UNIVERSIDAD AUTÓNOMA METROPOLITANA UNIDAD IZTAPALAPA

DEPARTAMENTO DE MATEMÁTICAS

TÉSIS DE MAESTRÍA:

Rendición de cuentas: Un modelo para detectar la corrupción en México

Tésis de maestría por Herrera Curiel José Roberto para obtener el grado de maestro en ciencias en la UAM-I

Supervisada por:
Pérez Salvador Blanca R.

México, D.F., a 24 de Abril del 2017

Dedicado a
mis padres

Agradecimientos

Agradezco a la Universidad Autónoma Metropolitana Unidad Iztapalapa, la cual me brindo su apoyo como institución en todo momento, me abrió sus puertas y me ofrecío la oportunidad de crecer y desarrollarme intelectualmente, todo esto con un marco de prestigio y calidad.

Agradezco a mis padres, piezas fundamentales en todo momento para poder llevar a cabo mis metas, por su dedicación, formación y afecto que siempre me han brindado.

Índice general

Página

Objetivos

El propósito de este proyecto es analizar la problemática y realizar una propuesta de un modelo que utilice patrones, tendencias o tipologías actuales del comportamiento de funcionarios que utilizan su cargo o las instituciones de gobierno para obtener beneficios particulares distintos a las funciones específicas de su cargo; de tal modo que ayude a dar un acercamiento en la medición de los niveles de corrupción de los servidores públicos en el país. Este trabajo posee 3 objetivos principales.

- Definir un modelo conceptual sobre conductas de corrupción derivado de casos reconocidos o identificados en México. En este paso se va revisar información documental de carácter nacional e internacional sobre el fenómeno de corrupción.
- Construir una base de datos con información de operaciones financieras para identificar o simular el comportamiento de interés. Identificar la información de acceso público de datos de las personas políticamente expuestas (PEPs) que ejercen cargos públicos que pudieren caer en actos u omisiones que impliquen un comportamiento de corrupción tales como la malversación de fondos. Así como simular aquella información que aunque exista no es pública pero que si puede tenerse acceso desde las instituciones gubernamentales o privadas, quienes son responsable de resguardarla. Las variables o indicadores deberán cubrir criterios conceptuales teóricos definidos en la primera parte y también criterios estadísticos.
- Definir un modelo matemático o estadístico en términos de indicadores que permita seleccionar sujetos con posibles elementos de corrupción para que sean sometidos a investigación.

Capítulo 1

Introducción

La corrupción es un fénomeno que afecta los intereses económicos, políticos y sociales de México, en la actualidad presenta un grave aumento en todos los niveles, órdenes y rubros de nuestra vida cotidiana, especialmente, en la Administración Pública. La corrupción, al igual que el crimen organizado, ha ido evolucionado y trascendido las fronteras de los países que conforman este ya globalizado mundo, surgiendo con esto, nuevas formas y figuras delictivas, más técnicas, especializadas y complejas, apreciándose prácticas corruptas ya no sólo en el ámbito público del país, sino en el sector privado nacional e internacional. La comunidad internacional, de la cual México forma parte, conscientes de esta situación, así como de los devastadores efectos que ocasiona la corrupción en sus propias Naciones, a través de diversos organismos e instituciones, han adoptado políticas, acciones y estrategias anticorrupción, para el combate de este mal, así como de sus grandes aliados, el crimen organizado y el lavado de dinero, esto, mediante convenios, acuerdos y compromisos en la materia. Sin embargo, la percepción y realidad en nuestro país, es que no obstante lo anterior, son pocos o nulos los resultados que se ven para frenar los excesos e impunidad de funcionarios o servidores públicos en prácticas corruptas, así como su participación en actos o conductas delictivas, obligando a replantear un mejor y eficaz combate a este mal, a través de medidas efectivas en torno a los tentáculos de la misma.

1.1. Antecedentes

La corrupción ha sido un común denominador en México y el resto del mundo, su existencia no depende del nivel de desarrollo de este, pues en países más

desarrollados que en el nuestro la corrupción existe; de la misma forma, no se puede asegurar que sea un fenómeno privativo de México, y desde luego no podemos asociarla sólo al pasado reciente, la corrupción está y ha estado presente en regímenes democráticos y no democráticos, en transición o en proceso de consolidación democrática en los países de todo el mundo.

La historia de la corrupción en México o en el mundo entero ilustra de alguna manera que ha estado presente desde hace ya mucho tiempo. Bien se podría decir, que desde que existe una organización humana, ó algún grado de institucionalidad y normas, se encuentra a alguien dispuesto a violar ese orden a cambio de obtener beneficios extraposicionales. Un recorrido por el México precolombino, el de la colonia, el independiente, el de la reforma, el de la revolución, el posrevolucionario o el actual nos darían más de un ejemplo de lo que aquí se comenta; Fray Bernardino de Sahagún comenta cómo en la selección de jueces en el México prehispánico se evitaba que los electos tomaran dádivas o actuasen parcialmente, como señala Cárdenas (2005). Por otro lado Busquet (2005), afirma que uno de los grandes problemas de México fue que al volverse una nación independiente, nació sin tener bases institucionales sólidas. Casi todo el siglo XIX se caracterizó por ser un periodo de inestabilidad política. No es hasta el llamado Porfiriato cuando gobernó Porfirio Díaz, que México vive un periodo de paz y prosperidad económica. Porfirio Díaz logró apaciguar a la competencia política y mantuvo contentos a sus colaboradores. La corrupción entre su gente era algo tolerado por Díaz. Ésta tenía como objetivo asegurar la lealtad de sus funcionarios; era un costo necesario para mantener la estabilidad política que necesitaba el país.

En nuestra independencia Aldama y Allende tenian grandes dificultades con Hidalgo, pues el valiente y talentoso cura no sólo permitía que la gente se dedicara al saqueo, si no que incluso parecía propiciarlo, ya independientes con Santa Anna que ocupó la presidencia 11 veces , la corrupción alcanzó dimensiones de delirio, jugador y enamorado , fue el primer Presidente que inaguró el estilo de manejar la hacienda pública como si fuera su caja chica, no en balde le apodaba El quince uñas‘ Gonzáles (1993).

La tradición del país en materia de corrupción es de mas prosapia que nuestra tradición tequilera, hoy en dia han surgido mas tipos de corrupción, pero la constante de deshonestidad ha sido la misma, de ella no se salva ni uno de nuestros héroes más limpios, Madero. Al prócer le dio por el nepotismo, dos de sus tíos encabezaban las secretarías de Desarrollo y de Economiá; un primo estaba en la Secretaría de la Defensa, su hermano Gustavo dirigía su partido, y su otro hermano Emilio comandaba las fuerzas armadas del norte, la nómina familiar era amplia Gonzáles, (2005).

Cómo estos ejemplos hay muchos más escritos y reportados en la literatura

actual, como ya mencionamos el problema de la corrupción en México y el resto del mundo, no se remonta a unos cuantos siglos antes del periodo actual. Todo lo mencionado anteriormente indica que la corrupción ha estado presente desde hace ya un largo tiempo, en nuestro país hasta nuestros días, el pasado, las encuestas y los dichos pupulares, documentan de algúna manera el hecho de que la política y sus oficiantes han sido las grandes causas de la corrupción en México.

1.2. Definición y Tipos de Corrupción

Para entender cualquier problema primero es necesario definirlo e identificarlo, esto es especialmente complejo en el caso de la corrupción por dos motivos. En primer lugar porque la corrupción engloba numerosas conductas siempre enunciadas pero casi nunca bien definidas y tipificadas en la ley. Segundo, porque siendo conductas apartadas de la ley y merecedoras de un castigo, se practican a la sombra o de manera clandestina. Los que la ejercen de manera cotidiana o los que la cultivan como forma de vida intentan ocultar sus huellas y desaparecer el cuerpo del delito Casares (2015). El concepto de corrupción supone importantes problemas de definición. La corrupción es un fenómeno muy complejo con múltiples causas y efectos, que fluctúan desde el simple acto de un pago ilícito hasta el funcionamiento endémico del sistema económico y político-social. El problema de la corrupción ha sido considerado no sólo como un problema estructural, sino también moral y cultural. Por tanto, las definiciones que existen sobre corrupción van desde términos generales de mal uso del poder público y "deterioro moral", hasta definiciones legales estrictas, que describen este fenómeno como un mero acto de extorsión que involucra a algún servidor público y una transferencia de recursos. Por lo tanto, el estudio de la corrupción ha sido multidisciplinario y disperso, y ha fluctuado desde los modelos teóricos universales hasta las descripciones detalladas de escándalos de corrupción individual.

En la literatura especializada existen varias definiciones de corrupción. Los investigadores de este fenómeno se han ocupado, en parte, en clasificar las diferentes formas de corrupción, con el objeto de hacer más operable este concepto y facilitar su análisis. Existen, por lo tanto, varias sugerencias de cómo definir este fenómeno y cómo clasificarlo en subfenómenos: La palabra corrupción proviene del adjetivo corruptus, que en latín significa estropeado, descompuesto o destruido. De acuerdo con el Concise Oxford English Dictionary, un significado de corromper en el contexto social es sobornar, y corrupción equivale a "deterioro moral". Esta definición ni la etimología latina de la palabra restringen la noción de corrupción al sector público. De modo que la corrup-

ción también puede ocurrir en la esfera privada. Destacadas organizaciones internacionales adoptan un definición igualmente incluyente de corrupción. La oficina de las Naciones unidas sobre Drogas y Crimen subraya que la corrupción "puede ocurrir en los dominios público y privado", su programa Global contra la Corrupción define la corrupción como el "abuso del poder para beneficio privado" e incluye al sector público y al privado. En forma similar el Banco Mundial no considera que la corrupción limite al sector público. Para Trasparencia Internacional la corrupción se define, operativamente como "el mal uso del poder otorgado para beneficio privado". Esta definición también incluye a los individuos de los sectores privado y público. Pero entre economistas predomina un consenso diferente, en su artículo de revisión, Jain (2001) asegura "hay consenso en que la corrupción se refiere a los actos en los que el poder del cargo público se usa para beneficio personal de una manera que contravienen las reglas del juego". es decir el poder de un cargo público puede ser usado para facilitar la corrupción.

En suma, no existe un consenso claro sobre qué se entiende por corrupción. Algunas definiciones buscan dar un significado formal y amplio, mientras que otras no están diseñadas para definir la conducta corrupta per se, pero se elaboran para aislar aquellas actividades que son el tema de interés del investigador. Cada definición representa un nivel de análisis distinto y, por ende, una manera diferente de entender el fenómeno. Definir el concepto de corrupción es, sin duda, uno de los problemas más importantes que enfrentan los interesados en este fenómeno. La tarea es engañosa dada la actitud hacia el tema, pues lo que algunos consideran como corrupción puede no serlo para otros. Este problema de definición se basa en el hecho de que el término no tiene sentido sin un referente de comparación.

Para este trabajo adoptaremos la definición de Transparencia Internacional, la cual es para nuestros fines la mejor, en el sentido que abarca al sector privado y a los individuos inmersos en estas instituciones. Una vez establecida la definición que usaremos, debemos de considerar los tipos de corrupción que tomaremos en cuenta; hay un sinfin de delitos que son considerados como corrupción, sin embargo, usaremos la siguiente clasificación establecida por Gonzáles (2005) : 1) soborno de funcionarios públicos nacionales. 2) Tráfico de influencias. 3) Abuso de influencias. 4) Enriquecimiento ilícito. 5) Encubrimento. 6) Obstrucción de justicia.

1.3. Cómo se Mide la Corrupción

La corrupción es el típico ejemplo de un fenómeno que es posible observar pero no cuantificar, en tanto que "no puede existir estadística alguna sobre un fenómeno cuya naturaleza es ambigua", por tanto el principal problema de medir la corrupción es que, es quizás el crimen menos reportado que existe. Un acto de corrupción es perpetrado generalmente con gran secrecía. Todas las partes involucradas en la transacción corrupta (el que da el soborno y el que lo recibe) están usualmente satisfechos con el resultado y reconocen las posibles consecuencias negativas que resultarían de revelar su propio papel en dicha conducta criminal, incluso si no se encuentran satisfechos. Mientras tanto, las víctimas de la corrupción, que son usualmente el público en general y la sociedad en su conjunto, se encuentran inconscientes de los actos específicos de la corrupción o están acostumbrados de manera tal que se vuelven indiferentes a ella. Dada esa secrecía y los intereses comunes entre las partes involucradas, los niveles de corrupción son estremadamente dificiles de medir, sin embargo, en los últimos años hemos sido testigos de la publicación de varios índices que intentan medir el nivel de corrupción, de opacidad o de transparencia de diferentes países en todo el mundo.

En Castillo se analiza la diferencia entre indicador e índice; los indicadores son parámetros de medición que reflejan el comportamiento observado de un fenómeno. Representan medidas sobre aspectos no directamente mensurables, como lo son muchas de las actividades y propósitos gubernamentales: salud, educación, bienestar social, etcétera. Formalmente hablando, los indicadores se expresan en términos de ecuaciones, donde el número de la categoría observada es dividido entre el total del universo de referencia. Por ejemplo, un indicador puede ser la tasa de fecundidad. En tal caso, el número de nacidos vivos se divide entre el número de mujeres de edad fértil (15-49años), multiplicado por cien. Los índices, por su parte, son el resultado de la agregación de datos que se obtienen de los indicadores.

Una vez hecha la distinción entre indicador e índice, veamos como es que la corrupción es medida. Definir a la corrupción resulta ser un ejercicio complejo en demasia, medirla lo es aún más, descubrir un acto de corrupción que por definición busca ser encubierto requiere, además de voluntad, de recursos y capacidades de investigación importantes. Una vez descubiertos, los actos pueden ser clasificados y contabilizados pero ahí donde reinan y gobierna la opacidad, la complicidad y la impunidad de estos actos,una medición precisa y certera de este delito es prácticamente imposible, para corregir estas dificultades y tener un acercamiento más preciso al fenómeno de la corrupción se han desarrollado distintos indicadores cuyo objetivo es aproximarse al número de casos reales

de corrupción así como a las actitudes y valores de la ciudadanía y de las autoridades. Ante la dificultad o incluso imposibilidad de conocer exactamente el número de actos de corrupción cometidos, se han desarrollado metodologías alternativas para su medición, la mayoría de los estudios corresponden a una de tres categorías de encuestas Alcasar, (2015):

- de percepción sobre la extensión y frecuencia de la corrupción
- sobre la participación o exposición a una conducta clasificada como acto de corrupción
- de actitudes y valores frente los actos de corrupción propios o de otros

Junto a estos estudios coexisten aquellos de investigación participativa o experimental, los que recopilan, dan seguimiento y clasifican los actos de corrupción a partir de las investigaciones, expedientes abiertos y/o el número de condenas y, desde luego, los estudios de caso a partir de los cuales se desentrañan los mecanismos finos de la corrupción en un país.

Organización/Publicación	Metodología	Indicador/Medida	Rango
Índice de Percepción de la Corrupción *Transparencia Internacional*	Recopilación de resultados de encuestas elaboradas en más de 140 países	Percepción de niveles de corrupción según ciudadanos, empresarios y analistas	Altamente corrupto (0) Ausencia de corrupción (100)
Barómetro Global de la Corrupción *Transparencia Internacional*	Una encuesta aplicada a más de 114,000 participantes de 107 países	Experiencias directas de corrupción y percepción de la corrupción en las principales instituciones del país	Varía según la pregunta
Índice de Competitividad Global *Foro Económico Mundial*	Análisis institucional, legislativo y encuestas de opinión	Tres subíndices (i) percepción de la corrupción (ii) Leyes anti corrupción (iii) prácticas anti corrupción	Peor (1) mejor (7)
Índice de Fuentes de Soborno *Transparencia Internacional*	Encuesta aplicada a más de 300 presidentes de empresas en el mundo.	Percepción de la probabilidad de que empresas de cierta nacionalidad estén dispuestas a pagar sobornos en el exterior	Poca probabilidad (0) Alta probabilidad (10)
Latinobarómetro	Aplicación anual de más de 20,000 encuestas en 18 países de América Latina	Frecuencia y calidad institucional en el combate a la corrupción	Varía según la pregunta
Reporte de Integridad Global *Global Integirty*	Encuesta a redes de expertos y periodistas acerca de más de 300 acciones directamente relacionadas con la corrupción	Evaluación del marco anti-corrupción con base a trámites y actividades específicas	Varía según la pregunta
Indicadores Globales de Gobernabilidad *Banco Mundial*	Recopilación de encuestas a líderes y expertos en instituciones de gobierno	Incluye un indicador de *Control de la Corrupción,* que mide la efectividad de las instituciones y las tradiciones para frenar actos de corrupción	Bajo (0) Alto (1)
Índice de Estado de Derecho *World Justice Project*	Elaboración de encuestas a ciudadanos, expertos y líderes.	Incluye un indicador de *percepción de la corrupción* en el poder ejecutivo, legislativo, judicial y fuerzas de seguridad pública	Malo (0) Bueno (1)

Figura 1.1: Principales índicadores con los que se míde la corrupción

En el cuadro anterior (Alcasar, 2015) se detallan los indicadores de medición de la corrupción que son más conocidos y utilizados así como el organismo encargado de elaborarlos y difundirlos. Son indicadores imperfectos pero permiten sistematizar la información disponible, dar seguimiento a su evolución, comparar el comportamiento de distintos países y avanzar en la agenda pública para su combate.

Como vimos arriba, existen indicadores sobre corrupción basados en la percepción, pero como ofrecen datos que abarcan a una sociedad en su totalidad, son de poca ayuda cuando se busca conocer los efectos específicos de una medida preventiva o de control. El más destacado de estos índices es el índice de Percepción de la Corrupción de *Transparencia Internacional* (IPC) el cual ha aparecido anualmente desde 1995 y podriá ser descrito como un índice de índices. Es decir, el IPC se presenta como un promedio de varias encuestas, las cuales varían de un mínimo de tres hasta más de una docena. Dichas encuestas han sido aplicadas previamente y contienen preguntas sobre los niveles de corrupción que posee un determinado país. En algunos casos, el muestreo incluye a los miembros de una sociedad en específico; en otros, la percepción que se capta es la de los empresarios extranjeros o la de los expertos. El IPC ha sido extensamente usado por los investigadores (Michael, 2005).

