustración 3 - Principales actividades para la revisión de actividades ilícitas y la prevención de lavado de dinero y el combate al terrorismo



Fuente: Elaboración propia

En la Ilustración 3 podemos ver como la actividad de revisión de nombres está ligada tanto al inicio de la captura de un nuevo cliente o actualización de sus datos, así como en el monitoreo de transacciones recurrentes y actividades que pudiesen llegar a ser sospechosas en caso de que incumplan con algún mandato estipulado en la Ley LFPIORPI (Gobierno de México, 2012).

El servicio de manejo de casos (CMS, por sus siglas en inglés), no es otra cosa más que una herramienta de software en el cual se pueda concentrar todos aquellos casos o alertas que hayan sido detectadas en el proceso de monitoreo de actividades sospechosas, y que puedan ser susceptibles a investigación por parte de las autoridades bancarias. Es de saber que no contar con un sistema de estas características se traduciría en tiempo invertido para hacer funcionar los 3 elementos claves en la lucha contra el lavado de dinero, ya que hacer seguimiento

de una actividad no sería tan transparente, y faltaría una parte importante dentro de este tipo de sistemas, la ponderación general de las alertas que se hayan levantado en los diferentes rubros para un mismo cliente.

1.2.1.7 Revisión de Listas Negras

Cada persona o entidad financiera que realice algunas de las actividades vulnerables (Ilustración 2) debe por ley monitorear todo de tipo de transacciones y transferencias haciendo uso de los estatutos estipulados en la Ley LFPIORPI (Ley Federal para la Prevención e Identificación de Operaciones con Recursos de Procedencia Ilícita), además, debe por mandato revisar a la persona o empresa que está realizando este tipo de movimientos financieros, todo esto haciendo una búsqueda de su nombre contra la lista negra de la SHCP, basada en la lista de negra de OFAC.

Las búsquedas de nombres pueden arrojar distintos resultados, generar alertas o inclusive llegar a detener una transacción dependiendo del tipo de resultado que arroje la búsqueda. Estas búsquedas se hacen sobre listas que a su vez tienen que estar actualizadas si no diario, si de manera muy regular.

Para poder hacer frente al creciente uso de la tecnología y medios digitales para realizar transacciones o transferencias electrónicas es necesario contar con software especializado que ayude con las búsquedas de nombres entre un volumen elevado de transacciones (para el caso de una institución bancaria) al día. Software que esté basado en algoritmos que permitan buscar si no de forma precisa, si aproximada y permitan hacer una pre-selección de posibles actividades ilícitas.

1.2.1.8 Búsqueda de Nombres

El proceso de coincidencia de nombres, es la comparación de dos conjuntos de datos para identificar entradas similares entre dos listas, si no exactas, al menos aproximadas y que permitan poner atención solo a las entradas que resulten relevantes o que tengan alguna relación entre sí.

Los algoritmos deben de ser capaces de detectar coincidencias entre dos listas, algunos de ellos fonéticas o por medio de comparación de distancias entre caracteres. Tanto los algoritmos fonéticos como de distancia entre caracteres son ampliamente usados hoy en día en conjunto con algoritmos de lógica difusa para tener un mejor rendimiento y precisión a la hora de realizar dichas búsquedas (Ramachandran, 2014).

1.2.2 El Lavado de Dinero, sus Leyes, sus Sanciones y sus Números

1.2.2.1 Cronología de la Lucha Contra el Lavado de Dinero

La lucha contra el lavado de dinero comenzó en los años 70's con la formulación de la ley del Secreto Bancario (BSA, Bank Secrecy Act, por sus siglas en inglés). Desde entonces las regulaciones y leyes han ido evolucionando para hacerle frente a este problema (ver Tabla 1).

Tabla 1 - Evolución de las leyes en el mundo para combatir el lavado de dinero y el financiamiento al terrorismo.

Año	Ley o Grupo	Definición de Responsabilidades	Contenido o Comunicados Importantes
1970	Ley del Secreto Bancario	El principal estatuto regulatorio de Estados Unidos para combatir el lavado de dinero, promulgada en 1970 y mejorada por la ley USA PATRIOT en 2001	Le requiere a los bancos monitorear transacciones en efectivo, o reportes de transacciones por mas de $10,000 dólares, y realizar un reporte de actividades sospechosas así como mantener un historial de varias transacciones financieras
1974	Creación del comité de Basilea para la supervisión bancaria	Establecida por los gobernadores de bancos centrales del G10. Incrementa los estándares de supervisión a nivel mundial	Debida diligencia para la papelería bancaria (2001). Intercambio de registros financieros entre distintas jurisdicciones para el combate al financiamiento al terrorismo (2002). Guía general para la apertura de cuentas e identificación del cliente (2003). guía de manejo de riesgo del conocimiento al cliente mejorada (2004).
1986	Ley para Control de Lavado de Dinero	Primera ley en el mundo en hacer del lavado de dinero un crimen	Esta ley criminaliza el lavado de dinero, introdujo confiscación penal y civil por violaciones a la ley BSA.
1989	Grupo de Acción Financiera	Un cuerpo intergubernamental con 34 países miembros y 2 organizaciones internacionales establecidas por el G7 para desarrollar políticas contra el lavado de dinero y el financiamiento al terrorismo	Estipula 40 recomendaciones contra el lavado de dinero y el financiamiento al terrorismo
1991	Unión Europea	Las directivas de la Unión Europea para combate al lavado de dinero requieren que los miembros de esta unión tomen medidas para la prevención de estos delitos.	1ra directiva de la UE para la prevención del uso de los sistemas financieros para el propósito del lavado de dinero (1991). 2da directiva (2001). 3ra directiva (2005).
2001	Ley USA PATRIOT	Estipulada el 26 de Octubre del 2001, trajo consigo mas de 50 mejoras a la Ley BSA.	Esta ley subió significativamente la apuesta y la carga administrativa regulatoria de las instituciones de Estados Unidos , y ha servido como motor de la regulación del lavado de dinero para otros países.
2002	Grupo Wolfberg	Asociación de 11 bancos a nivel mundial. Dedicado a desarrollar estándares para los controles del lavado de dinero en las instituciones financieras.	Principios del anti-lavado de dinero de Wolfberg para la banca privada (2012) La guía de supresión del financiamiento al terrorismo (2002) Principios de lucha contra el lavado de dinero en la banca corresponsal (2014)
2004	Grupo Egmont	Red informal de unidades de inteligencia financiera	Estatuto de propósito (2004). Principio para el intercambio de información entre unidades de inteligencia financiera para casos de lavado de dinero (2013). Mejores practicas para el intercambio de información entre unidades de inteligencia financiera (2004).
2012	Ley LFPIORPI	Ley Federal para la Prevención e Identificación de Operaciones con Recursos de Procedencia Ilícita. Promulgada por el entonces presidente de México Felipe Calderón para el combate al lavado de dinero y financiamiento al terrorismo.	Contempla regulaciones para la Alta de nuevos clientes en instituciones financieras o que su rubro se encuentre dentro de las estipuladas como actividades vulnerables Identificación de clientes y usuarios ya dados de alta antes de la promulgación de la Ley Presentación de avisos e informes en tiempo oportuno al SAT acerca de actividades sospechosas.

Fuente: Elaboración Propia

1.2.2.2 Normativa en México para el Combate al Lavado de Dinero

El 17 de octubre de 2012 se publicó en el Diario Oficial de la Federación la Ley Federal para la Prevención e Identificación de Operaciones con Recursos de Procedencia Ilícita (LFPIORPI). La Ley tiene por objeto proteger al sistema financiero y la economía nacional, estableciendo medidas y procedimientos para prevenir y detectar actos u operaciones que involucren recursos de procedencia ilícita, de acuerdo con la Ley, diversas actividades no financieras son consideradas vulnerables (SHCP, 2015).

1.2.2.3 Estimación de Costo Económico del Lavado de Dinero en México

De acuerdo con un estudio realizado por Pedroza, (Pedrosa Leyva, 2013) las instituciones del sistema financiero mexicano están obligadas a reportar a la UIF (Unidad de Inteligencia Financiera de la SHCP) cualquier operación que detecten o realicen y que se ubique en alguno de los siguientes supuestos:

- *Operaciones relevantes:* cuando una transacción es superior a los 10 mil dólares.
- *Operaciones inusuales:* si no coincide con el patrón habitual de comportamiento transaccional del cliente.
- *Operaciones preocupantes:* en la que interviene un representante de la institución financiera y pudiera contravenir cualquier disposición legal.

Pedroza menciona en su artículo que en base a un modelo de estudio hecho por Argentiero *et al.* (Pedrosa Leyva, 2013), se estimó la cantidad de recursos monetarios producidos por actividades ilegales que son potencialmente objeto de lavado de dinero en México. En él se asume la existencia de dos sectores (formal e informal), donde actúan tres agentes (empresas, hogares y gobierno).

De acuerdo a los resultados del modelo, el valor estimado promedio de lavado de dinero en México entre el segundo trimestre de 1993 y el 2009 equivalió a 1.688% del producto interno bruto (PIB) como se muestra en la siguiente la Tabla 2.

Tabla 2- Frecuencia de lavado de dinero nacional anual - México 1993-2008

Periodo	Lavado de Dinero (pesos)	Lavado de Dinero (% PIB)
1993	$ 33,497,223,489.54	0.56%
1994	$ 32,936,322,690.71	0.52%
1995	$ 32,918,308,522.76	0.56%
1996	$ 51,575,307,120.26	0.82%
1997	$ 88,708,758,911.61	1.32%
1998	$ 114,867,920,322.12	1.67%
1999	$ 189,066,676,321.15	2.61%
2000	$ 159,273,323,733.36	2.10%
2001	$ 193,326,506,986.16	2.58%
2002	$ 181,576,440,055.55	2.38%
2003	$ 137,300,949,218.71	1.77%
2004	$ 137,694,782,652.44	1.69%
2005	$ 163,857,627,633.37	1.95%
2006	$ 147,034,644,137.50	1.68%
2007	$ 160,083,919,910.03	1.76%
2008	$ 145,687,086,968.02	1.63%

Fuente: Lavado de dinero en México. Estimación de su magnitud y análisis de su combate a través de la inteligencia financiera (Pedrosa Leyva, 2013).

En la Tabla 3 podemos ver la cantidad de operaciones inusuales que fueron reportadas al SAT por entidad federativa del 2004 al 2008. Así mismo en la Tabla 4 podemos ver qué porcentaje del PIB equivalen las cantidades antes mencionadas en la Tabla 2 por entidad federativa.

Tabla 3 - Reporte de operaciones inusuales por Entidad Federativa

	2004	2005	2006	2007	2008
Aguascalientes	38,640	63,902	67,051	70,929	69,174
Baja California	139,542	227,909	233,294	251,880	263,368
Baja California Sur	22,777	44,650	54,573	58,650	56,300
Campeche	21,462	33,967	37,677	33,620	38,489
Coahuila	74,043	116,768	148,071	171,696	178,681
Colima	18,675	31,499	39,481	39,172	43,759
Chiapas	47,242	87,138	103,775	120,154	132,877
Chihuahua	96,667	160,527	177,571	192,503	184,942
Distrito Federal	668,504	1,086,588	1,161,370	1,180,837	969,256
Durango	33,305	53,757	60,772	62,367	66,188
Guanajuato	139,536	214,221	243,751	255,905	268,369
Guerrero	80,779	127,802	119,009	125,828	134,107
Hidalgo	57,572	86,281	88,815	90,776	95,537
Jalisco	222,732	376,355	411,673	445,692	494,387
México	158,113	249,248	273,409	300,436	317,539
Michoacán	113,545	184,471	202,881	214,187	225,236
Morelos	36,472	51,025	61,481	76,697	74,011
Nayarit	19,075	33,151	37,342	44,967	49,076
Nuevo León	180,768	304,326	380,138	437,774	459,553
Oaxaca	70,607	113,761	119,469	127,537	137,284
Puebla	121,126	187,388	204,112	209,241	219,665
Querétaro	44,252	75,219	84,831	89,352	91,492
Quintana Roo	77,378	161,210	132,694	138,926	141,731
San Luis Potosí	61,955	97,990	108,636	106,162	105,743
Sinaloa	77,574	122,142	156,387	200,087	232,213
Sonora	71,464	141,433	139,399	147,369	137,101
Tabasco	52,903	90,351	102,590	104,081	121,911
Tamaulipas	114,246	164,013	178,023	174,872	185,743
Tlaxcala	15,239	24,341	24,218	23,575	27,417
Veracruz	144,303	228,342	277,250	313,467	334,127
Yucatán	55,759	93,920	109,533	114,037	124,613
Zacatecas	32,985	48,213	54,058	48,955	52,614
Total	**3,109,239**	**5,081,907**	**5,593,334**	**5,971,731**	**6,032,504**

Fuente: Lavado de dinero en México. Reporte de operaciones inusuales en comparación con el PIB (INEGI, 2015).

Tabla 4- Lavado de Dinero por Entidad Federativa en relación al PIB.

Entidad Federativa	2002	2003	2004	2005	2006	2007	2008
Aguascalientes	0.002%	0.003%	0.003%	0.005%	0.006%	0.011%	0.012%
Baja California	0.085%	0.078%	0.120%	0.127%	0.125%	0.164%	0.221%
Baja California Sur	0.028%	0.023%	0.017%	0.026%	0.025%	0.020%	0.013%
Campeche	0.010%	0.014%	0.016%	0.018%	0.011%	0.011%	0.004%
Coahuila	0.050%	0.029%	0.030%	0.035%	0.031%	0.037%	0.025%
Colima	0.037%	0.040%	0.038%	0.034%	0.033%	0.023%	0.009%
Chiapas	0.108%	0.090%	0.059%	0.061%	0.048%	0.045%	0.060%
Chihuahua	0.047%	0.033%	0.034%	0.044%	0.054%	0.051%	0.037%
Distrito Federal	0.216%	0.143%	0.103%	0.115%	0.082%	0.170%	0.349%
Durango	0.048%	0.042%	0.023%	0.040%	0.029%	0.024%	0.010%
Guanajuato	0.070%	0.042%	0.050%	0.061%	0.053%	0.074%	0.081%
Guerrero	0.122%	0.085%	0.078%	0.054%	0.037%	0.021%	0.011%
Hidalgo	0.037%	0.031%	0.033%	0.061%	0.044%	0.022%	0.008%
Jalisco	0.145%	0.119%	0.126%	0.138%	0.133%	0.135%	0.100%
México	0.168%	0.109%	0.129%	0.150%	0.157%	0.148%	0.059%
Michoacán	0.052%	0.048%	0.045%	0.061%	0.062%	0.086%	0.081%
Morelos	0.083%	0.052%	0.030%	0.029%	0.019%	0.013%	0.017%
Nayarit	0.035%	0.032%	0.030%	0.034%	0.034%	0.041%	0.021%
Nuevo León	0.042%	0.023%	0.025%	0.025%	0.025%	0.035%	0.045%
Oaxaca	0.173%	0.102%	0.079%	0.080%	0.046%	0.025%	0.014%
Puebla	0.166%	0.104%	0.091%	0.108%	0.082%	0.063%	0.032%
Querétaro	0.033%	0.031%	0.030%	0.049%	0.043%	0.038%	0.020%
Quintana Roo	0.047%	0.050%	0.064%	0.058%	0.039%	0.019%	0.005%
San Luis Potosí	0.097%	0.081%	0.071%	0.076%	0.053%	0.041%	0.018%
Sinaloa	0.057%	0.040%	0.052%	0.057%	0.057%	0.074%	0.056%
Sonora	0.078%	0.064%	0.055%	0.088%	0.077%	0.126%	0.111%
Tabasco	0.018%	0.015%	0.018%	0.020%	0.016%	0.024%	0.022%
Tamaulipas	0.132%	0.087%	0.084%	0.075%	0.059%	0.077%	0.116%
Tlaxcala	0.038%	0.023%	0.021%	0.021%	0.017%	0.011%	0.005%
Veracruz	0.083%	0.056%	0.060%	0.092%	0.091%	0.073%	0.035%
Yucatán	0.047%	0.045%	0.056%	0.054%	0.039%	0.025%	0.016%
Zacatecas	0.023%	0.033%	0.025%	0.051%	0.051%	0.035%	0.015%
TOTAL	**2.376%**	**1.765%**	**1.694%**	**1.946%**	**1.679%**	**1.762%**	**1.630%**

Fuente: Lavado de dinero en México. Reporte de operaciones inusuales en comparación con el PIB (INEGI, 2015).

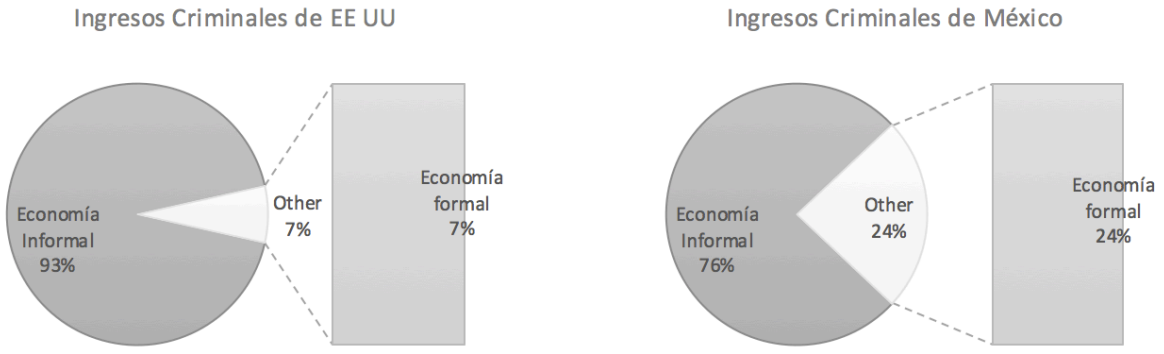
La Secretaría de Hacienda y Crédito Público identifica, por igual, que en el sistema financiero mexicano se registra un excedente de 10 mil millones de dólares al cierre del año fiscal, que presuntamente provienen de actividades ilícitas. Así mismo el Centro Nacional de Inteligencia sobre Narcótico del Departamento de Justicia de

los Estados Unidos de América (NDIC, por sus siglas en inglés) dice desconocer la cantidad exacta de dinero que abandona el país, aunque estima que cerca de 39 mil millones de dólares se lavan fuera de sus fronteras, acción que desarrollan sobre todo organizaciones criminales de Colombia y México, siendo México el país que lava en más volumen (CESOP, 2012).

No Money Laundering, una ONG estadounidense, estima que en México los cárteles del narcotráfico obtienen ganancias de alrededor de 5% del producto interno bruto (PIB), cifra que ascendería a poco más de 59 mil 500 millones de dólares, si se considera el PIB nominal del segundo trimestre de 2015 y del tipo de cambio al cierre de junio del mismo año (INEGI, 2015).

Tomando en consideración las notas de Scheider y Enste, en su publicación, La Sombra de la Economía y la Corrupción Alrededor del Mundo, Nuevos Estimados para 145 Países (*Shadow Economies and Corruption All Over the World: New Estimates for 145 countries*), se precisa la distribución de ganancias criminales según el tipo de economía en Estados Unidos y México (CESOP, 2012).

Ilustración 4 - Distribución de ganancias criminales por tipo de economía en Estados Unidos y México.

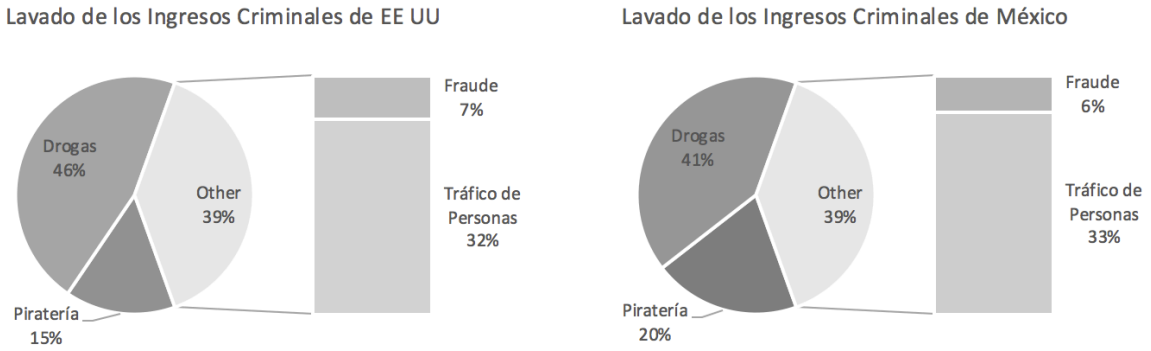


Fuente: Lavado de Dinero. Indicadores y Acciones Binacionales (CESOP, 2012).

La Integridad Financiera Global y la Universidad de Columbia en Nueva York estiman, a partir de publicaciones del Instituto de Economía Internacional en

Washington D.C., la distribución del blanqueo de dinero según la actividad criminal en Estados Unidos y México (CESOP, 2012).

Ilustración 5- Distribución de blanqueo de dinero según actividad criminal en Estados Unidos y México.



Fuente: Lavado de Dinero. Indicadores y Acciones Binacionales (CESOP, 2012).

Los resultados contra el combate al lavado de dinero que se presentaron en el Quinto Informe de Gobierno presentado en el 2011 por el entonces presidente de México, Felipe Calderón, se muestran en la Tabla 5.

Tabla 5 - Resultados del combate a las operaciones con recursos de procedencia ilícita (Lavado de Dinero), 2007-2011.

Concepto	Datos anuales				Enero-julio		
	2007	2008	2009	2010	2010	2011 p/	Variación % anual
Dinero asegurado							
Pesos mexicanos (Miles)	11,425.60	28,394.80	48,113.50	19,492.50	2,612.26	10,728.37	310.7
Dólares americanos (Miles)	17,491.20	71,641.30	56,122.10	24,662.17	22,030.71	7,034.66	-68.1
Averiguaciones previas iniciadas	**199**	**276**	**245**	**305**	**180**	**173**	**-3.9**
Averiguaciones previas despachadas	**160**	**210**	**253**	**254**	**138**	**120**	**-13.0**
Averiguaciones previas consignadas	54	67	50	83	46	34	-26.1
Incompetencias	25	39	94	55	37	26	-29.7
No ejercicio de la acción penal	22	14	17	17	8	22	175.0
Reservas	35	39	56	52	29	16	-44.8
Acumulaciones	24	51	36	47	18	22	22.2
Órdenes de aprehensión libradas	**48**	**29**	**27**	**60**	**35**	**15**	**-57.1**
Procesos penales iniciados	**34**	**62**	**62**	**43**	**35**	**32**	**-8.6**
Número de personas contra las que se ejercitó acción penal	**59**	**84**	**220**	**252**	**120**	**81**	**-32.5**
Sentencias condenatorias	**4**	**28**	**19**	**13**	**8**	**11**	**37.5**
Total de detenidos	**131**	**128**	**152**	**114**	**78**	**71**	**-9.0**
Organizaciones delictivas desarticuladas	**2**	**0**	**10**	**6**	**3**	**0**	**n.a.**

p/ Cifras preliminares.
n.a. No aplicable.

Fuente: Procuraduría General de la República. Unidad Especializada en Investigación de Operaciones con Recursos de Procedencia Ilícita y de Falsificación o Alteración de la Moneda. (Calderón Hinojosa, 2011)

Así mismo los últimos datos mostrados en el Tercer Informe de Gobierno del Presidente en turno, Enrique Peña Nieto muestran los siguientes resultados en el combate al lavado de dinero, véase la Tabla 6.

Tabla 6- Resultados del combate a las operaciones con recursos de procedencia ilícita (Lavado de Dinero), 2015.

