UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE FILOSOFÍA Y LETRAS

COLEGIO DE GEOGRAFÍA

APLICACIÓN DE LA MINERÍA DE DATOS EN GEOGRAFÍA

T E S I S

PARA OBTENER EL TÍTULO DE:

LICENCIADA EN GEOGRAFÍA

PRESENTA:

Leyva Jiménez Alejandra

ASESOR DE TESIS:

Mtro. Jaime Morales

CIUDAD UNIVERSITARIA, CD. MX., 2019

Agradecimientos

A mi mamá, por su amor incondicional; a Yanina, por ser nuestro equilibrio y compañía más sincera; a Erika y Omar, por cada una de sus oraciones; a Gaby y Víctor, por su guía y respaldo; al maestro Jaime, por su cobijo y apoyo; a Socorrito, por su amistad y cariño.

A mi familia y amigos, gracias por guiar, sostener y acompañar mi camino.

Por sus valiosas observaciones, a la Mtra. Bere Álvarez y al Mtro. Armando García de León. A mis profesores de la licenciatura, en especial, por su cariño y compromiso hacía la docencia y su amor a la geografía, a Tobyanne Berenberg, Jaime Morales, Mario Casasola y Macario Arredondo.

Contenido

Índice de figuras

Índice de tablas

Índice de gráficos

INTRODUCCIÓN

En la presente investigación se abordará la aplicación que tiene la minería de datos en el análisis de información en geografía. La minería de datos ha cobrado importancia en los últimos años debido a dos aspectos principales; el desarrollo de la computación y la informática y, por la generación y almacenamiento continuo de información. Cualquier actividad humana es capaz de dejar un registro almacenable y susceptible a analizarse, así también, un fenómeno u hecho físico es proclive a generar grandes bancos de datos. Ejemplo de lo anterior son las transacciones bancarías, los registros en un aeropuerto o una carretera de peaje o los datos de humedad, temperatura, precipitación o viento que se almacenan en una estación atmosférica.

La minería de datos se define como el proceso que permite interactuar y analizar grandes volúmenes de información mediante un conjunto de técnicas y métodos y el desarrollo de algoritmos y automatización de procedimientos; su objetivo consiste en descubrir la información oculta de una base de datos para posteriormente, a partir de ella, extraer conocimiento. Su aplicación se encuentra en ámbitos muy variados como el análisis de una obra literaria, campañas electorales o publicitarias, enfermedades, servicios sociales, prevención de desastres o elaboración de escenarios.

El nexo entre la minería de datos y la geografía se halla en los fundamentos de la geografía automatizada. Gran parte de la información que se genera en distintos ámbitos de la vida diaria tiene una referencia espacial que permite ser representada y caracterizada en un plano cartográfico, esta característica permite también que los datos puedan analizarse mediante el uso de tecnologías de la información geográfica; es precisamente en la extracción de información de un conjunto de datos mediante la automatización de procedimientos en donde se halla el vínculo entre ambas.

Si consideramos la revaloración de la dimensión espacial en los últimos años dentro de las ciencias sociales y el desarrollo constante de las innovaciones tecnológicas, la geografía tiene la oportunidad de posicionarse como especialista en el tratamiento e interpretación de grandes volúmenes de información geoespacial. Así también, la minería de datos brinda la oportunidad a la geografía de abordar problemáticas desde otros enfoques, obtener datos más allá de las fuentes oficiales y abordar la realidad desde su dinamicidad. Cabe destacar que en la presente investigación se hace hincapié en la importancia del saber teórico del área del conocimiento desde donde se aborda la problemática a tratar y se considera al método como un medio para analizar un conjunto de información y no como un fin.

Hipótesis

Se pueden aplicar técnicas de minería de datos, como el Análisis de Correspondencia Simple, para la extracción y análisis de información de un conjunto de datos para distintas áreas del conocimiento en geografía.

Objetivo general.

Aplicar una técnica de minería de datos para la extracción y análisis de información en el campo de la geografía social, geografía de la población y geografía de los riesgos.

Objetivos particulares.

1. Elaborar un marco teórico en torno a la minería de datos y su vínculo con la geografía.
2. Exponer las características y el procedimiento de una técnica en minería de datos: Análisis de Correspondencias Simple.
3. Mostrar la aplicación del Análisis de Correspondencias Simple en tres áreas del conocimiento en geografía.

Para cumplir con el objetivo, la investigación se estructuró en tres apartados. En el primero, se elaboró un marco teórico en torno a la minería de datos, en este capítulo se exponen los conceptos básicos, técnicas de las cuales se apoya, su método general y algunas aplicaciones. Posteriormente se hace un breve recorrido histórico por el desarrollo de la geografía cuantitativa y la geografía automatizada para finalmente, explicar el vínculo entre la minería de datos y la geografía.

El segundo capítulo tuvo por objetivo exponer las características de una técnica en el tratamiento de variables discretas en la minería de datos: el análisis de correspondencias simple. En el desarrollo del presente capítulo se hizo uso de un software para la ejemplificación del método y el análisis de las salidas de resultados. Si bien se incluyen varias expresiones matemáticas, su incorporación se consideró necesaria para una mejor comprensión del método.

El tercer capítulo demostró la aplicación del análisis de correspondencias simple en tres áreas del conocimiento en geografía. Para ello, se seleccionaron tres bases de datos, la primera relacionada con la muerte de periodistas en el periodo comprendido entre 1994 y 2019; la segunda con la migración interna de jornaleros agrícolas a los campos de cultivo de la región norte y noroeste, centro y occidente y sureste del país; y finalmente, la tercera vinculada a la ocurrencia de fenómenos hidrometeorológicos y geológicos en México como son las lluvias, ciclones tropicales, inundaciones, sequías, nevadas, bajas temperaturas y sismos.

CAPITULO 1: APROXIMACIONES TEÓRICAS A LA MINERÍA DE DATOS Y SU VÍNCULO CON LA CIENCIA GEOGRÁFICA

Diariamente se generan grandes volúmenes de datos de manera intencional e involuntaria, cuando realizamos una búsqueda en la Internet, una transferencia bancaría, una llamada telefónica, una compra, cuando abordamos un transporte o nos movemos de un lugar a otro, se está generando información. Periódicamente se registran y actualizan datos relacionados con variables atmosféricas, accidentes viales, delitos, epidemias, nacimientos, compra – venta de productos agrícolas, movilidad, consumo de narcóticos, audiencias radiofónicas, monitoreo de tránsito aéreo, etc. La información que se genera día con día permite monitorear, observar, investigar, planificar, gestionar y pronosticar distintos escenarios y acontecimientos del ámbito social, ambiental, económico, político o de negocios (Molina, 2002).

Sin embargo, los datos generados, recolectados y almacenados en un repositorio, por sí mismos, no tienen mayor relevancia si no se cultiva la habilidad para transformarlos en información que permita comprender con mayor profundidad un fenómeno, resolver problemas o tomar decisiones; toda vez que el "valor táctico y estratégico" de los datos reside en su análisis y en la capacidad humana de leerlos, comprenderlos, darles sentido y descifrarlos en su contexto, "los datos en bruto" carecen de valor en cuanto a los beneficios que pudieran tener en su aplicación (Riquelme, Ruíz y Gilbert, 2006:12, Beltrán, 2003:9).

La necesidad de desarrollar y adecuar métodos para comprender e interactuar con grandes volúmenes de datos deviene del crecimiento opuesto entre la capacidad de recolección y almacenamiento de los datos y la capacidad humana para procesarlos, manipularlos, analizarlos y representarlos (Riquelme, Ruíz y Gilbert, 2006:12, Molina, 2002). Ante ello, la curiosidad e interés por comprender de manera sistemática un gran conjunto de datos ha permitido el desarrollo de conceptos y métodos relacionado con la minería de datos (Ballesteros, Iñiguez y Velazco, 2018: 343).

Hasta este punto, surge una interrogante ¿Qué factores han permitido el almacenamiento masivo de información? De acuerdo con Hernández, Ramírez y Ferri (2004:3), Beltrán (2003:11) y Molina (2002:2) entre las principales causas se encuentran las siguientes:

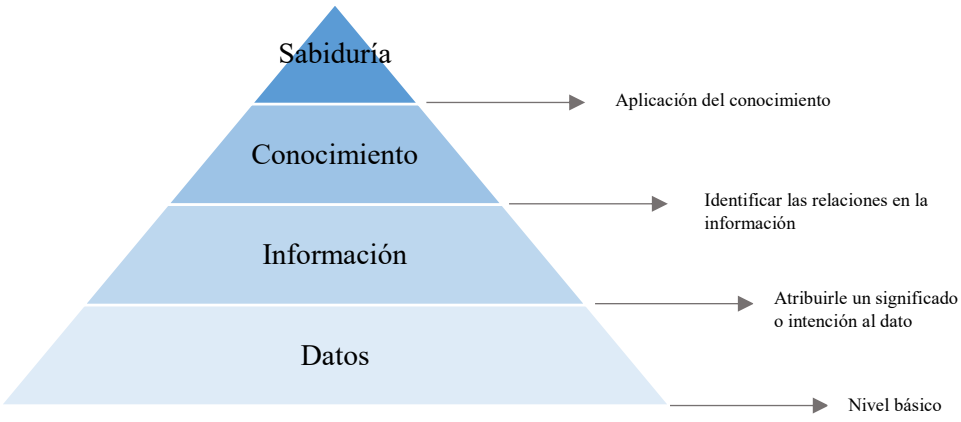
- Desarrollo tecnológico de las ciencias computacionales
- El incremento en el procesamiento y velocidad de los equipos de cómputo
- Disminución de los costes en los sistemas de cómputo y almacenamiento

- Desarrollo y ajuste de métodos y técnicas de recolección y administración de bases de datos (BD)
- Mejora en la eficiencia de procesos de captura, acopio, distribución y transferencia de BD
- Desarrollo y mejora de software especializado en BD
- Mejoramiento en la percepción sobre la confiabilidad en la transmisión de datos

De acuerdo con Hernández, Ramírez y Ferri, (2004:3) el creciente interés alrededor de las grandes bases de datos versa sobre una sola causa; la necesidad, apreciación y reconocimiento del valor potencial e indiscutible de los datos almacenados en los sistemas de información. La relativa facilidad para acceder a grandes y variados conjuntos de datos, así como la difusión generalizada de herramientas y plataformas informáticas, ha modificado la manera en cómo interactuamos y nos comunicamos con ellos. Respondiendo a la necesidad apremiante de extraer información relevante que a simple vista los datos no muestran y de disponer de información precisa, la minería de datos se ha consolidado como un conjunto de técnicas orientada a comprender en su totalidad la estructura de una base de datos (Pérez y Santín, 2007:1).