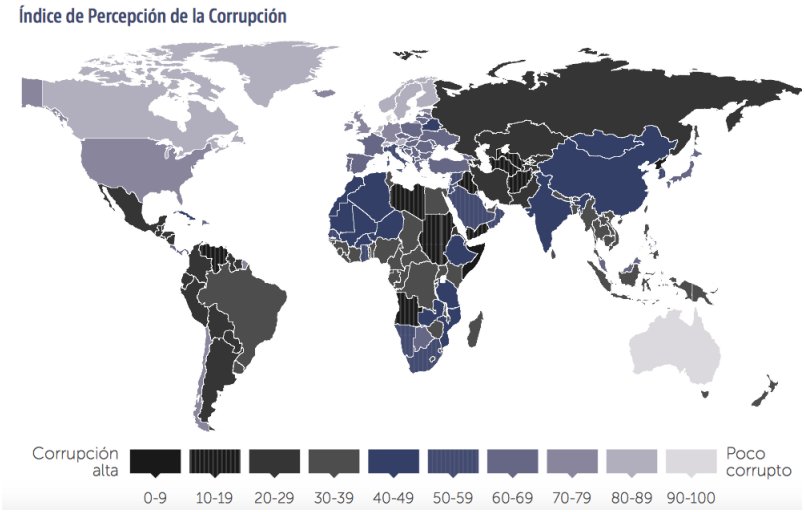


Figura 1.2: Transparencia Internacional 2014. El Índice de Percepción de la Corrupción de Transparencia Internacional otorga a cada país una calificación de 0 a 100 donde 0 es altamente corrupto y 100 altamente limpio. Utilizando las calificaciones de todos los países elabora un ranking mundial de percepción de la corrupción. El mapa muestra a los países en tonos distintos según la calificación otorgada, no su lugar en el ranking.

Los resultados arrojados por el IPC revelan que la corrupción es un problema de alcances globales aunque sus niveles son muy dispares. El último mapa elaborado por Transparencia Internacional, con los resultados del IPC 2014 en 174 países es ilustrativo.

Ahora bien, los niveles de percepción de la corrupción en México son muy alarmantes y los pocos o muchos intentos para reducirlos han sido un total fracaso. En año 2014 México obtuvo una puntuación de 35 puntos de 100 posibles y el lugar número 103 de un total de 175 países según Transparencia Internacional. Está situación no es nada alentadora para los mexicanos, ni cuando se le compara con países miembros de Organismos Internacionales a los que pertenece ni tampoco cuando se le compara con países de características similares. Acontinuación se presenta una tabla sobre las puntuaciones obtenidas por México en el IPC en estos últimos años.

Año	**Score**
2008	36
2009	33
2010	31
2011	30
2012	34
2013	34
2014	35

Cuadro 1.1: Puntuación de los últimos 7 años de México, en el IPC de *Tranparencia Internacional*, donde 100 significa mínimo nivel corrupción y 0 máximo nivel de corrupción

El índice (IPC) muestra que la corrupción en México ha sido persistente, por lo cúal es necesario hacer algo para frenar el avance de este delito, a sí como también es necesario conocer las causas y elementos que motivan su crecimiento.

Capítulo 2

Midiendo la corrupción con ecuaciones estructurales.

En este capítulo se hace una revision cuidadosa de un modelo ya existente que relaciona la corrupción y economía de las sombras por medio de ecuaciones estructurales, tambíen se analiza la teoría basica de aprendizaje no supervisado, este último dará la pauta para el modelo que se desea desarrollar, para la clasificación y deteccion de los funcionarios corruptos en nuestro país, comenzaremos con el modelo de corrupción y economía sumergida.

2.1. Modelos de ecuaciones estructurales

Los modelos de ecuaciones estructurales (MES) son una metodología estadística que utilizan un enfoque confirmatorio del análisis multivalente aplicado a una teoría estructural relacionada con un fenómeno determinado. Lo que se intentan conseguir con estos modelos de ecuaciones estructurales, es el estudio de las relaciones causales entre los datos que sean directamente observables asumiendo que estas relaciones existentes son lineales. Este tipo de modelos constituye una de las herramientas más poderosas para los estudios de relaciones causales sobre datos no experimentales cuando las relaciones son del tipo lineal. Esto hace que se haya convertido en una herramienta popular y generalmente aceptada para probar fundamentos teóricos en un gran número de disciplinas, en partícular en las ciencias sociales y del comportamiento que suelen enfrentarse a procesos cuya teoría es relativamente pobre, y suelen carecer de medios para controlar experimentalmente la recolección de información.

Existen diferentes tipos de variables que juegan un papel muy importante en los (MES), enseguida se hace la distinción de estás, según sea su medición o el papel que realizan dentro del modelo, Gómez (2011).

- **Variable latente:** también llamadas factores o variables no observadas. Por lo general son el objeto de interés en el análisis, conceptos abstractos que pueden ser observados indirectamente a través de sus efectos en los indicadores o variables observadas.
- **Variable observada:** (o también denominada de medidas o indicadoras), son aquellas variables que pueden ser medidas y cuantificadas.

Dentro de las variables latentes se tiene la siguiente clasificación

- **Exógenas:** son variables latentes de alguna forma independientes, ya que, afectan a otras variables pero no reciben ningún efecto de ninguna de ellas. Estas variables se pueden detectar en la gráficas porque no entra ninguna de las flechas a estás variables.
- **Variable endógena:** por el contrario a las exógenas, estas son consideradas como variables latentes dependientes, ya que reciben el efecto de otras variables, es decir, en las gráficas son las variables a las que llegan las flechas. Estas variables están afectadas por un término de perturbación o de error.
- **Variable error:** este tipo de variable tiene en cuenta todas aquellas fuentes de variación que no están consideradas dentro de el modelo, como puede ser en la medición de las variables. Notemos que son variables de tipo latente al no ser observables.

Representación

Los sistemas de ecuaciones estructurales pueden representarse de forma gráfica por medio de diagramas causales (path diagrams). Esta técnica se sirve de grafos que reflejan el proceso haciendo estos diagramas acordes con las ecuaciones, el proceso de su elaboración es el siguiente, ver Bruce (2003).

1.- Las relaciones entre las variables es indicado por una flecha cuya dirección es desde la variable causa hacia la variable efecto. Cada una de estas relaciones está afectada por un coeficiente, que indica la magnitud del efecto entre ambas variables, si entre dos variables no se ha especificado ninguna relación (flecha) se da por entendido que su efecto ha de ser nulo.

2.- La relación entre dos variables endógenas o de dos términos de perturbación sin una interpretación causal, se representa con una flecha bidireccional que une a ambas variables, y el parámetro asociado se indica con una varianza.

3.- Las variables observables por lo general son enmarcadas en los diagramas mediantes cuadrados sín embargo a veces este tipo de figuras son omitidas y las variables quedan representadas solo por su nombre y las variables latentes son representadas con círculos u óvalos.

4.- Los parámetros del modelo se representan sobre la flecha correspondiente.

En conclusión tenemos los siguientes puntos que debemos llevar acabo para la representación por medio de diagramas de las ecuaciones estructurales:

*Las variables observables se representas encerradas en rectángulos o cuadrados.

*Las variables no observables se representan encerradas en óvalos o círculos.

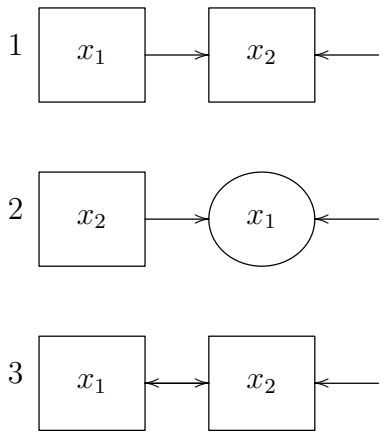
*Los errores se representan sin círculos ni rectángulos.

*Las relaciones bidireccionales se representan como líneas curvas terminadas en flechas en cada extremo.

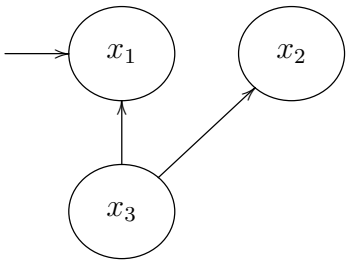
Relación entre las variables

En lo siguiente consideraremos a x_1, x_2 , x_3 variables de cualquier tipo de las mencionadas anteriormente.

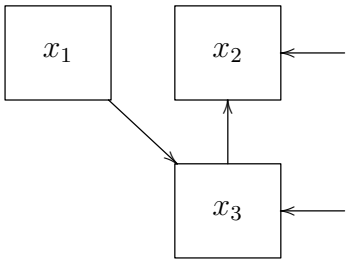
Diremos que x_1, x_2 están relacionadas directamente si x_1 causa x_2 "ver 1" (en este caso se asumiria un modelo de regresión de x_2 a x_1) ó si x_2 causa x_1 "ver 2" (en este caso se asumiria un modelo de regresión de x_1 sobre x_2), aunque estas también pueden ser reciprocas, en este caso la causalidad será bidireccional "ver 3", ver las suguientes figuras.



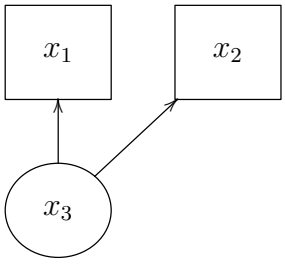
Diremos que x_1, x_2 están relacionadas espureamente si ambas tienen una causa común a la variable interveniente x_3, ver figura siguiente.



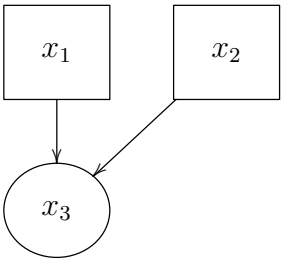
Diremos que x_1, x_2 están relacionadas indirectamente, si ambas están relacionadas a una tercera variable interviniente x_3, ver figura siguiente.



En el modelo de ecuaciones estructurales existen basicamente dos tipos de indicadores: (que como vimos son variables observables) los indicadores reflexivos que dependen de las variables latentes y los indicadores formativos los cuales causan a las variables latentes, cuya representación en el modelo es el siguiente, ver Gómez (2011).



Indicador reflexivo



Indicador formativo

Una vez revisado el tipo de escritura concerniente a las ecuaciones estructurales, tenemos la herramienta necesaria para analizar el modelo de ecuaciones estructurales.

Formalmente, el (MES) se compone de dos partes: el modelo de ecuaciones estructurales y el modelo de medición, Bollen (1989) . El modelo de ecuación

estructural puede ser representado por:

$$\eta = \mathbf{B}\eta + \mathbf{\Gamma}\mathbf{x} + \zeta$$

Donde cada x_i, $i = 1, ..., q$ en forma vectorial $\mathbf{x} = (x_1, x_2, ..., x_q)'$ es una causa potencial de una de las variables latentes contenidas en el vector η. Los coeficientes individuales en la matriz $\mathbf{\Gamma}$ describen las relaciones entre las variables latentes y sus causas. Cada variable latente es determinada por un conjunto de causas exógenas. Los términos de error en el vector ζ representan las componentes inexplicables. La matriz de covarianza la cual es abreviada por Ψ, Φ de $q \times q$ es matriz de covarinzas causales. La matriz de coeficientes $\mathbf{B}$ muestra la influencia de las variables latentes entre sí.

El modelo de medición vincula la variable latente a sus múltiples indicadores observables, es decir, se supone que la variable latente esta determina por sus indicadores. El modelo de medición proporciona información que los modelos de un solo indicador no. Esto es especificado por:

$$\mathbf{y} = \mathbf{\Lambda}\eta + \epsilon$$

En la que $\mathbf{y} = (y_1, y_2, ..., y_p)'$ es el vector de indicadores para las variables latentes contenidas en η, $\mathbf{\Lambda}$ es una matriz de coeficientes de regresión y ϵ es un vector $p \times 1$ de perturbaciones de ruído blanco, cuya matriz de covarianza de $p \times p$ está dada por Θ_ϵ. Los parámetros del modelo se estiman utilizando la información contenida en las variables observables de las matrices de varianza y covarianza. Por lo tanto, el objetivo del procedimiento de estimación es encontrar los valores de los parámetros y covarianzas que producen una estimación para el modelo de ecuaciones estructurales la matriz de covarianza $\Sigma(\Theta)$,$\widehat{\Sigma}(\Theta) = \Sigma\widehat{\Theta}$, que corresponda más estrechamente a la covarianza de la muestra de la matriz de las causas y los indicadores observados.

Para abordar este modelo se necesitará una herramienta estadística llamada análisis factorial, él cual es congruente con indicadores reflexivos del modelo de medida.

2.2. Análisis Factorial

Análisis factorial es un nombre genérico que se da a una clase de métodos estadísticos multivariantes cuyo propósito principal es definir la estructura subyacente en una matriz de datos. Generalmete hablando, aborda el problema de cómo analizar la estructura de las interrelaciones (correlaciones) entre un gran número de variables (por ejemplo, las puntuaciones de prueba, artículos

de prueba, respuestas de cuestionarios) con la definición de una serie de dimensiones subyacentes comunes, conocidas como factores. Con el análisis factorial, se identifica primero las dimensiones separadas de la estructura y entonces se determina el grado en que se justifica cada variable por cada dimensión. Una vez que se ha determinado estas dimensiones y la explicación de cada variable, se puede lograr los dos objetivos principales para el análisis factorial, que son: el resumen y la reducción de datos, Anderson (2010). A la hora de resumir los datos, con el análisis factorial se obtienen unas dimensiones subyacentes que, cuando son interpretadas y comprendidas , describen los datos con un número de conceptos mucho más reducido que las variables individuales originales. En conclusión Análisis factorial es un método estadístico que tiene como objetivo principal determinar si un conjunto de variables puede ser explicado por un número reducido de variables, llamadas factores (variables latentes), los cuales representarán a las variables originales, con una pérdida mínima de información, por tanto el análisis factorial es una técnica de reducción de dimensionalidad de datos. Como ejemplo supongamos que deseamos medir la capacidad mental (la cúal medimos con diferentes tipos de pruebas) de un individuo para procesar información y resolver problemas específicos, nos podriamos preguntar si existen algunos factores los cuales no sean necesariamente observables, que expliquen el conjunto de resultados observados con un pequeño error, el conjunto de estos factores es a los que llamamos inteligencia.

El análisis factorial es un método que se relaciona con componentes principales, sin embargo hay ciertas diferencias entre estos dos métodos, la diferencia primordial es la siguiente; los componentes principales se desarrollan para explicar las varianzas de las variables, mientras que en el análisis factorial construye factores con el fin de analizar y explicar las covarianzas y correlaciones que existen entre las variables.

Modelo

Para desarrollar el modelo factorial seguimos a Rencher (2002), supondremos que $\mathbf{y} \in \mathcal{R}^p$ es un vector de datos u observaciones, $\mathbf{v} \in \mathcal{R}^p$ es el vector de medias de $\mathbf{y}$, que $\mathbf{f} \in \mathcal{R}^m$ es un vector de factores (variables latentes) no observables que sigue una distribución $N_m(\mathbf{0}, \mathbf{I})$, esto implica que cada factor considerado, es una variable con media cero y varianza 1, independientes entre si, con distribucion normal univariada; que $\Delta \in M_{p\times m}(\mathcal{R})$ matriz de constantes desconocidas (llamada matriz de carga) donde $m < p$ que explica como los factores $\mathbf{f}$ afectan a los datos observados $\mathbf{y}$; $\mathbf{u} \in \mathcal{R}^p$ es un vector de errores no observables que sigue una distribución $N_p(\mathbf{0}, \psi)$, donde ψ es diagonal y finalmente que los factores $\mathbf{f}$ y los errores $\mathbf{u}$ no están correlacionados. Con

estas hipótesis el modelo factorial es de la siguiente manera:

$$\mathbf{y} = \mathbf{v} + \Delta\mathbf{f} + \mathbf{u} \tag{2.1}$$

Dado que $\mathbf{y}$ es suma de variables que se distribuyen normal, de la ecuación anterior tenemos que $\mathbf{y} \sim N_p(\mathbf{v}, \mathbf{V})$ ver Ash (2000), donde $\mathbf{V}$ es la matriz de covarianza de $\mathbf{y}$.

De la ecuación (2.1) obtenemos que para una muestra de tamaño n cada dato $y_{i,j}$ se puede expresar como:

$$y_{i,j} = v_j + \lambda_{j1}f_{1i} + ... + \lambda_{jm}f_{mi} + u_{ij} \quad i = 1, ..., n \quad j = 1, ..., p \tag{2.2}$$

Donde $y_{i,j}$ es valor de la i-esima muestra en la j-esima variable, colocando todas las escuaciones para todas las observaciones, la matriz de datos $\mathbf{Y}$ puede escribirse de la siguiente forma:

$$\mathbf{Y} = \mathbf{1}\mathbf{v}' + \mathbf{F}\Delta' + \mathbf{U} \tag{2.3}$$

Donde $\mathbf{1} \in \mathcal{R}^n$, vector de unos, $\mathbf{F} \in M_{n\times m}(\mathcal{R})$ matriz que contiene los m factores para los n elementos de la muestra , $\Delta' \in M_{m\times p}(\mathcal{R})$ es la traspuesta de la matriz de carga Δ y finalmente $\mathbf{U} \in M_{n\times p}(\mathcal{R})$ es la matriz de perturbaciones de las n muestras.