Concepto	Septiembre 2014-julio 2015 p/
Dinero asegurado 2/	
Pesos mexicanos (Millones)	167.6
Dólares americanos (Millones)	14
Averiguaciones previas iniciadas	202
Averiguaciones previas despachadas	135
Averiguaciones previas consignadas	56
Incompetencias	17
No ejercicio de la acción penal	24
Reservas	16
Acumulaciones	21
Órdenes de aprehensión libradas	35
Procesos penales iniciados	15
Número de personas contra las que se ejercitó acción penal	189
Sentencias condenatorias	14
Total de detenidos	15
Organizaciones delictivas desarticuladas	1

1/ Resultados de la Unidad Especializada en Investigación de Operaciones con Recursos de Procedencia Ilícita, Falsificación o Alteración de Moneda de la Subprocuraduría Especializada en Investigación de Delincuencia Organizada.

2/ Total de dinero asegurado en efectivo y cuentas bancarias.

p/ Cifras preliminares.

Fuente: Procuraduría General de la República. (Peña Nieto, 2015)

La Unidad de Inteligencia Financiera, de la SHCP recibió 13.7 millones de reportes de operaciones y avisos de sujetos obligados de los sectores financiero y no financiero entre septiembre de 2014 y julio 2015, con la finalidad de detectar y denunciar operaciones con recursos de procedencia ilícita y financiamiento al terrorismo. Al respecto destaca (Peña Nieto, 2015):

- Formuló 69 denuncias ante la PGR (Procuraduría General de la República), que involucran a 559 sujetos por probable comisión del delito de operaciones con recursos de procedencia ilícita, logrando el aseguramiento

de 260 millones de pesos y 115.1 miles de dólares de los Estados Unidos de América.

- Se incluyeron 798 personas en la Lista de Personas Bloqueadas, ente las cuales 669 son internacionales y 129 son nacionales, en relación a estos últimos, se han bloqueado saldos por 279.4 millones de pesos y 4.9 millones de dólares.

1.3 Planteamiento del Problema

Los algoritmos utilizados hoy en día para llevar a cabo una búsqueda de coincidencia de nombres, son algoritmos fonéticos (basados en la pronunciación de las palabras) y algoritmos de cálculo de distancia (también conocidos como reconocimiento de patrones), existen diferencias en sus usos y aplicaciones, y también adaptaciones de estos algoritmos según su utilización y contexto, algunas de las aplicaciones para estos algoritmos son para su utilización en la búsqueda de nombres dentro de las redes sociales, para hacer emparejamiento de individuos, reconocimiento geográfico de personas en base a la pronunciación de su nombre y búsqueda genealógica de familias y apellidos, véase (Lait & Randell, 1995) para futuras referencias.

En lo que respecta al lavado de dinero y el combate al terrorismo, ha habido ajustes hacia los algoritmos más usados, principalmente los de tipo fonético, al estar relacionado con tareas aduanales, en donde el reconocimiento de una persona por la forma en la que se pronuncia su nombre es importante para el monitoreo de entradas a un país. El algoritmo más utilizado para estos casos es el algoritmo soundex (David et al., 2012).

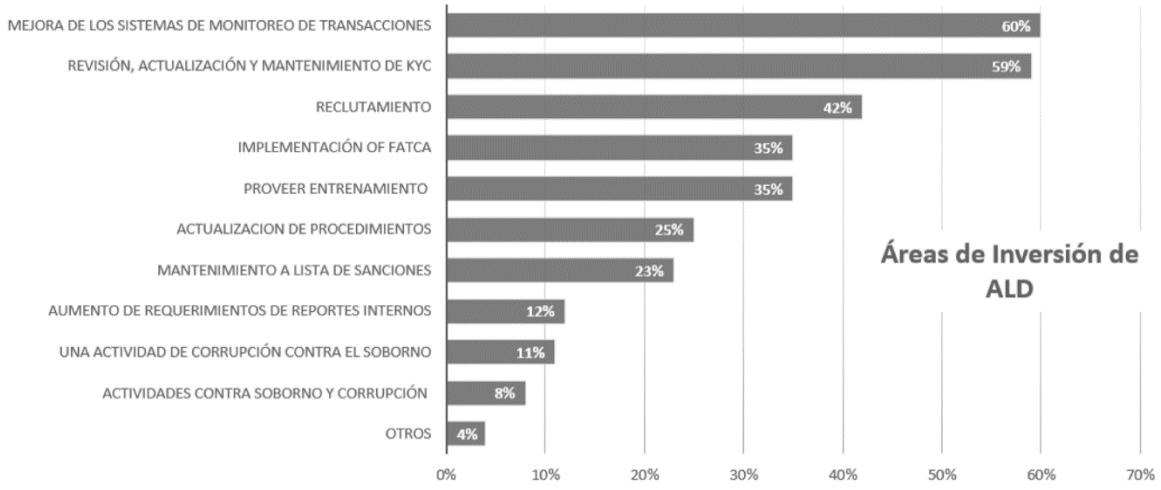
El algoritmo (soundex) junto con otros 2 algoritmos adaptados para la búsqueda de nombres como lo son el metaphone (Ramachandran, 2014) y el NYSIIS (Sistema de Identificación e Inteligencia del Estado de Nueva York, por sus siglas en inglés), son algoritmos que utilizan el lenguaje inglés para el reconocimiento fonético. Si

bien en antiguos estudios se ha probado su precisión, no se cuentan con suficientes referencias para decir que sean algoritmos que funcionen de la misma manera con nombres en lenguaje en español, es decir, que las reglas gramaticales que se encuentran de manera intrínsecas programadas en estos algoritmos, podrían no encajar a la perfección con el uso de caracteres especiales y pronunciaciones especificas del lenguaje en español, como lo son la letra "ñ" o la omisión de la pronunciación de la letra "h" al inicio de una palabra.

El problema se puede definir a partir de las siguientes afirmaciones:

1. De acuerdo con datos provenientes del Comité de Seguridad Nacional de los Estados Unidos de América, un sistema actual para la detección de personas que quieren ingresar al país y que se encuentran en listas negras, tiene una tasa efectiva de un 56% de precisión (Select Committee on Homeland Security, 2004). Según estima PwC (Price Waterhouse Cooper por sus siglas en inglés), esta tasa de falsos positivos es más alta que la reportada por el Comité de Seguridad Nacional, llegando hasta una tasa de entre 90% y 95% de falsos positivos. La mayoría de estos casos debido a una incorrecta configuración de los sistemas de detección con los que cuentan las instituciones financieras o comerciales que hacen estas revisiones (PWC, 2010).
2. Acorde a una encuesta publicada por KPMG en el 2014, las principales áreas en las que las instituciones financieras y comercios están invirtiendo para combatir el lavado de dinero son (KPMG, 2014):
    - Mejora de los sistemas de monitoreo de transacciones
    - Revisión, actualización y mantenimiento de sistemas de Conocimiento del Cliente (KYC)
    - Reclutamiento de personal capacitado en el área de prevención de crimen financiero

Ilustración 6 - Resultado de la encuesta realizada por KPMG con datos recopilados en el 2014



Fuente: Global Antimoney Laundering Survey (KPMG, 2014)

3. Existen hoy en día estudios de algoritmos fonéticos adaptados a distintos lenguajes como lo son:
    - Algoritmo Daitch-Mokotoff Soundex, adaptado para nombres de Europa del Este.
    - Soundex, Metaphone, Double Metaphone, adaptado para el lenguaje inglés.
    - Beider-Morse Phonetic, adaptado para nombres judíos (Beider & Morse, 2010)
    - Indian Soundex, adaptado para el lenguaje Indú (Shah, 2014)

En estas investigaciones, se comprueba la necesidad de tener algoritmos precisos y adaptados al lenguaje de la zona geográfica donde se hace el monitoreo, para poder reducir así los falsos positivos, que vimos anteriormente que generan re-trabajo y representan costos para las instituciones y provoca frustración en las personas que han sido detectadas falsamente como terroristas.

Es por esto que el presente documento se concentró en el primer rubro arrojado por la encuesta realizada por KPMG (véase Ilustración 6), el cual está relacionado a la mejora de los sistemas de monitoreo de transacciones.

Particularmente se abordó el estudio uno de los sistemas de monitoreo llamado *revisión de listas negras (WLF)*, utilizado para contrastar una lista de nombres de personas latinas contra la lista de OFAC.

Con la finalidad de poder revisar la efectividad de tres de los algoritmos más utilizados hoy en día para la búsqueda de nombres sobre este tipo de listas negras, los cuales son:

1. **Soundex**
2. **Metaphone**
3. **NYSIIS**

Se utilizó la lista de OFAC como referencia y no la lista avalada por la SHCP de México debido a que esta última no es de dominio público y se sabe que tiene sus orígenes en la de OFAC (Videgaray Caso, 2014).

En resumen, se analizarán los tres algoritmos fonéticos mencionados anteriormente (soundex, metaphone, NYSIIS) contra una lista de nombres hispanos, y se contrastó dicha lista con la lista negra de OFAC (SDN), la intención fue medir la efectividad y velocidad de ejecución de los algoritmos y revisar su comportamiento contra una lista de nombres que contienen aspectos ortográficos que no fueron considerados en los algoritmos al momento de su elaboración.

1.4 Justificación

1.4.1 Impacto de Investigación

Se sabe que no se cuenta con un algoritmo fonético que este perfectamente adaptado a la búsqueda de nombres en el lenguaje español, es por eso que medir la efectividad de los algoritmos mencionados con anterioridad tiene relevancia, esto para poder determinar cuál de ellos tiene un mejor desempeño y precisión en la búsqueda y comparación de los nombres contra la lista de OFAC.

Tales resultados sustentan el estudio, y permiten revisar coincidencias insertando nombres de personas dentro de los datos a evaluar, identificadas por cometer actos de lavado de dinero o financiamiento al terrorismo. Por último la investigación sienta bases para continuar con la investigación y construcción de un algoritmo fonético basado en la pronunciación de nombres en español para combatir el lavado de dinero y el combate al terrorismo.

1.4.2 Impacto Operativo

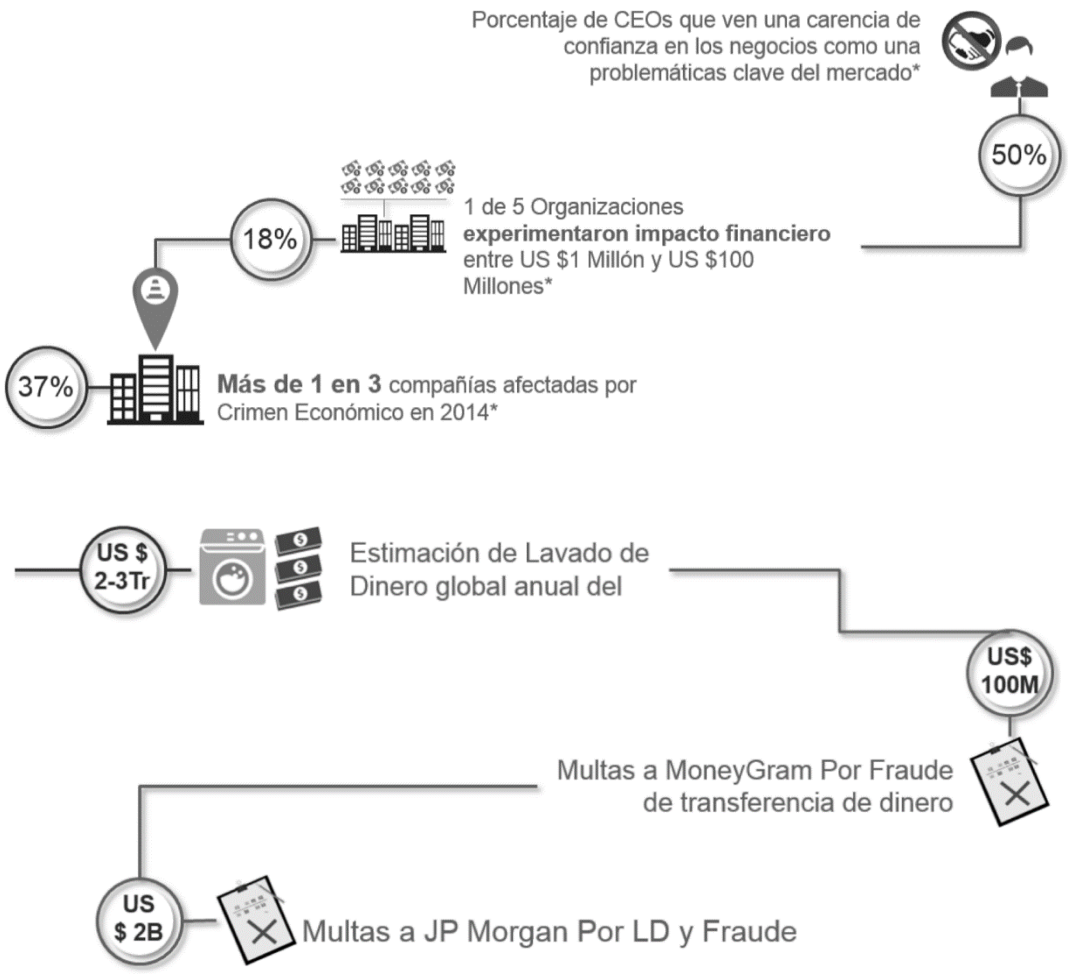
Considerando una tasa de falsos positivos de hasta un 50% (Ramachandran, 2014); significaría el uso de recursos humanos y económicos para poder procesar a las personas que están siendo identificadas de forma errónea, o en el caso de los bancos, en el bloqueo de transacciones o transferencias, lo cual sería una perdida directa financiera para la institución. Este caso es mejor ejemplificado, si consideramos que en el 2015 hubo una afluencia de pasajeros de 38 millones en el aeropuerto de la Ciudad de México (SCT, 2016), de los cuales 12 millones son pasajeros provenientes de otros países y que por normativa tienen que ser revisados contra listas negras, podemos suponer que alrededor de 5 millones de esos individuos podrían ser catalogados de forma errónea; esto se transformaría en tiempo que el individuo que viaja perdería en revisiones, tiempo que el personal de

revisión se toma para determinar que una persona es en efecto libre de alguna persecución legal, entre otros factores que se ven afectados por cada persona detectada en el escaneo de nombres.

1.4.3 Impacto Económico

La detección de falsos positivos como la admisión de falsos negativos suponen impactos de diferente índole económico, como se muestra en la Ilustración 7.

Ilustración 7 - Tomada de la encuesta de KPMG donde refleja los costos por gastos de incumplimiento



Fuente: The Cost of Compliance 2013, (KPMG Cost of Compliance, 2013)

Estas multas por concepto derivado del lavado de dinero, son algunas de las multas representativas a nivel internacional, pero dejan en claro la importancia de contar con mecanismos que apoyen en la detección de actividades sospechosas, que sean eficaces y que sustenten la inversión en la investigación del tema y el análisis más detallado de los algoritmos para la reducción, en este caso, de los falsos positivos.

1.5 Objetivos de la Investigación

A continuación, se describe el objetivo general de la investigación de tesis, así como los objetivos específicos que se pretenden alcanzar.

1.5.1 Objetivo General

Realizar un análisis comparativo de algoritmos de búsqueda de coincidencia de nombres por variación fonética contra la lista de OFAC para determinar su eficacia contra una lista de nombres en español.

1.5.2 Objetivos Específicos

- Elaborar pruebas unitarias de los algoritmos soundex, metaphone y NYSIIS.
- Realizar un análisis estadístico con los datos obtenidos de las pruebas unitarias.
- Determinar la eficiencia y eficacia de los tres algoritmos evaluados
- Determinar que algoritmo mostro el mejor desempeño comparándolo contra una lista de nombres en español.
- Identificar que algoritmo de entre los tres evaluados presenta un índice menor de falsos positivos después de su análisis.

- Especificar las razones por las cuales se optaría por seleccionar algún algoritmo de los probados para la búsqueda de nombres en español en una lista negra.

CAPÍTULO 2: MARCO TEÓRICO

2.1 Sanciones Administrativas en México

Derivado del compromiso adoptado por México en el ámbito internacional y como miembro del Grupo de Acción Financiera (GAFI, por sus siglas en inglés), las autoridades de México, por medio de la Secretaría de Hacienda y Crédito Público (SHCP), han establecido estándares aplicables a instituciones financieras y no financieras para el combare al lavado de dinero y financiamiento al terrorismo.

Estipulada en la Ley Federal para la Prevención e Identificación de Operaciones con Recursos de Procedencia Ilícita (Gobierno de México, 2012), publicada el 17 de octubre de 2012, establece las obligaciones para las actividades económicas consideradas como actividades vulnerables, ver Ilustración 2. Estas obligaciones en resumen son las siguientes (PWC, 2013):

- Identificar y conocer a los clientes y usuarios con los que se realicen actividades vulnerables
- Proteger y evitar la destrucción u ocultamiento de la información que sirva de soporte a la actividad vulnerable.
- Designar ante la SHCP a un representante encargado del cumplimiento de las obligaciones.
- Brindar las facilidades necesarias para que se lleven a cabo las visitas de verificaciones.
- Presentar los avisos en la SHCP en los tiempos y bajo la forma prevista en la ley.
- Abstenerse de realizar cualquier acto u operación tipificada como actividad vulnerable, cuando sus clientes o usuarios se nieguen a proporcionar información o documentación relacionada con su identificación y conocimiento del mismo.

Basado en las obligaciones antes descritas y en consecuencia, las sanciones que deriven en algún tipo de incumplimiento de ley pudiesen ser (SHCP, 2015):

Se impondrá una sanción de 200 y hasta 2,000 días de Salario Mínimo Vigente en el Distrito Federal (SMVDF) en caso de:

- No implementar una política de identificación y conocimiento del cliente (conocido también como KYC – por sus siglas en inglés)
- No guardar y proteger la información que soporte la actividad vulnerable.
- No respetar plazos y formas de presentación de los avisos.
- Se impondrá una sanción de entre 10,000 y hasta 65,000 días de SMVDF en caso de:
    - Se de omisión en la emisión y presentación de avisos ante la SHCP.
    - Se compruebe participación en actos u operaciones prohibidos en términos de uso de efectivo.
- En caso de los federativos públicos serán sujetos a una multa de 2,000 hasta 10,000 SMVDF en caso de no cumplir con sus obligaciones respectivas en materia de prevención de lavado de dinero (PWC, 2013).

2.1.1 Umbrales de Aviso

Otra de las obligaciones de quienes realizan Actividades Vulnerables, es la presentación de Avisos a la SHCP sobre las operaciones que sus clientes o usuarios lleven a cabo por un monto superior al establecido en la Ley LFPIORPI. De similar manera que con la obligación de identificación, en algunas actividades el Aviso se presenta por la simple realización de la actividad, mientras que en otros existe un umbral de Aviso, mostrados en la Tabla 7 (SHCP, 2015).

Tabla 7 – Umbrales de aviso que estipula la SHCP para monitorear actividades que puedan ser consideradas de alto riesgo.

Actividad	Umbral de Identificación	Umbral de aviso
Juegos con apuesta, concursos y sorteos	325 SMVDF ($23,738)	645 SMVDF ($47,110.80)
Tarjetas de crédito o de servicios	805 SMVDF ($58,797.20)	1,285 SMVDF ($93,856.40)
Tarjetas prepagadas	645 SMVDF ($47,110.80)	645 SMVDF ($47,110.80)
Cheques de viajero	Siempre	645 SMVDF ($47,110.80)
Préstamos o créditos, con o sin garantía	Siempre	1,605 SMVDF ($117,229.20)
Servicios de construcción, desarrollo o comercialización de bienes inmuebles	Siempre	8,025 SMVDF ($586,146)
Comercialización de piedras y metales preciosos, joyas y relojes	805 SMVDF ($58,797.20)	1,605 SMVDF ($117,229.20)
Subasta y comercialización de obras de arte	2,410 SMVDF ($176,026.40)	4,815 SMVDF ($351,687.60)
Distribución y comercialización de todo tipo de vehículos (terrestres, marinos, aéreos)	3,210 SMVDF ($234,458.40)	6,420 SMVDF ($468,916.80)
Servicios de blindaje (vehículos y bienes inmuebles)	2,410 SMVDF ($176,026.40)	4,815 SMVDF ($351,687.60)
Transporte y custodia de dinero o valores	Siempre	3,210 SMVDF ($234,458.40)
Derechos personales de uso y goce de bienes inmuebles	1,605 SMVDF ($117,229.20)	3,210 SMVDF ($234,458.40)
Recepción de donativos por parte de organizaciones sin fines de lucro	1,605 SMVDF ($117,229.20)	3,210 SMVDF ($234,458.40)

Fuente: Secretaría de Hacienda y Crédito Público. (SHCP, 2015)

2.2 Herramientas para Combatir el Lavado de Dinero y el Financiamiento al Terrorismo.

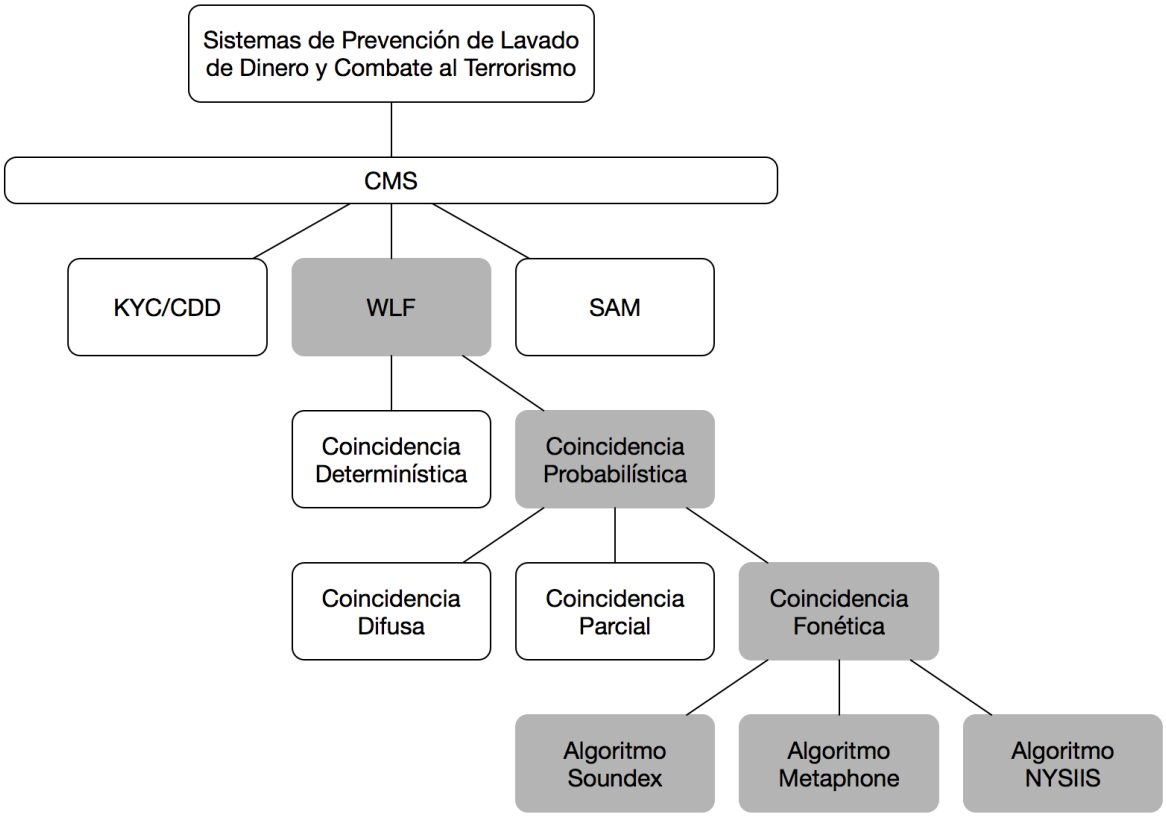
Las actividades sospechosas deben de ser monitoreadas con herramientas de software que permitan seguir el paso en tiempo real de operaciones que pudieran derivar en el blanqueo de capitales o que permitan financiar actividades delictivas como el terrorismo. Estas herramientas deben de ser capaces de procesar grandes cantidades de información provenientes de transacciones, transferencias u operaciones en ventanilla de alguna entidad financiera o de alguna entidad que se encuentre trabajando con alguna de las actividades consideradas como vulnerables (SHCP, 2015).