Modelo: Datos – Información – Conocimiento – Sabiduría (Figura 1)**.** Para comprender mejor la importancia de los datos es necesario contextualizar su relevancia en el proceso de la generación de conocimiento. Como punto inicial, los datos son la materia prima que al momento de atribuírsele un significado, se convierten en información, a ésta le sumaremos un valor agregado que es la interpretación que le da el investigador o analista para explicar un modelo o comportamiento de la realidad; cuando se analizan las relaciones existentes en la información, se está generando conocimiento. Finalmente, la sabiduría deviene en la capacidad de aplicar el conocimiento en el momento preciso y de acuerdo a las exigencias de la situación enfrentada (Hernández, Ramírez y Ferri, 2004:3, Molina, 2002:2).

Figura 1: Pirámide del conocimiento



Elaboración propia con base en Olvera (2014)

1.1. MINERÍA DE DATOS

En el ámbito académico, la minería de datos es una etapa de un proceso más grande conocido como extracción de conocimiento a partir de bases de datos o KDD por sus siglas en inglés (Knowledge Discovery in Database). Su utilización se remonta a finales de 1980 y congrega distintas áreas del conocimiento como la estadística, informática, bases de datos, el aprendizaje automático y la inteligencia artificial. Descubrir el conocimiento que guarda una base de datos requiere de al menos tres procesos; la búsqueda de patrones, el establecimiento de modelos y la interpretación de los datos. (Hernández, Ramírez y Ferri, 2004:13-19, Beltrán, 2003:7-12). A continuación, se presentan las etapas que conforman este proceso, siendo la segunda, el elemento central a considerar en el desarrollo del presente capítulo:

1. Recolección y preparación de datos
2. Minería de datos
3. Validación, evaluación e interpretación de la información obtenida
4. Difusión
5. Seguimiento

En la literatura, la Minería de Datos (MD) se define como el proceso de análisis que vinculado a un conjunto de técnicas, métodos y algoritmos, permite seleccionar, condensar, inspeccionar, transformar, modelizar y evaluar grandes volúmenes de información tanto estructurada como no estructurada, se aglomera en bases de datos masivas y complejas recopiladas en un lapso de tiempo de manera periódica y se almacena en diferentes formatos (Pérez y Santín, 2007:7, Hernández, Ramírez y Ferri, 2004: 5). El objetivo de la minería de datos es descubrir conocimiento de un gran conjunto de datos mediante técnicas y métodos automatizados.

Hacer referencia a “grandes volúmenes de datos” pudiera parecer ambiguo, sin embargo, cuando nos referimos a ello, hacemos alusión a todo aquel conjunto de datos que rebasa la capacidad humana de recolección, tratamiento, análisis, procesamiento y representación de las herramientas que comúnmente son utilizadas en la gestión de datos (Beltrán 2015:102). Arcila, Barbosa y Cabezuelo (2016:627) anotan que en muchas ocasiones las grandes bases de datos con las que trabaja la minería de datos no fueron recadas con tal propósito, sin embargo, por sus características y complejidad, la MD figura como un mecanismo adecuado y útil para su análisis.

Entre los principales atributos que caracterizan a la minería de datos podemos encontrar los siguientes (Gutiérrez, García y Salas, 2016:5, Gutiérrez, 2018:199):

1. Volumen. Interacción con grandes bases de datos (Terabytes – Petabytes)
2. Velocidad. Reducido periodo de tiempo en que los datos son producidos y procesados
3. Variedad. Heterogeneidad del tipo, formato y fuente de datos
4. Exhaustividad. Cobertura de los datos
5. Resolución. Escala temporal y espacial
6. Flexibilidad. Diversidad de los usos de una base datos

La aplicación de la minería de datos en distintas áreas del conocimiento es variada, se encuentra en distintos ejemplos que van desde el análisis de redes, microorganismos, sociedades, textos literarios, etc. (Figura 2).

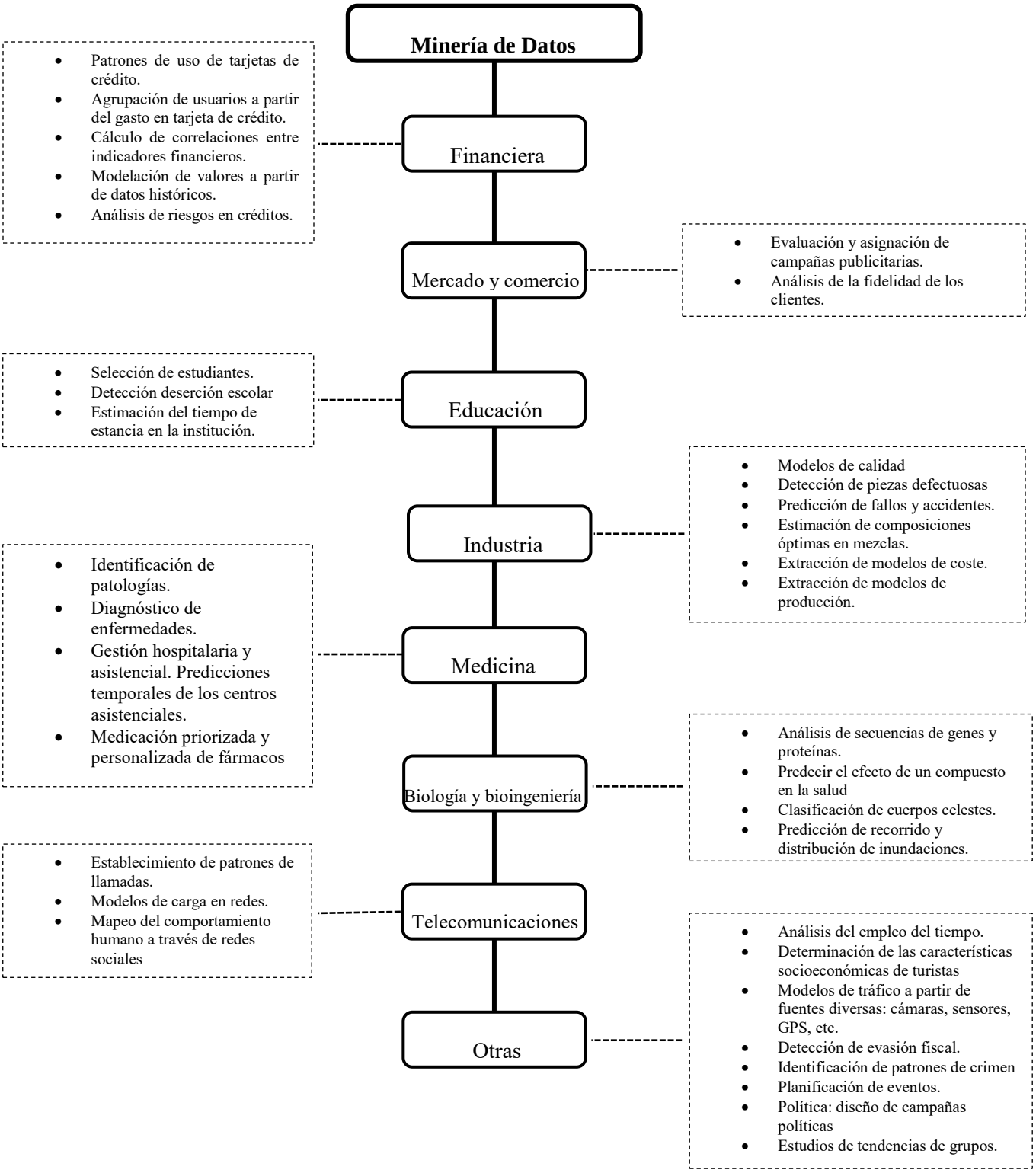
¿Cómo logra la minería de datos interactuar con grandes volúmenes de datos? La minería de datos se auxilia de procesos o algoritmos computacionales iterativos automáticos y semiautomáticos (Ballesteros, Iñiguez y Velazco, 2018:345). Se apoya de diferentes y muy variadas áreas del conocimiento como son; el razonamiento matemático, análisis estadístico, inteligencia artificial, bases de datos, computación gráfica, aprendizaje automático, procesamiento masivo, gestión y soporte de toma de decisiones (Beltrán, 2003:19, Molina, 2002:3).

La minería de datos posee un abanico de potencialidades permitiendo entre otros aspectos:

1. Filtrar, extraer, analizar, visualizar y modelar la información oculta y relevante de valor para el investigador (Ballesteros, Iñiguez y Velazco, 2018:340-342).
2. Caracterizar y cuantificar asociaciones o correlaciones significativas (Ballesteros, Iñiguez y Velazco, 2018:344, Beltrán, 2003:19).
3. Identificar y describir patrones desconocidos y útiles a veces ocultos (Riquelme, Ruíz y Gilbert 2006:12).
4. Descubrir tendencias, desviaciones y regularidades (Beltrán, 2003:18).
5. Explorar, analizar y comprender anomalías o comportamientos atípicos (Pérez y Santín, 2007:2).
6. Obtener trayectorias ocultas y estructuras que no siempre resultan aparentes (Beltrán, 2003:18).
7. Describir y explicar el comportamiento de los datos (Ballesteros, Iñiguez y Velazco, 2018:340).
8. Descubrir y proporcionar conocimiento útil para el soporte de la toma y dirección de decisiones y acciones y que contribuya en la reducción del sesgo y la incertidumbre (Riquelme, Ruíz y Gilbert, 2006:12, Beltrán, 2003:18 Molina 2002:3).