Una propiedad importante de la matriz de carga es la siguiente: la matriz de carga Δ contiene las covarianzas entre los factores y las variables observadas, en efecto, si multiplicamos la ecuación (2.1) por $\mathbf{f}'$ por la derecha y después tomamos la esperanza tenemos que:

$$E[(\mathbf{y} - \mathbf{v})\mathbf{f}'] = \Delta E[\mathbf{f}\mathbf{f}'] + E[\mathbf{u}\mathbf{f}'] = \Delta \tag{2.4}$$

ya que por hipótesis $E[\mathbf{u}\mathbf{f}'] = 0$ y $E[\mathbf{f}\mathbf{f}'] = \mathbf{I}$, por otro lado la matriz de covarianzas de los datos u observaciones satisface según la ecuación (2.1) la siguiente propiedad

$$\begin{aligned} \mathbf{V} = E[(\mathbf{y} - \mathbf{v})(\mathbf{y} - \mathbf{v})'] &= \Delta E[\mathbf{f}\mathbf{f}']\Delta' + E[\mathbf{u}\mathbf{u}'] \\ &= \Delta\mathbf{I}\Delta' + E[\mathbf{u}\mathbf{u}'] \\ &= \Delta\Delta' + \psi \end{aligned} \tag{2.5}$$

Esta descomposición de la matriz de varianza $\mathbf{V}$ es muy importante ya que es la suma de dos matrices, la primera matriz $\Delta\Delta' \in M_{m\times m}(\mathcal{R})$ es una matriz simétrica de rango $m < p$, esta matriz representa la parte común al conjunto de las variables y depende de forma única de las covarianzas entre los factores y las variables observadas; la segunda es la matriz ψ que como supusimos es

diagonal y contiene la parte específica de cada variable, que es independiente del resto.

La descomposición de (2.5) implica que las varianzas de las variables observadas tiene la siguiente forma:

$$\sigma_i^2 = \sum_{j=1}^{m} \lambda_{ij}^2 + \psi_i^2 \quad i = 1, ..., p \tag{2.6}$$

En la representación anterior de la varianza, si hacemos $h_i^2 = \sum_{j=1}^{m} \lambda_{ij}^2$ (llamada comunalidad) que no es más que la suma de los efectos de los factores , tenemos que:

$$\sigma_i^2 = h_i^2 + \psi_i^2 \quad i = 1, ..., p \tag{2.7}$$

En el análisis factorial tenemos que, tanto la matriz de carga como los factores son no observables, lo cual produce algunos problemas de indeterminación en el modelo en el siguiente sentido:

- Un conjunto de datos puede explicarse tanto por factores correlacionados como no correlacionados con la misma precisión
- Los factores con los que se pueden explicar las variables observadas no son únicos

A continuación análizamos estos dos tipos de indeterminaciones, para la primera indeterminación, notemos que si $\mathbf{f}$ son factores no correlacionados y $\mathbf{H}$ es una matriz no singular, entonces el modelo factorial en (2.1) es igual a

$$\mathbf{y} = \mathbf{v} + \Delta\mathbf{H}\mathbf{H}^{-1}\mathbf{f} + \mathbf{u}$$

si hacemos $\Delta^* = \Delta\mathbf{H}$ que será la nueva matriz de carga y llamando a $\mathbf{f}^* = \mathbf{H}^{-1}\mathbf{f}$ a los nuevos factores entonces

$$\mathbf{y} = \mathbf{v} + \Delta^*\mathbf{f}^* + \mathbf{u}$$

donde los factores $\mathbf{f}^* \sim N(\mathbf{0}, \mathbf{H}(\mathbf{H^{-1}})')$ por lo cual están correlacionados, de la misma forma si comenzamos nuestro análisis apartir de factores correlacionados, digamos $\mathbf{f} \sim N(\mathbf{0}, \mathbf{V_f})$ con $\mathbf{A}$ tal que $\mathbf{V_f} = \mathbf{A}\mathbf{A}'$ (la matriz $\mathbf{A}$ existe siempre que $\mathbf{V_f}$ sea definida positiva) así $\mathbf{A^{-1}}\mathbf{V_f}(\mathbf{A^{-1}})' = \mathbf{I}$ escribiendo

$$\mathbf{y} = \mathbf{v} + \Delta\mathbf{A}\mathbf{A^{-1}}\mathbf{f} + \mathbf{u}$$

haciendo $\Delta^* = \Delta\mathbf{A}$ y $\mathbf{f}^* = \mathbf{A^{-1}}\mathbf{f}$ entonces el modelo se convierte a uno nuevo pero con factores no correlacionados.

Ahora, para la segunda indeterminación, tenemos que si $\mathbf{H}$ es una matriz ortogonal entonces los modelos

$$\mathbf{y} = \mathbf{v} + \Delta\mathbf{f} + \mathbf{u}$$
$$\mathbf{y} = \mathbf{v} + (\Delta\mathbf{H})(\mathbf{H}'\mathbf{f}) + \mathbf{u}$$

son equivalentes o indistinguibles, los dos modelos poseen factores no correlacionados con matriz de covarianza igual a la identidad, por tanto, el modelo factorial está indeterminado bajo rotaciones; ambas indeterminaciones en el modelo se pueden anular poniendo algunas restricciones sobre la matriz de carga que veremos a continuación.

Debido a que el modelo factorial está indeterminado bajo rotaciones entonces el modelo no estará identificado, lo que supone en principio que aunque podamos observar toda la población y tanto como las medias $\mathbf{v}$ y la matriz de covarianzas $\mathbf{V}$ sean totalmente conocidas no podremos determinar la matriz de carga Δ de forma única, por tanto impondremos algunas restricciones a la matriz de carga y así pueda ser determinada de forma única; las restricciones serán las siguientes:

- **Restricción 1:** $\Delta'\Delta = \mathbf{D}$ (matriz diagonal)
- **Restricción 2:** $\Delta'\psi^{-1}\Delta = \mathbf{Z}$ (matriz diagonal)

se puede mostrar que cualquiera de las dos restricciones impuestas arriba conducen a unicidad en la matriz de carga, si suponemos que se satisface la restricción 1, postmultiplicamos la ecuación (2.5) por la matriz Δ y entonces podemos escribir

$$(\mathbf{V} - \psi)\Delta = \Delta\mathbf{D} \tag{2.8}$$

donde deducimos que las columnas de la matriz Δ son los vectores propios de la matriz $\mathbf{V} - \psi$, que tiene como valores propios a los elementos de la matriz diagonal $\mathbf{D}$; por otro lado si suponemos que se satisface la restricción 2, postmultiplicando la ecuación (2.5) por el término $\psi^{-\mathbf{1}}\boldsymbol{\Delta}$ tenemos

$$\mathbf{V}\psi^{-\mathbf{1}}\Delta - \Delta = \Delta\mathbf{Z} \tag{2.9}$$

ahora premultiplicando por la matriz $\psi^{-\mathbf{1/2}}$, realizando algunas descompocisiones y operaciones tenemos

$$\psi^{-\mathbf{1/2}}\mathbf{V}\psi^{-\mathbf{1/2}}\psi^{-\mathbf{1/2}}\Delta = \psi^{-\mathbf{1/2}}\Delta(Z + I) \tag{2.10}$$

de aqui podemos observar que la matriz $\psi^{-\mathbf{1/2}}\mathbf{V}\psi^{-\mathbf{1/2}}$ tiene como vectores propios $\psi^{-\mathbf{1/2}}\Delta$ con valores propios $(\mathbf{Z} + \mathbf{I})$.

Ahora para verificar que el sistema sea determinado en el sentido usual debe haber un número de ecuaciones igual o mayor que el de incognitas. Por lo tanto una condición suficiente para que el sistema sea determinado, es la siguiente

$$(p-m)^2 \geq p+m$$

donde m es el número de factores y p el número de variables.

Estimación

Una vez análizado el modelo y ya que hemos sacado a la luz propiedades importantes de este, es hora de dar un método de estimación de la matriz de carga, usaremos el método de factor principal el cúal es basado en componentes principales, esté método tiene la ventajas sobre otros; ya que la dimensión del sistema puede identificarse de forma aproximada, también evita el tener que resolver ecuaciones de máxima verosimilitud. Sus fundamentos son los siguientes, supongamos por un momento que podemos obtener una estimación inicial de la matriz de varianzas de los errores $\hat{\psi}$, entonces

$$\mathbf{S} - \hat{\psi} = \Delta\Delta'$$

Donde $\mathbf{S}$ es la matriz de varianza muestral; como $\mathbf{S}$ es simétrica y $\hat{\psi}$ es diagonal entonces $\mathbf{S} - \hat{\psi}$ será simétrica y por tanto siempre podra descomponerse como;

$$\mathbf{S} - \hat{\psi} = \mathbf{HGH'} = (\mathbf{HG^{1/2}})(\mathbf{HG^{1/2}})' \tag{2.11}$$

donde $\mathbf{H}$ es una matriz cuadrada diagonal de orden p y ortogonal, que contiene los eigenvalores de la matriz $\mathbf{S} - \hat{\psi}$. Por tanto el modelo factorial establece que $\mathbf{G}$ debe ser diagonal de la siguiente forma

$$\mathbf{G} = \begin{bmatrix} \mathbf{G_{1m\times m}} & \mathbf{0_{m\times(p-m)}} \\ \mathbf{0_{(p-m)\times m}} & \mathbf{0_{(p-m)\times(p-m)}} \end{bmatrix} \tag{2.12}$$

debido a que $\mathbf{S} - \hat{\psi}$ tiene rango m, si llamamos $\mathbf{H_1}$ a la matriz de $p \times m$ que contiene los vectores propios asociados a los valores propios no nulos de $\mathbf{G_1}$ entonces podemos tomar como estimador de la matriz de carga Δ la matriz:

$$\hat{\Delta} = \mathbf{H_1 G_1^{1/2}} \tag{2.13}$$

con $\hat{\Delta} \in M_{p\times m}(\mathcal{R})$ y esto resolveria el problema de la estimación, notemos que

$$\hat{\Delta}\hat{\Delta}' = \mathbf{G_1^{1/2} H_1' H_1 G_1^{1/2}} = \mathbf{G_1} \tag{2.14}$$

donde la matriz $\mathbf{G_1}$ es diagonal, debido a que los vectores propios de matrices simétricas son ortogonales y así $\mathbf{H_1' H_1} = \mathbf{I_m}$, en general y en la practica la estimación de la matriz de carga se lleva a cabo de la siguiente manera iterativa:

- **I**. Comenzar de una estimación inicial de la matriz $\hat{\Delta}_i$ o de la matriz $\hat{\psi}_i$ con $\hat{\psi}_i = diag(\mathbf{S} - \hat{\Delta}\hat{\Delta}')$.
- **II**. Calcular la matriz $\mathbf{Q}_i = \mathbf{S} - \hat{\psi}_i$ la cual resulta simétrica y cuadrada.
- **III**. Obtener la descomposición espectral de la forma siguiente

$$\mathbf{Q}_i = \mathbf{H_{1}}_i\mathbf{G_{1}}_i\mathbf{H_{1}}_i' + \mathbf{H_{2}}_i\mathbf{G_{2}}_i\mathbf{H_{2}}_i' \tag{2.15}$$

  con la matriz $\mathbf{G_1}_i$ conteniendo los m mayores valores propios de la matriz $\mathbf{Q}_i$ y $\mathbf{H_1}_i$ sus respectivos valores propios, mientras tanto la matriz $\mathbf{G_2}_i$ contendrá los valores propios restantes de manera que estos valores sean pequeños y similares entre sí, puede haber algunos problemas con la matriz $\mathbf{Q}_i$, ya que puede resultar que no sea definida positiva y algunos de sus valores propios ser negativos, sin embargo si son muy cercanos a a cero, los tomaremos como cero.

- **IV**. Tomar $\hat{\Delta}_{i+1} = \mathbf{H_1}_i\mathbf{G_1}_i^{\mathbf{1/2}}$ y volver a comenzar el procesos desde **I**, continuar iterando hasta que, dado $\epsilon > 0$ tengamos que $\|\Delta_{j+1} - \Delta_j\| < \epsilon$ para j suficientemente grande.

con este proceso que acabamos de explicar, los estimadores obtenidos serán consistentes, pero no eficientes, tampoco serán invariantes bajo transformaciones lineales. En todo este análisis hecho para obtener la estimación de la matriz de carga, hicimos la suposición que teniamos una estimacion para la matriz ψ a saber $\hat{\psi}$, entonces debemos especificar como es que se puede suponer en principio el estimador $\hat{\psi}$, esté problema es llamado la estimación de comunalidades.

Estimación de comunalidades

Estimar los términos $\hat{\psi}_i$ es una forma equivalente a definir los valores para los términos diagonales, h_i^2, de $\Delta\Delta'$, ya que, como sabemos, $h_i^2 = s_i^2 - \hat{\psi}_i^2$, por lo cual existen dos formas de hacerlo:

- **1.** Tomar $\hat{\psi}_i = 0$
- **2.** Tomar $\hat{\psi}_i = 1/s^*_{jj}$ donde s^*_{jj} es el elemento diagonal de la matriz de presición $\mathbf{S^{-1}}$

si tomamos **1.** será equivalente a extraer los componentes principales de $\mathbf{S}$ y supone tomar a $\hat{h_i^2} = s_i^2$; sí tomamos **2.** tendrémos que $\hat{h_i^2} = s_i^2 - s_i^2(1 - R^2) =$

$s_i^2 R_i^2$, en donde el valor R_i^2 se conoce como el coefieciente de correlación múltiple entre la variable x_i y el resto de las variables.

Más generalmente el método de estimación visto hasta ahora (factor principal) es equivalente a maximizar la función siguiente:

$$F = tr(\mathbf{S} - \Delta\Delta' - \psi)^2 \tag{2.16}$$

Ya que está función se puede escribir de la sigiuente manera

$$F = \sum_{i=1}^{p}\sum_{i=1}^{p}(s_{ij} - v_{ij}) \tag{2.17}$$

donde los valores v_{ij} en la ecuación son los elementos de la matriz $\mathbf{V} = \Delta\Delta' + \psi$ ahora bien, por la descomposición espectral, tenemos que dada una matriz $\mathbf{S}$ cuadrada, simétrica y no negativa la mejor aproximación en el sentido de mínimos cuadrados (2,17) mediante una matriz de rango m, $\mathbf{AA'}$ se obtiene tomando a $\mathbf{A} = \mathbf{HD^{1/2}}$, donde $\mathbf{H}$ contiene los vectores propios y $\mathbf{D^{1/2}}$ las raíces de los valores propios de la matriz $\mathbf{S}$, que es lo que hace el método de factor principal.

Existen otros métodos de estiamación para encontrar los valores de los parámetros de la matriz de carga, por ejemplo; máxima verosimilitud los cuales se obtendrían maximizando la siguiente función:

$$\mathbf{L}(\Delta, \psi) = -\frac{n}{2}(log|\Delta\Delta' + \psi| + tr(\mathbf{S}(\Delta\Delta' + \psi))^{-1}) \tag{2.18}$$

y de la misma forma como en caso de factor principal, existe un método iterativo que se basa en Newton-Rapshon para encontrar la estimación.

Nostros preferimos el método de factor principal ya que para el problema que necesitamos resolver es suficiente con esté método.

La correspondencia del modelo de medida y análisis factorial

En el análisis factorial ver Anderson (2010), cada variable individual es "explicada" por su ponderación en cada factor. El objetivo es representar lo mejor posible todas las variables en un número reducido de factores, es decir, los factores referidos a "dimensiones subyacentes" de los datos, que después tendremos que interpretar y clasificar. El análisis factorial comunmente se clasifica como una técnica exploratoria por que no existen restricciones sobre las cargas de las variables. Cada variable tiene una carga sobre cada factor. El valor de cada factor (puntuación del factor) se calcula mediante las cargas sobre cada

factor por ejemplo: $factor\,1 = \lambda_{11}y_1 + ... + \lambda_{51}y_5$ ver Anderson (2010) donde y_1 hasta y_5 son los valores efectivos de cada variable. También el valor predictor para cada variable se calcula mediante las cargas de la variable para cada factor. Sin embargo cada variable tiene una carga factorial; por lo tanto cada factor es siempre una composición de todas sus variables, aunque sus cargas varíen en magnitud. Por consiguiente, un factor es en realidad un constructo latente, definido por las cargas de todas sus variables.

Análisis factorial: cargas factoriales				modelo de medida		
Varible	Factor 1	Factor 2	Factor 3	Constructo A	Constructo B	Constructo C
x_1	λ_{11}	λ_{12}	λ_{13}	λ_1		
x_2	λ_{21}	λ_{22}	λ_{23}	λ_2		
x_3	λ_{31}	λ_{32}	λ_{33}		λ_3	
x_4	λ_{41}	λ_{42}	λ_{43}		λ_4	
x_5	λ_{51}	λ_{52}	λ_{53}			λ_5

Cuadro 2.1: Ejemplo

Para especificar el modelo de medida, hacemos la transición desde el análisis factorial, en el que el investigador, no tiene control sobre qué variables describen cada factor, a un modo confirmatorio, en el que el investigador especifica que variables definen cada constructo. Las variables observadas que obtenemos de los encuestados se denominan en este caso indicadores en el modelo de medida, por que los utilizamos para medir o "indicar", los constructos (factores) latentes. Supongamos en el ejemplo de arriba que y_1 y y_2 son indicadores del constructo A, y_3 e y_4 son indicadores del constructo B y que y_5 es un único indicador del constructo C. El modelo de medida entonces quedará indicado como en el ejemplo del cuadro anterior.

La diferencia de está configuración de cargas entre análisis factorial y el modelo de medida, es el reducido número de de ponderaciones. En el modelo explicativo del análisis factorial, el investigador no puede controlar las ponderaciones. En en modelo de medida sin embargo el investigador tiene un control completo sobre las variables descritas por cada constructo.

Una vez resuelto el modelo de medida por el método de análisis factorial se procede hacer regresión en el modelo de ecuaciones estructurales para obtener los puntajes de las variables latentes.

2.3. Economía sumergida y corrupción: Un enfoque con el modelo de ecuaciones estructurales

En este apartado abordaremos el modelo propuesto por Buehn (2009), esté plantea ecuaciones estructurales (MES) para modelar la corrupción y la economía sumergida como dos variables latentes distintas y explora su relación usando las estructuras de covarianza entre las causas e indicadores observables de estas variables latentes. Antes de analizar este modelo veamos que es la economía sombria o de las sombras.