Es por esto que las principales áreas de monitoreo se han dividido por tu tipo de actividad, y existen distintas formas de revisar esas actividades. Se muestra un diagrama en la Ilustración 8 que ejemplifica el tipo de actividad, las técnicas de búsqueda y los algoritmos más utilizados específicamente para la búsqueda de nombres sobre listas negras (WLF, por sus siglas en inglés).

En este diagrama (Ilustración 8), se aprecia que existen diferentes técnicas utilizadas para la búsqueda de nombres dentro de listas negras, estas técnicas pueden estar basadas en coincidencias determinísticas (reglas estipuladas), en las cuales es necesario saber exactamente lo que se quiere comparar y la coincidencia debe de cumplir cabalmente el criterio o regla evaluado; o coincidencias probabilísticas, las cuales son coincidencias no necesariamente exactas pero muy aproximadas que permiten encontrar semejanzas entre pares de cadenas más rápido y de una forma dinámica.

En este estudio, se utilizaron algoritmos de búsqueda aproximada (difusa, parcial o fonética), ya que pueden dar resultados en menor tiempo y no requieren de reglas precisas para poder encontrar una coincidencia, a diferencia de las técnicas determinísticas (véase capítulo 2.3.3).

Ilustración 8 - Diagrama que muestra el tipo de actividad a monitorear y los mecanismos de búsqueda que se utilizan para satisfacer la demanda de dicha actividad.



Fuente: Elaboración Propia

2.2.1 Listas Negras

En México la SHCP y en Estados Unidos el Departamento de Tesorería mantienen listas de personas o compañías con las que todas las personas tienen permiso hacer tratos o negocios. Estas listas son actualizadas constantemente (algunas listas son diarias, otras semanales o mensuales), y contienen miles de nombres de personas que viven principalmente en el país donde la lista es mantenida, así mismo, cientos de personas o compañías consideradas de alto riesgo para hacer negocios con ellas (Ramachandran, 2014).

Realizar algún negocio con alguna persona o empresa que se encuentran registradas dentro de esta lista puede resultar en multas, encarcelamiento o perdida de reputación financiera, según sea el caso.

Algunas de las listas negras que se utilizan a nivel mundial son:

- OFAC (Office of Foreign Assets Control, por su siglas en inglés)
- FBI (Federal Bureau of Investigation, por sus siglas en inglés)
- Interpol

En México se cuenta con la lista de la SHCP de reciente creación (Videgaray Caso, 2014), y que está basada en la lista de la OFAC con algunas adiciones.

2.2.1.1 La Lista de OFAC

La lista de OFAC contiene una lista de individuos, grupos y entidades sujetas a sanciones económicas por el departamento de Tesorería de los Estados Unidos. La OFAC administra y hace cumplir el programa de sanciones económicas principalmente contra los países y grupos de individuos, como los terroristas o traficantes de drogas. Las sanciones pueden ser globales o selectivas, utilizando el bloqueo de activos y comercios, con el fin de hacer valer las leyes de política exterior. La lista de OFAC es mantenida periódicamente y cubre (OFAC List Search, 2016):

- Individuos o entidades localizadas en todo el mundo que pertenecen o son controladas, o actúan en vez de, el gobierno o país sancionado.
- Terroristas y traficantes de drogas especialmente designados.
- Organizaciones terroristas.
- Individuos identificados como implicados en la proliferación de armas de destrucción masiva.
- Compañías, bancos y empresas particulares que a simple vista parecieran no estar relacionadas con los objetivos de sanción que representan.

2.2.1.2 Retos para Mantener Actualizadas las Listas Negras

Existen distintos tipos de listas negras publicadas por gobiernos, cuerpos políticos, económicos y de cumplimiento de ley. También hay listas publicadas por fuentes comerciales, como la lista de PEP (Politically Exposed Person, por su siglas en inglés), listas internas creadas por las compañías mismas y que requieren de revisión (PwC Sanctions Revisited, 2014).

2.2.1.3 Tipos de Listas Negras

- La lista de OFAC (que contiene fuentes como SDN, la lista del consejo legislativo de Palestina, etc..).Listas especificas por países (por ejemplo, la lista de la SHCP para México, o la lista de sanciones de China y Japón).
- Listas de regiones específicas (por ejemplo, la lista de Asia, la lista de la Unión Europea, etc..).
- Listas de sanciones específicas de negocios (ejemplo, la lista de Mercadeo Especifico de Exportadores).
- Listas Internas, donde los bancos o comercios generan sus propias listas en base a clientes de alto riesgo que no han logrado encontrar en alguna otra lista.
- Alguna otra lista oficial como la de los más buscados de la Interpol.

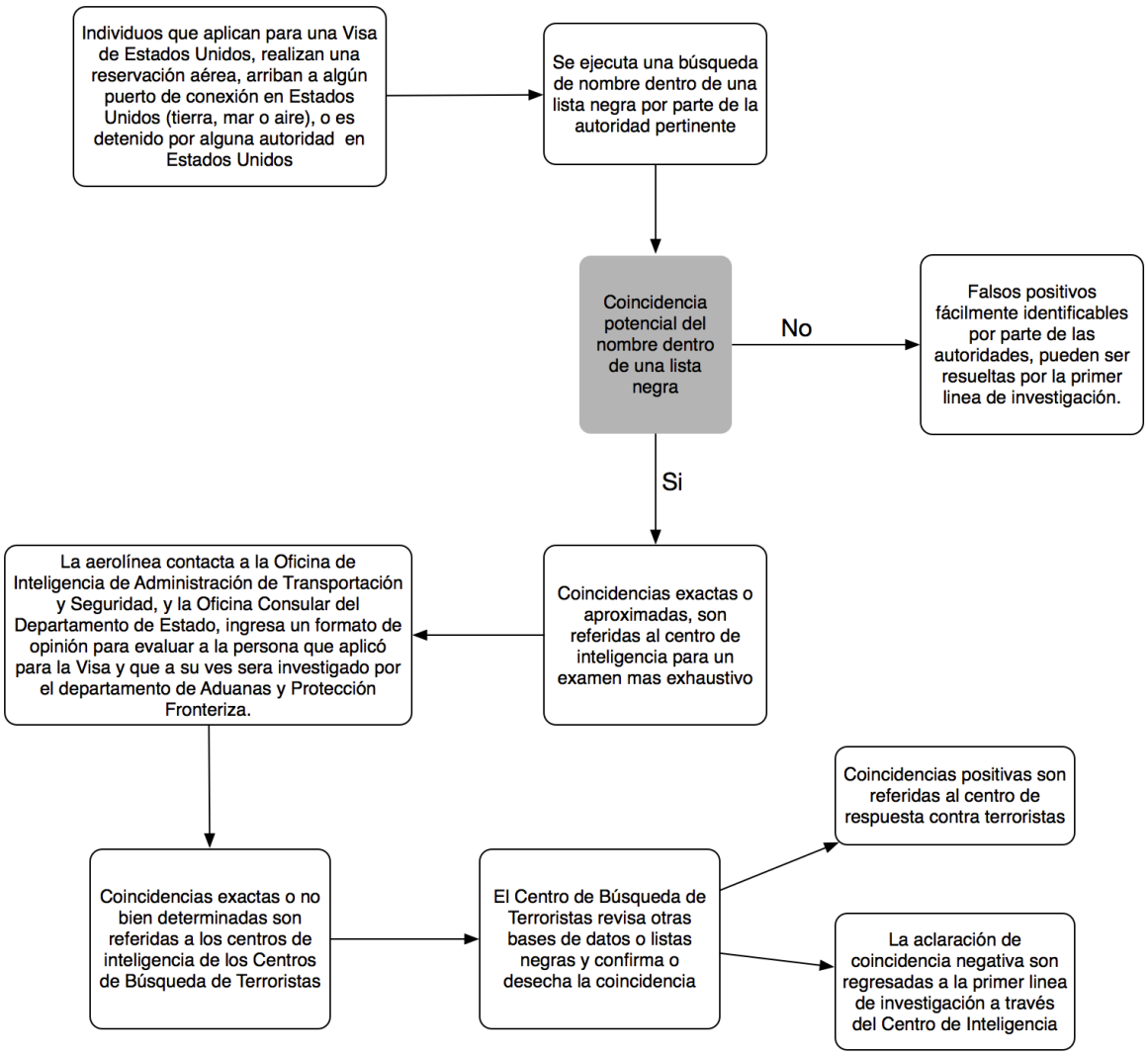
2.2.2 Búsqueda de Nombres sobre Listas Negras

El proceso de búsqueda de un nombre sobre alguna lista negra es disparado por distintos tipos de actividades y es mostrado en la Ilustración 9 (para un ejemplo en el que una persona necesita entrar a un país y pasa por el proceso aduanal), estas pueden ser:

- Entradas de individuos a un país (ver Ilustración 9)
- Alta de clientes en una institución financiera
- Contrato de compra-venta con un nuevo proveedor dentro de una cadena de suministros

- Pagos que excedan los umbrales estipulados por el gobierno del país donde se haga el monitoreo de actividades sospechosas, ver Tabla 7.

Ilustración 9 - Vista general del proceso de búsqueda de nombres sobre una lista negra, ejemplo relacionado a la entrada de un individuo a un país.



Fuente: Revisión de Listas Negras contra el Terrorismo. (Government Accountability Office, 2008).

Como se puede observar en la Ilustración 9, para poder corroborar una coincidencia parcial de nombre se requiere de hacer revisiones con diferentes entidades, en algunos casos, la coincidencia parcial es errónea lo cual se conoce como *falso positivo*.

2.2.3 Falsos Positivos

Un falso positivo es el caso donde una transacción es bloqueada y es asociada con un cliente genuino, la razón del bloqueo es porque el nombre del cliente tiene similitud con algún nombre contenido en una lista negra.

Los falsos positivos pueden contribuir al exceso de trabajo administrativo, debe de ser investigado y corroborado para poder así ser descartado (Department of the Treasury, 2015).

2.2.3.1 Mitigación de Falsos Positivos

Los falsos positivos deben de ser mitigados para evitar re-trabajo en la medida de lo posible, esto deberá realizarse sin crear *falsos negativos* (casos en donde se dejan pasar las transacciones de terroristas o individuos dentro de una lista negra). Para tal caso Ramachandran (Ramachandran, 2014) propone los pasos:

- *Revisar las sentencias de forma separada.* Los algoritmos deben de separar los nombres, direcciones y ciudades de cualquier otro dato. No se deberán combinar con ningún otro dato dentro de la búsqueda. De lo contrario un nombre como José González que vive en la calle Martínez, podría hacer coincidencia con el nombre bloqueado José Martínez González.
- *Revisar las cuentas de forma separada.* Los algoritmos deben separar los nombres personales y los nombres de corporaciones, por ejemplo, Mayan King Limited que se encuentra en la lista de SDN de la OFAC, podría hacer coincidencia con el nombre Maya.
- *Revisar los contenedores de forma separada y solo revisar los nombres de contenedores con contenedores.* El transporte de contenedores ha tenido recientemente un especial foco de atención dado que en ellos se puede transportar petróleo, aceite, así como uranio y otros químicos que podrían ser usados para la fabricación de armas de destrucción masiva.
- *Identificar que campos son los que se van a revisar*. En un mensaje, el bloque de la cabecera y el bloque de texto o el cuerpo contienen la mayoría de

los datos del mensaje. Se tiene que identificar y mapear de forma correcta aquellos campos que son específicos y únicos. La lista de OFAC contiene nombres de personas, de países, números de pasaporte, estos elementos son únicos, es por esto que deben de ser los campos en los que se tiene que realizar la búsqueda.

- *Intentar excluir nombres de locaciones.* Las direcciones no son información única para identificar a una persona, y, por lo tanto, tomarlas en cuenta para realizar la búsqueda de nombres incrementa la posibilidad de obtener falsos positivos. Por ejemplo, una dirección como calle Havana, abre la posibilidad de confundir el nombre con alguien de Cuba o Illinois donde existe un nombre de calle similar. Se tiene que intentar excluir de las búsquedas nombres de direcciones, estados, provincias, territorios, códigos postales (ya que los números podrían llegar a coincidir parcialmente con los números de los pasaportes).
- *Crear reglas para los nombres comunes.* Nombres como Jim/Jimmy/, David, José/Joseph, son comunes, es por eso que las reglas deben de ser específicas para poder mitigar así los falsos positivos.
- ***Utilización de algoritmos específicos para los lenguajes.*** Hoy en día los algoritmos más utilizados de tipo fonético para la búsqueda de nombres sobre listas negras, están basados en el lenguaje inglés, es por esto que se han hecho intentos por adaptar las reglas gramaticales de dichos algoritmos los distintos lenguajes de diferentes países, para poder hacer búsquedas con un incremento en su precisión y con esto ayudar con el decremento de falsos positivos.

2.3 Algoritmos Utilizados para la Búsqueda de Nombres

En la Ilustración 8 se pueden observar los diferentes mecanismos de búsqueda, así como 3 algoritmos fonéticos que son utilizados hoy en día para la búsqueda de nombres en listas negras como la lista de OFAC.

Tipos de algoritmos utilizados para la búsqueda y coincidencia de nombres en una lista negra

2.3.1 Coincidencia Determinística

La coincidencia de nombres determinística, también llamada la coincidencia basada en reglas, utiliza una combinación de algoritmos y reglas de negocio para poder determinar cuándo dos o más registros coinciden a través de identificadores únicos como lo son el SSN (Social Security Number, por sus siglas en inglés), número de pasaporte o en el caso de México el CURP. Este tipo de búsquedas coinciden de forma perfecta con ciertos criterios o campos; genera también uniones entre otros campos haciendo uso de los identificadores únicos a lo largo de todo el conjunto de datos a evaluar. Esta es la más fácil y rápida estrategia de unión de campos, sin embargo, los sistemas basados en coincidencia determinística, tienen en una precisión menor comparados con los sistemas no deterministicos o también llamados probabilísticos. Este tipo de sistemas encajan mejor para trabajar con un numero de datos relativamente pequeño (menos de 2 millones de datos) (Miller & Arehart, 2008).

2.3.1.1 Coincidencia Directa

Una relación de coincidencia entre dos registros es directa cuando esos dos registros tienen dicha coincidencia basada en una regla vigente. Cuando un nombre de cliente coincide de manera exacta con el nombre en una lista sancionada o lista negra, las instituciones financieras deben de reaccionar apropiadamente para cumplir con la ley que rija el país donde ese nombre este siendo evaluado, en el caso de estados unidos, con la ley USA PATRIOT y los requerimientos de OFAC. No cumplir con estas leyes puede resultar en multas y penalizaciones así como daño a la reputación financiera (Ramachandran, 2014).

Tabla 8 - Ejemplo de coincidencia de nombres directa

Objetivo a Buscar	Objetivo Sancionado	Decisión de la Coincidencia
David Carlos	David Carlos	Coincide
Osama Bin Laden	Osama Bin Laden	Coincide
Joaquín Loera	Joaquín Loera	Coincide

Fuente: Elaboración Propia.

Tal como se muestra en la Tabla 8, donde los resultados arrojados por una búsqueda de nombres directa tienen una coincidencia exacta, es decir, los nombres no contienen ningún tipo de variación fonética o de caracteres.

2.3.2 Coincidencia Probabilística

La coincidencia probabilística se refiere a la comparación de distintos valores de campos entre dos registros, y asignando a cada campo un peso que indica que tan cerca los dos valores tienen coincidencia. La suma de los pesos de los campos individuales indica la cercanía o la coincidencia entre dos registros. La coincidencia probabilística efectúa análisis estadísticos de los datos y determina la frecuencia de los distintos elementos. Luego aplica ese mismo análisis para determinar un peso para la coincidencia, de la misma forma que un usuario puede determinar la relevancia que tiene alguna tupla o registro en particular.

La búsqueda de nombres por coincidencia probabilística toma relevancia en casos en los cuales se tiene información la cual es afectada por, aspectos multiculturales, nombres, captura de datos hecha por diferentes sistemas de captura o instituciones y prácticas culturales para asignación de nombres (Miller & Arehart, 2008).

2.3.3 Tipos de Coincidencia de los Algoritmos Probabilísticos

Los sistemas de captura y de ingreso de información a los diferentes sistemas se ven afectados por la falta de cuidados al momento de ingresar la información, esto conlleva a tener sistemas de datos no correctamente estructurados o con errores ortográficos, abreviaciones, letras eliminadas u omitidas, letras extras, letras cambiadas por otras, palabras divididas, palabras juntas, etc., es por esto que para los diferentes tipos de errores y modificaciones culturales, se han desarrollado distintos algoritmos y tipos de búsqueda, algunos hechos específicamente para atender algún tipo de error en particular, otros para revisar pronunciación de nombres (algoritmos fonéticos) (Rajkovic & Dragan, 2007).

2.3.3.1 Coincidencia Difusa

La coincidencia difusa es la implementación de un proceso algorítmico llamado lógica difusa (fuzzy logic, comúnmente conocido en inglés) para determinar la similitud que existe entre elementos de datos como lo son el nombre de negocio, nombre personal o información acerca de la dirección. La función de la lógica difusa permite que el algoritmo detecte y evaluar coincidencias cercanas más que coincidencias exactas. Dependiendo del algoritmo, este podría considerar nombres alternativos, como lo son "José" o "Joseph". Los nombres (personas, lugares o entidades) tendrían una coincidencia sencilla si los datos fuesen consistentes; sin embargo, los individuos que se dedican al lavado de dinero utilizan diferentes técnicas para sobrepasar los filtros de detección impuestos.

En la Tabla 9 observamos un ejemplo de coincidencia por medio de la lógica difusa, vemos como nombres similares en tamaño de caracteres o tipo de caracteres podrían provocar una coincidencia que sumada a otros algoritmos darían un valor o un peso específico de dicha coincidencia y podría ser necesario una segunda revisión del individuo a evaluar.

Tabla 9 - Ejemplo de coincidencia por medio de lógica difusa

Objetivo a Buscar	Objetivo Sancionado	Decisión de la Coincidencia
Peter	Petr	Coincidencia a través de Lógica Difusa
Qadir	Kadar	Coincidencia a través de Lógica Difusa
Rahim	Raheem	Coincidencia a través de Lógica Difusa

Fuente: Elaboración Propia.

2.3.3.2 Coincidencia Parcial

Existen algoritmos que ayudan a encontrar lo que se conoce como coincidencias parciales, este tipo de algoritmos reportan posibles coincidencias cuando la información de un cliente es la misma o similar a la información que se encuentra en una lista negra. Dos registros muestran este tipo de relación cuando alguno de los elementos del primer registro (no todos) coinciden con alguno de los elementos del segundo registro (no todos). Un ejemplo típico podría ser los registros correspondientes a los de un padre y su hijo que viven en la misma dirección y que podrían tener el mismo apellido, incluso el mismo nombre o el mismo número de teléfono local, pero elementos del registro como número de teléfono móvil, número de seguro social, CURP, correo electrónico y algún otro campo podrían no coincidir.

Las coincidencias presentan un reto para los diferentes algoritmos de búsqueda, porque tienen que involucrar más variables a considerar y esto, podría significar un incremento en el tiempo que toma el algoritmo en encontrar una coincidencia. Un ejemplo de este tipo de coincidencia se muestra en la Tabla 10, donde existe un registro similar al registro que se encuentra sancionado, pero hay diferencias de nombre.

Tabla 10 - Ejemplo de coincidencia de nombres parciales

Objetivo a Buscar	Objetivo Sancionado	Decisión de la Coincidencia
John Paul Castro	John Peter Castro	Coincide Parcialmente - Paul ≠ Peter
David Jol Chung	Daniel Jol Chung	Coincide Parcialmente - David ≠ Daniel
John Longman	Emily Longman	Coincide Parcialmente - John ≠ Emilly

Fuente: Elaboración Propia.

2.3.3.3 Coincidencia Fonética

La coincidencia fonética es el proceso de hacer coincidir información o datos utilizando algoritmos o funciones que han sido creados específicamente para concentrarse en como la palabra es pronunciada más que como está escrita. Un algoritmo fonético hace coincidir dos palabras diferentes que tienen una pronunciación similar con un mismo código, lo cual permite una emplear una técnica de indexación y comparación basado en similitudes fonéticas.

Existen palabras que tienen distinta forma de escritura, pero una pronunciación similar y deberían de coincidir, por ejemplo, Sofía y Sophia, Reynold y Renault, etc., por lo tanto un motor de búsqueda es requerido para construir conexiones basadas en diferentes reglas de transformación fonética.

Los algoritmos de coincidencia fonética son comúnmente acompañados por un algoritmo de lógica difusa para incrementar la precisión de la búsqueda y la rapidez para obtener resultados, en la Tabla 11 se muestran algunos ejemplos de nombres que se pronuncian de forma similar (Shah, 2014).

Tabla 11 - Ejemplo de coincidencia fonética

Objetivo a Buscar	Objetivo Sancionado	Decisión de la Coincidencia
Sofia	Sophia	Lógica Difusa
Reynolds	Renaults	Lógica Difusa
Smith	Smyth	Lógica Difusa

Fuente: Elaboración Propia.

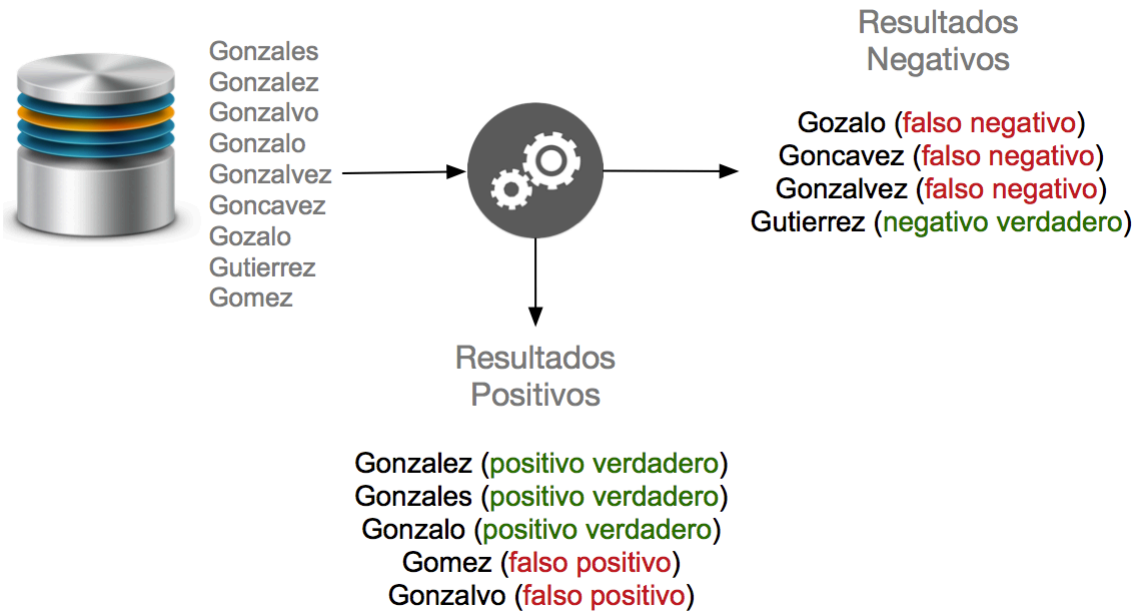
La mayoría de los algoritmos fonéticos están construidos y adaptados para el lenguaje inglés.