9. Predecir y modelar comportamientos futuros (Pérez y Santín, 2007:2).

Figura 2: Aplicaciones de la Minería de Datos



Elaboración propia con base en Hernández, Ramírez y Ferri, 2004:16-17

El objetivo medular de la minería de datos es la extracción de información almacenada en una gran cantidad de datos. El procesamiento de grandes bases de datos exige la automatización de un conjunto de operaciones de manera secuencial para la obtención, captura, almacenamiento, limpieza, análisis, predicción y presentación de la información. La automatización permite estructurar la información que, en posterior, pretende ser analizada, se espera que los datos se conviertan en información y conocimiento para retroalimentar acciones o decisiones en torno a una problemática (Ballesteros, Iñiguez y Velazco, 2018:340-343, Hernández, Ramírez y Ferri, 2004:14).

La ventaja competitiva de la MD descansa en la capacidad de analizar una gran cantidad de datos. Juanes (2014:94) menciona que la base del conocimiento son los datos, al darle sentido a estos se obtiene información y ésta al ser utilizada, se convertirte en conocimiento. Conocer el momento oportuno para disponer de dicho conocimiento nos refiere a la sabiduría o inteligencia.

1.2. MODELOS DE MINERÍA DE DATOS

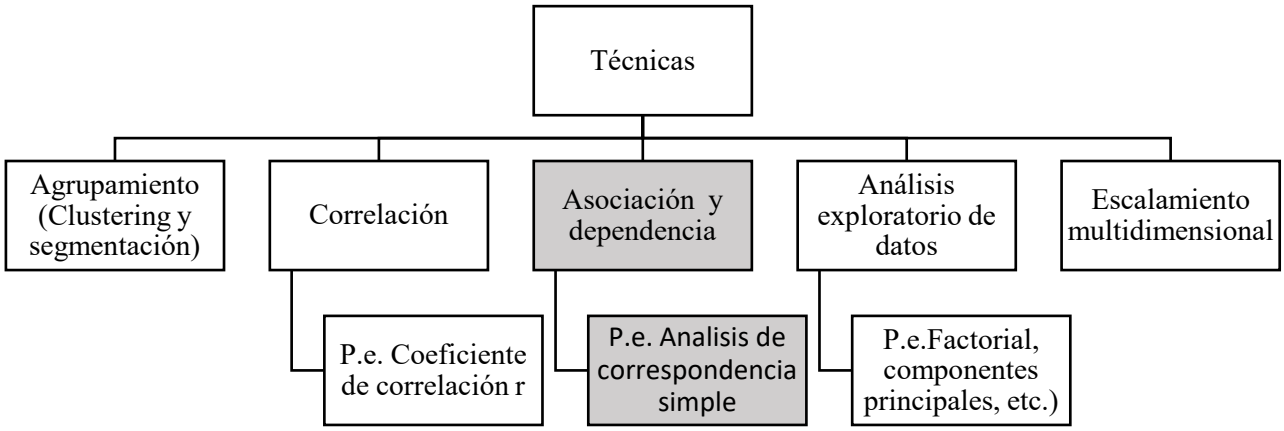
En minería de datos podeos encontrar dos principales modelos: el descriptivo y el predictivo, cada uno de ellos persigue objetivos distintos, se nutre de conjuntos de datos particulares y utiliza diferentes técnicas de procesamiento.

1.2.1. Modelo descriptivo de minería de datos

El modelo descriptivo en la minería de datos tiene por objetivos identificar patrones, tendencias o reglas, explorar y describir las propiedades y características, proporcionar información sobre las relaciones y resumir los datos. Lo anterior permite robustecer el conocimiento sobre un conjunto de datos y descubrir información o rasgos desconocidos antes del análisis (Hernández, Ramírez y Ferri, 2004:12, Beltrán, 2003:31). Por ejemplo, se pretende identificar el patrón de comportamiento de un conjunto de municipios con alto índice de delincuencia con el fin de planificar y ejecutar un plan de acción que reduzca de manera significativa la incidencia de hechos delictivos. Es importante destacar que los modelos predictivos no tienen por objetivo hacer conjeturas o pronosticar a través de los datos, únicamente, mediante ellos, pueden proponerse o generarse hipótesis (Beltrán, 2003:24).

Las técnicas asociadas a los modelos descriptivos generalmente se basan en la generación de patrones. Entre las principales técnicas se encuentran las siguientes:

Figura 3: Técnicas. Modelos descriptivos



Elaboración propia con base en Pérez y Santín, 2007:6-8, Hernández, Ramírez y Ferri, 2004:27

1.2.2. Modelo predictivo de minería de datos

Los modelos predictivos tienen por objetivo estimar o atribuir un valor o una característica a un objeto con base en el conocimiento extraído de datos anteriores. Mediante el desarrollo de un algoritmo supervisado o predictivo se pretende asignar un valor a una variable (objetivo o dependiente) usando un dato conocido de otra variable (variable independiente o predictiva). Para poder llevar a cabo la tarea anterior, el algoritmo considera qué atributos se relacionan a la etiqueta de un objeto para asociarlos a posteridad con otro (Hernández, Ramírez y Ferri, 2004:12, Beltrán, 2003:31). Por ejemplo, con base en el patrón de comportamiento del cuatrimestre anterior de los municipios con mayor índice de delincuencia se identificaron al menos cinco variables que inciden directamente en dicho fenómeno, se pretende identificar a los municipios que aún no presentan altos índices de delincuencia pero que, con base en el estudio anterior, se aproximan a los parámetros establecidos, todo ello con el propósito de diseñar e implementar un conjunto de acciones preventivas.

Las técnicas asociadas a los modelos predictivos se basan en un conocimiento profundo de los datos, las principales son la clasificación y la regresión, sin embargo, también encontramos otras técnicas como las series de tiempo, análisis de ANOVA, el análisis discriminante, los métodos bayesianos, arboles de decisión y redes neuronales (Pérez y Santín, 2007:6-8, Hernández, Ramírez y Ferri, 2004:26).

1.3. METODOLOGÍA GENERAL EN EL PROCESO DE MINERÍA DE DATOS

Frente a la tarea de analizar una gran cantidad de datos se vuelve apremiante el desarrollo, adecuación y fortalecimiento de métodos que permitan imprimir la mayor cantidad de información y

conocimiento posible. El proceso general de la minería de datos, demanda el desarrollo de un algoritmo que garantice el descubrimiento, reconocimiento y extracción de patrones de información de un conjunto de datos que, analizados mediante otros métodos, podrían permanecer ocultos (Arcila, Barbosa y Cabezuelo, 2016:625).

A continuación, se describen las etapas del proceso general de la minería de datos:

1. Definición de objetivos. Esta etapa es fundamental ya que establece hacía donde debe dirigirse el resto del proceso. Es clave para la selección del modelo y tarea de minería de datos y en ella se determinan los resultados esperados (Molina, 2002:4).
2. Identificación de los datos. Se definen los tipos de datos que se requieren, su escala de medida, ubicación y accesibilidad (Beltrán, 2003).
3. Construcción y diseño de la base de datos. En esta etapa se organizan y ordenan los datos en el formato más apropiado. En algunas ocasiones, esta etapa no será necesaria ya que las bases de datos pudieran estar disponibles previo a la definición de los objetivos, es ese caso la etapa siguiente será de gran importancia (Beltrán, 2003).
4. Selección y procesamiento de los datos. Seleccionar los datos que serán parte del análisis, suprimir variables secundarias conservando aquellas que aporten varianza a la investigación y establecer la estrategia para el tratamiento de datos incorrectos, redundantes o extremos (Beltrán, 2003). Esta etapa tiene por objetivo eliminar los datos o variables que puedan provocar un error y de acuerdo a Molina (2002:4) representa el 70% del proceso.
5. Determinación o elección del modelo. Considerando los objetivos planteados en la primera etapa y la escala de medida de los datos, se define el modelo más apropiado para el análisis. (Riquelme, Ruíz y Gilbert 2006:15, Hernández, Ramírez y Ferri, 2004: 25 – 29, Molina, 2002:4).
6. Elección del algoritmo o tarea de minería de datos. La determinación del algoritmo se define de acuerdo con el modelo en minería de datos, los objetivos planteados y la escala de medida. Las tareas desarrolladas en minería de datos tienen objetivos muy específicos que, aparejados con los objetivos de la investigación, permiten coherencia en el proceso (Hernández, Ramírez y Ferri, 2004:25 – 29)
7. Extracción de información. Se aplica el algoritmo y si es necesario se adecua la secuencia a la situación actual de los datos. En esta etapa se espera extraer la mayor cantidad de información y conocimiento posible y aunque pudiera considerarse la etapa más importante, el desarrollo meticuloso de los pasos anteriores define en gran medida que los resultados sean satisfactorios (Beltrán, 2003).

8. Evaluación y análisis de resultados. Se procede a validar y cotejar la coherencia de los valores obtenidos. Si no se ha llegado a la información esperada se procede a verificar la consistencia de los pasos anteriores (Beltrán, 2003, Molina, 2002:4).

1.4. TÉCNICAS DE MINERÍA DE DATOS

Las técnicas en minería de datos están ligadas a los modelos, si bien existen una gran cantidad de ellos, a continuación, se presentan aquellas que por su uso y aplicación resultan de mayor utilidad.