La economía sombria es un fenómeno económico no observable, y no existe consenso en cuanto su definición. Por ejemplo, Smith (1994) lo define como "la producción basada en el mercado de bienes y servicios, ya sea legal o ilegal, que escapa a la detección en las estimaciones oficiales del PIB".

En el modelo propuesto por Buehn (2009) se toman en cuenta las siguientes variables en el modelo estructural; economía sumergida y a la corrupción, el vector η es un vector de dimensión 2, que contiene a la economía sumergida y a la corrupción.

Las causas de la economía sumergida, determinada por evidencia empírica y teórica, según estos autores son: x_1 = consumo del gobierno, x_2 = transferencias y subsidios, x_3 = tamaño del gobierno, x_4 = regulaciones del mercado laboral, x_5 = regulación de negocios y x_6 = tasa de desempleo. Además de las variables causales que determinan el tamaño y el desarrollo de la economía sumergida se utilizan tres variables indicadoras que la hacen visible, estas son: y_1 = tasa de M0 a M1, y_2 = tasa de crecimiento real del PIB y y_3 = tasa de actividad laboral.

De igual manera la selección de las causas y los indicadores de corrupción se basaron en hallazgos previos de la literatura teórica y empírica relevante, Buenh & Schneider (2009). Las causas de la corrupción propuestas son: x_7 = efectividad del gobierno, x_8 = reglas de la ley, x_9 = matriculación escolar, x_{10} = costos de la burocracia. y x_{11} = libertad físcal, por otro lado sus indicadores son: y_4 =PIB real per cápita, y_5 = pago de sobornos, y_6 = independencia judicial, y_7 = ausencia de la Corrupción.

El diagrama propuesto por el autor es el siguiente:

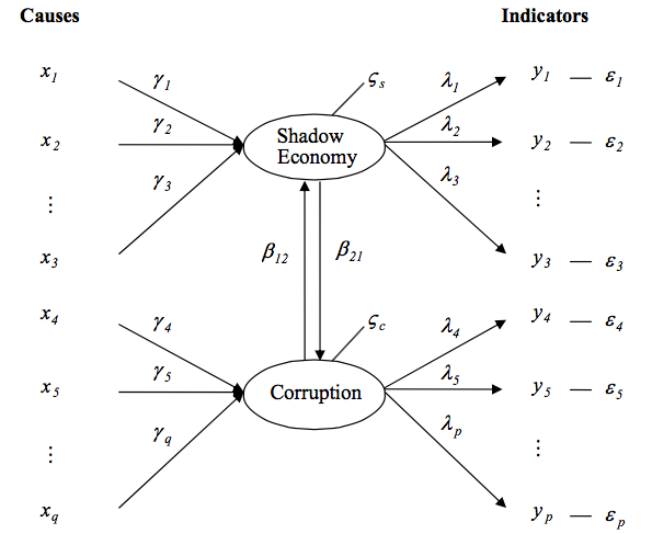


Figura 2.1: La economía sumergida y la corrupción son variables no observables, o latentes. Las variables x son causantes de la variables latentes y son observables, finalmente las variables y son indicadores que también son observables.

La forma matricial del modelo extructural es:

$$\begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} 0 & \beta_{12} \\ \beta_{21} & 0 \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} + \begin{bmatrix} \gamma_1 \dots \gamma_6 & 0 \dots 0 \\ 0 \dots 0 & \gamma_7 \dots \gamma_{11} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ \vdots \\ x_{11} \end{bmatrix} + \begin{bmatrix} \zeta! \\ \zeta_2 \end{bmatrix}$$

La forma matricial del modelo de medida es:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 \\ \lambda_2 & 0 \\ \lambda_3 & 0 \\ 0 & \lambda_4 \\ 0 & \lambda_5 \\ 0 & \lambda_6 \\ 0 & \lambda_7 \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \end{bmatrix}$$

Este modelo se aplicó a los datos obtenidos de 51 países tomados en muestras anuales, desde el año 2000 hasta el año 2005, Buehn (2009) dando los siguientes resultados:

Corridas		**1**		**2**		**3**		**4**		**5**
variables latentes	SE	C	SE	C	SE	C	SE	C	SE	C
causas										
Regulación de negocios	.18** (2)		.13* (1.84)		.21** (2.18)		.18* (2.02)		.18** (1.98)	
Desempleo	.19** (.98)				.16* (1.78)		.17* (1.93)		.2** (2.02)	
Transferencias y subsidios	.09 (1.16)		.05 (1.09)		.11 (1.35)		.09 (1.22)		.09 (1.15)	
Consumo de gobierno	.16** (1.98)		.11* (1.76)				.15* (1.91)		.17* (2.05)	
Regulaciones de mercado laboral			.22** (2.05)							
Tamaño de gobierno					.14* (1.66)					
Efectividad de gobierno		-.22*** (3.13)		-.15** (2.25)		-.2*** (2.66)		-.23*** (3.36)		-.21*** (3.01)
Libertad fiscal		-.15*** (2.48)		-.09* (1.81)		-.15*** (2.27)		-.14** (2.37)		-.17*** (2.68)
Costo burcratico		.42*** (5.15)		.34** (2.95)		.41*** (4.29)		.40*** (4.79)		.45*** (5.52)
Reglas de la ley		-.01 (.10)		.01 (.19)		-.01 (.09)				-.02 (.38)
Costo burcratico		.42*** (5.15)		.34** (2.95)		.41*** (4.29)		.40*** (4.79)		.45*** (5.52)
Matriculación escolar										.06 (1.01)

.

Corridas		**1**		**2**		**3**		**4**		**5**
variables latentes	SE	C	SE	C	SE	C	SE	C	SE	C
indicadores										
crecimiento GDP	-.51		-.47		-.46		-.5		-.51	
Tasa de M0 a M1	.41** (4.15)		-.44*** (4.02)		-.43*** (4.04)		-.41*** (4.13)		-.4*** (4.15)	
Real GDP per capita		-.78		-.75		-.74		-.78		-.77
Sobornos		.15* (1.73)		.16** (1.99)		.16* (1.95)		.15** (1.74)		.14* (1.71)
Independencia judicial		-.06 (.73)		-.08 (.99)		-.07 (.8)		-.06 (.71)		
Libertad de corrupción		-.01 (.10)		.01 (.19)		-.01 (.09)				-.02 (.38)

En las tablas de arriba los coeficientes de senderos son los elementos de los segundos renglones que están en parentesis dentro de la tabla, Donde * es la significancia al nivel del 10 %, ** es la significancia al nivel del 5 %,*** es la significancia al nivel del 1 %.

En la siguiente tabla observamos los coeficientes de senderos que son los elementos de los segundos renglones que están en parentesis dentro de la tabla, es decir el coeficiente de sendero de $E.S. \longrightarrow C$ correspondiente a la primera corrida del modelo es el valor 4.23, y el coeficiente de sendero de $C \longrightarrow E.S.$ de la primera corrida del modelo es el valor 2.64, obtenidos apartir de un muestreo de 17 variables correspondientes a 51 países, Donde *** es la significicancia al nivel del 1 %.

Variables latentes	**1**	**2**	**3**	**4**	**5**
E.S. → C	.68***	1.07***	.81***	.69***	.67***
	(4.23)	(4.34)	(3.98)	(4.19)	(4.23)
C → E.S.	.42***	.43***	.37***	.47***	.39***
	(2.64)	(2.7)	(2.27)	(2.95)	(2.50)

Se puede observar una relación positiva, robusta y estable entre las variables, lo cuál indica que altos niveles de economía sumergida están ligados a altos niveles de corrupción, el modelo revela que una gran economía sumergida está ligada a los altos niveles de corrupción Buehn (2009). Por lo tanto, la relación empírica entre la corrupción y la economía sumergida existe, Buehn (2009).

Capítulo 3

Aprendizaje no supervisado

El aprendizaje no supervisado es un concepto muy profundo que puede ser abordado desde diferentes perspectivas, como la psicología y la ciencia cognitiva de la matemática. Muy a menudo se le llama "aprender sin maestro". A continuación explicaremos en que consiste el aprendizaje supervisado y no supervisado, cuales son las diferencias que existen entre ellos y los problemas específicos que aborda.

Por lo regular en la estadística y en muchas ramas de la matemática lo que se busca es predecir los valores de una o más variables, llamadas respuesta $Y = (Y_!, ..., Y_M)$ para un cierto conjunto de variables de entrada ó predictivas $X^T = (X_1, ..., X_P)$, denotamos por $X_i^T = (x_{i1}, ...x_{ip})$ a la entradas del i-esimo caso de entrenamiento, tomemos a y_i como una medida de respuesta, entonces las predicciones se basan en la muestra de entrenamiento $(x_1, y_1), ..., (x_N, y_N)$, si todas las variables son conocidas, se dice que el aprendizaje es supervisado ó "el aprendizaje es con un maestro". Bajo esta metáfora del "estudiante"presenta una respuesta y_i para cada x_i en la muestra de entrenamiento, y el supervisor o "maestro" ofrece ya sea la respuesta correcta y/o un error asociado con la respuesta del estudiante.

Si suponemos que (X, Y) son variables aleatorias que están representadas por alguna densidad de probabilidad conjunta, digamos $Pr(X, Y)$, entonces el aprendizaje supervisado se puede caracterizar formalmente como un problema de estimación de la densidad, donde uno se refiere a propiedades determinantes de la densidad condicional $Pr(Y|X)$. Recordemos que la distribucion conjunta y la distribución condional se relacionan por $Pr(X, Y) = Pr(Y|X) \cdot Pr(X)$, donde $Pr(X)$ es la densidad marginal conjunta de los valores de X por sí solos. En el aprendizaje supervisado $Pr(X)$ no presenta algún interés directo. Uno

está interesado principalmente en las propiedades de la densidad condicional $Pr(Y|X)$. Ya que a menudo Y es la variable de interés.

Por otro lado el aprendizaje no supervisado o "aprendizaje sin maestro", tiene el objetivo de inferir directamente las propiedades de la densidad conjunta $Pr(X)$ del vector aleatorio $X \in \mathbb{R}^p$ sin la ayuda de un supervisor o maestro que proporcione respuestas o grados de error para cada observación correcta directamente de la base de datos. La dimension de X es a veces mucho mayor que en el aprendizaje supervisado, y las propiedades de interés son amenudo más complejas. Estos factores están algo mitigados por el hecho de que X representa todas las variables en estudio; no se requiere para inferir cómo las propiedades de $Pr(X)$ cambian, condicionadas a los valores cambiantes de otro conjunto de variables.

Cuando (p $\leq$ 3), podemos estimar directamente la densidad $Pr(X)$ para todos los valores de X, mediante la frecuencia relativa. SI la dimensión de X es alta (p $\geq$ 3) esté método falla, dado que si tenemos muchos posibles valores de X la probabilidad para cada uno de estos valores es pequeña y por lo tanto conviene considerar conjuntos de valores de X que tengan probabilidades relativamente altas. Componentes principales, escalamiento multidimensional, mapas auto-organizados, y curvas principales, por dar un ejemplo, tratan de identificar las variedades de baja dimensionalidad dentro del espacio X, que destacan por su gran densidad de datos. Esto proporciona información sobre las asociaciones entre las variables. Las reglas de asociación intentan construir descripciones simples (reglas conjuntivas) que describen las regiones de alta densidad, en el caso especial de datos binarios, de muy altas dimensiones. Está técnica es presentada en la siguiente subsección.

3.1. Reglas de asociación

El análisis de reglas de asociación se ha convertido en una de las herramientas más populares para la minería de bases de datos comerciales. Supongamos que tenemos P vectores $X = (X_1, X_2, ..., X_p)$ correspondientes a N individuos en una base de datos, supongamos que $p = 4$ y que las variables X_i son las siguientes; $X_1 = Estado\,civil$, $X_2 = Sexo$, $X_3 = Edad$, $X_4 = Percepcion\,de\,ganacias$, entonces $X = (\,Estado\,civil,\ Sexo, Edad, Percepcion\,de\,ganacias)$ cuyas entradas toman los valores $X_1 \in \{\,Soltero, Casado, Divorciado, Viudo\}$, $X_2 \in \{\,Masculino, Femenino\}$, $X_3 \in \{1-20, 21-45, 46-100\}$, $X_4 \in \{\leq 120000,\ \geq 120001\}$ es decir, por cada individuo en nuestra base de datos se posee la informacioñ referente al $Estado\,civil$, $Sexo$, $Edad$ y $Percepcion\,de\,ganancias$, entonces el objetivo general de las reglas de asociación es encontrar conjun-

tos de coordenadas del vector X cuya ocurrencia sea alta, en otras palabras aquellas entradas del vector que en la base de datos tengan alta frecuencia de ocurrir.

Nuestra base de datos entonces posee un total de 5 columnas ordenadas de la siguiente manera; columna 1= individuo, columna 2 = estado civil, columna 3 = sexo, columna 4 = edad, columna 5 = percepción de ganancias, ahora el primer paso en las reglas de asociación, es transformar la base de datos original. Esto se hará de la siguiente manera; se sustituira la j-esima columna asociada a la variable X_j del vector X por tantas columnas como el cardinal de valores posibles que tome la j-esima columna. En nuestra base original tenemos un total de 5 columnas, nos fijamos en la columna uno "$Estado\,civil$" que corresponde a la primera entrada del vector X, entonces esta columna se sustituira por un total de 4 columnas, ya que $X_1 \in \{\, Soltero, Casado, Divorciado, Viudo\}$, la columna dos "$Sexo$" correspondiente a la segunda entrada del vector X será sustituida por un total de 2 columnas debido a que $X_2 \in \{Masculino, Femenino\}$ y asi sucesivamente, al final remplazaremos las 5 columnas originales por un total de 12 columnas.

De manera más general dado $X = (X_1, X_2, ..., X_p)$, para cada X_j sea S_j el conjunto de valores posibles que puede tomar la j-esima entrada del vector X, esto es $X_j \in S_j$, y así en nuestra base tenemos que.

$$\begin{aligned} X_1 &\in \{Soltero, Casado, Divorciado, Viudo\} = S_1 \\ X_2 &\in \{Masculino, Femenino\} = S_2 \\ X_3 &\in \{1-20, 21-45, 46-100\} = S_3 \\ X_4 &\in \{\leq 120000, \geq 120001\} = S_4 \end{aligned}$$

De ahora en adelante llamaremos atributo a cualquier posible valor que pueda tomar cualquier entrada del vector X, (note que cada atributo corresponde a una columna en nuestra nueva base de datos) entonces tenemos que el número total de atributos es:

$$K = \sum_{j=1}^{p} |S_j| \tag{3.1}$$

Donde $|S_j|$ es el número de distintos valores alcanzados por la entrada X_j, de la discusión anterior observamos que el número total de columnas en la base de datos nueva es $K + 1$.

El siguiente paso es crear variables ficticias apartir de las variables originales, el método es el siguiente; por cada atributo i crearemos una variable ficticia Z_i que tomara los siguientes valores $Z_i = 1$ si nuestro j-ésimo individuo en la base de datos tiene el atributo i y $Z_i = 0$ si nuestro individuo no posee el atributo i. así el número total de variables fícticias creadas coincide con el valor

de K, todo lo anterior se resume en transformar una base de datos cualquiera en una base de datos que toma valores binarios. Las siguientes tablas ilustran lo explicado arriba.

Individuo	**Estado civil**	**Sexo**	**Edad**	**Percepcion de ganancias**
1	Soltero	Femenino	18	130000
2	Divorciado	Masculino	36	85000
3	Casado	Masculino	27	95000
4	Divorciado	Femenino	50	101000
5	Casado	Masculino	38	70000
6	Divorciado	Femenino	40	155000
7	Viudo	Masculino	42	145000
8	Soltero	Femenino	60	160000
9	Divorciado	Femenino	53	92000
10	Divorciado	Masculino	43	85000
11	Soltero	Femenino	20	125000
12	Divorciado	Femenino	65	70000

Cuadro 3.1: Esta tabla muestra la información de 12 personas en la base de datos original, tiene un total de 5 columnas.

Individuo	Soltero	Casado	Divorciado	Viudo	Masculino	Femenino	1-20	21-45	46-100	≤120000	≥120001
1	1	0	0	0	0	1	1	0	0	0	1
2	0	0	1	0	1	0	0	1	0	1	0
3	0	1	0	0	1	0	0	1	0	1	0
4	0	0	1	0	0	1	0	0	1	1	0
5	0	1	0	0	1	0	0	1	0	1	0
6	0	0	1	0	0	1	0	0	1	0	1
7	0	0	0	1	1	0	0	1	0	0	1
8	1	0	0	0	0	1	0	0	1	0	1
9	0	0	1	0	0	1	0	0	1	1	0
10	0	0	1	0	1	0	0	1	0	1	0
11	1	0	0	0	0	1	1	0	0	0	1
12	0	0	1	0	0	1	0	0	1	1	0

Cuadro 3.2: Base de datos modificada, con 12 columnas totales"$K = 12$".

Análisis de canasta de mercado

Considere un supermercado con una gran colección de artículos. Algúnas decisiones empresariales típicas en la administración del supermercado son, ¿qué poner a la venta?, ¿cómo diseñar cupones?, ¿cómo colocar la mercancía en estantes con el fin de maximizar la ganacia?, etc. Análisis de datos de transacciones pasadas es un enfoque comúnmente utilizado con el fin de mejorar la calidad de tales decisiones. Hasta hace poco, sin embargo, sólo los datos globales sobre las ventas acumuladas durante algun periodo de tiempo (un dia, una semana, un mes, etc...) estaban disponibles en la computadora. Progresos en la tecnología de código de barras han hecho posible almacenar los llamados datos de cesta, que almacena la cesta de artículos comprados en un una base de datos por transaccion. Transacciones de tipo de datos de canasta no necesariamente constarán de artículos comprados juntos en el mismo punto de tiempo. Puede consistir en artículos comprados por un cliente durante un período de tiempo.