2.4 Funcionamiento de la Coincidencia Fonética

La coincidencia fonética realiza la operación de coincidencia basada en la pronunciación de las palabras. Para entender mejor el funcionamiento, discutiremos el ejemplo de una base de datos que contiene apellidos como: González, Gonzales, Gonzalvo, Gonzalo y Gómez. Supongamos que queremos buscar el apellido de Gonzalvez. Las coincidencias que las búsquedas arrojan son llamados resultados positivos, y aquellos que no son tomados en cuenta son llamados negativos.

Los resultados positivos que son tomados en cuenta o que tienen relevancia, son llamados "positivos verdaderos", y los que no tienen relevancia llamados "falsos positivos" (Shah, 2014).

Ilustración 10 - Funcionamiento de un algoritmo fonético



Fuente: Elaboración propia

Como lo muestra la Ilustración 10 en lo que es un ejemplo de coincidencia fonética, podemos ver como hay:

- Resultados relevantes (positivos verdaderos).
- No relevantes (falsos positivos).
- Resultados rechazados que son relevantes (falso negativo).
- Resultados rechazados no relevantes (negativo verdadero).

Los resultados son catalogados como relevantes o no relevantes por medio del uso de técnicas fonéticas, y un poco de tratamiento manual, es decir, cuando una alerta es disparada se tiene que investigar más a detalle para poder confirmar que en efecto es una coincidencia verídica y valida.

Existen muchas investigaciones en la rama de los algoritmos fonéticos, algoritmos basados en la pronunciación de las palabras (Beider & Morse, 2010), algoritmos basados en la edición de la distancia (Amón, Moreno, & Echeverri, 2012) y algoritmos basados en patrones (Christen, 2006).

Dentro de los algoritmos fonéticos basados en la pronunciación de las palabras existen algoritmos como: *soundex, metaphone, double metaphone, NYSIIS, entre otros*. Para esta investigación se consideraron los algoritmos soundex, metaphone y NYSIIS para observar su comportamiento con nombres basados en la lengua española, esto por considerarse 3 de los algoritmos más utilizados para la búsqueda de nombres en diferentes motores de coincidencia fonética (Ramachandran, 2014).

2.4.1 Algoritmo Soundex

Soundex es un algoritmo de codificación fonética propuesto en el año 1912 por Robert Russell (Shah, 2014), que convierte una palabra en un código. El código soundex consiste en sustituir las consonantes de la palabra afectada por un número, si es necesario, se agregan ceros al final del código para conformar un código de 4 dígitos. Soundex elige una clasificación de los caracteres con base en el lugar de la articulación de la lengua inglesa. La Tabla 12 presenta las equivalencias usadas por el algoritmo soundex.

Tabla 12 - Equivalencias del algoritmo soundex.

Dígito	Caracteres
1	B,F,P,V
2	C,G,J,K,Q,S,X,Z
3	D,T,
4	L
5	M,N,
6	R

Fuente: Algoritmo Fonético para la Detección de Cadenas de Texto Duplicadas en el Idioma Español. (Amón et al., 2012)

Debido a que, en el idioma inglés, las letras A, E, I, O, U, H, W y Y no hacen diferenciación fonética, son descartadas. Adicionalmente existen otras reglas complementarias para la codificación de las letras dobles (si el texto tiene letras dobles, estas deben ser tratadas como una sola letra) y para letras a lado y lado

con el mismo código (si el texto tiene diferentes letras lado a lado que tienen el mismo número en la guía de codificación soundex, estas deben ser tratadas como una sola letra) (Harris & Ross, 2006).

Por ejemplo: Giraldo se codifica G643 (G, 6 por la R, 4 por la L, 3 por la D, las otras letras se descartan). Juan se codifica J500 (J, 5 por la N, las otras letras se descartan y se agregan dos ceros) (Amón et al., 2012).

2.4.2 Algoritmo Metaphone

El algoritmo metaphone fue publicado por primera vez en 1990 por Lawrence Phillips y fue creado para poder indexar palabras con pronunciación en inglés, como un intento de mejorar lo que ya se tenía con el algoritmo soundex. Este algoritmo considera variaciones e inconsistencias que se tienen en la pronunciación de algunas palabras en inglés para producir así una codificación más precisa.

El algoritmo metaphone utiliza 16 símbolos de consonantes '0BFHJKLMNPRSTWXY', donde el '0' representa la "th", la 'X' representa "sh" o "ch", y las otras letras representan las de consonantes usuales en el lenguaje inglés. Las vocales 'AEIOU' son también utilizadas, pero solo al principio del código, a continuación la Tabla 13 muestra la mayoría de las reglas de la implementación original (Snae, 2007).

Tabla 13 - Reglas del algoritmo metaphone.

1	Remover todos los vecinos repetidos exceptuando la letra C.
2	El inicio de la palabra deberá ser transformado utilizando las siguientes reglas:
	KN → N
	GN → N
	PN → N
	AE → E
	WR → R
3	Remover la letra B al final si se encuentra después de la M.
4	Reemplazar la letra C utilizando las siguientes reglas:
	Con X: CIA → XIA, SCH → SKH, CH → XH
	Con S: CI → SI, CE → SE, CY → SY
	Con K: C → K
5	Reemplazar la letra D utilizando las siguientes reglas:
	Con J: DGE → JGE, DGY → JGY, DGI → JGY
	Con T: D → T
6	Reemplazar GH → H, exceptuando si es al final o antes de una vocal.
7	Reemplazar GN → N y GNED → NED, si están al final.
8	Reemplazar G utilizando las siguientes reglas:
	Con J: GI → JI, GE → JE, GY → JY
	Con K: G → K
9	Reemover todas las letras H después de una vocal pero no antes de una vocal.
10	Realizar las siguientes transformaciones utilizando las siguientes reglas:
	CK → K
	PH → F
	Q → K
	V → F
	Z → S
11	Reemplazar S con X:
	SH → XH
	SIO → XIO
	SIA → XIA
12	Reemplazar la letra T utilizando las siguientes reglas:
	Con X: TIA → XIA, TIO → XIO
	Con 0: TH → 0
	Remover: TCH → CH
13	Transformar WH → W al inicio. Remover W si no hay una vocal después de ella.
14	Si la letra X esta al inicio, entonces reemplazar X → S, en caso contrario, reemplazar X → KS
15	Remover todas las letras Y que no estén antes de una vocal
16	Remover todas las vocales, exceptuando las vocales que estan al inicio de la palabra

Fuente: Caverphone : Phonetic Matching algorithm. (Hood, 2012)

Algunos ejemplos de codificación de nombres del algoritmo metaphone son:

- **FXPL** → Fishpool, Fishpoole.
- **JLTL** → Gallately, Gelletly.
- **LWRS** → Lowers, Lowerson.
- **MLBR** → Mallabar, Melbert, Merlbourn, Melbourne, Melburg, Melbury, Milberry, Milborn, Milbourn, Milburne, Mulbry.

- **SP** → Saipy, Sapey, Sapp, Sappy, Spey, Seppey, Sopp, Zoppie, Zoppo, Zupo, Zuppa.

Por ejemplo: Tomando el apellido González, se aplica la regla 8 lo cual nos da la primera letra codificada, la letra K, después al no presentarse ninguna combinación de las mencionadas en las reglas del algoritmo (esto también por no ser un apellido que cumpla reglas gramaticales del lenguaje inglés), simplemente se mantienen las consonantes, por lo que el resultado final codificado es: **KNSLS**, para la primer S se aplicó la regla 10, sustituyendo la Z por la S.

2.4.3 Algoritmo NYSIIS

El algoritmo NYSIIS (New York State Identification and Intelligence Algorithm, por sus siglas en inglés) es un algoritmo que ve la luz en 1970, mediante un proyecto liderado por Robert L. Taft y publicado en el artículo titulado "Name Search Techniques" (Harris & Ross, 2006). En este documento se explica el tipo de diseño empírico usado para el desarrollo de este algoritmo. La Tabla 14 muestra las reglas que el algoritmo NYSIIS sigue para su ejecución.

Tabla 14 - Reglas del algoritmo NYSIIS.

1	Se transforma el inicio de la palabra usando las siguientes reglas:
	MAC → MCC
	KN → N
	K → C
	PH, PF → FF
	SCH → SSS
2	Se transforma el final de la palabra usando las siguientes reglas:
	EE → Y
	IE → Y
	DT, RT, RD, NT, ND → D
3	Se transforman todas las letras exeptuando la primera usando las siguientes reglas:
	EV → AF
	A, E, I, O, U → A
	Q → G
	Z → S
	M → N
	KN → N
	K → C
	SCH → SSS
	PH → FF
	Depues de una vocal: remover H y transformar W → A
4	Remover la S al final
5	Transformar AY al final en → Y
6	Remover A al final
7	Truncar la palabra a 6 letras (este último paso es opcional)

Fuente: Caverphone : Phonetic Matching algorithm. (Hood, 2012)

Algunos ejemplos de codificación de nombres del algoritmo NYSIIS son:

- **FXPL** → Fishpool, Fishpoole.
- **JLTL** → Gellately, Gelletly.
- **LWRS** → Lowers, Lowerson.
- **MLBR** → Mallabar, Melbert, Melbourn, Melbourne, Melburg, Melbury, Milberry, Milborn, Milbourn, Milbourne, Milburn, Milburne, Millberg, Mulberry, Mulbery, Mulbry.
- **SP** → Saipy, Sapey, Sapp, Sappy, Sepey, Seppey, Sopp, Zoppie, Zoppo, Zupa, Zupo, Zuppa.

Por ejemplo: seleccionando el apellido Cámara, al no haber una regla específica para el inicio de una palabra con la letra C, se deja la C, la letra á acentuada, se transforma en A hacienda uso de la regla 3, aquí es donde podría haber una diferenciación, al ser las palabras acentuadas un rasgo no incluido en las reglas

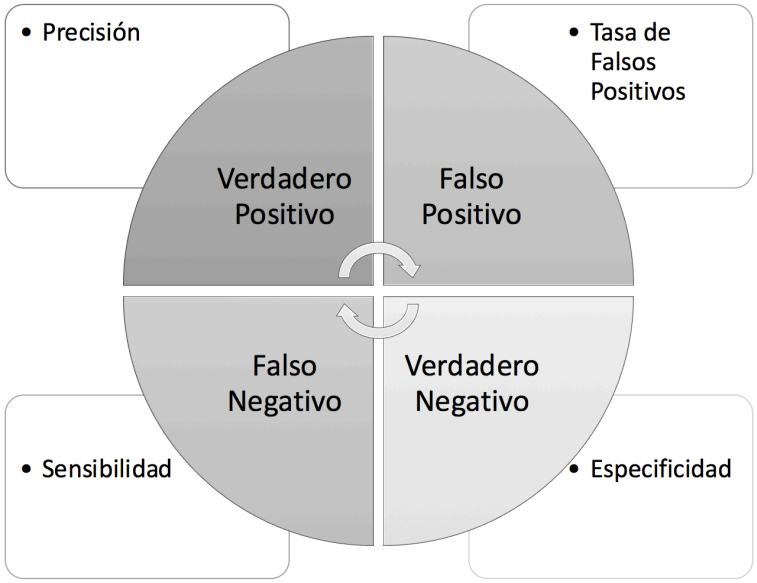
gramaticales del lenguaje inglés y que da una pronunciación diferente en el lenguaje en español. Por último se remueve la letra A del final de la palabra utilizando la regla 6, la palabra final nos queda: **CAMAR**.

2.5 Medidas de Calidad de Enlace de Datos

En esta sección se presentan diferentes medidas utilizadas para poder analizar la información que resulta de la comparación de dos conjuntos de datos. Se asume en este método de enlace de la información que al comparar dos elementos de conjuntos de datos distintos podemos obtener lo que conocemos como coincidencia positiva (en este documento le llamaremos incidencia) y la no conciencia (Goiser & Christen, 2007).

En la elaboración de este documento se asume, que se conocen las coincidencias positivas y las no coincidencias en el universo de datos dados, dicho de otra forma, qué pares comparados en efecto son coincidentes (verdaderos positivos) y cuales no lo son (verdaderos negativos), y así, podemos determinar los falsos positivos y los falsos negativos. Estos elementos dan como resultado una matriz de confusión como se muestra a continuación (Ilustración 11).

Ilustración 11 - Matriz de confusión de registro de clasificación de pares



Fuente: Quality and Complexity Measures for Data Linkeage and Deduplication. (Goiser & Christen, 2007)

En esta tabla podemos ver las siguientes medidas con su correspondiente nomenclatura a ser utilizada en este documento:

- **Verdaderos positivos** (**TP**, por sus siglas en inglés, True Positives)
- **Verdaderos Negativos** (**TN**, por sus siglas en inglés, True Negatives)
- **Falsos Positivos** (**FP**, por sus siglas en inglés, False Positives)
- **Falsos Negativos** (**FN**, por sus siglas en inglés, False Negatives)

Las medidas de calidad originadas de estos valores son tomadas del trabajo de Goiser y Christen (Goiser & Christen, 2007) por su estudio en la comparación de datos entre distintos conjuntos de información:

Exactitud (acc). Indica la capacidad que tiene el algoritmo de detectar coincidencias entre los dos conjuntos de datos a evaluar, a menudo se utiliza de la misma forma que la precisión, la diferencia es que la exactitud toma en cuenta todos los valores de la matriz de confusión (Ilustración 11) y está determinada por la siguiente formula:

$$acc = \frac{|TP| + |TN|}{|TP| + |FP| + |TN| + |FN|}$$

Precisión (*prec*). También llamada *valor positivo predictivo*, es la proporción de coincidencias clasificadas que son coincidencias verdaderas y se encuentra representada por la siguiente ecuación:

$$prec = \frac{|TP|}{|TP| + |FP|}$$

Sensivilidad (rec). También conocida como *recall o tasa de verdaderos positivos,* esta es la proporción de coincidencias actuales que han sido clasificadas correctamente, y la fórmula para su cálculo es la siguiente:

$$rec = \frac{|TP|}{|TP| + |FN|}$$

Especificidad (spec). (que es la tasa de verdaderos negativos), este valor podría ser alterado si se cuenta con una tasa muy alta de verdaderos negativos, normalmente originado por que el conjunto de datos ha sido ya tratado previamente para quitar incongruencias. Su fórmula se describe a continuación:

$$spec = \frac{|TN|}{|TN| + |FP|}$$

Tasa de Falsos Positivos (fpr). Se considera la inversa de la especificidad y nos ayuda a determinar la tasa actual de falsos positivos de una evaluación dada. Está determinada por la formula siguiente:

$$fpr = \frac{|FP|}{|TN| + |FP|}$$

o también por:

$$fpr = 1 - spec$$

F-measure. También conocida como la media armónica entre la sensibilidad y la precisión, esta tendrá un valor elevado cuando tanto la sensibilidad como la precisión tengan valores elevados y se encuentra determinada por la formula siguiente:

$$f - meas = 2\left(\frac{prec \times rec}{prec + rec}\right)$$

CAPÍTULO 3: METODOLOGÍA

En este capítulo se abordarán los temas referentes al diseño metodológico que se utilizó para poder probar los algoritmos de una forma sistemática y de acorde a lo que se buscaba del objetivo primario, así como, de los objetivos secundarios. se definirá y explicara el conjunto de datos que se utilizó para las pruebas y las pruebas estadísticas que se utilizaron para la comprobación de los resultados.

3.1 Definición de la Problemática

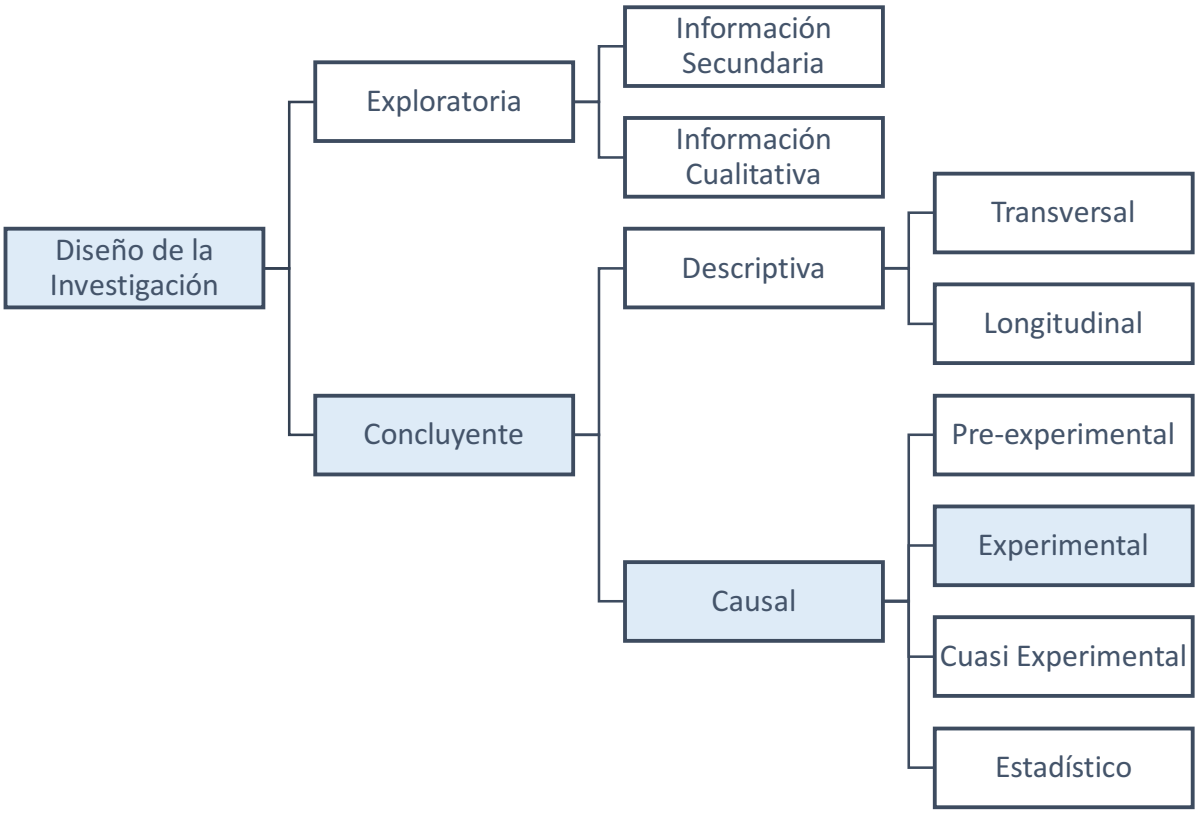
Hoy en día existen diferentes métodos para poder hacer una revisión de nombres sobre listas negras con el motivo de prevenir el lavado de dinero y combatir el terrorismo.

Se utilizan desde algoritmos de lógica difusa, hasta algoritmos fonéticos y de evaluación de distancia entre caracteres (Christen, 2006), normalmente de manera combinada para poder tener resultados más exactos y de manera más rápida. Sin embargo, esto no es suficiente, ya que con los métodos actuales se tienen cifras de hasta un 50% de falsos positivos (Ramachandran, 2014); esto puede significar, el bloqueo de transacciones o la detención de migrantes (en los aeropuertos o aduanas) de forma errónea, así como el uso de recursos humanos y económicos para el procesamiento de estos casos detectados erróneamente.

3.2 Tipo de Investigación

El tipo de investigación que se llevó a cabo en la elaboración de este trabajo, es como se muestra en la Ilustración 12, de tipo experimental.

Ilustración 12- Tipo de Investigación Experimental



Fuente: Metodología de la Investigación. (Hernández, Fernández, & Baptista, 1991)

Esto es porque el método de selección de los datos se da de manera aleatoria y porque el tanto el tratamiento de datos como las observaciones de los mismos se realizan con grupos de datos seleccionados en un universo de datos aleatorizado, no existe ningún patrón de datos u ordenamiento sobre el tratamiento de los mismos.

3.3 Diseño del Experimento

3.3.1 Herramientas

Con el fin de poder analizar los diferentes algoritmos, era necesario contar con las herramientas mencionadas a continuación:

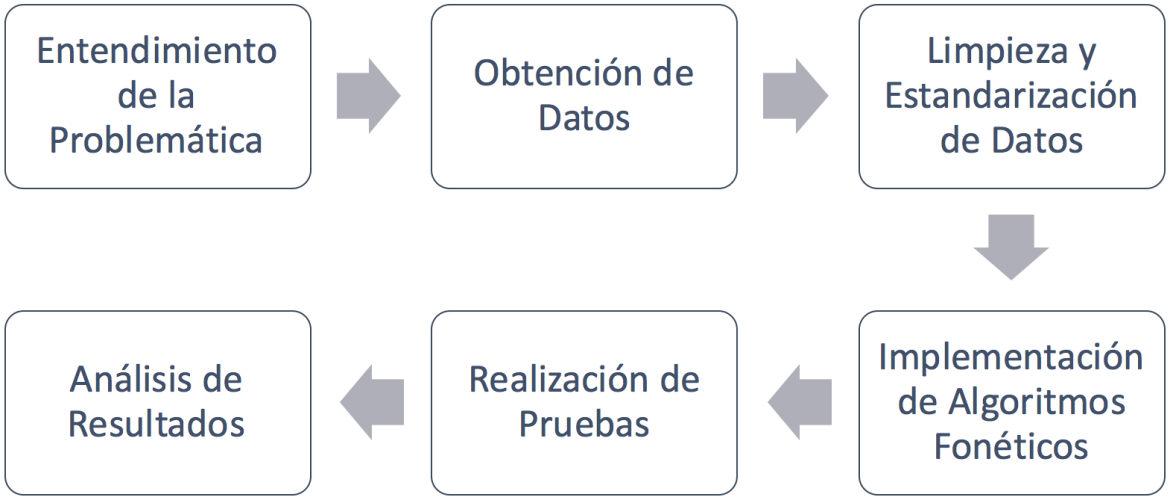
1. Conjunto de nombres a analizar.
2. Lista negra de OFAC (Para este experimento se utilizó la lista del día 30/03/2016).
3. Algoritmos a evaluar (soundex, metaphone y NYSIIS).
4. Aplicación informática donde se utilicen los algoritmos y hagan uso de la lista negra y de la lista de nombres a evaluar.
5. Software de análisis estadístico (SPSS de IBM).

El experimento se llevó a cabo en una computadora con sistema operativo Mac OS X – El Capitán, con un procesador Intel Core i5 y 8GB de RAM.

3.3.2 Método de Análisis

Los pasos que se siguieron para realizar el experimento están descritos en la Ilustración 13 y se mencionan a continuación.

Ilustración 13 - Pasos en forma de proceso que fueron seguidos para el desarrollo de la metodología de este documento de tesis



Fuente: Elaboración Propia.

3.3.2.1 Entendimiento de la Problemática

Las instituciones financieras y cuentan con software para la prevención del lavado de dinero y el combate al terrorismo, estos métodos incluyen búsquedas de patrones, cadenas, nombres y relaciones entre diferentes identificadores de distintos conjuntos de datos.

Entre los algoritmos utilizados se encuentran aquellos de tipo fonético, que focalizan su uso en la pronunciación de las palabras, más que en su tamaño, composición o tipo.

Se sabe por medio de distintos estudios (Rajkovic & Dragan, 2007), que la mayoría de los algoritmos fonéticos están desarrollados para su mejor funcionamiento en el idioma inglés, y sin embargo son pocos los esfuerzos que se han llevado a cabo para su adaptabilidad a otros lenguajes.

Este experimento realizó una comparativa de 3 algoritmos fonéticos con una lista de nombres hispanos. Las comparaciones iniciales fueron hechas contra la lista SDN de la OFAC publicada el día 30 de marzo del 2016. Todo esto con el fin de sentar

bases para un estudio posterior y adaptación de estos algoritmos a ser usados en regiones donde las personas tienen apellidos hispanos en su mayoría.