1.4.1. Técnicas de minería de datos para modelos descriptivos

a. Agrupamiento.

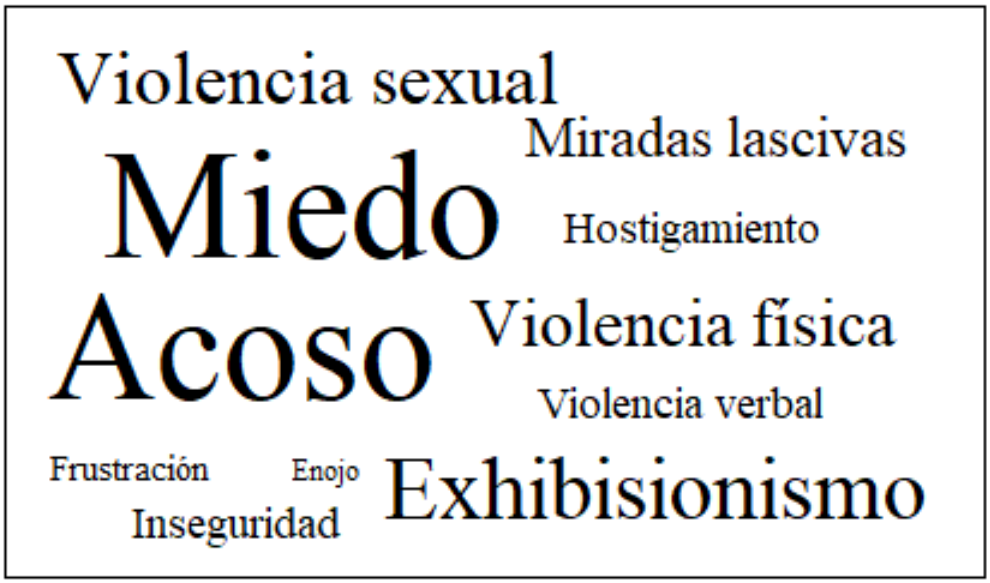
La agrupación tiene como finalidad la obtención de grupos a partir de los atributos de los datos, se fundamenta en la maximización de similitudes entre los individuos o elementos que conforman un grupo y la minimización de similitudes entre los elementos del resto de los grupos (Jaramillo, 2016:5-6). Los grupos son subconjuntos que permiten resumir y describir las características de los elementos del universo de modo que los objetos pertenecientes a un grupo, presentan características semejantes entre sí y diferentes a los objetos de los otros grupos, cada grupo tiene la cualidad de ser exhaustivo y excluyente, es decir, todo objeto pertenece a un grupo y corresponde a uno solo (Beltrán, 2003:25).

En otros ámbitos, la agrupación puede ser sinónimo de segmentación toda vez que ésta última, divide o segmenta un conjunto de datos en grupos, en el caso contrario, la agrupación no es equivalente a la clasificación puesto que ésta parte de la existencia de una etiqueta que asigna una característica a un objeto, mientras que el agrupamiento asigna la etiqueta después del ejercicio (Hernández, Ramírez y Ferri, 2004:26).

Por ejemplo.

Un grupo de investigadoras desea mapear la violencia que viven las mujeres en un campus universitario. Para ello se ha invitado a compartir en la página oficial de Facebook de la universidad, las experiencias de violencia que han vivido las universitarias, docentes y trabajadoras en las instalaciones, la publicación debe acompañarse del hashtag #libresdeviolenciaUNP más el lugar donde ocurrió la incidencia. A continuación, se pretende que la información compartida pase por un algoritmo que de manera automática agrupe las palabras contenidas en los comentarios de acuerdo al lugar donde se vivió la agresión y a la postre, se genere un gráfico de nube de palabras o etiquetas con referencia al sitio.

Figura 4: Nube de palabras para el sitio A: Estacionamiento.



Elaboración propia

b. Correlaciones

El análisis de correlaciones tiene por objetivo corroborar la existencia o ausencia de similitud y su dirección entre los valores de una o más variables continuas medidas generalmente en escala de intervalo o razón. Este método es de gran interés cuando se desea conocer si existe relación entre un par o un grupo de variables, un ejemplo clásico describe la relación entre la diminución de la temperatura y el aumento del chirrido de un grillo por minuto, sin embargo, su aplicación en distintas áreas del conocimiento es variada.

Una de las medidas más populares es el coeficiente de correlación Pearson, el valor de ésta medida comprende un rango de -1 a 1, valores próximos a -1 indican una correlación negativa, es decir, a medida que disminuye el valor de la variable A, aumenta el valor de la variable B o viceversa, en el caso contrario, valores cercanos a 1 indican la existencia de una correlación positiva, es decir, a medida que aumentan o disminuyen los valores de la variable A, también lo hacen los de la variable B. Se dice que una variable está perfectamente correlacionada cuando los valores de r son iguales a -1 o 1, cuando el valor es igual a 0 se dice que no hay correlación lineal, es decir, las variables no están relacionadas entre sí (Hernández, Ramírez y Ferri, 2004:25).

Por ejemplo.

Se desea conocer la relación que existe entre los niveles de contaminación de una urbe y la población. Para afirmar que existe asociación entre las variables se procederá a realizar un análisis de correlaciones, los datos se obtendrán de distintas fuentes, para el caso de los niveles de contaminación, se tomarán los registros diarios que generan los sensores destinados a medir los gases contaminantes, en lo referente a la densidad del uso del espacio se tomarán los datos de los registros generados por los teléfonos móviles. De esta manera se podrá afirmar en qué medida el aumento o disminución de contaminantes está relacionada con la densidad de población, de manera adicional, la naturaleza de los datos, permitirían conocer esta asociación tomando en cuenta las variaciones durante el día y su distribución en cada área de la urbe (Gutiérrez, 2018:201).

c. Reglas de asociación.

Las reglas de asociación tienen por objetivo hallar o descartar la existencia de relaciones o correlaciones tácitas entre los atributos de los objetos de una base de datos. Una diferencia con el método anterior es que las reglas de asociación atienden principalmente variables medidas en escala nominal u ordinal. Basado en reglas de asociación, la finalidad última del método es el reconocimiento de patrones o tendencias en los datos, de modo que cuando un atributo cualquiera toma un valor A, el valor de un atributo distinto se asociará a un valor de B. Hasta este punto es importante mencionar que las reglas de asociación no necesariamente implican una relación causa – efecto (Jaramillo, 2016:5- 6 Hernández, Ramírez y Ferri, 2004:27).

Por ejemplo.

Se desea formular un modelo que explique el ritmo y uso del espacio en una ciudad considerando la hora del día y los días de la semana. Para realizar el análisis de asociación se considerarán dos variables, la primera está asociada a la movilidad de la población y se mide en escala absoluta o de conteo, para ello se tomarán en cuenta los registros georreferenciados de los usuarios activos en Twitter. La segunda variable medida es escala nominal, pretende medir el uso del espacio, para ello se tomarán en cuenta los usos de suelo provenientes de fuente oficiales (Tabla 1). De esta manera se pretende conocer la dinámica de una ciudad considerando las asociaciones existentes que marcan los ritmos y usos del espacio a lo largo del día y durante la semana, por ejemplo, se establecen los horarios y lugares cumbre de los usuarios tomando en cuenta su ubicación, la hora y el día (Gutiérrez, 2018:201).

Tabla 1: Uso del espacio en la Ciudad De México

		Frecuencia de usuarios activos en Twitter							
Uso de suelo	Horario	Lunes	Martes	Miércoles	Jueves	Viernes	Sábado	Domingo	**Total**
Residencial	Mañana	f_{11}	f_{12}	⋯	⋯	⋯	⋯	⋯	f_{1+}
	Tarde	f_{21}	f_{22}	⋯	⋯	⋯	⋯	⋯	⋮
	Noche	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Comercial	Mañana	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	Tarde	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	Noche	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Servicios	Mañana	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	Tarde	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	Noche	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Especiales	Mañana	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	Tarde	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	Noche	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Total		f_{+1}	f_{+2}	⋯	⋯	⋯	⋯	⋯	N

1.4.2. Técnicas de minería de datos para modelos predictivos

a. Clasificación

Del conjunto de métodos en minería de datos, por su potencial y aplicación en distintas áreas del conocimiento, la clasificación es una de las técnicas más utilizadas, su objetivo consiste en clasificar o mapear un conglomerado de datos a partir de un conjunto de clases definidas con anterioridad. Cada registro permite la asignación de una clase o etiqueta, es decir, a cada elemento de una base de datos corresponde una clase definida por los valores de sus atributos. La finalidad última de esta tarea es la de predecir, con el mayor grado de precisión, la clase de un objeto (Jaramillo, 2016:5-6, Hernández, Ramírez y Ferri, 2004:26).

Por ejemplo.

El departamento de inteligencia y logística del escuadrón de emergencia de una ciudad, desea clasificar, de acuerdo a su orden, las calles y avenidas de la urbe. Los datos se obtendrán de los sensores colocados en las aristas de los cuadrantes de las manzanas que diariamente registran en un intervalo de cinco minutos, datos referentes a la frecuencia, velocidad y volumen de los vehículos motorizados que circulan por las calles. La clasificación de la circulación en “rápida”, “moderada” o “lenta” tomando en cuenta la hora y el día de la semana, permitirá adecuar un algoritmo que indique la ruta más óptima entre la estación de emergencia y el lugar del siniestro.

b. Regresión.

La regresión tiene por objetivo obtener una función que permita calcular el valor real de una instancia faltante, basándose en el resto de los datos y considerando la relación existente entre las variables. A diferencia de la tarea anterior que predecía la etiqueta de clase de un objeto, la regresión permite predecir su valor numérico, esta tarea se basa en un registro histórico de la variación de los datos. La finalidad última de la regresión es la de reducir al máximo el error entre el valor proyectado y el valor verdadero (Jaramillo, 2016:5-6, Hernández, Ramírez y Ferri, 2004:26).