Ejemplos incluyen compras mensuales de los miembros de un club de lectura o una Club de música.

Comúnmente el ánalisis de la canasta de mercado esta vinculado a transacciones almacenadas en las bases de datos de las cajas registradoras de una tienda, estas transacciones son realizadas por clientes de la tienda. El objetivo del análisis de canasta de mercado es determinar que tipos de artículos son comprados con mayor frecuencia por los clientes, y además que combinaciones de estos artículos son más frecuentes en las canastas de los clientes. Más generalmente el objetivo de la canasta de mercado es descubrir conocimiento apartir de un conjunto de transacciones. En nuestro caso consideraremos a las transacciones como los individuos en nuestra base de datos y los artículos registrados serán los atributos asociados a los individuos en nuestra base.

Así el objetivo de la canasta de mercado se convierte en encontrar un subconjunto de números enteros $\mathcal{K} \subset \{1, ..., K\}$ tal que:

$$\mathbf{Pr}\left[\bigcap_{k\in\mathcal{K}}(Z_k = 1)\right] = \mathbf{Pr}\left[\prod_{k\in\mathcal{K}} Z_k = 1\right] \tag{3.2}$$

sea grande. Esta es la formulación estándar del problema cesta de mercado. El conjunto $\mathcal{K}$ se denomina "conjunto de objetos ó atributos". El número de variables Z_k en el conjunto de objetos se llama su "tamaño". La estimación del valor de (3.2) se toma como la fracción de observaciones en la base de datos para los que la conjunción es verdadera, en otras palabras; el valor de (3.2) es estimado como la proporción de individuos dentro de la base de datos que posean los atributos dados por el conjunto $\mathcal{K}$, es decir.

$$\hat{\mathbf{Pr}}\left[\prod_{k\in\mathcal{K}} Z_k = 1\right] = \frac{1}{N}\sum_{i=1}^{N}\prod_{k\in\mathcal{K}} Z_{ik} \tag{3.3}$$

en (3.3) z_{ik} se refiere al valor de la variable Z_k para el i-ésimo individuo en la base de datos, al valor de (3.3) se le llama el "soporte" o "prevalencia" del conjunto de atributos $\mathcal{K}$ y se denota cómo $T(\mathcal{K})$. A aquellos individuos en la base de datos para los cuales se tenga que $\prod_{k\in\mathcal{K}} Z_{ik} = 1$ significará que poseen todos los atributos de el conjunto $\mathcal{K}$. En la minería de reglas de asociación se especifica un soporte t inferior, y uno busca todo los conjuntos de elementos del conjunto $\mathcal{K}_l \subset \{1, 2, 3, ...K\}$ que se pueden formar a partir de las variables $Z_1, ..., Z_K$ con el soporte en la base de datos mayor que este límite inferior t.

$$\{\mathcal{K}_l | T(\mathcal{K}_l) > t\} \tag{3.4}$$

Veamos como se aplica lo discutido arriba en nuestra base. Supongamos que deseamos todas los conjuntos de atributos que tienen prevalencia $T(\mathcal{K}) >. 15$,

recordemos que $K = 12$ (número de columnas en la nueva base) y $N = 12$ (número de individuos), si se observa detinidamente el cuadro 3.2 y se aplica (3.3) obtendremos que los conjuntos de atributos que satisfacen $T(\mathcal{K}) > t$ son $\{soltero, femenino, 1-20, \geq 120001\}$, $\{casado, masculino, 21-45, \leq 120000\}$, $\{divorciado, masculino, 21-45, \leq 120000\}$ los cuales tienen una prevalencia $T(\mathcal{K}) =. 16666$. Hasta ahora hemos visto el enfoque de Friedman (2008), sin embargo en Rakesh (2010) y Pang-Ning (2006) el tema es manejado desde un punto de vista más tradicional.

Nota: En nuestra base de datos nuestras categorias son excluyentes, y con excluyentes hacemos referencia a lo siguiente, si un individuo en la base es soltero, entonces en la columna correspondiente "columna 1" tendra un uno y por tanto tres ceros correspondientes a las columnas siguientes las cuales pertenecen a la categoria de estado civil de ese individuo, que son Casado, Divorciado y Viudo, sín embargo en las reglas de asociación, más generalmente, en la canasta de mercado no tiene por que ser así.

Algoritmo Apriori

La solución al problema en (3.4) se puede obtener para bases de datos muy grandes siempre y cuando el umbral t sea alto o el número de atributos sea pequeño, ya qué (3.4) se compone de sólo una pequeña fracción de todos los posibles conjuntos de elementos del conjunto potencia 2^K. El algoritmo "A priori" explota las propiedades de que: para un umbral Friedman (2008) de prevalencia t dado , se satisface:

- El cardinal de $|\mathcal{K}|T(\mathcal{K}) > t|$ es relativamente pequeño.
- Cualquier conjunto de objetos L que consiste en un subconjunto de los elementos de $\mathcal{K}$ debe tener soporte mayor que o igual que la de $\mathcal{K}$, es decir, sí $L \subset \mathcal{K} \Longrightarrow T(L) \geq T(\mathcal{K})$.

El algoritmo A priori Troncoso (2002) funciona de la siguiente manera: el primer paso consiste en simplemente determinar los items que existen en la base de datos y crear el conjunto de artículos de tamaño uno cuyo soporte exceda al umbral t, en el n-esimo paso $n \geq 1$ se calcula el soporte los subconjuntos $\mathcal{K}_l \subset \{1, 2, 3, ..., K\}$ de n elementos para determinar si estos sobrepasan el umbral t ($T(\mathcal{K}_l) > t$) con la condición adicional de que los subconjutos con $n-1$ elementos de $\mathcal{K}_l$ tambien sobrepasen dicho umbral t. Una de las ventajas de el algoritmo Apriori es que requiere sólo una pasada de los datos para cada valor de $|\mathcal{K}|$, que es crucial, ya que se supone que los datos no se pueden instalar en

la memoria principal de una computadora. Si los datos son suficientemente escasos (o si el umbral t es lo suficientemente alto), entonces el proceso terminará en un tiempo razonable, incluso para grandes conjuntos de datos.

Cada conjunto de artículos $\mathcal{K}$ de alto soporte (establecido) devuelto por el algoritmo Apriori es echado en un conjunto de "reglas de asociación". El conjunto $\mathcal{K}$ se dividira en dos subconjuntos ajenos y disjuntos A y B. Esto es, $A \cup B = \mathcal{K}$ y $A \cap B = \emptyset$. Escribimos la relación entre A y B como $A \Rightarrow B$ con A y B como arriba, en este caso al conjunto A le llamaremos "antecedente" y el conjunto B será llamado el "consecuente" de la "regla de asociación". Esta relación significa que cada vez que se tienen los atributos de A con alta probabilidad tambien se tienen los atributos de B, por ejemplo en nuestra base de datos si un individuo es del sexo $Masculino$ y gana menos de \$120000 es muy probabable que su $Edad$ este entre $21-45$ es decir $A = \{Masculino, \leq 120000\} \Rightarrow B = \{21-45\}$

El "soporte" de la regla $A \Rightarrow B$, $T(A \Rightarrow B)$ es la fracción de individuos que poseen los atributos de A y B (antecedente y consecuente), que no es más que el soporte del conjunto de elementos $\mathcal{K}$ de la que se derivaron, esto es $T(A \Rightarrow B) = T(\mathcal{K}) = T(A \cup B)$,

Para un conjunto $W \subset \{1, 2, 3..., K\}$ definimos:

$$E_W = \{x \in \ell : x\, tiene\, los\, atributos\, de\, W\}$$

Entonces $\mathbf{Pr}(E_W)$ es entendida como la probabilidad de que un sujeto de nuestra base de datos elegido aleatoriamente tenga los atributos contenidos en el conjunto W, así con está nueva fomulación tenemos que

$$\hat{\mathbf{Pr}}(E_W) = T(W) \tag{3.5}$$

Consecuentemente

$$T(A \Rightarrow B) = T(A \cup B) = \hat{\mathbf{Pr}}(E_A \cap E_B) \tag{3.6}$$

Definimos la confianza de la relación $A \Rightarrow B$ como el soporte de la relación $A \Rightarrow B$ entre el soporte de A esto es:

$$C(A \Rightarrow B) = \frac{T(A \Rightarrow B)}{T(A)} \tag{3.7}$$

Que por (3.6) puede ser escrito como:

$$C(A \Rightarrow B) = \frac{\hat{\mathbf{Pr}}(E_A \cap E_B)}{T(A)} \tag{3.8}$$

La confianza de la relación $A \Rightarrow B$ puede ser entendida como el porcentaje de individuos que poseen los atributos del antecedente (A) tambíen poseen los atributos del consecuente (B).

El "lift" o levantamiento de la relación $A \Rightarrow B$ es definido como la confianza de la relación $A \Rightarrow B$ entre el soporte de B, esto es:

$$L(A \Rightarrow B) = \frac{C(A \Rightarrow B)}{T(B)} \tag{3.9}$$

Que por (1.8) puede ser reescrita de la siguiente manera:

$$L(A \Rightarrow B) = \frac{\hat{\mathbf{Pr}}(E_A \cap E_B)}{T(A)T(B)} \tag{3.10}$$

El "lift" o levantamiento de la relación $A \Rightarrow B$ representa la independencia entre el antecedente y el consecuente; un valor igual a 1 indica que el antecedente y el consecuente son independientes. Un valor mayor a uno significa que están positiviamente correlacionados. Un valor menor que uno indica que están negativamente correlacionados.

Supongamos que $\mathcal{K} = \{Masculino, \leq 120000, 21-45\}$ con $A = \{Masculino, \leq 120000\}$ y $B = \{21 - 45\}$ entonces un soporte de .05 en la relación $A \Rightarrow B$ significa que el 5 % de nuestra base de datos posee estos atributos ($Maculino, \leq 120000, 21-45$) conjuntamente. Una confianza de .7 para está relación significa que el 70 % de nuestra base de datos que poseen los atributos $Masculino$ y ≤ 120000 también tienen el atributo $21 - 45$. Ahora si el 60 % de nuestra base de datos es de sexo $Masculino$ estó nos dirá que la relación tiene un levantamiento de 1.666.

Todo esté análisis es muy útil en empresas y supermercados Rakesh (2010), ya que ellos pueden estar interesados en lo siguiente:

- Encontrar todas las reglas que tienen "Coca-Cola Light" como consecuente. Estas reglas pueden ayudar a planificar lo que la tienda debe hacer para impulsar la venta de Coca-Cola Light.
- Encontrar todas las reglas que tienen "salchichas" en el antecedente y "mostaza.[en] el consecuente. Esta consulta se puede expresar alternativamente como una solicitud de los elementos adicionales que tienen que ser vendidos junto con la salchicha con el fin de hacer que sea muy probable que la mostaza también se venda.
- Encontrar las "mejores" k reglas que tienen donas en el consecuente. Aquí, el "mejor" puede formularse en términos de los factores confianza de las reglas, o en términos de su soporte, es decir, la fracción de transacciones que satisfacen la regla, Troncoso (2002).

3.2. FCA

Aqui proporcionamos una pequeña reseña sobre el análisis de conceptos formales (FCA), introduciremos los fundamentos matemáticos más importantes derivados su teoría.

El FCA es una herramienta matemática relativamente reciente para la formalización del conocimiento conceptual. Es una teoría de formación de conceptos derivada de teoría de retículos y conjuntos ordenados que proporciona un modelo matemático para el análisis de jerarquías conceptuales, es también un método de análisis de datos con creciente popularidad a través de varios dominios, analiza los datos que describen la relación entre un conjunto determinado de objetos y un determinado conjunto de atributos. Estos datos aparecen comúnmente en muchas áreas de la actividad humana. El FCA produce dos tipos de salida de los datos de entrada. El primero es un concepto de retículo. Un concepto de retículo es una colección de conceptos formales en los datos que son jerárquicamente ordenados por una relación subconcepto-superconcepto. Los conceptos formales son grupos partículares que representan conceptos humanos y naturales, como el "coche con todos los sistemas de transmisión de la rueda", "organismo vivo en el agua", "número divisible por 3 y 4", etc. La segunda salida del FCA es un colección de los llamados atributos de implicaciones. Un atributo de implicación describe una dependencia particular que es válida en los datos como "todo número divisible por 3 y 4 es divisible por 6", "se retiró cada encuestado con edad mayor de 60", etc Ganter (2002).

Contextos Formales y Conceptos Formales

Debido a que un concepto puede tener un gran número de instancias ya que estas pueden ser un conjunto prácticamente ilimitado de propiedades o atributos compartidos, habitualmente se trabaja sobre un contexto específico dentro del cual están limitados tanto el conjunto de objetos como el de atributos. El modelo matemático que representa la relación entre los objetos y los atributos se denomina contexto formal y se define como:

Definición 3.1. *Un contexto (formal) es una tripleta* $\aleph := (G, M, I)$ *donde* G *es un conjunto cuyos elementos son llamados objetos,* M *es un conjunto cuyos elementos son llamados atributos e* I *es una relación binaria entre* G *y* M *i.e.* $(I \subset G \times M), (g, m) \in I$ *se lee como "el objeto* g *tiene el atributo* m*".*

Los elementos g de G representan los objetos o entidades del contexto, mientras que los elementos m de M representan los atributos o características que los objetos pueden tener asociados. La relación $g\,I\,m$ afirma que "el objeto g tiene

el atributo m" o, de forma equivalente, que "el atributo m se aplica sobre el objeto g". Ahora podemos definir lo que es un concepto formal, ver Uta P.

Note que esta definición de contexto formal es muy general y no hace restricciones sobre la naturaleza de los objetos y los atributos, nosotros podemos considerar objetos fisicos ó personas, numeros, procesos, estructuras, etc,...,etc.

Definición 3.2. *Para $A \subseteq G$, sea A'=$\{m \in M : \forall g \in A \mid (g,m) \in I\}$ y para $B \subseteq M$, sea B'=$\{g \in G : \forall m \in B \mid (g,m) \in I\}$ un concepto (formal) de un contexto (formal) (G, M, I) es un par (A, B) con $A \subseteq G$, $B \subseteq M$, $A' = B$ y $B' = A$, los conjuntos A y B son llamados, la **extensión** y la **intensión** de un concepto formal (A, B), respectivamente.*

La definición anterior lo que nos dice es que A' es el conjunto de todos lo atributos del conjunto M que se aplican sobre todos y cada uno de los objetos de A. De igual modo, dado un subconjunto B del conjunto de atributos M, B' denota el conjunto de objetos pertenecientes a G sobre los que se aplican todos los atributos de B. De acuerdo a la teoría del FCA, un concepto (formal) viene definido por dos partes bien diferenciadas:

- Su extensión, que hace referencia al conjunto de objetos (entidades o instancias) pertenecientes al concepto.
- Su intensión o comprensión, que hace referencia a todos los atributos (propiedades o caracteríticas) que comparten todos los objetos considerados

Si un objeto y un atributo pertenecen a un mismo concepto, entonces se puede afirmar que el objeto tiene ese atributo concreto, es decir, dentro de un concepto la extensión se relaciona con la intensión mediante una relación de incidencia entre los objetos y los atributos, que son recíprocamente dependientes.

Tenemos los siguientes hechos simples ver Radim (2008)

Proposición 3.1. *Para subconjuntos $A, A_2, A_3 \subseteq G$ y $B, B_1, B_2 \subseteq M$, tenemos que*

1. Si $A_1 \subseteq A_2 \Longrightarrow A_2' \subseteq A_1'$,

2. $A \subseteq A''$,

3. $A''' = A'$,

4. Si $B_1 \subseteq B_2 \Longrightarrow B_2' \subseteq B_1'$,

5. $B \subseteq B''$.

6. $B''' = B'$,

Definición 3.3. *(Cerrado) Sea $A \subseteq X$, A(un subconjunto de objectos) es cerrado si cumple que $A'' = A$. Análogamente, para B(el subconjunto de atributos), es decir, sea $B \subseteq Y$, B es cerrado si cumple que $B'' = B$.*

La figura (3.1) muestra el contexto formal correspondiente al conjunto de planetas del Sistema Solar descritos por un conjunto de atributos. El conjunto de atributos seleccionado para describir los planetas (objetos del contexto formal) son los siguientes: su tamaño (que podrá ser pequeño, mediano o grande), su distancia al Sol (que podrá ser cercana o lejana), y, finalmente, si posee luna propia o no. De este modo, el contexto formal $\aleph := (G, M, I)$ vendría descrito por el conjunto de planetas, el conjunto de características utilizadas para realizar la descripción del dominio que acabamos de describir y el conjunto de relaciones entre objetos y atributos que estarían representadas en el cuadro mediante una x, ver Recuero (2008)

	TP	TM	TG	DC	DL	LS	LN
Mercurio	×			×			×
Venus	×			×			×
Tierra	×			×		×	
Marte	×			×		×	
Júpiter			×		×	×	
Saturno			×		×	×	
Urano		×			×	×	
Neptuno		×			×	×	
Plutón	×				×	×	

Figura 3.1: Contexto formal correspondiente a los planetas del Sistema Solar descritos por un conjunto de atributos. Las siglas se corresponden con TP=Tamaño pequeño, TM=Tamaño mediano, TG=Tamaño grande, DC=Distancia al sol cercana, DL=Distancia al sol lejana, LS=Posee luna, LN=No posee luna

En general, la definición de concepto formal impone restricciones importantes, haciendo que el número real de conceptos correspondientes a un contexto concreto sea bastante pequeño en comparación con todos los pares extensión-intensión posibles. Además, podemos afirmar que este número crece de manera lineal con el número de objetos dentro del contexto.

Volviendo al ejemplo presentado en la figura (3.1) el par:

$$((Pluton, Jupiter, Saturno, Urano, Neptuno), (DL, LS)) \tag{3.11}$$

sería un concepto formal dado que cumple:

$$(Pluton, Jupiter, Saturno, Urano, Neptuno)' = (DL, LS) \tag{3.12}$$

y

$$(DL, LS)' = (Pluton, Jupiter, Saturno, Urano, Neptuno) \tag{3.13}$$

Sin embargo, el par:

$$((Tierra, Saturno), (DL, LS)) \tag{3.14}$$

no sería un concepto formal dado que:

$$(Tierra, Saturno)' = \emptyset \tag{3.15}$$

que no coincide con la intensión (DL, LS).