3.3.2.2 Obtención de Datos

3.3.2.2.1 Conjunto de Datos de Nombres

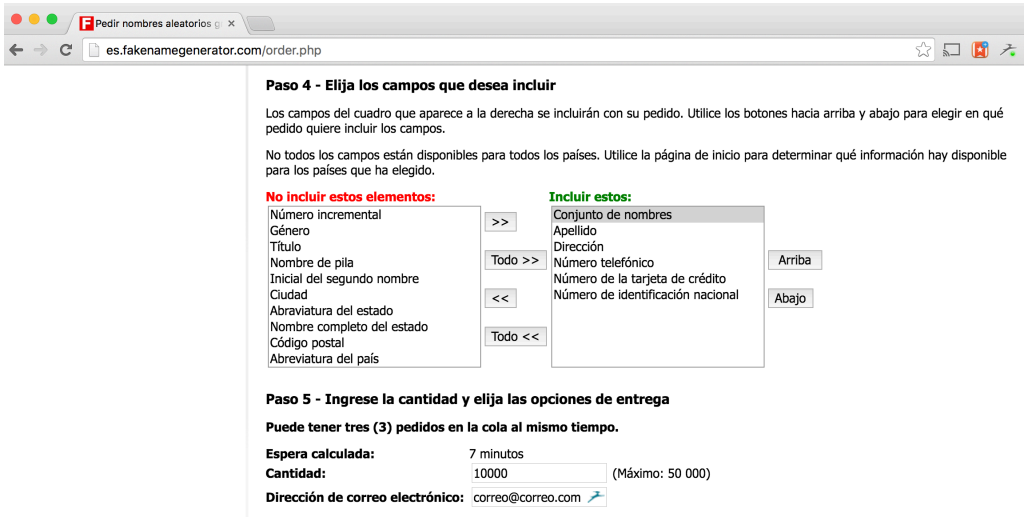
Mediante la herramienta *Web Fake Name Generator* (Works, 2016), se generaron aleatoriamente conjuntos de datos compuestos por registros con atributos como nombre, apellido, dirección y teléfono. Esta herramienta prueba su eficacia en el documento de análisis de algoritmos de Ivan Amón (Amón et al., 2012) para generar cadenas para su posterior uso en comparaciones.

Los conjuntos de datos generados fueron, un conjunto de datos de 10,000 nombres hispanos y un conjunto de datos de 10,000 nombres anglosajones. La Ilustración 14 y la Ilustración 15, muestran el proceso de creación de la lista, en este caso la de nombres hispanos.

Ilustración 14 - Primera parte del formulario para obtener el conjunto de datos de nombres de la página Web Fake Name Generator.

Fuente: Página web de Web Fake Name Generator. (Works, 2016)

Ilustración 15 - Segunda parte del formulario para obtener el conjunto de datos de nombres de la página Web Fake Name Generator.



Fuente: Página web de Web Fake Name Generator. (Works, 2016)

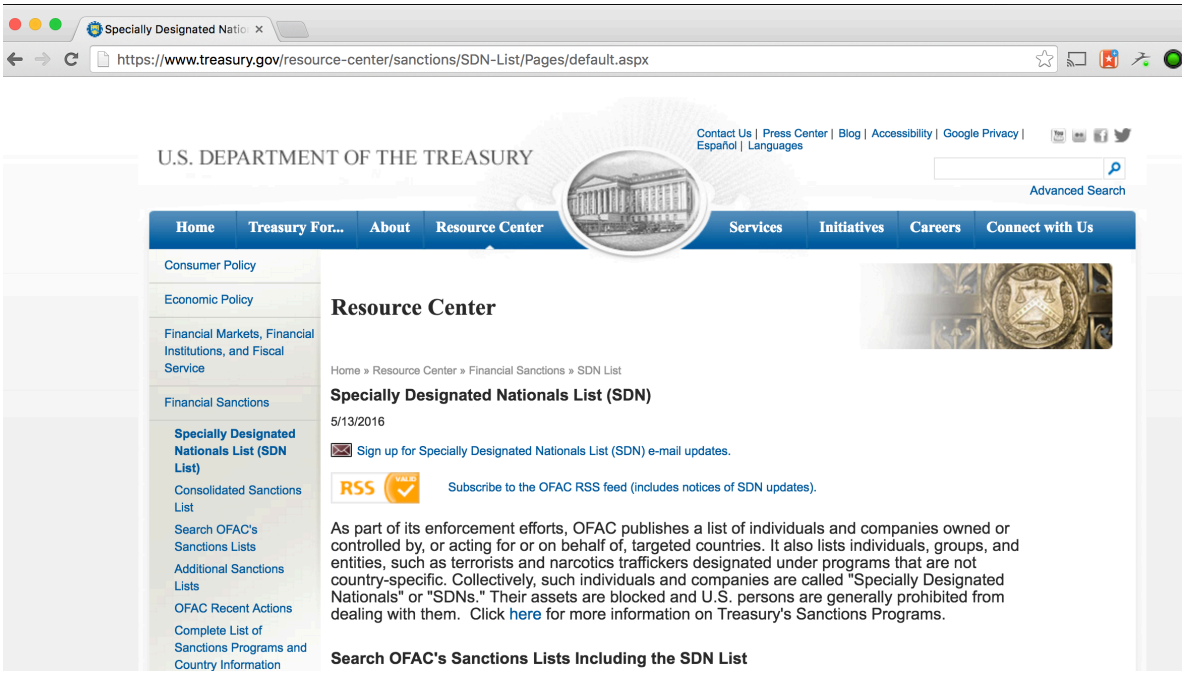
3.3.2.2.2 Obtención de la Lista Negra de OFAC (SDN)

La lista SDN se obtuvo al ingresar a la página del Departamento del Tesoro de Estados Unidos (OFAC SDN List, 2016), específicamente en la sección de *Data Center*. Esta lista sirvió para poder llevar a cabo el contraste entre la lista de nombres obtenida anteriormente en la página de *Web Fake Name Generator.*

Existen muchos tipos de listas en la página de la OFAC, pero era de un interés particular, aquella que contuviese nombres de empresas y nombres de personas para poder tener dos campos de datos comparables.

La Ilustración 16 muestra la página inicial para descarga de la lista SDN.

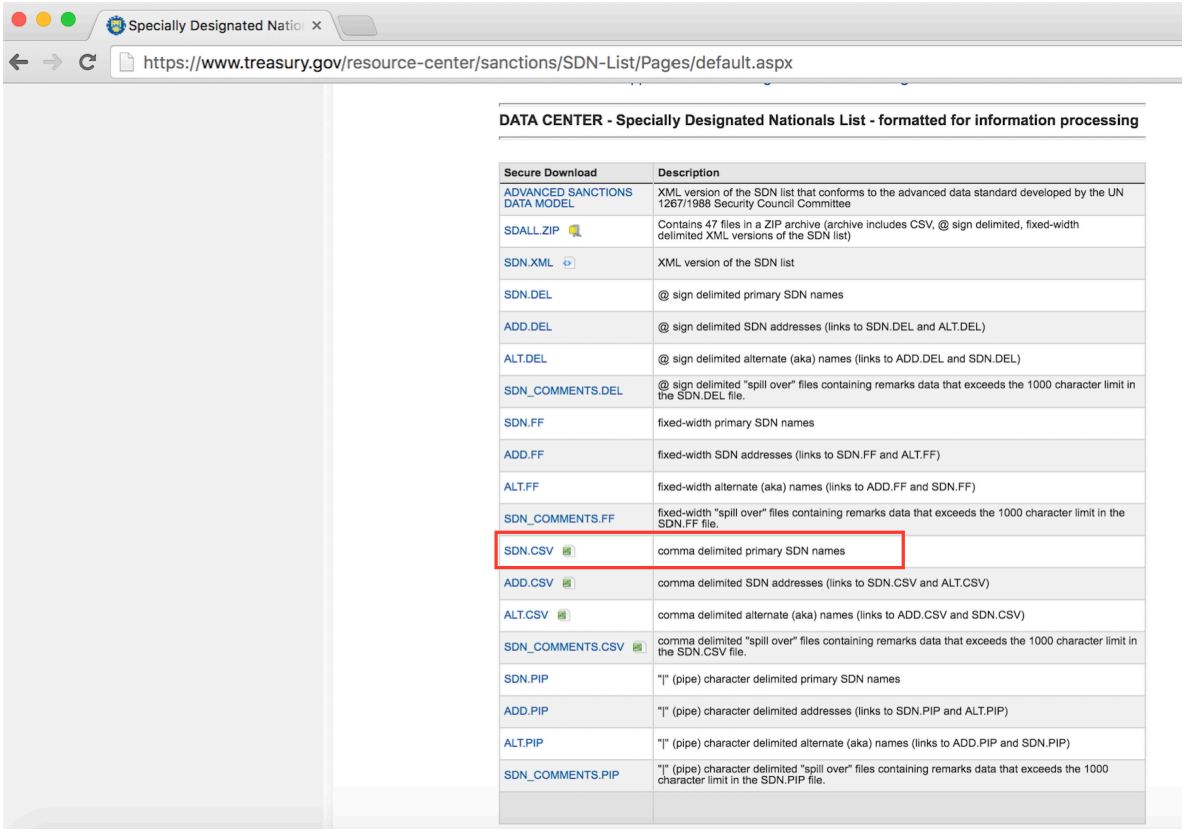
Ilustración 16 - Página principal de descarga de la lista de OFAC (SDN).



Fuente: Página web del Departamento del Tesoro de Estados Unidos. (OFAC SDN List, 2016).

y en la Ilustración 17 se hace referencia a los distintos tipos de listas que pueden ser encontradas en esta página web y cuál fue la lista seleccionada para las comparaciones.

Ilustración 17 - Tipo de archivos disponibles para descarga en la página de OFAC.



Fuente: Página web del Departamento del Tesoro de Estados Unidos. (OFAC SDN List, 2016).

La razón por la que se optó por seleccionar el archivo SDN.csv para descarga es porque este contiene nombres catalogados como de alto riesgo para hacer negocios con ellos, tanto de empresas como de personas. Así mismo vienen en el formato necesario para poder llevar a cabo su procesamiento (mismo formato que el archivo de nombres a comparar).

3.3.2.3 Limpieza y Estandarización de Datos

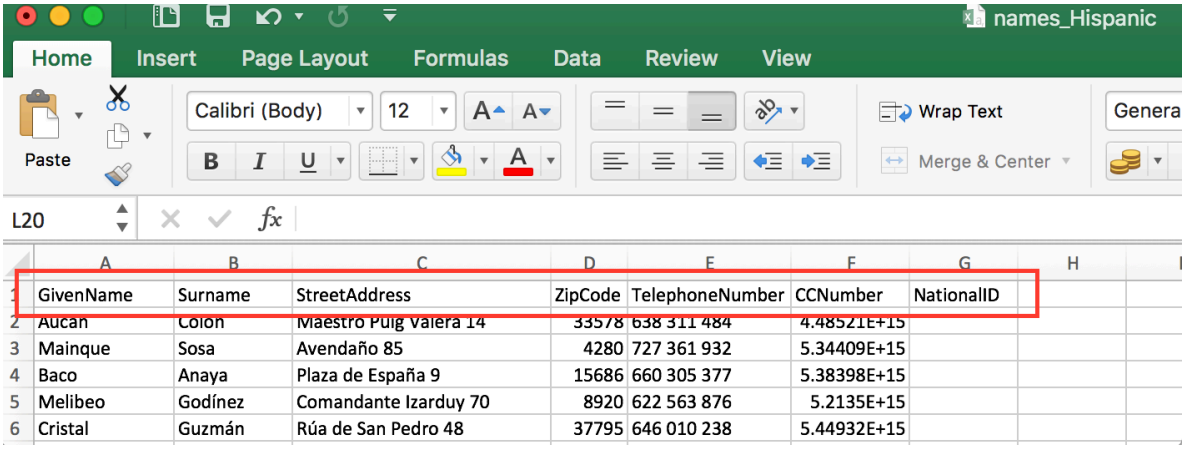
Una vez obtenidos los datos en formato .csv, era necesario estandarizarlos para su uso posterior con los algoritmos fonéticos. Para este fin y dado el formato inicial de la lista de nombres generada con el sitio *Web Fake Name Generator,* en el cual tanto los apellidos como los nombres se encontraban separado por comas, y en comparación con la lista SDN que se encuentra estandarizada para su utilización

con los nombres y apellidos como un solo campo, fue necesario adaptar la lista de nombre a comparar.

Dado que la lista de SDN está diseñada en un formato que ya es utilizado por miles de empresas a nivel mundial, resultaba coherente adaptar la lista de nombres a evaluar, la limpieza de datos contemplo los siguientes elementos:

1. *Limpieza de cabecera sobre el archivo de nombres a evaluar*. **(**Ilustración 18**)**

Ilustración 18 - Ejemplo de limpieza de cabecera en el archivo de nombres a evaluar.

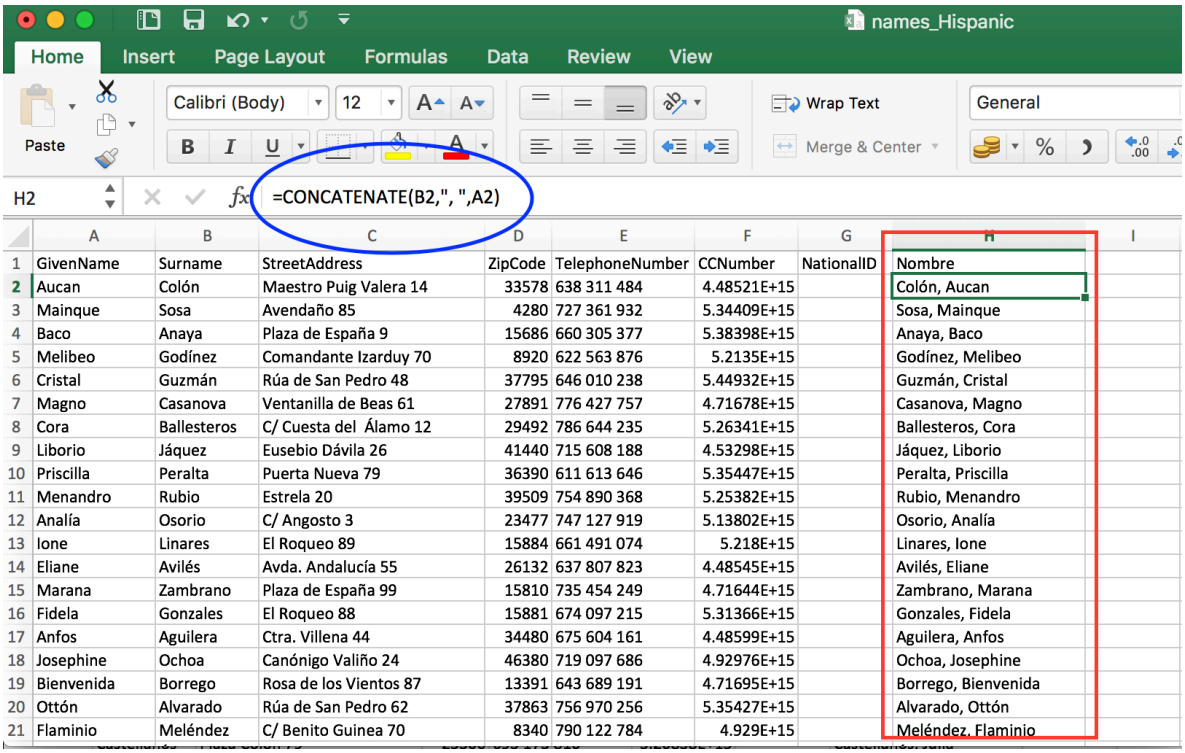


	A	B	C	D	E	F	G	H
1	GivenName	Surname	StreetAddress	ZipCode	TelephoneNumber	CCNumber	NationalID	
2	Aucan	Colon	Maestro Puig Valera 14	33578	638 311 484	4.48521E+15		
3	Mainque	Sosa	Avendaño 85	4280	727 361 932	5.34409E+15		
4	Baco	Anaya	Plaza de España 9	15686	660 305 377	5.38398E+15		
5	Melibeo	Godínez	Comandante Izarduy 70	8920	622 563 876	5.2135E+15		
6	Cristal	Guzmán	Rúa de San Pedro 48	37795	646 010 238	5.44932E+15		

Fuente: Elaboración Propia.

2. *Unión en un solo campo del nombre y el apellido*. **(**Ilustración 19**)**

Ilustración 19 - Ejemplo donde se muestra como juntar los campos de nombre y apellido en una sola columna por medio de la función CONCATENATE de Excel.

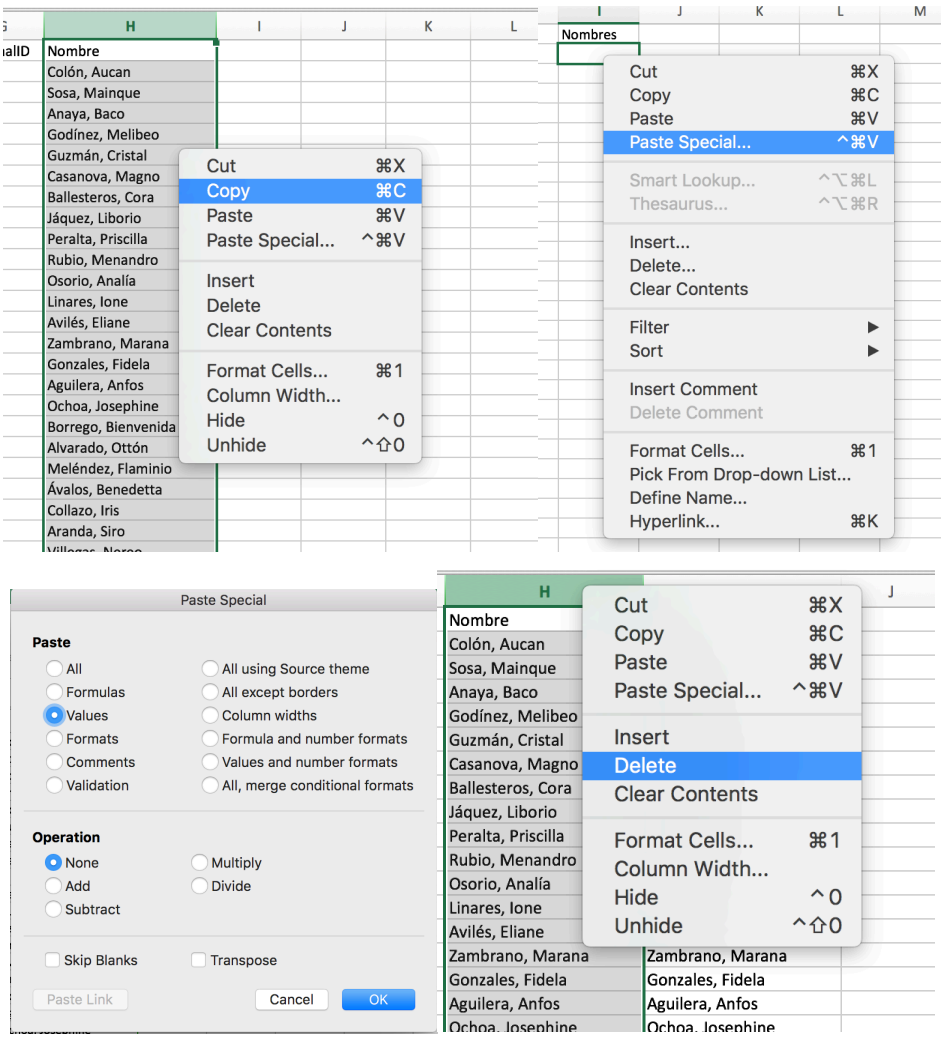


	A	B	C	D	E	F	G	H	I
1	GivenName	Surname	StreetAddress	ZipCode	TelephoneNumber	CCNumber	NationalID	Nombre	
2	Aucan	Colón	Maestro Puig Valera 14	33578	638 311 484	4.48521E+15		Colón, Aucan	
3	Mainque	Sosa	Avendaño 85	4280	727 361 932	5.34409E+15		Sosa, Mainque	
4	Baco	Anaya	Plaza de España 9	15686	660 305 377	5.38398E+15		Anaya, Baco	
5	Melibeo	Godínez	Comandante Izarduy 70	8920	622 563 876	5.2135E+15		Godínez, Melibeo	
6	Cristal	Guzmán	Rúa de San Pedro 48	37795	646 010 238	5.44932E+15		Guzmán, Cristal	
7	Magno	Casanova	Ventanilla de Beas 61	27891	776 427 757	4.71678E+15		Casanova, Magno	
8	Cora	Ballesteros	C/ Cuesta del Álamo 12	29492	786 644 235	5.26341E+15		Ballesteros, Cora	
9	Liborio	Jáquez	Eusebio Dávila 26	41440	715 608 188	4.53298E+15		Jáquez, Liborio	
10	Priscilla	Peralta	Puerta Nueva 79	36390	611 613 646	5.35447E+15		Peralta, Priscilla	
11	Menandro	Rubio	Estrela 20	39509	754 890 368	5.25382E+15		Rubio, Menandro	
12	Analía	Osorio	C/ Angosto 3	23477	747 127 919	5.13802E+15		Osorio, Analía	
13	Ione	Linares	El Roqueo 89	15884	661 491 074	5.218E+15		Linares, Ione	
14	Eliane	Avilés	Avda. Andalucía 55	26132	637 807 823	4.48545E+15		Avilés, Eliane	
15	Marana	Zambrano	Plaza de España 99	15810	735 454 249	4.71644E+15		Zambrano, Marana	
16	Fidela	Gonzales	El Roqueo 88	15881	674 097 215	5.31366E+15		Gonzales, Fidela	
17	Anfos	Aguilera	Ctra. Villena 44	34480	675 604 161	4.48599E+15		Aguilera, Anfos	
18	Josephine	Ochoa	Canónigo Valiño 24	46380	719 097 686	4.92976E+15		Ochoa, Josephine	
19	Bienvenida	Borrego	Rosa de los Vientos 87	13391	643 689 191	4.71695E+15		Borrego, Bienvenida	
20	Ottón	Alvarado	Rúa de San Pedro 62	37863	756 970 256	5.35427E+15		Alvarado, Ottón	
21	Flaminio	Meléndez	C/ Benito Guinea 70	8340	790 122 784	4.929E+15		Meléndez, Flaminio	

Fuente: Elaboración Propia.