Por ejemplo.

Un grupo de investigadores de la Universidad Nacional de Educación a Distancia, desea implementar un modelo de investigación educativa cuyo objetivo consiste en pronosticar la probabilidad de éxito académico en los estudiantes de nuevo ingreso. Para llevar a cabo la tarea anterior se aplicará un modelo de regresión lineal múltiple, entre el conjunto de variables que los investigadores han definido se encuentran el sexo, promedio de bachillerato, escolaridad de los padres, ingreso, estado civil y ubicación geográfica. Al finalizar el algoritmo, se pretende en una primera etapa definir los factores que mayor influencia tienen en la deserción escolar y con ello implementar acciones para reducirla.

En una etapa siguiente, se pretende que, habiendo ajustado el algoritmo para cada perfil y facultad, se haga un seguimiento periódico que permita atender el rezago educativo de modo personalizado (Zaldivar, *et.al*., 2011).

1.4.3. Otras técnicas de minería de datos

De acuerdo a Beltrán (2003:25-26) existen otras tareas en minería de datos que frecuentemente son de gran utilidad en la exploración de los datos.

a. Condensación o descripción de conceptos.
El objetivo de esta tarea es la describir o resumir un conjunto de datos mediante la extracción de una submuestra.

b. Detección de deviaciones, casos extremos o anomalías.
El objetivo de esta tarea es la de identificar aquellos datos cuyo comportamiento se encuentra muy alejado de la media o de registros anteriores.

c. Modelado de dependencias.
El objetivo de esta tarea es la de construir un modelo matemático que sea capaz de identificar la existencia de dependencias entre variables. Se pueden identificar dos principales niveles; estructural

y numérico, el primero se asocia con variables discretas y a la identificación de patrones mediante el método gráfico, mientras que el segundo nivel se asocia con variables continuas y al uso de escalas numéricas para describir las dependencias.

1.5 VÍCULO DE LA MINERÍA DE DATOS CON LA GEOGRAFÍA

La relación entre la minería de datos y la ciencia geográfica puede encontrarse en un primer momento en los fundamentos de la geografía cuantitativa y, en su posterior a las aplicaciones y auge en los últimos años de la geografía automatizada.

1.5.1. Geografía cuantitativa

Entre 1950 y 1970, sucesos como la segunda guerra mundial y el periodo de posguerra marcaron el desarrollo científico y tecnológico de la época; influenciado por las ciencias matemáticas, el paradigma cuantitativo imperó en las tendencias conceptuales y metodológicas que se discutían en los entornos académicos (Buzai, 2005:6-7). La revolución cuantitativa sostenía que la ciencia debía ocuparse en la formulación de hipótesis, que al ser confrontadas con la realidad transitarían hacía el establecimiento de teorías o leyes, la realización del objetivo marcado podría apoyarse de distintos métodos entre los que destacaban los métodos cuantitativos (Ramírez y Claret, 2015 en Buzai, *et al*., 2015:105).

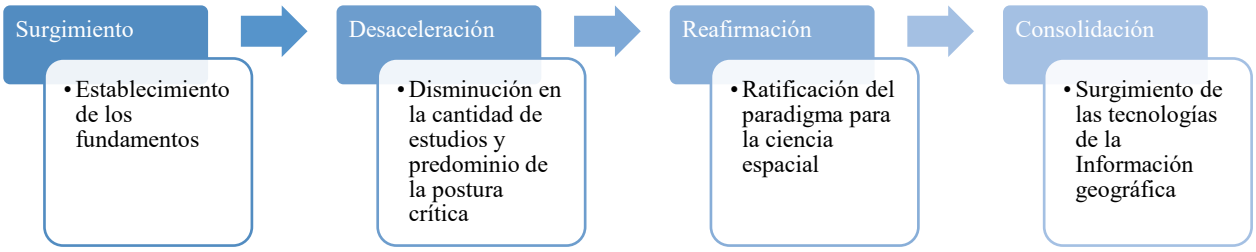
Apoyados en conceptualizaciones matemáticas, los métodos cuantitativos figuraron como un mecanismo obligatorio en la formulación de modelos cuyo fin se situaba en la "construcción y transmisión de conocimiento" (Buzai, 2005:7), así también como en mantener una característica apreciada en la discusión científica: la objetividad, cuyos atributos resaltaban la importancia de obtener resultados similares después de repetir un mismo procedimiento *n* número de veces. De esta manera, como señala Tulla (1993:168) los métodos cuantitativos podrían aplicarse en dos niveles, el primero referido a su utilización como instrumento de análisis y el segundo, a la postulación de leyes o modelos conceptuales a partir de ellos.

Influenciada por el paradigma de la época, la geografía cuantitativa se ocupó del establecimiento de principios, modelos o leyes para explicar los patrones de distribución espacial (Figura 5). El análisis del espacio geográfico a partir de la apreciación empírica y la medición de la realidad para la identificación de patrones de distribución, permitía a través de la generalización, formular tratados o leyes a fin de dar respuesta a dos grandes interrogantes; el porqué del orden territorial y el cómo de su gestión y planificación (Buzai, *et al*, 2015:25). El territorio, punto de partida para el análisis de la realidad, podía comprenderse mediante métodos cuantitativos a través del planteamiento de hipótesis

y métodos estadísticos o bien, para contrastar la teoría con la realidad a fin de validar las afirmaciones teóricas (Baxendale, 2015:46).

En términos generales, el paradigma cuantitativo en geografía apuntaló por expresar en modelos o leyes los acontecimientos geográficos de una manera mensurable y lo más exacta posible a fin de transcender el quehacer descriptivo que por muchos años había acompañado a la disciplina (Ramírez y Claret, 2015:105).

Figura 5: Evolución de la geografía cuantitativa



Elaboración propia con base en Buzai y García de León, 2015:31-32.

Para algunos autores como Buzai, la Teoría de los Sistemas Complejos (TGS), permitía desde una perspectiva teórica, comprender la realidad, el objetivo de esta se situaba en hallar afirmaciones que pudieran generalizarse a distintos sistemas en diferentes escalas, por su parte la Teoría de los Sistemas Complejos (TSC) buscaba establecer generalizaciones aplicadas a cada uno de los niveles que conformarían los sistemas. La TSC permitía aproximarse a la realidad mediante el análisis de la estructura de los sistemas definidos a partir de los insumos extraídos de esta: los datos; la interpretación a partir de ellos: los observables, y la interconexión entre ambos: los hechos. La realidad, objeto de interés geográfico, se comprendía como una totalidad estratificada por niveles de organización en donde acontecían dinámicas específicas. Desde este enfoque, cada paradigma de la geografía se interesaría por un nivel de la realidad, en el caso particular de la geografía cuantitativa, este se abocó al nivel espacial (Buzai *et al*, 2015:31).

El análisis de un sistema complejo como lo es la realidad, demandaría de un conglomerado diverso y robusto de enfoques teóricos que por sí mismos, serían capaces de explicar una parte o nivel de esta, pero ninguno de ellos conveniente para explicar la totalidad, un planteamiento teórico o método en un nivel de la realidad podría presentarse como adecuado, pero poco apropiado en otro. Cada paradigma de la geográfica explicaría una parte de la realidad, la complementariedad entre diferentes enfoques permitiría una explicación más fina de ella (Baxendale, 2015:45-46).

Los estudios realizados desde el paradigma cuantitativo se interesaron principalmente por la organización, la gestión y la planificación del territorio, sin embargo, hablar en términos de lo cuantitativo en cuanto al método se refiere, no implicó a priori una discusión teórica con base en las ciencias exactas. De acuerdo a Buzai y García de León (2015:48), la geografía cuantitativa contempló en su análisis toda manifestación espacial producto de las relaciones sociales y las desigualdades territoriales que de ellas se desprendían, cuantificar su comportamiento permitió proponer mecanismos de acción para reducirlas y contribuir al bienestar de la población.

En esta discusión, Buzai (2005:10) anota que la geografía cuantitativa no desarrolló su quehacer en torno y en oposición al sistema político, más bien trabajó al margen de cualquiera que este fuera, con el único fin de contribuir mediante el estudio, análisis y comprensión de la realidad a mejorar las condiciones de vida de la sociedad: "Las posibilidades de llevar a la práctica las soluciones encontradas generalmente no están en mano del científico o profesional geógrafo que realiza el estudio, sino en el nivel político, en donde deberían ser utilizados los informes generados a fin de tomar decisiones incorporando bases de fundamentación espacial".

De acuerdo a Montes (2015:62), la estructura metodológica que debería seguir una investigación de corte cuantitativo se compone de cuatro fases:

1. Descriptiva. ¿Cómo es el objeto geográfico?

Procesamiento de la información. Apoyándose de la estadística descriptiva, el objetivo de esta etapa es conocer la estructura de los datos.

2. Explicativa. ¿En dónde está y por qué se ubica ahí?

Demostración teórica. Apoyándose de la estadística inferencial, el objetivo de esta etapa es explicar el comportamiento y estructura de los datos de la etapa anterior a partir de un marco teórico. Con base en los patrones de repetición observados, se procede a la propuesta de modelos teóricos explicativos y a la simulación de posibles escenarios.

3. Contrastiva.

Validación teórica. Apoyándose de la estadística inferencial o la constatación en campo, el objetivo de esta etapa es contrastar los modelos propuestos en la etapa anterior con la realidad observada

4. Aplicativa

Diseño, ejecución y evaluación del proyecto de intervención del objeto geográfico.