Veamos que de hecho, en está estructura tan rica como lo son los conceptos formales, se puede definir un orden.

Jerarquia conceptual

Los conceptos formales pueden ser (parcialmente) ordenados de una forma natural, una vez más la definición es inspirida por la forma en que normalmente ordenamos conceptos en jerarquia, subconcepto-superconcepto. por ejemplo cerdo es un subconcepto de un mamifero, por que cada cerdo es un mamifero, transfiriendo esto a a los conceptos formales, la definición natural es la siguiente:

Definición 3.4. *(Orden) Sean $c_1 = (A_1, B_1)$ y $c_2 = (A_2, B_2)$ dos conceptos pertenecientes al contexto $\aleph := (G, M, I)$ diremos que c_1 es un subconcepto de c_2 (y equivalentemente que c_2 es un superconcepto de c_1) sii $A_1 \subseteq A_2$. usaremos el signo $\leq$ para expresar está relación y asi tenemos*

$$c_1 \leq c_2 :\Leftrightarrow A_1 \subseteq A_2$$

Definición 3.5. *La colección de todos los conceptos formales de un contexto formal (G, M, I), se llama un concepto de reticulo, otra noción fundamental en FCA, esto es $\underline{\mathcal{B}}(G, M, I) = \{(A, B) | A' = B, B = A'\}$*

La relación binaria definida anteriormente en el conjunto $\underline{\mathcal{B}}(G, M, I)$ cumple las propiedades reflexiva, transitiva y antisimétrica, es decir el concepto de retículo forma un conjunto parcialmente ordenado con el orden dado anteriormente. corresponde con la idea de que un concepto siempre tiene una extensión más pequeña y una intensión más grande que cualquiera de sus superconceptos.

Definición 3.6. *Un conjunto parcialmente ordenado es un par $(P, \leq)$ donde P es un conjunto y $\leq$ es una relación binaria en P (i.e. $\leq$ es un subconjunto de $P \times P$) el cual es:*
1. reflexivo : $x \leq x$ para toda $x \in P$
2. antisimétrico : $x \leq y \wedge x \neq y \Longrightarrow \neg y \leq x$ para todo x, $y \in P$
3. transitivo : $x \leq y \wedge y \leq z \Longrightarrow x \leq z$ para todo x, y, $z \in P$

Ejemplos de conjuntos parcialmente ordenados son:

- El conjunto de todos los números reales $\mathbb{R}$ junto con el orden usual $\leq$.
- Todo arbol es un conjunto parcialmente ordenado.
- Para un conjunto M, el conjunto potencia $P(M)$ con la inclusión de conjuntos $\subseteq$ es también un orden parcial.

La relación de subtipo-supertipo definida sobre el conjunto $\underline{\mathcal{B}}$(G,M,I) se denomina generalización-especialización. De acuerdo al ejemplo presentado anteriormente, una posible relación de generalización especialización entre alguno de los conceptos obtenidos sería $c_1 \geq c_2$ donde:

$$c_1 = (Pluton, Jupiter, Saturno, Urano, Neptuno), (DL, LS))$$

$$c_2 = ((Pluton), (TP, DL, LS))$$

Los elementos de un conjunto ordenado se pueden comparar. Dados dos elementos p y q pertenecientes a un conjunto ordenado P, se afirma que p y q son comparables si $p \leq q$ ó $q \leq p$, en cualquier otro caso se dice que p y q son no comparables.

Dentro de un conjunto ordenado $(P, \leq)$ se puede definir una relación de vecindad $\prec$ entre sus elementos. Dados dos elementos p y q pertenecientes a un conjunto ordenado $(P, \leq)$, se dice que p es el vecino inferior de q o, de igual modo, que q es el vecino superior de p si y sólo si se cumple: a) $p \leq q$ y $p \neq q$ y; b) para todo elemento r perteneciente a P si $p \leq r \leq q$ entonces $r = p$ o $r = q$. Si $(P, \leq)$ es un conjunto ordenado finito, la relación de orden $\leq$ se encuentra directamente determinada por la relación de vecindad que se acaba de definir.

Volviendo al ejemplo anterior, $c_1 \succ c_2$, dado que no existe ningún otro concepto formal intermedio r tal que $c_1 \geq r \geq c_2$.

El conjunto ordenado $(P, \leq)$ que acabamos de definir puede representarse mediante un diagrama de lineas o diagrama de Hasse ver Pichardo. En un diagrama de este tipo, los elementos de P se representan mediante pequeños círculos

de tal forma que, cuando un elemento p es vecino inferior de otro elemento q (cuando $p \prec q$), el elemento q se representa por encima del elemento p y ambos quedan unidos por una linea que no pasa por ningún otro elemento del conjunto. A partir de un diagrama de este tipo es posible deducir cualquier relación $v \leq w$ existente entre los elementos del conjunto siguiendo la sucesión de lineas ascendentes que van desde el nodo que representa al elemento v hasta el nodo que representa al elemento w.

En el caso concreto del conjunto $\underline{\mathcal{B}}$(G,M,I), cada nodo representa un concepto y se etiqueta con el conjunto de objetos y de atributos que lo definen. Los objetos se situán en la parte inferior derecha del nodo, mientras que los atributos se disponen en su parte superior derecha. Por convenio, los nombres de los objetos se escribirán sólo en los nodos correspondientes a los conceptos objeto generados a partir de cada uno de los objetos y los nombres de los atributos se escribirán en la parte superior derecha de los nodos que representen conceptos atributo generados a partir de cada uno de los atributos. Todos los nodos que se encuentren por encima de un nodo que contenga al objeto $g \in G$ también contendrañ al objeto g en su extensión. De igual forma, todos los nodos que se encuentren por debajo de un nodo que contenga al atributo $m \in M$ también contendrán al atributo m en su intensión. La forma de leer la extensión de un concepto dentro del diagrama es tomando como extensión del concepto todos los objetos que definan al propio nodo y a cualquier otro nodo que se encuentre en el camino descendente desde el nodo hacia la parte inferior del retículo. De igual forma, la intensión de un concepto la podremos obtener tomando los atributos del propio nodo y los de todos los nodos que se encuentren en el camino ascendente desde el nodo hacia la parte superior del retículo. La siguiente figura presenta el diagrama de Hasse correspondiente al contexto formal presentado anteriormente .

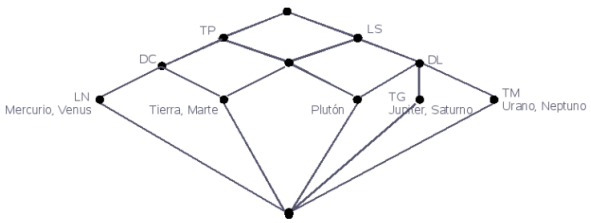


Figura 3.2: Diagrama de Hasse correspondiente al contexto formal presentado en la tabla (3.1)

3.3. Aplicación

Uno de nuestros objetivos de la tesis "Construir una base de datos con contenido demográfico y de operaciones financieras y de actividades no financieras para identificar o simular el comportamiento de interés. Identificar la información de acceso puúblico de datos de las personas políticamente expuestas (PEPs) que ejercen cargos públicos que pudieren caer en actos u omisiones que impliquen un comportamiento de corrupción tales como la malversacioón de fondos" por lo cuál en este trabajo se construyó una base de datos que incluye sujetos que realmente están inmiscuidos en el delito de corrupcioón, esta base incluye su edad, el estatus de su caso, etc... ver ápendice A.

Por otro lado, el tema primordial de este apartado es la aplicación de métodos no supervisados a una base datos que simularemos. los atributos o variables que consideraremos en esta base se puede obtener de la información que recibe hacienda de sus sujetos obligados sobre las operaciones financieras o de actividades o profesiones, en las secciones siguientes haremos una revisión de los elementos que usaremos para generar dicha base y finalmente se dará una conclusión acerca de los métodos utilizados para la indentificación de funcionarios que pudieran estar realizando actos de corrupción en el ejercicio de sus funciones en nuestro país. La razón de utilizar una base de datos obtenida por simulación es porque es información de carácter confidencial no disponible al público en general y es de resguardo de las instituciones, pero para el propósito de este trabajo nos basta conocer qué tipo de información poseen las fuentes de datos.

3.3.1. Variables incluidas en la base de datos

Una de las motivaciones para la comisión del delito de corrupción es la malversación de fondos, las cuotas por favores, los beneficios económicos, esto es, se generan capitales que no pueden justificarse con sus ingresos lo que conduce al ocultamiento de recursos. De ahí que un modo de ocultar los recursos ilícitos producto de la corrupción sea transfórmalos en recursos lícitos inyectando estos recursos al sistema financiero no sólo mexicano sino internacional, o transformarlos es bienes o servicios.

El gobierno mexicano a través de la Secretaria de Hacienda y Crédito Público (SHCP) ha realizado esfuerzos por legislar leyes y obligaciones para sujetos que pudieran verse insmiscuidos en el ocultamiento o transformación de recursos de procedencia ilícita. Derivado de estas leyes la SHCP recibe reportes de operaciones financieras, avisos de operaciones con actividades relacionadas al

sistema financiero. Cabe mencionar que muchas de estas medidas son aplicadas por recomendaciones de organismos internacionales. Los reportes que recibe de los sujetos obligados son de vital importancia para detectar actos delictivos previstos en los artículos 139 o 148 Bis del Código Penal Federal que tiene que ver con lavado de dinero [30].

Las entidades financieras que tienen que presentar estos reportes son, por mencionarpor mencionar algunos

- Instituciones de Banca Múltiple
- Instituciones de Banca de Desarrollo
- Sociedades Financieras de Objeto Limitado
- Casas de Bolsa
- Sociedades Operadoras de Sociedades de Inversión
- Sociedades Distribuidoras de Acciones de Sociedades de Inversión
- Arrendadoras Financieras
- Empresas de Factoraje Financiero
- Almacenes Generales De Depósito
- Uniones de Crédito
- Sociedades de Ahorro y Préstamo
- Casas de Cambio
- Sociedad Financieras de Objeto Múltiple
- Entidades de Ahorro y Crédito Popular
- Centros Cambiarios
- Transmisores de Dinero
- Instituciones de Seguros
- Instituciones de Fianzas

ver [31], estas instituciones realizan y envÃan a la SHCP tres tipos de reportes principales, que corresponden al tipo de operación realizada, las operaciones son las siguientes:

- Operación inusual: es la operación, actividad o comportamiento de una persona que no concuerde con su patrón habitual de realización de transaciones.

- Operación preocupante: es la operación, actividad o comportamiento de los directivos, funcionarios, empleados y apoderados de la persona obligada, sin justificar razonablemente gastos superiores al de sus posibilidades.

- Operación relevante: es la operación que se efectúa por un monto mayor o igual 10,000 dólares, ya sea con billetes o con monedas metalicas de curso legal en los Estados Unidos Mexicanos o en cualquier otro país [32]; en [33] se presentan los campos que deben ser llenadas por dichas entidades; en [34] se presenta el marco jurídico en materia de prevención y combate de los delitos de operaciones con recursos de procedencia ilícita y de terrorismo y su financiamiento, ahí mismo están los formatos oficiales para los reportes antes mencionados.

En [35] se establece diversas actividades no financieras consideradas vulnerables, como, juegos y sorteos, compra venta de inmuebles, vehículos (aéreos, maritimos, terrestre), joyas, obras de arte, tarjeta de prepago y cupones. Los sujetos que realicen estas actividades están obligados a emitir avisos a la Secretaría de Hacienda y Crédito Público de operaciones para prevenir o detectar posibles actos u operaciones de lavado de dinero.

El siguiente diagrama de flujo muestra explicitamente como es que activamente funcionan las Leyes Federales para la Prevención e Identificación de Operaciones con Recursos de Procedencia Ilícita, estás leyes son fundamentales para todas aquellas instituciones y entidades cuyo principal objetivo es detectar esté tipo de operaciones por medio de la aplicación de la misma.

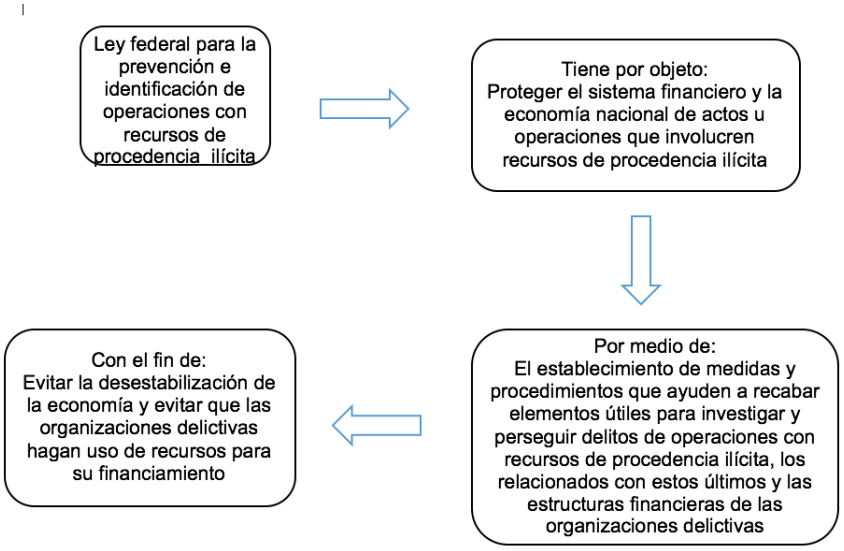


Figura 3.3: Explicación de la Ley Federal para la prevención de operaciones con recursos de procedencia ilícita

Podemos construir una diversidad de variables que pueden obtenerse de la información financiera o de actividades vulnerables que puedan ayudar a identificar algun elemento del ocultamiento o utilización de recursos procedentes de la corrupción.

3.3.2. Generación de base de datos

En esta sección mostraremos detalladamente la construcción de una base de datos generada por simulación, así como la aplicación de los métodos revisados en el capítulo anterior tendientes a detectar funcionarios corruptos.

Simulamos una base de datos con varibles (atributos) que nos parece pudieran ser más relevantes en nuestro problema de identificación de sujetos corruptos, haciendo la aclaración que nos se pretenden duplicar los datos reales. Las variables que tomamos en cuenta son las siguientes; X_1 =monto total por depositos de operaciones relevantes "Monto_dep_rel", X_2 =monto promedio por operaciones relevantes "Monto_prom_rel", X_3 =número de reportes por compra de dólares de operaciones relevantes "Monto_comp_dol", X_4 =monto por la compra de inmuebles de alto valor "Monto_inmueble", X_5 =número de inmuebles adquiridos de alto valor "Num_inmuebles", y X_6 =sexo "Sexo". Se generarón 10000 registros correspondientes a la información de 10000 sujetos, donde cada registro X consta de las variables anteriores, es decir, cada registro es de la siguiente forma $X = (X_1, X_2, X_3, X_4, X_5, X_6)$. Elegímos a la variable Monto

total por depósito de operaciones relevantes "Monto_dep_rel" como variable principal porque consideramos que es la variable de mayor peso y que incide en la mayoria de los valores de las variables restantes que se tomarón en cuenta

Ya que en la realidad los valores más pequeños de la variable de interés "Monto_dep_rel" son más probables que los valores grandes, se considero utilizar diferentes probabilidades para la generación de esta variable Definimos x_{1i} como la $i-$ésima entrada de la variable X_1 "Monto_dep_rel" es decir la primera entrada del $i-$ésimo registro $1 \leq i \leq 10000$, la cual fue generada de la siguiente forma: se genera un número aleatorio "n_1" (entero) dentro del intervalo $[0, 20)$ y:

- si $n_1 \in \{0, .., 13\}$ entonces $x_{1i} \in [4000, 12000) = s_{11}$
- si $n_1 \in \{14, ..., 16\}$ entonces $x_{1i} \in [12000, 102000) = s_{12}$
- si $n_1 \in \{17, 18, 19\}$ entonces $x_{1i} \in [102000, 302000) = s_{13}$
  para cada i.

Donde $x_{1i} \in [a, b)$ significa que x_{1i} toma un valor aleatorio entre $[a, b)$. Así, el vector x_{1i}, $0 \leq i < 10000$ es como dice su definición "El monto total de los depósitos hechos por el $i-$ésimo sujeto" , una vez hecho esto usaremos el método de variables ficticias para crear tantas variables "ficticias" como categorías haya; observemos que $X_1 \in S_1$ donde $S_1 = \{s_{11}, s_{12}, s_{13}\}$ y s_{1j} corresponde a la $j-$ésima categoría de la primera variable, en este caso hay 3 categorías por lo cual se crearon 3 variables ficticias; $z_{11i}, z_{12i}, z_{13i}$, que se llenarón de la siguiente manera:

- si $x_{1i} \in s_{11}$ entonces $z_{11i} = 1$, $z_{12i} = 0$ y $z_{13i} = 0$
- si $x_{1i} \in s_{12}$ entonces $z_{11i} = 0$, $z_{12i} = 1$ y $z_{13i} = 0$
- si $x_{1i} \in s_{13}$ entonces $z_{11i} = 0$, $z_{12i} = 0$ y $z_{13i} = 1$

El siguiente cuadro muestra un ejemplo de los valores que toma la primera variable:

i	n_1	categoría x_1	x_1	z_{11}	z_{12}	z_{13}
1	17	2	79900	0	1	0
2	15	2	89890	0	1	0
3	3	1	8123	1	0	0
4	10	1	4420	1	0	0
5	15	2	22085	0	1	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Es de hacerse notar el hecho de que las categorías tienen diferentes pesos en los valores de la variable x_{1i}, es decir, más del 50 % de los individuos realizan un monto menor a 12000, más de el 20 % realiza un monto entre 12000 y 102000, y menos del 15 % un monto de entre 102000 y 302000.