3. *Utilización de las variables ya sin formulas en Excel, para su transformación a formato .csv, y luego la eliminación de esa columna con fórmulas, para dejar solo los valores de nombres ya conjuntados con los apellidos.* **(**Ilustración 20**)**

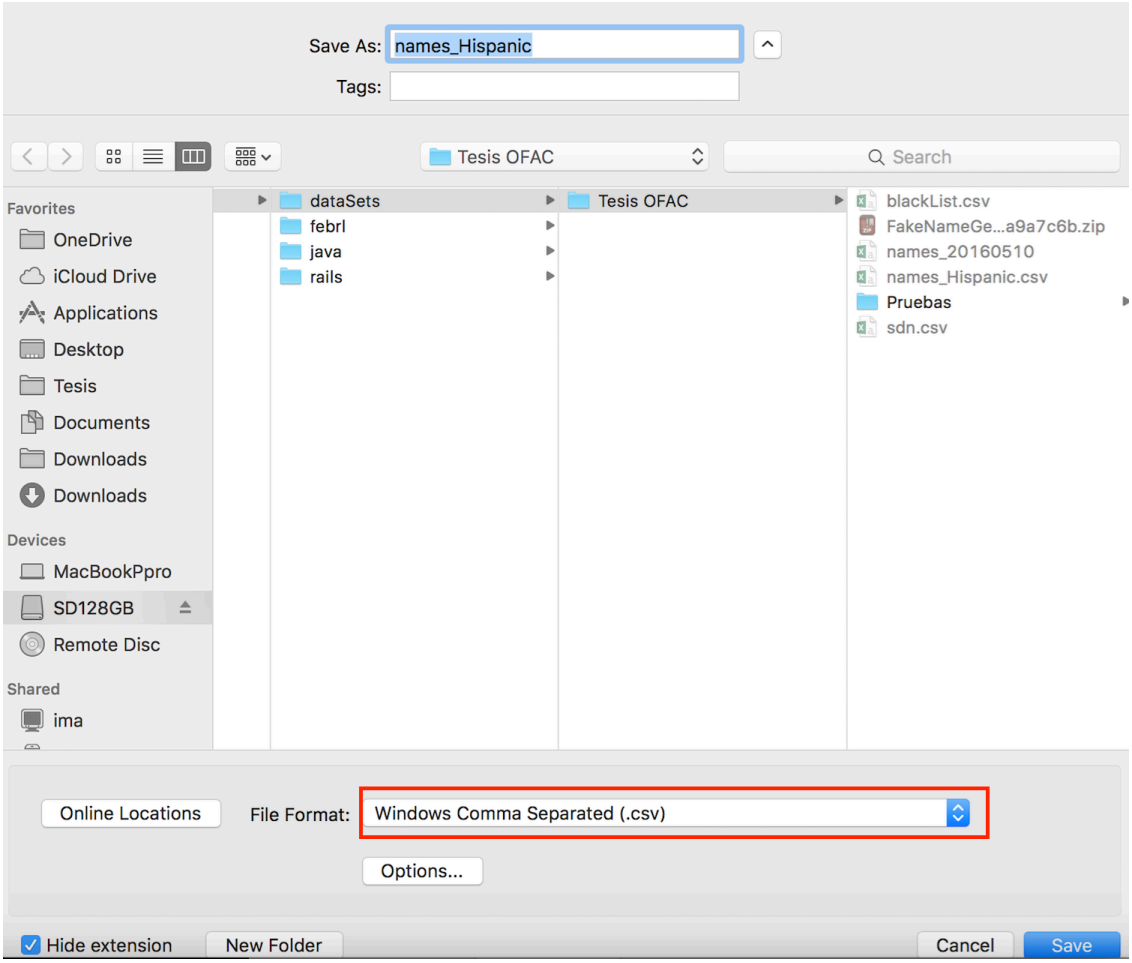
Ilustración 20 - Ejemplo de cómo pegar los valores ya sin formula en la columna de "nombres", para que posteriormente el archivo pueda ser guardado con extensión .csv.



Fuente: Elaboración Propia.

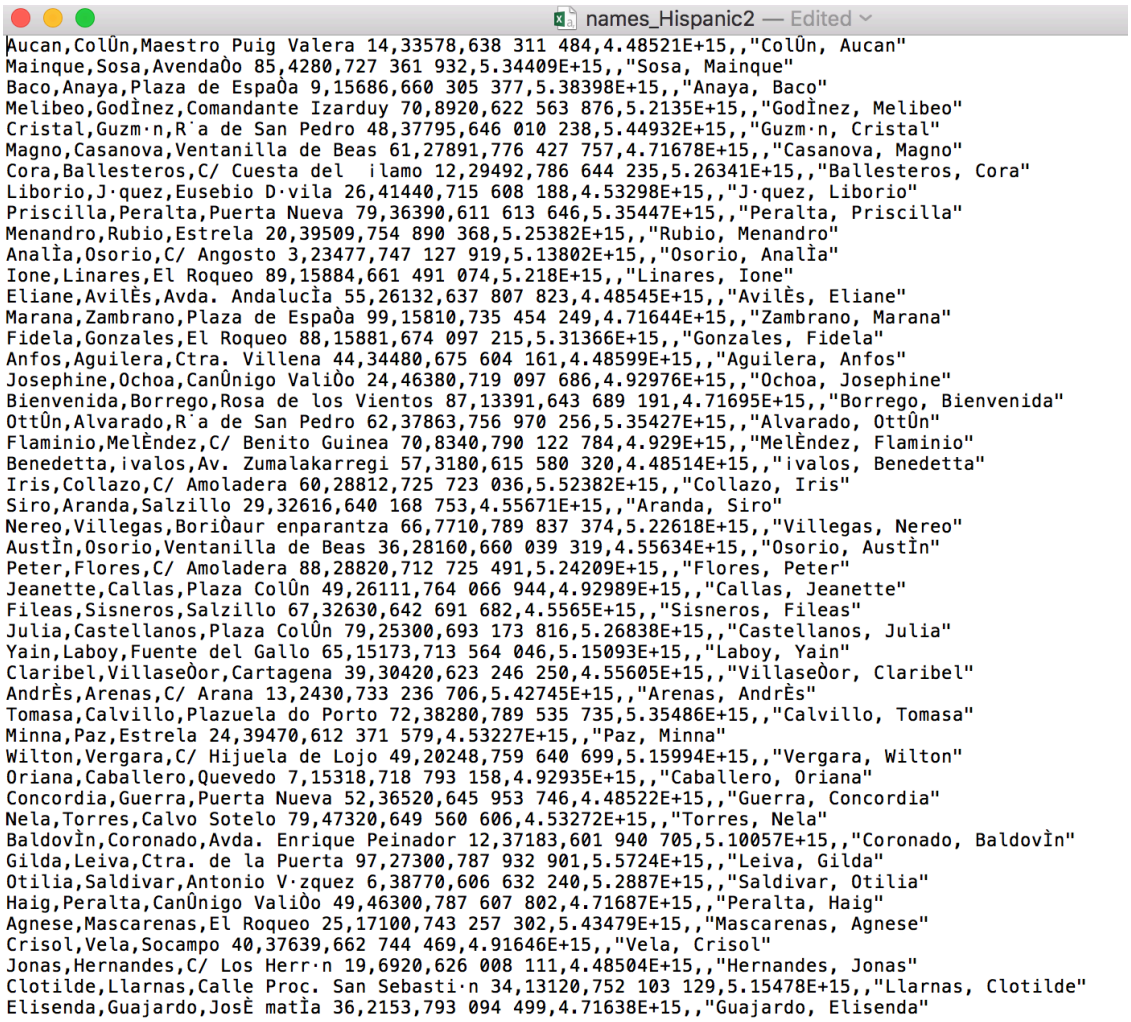
4. *Guardar el archivo en formato .csv*. **(**Ilustración 21 **e** Ilustración 22**)**

Ilustración 21 - Ejemplo de pantalla donde se guarda el archivo final con extensión .csv.



Fuente: Elaboración Propia.

Ilustración 22 - Ejemplo de archivo con extensión .csv abierto desde un editor de texto.



Fuente: Elaboración Propia.

3.3.3.4 Implementación de los Algoritmos Fonéticos

Para poder analizar la información con los algoritmos fonéticos, era necesario contar con algún software con el cual se pudiera hacer uso de ellos y al mismo tiempo que permitiese tener una salida acorde para poder analizar los datos.

3.3.3.4.1 Librería de Java con las Implementaciones de los Algoritmos Fonéticos

Los paquetes que se utilizaron para el análisis de los datos, están contenidos en la librería de Java llamada *org.apache.commons.codec.language,* (Apache, 2014).

Este paquete cuenta con los siguientes codificadores fonéticos señalados en la Tabla 15, de los cuales se utilizaron las implementaciones de soundex, NYSIIS y Methaphone, al ser estos tres algoritmos la base de más implementaciones y de los algoritmos más utilizados para la búsqueda de coincidencias fonéticas.

Tabla 15 - Tabla que contiene las clases del paquete de codificación fonética de java.

Clase	Descripción
AbstractCaverphone	Codifica una cadena a tipo Caverphone
Caverphone	No se usa más
	La versión 1.5 será removida y ahora se usará la versión 2.0
Caverphone1	Codifica una cadena a tipo Caverphone 1.0
Caverphone2	Codifica una cadena a tipo Caverphone 2.0
ColognePhonetic	Codifica una cadena a tipo Cologne Phonetic.
DaitchMokotoffSoundex	Codifica una cadena a tipo Soundex de Daitch Mokotoff.
DoubleMetaphone	Codifica una cadena a tipo Doble Metaphone.
MatchRatingApproachEncoder	Algoritmo desarrollado por la Aerolínea Western.
Metaphone	Codifica una cadena a tipo Metaphone
Nysiis	Codifica una cadena a tipo NYSIIS
RefinedSoundex	Codifica una cadena a tipo Soundex Refinado
Soundex	Codifica una cadena a tipo Soundex

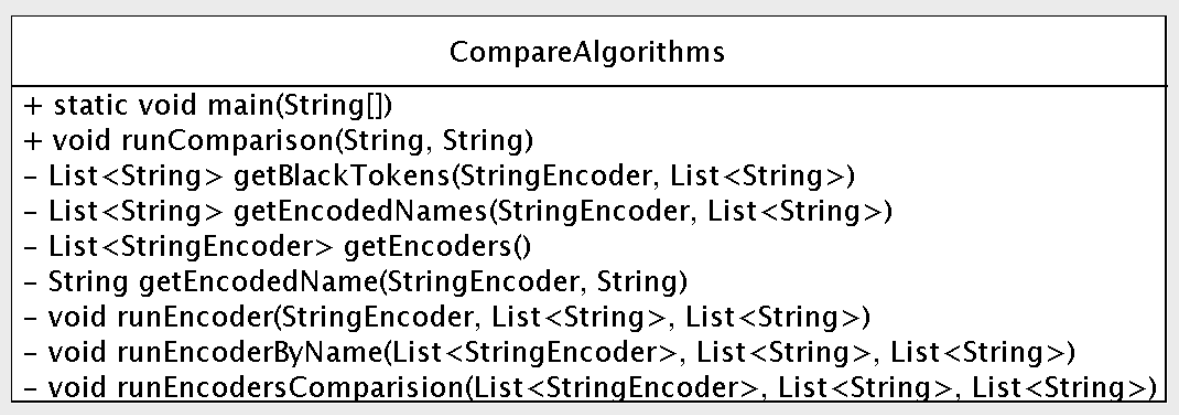
Fuente: Paquetes de Apache Org. (Apache, 2014).

3.3.3.4.2 Implementación de los Algoritmos Fonéticos

Para poder hacer uso del paquete de java que contiene la implementación de los algoritmos fonéticos, fue necesario desarrollar un código que permitiese leer como datos de entrada la lista de nombres a evaluar y la lista de sdn.

Este código fue desarrollado usando la plataforma de desarrollo llamada Eclipse Spring (Pivotal, 2016) y consta de un archivo principal donde se encuentra la implementación del código y dos archivos secundarios que apoyan en la medición de tiempo de ejecución y la creación de un arreglo que contiene la lista de nombres a evaluar, una lista de los codificadores a utilizar (esta implementación permite utilizar todo la pila de algoritmos fonéticos que contiene el paquete de implementación principal) y la lista negra de nombres (SDN). En la Ilustración 23 se muestra la clase principal que hace uso de los algoritmos fonéticos.

Ilustración 23 - Representación en UML del algoritmo elaborado para poder utilizar los algoritmos fonéticos en java.



Fuente: Elaboración Propia.

La salida de la implementación que hace uso de los algoritmos fonéticos contiene 2 datos principales, que permitieron recolectar datos para analizarlos. Por un lado, se obtiene el tiempo de comparación por cada uno de los algoritmos y, por otro lado, un archivo de salida con extensión .csv, que contiene la siguiente estructura de archivo (Ilustración 24).

Ilustración 24 - Ejemplo de salida del algoritmo que implementa las clases fonéticas para comparar el archivo de nombres a evaluar con la lista SDN.

	A	B	C
1	GALINDO MARTINEZ, Fernando Alberto	KLNT	FALSE
2	HERRENO BARRERA, Alejandro	HRNB	FALSE
3	MOLANO TORRES, Deysi Yamile	MLNT	TRUE

Fuente: Elaboración Propia.

La primera columna contiene el nombre que se buscó en la lista SDN, la segunda columna contiene la codificación fonética correspondiente al algoritmo seleccionado y la tercera columna la incidencia, siendo esta verdadera o falsa según el resultado de la búsqueda.

3.3.3.4.3 Estandarización de Salida de Datos de Algoritmo

Para poder utilizar la información obtenida del algoritmo que implementa la búsqueda fonética, fue necesario formatear esta información y al mismo tiempo empezar a obtener las incidencias o variables para el futuro análisis, es por esta razón que se hicieron las siguientes adecuaciones al archivo que contenía la salida de la aplicación señalados en la Ilustración 25.

Ilustración 25 - Salida de la aplicación que implementa los algoritmos fonéticos, organizando la información por tipo de algoritmo utilizado

B	C	D	E	F	G	H	I	J
Codificación	Incidencia	Cadena Comparada	Método de Codificación	Corrida	Falso Positivo	Falso Negativo	Positivos Verdaderos	Negativos Verdaderos
KSMN	1	GUZMAN LOERA, Joaquin	Metaphone	1	0	0	1	0
KHNL	0	NOT FOUND	Metaphone	1	0	1	0	0
HTMR	0	NOT FOUND	Metaphone	1	0	0	0	1
NJLS	0	NOT FOUND	Metaphone	1	0	0	0	1
MRMN	1	MIRAMONTES GUTIERREZ, Ofelia Margarita	Metaphone	1	1	0	0	0
PLRL	0	NOT FOUND	Metaphone	1	0	0	0	1
BRNH	0	NOT FOUND	Metaphone	1	0	1	0	0
RTNK	1	RUDENKO, Miroslav Vladimirovich	Metaphone	1	1	0	0	0
SBRS	1	ESSABAR, Zakarya	Metaphone	1	1	0	0	0

Fuente: Elaboración Propia.

Donde se agregaron al archivo de análisis los siguientes campos:

Cadena Comparada: Este campo contiene dos distintos valores según si el algoritmo pudo encontrar o el nombre a buscar dentro de la lista SDN. Si pudo encontrarlo contendrá la cadena localizada, si no lo encontró mostrará un mensaje de "NOT FOUND".

Método de Codificación: Este campo se encarga de señalar que método se usó para realizar la comparación.

Corrida: Para efecto de tener un muestreo adecuado de datos se realizaron 15 pruebas con 20 datos cada una de ellas, este campo (en el archivo de Excel la columna F) señala en que prueba se obtuvo el resultado señalado.

Falso Positivo: Incidencia de falsos positivos en la búsqueda de nombres

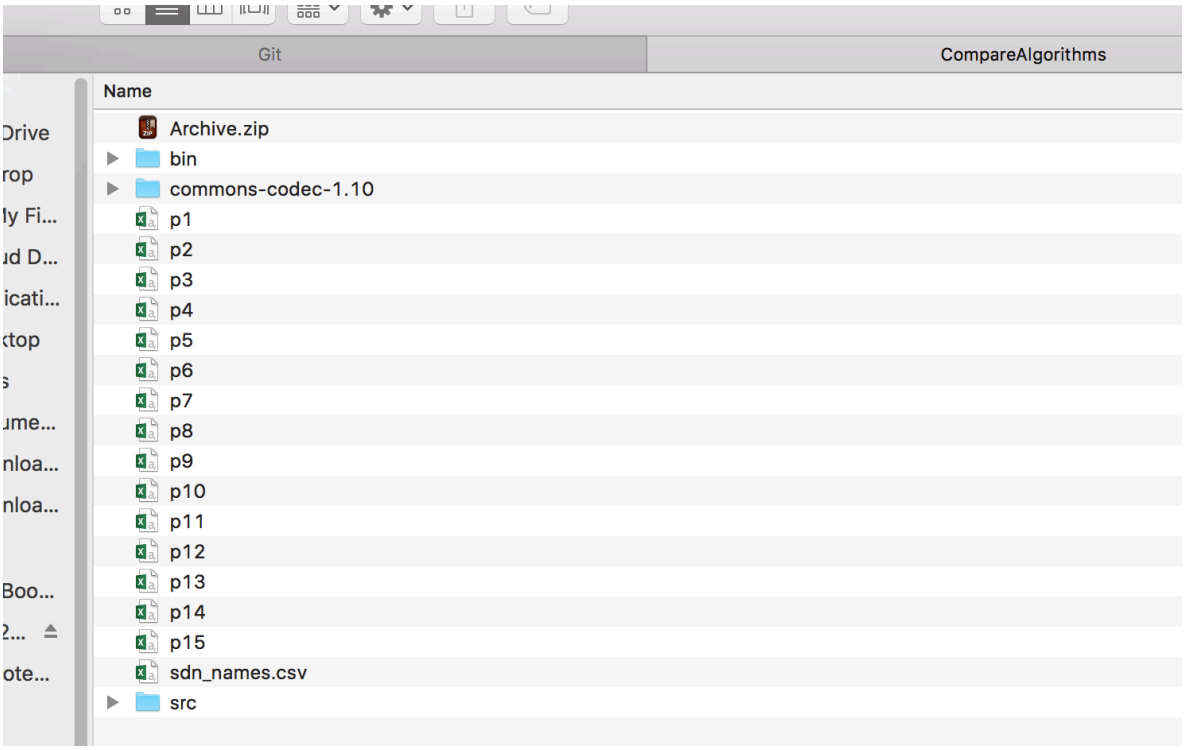
Falso Negativo: Valor que determina cuando un elemento del conjunto de datos fue falsamente clasificado como "no encontrado, o NOT FOUND".

Positivos Verdaderos: Valor que es en efecto una coincidencia legitima.

Negativos Verdaderos: Elemento que fue clasificado de forma correcta como "no encontrado, o NOT FOUND".

Tiempo de ejecución: Se tomaron mediciones de tiempo al inicio y termino de la comparativa de los datos. Para este fin cada conjunto de datos de 20 elementos se dividió en 15 archivos diferentes y se cargaron al inicio del programa, esto con el fin de que el proceso de carga de información inicial no interfiriese en los tiempos de ejecución iniciales y arrojara valores distintos a las subsecuentes evaluaciones. Estos ajustes se muestran a continuación en la Ilustración 26 e Ilustración 27.

Ilustración 26 - División de los conjuntos de datos por archivos



Fuente: Elaboración Propia.

Ilustración 27 - Método de carga de los archivos al inicio la ejecución.

CompareAlgorithms.java

```
1  package test;
2
3  import java.util.ArrayList;
17
18 public class CompareAlgorithms {
19
20     public static void main(String[] args) throws EncoderException {
21         CompareAlgorithms comAlg = new CompareAlgorithms();
22         Map<String, String> files = new HashMap<>();
23         files.put("p1.csv", "sdn_names.csv");
24         files.put("p2.csv", "sdn_names.csv");
25         files.put("p3.csv", "sdn_names.csv");
26         files.put("p4.csv", "sdn_names.csv");
27         files.put("p5.csv", "sdn_names.csv");
28         files.put("p6.csv", "sdn_names.csv");
29         files.put("p7.csv", "sdn_names.csv");
30         files.put("p8.csv", "sdn_names.csv");
31         files.put("p9.csv", "sdn_names.csv");
32         files.put("p10.csv", "sdn_names.csv");
33         files.put("p11.csv", "sdn_names.csv");
34         files.put("p12.csv", "sdn_names.csv");
35         files.put("p13.csv", "sdn_names.csv");
36         files.put("p14.csv", "sdn_names.csv");
37         files.put("p15.csv", "sdn_names.csv");
38         for (String peopleToCheck : files.keySet()) {
39             comAlg.runComparison(peopleToCheck, files.get(peopleToCheck));
40         }
41     }
```

Fuente: Elaboración Propia.

Una vez obtenido los valores mostrados en la Ilustración 25, era necesario obtener los valores que se necesitarían para el análisis de media que se llevaría a cabo subsecuentemente. Para este fin se hicieron los cálculos de las variables o medidas de calidad revisadas en el marco teórico (véase Capitulo 2.5, Medidas de Calidad de Enlace de Datos).

1. Exactitud
2. Precisión
3. Sensibilidad
4. Tasa de Falsos Positivos

5. Tiempo de Ejecución
6. Especificidad
7. F-Measure

3.3.3.4.4 Determinación del índice de Falsos Positivos

Para poder determinar el índice de falsos positivos se revisó manualmente en la lista de la OFAC los nombres tomados para las muestras, para ver cuál de ellos efectivamente estaba identificado como tal y cual era un falso positivo por su evaluación fonética.

Con esta comparativo manual se pudo obtener los valores de verdaderos positivos, falsos positivos y por extrapolación los verdaderos negativos y falsos negativos.

3.3.3.5 Metodología para el Análisis de las Pruebas

3.3.3.5.1 Determinación de las Variables Estadísticas

La variable dependiente está determinada por el número de aciertos o coincidencias fonéticas verdaderas.

La variable independiente está determinada por el tipo de algoritmo utilizado, este afecta a la variable dependiente conforme se utilice un algoritmo distinto de codificación fonética.

3.3.3.5.2 Determinación del Método Estadístico

Para poder llevar a cabo el experimento se utilizó el método estadístico de ANOVA, para la comparación de las medias de las distintas variables a evaluar (exactitud, precisión, sensibilidad, tasa de falsos positivos y tiempo de ejecución). Se crearon 3 grupos de datos con los resultados de las comparaciones entre algoritmos. La comparación posterior se llevó a cabo con la prueba de DUNCAN (Valdivieso, Valdivieso, & Valdivieso, 2011), esto con el fin de determinar en los casos en los que hubo diferencia en las medias, cual fue la causa, a esto se le conoce como análisis posterior.

Se seleccionaron 15 grupos de 20 nombres para poder tener una muestra representativa y un análisis de varianza que permita contrastar la hipótesis

correctamente. El experimento con la prueba de la ANOVA se describe a continuación:

Prueba de análisis de medias (ANOVA)

El experimento tiene el siguiente modelo estadístico.

$$H_0 = \mu_1 = \mu_2 = \mu_3 = \mu_k$$

$$H_a = \text{por lo menos 2 } \mu_s \text{ diferentes}$$

Donde $\mathbf{H_0}$ es la definición de la hipótesis nula, $\mathbf{H_a}$ es la hipótesis alternativa, $\boldsymbol{\mu}$ es la media de los aciertos por grupo de algoritmos comparados, **k**, es el número de grupos evaluados, en este caso 3.

Si la significancia de **F** es menor que 0.05 se RECHAZA $\mathbf{H_0}$ y se dice que no hay suficiente evidencia para decir que existe una diferencia entre los 3 algoritmos fonéticos para la búsqueda de nombres.

Cada uno de los 15 grupos de 20 nombres se probó corriendo las comparaciones con los algoritmos fonéticos y la lista de OFAC, los datos arrojados por el algoritmo fueron capturados para su posterior uso con las pruebas estadísticas. Véase la Tabla 16 para los valores del algoritmo Soundex, la Tabla 17 para los valores del algoritmo Metaphone y los valores mostrados en la Tabla 18 para el algoritmo NYSIIS.

Tabla 16 - Conjunto de datos estandarizados y homologados del algoritmo Soundex, conteniendo las variables a evaluar

Soundex											
Corridas	FP	FN	TP	TN	Tiempo (ms)	Exactitud	Presición	Sensibilidad	Tasa de Falsos Positivos	Especificidad	F-Measure
1	8	0	5	7	2	0.6	0.4	1.0	0.5	0.5	0.6
2	8	0	6	6	2	0.6	0.4	1.0	0.6	0.4	0.6
3	4	5	4	7	3	0.55	0.5	0.4	0.4	0.6	0.5
4	8	3	2	7	4	0.45	0.2	0.4	0.5	0.5	0.3
5	8	0	3	9	2	0.6	0.3	1.0	0.5	0.5	0.4
6	11	2	2	5	2	0.35	0.2	0.5	0.7	0.3	0.2
7	9	1	7	3	6	0.5	0.4	0.9	0.8	0.3	0.6
8	9	1	2	8	4	0.5	0.2	0.7	0.5	0.5	0.3
9	11	0	6	3	2	0.45	0.4	1.0	0.8	0.2	0.5
10	10	0	7	3	5	0.5	0.4	1.0	0.8	0.2	0.6
11	8	2	3	7	3	0.5	0.3	0.6	0.5	0.5	0.4
12	12	1	0	7	6	0.35	0.0	0.0	0.6	0.4	0.0
13	7	3	3	7	4	0.5	0.3	0.5	0.5	0.5	0.4
14	8	2	7	3	8	0.5	0.5	0.8	0.7	0.3	0.6
15	5	2	4	9	4	0.65	0.4	0.7	0.4	0.6	0.5

Fuente: Elaboración Propia.