El uso de técnicas asociadas al método cuantitativo para el análisis espacial, más que un desarrollo tecnológico, representó una manera distinta de aprehender, pensar y representar el espacio geográfico (Buzai y García de León, 2015:30). Éstas técnicas, no se consideraron como el fin último de la geografía cuantitativa, sino un medio para analizar el entorno geográfico, pues como anota Tulla (1993:177); "las formulaciones teóricas utilizan técnicas y métodos cuantitativos y no viceversa". Habiendo definido el lugar que ocupaba y los objetivos del conocimiento numérico en geografía, este se posicionó como fundamental en dos sentidos, por una parte, el de auxiliar la toma de decisiones en cuanto a la elección del método o conjunto de técnicas, y por el otro, el de brindar un fundamento para justificar o desechar el establecimiento de parámetros necesarios para ciertos cálculos teniendo presentes los efectos que estos pudiesen tener en los resultados (Moreno, 2015:24).

1.5.2. Geografía automatizada

La aproximación al conocimiento transita por una serie de etapas que parten de la observación, la identificación, registro y medición de los objetos geográficos, el mapeo de los patrones de comportamiento y la verificación de la información; en cada una ellas se vincula la necesidad apremiante de contar con un conjunto de herramientas e instrumentos referentes al terreno de la observación y la medición. En este escenario, las tecnologías de la información geográfica se situaron como un conjunto de técnicas capaces de aprehender y registrar objetos de la realidad espacial potencializando así su desarrollo y presencia en los estudios geográficos (Montes, 2015:57).

La relativa facilidad para producir datos provocó en algunas ocasiones que los métodos tradicionales de análisis fueran insuficientes para analizar grandes volúmenes de información, Buzai citando a Stotman expone un ejemplo por demás apropiado al anotar que después de finalizar su primer recorrido, el satélite artificial LANDSAT 1, en su búsqueda por explorar el planeta Tierra, produjo la misma cantidad de información que hasta el siglo XV habían reunido los geógrafos y para su segundo recorrido lo equivalente a lo conocido hasta el siglo XIX. Si bien esta oleada de datos pudo representar un cuestionamiento a las capacidades técnicas tradicionales, no lo fue para las capacidades de raciocinio humano que ante un nuevo escenario se verían apremiadas a explorar, mediante la creatividad y el ingenio, nuevas formas para analizar la información (Buzai, 2015:6). Las fórmulas que hasta antes de la década de los años cincuenta del siglo XX se había desarrollado de manera analógica transitaron hacía un formato digital para la incorporación, manejo y análisis de la información mediante nuevos procedimientos apoyados en los desarrollos tecnológicos.

Con base en la propuesta paradigmática que hace Jim Gray sobre la evolución de la ciencia, desde un enfoque cuantitativo, Gutiérrez (2018:202) identifica cuatro paradigmas en el desarrollo de la geografía:

1. Empírico. Se asocia a los estudios descriptivos comprendidos entre la antigüedad y la geografía regional francesa.
2. Teórico. Corresponde el establecimiento de modelos y generalizaciones hechas desde la geografía teorética y cuantitativa.
3. Computacional. Comprende el análisis de los sistemas espaciales complejos a partir del desarrollo de los Sistemas de Información Geográfica.
4. Ciencia de la exploración de datos. Extracción de conocimiento de datos geográficos masivos producidos de manera continua (Minería de datos).

Por otro lado, Buzai (2005:5) identifica tres principales enfoques en los estudios geográficos que se han consolidado en los últimos años. El primero está vinculado a la ecología del paisaje, el segundo a la geografía posmoderna y el tercero, de intereses para el tema que nos atañe, a la geografía automatizada, cuyo sustento se encuentra en el desarrollo de las "geo- tecnologías" asociadas a los avances tecnológicos en el quehacer geográfico.

Influenciada por el desarrollo de las tecnologías informáticas, la geografía incorporaría de éstas, conceptos y métodos que marcarían la gestación de un nuevo enfoque disciplinario. El desarrollo de las tecnologías de la información geográfica en la década de los 70s del siglo pasado, sentó las bases tecnológicas de la automatización en las tareas geográficas por medio de mecanismos computacionales, es a principios de la década de 1980 cuando surge un debate en torno a la viabilidad del uso de tecnologías computacionales como mecanismo para automatizar los métodos tradicionales en la ciencia geográfica. La incorporación de las tecnologías de la información geográfica en los últimos años ha facilitado la obtención, tratamiento y análisis de datos territoriales, sin bien no como ley universal, en la mayoría de las ocasiones las innovaciones tecnológicas han posibilitado la obtención de datos geográficos de manera económica, accesible y en un lapso menor de tiempo (Buzai, 2005:6-8, Bosque, 2015:168, Oropeza y Díaz, 2007:72).

La geografía automatizada retoma en su discusión afirmaciones del paradigma cuantitativo sin representar una fiel reproducción de este, Buzai (2005:7) sugiere la presencia de un nuevo enfoque basado en las Tecnologías de la Información Geográfica (TIG) que vislumbra una manera distinta de mirar y modelar la realidad. La incorporación de las TIG al quehacer geográfico, no representa únicamente una nueva manera de hacer, sino también del ser en geografía, representando una forma distinta y única de aprehender, observar y analizar el mundo.

El surgimiento de la geografía automatizada permite incorporar al análisis espacial procedimientos apoyados en tecnologías digitales, nombradas también geotecnologías, a fin de favorecer la

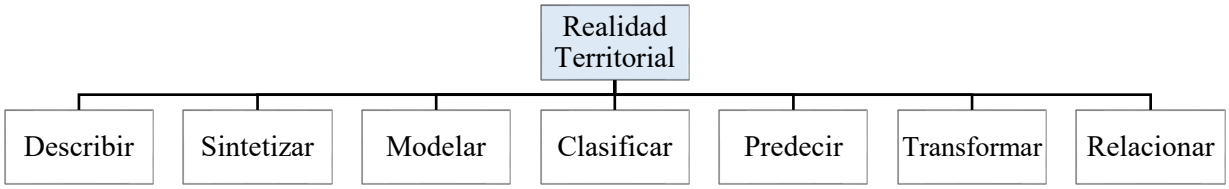
automatización de los procesos en la obtención, procesamiento, análisis y presentación de datos geográficos (Buzai y Ruíz, 2012:96). Su desarrollo fue posible gracias al uso del computador, las tecnologías de la información, las comunicaciones y fundamentalmente el uso del intelecto humano necesario para coordinar y potenciar los embalajes del sistema tecnológico (Buzai y García de León, 2015:45).

De acuerdo a Buzai (2001:29), es en la década de 1980 cuando se formaliza el debate en torno a los posibles aportes que traería el uso de las tecnologías de la información geográfica en la automatización de los procedimientos. Formuladas en disciplinas como la computación y las telecomunicaciones, el mismo autor (2005) conceptualiza las TIG como un recurso electrónico especializado que permite sistematizar un conjunto de métodos de análisis en distintas escalas y periodos de tiempo. Propiamente las TIG permiten entre otras cualidades:

1. Recolectar y medir características del territorio
2. Producir un formato digital de la realidad
3. Procesar y analizar computacionalmente datos geográficos
4. Analizar patrones espaciales

El conjunto de herramientas que conforman las TIG como programas cartográficos, de imágenes satelitales y estadísticos (ver figura 6):

Figura 6: Alcances de las TIG



Elaboración propia con base en Moreno (2015)

El uso de medios computacionales resulta ser una alternativa viable cuando se trata de grandes volúmenes de datos, cuando nos referimos a conjuntos de datos que concentran en su interior diferentes aspectos de índole geográfico se asume, por ejemplo, que cualquier objeto puede proyectarse digitalmente en términos geográficos como un punto, línea o polígono, localizarse a través de un par de coordenadas en el espacio, ubicarse en el tiempo y atribuírsele un conjunto de características o variables concernientes a su aspecto espacial (Buzai y Ruíz, 2012:96).

El uso de herramientas digitales permite proyectar o recrear una fracción de la totalidad del territorio en un plano digital siendo esta, una construcción paralela del espacio real, el objeto toma de la realidad

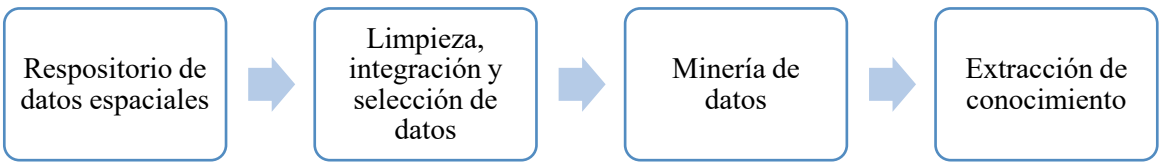
los aspectos más significativos o de interés para el investigador o agente de intervención con el fin de apoyar fundamentalmente el proceso de análisis, asimismo es de gran utilidad cuando se desea modelar una situación o experimentar los efectos de un suceso que en términos reales sería imposible realizar (Moreno, 2015: 21). Dicho objeto se presenta como un mecanismo moldeable y flexible cuando de aspectos territoriales se trata y deja ver dos aspectos importantes en el uso de las TIG para el análisis geográfico, por un lado, el uso en sí mismo de nuevas aplicaciones metodológicas y por el otro, las nuevas formas de pensar y proyectar la realidad en un plano digital (Buzai, 2015:5).

Frente a una realidad dinámica, global y en constante cambio, se dispone de un conjunto de recursos computacionales e informáticos que deben aprovecharse para extraer de ella una gran cantidad de datos que permitan darle un sentido a la estructura, dinámica y evolución de la realidad (Moreno, 2015:20).