Ahora explicaremos como es que se generó la segunda variable X_2 "Monto_prom_rel" y su dependencia con la variable X_1 "Monto_dep_rel". Definimos x_{2i} como la $i-$ésima entrada de la variable X_2 "Monto_prom_rel" es decir la segunda entrada del $i-$ésimo registro , esta variable refleja el número promedio de montos realizados por el sujeto i, $0 \leq i < 10000$, para dar valores a esta variable a cada sujeto se le generó un número aleatorio $n_2 \in \mathbb{Z}$ entre $0 - 19$ y conforme a la categoría que este obtuvo en la primera variable, le fue asignado el valor de la segunda variable, como lo muestra la siguiente tabla.

Categoría de la primera variable	Valores de i			
$s_{11} = [4000, 12000)$	$0 \leq i \leq 11$	$12 \leq i \leq 15$	$16 \leq i \leq 17$	$i = 19$
$s_{12} = [12000, 102000)$	$0 \leq i \leq 8$	$9 \leq i \leq 13$	$14 \leq i \leq 17$	$18 \leq i \leq 19$
$s_{13} = [102000, 302000)$	$0 \leq i \leq 1$	$2 \leq i \leq 5$	$6 \leq i \leq 11$	$12 \leq i \leq 19$
Categoría asignada a la segunda variable	$s_{21} = \{0, 1, 2\}$	$s_{22} = \{3, 4\}$	$s_{23} = \{5, 6\}$	$s_{23} = \{7, ..., 13\}$

Cuadro 3.3: Categoría que se asigna a la segunda variable dependiendo el valor de la categoría s_{1j} de la primera variable

Por ejemplo, si el i-esimo sujeto tiene la categoría $s_{11} = [4000, 12000)$ de la primera variable y el número aleatorio n_2 está dentro del intervalo 0-11, entonces se le asigno un número aleatorio en $s_{21} = \{0, 1, 2\}$ que corresponde al número promedio de montos por depósitos que el individuo realizó; si el número aleatorio n_2 se encuentra dentro del intervalo 12-15, entonces se le asigno un número aleatorio en $s_{22} = \{3, 4\}$ que corresponde al número promedio de montos por depósitos que el individuos realizó; si el número aleatorio n_2 se encuentra dentro del intervalo 16-17, entonces se le asigno un número aleatorio en $s_{23} = \{5, 6\}$ que corresponde al número promedio de montos por depósitos que el individuos realizó y finalmente si el número aleatorio n_2 es igual a 19, entonces se le asigno un número aleatorio en $s_{24} = \{7, ..., 13\}$ que corresponde al número promedio de montos por depósitos que el sujeto realizó.

Nota: Como vimos la segunda variable "monto promedio por operaciones" depende de las categorías obtenidas de la primera variable "monto total por depósito" y esto es debido a que, si el individuo i realiza un monto total alto por depósitos es más probable que realice una mayor cantidad de numero de montos promedios por depositos y viceversa, esta información es la que esta contenida en la tabla 3.3.

De la misma forma se crean tantas variables ficticias como categorías haya; sea $S_2 = \{s_{21}, s_{22}, s_{23}.s_{24}\}$ donde s_{2j} corresponde a la $j-$ésima categoría de la segunda variable, por lo cual hay 4 categorías posibles, entonces se crearon un total de 4 variables ficticias: $z_{21i}, z_{22i}, z_{23i}, z_{24i}$ correspondientes a la segunda variable, que serán llenadas de la siguiente manera:

- si $x_{2i} \in \{0, 1, 2\}$ entonces $z_{21i} = 1$, $z_{22i} = 0$, $z_{23i} = 0$ y $z_{24i} = 0$
- si $x_{2i} \in \{3, 4\}$ entonces $z_{21i} = 0$, $z_{22i} = 1$, $z_{23i} = 0$ y $z_{24i} = 0$
- si $x_{2i} \in \{5, 6\}$ entonces $z_{21i} = 1$, $z_{22i} = 0$, $z_{23i} = 1$ y $z_{24i} = 0$
- si $x_{2i} \in \{7, .., 13\}$ entonces $z_{21i} = 1$, $z_{22i} = 0$, $z_{23i} = 0$ y $z_{24i} = 1$

El siguiente cuadro muestra un ejemplo de los valores que toma la segunda variable:

i	n_1	n_2	categoría x_1	categoría x_2	x_1	x_2	z_{11}	z_{12}	z_{13}	z_{21}	z_{22}	z_{23}	z_{24}
1	17	4	2	1	79900	1	0	1	0	1	0	0	0
2	15	0	2	1	89890	2	0	1	0	1	0	0	0
3	3	17	1	3	8123	5	1	0	0	0	0	1	0
4	10	16	1	3	4420	6	1	0	0	0	0	1	0
5	15	3	2	1	22085	0	0	1	0	1	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Ahora explicaremos la generación de la tercera variable X_3 "Monto_comp_divisa" que es el monto acumulado por compra de dólares y su dependencia con la variable X_2 "Monto_prom_rel". Definimos x_{3i} como la $i-$ésima entrada de la variable X_3, es decir la tercera entrada del $i-$ésimo registro con $0 \leq i < 10000$, para dar valores a esta variable a cada sujeto se le generó un número aleatorio $n_3 \in \mathbb{Z}$ entre $0 - 19$ y conforme a la categoría que tuvo el individuo en la segunda variable, fue asignado el valor de la tercera variable, como lo muestra la siguiente tabla.

Categoría de la segunda variable	Valores de i		
$s_{21} = \{0, 1, 2\}$	$0 \leq i \leq 9$	$10 \leq i \leq 15$	$16 \leq i \leq 19$
$s_{22} = \{3, 4\}$	$0 \leq i \leq 7$	$8 \leq i \leq 16$	$17 \leq i \leq 19$
$s_{23} = \{5, 6\}$	$0 \leq i \leq 7$	$8 \leq i \leq 14$	$15 \leq i \leq 19$
$s_{24} = \{7, ..., 13\}$	$0 \leq i \leq 5$	$6 \leq i \leq 9$	$10 \leq i \leq 19$
Categoría asignada a la tercera variable	$s_{31} = [0, 300)$	$s_{32} = [300, 2800)$	$s_{33} = [2800, 3800)$

Cuadro 3.4: Categoría que se asigna a la tercera variable dependiendo el valor de la categoría s_{2j} de la segunda variable

Por ejemplo, si el i-esimo sujeto tiene la categoría $s_{23} = \{5, 6\}$ de la segunda variable y el número aleatorio n_2 está dentro del intervalo 0-7, entonces a la tercera variable se le asignó un número aleatorio en $s_{31} = [0, 300)$ que corresponde al monto acumulado por compra de dólares que el individuo realizó; si el número aleatorio n_2 se encuentra dentro del intervalo 8-14, entonces a la tercera variable se le asignó un número aleatorio en $s_{32} = [300, 2800)$ que corresponde al monto acumulado por compra de dólares que el individuo hizo; por último si el número aleatorio n_2 se encuentra dentro del intervalo 15-19, entonces a la tercera variable se le asignó un número aleatorio en $s_{33} = [2800, 3800)$ que corresponde al monto acumulado por compra de dólares que el individuo realizó.

Nota: Como vimos la tercera variable "monto por compras de dólares" depende de las categorías obtenidas de la segunda variable "número de montos promedio por deposito " y esto es debido a que, si el individuo i realiza un monto promedio alto por operaciones es más probable que realice una mayor cantidad de monto por compra de dólares y viceversa, esta información es la que esta contenida en la tabla 3.4.

Procedemos a la creación de variables ficticias; sea $S_3 = \{s_{31}, s_{32}, s_{33}\}$ donde s_{3j} corresponde a la $j-$ésima categoría de la tercera variable, en este caso hay 3 categorías posibles que puede tomar el valor de la variable X_3 por tanto, se crearón 3 variables: $z_{31i}, z_{32i}, z_{33i}$ y fueron llenadas de la siguiente forma:

- si $x_{3i} \in [0, 300)$ entonces $z_{31i} = 1$, $z_{32i} = 0$ y $z_{33i} = 0$
- si $x_{3i} \in [300, 2800)$ entonces $z_{31i} = 0$, $z_{32i} = 1$ y $z_{33i} = 0$
- si $x_{3i} \in [2800, 3800)$ entonces $z_{31i} = 0$, $z_{32i} = 0$ y $z_{33i} = 1$

El siguiente cuadro muestra un ejemplo de los valores que toma la tercera variable:

i	n_2	n_3	categoría x_2	categoría x_3	x_2	x_3	z_{21}	z_{22}	z_{23}	z_{24}	z_{31}	z_{32}	z_{33}
1	4	17	1	3	1	3657	1	0	0	0	0	0	1
2	0	3	1	1	2	44	1	0	0	0	1	0	0
3	17	13	3	2	5	1431	0	0	1	0	0	1	0
4	16	14	3	2	6	520	0	0	1	0	0	1	0
5	3	14	1	2	0	2299	1	0	0	0	0	1	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Explicación de la generación de la cuarta variable X_4 "Monto_inmueble" y dependencia con la variable X_1 "Monto_dep_rel". Sea x_{4i} la i −ésima entrada de la variable X_4, es decir la cuarta entrada del $i-$ésimo registro, que no es más que es el monto de inmuebles de alto valor adquirido por el $i-$ésimo sujeto $0 \leq i < 10000$, para dar valores a esta variable a cada sujeto se le generó un número aleatorio $n_4 \in \mathbb{Z}$ entre $0 - 19$ y conforme a la categoría que tiene el individuo en la primera variable, fue asignado el valor de la cuerta variable, como lo muestra la siguiente tabla.

Categoría de la primera variable	Valores de i	
$s_{11} = [4000, 12000)$	$0 \leq i \leq 16$	$17 \leq i \leq 19$
$s_{12} = [12000, 102000)$	$0 \leq i \leq 14$	$15 \leq i \leq 19$
$s_{13} = [102000, 302000)$	$0 \leq i \leq 9$	$10 \leq i \leq 19$
Categoría asignada a la cuarta variable	$s_{41} = 0$	$s_{42} = [3000, 90000)$

Cuadro 3.5: Categoría que se asigna a la cuarta variable dependiendo el valor de la categoría s_{1j} de la primera variable

Por ejemplo, si el i-esimo sujeto tiene la categoría $s_{12} = [12000, 102000)$ de la primera variable y el número aleatorio n_4 está dentro del intervalo 0-14, entonces se le asignó el valor $s_{41} = 0$ a la cuarta variable, que corresponde al monto por la compra de inmueblesde alto valor que el individuo realizó; si el número aleatorio n_2 se encuentra dentro del intervalo 15-19, entonces se le asigno a la cuarta variable un número aleatorio en $s_{42} = [3000, 90000)$ que corresponde al monto por la compra de inmuebles de alto valor que el individuo realizó.

Nota: Como vimos la cuarta variable "monto por la compra de inmueblesde alto valor" depende de las categorías obtenidas de la primera variable " monto total por depósitos" y esto es debido a que, si el individuo i realiza un monto total alto de depósitos es más probable que realice compras de inmuebles de alto valor y viceversa, esta información es la que esta contenida en la tabla 3.5.

De la misma forma creamos un total de 2 variables ficticias z_{41i}, z_{42i} ya que $S_4 = \{s_{41}, s_{42}, \}$, cuyo llenado es el siguiente:

- si $x_{4i} = \{0\}$ entonces $z_{41i} = 1$ y $z_{42i} = 0$

- si $x_{4i} \in [30000, 90000)$ entonces $z_{41i} = 0$ y $z_{42i} = 1$

El siguiente cuadro muestra un ejemplo de los valores que toma la cuarta variable

i	n_1	n_4	categoría x_1	categoría x_4	x_1	x_4	z_{11}	z_{12}	z_{13}	z_{41}	z_{42}
1	17	9	2	1	79900	0	0	1	0	1	0
2	15	9	2	1	89890	0	0	1	0	1	0
3	3	2	1	1	8123	0	1	0	0	1	0
4	10	14	1	1	4420	0	1	0	0	1	0
5	15	9	2	1	22085	0	0	1	0	1	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

A continuación mostramos como se generó la variable X_5 "Num_inmuebles" y la dependencia con la variable X_4 "Monto_inmueble". Sea x_{5i} la $i-$ésima entrada de la variable X_5, es decir la quinta entrada del $i-$ésimo registro, que es el numero de inmuebles de alto valor adquiridos por el $i$ésimo sujeto $0 \leq i < 10000$, para dar valores a esta variable para a sujeto se le generó un número aleatorio $n_5 \in \mathbb{Z}$ entre $0 - 19$ y conforme a la categoría que tiene el individuo en la cuarta variable, fue asignado el valor de la quinta variable, como lo muestra la siguiente tabla.

Categoría de la cuarta variable	Valores de i	
$s_{41} = 0$	$0 \leq i \leq 19$	-
$s_{42} = [30000, 90000)$	-	$0 \leq i \leq 19$
Categoría de la quinta variable	$s_{51} = 0$	$s_{52} = \{1, 2\}$

Cuadro 3.6: Categoría que se asigna a la quinta variable dependiendo el valor de la categoría s_{4j} de la cuarta variable

Por ejemplo, si el i-esimo sujeto tiene la categoría $s_{42} = [30000, 90000)$ de la cuarta variable y el número aleatorio n_4 está dentro del intervalo 0-19, entonces se le asigno el valor $s_{52} = \{1, 2\}$ que corresponde al número de inmuebles adquiridos de alto valor que el individuo realizó.

Nota: Como vimos la quinta variable "número de inmuebles adquiridos de alto valor" depende de las categorías obtenidas de la cuarta variable " monto por la compra de inmuebles de alto valor", esta información es la que esta contenida en la tabla 3.6.

Creamos dos variables ficticias: z_{51i}, z_{52i} ya que $S_5 = \{s_{51}, s_{52}\}$ que toman los valores suguientes:

- si $x_{5i} = 0$ entonces $z_{51i} = 1$ y $z_{52i} = 0$

- si $x_{5i} \neq 1$ entonces $z_{51i} = 0$ y $z_{52i} = 1$

Cuya representación se ejemplifica abajo:

i	n_4	n_5	categoría x_4	categoría x_5	x_4	x_5	z_{41}	z_{42}	z_{51}	z_{52}
1	9	1	1	1	0	0	1	0	1	0
2	9	3	1	1	0	0	1	0	1	0
3	2	11	1	1	0	0	1	0	1	0
4	14	17	1	1	0	0	1	0	1	0
5	9	14	1	1	0	0	1	0	1	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Finalmente la sexta variable X_6 "Sexo". Sea x_{6i} la $i-$ésima entrada de la variable X_6, es decir la sexta entrada del $i-$ésimo registro, que es el sexo del $i-$ésimo sujeto $0 \leq i < 10000$, para esta variable se generó un número aleatorio $n_6 \in \{0, ..., 19\}$ y se dierón los siguientes valores:

- si $n_6 \in [0, 10)$ entonces $x_{6i} = masculino = s_{61}$
- si $n_6 \in [10, 20)$ entonces $x_{6i} = femenino = s_{62}$

Note que esta variable es independiente de las demás variables. Creamos 2 variables ficticias: x_{61i}, x_{62i} pues $X_6 \in S_6$ donde $S_6 = \{s_{61}, s_6\}$, que fuerón llenadas como se muestra a continuación:

- si $x_{6i} \in [0, 10)$ entonces $z_{61i} = 1$ y $z_{62i} = 0$
- si $x_{6i} \in [10, 20)$ entonces $z_{61i} = 0$ y $z_{62i} = 1$

Con la siguiente representación

n_6	x_6	z_{61}	z_{62}
10	$Masculino$	1	0
5	$Masculino$	1	0
17	$Masculino$	1	0
3	$Masculino$	1	0
⋮	⋮	⋮	⋮

A continuación se puede ver la tabla que muestra parte de las variables generadas, esta tabla como tal, es parte de la base de datos que se simuló.

Sujeto i	x_1	x_2	x_3	x_4	x_5	x_6	z_{11}	z_{12}	z_{13}	z_{21}	z_{22}	z_{23}	z_{24}	z_{31}	z_{32}	z_{33}	z_{41}	z_{42}	z_{51}	z_{52}	z_{61}	z_{62}
1	79900	1	3657	0	0	*Masculino*	0	1	0	1	0	0	0	0	0	1	1	0	1	0	1	0
2	89890	2	44	0	0	*Masculino*	0	1	0	1	0	0	0	1	0	0	1	0	1	0	1	0
3	8123	5	1431	0	0	*Masculino*	1	0	0	0	0	1	0	0	1	0	1	0	1	0	1	0
4	4420	6	520	0	0	*Masculino*	1	0	0	0	0	1	0	0	1	0	1	0	1	0	1	0
5	22085	0	2299	0	0	*Masculino*	0	1	0	1	0	0	0	0	1	0	1	0	1	0	1	0
6	5235	2	2134	0	0	*Femenino*	1	0	0	1	0	0	0	0	1	0	1	0	1	0	0	1
7	5136	5	253	0	0	*Masculino*	1	0	0	0	0	1	0	1	0	0	1	0	1	0	1	0
8	9097	13	3396	0	0	*Femenino*	1	0	0	0	0	0	1	0	0	1	1	0	1	0	0	1
9	13119	11	3289	59801	1	*Masculino*	1	0	0	0	0	0	1	0	0	1	0	1	0	1	1	0
10	4188	1	1312	0	0	*Masculino*	1	0	0	1	0	0	0	0	1	0	1	0	1	0	1	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Cuadro 3.7: Base de datos simulada

De ahora en adelante llamaremos atributo a cualquier posible valor que pueda tomar cualquier entrada del vector X, (note que cada atributo corresponde a una columna en nuestra nueva base de datos simulada) entonces tenemos que el número total de atributos es:

$$K = \sum_{j=1}^{p} |S_j| = 16$$

ver (3.1).