Tabla 17 - Conjunto de datos estandarizados y homologados del algoritmo Metaphone, conteniendo las variables a evaluar

Metaphone											
Corridas	FP	FN	TP	TN	Tiempo (ms)	Exactitud	Presición	Sensibilidad	Tasa de Falsos Positivos	Especificidad	F-Measure
1	5	7	1	7	6	0.4	0.17	0.13	0.42	0.58	0.14
2	3	9	2	6	21	0.4	0.40	0.18	0.33	0.67	0.25
3	4	7	2	7	4	0.45	0.33	0.22	0.36	0.64	0.27
4	4	8	1	7	10	0.4	0.20	0.11	0.36	0.64	0.14
5	4	6	1	9	3	0.5	0.20	0.14	0.31	0.69	0.17
6	4	6	5	5	15	0.5	0.56	0.45	0.44	0.56	0.50
7	7	9	1	3	19	0.2	0.13	0.10	0.70	0.30	0.11
8	4	7	1	8	7	0.45	0.20	0.13	0.33	0.67	0.15
9	7	7	3	3	4	0.3	0.30	0.30	0.70	0.30	0.30
10	6	8	3	3	18	0.3	0.33	0.27	0.67	0.33	0.30
11	4	7	2	7	9	0.45	0.33	0.22	0.36	0.64	0.27
12	8	3	2	7	13	0.45	0.20	0.40	0.53	0.47	0.27
13	4	8	1	7	15	0.4	0.20	0.11	0.36	0.64	0.14
14	6	9	2	3	18	0.25	0.25	0.18	0.67	0.33	0.21
15	3	5	3	9	14	0.6	0.50	0.38	0.25	0.75	0.43

Fuente: Elaboración Propia.

Tabla 18 - Conjunto de datos estandarizados y homologados del algoritmo NYSIIS, conteniendo las variables a evaluar

NYSIIS											
Corridas	FP	FN	TP	TN	Tiempo (ms)	Exactitud	Presición	Sensibilidad	Tasa de Falsos Positivos	Especificidad	F-Measure
1	1	10	2	7	17	0.45	0.67	0.17	0.13	0.88	0.27
2	1	12	1	6	22	0.35	0.50	0.08	0.14	0.86	0.13
3	2	10	1	7	17	0.4	0.33	0.09	0.22	0.78	0.14
4	0	10	3	7	39	0.5	1.00	0.23	0.00	1.00	0.38
5	1	8	2	9	14	0.55	0.67	0.20	0.10	0.90	0.31
6	2	12	1	5	16	0.3	0.33	0.08	0.29	0.71	0.13
7	2	14	1	3	65	0.2	0.33	0.07	0.40	0.60	0.11
8	1	9	2	8	42	0.5	0.67	0.18	0.11	0.89	0.29
9	2	11	4	3	16	0.35	0.67	0.27	0.40	0.60	0.38
10	1	12	4	3	32	0.35	0.80	0.25	0.25	0.75	0.38
11	0	11	2	7	33	0.45	1.00	0.15	0.00	1.00	0.27
12	1	12	0	7	57	0.35	0.00	0.00	0.13	0.88	0.00
13	0	12	1	7	23	0.4	1.00	0.08	0.00	1.00	0.14
14	2	12	3	3	54	0.3	0.60	0.20	0.40	0.60	0.30
15	0	6	5	9	19	0.7	1.00	0.45	0.00	1.00	0.63

Fuente: Elaboración Propia.

Prueba posterior de análisis de medias (DUNCAN)

La prueba de rango múltiple de DUNCAN se utilizó para poder determinar la diferencia entre pares de medias después de que se ha rechazado la hipótesis nula (H_0) en el análisis de la varianza. Esta diferencia indicó que conjunto de datos tenían medias similares, y que conjunto de datos eran diferentes y que valores representaban esas diferencias.

Todas estas evaluaciones se llevaron a cabo con el programa SPSS de IBM, utilizando la prueba de medias y después la prueba de ANOVA de una vía con un análisis posterior de DUNCAN.

3.3.3.6 Análisis de Resultados

El análisis de resultados se describe en el capítulo 4. En este paso se toman los resultados de las corridas de los algoritmos, se hacen los análisis de varianzas y se prepara la información para formular las conclusiones.

CAPÍTULO 4: RESULTADOS

En este capítulo se muestran los resultados obtenidos de los análisis estadísticos de las diferentes variables consideradas.

4.1 Prueba de ANOVA de una Vía con un Análisis Posterior de DUNCAN

Si consideramos las siguientes hipótesis evaluando las medias de los valores de exactitud, precisión, sensibilidad, especificidad, tasa de falsos positivos, tiempo de ejecución y armonía entre la precisión y la sensibilidad (F-Measure), de los algoritmos fonéticos soundex, metaphone y NYSIIS, se puede decir que:

$$H_0 = \mu_1 = \mu_2 = \mu_3 = \mu_k$$

$$H_a = \text{por lo menos 2 } \mu_s \text{ diferentes}$$

4.1.1 Exactitud

En la prueba de ANOVA se puede observar que el valor de la significancia (Sig.) de F es menor a 0.05 (0.016, ver la Ilustración 28), por lo tanto, se rechaza H_0.

En el análisis posterior utilizando la prueba de DUNCAN vemos que no hay diferencia significativa entre el algoritmo metaphone y NYSIIS, pero si hay diferencia significativa con el algoritmo soundex. Se concluye que el algoritmo soundex tiene un valor más grande de *exactitud*, y que los algoritmos metaphone y NYSIIS son parecidos en este rubro, y por lo tanto podemos afirmar que el algoritmo soundex es el algoritmo con un mejor comportamiento al momento de detectar incidencias (de cualquier tipo, verdaderas o falsas) en una búsqueda de nombres.

Ilustración 28 - Resultado arrojado por la prueba de ANOVA sobre la comparación de medias de la exactitud de los algoritmos fonéticos evaluados

```
ONEWAY Exactitud BY Metodo
  /MISSING ANALYSIS
  /POSTHOC=DUNCAN ALPHA(0.05).
```

Oneway

```
[DataSet1] /Volumes/SD128GB/home/KarOn/Lab/Repository/Git/dataSets/Tesis OFAC/Prueba_2.sav
```

ANOVA

Exactitud (Accuracy)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.100	2	.050	4.554	.016
Within Groups	.463	42	.011		
Total	.563	44			

Post Hoc Tests

Homogeneous Subsets

Exactitud (Accuracy)

Duncan[a]

Método Fonético Utilizado	N	Subset for alpha = 0.05	
		1	2
Metaphone	15	.4033	
NYSIIS	15	.4100	
Soundex	15		.5067
Sig.		.863	1.000

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 15.000.

Fuente: Elaboración Propia.

4.1.2 Precisión

En la prueba de ANOVA se puede observar que el valor de la significancia (Sig.) de F es menor a 0.05 (0.000, ver la Ilustración 29), por lo tanto, se rechaza H_0.

En el análisis posterior utilizando la prueba de DUNCAN vemos que no hay diferencia significativa entre el algoritmo metaphone y soundex, pero si hay diferencia significativa con el algoritmo NYSIIS. Se concluye que el algoritmo NYSIIS tiene un valor más grande de *precisión*, y que los algoritmos metaphone y soundex

son parecidos en este rubro y por lo tanto se puede afirmar que el algoritmo NYSIIS tiene el mejor desempeño de los algoritmos evaluados en esta prueba para la detección de coincidencias verdaderas, considerando incidencias verdaderas positivas e incidencias verdaderas negativas.

Ilustración 29 - Resultado arrojado por la prueba de ANOVA sobre la comparación de medias de la precisión de los algoritmos fonéticos evaluados

```
ONEWAY Precision BY Metodo
  /MISSING ANALYSIS
  /POSTHOC=DUNCAN ALPHA(0.05).
```

Oneway

ANOVA

Precisión (Positive Predictive Value)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1.110	2	.555	13.457	.000
Within Groups	1.732	42	.041		
Total	2.842	44			

Post Hoc Tests

Homogeneous Subsets

Precisión (Positive Predictive Value)

Duncan[a]

Método Fonético Utilizado	N	Subset for alpha = 0.05	
		1	2
Metaphone	15	.2867	
Soundex	15	.3267	
NYSIIS	15		.6380
Sig.		.592	1.000

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 15.000.

Fuente: Elaboración Propia.

4.1.3 Sensibilidad

En la prueba de ANOVA se puede observar que el valor de la significancia (Sig.) de F es menor a 0.05 (0.000, ver la Ilustración 30), por lo tanto, se rechaza H_0.

En el análisis posterior utilizando la prueba de DUNCAN vemos que no hay diferencia significativa entre el algoritmo metaphone y NYSIIS, pero si hay diferencia significativa con el algoritmo soundex. Se concluye que el algoritmo soundex tiene un valor más grande de *sensibilidad*, y que los algoritmos metaphone y NYSIIS son parecidos en este rubro y por lo se afirma que el algoritmo soundex tiene el mejor desempeño para la detección de valores verdaderos positivos, comparado con NYSIIS y metaphone.

Ilustración 30 - Resultado arrojado por la prueba de ANOVA sobre la comparación de medias de la sensibilidad de los algoritmos fonéticos evaluados

```
SAVE OUTFILE='/Users/karon/Cloud/OneDrive/Documentos/Tesis/Pruebas/Prueba_2.sav'
  /COMPRESSED.
ONEWAY Sensibilidad BY Metodo
  /MISSING ANALYSIS
  /POSTHOC=DUNCAN ALPHA(0.05).
```

Oneway

```
[DataSet1] /Users/karon/Cloud/OneDrive/Documentos/Tesis/Pruebas/Prueba_2.sav
```

ANOVA

Sensibilidad (Recall)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2.583	2	1.291	33.598	.000
Within Groups	1.614	42	.038		
Total	4.197	44			

Post Hoc Tests

Homogeneous Subsets

Sensibilidad (Recall)

Duncan[a]

Método Fonético Utilizado	N	Subset for alpha = 0.05	
		1	2
NYSIIS	15	.1667	
Metaphone	15	.2213	
Soundex	15		.7000
Sig.		.449	1.000

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 15.000.

Fuente: Elaboración Propia.

4.1.4 Especificidad

En la prueba de ANOVA se puede observar que el valor de la significancia (Sig.) de F es menor a 0.05 (0.000, ver la Ilustración 31), por lo tanto, se rechaza H_0.

En el análisis posterior utilizando la prueba de DUNCAN se puede apreciar que existe diferencia significativa entre los 3 algoritmos, y se concluye que el algoritmo NYSIIS tiene un valor más alto de especificad, y por tanto se dice que el algoritmo NYSIIS es el algoritmo con un mejor comportamiento cuando se trata de detectar coincidencias verdaderas negativas.

Ilustración 31 - Resultado arrojado por la prueba de ANOVA sobre la comparación de medias de la especificidad de los algoritmos fonéticos evaluados

```
ONEWAY Especificidad BY Metodo
  /MISSING ANALYSIS
  /POSTHOC=DUNCAN ALPHA(0.05).
```

Oneway

ANOVA

Especificidad

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1.321	2	.661	30.624	.000
Within Groups	.906	42	.022		
Total	2.227	44			

Post Hoc Tests

Homogeneous Subsets

Especificidad

Duncan[a]

Método Fonético Utilizado	N	Subset for alpha = 0.05		
		1	2	3
Soundex	15	.4200		
Metaphone	15		.5473	
NYSIIS	15			.8300
Sig.		1.000	1.000	1.000

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 15.000.

Fuente: Elaboración Propia.

4.1.5 Tasa de Falsos Positivos

En la prueba de ANOVA podemos ver que el valor de la significancia (Sig.) de F es menor a 0.05 (0.000, ver la Ilustración 32), por lo tanto, se rechaza H_0.

En el análisis posterior utilizando la prueba de DUNCAN se puede apreciar que existe diferencia significativa entre los 3 algoritmos, y se concluye que el algoritmo soundex tiene un valor más alto de falsos positivos, por lo que se comprueba que tiene el peor desempeño de los tres algoritmos evaluados, no así, el algoritmo NYSIIS que muestra una tasa de falsos positivos baja comparada con metaphone y soundex y dado este valor se dice que tuvo el mejor desempeño evaluando esta variable.

Ilustración 32 - Resultado arrojado por la prueba de ANOVA sobre la comparación de medias de la tasa de falsos positivos de los algoritmos fonéticos evaluados

```
ONEWAY TasaFalsosPositivos BY Metodo
  /MISSING ANALYSIS
  /POSTHOC=DUNCAN ALPHA(0.05).
```

Oneway

ANOVA

Tasa de Falsos Positivos

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1.348	2	.674	30.199	.000
Within Groups	.937	42	.022		
Total	2.285	44			

Post Hoc Tests

Homogeneous Subsets

Tasa de Falsos Positivos

Duncan[a]

Método Fonético Utilizado	N	Subset for alpha = 0.05		
		1	2	3
NYSIIS	15	.1713		
Metaphone	15		.4527	
Soundex	15			.5867
Sig.		1.000	1.000	1.000

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 15.000.

Fuente: Elaboración Propia.

4.1.5 Tiempo de Ejecución

En la prueba de ANOVA se puede observar que el valor de la significancia (Sig.) de F es menor a 0.05 (0.000, ver la Ilustración 33), por lo tanto, se rechaza H_0.

En el análisis posterior utilizando la prueba de DUNCAN se puede apreciar que existe diferencia significativa entre los 3 algoritmos, y se concluye que el algoritmo soundex es el algoritmo más rápido en cuestión de tiempo de ejecución de los 3 y,

por lo tanto, se afirma que el algoritmo soundex es el algoritmo más rápido en términos de tiempo de ejecución de los 3 algoritmos evaluados.

Ilustración 33 - Resultado arrojado por la prueba de ANOVA sobre la comparación de medias de los tiempos de ejecución de los algoritmos fonéticos evaluados

```
ONEWAY TiempoEjecucion BY Metodo
  /MISSING ANALYSIS
  /POSTHOC=DUNCAN ALPHA(0.05).
```

Oneway

ANOVA

Tiempo de Ejecución de los Algoritmos

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	5900.933	2	2950.467	27.390	.000
Within Groups	4524.267	42	107.721		
Total	10425.200	44			

Post Hoc Tests

Homogeneous Subsets

Tiempo de Ejecución de los Algoritmos

Duncan[a]

Método Fonético Utilizado	N	Subset for alpha = 0.05		
		1	2	3
Soundex	15	3.8000		
Metaphone	15		11.7333	
NYSIIS	15			31.0667
Sig.		1.000	1.000	1.000

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 15.000.

Fuente: Elaboración Propia.

4.1.6 Armonía entre la Precisión y la Sensibilidad (F-Measure)

En la prueba de ANOVA podemos ver que el valor de la significancia (Sig.) de F es menor a 0.05 (0.002, ver la Ilustración 34), por lo tanto, se rechaza H_0.

En el análisis posterior utilizando la prueba de DUNCAN vemos que no hay diferencia significativa entre el algoritmo metaphone y NYSIIS, pero si hay diferencia significativa con el algoritmo soundex. Se concluye que el algoritmo soundex tiene un valor más alto de armonía entre la precisión y la sensibilidad comparado con los algoritmos metaphone y NYSIIS que probaron ser similares en este rubro, por lo que se concluye que el algoritmo soundex tiene un mejor balance entre precisión y sensibilidad en comparación con NYSIIS y metaphone.

Ilustración 34 - Resultado arrojado por la prueba de ANOVA sobre la comparación de medias de la armonía entre la precisión y la sensibilidad de los algoritmos fonéticos evaluados

```
ONEWAY FMeasure BY Metodo
  /MISSING ANALYSIS
  /POSTHOC=DUNCAN ALPHA(0.05).
```

Oneway

ANOVA

F-Measure (Armonía entre la Precision y la Sensitividad

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.336	2	.168	7.470	.002
Within Groups	.946	42	.023		
Total	1.282	44			

Post Hoc Tests

Homogeneous Subsets

F-Measure (Armonía entre la Precision y la Sensitividad

Duncan[a]

Método Fonético Utilizado	N	Subset for alpha = 0.05	
		1	2
Metaphone	15	.2433	
NYSIIS	15	.2573	
Soundex	15		.4333
Sig.		.800	1.000

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 15.000.

Fuente: Elaboración Propia.

4.2 Interpretación de Resultados

Tomando en cuenta los resultados obtenidos de las pruebas estadísticas, se creó la siguiente Tabla 19, en la cual se resaltan los valores más representativos de la comparativa de medias entre los 3 algoritmos fonéticos.

Tabla 19 - Tabla comparativa de resultados de los algoritmos fonéticos evaluados con las variables consideradas para la prueba de medias

Variables	Soundex	Methaphone	NYSIIS
Exactitud	**0.5067**	0.4033	0.41
Presición	0.3267	0.2867	**0.638**
Sensibilidad	**0.7**	0.2213	0.1667
Tasa de Falsos Positivos	0.5867	0.4527	**0.1713**
Especificidad	0.42	0.5473	**0.83**
F-Measure	**0.433**	0.2433	0.2573
Tiempo de Ejecución	**3.8**	11.7333	31.0667

Fuente: Elaboración Propia.

Se observa que el algoritmo soundex tiene un mejor desempeño en cuestión de tiempo de ejecución, superando al algoritmo metaphone por casi 4 veces en velocidad y hasta 10 veces más rápido que el algoritmo NYSIIS.

Si consideramos un alto índice de comparaciones diarias para un banco o una aduana en algún aeropuerto, el algoritmo soundex podría arrojar resultados de una forma más pronta.

Sin embargo, si revisamos el índice de falsos positivos que el algoritmo soundex arroja, vemos que es el más elevado, este presenta 5 veces más falsos positivos que el algoritmo NYSIIS, en donde claramente podemos ver que el algoritmo NYSIIS tiene mejor desempeño, y tomando esto en consideración, NYSIIS es más lento, pero más certero al momento de realizar alguna coincidencia; metaphone presenta resultados muy parecidos a soundex, por lo que en este rubro no destaca.

De nueva cuenta si queremos precisión en la detección de resultados, donde esta precisión toma en consideración, los verdaderos positivos y los verdaderos negativos, el algoritmo NYSIIS tiene los mejores resultados al tener un índice menor de falsos positivos y un factor de predicción de verdaderos positivos más alto (precisión)

Una interpretación general nos dice que el algoritmo NYSIIS nos ayudaría a disminuir la tasa de falsos positivos en comparación con el algoritmo metaphone o soundex, que resultan ser más rápidos, pero menos precisos.

Por último, vemos que el balance armónico entre la precisión y la sensibilidad (tasa de verdaderos positivos) es más alta con el algoritmo soundex, esto nos dice que su factor de predicción de valores verdaderos positivos en comparación con la tasa de verdaderos positivos es alto, en comparación con NYSIIS y metaphone.

Metaphone obtiene los peores resultados en la comparación, al no sobresalir en ningún rubro sobre soundex y NYSIIS. En cambio, NYSIIS destaca en precisión y especificidad (tasa de verdaderos negativos). Soundex por su parte tiene mejores resultados en velocidad de ejecución y armonía entre precisión y sensibilidad, pero no lo hace muy bien al momento de detectar falsos positivos.

CAPÍTULO 5: CONCLUSIONES

Tenemos por un lado un algoritmo que tiene la tasa de falsos positivos más baja de los 3 (NYSIIS), pero que en cuestiones de tiempos de ejecución es 10 veces más lento que el algoritmo más rápido evaluado en esta prueba (soundex), y por otro lado tenemos un algoritmo que si bien no sobresale en ningún rubro en particular (metaphone), no se mantiene lejos de los otros dos algoritmos. En las pruebas estadísticas siempre estuvo cerca de los mejore valores, sin alcanzarlos, pero no se comportó de manera antípoda como lo hiciesen el soundex o el NYSIIS.

Se puede decir que soundex presenta una alternativa más rápida y equilibrada en cuestiones de búsqueda de nombres en español sobre la lista negra de OFAC, pero a cambio maneja un índice de falsos positivos por encima de los otros dos algoritmos evaluados, en cuestiones de cumplimiento, no puede sacrificarse esta variable por ser más rápido en ejecución, por lo que para fines de utilización, no recomendaría su uso para trabajar con una lista de nombres hispanos.

En el experimento vimos que NYSIIS tiene mejor tasa de falsos positivos cuando se procesan listas de nombres en español, al ser la más baja, así mismo su valor de especificidad es el más alto, es decir, tiene una tasa de verdaderos negativos más alta. Con estos dos valores sin considerar el del tiempo, considero que NYSIIS tiene un mejor desempeño en general. El factor de tiempo de ejecución pienso que puede ser mejorado si se emplea en conjunto con algún algoritmo de lógica difusa que ayude a buscar sobre menos campos, solamente los valores identificados como similares por este algoritmo y luego haciendo la evaluación fonética.

Fueron evaluados los objetivos específicos y se llegó a las siguientes conclusiones:

- Se elaboraron las pruebas unitarias de los algoritmos soundex, metaphone y NYSIIS, donde se obtuvieron datos que fueron procesador y analizados posteriormente.

- Una vez obtenidos los datos, se realizó un análisis estadístico con la prueba de ANOVA de una vía para poder hacer una comparación de medias y posteriormente una prueba de DUNCAN para poder ver donde estaban las diferencias entre los algoritmos
- Los algoritmos fueron comparados en sus diferentes variables en cuestiones de tiempo y precisión para encontrar semejanzas entre palabras codificadas con la lista de nombres y la lista negra de OFAC.
- Se hizo la selección de un algoritmo que cumplía con los requisitos de ser efectivo al momento de detectar coincidencias contra una lista de nombres en español.
- De la misma manera ese algoritmo seleccionado tenía que cumplir con una característica principal, que era la de disminuir el índice de falsos positivos.
- Al final se hizo un balance de las variables y la eficacia de los algoritmos y se optó por seleccionar el algoritmo NYSIIS como la mejor opción. Si bien en términos de velocidad no tenía el mejor desempeño, su disminución de falsos positivos y precisión para detectar verdaderos positivos fueron los índices más elevados de los 3.

5.1 Trabajo a Futuro

Existen adaptaciones personales a ciertos algoritmos fonéticos para la búsqueda de nombres en español (Amón et al., 2012) hoy en día, pero no se han hecho pruebas para búsquedas en conjunto con algoritmos de lógica difusa y de búsqueda parcial con el objetivo de ser utilizados para la detección de casos anómalos comparados contra alguna lista negra, como pudiera ser la de OFAC, FBI u alguna otra.

Sería interesante hacer una análisis u adaptación de estos algoritmos fonéticos a más de un solo lenguaje, en este caso y para este trabajo en español, ya que podría mejorar el índice de detección de los algoritmos que se usan hoy en día

para la detección de nombres en aduanas, bancos u algún otro comercio que sea auditado para su revisión de lavado de dinero.

También la utilización de nuevas técnicas de manejo de datos o de agrupación de los mismos en conjunto con algoritmos fonéticos podrían arrojar resultados que aún no conocemos y que pudiesen ser mejores de los que sabemos hoy en día, técnicas como machine language o de minería de datos son alternativas no tan fuertemente exploradas en este rubro.

Otro tema interesante es el uso de base de datos no relacionales para el manejo de mucha información en conjunto con los algoritmos fonéticos. Los bancos tienen que procesar información todo el tiempo y en estos momentos con el crecimiento de las transacciones y transferencias electrónicas y el uso de dispositivos móviles para accesar a la banca en línea, se incrementa el número de transacciones diarias, al ya no ser un impedimento tener que estar físicamente en un banco para depositarle a alguien o comprar algo por internet. Este tipo de actividades incrementan los factores de riesgo de lavado de dinero y como consecuencia del combate al terrorismo, pasando también por aspectos no tratados en este documento como el fraude o suplantación de identidad.