1.5.3 Geografía automatizada y minería de datos

La relación entre la minería de datos y la ciencia geográfica puede encontrarse en los fundamentos de la geografía automatizada. En nuestro contexto tecnológico es común que las actividades humanas dejen un rastro digital con una referencia espacial y temporal. Los datos que se producen de manera masiva diariamente están relacionados, por ejemplo, con el registro automático de los teléfonos móviles, la actividad de tarjetas bancarias, los datos generados en tiempo real por medio de un GPS, el uso de redes sociales como Twitter, Facebook o Google, estadísticas sobre el uso de servicios públicos, imágenes registradas por cámaras de vigilancia, registros de los peajes en carreteras, etc. (Gutiérrez, García y Salas, 2016:6). Estos datos refieren a una diversidad de procesos, por ejemplo; la contaminación atmosférica puede analizarse a través de los registros que generan diariamente los sensores remotos de una estación climatológica o bien, la movilidad urbana y el uso del espacio puede comprenderse a partir del uso de aparatos móviles durante el transcurso del día, o también, el consumo en un centro de población podría analizarse a través del uso de tarjetas bancarias en zonas previamente segmentadas y en determinadas épocas del año (Gutiérrez, 2018:196).

Figura 7: Estructura de la minería de datos espaciales



Elaboración propia con base en Dueñas 2009:146

Muchos de los datos que se generan diariamente de manera automática por medio de sensores o dispositivos están georreferenciados y hacen alusión a personas y cosas (Bosque, 2015:168). De acuerdo con estimaciones de Beltrán (2003:22), el 80% de los datos que conforman una base datos, tiene una referencia espacial ya sea, por ejemplo, por medio de una dirección, un código postal o coordenadas. Por sus características, estos datos son susceptibles a cartografiarse y se caracterizan por tener un atributo que hace referencia a su geometría sea esta como un punto, línea o polígono y, a un atributo descriptivo (Gutiérrez, 2018:200; Dueñas 2009:149). Debido a que la gran mayoría de los datos tiene una referencia geográfica, estos pueden ser tratados mediante el uso de las tecnologías de la información geográfica, los beneficios que su uso representa es que a partir de la estructura de un conjunto de datos se pueden analizar las expresiones espaciales como un proceso y no como una imagen estática (Gutiérrez, 2018:197; Bosque, 2015:170).

Los objetivos de la minería de datos y la geografía automatizada se dirigen hacía la extracción de conocimiento de un gran número de datos mediante el análisis de los cambios, patrones, asociaciones, estructuras y procesos obtenidos a partir de un conjunto masivo de datos los cuales son analizados dentro de su contexto ambiental, socioeconómico y político y su análisis se apoya en la automatización de los procedimientos (Gutiérrez, 2018:200).

En un contexto en el cual, por un lado, las innovaciones tecnológicas se desarrollan a paso acelerado y por otro, en que la dimensión espacial se encuentra en un proceso de revaloración dentro de las ciencias sociales, la automatización de los procedimientos para analizar la realidad a partir de la información generada continuamente, representa una oportunidad para la geografía de posicionarse como especialista en el tratamiento e interpretación de información geoespacial y, si bien, la minería de datos se ha presentado atractiva principalmente para el sector empresarial o de los negocios, en los últimos años, el sector académico, ha encontrado en la minería de datos, una gama interesante de posibilidades para abordar, de manera innovadora y creativa, una problemática, o bien, aprehender y mirar desde otras perspectivas la realidad (Gutiérrez, García y Salas, 2016:2, Buzai, 2005:18).

La minería de datos ofrece al investigador o al responsable de la toma de decisiones la posibilidad de (Gutiérrez, García y Salas, 2016:17, Gutiérrez, 2018:200):

1. Abordar problemas de investigación a partir de otros enfoques
2. Descubrir procesos que mediante técnicas tradicionales no son de fácil apreciación
3. Aproximarse a procesos que con datos oficiales no eran de fácil acceso
4. Realizar estudios comparativos considerando la escala global de los datos

5. Disponer de información actualizada tomando en cuenta que la mayor parte de la información se genera de manera constante
6. Monitorear procesos de manera cotidiana
7. Trabajar con datos de “alta resolución espacial y temporal”
8. Complementar o validar la información proveniente de fuentes oficiales
9. Abordar problemáticas haciendo uso de la creatividad y la innovación

Analizar una gran cantidad de datos apoyándose de distintas técnicas en minería de datos va más allá del mero conocimiento técnico, el analista requiere partir de un marco teórico que le permita definir el método más adecuado, sustentar la lógica de operación y fundamentalmente, dirigir el análisis y dar sentido a los resultados que deriven del proceso (Gutiérrez, 2018:10). El tratamiento de grandes bancos de datos permite extraer de estos, información que, en posterior, podrá aplicarse en diversas áreas de la vida cotidiana; más allá del sector empresarial o de negocios, el adecuado manejo de un volumen grande de datos, representa una oportunidad para incidir de manera positiva en nuestro entorno pues los datos se presentan como un insumo valioso en la toma informada de decisiones y en la dirección de acciones.

CAPITULO 2: UN MÉTODO EN MINERÍA DE DATOS: ANÁLISIS DE CORRESPONDENCIA SIMPLE (ACS)

Comúnmente asociamos los métodos numéricos con variables continuas, menos populares son los métodos cuyo insumo se encuentra en las variables discretas, es por ello que se ha seleccionado el análisis de correspondencia simple para ejemplificar la aplicación de una técnica de minería de datos en geografía.

2.1 EL ANÁLISIS DE CORRESPONDENCIAS SIMPLE

Esta técnica fue desarrollada en la tradición francesa por el estadístico Jean Paul Benzécri (1963). El Análisis de Correspondencias (AC) se basa en el álgebra lineal, es un método de análisis multivariante descriptivo o exploratorio. Analiza la estructura de correlación o asociación de dependencia e independencia entre las filas y las columnas de un conjunto de variables y sus respectivas dimensiones (Pineda, 2003).

El AC determina la relación entre las categorías de dos o más variables en una representación dual gráfica mejor conocida como mapa de percepción o perceptual. En un mismo espacio vectorial sitúa la posición de los objetos o unidades bajo estudio (individuos, grupos sociales, demarcaciones geográficas, sectores, etc.) y las categorías de las variables de interés (características, cualidades, valoraciones, etc.) (Capula, 2004:52, Castillo, 2008:15).

El método resume una gran cantidad de información en dos o tres dimensiones de un conjunto de datos contenidos en una tabla de contingencia en donde las filas correspondes a los casos u objetos y las columnas a las variables. El tratamiento matemático de los datos permite proyectar en un espacio reducido, los puntos filas y los puntos columna de manera simultánea. En comparación con otros métodos de análisis, el AC además de comprender las relaciones de dependencia, permite analizar la estructura de asociación de las variables bajo estudio (Rodríguez y Mora, 2001:43, De la Fuente, 2011:1).

El AC parte de una matriz de datos rectangulares, esto es, una tabla de doble entrada o tabulación cruzada que relaciona dos o más variables en una misma tabla. Para este tipo de análisis generalmente se dispone de una tabla de contingencia con entradas no negativas (Castillo, 2008: 15). Las filas y las columnas de la matriz representan las categorías de cada una de las variables y las entradas corresponden a los conteos de la presencia o ausencia del atributo o característica (De la Fuente, 2011:1).

El AC se divide en análisis de correspondencias simple (ACS) y análisis de correspondencias múltiple (ACM) la diferencia entre ambos está marcada por el número de variables. El ACS analiza la relación entre dos variables y trabaja únicamente con tablas de contingencias de dos dimensiones o presentadas por pares. Por otra parte, el ACM puede incorporar al análisis más de dos variables (Chiapa, 2015:8, López y Facheli, 2015:73). Para el presente análisis nos enfocaremos únicamente al primero.

El AC es un caso particular del grupo de técnicas factoriales, la diferencia con otros métodos es el tipo de variable involucrada en el análisis, el AC analiza la información proveniente de variables discretas, mientras otras técnicas, como el análisis de componentes principales, lo hace con variables continuas. En términos generales, las técnicas de análisis factorial resumen, con base en la asociación o similitud, una gran cantidad de información perteneciente a una matriz de datos en un número reducido de dimensiones o factores que recuperan la totalidad de la varianza (Rodríguez y Mora, 2001:43, López y Facheli, 2015:73).

Antes de aplicar el método es fundamental corroborar que los datos se ajusten al modelo, dado las características del método el ajuste debe indicar que existe relación entre las variables. Para ello, y considerando el tipo de variable, se aplica el estadístico Chi- cuadrada a la tabla de contingencia, en esta prueba se contrasta la hipótesis nula (H_0) que afirma que existe interdependencia entre las variables, si el p-valor es menor que el nivel de significancia sugerido por el investigador se rechaza la hipótesis de independencia. Si la prueba indica que no hay relación entre las variables, el AC no procede (Rodríguez y Mora, 2001:44).

En términos generales, el análisis de correspondencias establece la asociación o proximidad de las categorías mediante la comparación de los perfiles definidos como las proporciones relativas o absolutas de las frecuencias. Cada valor de las celdas es un punto dotado de masa que asocia la relación entre los perfiles y los perfiles promedio, la comparación entre estos se puede hacer a partir de una medida de distancia o métrica (Chi-cuadrada, X^2). La reducción de la información en n dimensiones permite re- proyectar lo datos en un mismo espacio conservando la mayor variabilidad posible, a cada dimensión o eje le corresponde un valor propio el cual permite conocer la contribución relativa de cada categoría (Rodríguez y Mora, 2001:43, López y Facheli, 2015:74).

La interpretación de los resultados se auxilia de la representación gráfica en donde se proyectan los puntos fila y columna de una tabla cruzada en un espacio euclidiano generalmente de dos dimensiones. En el gráfico se representan todas las categorías de las variables como una nube de puntos, las coordenadas de cada uno de ellos se asocian a las similitudes entre categorías. De la representación cabe destacar el análisis de la concentración o dispersión de puntos, la lejanía o

cercanía entre ellos en un indicador de las relaciones de similitud o dependencia, por otro lado, una mayor distancia hacía el origen sugiere la concentración de la frecuencia en una celda especifica. Es preciso señalar que la representación gráfica es un componente del análisis, sin embargo, no representa su totalidad ya que este complementa los estadísticos de los que se apoya el AC (Rodríguez y Mora, 2001, 43-47).