Es importante señalar el hecho que las variables tomadas en consideración para la simulación se observarón en gran parte de los sujetos que realmente han sido procesados, culpados o señalados por cometer actos de corrupción. De la base de datos creada con un total de 10000 sujetos suponemos que 60 de estos han sido procesados por cometer el delito de corrupción. El objetivo de la simulación es identificar a aquellos sujetos en la base de datos que tienen los mismos atributos que poseen los 60 sujetos que han sido identificados como corruptos.

Notemos 2 cosas

1. El conjunto A' a lo más contiene 6 elementos, uno por cada variable

2. Puede ser de interés considerar aquellos atributos que sean comunes a un alto porcentaje de elementos de A, por ejemplo, al 90 o 95 porciento, de ellos en otras palabras atributos con soporte $T(\mathcal{K})$ alto. Y los individuos de la base de datos que tienen los mismos atributos de este subconjunto de A pueden también ser sospechosos de cometer actos de corrupción.

3.3.3. Resultados

De la base de datos de 10000 registros ya generada, se seleccionó un conjunto de 60 individuos, el conjunto A, que supusimos son corruptos procesados, a este conjunto se le aplicó el operador " $'$ " para obtener el conjunto de todos los atributos comunes de los individuos en A, el conjunto A', a este conjunto se le aplica nuevamente el operador " $''$ " , para obtener el conjuntoA'', que es el conjunto de individuos con atributos comunes a los 60 corruptos. Este conjunto de indidividuos, deberían ser investigados. El FCA arrojó como resultado un total de 315 sujetos contenidos en el conjunto " A'' " de los cuales 255 son posibles sujetos que cometen el delito de corrupción (esto por que $A \subseteq A''$).

Capítulo 4

Conclusiones

La identificación de sujetos corruptos es un tema complejo y delicado; en el sentido que hasta el momento no hay bibliografía que profundize y nos de metología o modelos para la identificación de sujetos de esta índole y delicado ya que cualquier asunción o señalamiento que se haga acerca de los sujetos en cuestión debe ser revisado meticulosamente.

Por un lado el modelo de ecuaciones estructurales es una metodología que establece la relación de dependencia entre las variables. Trata de integrar una serie de ecuaciones lineales y establecer cuáles de ellas son dependientes o independientes de otras, ya que dentro del mismo modelo las variables que pueden ser independientes en una relación pueden ser dependientes en otras. Este tipo de modelos aplicados al problema que abordamos resutaría beneficioso en demacia, ya que las entidades encargadas de peseguir este delito, podrían modelar la corrupción de un funcionario público como una variable latente vinculada tanto a variables observables como indicadores las cuales estarán en función del conocimiento empírico de cada organismo. Así mismo se podría establecer algún tipo de métrica sobre los puntajes de la variable latente propuesta, que de alguna manera logre medir en mayor o en menor forma el nivel de corrupción del sujeto y dar una medida o calificación que pueda servir para determinar que sujetos investigar si tenemos un universo basto de funcionarios que requieran ser evaluados o simplemente monitoriados. Y así proveer de una herramienta que ayude a atacar un problema tan grave en México.

Por otro lado se generó una base de datos simulada y se paleteo una propuesta de métodos no supervisados como herreamientas para la identificación de posibles sujetos corruptos.

El algoritmo realizado para la identificación lo consideramos como funcional

ya que nos permitió la identificación de posibles sujetos corruptos en nuestra base de datos simulada, por lo que la metodología podría ser aplicada en la información sobre movimientos financieros y no financieros u otra información relacionada al sujeto que pueda ser de interés analizar, tal como la información de la que dispone las Secretaria de Hacienda. Es claro que una vez que se tenga una medida de riesgo pueda validarse si verdaderamente existen elementos para determinar que los sujetos están incurriendo en actos de corrupción. Esto a su vez permitirá la retroalimentación o ajuste del modelo. Una vez más haciendo hincapié que despues de que los sjuetos hayan sido identificados por el método se deberá efectuar una investigación minuciosa y profunda para determinar si los sujetos que arroja el método en verdad incurren en el delito de corrupción. Si bien, los individuos que encontramos como sospechosos en nuestra base de datos no corresponden a casos reales, si podemos decir que la metodología estudiada si identifica a posibles sospechosos y que puede ser aplicada a bases reales con las que se puede medir su efectividad.

Los métodos no supervisados son métodos relacionales que dan como resultado conjuntos de objetos que son de gran interés en el ánalisis de grandes cantidades de información.

Los objetivos propuestos en un principio en este trabajo fuerón alcanzados, ya que se logro construir una base de datos con información y operaciones financieras para identificar o simular el comportamiento de interés, también se logro definir un modelo matemático o estadístico en términos de indicadores que permita seleccionar sujetos con posibles elementos de corrupción para que sean sometidos a investigación.

Además estas metodologías no son privativas solo para estudiar y atacar el problema de la corrupción sino cualquier otro problema social donde se tenga información que se considere relacional.

Apéndice A

Estadísticas descriptivas de sujetos corruptos de la base de datos obtenida obtenida mediante investigación en periódicos, revistas e Internet.

Se hizo una investigación en los medios noticiosos (periódicos, noticieros televisivos o internet) sobre las características de individuos que han sido, sentenciados, acusados o sospechosos de incurrir en el acto de corrupción. Esta información es pública y no pretendimos hacer esta investigación exhaustiva, los datos recabados dan una pauta del tipo de sujetos que cometen el delito de corrupción, la información fue obtenida de los principales períodicos de México como lo son: La Jornada, El Universal, El Diario de México, y noticieros de televisión

De la información recabada se observó que el 96 % de los sujetos corresponden al sexo masculino, este hecho está ligado a que la mayoría de los funcionarios públicos en México son varones; el 42 % de los sujetos tienen una edad entre 50 y 70 años, el 25 % de los sujetos tienen entre 40 y 50 años; el 36 % de los sujetos vinculados al delito de corrupción asumen o han asumido el cargo de gobernadores en las distintas entidades de la republica mexicana; el 15 % de los sujetos han tenido cargos en el gobierno del D.F, el 11 % ha sido acusado por malversación de fondos, mietras que el 17 % han sido sospechosos de cometer el delito de peculado

Cuadro A.1: Base de datos

Estatus	Caso	Sexo	Nom	F.N	C.P.O	L.C	E.C	O.D	D.C
sospechoso	1	f	Alejandra Sota		Vocera		DF	Presidencia Felipe Calderon	Malversación de fondos
sentenciado	2	m	Andrés Granier Melo	3/5/48	Gobernador	Villahermosa	Tabasco	Estado Tabasco	Malversación de fondos
sospechoso	3	m	Ángel Aguirre Rivero	2/20/56	Senador, gobernador		Guerrero	Estado Guerrero	
sospechoso	4	m	Arturo Montiel	10/15/53	Gobernador		Edo México	Estado de México	Enriquecimiento ilícito
sentenciado	5	m	Armando García Benhumea		Asesor	Tlalpan	Edo México	Estado de México	
sospechoso	6	m	Apolinar Mena Vargas		Secretario de communicaciones	Edo MÃ©xico	Edo México	Estado de México	
sospechoso	7	m	Carlos Hank Rhon	7/4/76	Alcalde	Tijuana	Baja California Norte	Municipio Tijuana	
sospechoso	8	m	Carlos Manuel Villalobos Organista	4/30/05	Tesorero Ascrito de la secretría de hacienda	Sonora	Sonora	Estado de Sonora	Enriquecimineto ilícito
sentenciado	9	m	Carlos Mateo Aguirre Rivero		Funcionario		Guerrero	Estado Guerrero	Malversación de fondos
sospechoso	10	m	Carlos Romero Deschamps	1/17/43	Presidente	CDMX	DF	Sindicato Trabajadores Petroleros de la República Mexicana (STPRM)	Enriquecimiento ilícito
sospechoso	11	m	Cesar Duarte JÁquez	4/14/47	Gobernador		Chihuahua	Estado Chihuahua	Peculado
sospechoso	12	m	Cesar Coli Carabias	4/11/47	director CEA	Guadalajara	Guadalajara	Estado de Guadalajara	
sentenciado	13	f	Elba Esther Gordillo	2/6/45	Presidenta	CDMX	DF	Sindicato Nacional de Trabajadores de la Educación (SNTE)	Malversación de fondos
sospechoso	14	m	Enrique Horcasitas Manjarrez,	68	Director del metro	CDMX	DF	Gobierno del D.F	
sospechoso	15	m	Enrique Peña Nieto	7/20/66	Presidente	México	México	México	
sospechoso	16	m	Fidel Herrera Beltrán	3/7/49	Gobernador		Veracruz	Estado Veracruz	
sospechoso	17	m	Genaro García Luna	7/10/68	Titular	CDMX	DF	Secretaria de Seguridad Pública SSP	
denunciado	18	m	Graco Luis Ramírez Garrido	6/26/49	Gobernador	Villahermosa	Tabasco	Estado de Villahermosa	
sospechoso	19	m	Guillermo Padrés Elías	7/29/69	Gobernador		Sonora	Estado Sonora	Peculado
denunciado	20	m	Humberto Moreira Valdés	7/28/66	Gobernador		Coahuila	Estado Coahuila	
sospechoso	21	m	Javier Alberto Martínez Verduzco		Director general de Control de Fondos y Pagaduría	Sonora	Sonora	Estado de Sonora	
denunciado	22	m	Javier Duarte Ochoa	9/14/73	Gobernador		Veracruz	Estado Veracruz	
sospechoso	23	m	Jorge Emilio GonzÁles Martínez	4/16/72	Presidente	CDMX	DF	Partido Verde	
sentenciado	24	m	Jesús Reyna GarcÃa	2/22/52	Gobernador interino		MichoacÁn	Estado Michoacán	
sentenciado	25	m	Jorge Joaquín Ignacio Serrano Limón		Presidente	CDMX	DF	PROVIDA AC	Peculado
denunciado	26	m	Jorge Torres López	2/20/54	Gobernador interino		Coahuila	Estado Coahuila	
denunciado	27	m	José Aroldo de Hoyos Morales		Gerente de Servicio a Proyecto Norte	CDMX	DF	PEMEX	
sospechoso	28	m	José Andrés de Oteyza	11/21/42	Director, vicepresidente			OHL	
denunciado	29	m	José Manuel Saiz Pineda	8/23/67	Secretario	Villahermosa	Tabasco	Finanzas Tabasco	Peculado y lavado de dlnero
denunciado	30	m	JosÃ© Germán Valenzuela Aganza		Coordinador de Supervisión del Instituto de Vivienda	Sonora	Sonora	Estado de Sonora	
sospechoso	31	m	Juan Sabines Guerrero	8/20/68	Gobernador		Chiapas	Estado Chiapas	Peculado
sospechoso	32	m	Juan Armando Hinojosa Cantú		Presidente			Grupo HIGA	
sospechoso	33	m	Luis Alberto Alvarez RAos		Policía Preventiva	Guanajuato	Guanajuato	Dirección General de Seguridad Ciudadana de Guanajuato	
sospechoso	34	m	Luis Alberto Sotelo González		Director de fondos y pagaduria	Sonora	Sonora	Estado de Sonora	
sentenciado	35	m	Luis Armando Reynoso Femat	8/15/58	Gobernador		Aguascalientes	Estado Aguascalientes	Peculado
sospechoso	36	m	Manlio Fabio Beltrones	8/30/52	Senador/Gobernador/Presidente del PRI		Sonora	Estado de Sonora	
sospechoso	37	m	Mario Cësar Cuen Aranda		Tesorero Ascrito de la secretaría de hacienda	Sonora	Sonora	Estado de Sonora	
sentenciado	38	m	Mario Medina Martínez		Servirdor Público	Tlalpan	Edo México	Estado de México	
sentenciado	39	m	Manuel Añorve Baños	9/15/57	Diputado local,presidente municipal/diputado federal	Acapulco	Guerrero	Municipio Acapulco	
sentenciado	40	m	Mario Villanueva Madrid	7/2/48	Gobernador		Quintana Roo	Estado Quintana Roo	Peculado
sospechoso	41	m	Narciso Agúndez Montaño	10/26/58	Gobernador		Baja California Sur	estado Baja California Sur	Peculado
sentenciado	42	m	Óscar Espinoza Villareal	11/23/53	Jefe DE Gobierno		DF	Distrito Federal	Peculado
sentenciado	43	m	Pablo Salazar Mendiguchia	10/9/54	Gobernador		Chiapas	Estado Chiapas	
sospechoso	44	m	Ramiro García Cantú		Empresario			Contratista petrolera grupo R	
sospechoso	45	m	Ricardo Monreal Ávila	9/19/70	Coordinador de la Agrupación Política Morena	CDMX	DF	Distrito Federal	
sospechoso	46	m	Roberto Romero Lopez		Secretario de gobierno		Sonora	Estado de Sonora	
denunciado	47	m	Rodrigo Medina de la Cruz	4/9/72	Gobernador		Nuevo León	Estado de Nuevo León	Desvio de fondos
denunciado	48	m	Rodolfo Ocampo		Director de intermunicipal de agua potable y alcantarillado	Guadalajara	Guadalajara	Estado de Guadalajara	Malversación de fondos
sospechoso	49	m	Rosario Robles Berlanga	1956	Diputada Federal, Jefa de Gobierno		DF	Distrito Federal	
denunciado	50	m	TomÁs Yarrington Ruvalcaba	3/7/59	Gobernador		Tamaulipas	Estado Tamaulipas	Lavado de dinero
Sentenciado	51	m	VÃctor Ignacio Hughes Alcocer		Subsecretario de finanzas		Guerrero	Estado de Guerrero	
denunciado	52	m	VÃctor Martínez Trujillo		Titular de sistemas de caminos de Nuevo leon	Nuevo León	Nuevo León	Estado de Nuevo León	

Cuadro A.2: Donde Nom=nombre del funionario, F.N=fecha de nacimiento, C.P.O=cargo público ocupado, L.C=localidad de cargo, E.C=estado de cargo, 0.D=organización o dependencia, D.C=descripción del caso

Bibliografía

[1] Anderson R & Babin B; Black W; Hair J., (2010).“*Multivariate Data Analysis: A global perspective*” , Prentice Hall, 7th edition, pp 91-150, 627-713.

[2] Ash R., (2000).“*Probability and Measure Theory*”, Academic Press, 2nd edition,pp 449-451

[3] Bollen K., (1989). “*Structural Equations with Latent Variables*”, New York, NY: John wiley & Sons. pp 1-170.

[4] Bruce H. & Adrian T. & Alexander V., (2003), “*Structural equation modeling: Applications in ecological and evolutionary biology*”, Cambridge University Press, pp 1-42.

[5] Buehn A. & Schneider F., (2009). “*Corruption and the Shadow Economy: A Structural Equation Model Approach*”, IZA Discussion Paper No. 4182.

[6] Busquet R., (2005), “*Factores que propiciarón la corrupción en México. Un análisis del soborno a nivel estatal*”,

[7] Cárdenas J; Mijangos M., (2005), “*Acerca del marco teórico de la corrupción.*

[8] Casares A., (2015). “*México: Anatomía de la corrupción*”, CIDE, Instituto Méxicano para la Competitividad A.C.

[9] Castillo A., “*Medición de la corrupción: Un indicador de la rendición de cuentas*”, Serie: Cultura de la Rendición de Cuentas, pp 5-57.

[10] Friedman J; Hastie T; Tibshirani R., (2008) “*The Elements of Statistcal Learning*, Second Edition, Springer Editorial, pages 485-515.

[11] Ganter B., (2002). "*Formal Analysis: Methods and Applications in Computer Science*", pp 1-22.

[12] Gómez C., (2011) "*Estimación de los modelos de ecuaciones estructurales, del índice mexicano de la satisfacción del usuario de programas sociales mexicanos, con la metodología de mínimos cuadrados parciales*", tesis.

[13] Gonzáles E., (1993) "*País de un solo hombre: El México de Santa Anna*, pp 25-57.

[14] Gonzáles E., (2005) "*La corrupción: patología colectiva*".

[15] Jain, A., (2001) "*Corruption: A Reeview*", Journal of Ecoomics Surveys, pp 71-83

[16] Lara A., (2014) "*Introducción a las ecuaciones estructurales en AMOS Y R*".

[17] Michael J., (2005) "*Es posible medir la corrupción, ¿pero podemos medir la reforma?*", Universidad Nacional Autónoma de México-Instituto de Investigaciones Sociales. Revista Mexicana de Sociología, año 67, núm. 2.

[18] Pang-Ning T, Steinbach M, Vipin K., (2006) "*Introduction to Data Mining*", Pearson Adison-Wesley, pp 71-103.

[19] Pichardo-Mendoza.,Ángel Tamariza-Mascúa, "*Álgebras Booleanas y Espacios Topológicos*".

[20] Rakesh A., Tomasz I., Arun S., (2010) "*Mining Association Rules between Sets of Items in Large Databases*"

[21] Radim B., (2008). "*Introduction to formal concept Analysis*", Departament of Sciencie, Palacky University, pp 10-15.

[22] Recuero J., (2008) "*Organización de Resultados de Búsqueda Medíante Análisis Formal de Conceptos*", tesís doctoral, pp 65-72

[23] Rencher A., (2002). "*Methods of Multivariate Analysis*", Brigham Young University: John wiley & Sons. pages 408-435.

[24] Smith P., (1994)"*Assessing the Size of the Underground Economy: The Canadian Statistical Perspectives*", pages 10-18

[25] Troncoso S., (2002) "*Descubrimiento y poda de reglas de asociación en el marco del análisis de canasta de mercado*".

[26] Uta P., (2008) "*Formal concept analysis in information science*".

[27] $www.unodc.org/unodc/en/corruption.html.$

[28] $www..org/news\_room/faq.$

[29] $http://www.transparency.org/research/cpi/overview.$

[30] $www.worldbank.com$

[31] $www.diputados.gob.mx > pdf$

[32] $www.www, cnbv.gob.mx > documentos$

[33] $www.shcp.gob.mx > LASHCP > nacional$

[34] $www.shcp.gob.mx > inteligencia1$

[35] $www, gob.mx/shcp/documentos/uif - marco - juridico$

[36] $www.sppld.sat.gob.mx/pld/interiores/leermas.html$