Definitivamente hay muchos temas de reciente estudio que son susceptibles a ser investigados y analizados para combatir el crimen financiero y el lavado de dinero, pero se requiere de información, pruebas y un estudio más profundo, tal vez con datos reales, con información proveniente de aduanas o de bancos y que permitan ver cómo se comportan nuevas técnicas utilizadas hoy en día para el manejo de cadenas, pero esta vez, para encontrar incidencias que puedan representar un manejo no adecuado de recursos o comúnmente conocido como recursos de procedencia ilícita.

BIBLIOGRAFÍA

ACAMS. (2011). AML Survey Results from Dow Jones Risk & Compliance & ACAMS. *The Dow Jones Anti-Money Laundering Survey with ACAMS*. Retrieved from http://www.acams.org/download-your-aml-resources/

Agnes, K. (2014). Legal Aspects of CyberSecurity in Emerging Technologies: Smart Grids and Big Data. In *Regulating eTechnologies in the European Union* (pp. 189–216). Springer International Publishing. Retrieved from http://link.springer.com.dibpxy.uaa.mx/chapter/10.1007/978-3-319-08117-5_10

Amón, I., Moreno, F., & Echeverri, J. (2012). Algoritmo Fonético para la Detección de Cadenas de Texto Duplicadas en el Idioma Español. *Revista Ingenierías Universidad de Medellín*, *11*(20), 9.

Apache. (2014). Java Language Codec Package. Retrieved from https://commons.apache.org/proper/commons-codec/apidocs/org/apache/commons/codec/language/package-summary.html

Basilea. (1988). *Declaración Del Comité De Autoridades De Supervisión Bancaria Del Grupo De Los Diez Y De Luxemburgo, Hecha En Basilea En diciembre De 1988, Sobre Prevención En La Utilización Del Sistema Bancario Para Blanquear Fondos De Origen Criminal* (p. 3). Basilea. Retrieved from http://www.pnsd.msssi.gob.es/pnsd/legislacion/pdfestatal/i47.pdf

Beider, A., & Morse, S. (2010, March). Phonetic Matching: A Better Soundex. Association of Professional Genealogists. Retrieved from http://stevemorse.org/phonetics/bmpm2.htm

Calderón Hinojosa, F. (2011). *Quinto Informe de Gobierno 2011* (Quinto Informe de Gobierno 2011) (p. 778). México. Retrieved from http://calderon.presidencia.gob.mx/informe/quinto/archivos/informe_de_gobierno/pdf/Quinto-informe-de-gobierno.pdf

CESOP. (2012, March). *Lavado de Dinero. Indicadores y Acciones Binacionales*. Presented at the Carpeta de Indicadores y Carpetas Sociales #17, México. Retrieved from

http%3A%2F%2Fwww3.diputados.gob.mx%2Fcamara%2Fcontent%2Fdownload%2F274232%2F852053%2Ffile%2FCarpeta-17-lavado-de-dinero.pdf

Christen, P. (2006a). A Comparison of Personal Name Matching: Techniques and Practical Issues. Department of Computer Science, The Australian National University. Retrieved from https://cs.anu.edu.au/people/Peter.Christen/publications/mcd2006.ps.gz

Christen, P. (2006b). A Comparison of Personal Name Matching: Techniques and Practical Issues. *The Australian National University*. Retrieved from https://digitalcollections.anu.edu.au/bitstream/1885/44521/3/TR-CS-06-02.pdf

Christen, P. (2007). Development and User Experiences of an Open Source Data Cleaning, Deduplication and Record Linkage System. *The Australian National University*, *11*(1), 10.

Christen, P. (2008). Automatic Record Linkage using Seeded Nearest Neighbour and Support Vector Machine Classification. The Australian National University. Retrieved from https://pdfs.semanticscholar.org/340d/01966221cbf6608365191d409d49c661927d.pdf

Christen, Peter. (2008). Febrl – An Open Source Data Cleaning, Deduplication and Record Linkage System with a Graphical User Interface. The Australian National University. Retrieved from http://users.cecs.anu.edu.au/~Peter.Christen/publications/kdd2008christen-febrl-demo.pdf

CICAD. (1990). *La Convención en Lavado, Registro, Embargo y Confiscación de los Productos del Crimen* (p. 78). Washington: CICAD. Retrieved from http%3A%2F%2Fwww.cicad.oas.org%2FLavado_Activos%2Fesp%2FGupoExpertos%2Fdocumentos%25202001-2005%2FEl%2520delito%2520de%2520lavado%2520de%2520activos%2520como%2520delito%2520autonomo.doc

CNBV. (2012). Certificacion de Auditores, Oficiales de Cumplimiento y demás Profesionales en Materia de Prevención de Lavado de Dinero y Financiamiento al Terrorismo. Secretaría de Hacienda y Crédito Público. Retrieved from

http://www.cnbv.gob.mx/PrevencionDeLavadoDeDinero/Documents/TemarioGuia-ObtencionCertificado-20150428.pdf

CNBV. (2015, June). Régimen Internacional PLD/FT. *SHCP*, 69.

CNBV. (n.d.). Comisión Nacional Bancaria de Valores - Prevención de Lavado de Dinero y Financiamiento al Terrorismo [Información]. Retrieved from http://www.cnbv.gob.mx/PrevencionDeLavadoDeDinero/Paginas/Descripci%C3%B3n.aspx

Córdoba Gutiérrez, A., & Palencia Escalante, C. (2001). *El Lavado de Dinero: Distorsiones Económicas e Implicaciones Sociales* (Primera). Instituto de Investigación Económica y Social Lucas Alamán, A.C. Primera Edición.

Dang Khoa Cao, P. D. (2012). Applying Data Mining in Money Laundering Detection for the Vietnamese Banking Industry. In *Intelligent Information and Database Systems* (pp. 207 – 216). Springer Berlin Heidelberg. Retrieved from http://link.springer.com.dibpxy.uaa.mx/chapter/10.1007/978-3-642-28490-8_22

Daniel Adeoyé Leslie. (2014). *Legal Principles for Combatting Cyberlaundering* (Vol. 19). Springer International Publishing. Retrieved from http://link.springer.com.dibpxy.uaa.mx/book/10.1007/978-3-319-06416-1

Danowski, J. (2011). Counterterrorism Mining for Individuals Semantically-Similar toWatchlist Members (p. 25). Presented at the Counterterrorism and Open Source Intelligence, Chicago, IL, USA: Springer-Verlag/Wien. http://doi.org/10.1007/978-3-7091-0388-3 12

David, P., Vilariño, D., Alemán, Y., Gómez, H., Loya, N., & Salazar-Jiménez, H. (2012). The Soundex Phonetic Algorithm Revisited for SMS Text Representation. *Springer*, 8.

Department of the Treasury. (2015, October 21). False Hit List Guidance. Department of the Treasury. Retrieved from https://www.treasury.gov/resource-center/sanctions/OFAC-Enforcement/Documents/false_hit.pdf

De Santics, F. M. (2014). **Football, Gambling and Money Laundering*. Springer International Publishing.

Dr. Jae-myong, K. (2006). *Supressing Terrorist Financing and Money Laundering*. Springer Berlin Heidelberg. Retrieved from http://link.springer.com.dibpxy.uaa.mx/book/10.1007/3-540-32519-0

Duen Horng Chau, & Christos Faloutsos. (2014). Fraud Detection Using Social Network Analysis, a Case Study. In *Encyclopedia of Social Network Analysis and Mining* (pp. 547–552). Springer New York. Retrieved from http://link.springer.com.dibpxy.uaa.mx/referenceworkentry/10.1007/978-1-4614-6170-8_284

Elizondo Gasparín, M. M. (2008). Las personas políticamente expuestas y el blindaje de las elecciones. *Instituto de Investigaciones Jurídicas de la UNAM*, 56.

FATF. (n.d.). FATF (The Financial Action Task Force) [Información]. Retrieved from http://www.fatf-gafi.org/

Fernandez Espejel, G., & Arellano Trejo, E. (2012, June). ¿Por qué legislar el combate al lavado de dinero? *Camara de Diputados*, 242.

GAFI. (1989). *Grupo de Acción Financiera Internacional sobre el Blanqueo de Capitales* (p. 2). Reunion G7. Retrieved from http://www.cnbv.gob.mx/CNBV/Documents/VSPP_GAFI.pdf

Galindo, J. (2006). *Fuzzy databases: Modeling, Design and Implementation* (Idea group publishing). Idea group publishing. Retrieved from http://www.boente.eti.br/boente2012/fuzzy/ebook/ebook-fuzzy-galindo.pdf

Gamboa, C. (2013). *"Lavado de Dinero" Estudio Teórico Conceptual, Derecho Comparado, Tratados Internacionales y de la Nueva Ley en la Materia en México*. Cámara de Diputados. Retrieved from http://www.diputados.gob.mx/sedia/sia/spi/SAPI-ISS-01-13.pdf

Gibson García, R. (2009). *Prevención De Lavado De Dinero Y Financiamiento Al Terrorismo* (Primera Edición). México: INACIPE. Retrieved from http://www.inacipe.gob.mx/stories/publicaciones/temas_selectos/Prevencion.lavado.dinero.pdf

Gobierno de México. Ley Federal Para La Prevención E Identificación De Operaciones Con Recursos De Procedencia Ilícita, DOF 17-10-2012 22 (2012). Retrieved from http://www.diputados.gob.mx/LeyesBiblio/pdf/LFPIORPI.pdf

Goiser, K., & Christen, P. (2007). Quality and Complexity Measures for Data Linkage and Deduplication. *Springer-Verlag Berlin Heidelberg 2007*, *1*, 25.

González Rodríguez, J. de J. (2009, April). El lavado de dinero en México, escenarios, marco legal y propuestas legislativas. Centro de Estudios Sociales y de Opinión Pública. Retrieved from www3.diputados.gob.mx/.../file/Lavado_dinero_Mexico_docto66.pdf

Government Accountability Office. (2008). *Terrorism Watch List Screening*. New York: Nova Science Publishers. Retrieved from https://books.google.com.mx/books?id=Ymnp40FOC_EC&pg=PA20&lpg=PA20&dq=watchlist+screening+algorithm&source=bl&ots=7xMuQVTJA3&sig=CZPavX_a6-0qbomLJIXxSBD5Uao&hl=en&sa=X&ved=0ahUKEwig69z7gYfLAhXLRyYKHY4VCPEQ6AEINjAC#v=onepage&q=watchlist%20screening%20algorithm&f=false

Harmeet Kaur Khanuja, & Dattatraya S. Adane. (2014). Forensic Analysis for Monitoring Database Transactions. In *Security in Computing and Communications* (Vol. 467, pp. 201–210). Springer Berlin Heidelberg. Retrieved from http://link.springer.com.dibpxy.uaa.mx/chapter/10.1007/978-3-662-44966-0_19

Harris, S., & Ross, J. (2006). *Beginning Algorithms* (Wiley Publishing, Inc., Vol. 1). Indianapolis: Wiley Publishing, Inc. Retrieved from http://mff.cz/data/ADS_BA.pdf

Hellman, D. (2011). *The Python Standard Library by Example*. Michigan: Addison-Wesley. Retrieved from https://github.com/manageyp/manageyp.github.com/blob/master/attachments/pdfs/The%20Python%20Standard%20Library%20by%20Example.pdf

Hernández, R., Fernández, C., & Baptista, P. (1991). *Metodología de la Investigación* (Vol. 1). Naulcalpan de Juarez, México: Hill Interamericana de México. Retrieved from http://www.dgsc.go.cr/dgsc/documentos/cecades/metodologia-de-la-investigacion.pdf

Hood, D. (2012). Caverphone : Phonetic Matching algorithm. *Matching Multiple Data Sources from New Zealand: The Experience of the Caversham Project*, *1*, 11.

Hsinchun, C. (2012a). *Dark Web* (Vol. 30). Springer New York. Retrieved from http://link.springer.com.dibpxy.uaa.mx/book/10.1007/978-1-4614-1557-2

Hsinchun, C. (2012b). Intelligence and Security Informatics (ISI): Research Framework. In *Dark Web* (Vol. 30, pp. 19–30). Springer New York. Retrieved from http://link.springer.com.dibpxy.uaa.mx/chapter/10.1007/978-1-4614-1557-2_2

Hyyrö, H. (2003, December 5). *Practical Methods for Approximate String Matching*. UNIVERSITY OF TAMPERE, DEPARTMENT OF COMPUTER SCIENCES. Retrieved from https://tampub.uta.fi/bitstream/handle/10024/67325/951-44-5840-0.pdf?sequence=1

INEGI. (2015). Producto Interno Bruto (PIB) - Trimestral. Retrieved from http://www.inegi.org.mx/est/contenidos/proyectos/cn/pibt/default.aspx

Innerhofer-Oberperfler, R. (2004, July 15). *Using Approximate String Matching Techniques to Join Street Names of Residential Addresses*. Free University of Bolzano-Bozen, Faculty of Computer Science. Retrieved from http://www.inf.unibz.it/dis/projects/ebz/theses/bscthesis-innerhoferoberperfler-04.pdf

Jaime, L. M., Sabu M, T., Danda B., R., & Di, J. (2014). *Security in Computing and Communications* (Vol. 467). Springer Berlin Heidelberg. Retrieved from http://link.springer.com.dibpxy.uaa.mx/book/10.1007/978-3-662-44966-0

Jeng-Shyang Pan, S.-M. C., & Ngoc Thanh Nguyen. (2012). *Intelligent Information and Database Systems*. Springer Berlin Heidelberg. Retrieved from http://link.springer.com.dibpxy.uaa.mx/book/10.1007/978-3-642-28490-8

Kock Wiil, U. (2011). *Counterterrorism and Open Source Intelligence*. SpringerWienNewYork. Retrieved from http://link.springer.com/chapter/10.1007/978-3-7091-0388-3_12

KPMG. (2014). *Global Anti-Money Laundering Survey 2014*. Encuesta. Retrieved from www.kpmg.com/Global/en/IssuesAndInsights/ArticlesPublications/global-anti-money-laundering-survey/Documents/global-anti-money-laundering-survey-v6.pdf

KPMG Cost of Compliance. (2013). *The Cost of Compliance 2013* (p. 32). KPMG. Retrieved from https://www.kpmg.com/dutchcaribbean/en/Documents/Publications/The-cost-of-compliance-v2.pdf

Lait, A. J., & Randell, B. (1995, March). An Assessment of Name Matching Algorithms. University of Newcastle upon Tyne. Retrieved from http://homepages.cs.ncl.ac.uk/brian.randell/Genealogy/NameMatching.pdf

Miller, K., & Arehart, M. (2008). Improving Watchlist Screening by Combining Evidence From Multiple Search Algoritms (pp. 106 –110). Presented at the Technologies for Homeland Security, 2008 IEEE Conference on, Waltham, MA: IEEE. http://doi.org/10.1109/THS.2008.4534432

Navarro, G., Baeza-Yates, R., Sutinen, E., & Tarhio, J. (2001). Indexing Methods for Approximate String Matching. IEEE. Retrieved from http://www.dcc.uchile.cl/~gnavarro/ps/deb01.pdf

Nhien-An Le-Khac, Sammer Markos, & M-Tahar Kechadi. (2009). A Heuristics Approach for Fast Detecting Suspicious Money Laundering Cases in an Investment Bank. *World Academy of Science, Engineering and Technology*, *3*, 5.

OCC. (2011, April 4). Supervisory Guidance On Model Risk Management. OCC 2011-2012. Retrieved from http://www.occ.gov/news-issuances/bulletins/2011/bulletin-2011-12a.pdf

OFAC List Search. (2016). *OFAC Sanctions List Search*. Retrieved from https://sanctionssearch.ofac.treas.gov/Details.aspx?id=1720

OFAC SDN List. (2016, May 13). Resource Center: Specially Designated Nationals List (SDN). Retrieved from https://www.treasury.gov/resource-center/sanctions/SDN-List/Pages/default.aspx

ONU. (1988). *Convención De Las Naciones Unidas Contra El Tráfico Ilícito De Estupefacientes Y Sustancias Psicotrópicas* (p. 44). Viena. Retrieved from http://www.poderjudicialyucatan.gob.mx/digestum/marcoLegal/08/2013/DIGESTUM08036.pdf

ONU. (2000, November 15). Convención de las Naciones Unidas contra la Delincuencia Organizada Transnacional. Retrieved from http://www.acnur.org/t3/fileadmin/scripts/doc.php?file=t3/fileadmin/Documentos/BDL/2002/1292

Patel, R. D., & Singh, D. K. (2013). Credit Card Fraud Detection & Prevention of Fraud Using Genetic Algorithm. *International Journal of Soft Computing and Engineering (IJSCE)*, *2*(6), 3.

Patman, F., & Shaefer, L. (2006). Is Soundex Good Enough for You? The Hidden Risks of Soundex-Based Name Searching. IBM. Retrieved from ftp://public.dhe.ibm.com/software/data/mdm/soundex.pdf

Pedrosa Leyva, E. (2013, August). Lavado de dinero en México. Estimación de su magnitud y análisis de su combate a través de la inteligencia financiera. *INEGI*, *4*, 12.

Peña Nieto, E. (2015). *Tercer Informe de Gobierno 2015* (Informe de Gobierno No. Tercero) (p. 640). México: Gobierno Federal. Retrieved from http://www.presidencia.gob.mx/tercerinforme/

Pivotal. (2016). Spring Pivotal. Retrieved from https://spring.io/tools/eclipse

PWC. (2010). From source to surveillance: the hidden risk in AML monitoring system optimization. PwC. Retrieved from https://www.pwc.com/us/en/anti-money-laundering/publications/assets/aml-monitoring-system-risks.pdf

PWC. (2013). *La ley federal para la prevención e identificación de operaciones con recursos de procedencia ilicita.* (p. 6). New York: PWC. Retrieved from http://www.pwc.com/mx/es/retos-sector-financiero/archivo/2013-07-ley-federal-operaciones-ilicita.pdf

PwC Sanctions Revisited. (2014). Financial Sanctions Revisited. PwC. Retrieved from https://www.pwc.com/us/en/financial-services/regulatory-services/publications/assets/global-financial-sanctions.pdf

Rainer, B. (2013). *The Economics of Information Security and Privacy*. Springer Berlin Heidelberg. Retrieved from http://link.springer.com.dibpxy.uaa.mx/book/10.1007/978-3-642-39498-0

Rajkovic, P., & Dragan, J. (2007). Adaptation and Application of Daitch-Mokotoff soundex algorithm on Serbian names. *XVII Conference on Applied Mathematics*, 12.

Ramachandran, R. (2014, July). OFAC Name Matching and False-Positive Reduction Techniques. Cognizant. Retrieved from http://www.cognizant.com/InsightsWhitepapers/OFAC-Name-Matching-and-False-Positive-Reduction-Techniques-codex1016.pdf

RamaKalyani, K., & UmaDevi, D. (2012). Fraud Detection of Credit Card Payment System by Genetic Algorithm. *International Journal of Scientific & Engineering Research*, *3*(7), 6.

Recomendaciones GAFI. Las recomendaciones del GAFI, GAFISUD 11 / II Plen 1 (2012). Retrieved from http://www.fatf-gafi.org/media/fatf/documents/recommendations/pdfs/FATF-40-Rec-2012-Spanish.pdf

Ross Anderson, Chris Barton, Rainer, B., Richard, C., Michel J. G., van E., Michael, L., … Stefan, S. (2013). Measuring the Cost of Cybercrime. In *The Economics of Information Security and Privacy* (Vol. 4, pp. 265–300). Springer Berlin Heidelberg. Retrieved from http://link.springer.com.dibpxy.uaa.mx/chapter/10.1007/978-3-642-39498-0_12

Schaidnagel, M., & Fritz Laux, I. P. (2013). DNA: An Online Algorithm for Credit Card Fraud Detection for Games Merchants (p. 6). Presented at the The Second International Conference on Data Analytics, IARIA.

Schatten, M., Kakulapati, V., & Cubrilo, M. (2010). Reasoning about Social Semantic Web Applications using String Similarity and Frame Logic, *1*, 11.

Schnell, R., Bachteler, T., & Bender, S. (2003). Record linkage using error-prone strings. 2003 Joint Statistical Meetings. Retrieved from https://www.amstat.org/sections/srms/Proceedings/y2003/Files/JSM2003-000833.pdf

Schott, P. (2007). *Guía de referencia para el antilavado de activos y la lucha contra el financiamiento del terrorismo*. Bogota, Colombia: Mayol Ediciones. Retrieved from http://siteresources.worldbank.org/INTAML/Resources/RefrenceGuideSpanish.pdf

SCT. (2016). Estadística Operacional de Aeropuertos. Retrieved from http://www.sct.gob.mx/transporte-y-medicina-preventiva/aeronautica-civil/estadisticas/estadistica-operacional-de-aeropuertos-airports-operational-statistics/

Select Committee on Homeland Security. (2004). *disrupting terrorist travel: safegaurding america's borders through information sharing*. US Government. Retrieved from https://books.google.com.mx/books?id=fiCMgkzYuAMC&pg=PA39&lpg=PA39&dq

=watch+list+false+positive+rate&source=bl&ots=IwIEqacK_5&sig=nZ6FfbHQaoihkfPkIteQoTiLubw&hl=es-419&sa=X&ved=0ahUKEwiU_P_88rLLAhVruoMKHaRNA9A4ChDoAQgoMAI#v=onepage&q=watch%20list%20false%20positive%20rate&f=false

Shah, R. (2014). Improvement of Soundex Algorithm for Indian Language Based on Phonetic Matching. *International Journal of Computer Science, Engineering and Applications*, *4*(3), 9. http://doi.org/10.5121/ijcsea.2014.4303

SHCP. (2015). Portal de Prevención de Lavado de Dinero [Gobierno]. Retrieved from https://sppld.sat.gob.mx/pld/interiores/leermas.html

Singh, S. (2013). Fraud Detection using Neural Network, 3.

Snae, C. (2007). A Comparison and Analysis of Name Matching Algorithms. *World Academy of Science, Engineering and Technology*, *1*(1), 5.

Srinivasa, S., & Mehta, S. (2014). *Big Data Analytics* (Vol. 1). Springer Berlin Heidelberg.

Tenorio, M. (2014). *Programa antilavado de dinero para las sofomes en México*. México. Retrieved from http://www.indifep.mx/assets/Presentacion2EQAICC.pdf

UNDP. (1995). *The Buenos Aires Plan of Action* (p. 35). New York: UNDP. Retrieved from http://ssc.undp.org/content/dam/ssc/documents/Key%20Policy%20Documents/BAPA.pdf

Valdivieso, C., Valdivieso, R., & Valdivieso, O. (2011). Determinación del Tamaño Muestral Mediante el Uso de Árboles de Decisión. Universidad Privada Bolmana. Retrieved from http://www.upb.edu/RePEc/iad/wpaper/0311.pdf

Vats, S., Dubey, S. K., & Pandey, N. K. (2013). Genetic algorithms for credit card fraud detection (p. 12). Presented at the International Conference on Education and Educational Technologies, India. Retrieved from http://www.europment.org/library/2013/rhodes/bypaper/EET/EET-03.pdf

Videgaray Caso, L. (2014, April). *Conferencia De Prensa Que Ofreció El Secretario De Hacienda, Dr. Luis Videgaray Caso, En La Embajada De México En Washington, D.C., Sobre Las Reuniones De Primavera Del Banco Mundial Y El Fondo Monetario Internacional, Así Como En La De Ministros De Finanzas Del G20*. Washington.

Retrieved from
http://www.shcp.gob.mx/Biblioteca_noticias_home/palabras_lvc_washington.pdf
Works, C. (2016). *Web Fake Generator*. Español, Corban Works LLC. Retrieved from
http://es.fakenamegenerator.com/