Un desarrollo adecuado de AC se compone de al menos cinco etapas (Aldás, 2000: 5-11):

1. Planteamiento de problema. Se establece el objetivo y se define el conjunto de variables.
2. Plan de análisis Se diseña el instrumento para la obtención de datos que posteriormente serán resumidos en una tabla de doble entrada, la naturaleza del método sugiere medir la presencia o ausencia de la característica y registrar la frecuencia en cada celda. Las categorías deben de ser exhaustivas y excluyentes entre sí.
3. Aplicabilidad de la técnica. Antes de proceder a la aplicación de AC se verifica la presencia de asociación entre las variables a través de la prueba estadística Chi-cuadrado (X^2).
4. Desarrollo del método. Se calcula la masa, la inercia y se define el número de factores o dimensiones representativos de los datos originales.
5. Interpretación de resultados. Se analizan los estadísticos que resultan del método y se procede a realizar el análisis gráfico.

El desarrollo del método para fines ilustrativos aplicado en una tabla de contingencias de pocas casillas podría parecer poco relevante y las asociaciones entre las variables podrían apreciase mediante la simple observación. Sin embargo, la aplicación del AC es particularmente útil cuando el conjunto de variables y categorías es numeroso, si bien los estadísticos de prueba en el análisis de tablas de contingencia permiten contrastar la hipótesis de independencia y medir el grado de asociación, lo resultados no permiten conocer la estructura de dependencia (López y Facheli, 2015:75).

El método de AC es de gran utilidad cuando disponemos de un conjunto de datos no métricos. De acuerdo a Capula (2004) y Chiapa (2015), la técnica permite:

- Representar las filas y las columnas, o bien, las categorías de una y otra variable en un mismo espacio
- Analizar la estructura de asociación de las variables y posteriormente la identificación de grupos
- Corroborar la dependencia de análisis comprobada en análisis anteriores
- Generar hipótesis en las etapas iniciales del diagnóstico

- Generar hipótesis que puedan referir investigaciones posteriores

De acuerdo a De la Fuente (2011:1) las aplicaciones del AC pueden ser múltiples y aplicarse a distintas áreas del conocimiento que van desde las ciencias matemáticas, biológicas, sociales y hasta las humanidades. En bien conocida la aplicación de esta técnica en el campo del mercadeo para el estudio de la preferencia de consumo o el posicionamiento de una empresa, sin embargo, también es útil en la identificación de patrones asociados a una epidemia, un rasgo psicológico, el comportamiento de un organismo vivo en biología y en el análisis de textos a partir de la repetición de una palabra y su relación con algún género literario. Así también, se puede aplicar está técnica para asociar grupos entre objetos sean personas o entidades geográficas considerando un grupo de variables.

2.1.1 Objetivos

El principal objetivo del análisis es descriptivo, como parte del grupo de técnicas factoriales entre sus objetivos está es el de resumir una gran cantidad de información en un número menor de dimensiones evitando la menor pérdida de información y conservando la mayor variabilidad posible. Concretamente el análisis de correspondencias tiene por objetivo analizar, a partir de una tabla de doble entrada, la estructura de la relación de semejanza o diferencia de las categorías de las variables involucradas (Capula, 2004:53).

Su objetivo final es el de resumir una masa de puntos, que posterior a un tratamiento matemático, puedan ser proyectados geométricamente en un espacio de menor dimensión de modo que la representación permita la localización, visualización e interpretación de la estructura de una gran matriz de datos. La representación gráfica conserva en la mayor medida posible la asociación de los puntos fila y los puntos columna y puede leerse como un mecanismo capaz de mostrar la estructura de una tabla de doble entrada inadvertida a simple vista (Saldaña, 2005, Saavedra, 2012).

2.1.2 Diseño

De acuerdo a Morales (2004:18) el diseño de la investigación del análisis de correspondencias es descriptivo e interdependiente. Descriptivo porque el objetivo es describir un evento mediante la comprensión de un conjunto de datos que permita obtener una visión general del fenómeno, generar hipótesis y detectar patrones. Interdependiente porque busca describir las interrelaciones entre las variables y su estructura. Su diseño, aunque cabalmente descriptivo, puede ser el punto de partida de un estudio inferencial (Saavedra, 2012:11).

2.1.3 Supuestos

El método no es rígido respecto a algún supuesto básico, sin embargo, es importante tener en cuenta los siguientes aspectos (Morales, 2004:19):

1. El método puede incluir dos o más variables en su análisis
2. Las variables de interés pueden medirse en cualquier escala, no obstante, el método se aplica a variables discretas medidas en escala nominal u ordinal.
   - Variables discretas: Agrupa a los objetos de acuerdo al número de categorías o eventos atribuidos a la característica a evaluar. Escalas de medida: nominal u ordinal.
   - Variables continuas: Los valores del atributo pueden tomar un valor cualquiera en una escala numérica continua. Escalas de medida: Intervalo o de razón.

   En caso de que las variables hayan sido medidas en escala de intervalo o razón éstas pueden analizarse mediante el AC siempre y cuando los datos sean recodificados, sin embargo, éste procedimiento asume una pérdida de información.
3. Los datos pueden representar relaciones lineales y no lineales.
4. El fundamental la presencia de asociación entre las variables.

2.2 ESTRUCTURA

En ACS se conforma por la matriz de doble entrada, tabla de frecuencias relativas, tabla de perfiles, Prueba Chi- cuadrada, espacio de perfil, masa e inercia total

2.2.1 Matriz o tabla de doble entrada

Se parte de una matriz o tabla de doble entrada de $I \times J$ casillas. El AC representa la asociación entre las categorías por medio un número reducido de dimensiones. El número de dimensiones es igual al $\text{mín}(I, J) - 1$. (López y Fachelli, 2015:76)

El análisis permite representar en un espacio de menor dimensión las categorías de las filas y las columnas. Sea una tabla $N(I, J)$, constituida por I filas y por J columnas representa como:

Tabla 3: Frecuencias absolutas $N(I, J)$

Y^X	Y_1	Y_2	$\cdots$	j	Total
X_1	n_{11}	n_{12}	$\cdots$	n_{1j}	n_{1+}
X_2	n_{21}	n_{22}	$\cdots$	n_{2j}	n_{2+}
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
i	n_{i1}	n_{i2}	$\cdots$	n_{ij}	n_{i+}
Total	n_{+1}	n_{+2}	$\cdots$	n_{+j}	n_{++}

Dónde:

X_1, X_2 = Niveles o categorías de la variable X

Y_1, Y_2 = Niveles o categorías de la variable Y

n_{ij} = Distribución de las frecuencias absolutas de cada casilla que se encuentran en renglón i-ésimo y la columna j-ésima

n_{i+} = Totales marginales fila. Total de elementos del primer y segundo renglón

n_{+j} = Totales marginales columna. Total de elementos de la primera y segunda columna

N= Total de casos de la muestra

2.2.2 Tabla de frecuencias relativas

La tabla de frecuencias relativas f_{ij} expresa las proporciones de las celdas respecto al total de casos:

Tabla 3: Frecuencias relativas $f_{(I,J)}$

Y^X	**1**	**2**	⋯	j	**Total**
1	f_{11}	f_{12}	⋯	f_{1j}	f_{1+}
2	f_{21}	f_{22}	⋯	f_{2j}	f_{2+}
⋮	⋮	⋮		⋮	⋮
i	f_{i1}	f_{i2}	⋯	f_{ij}	f_{i+}
Total	f_{+1}	f_{+2}	⋯	f_{+j}	N

La frecuencia relativa f_{ij} se determina por:

$$f_{ij} = \frac{n_{ij}}{N}$$

cumpliéndose que:

$$\Sigma_{i=1}^{I} + \Sigma_{j=1}^{J}\ f_{ij} = 1$$

El total marginal o la frecuencia relativa marginal de fila expresa el peso o masa del punto-fila i, se determina por:

$f_{i+} = \frac{n_{i+}}{N}$ o bien; $f_{i+} = \Sigma_{j=1}^{J} f_{ij}$ cumpliéndose la aseveración de que $\Sigma_{i=1}^{I}\ f_{i+} = 1$.

El total marginal o la frecuencia relativa marginal de columna expresa el peso o masa del punto – columna j, se determina por:

$f_{+j} = \frac{n_{+j}}{N}$ o bien $f_{+j} = \Sigma_{i=1}^{I} f_{ij}$ cumpliéndose la aseveración de que $\Sigma_{j=1}^{J} f_{+j} = 1$

Los totales marginales de las filas y las columnas corresponden al perfil medio en relación a la posición de las categorías de las filas y las columnas.

2.2.3 Tabla de perfiles

Definen las frecuencias relativas de las categorías de las variables obtenidas mediante la distribución condicional por fila o columna.

I. Perfil de fila (f_{ij}^F)

$f_{ij}^F = \frac{n_{ij}}{n_{i+}}$ o bien; $f_{ij}^F = \frac{f_{ij}}{f_{i+}}$

cumpliéndose que $\Sigma_{j=1}^{J} f_{ij}^F = 1$ dando lugar a la matriz F^F

$$F^F = \begin{bmatrix} \frac{f_{11}}{f_{1+}} & \frac{f_{12}}{f_{1+}} & \cdots & \frac{f_{1j}}{f_{1+}} \\ \frac{f_{21}}{f_{2+}} & \frac{f_{22}}{f_{2+}} & \cdots & \frac{f_{2j}}{f_{2+}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{f_{I1}}{f_{I+}} & \frac{f_{I2}}{f_{I+}} & \cdots & \frac{f_{IJ}}{F_{I+}} \end{bmatrix}$$

II. Perfil de columna $\left(f_{ij}^C\right)$

$f_{ij}^C = \frac{n_{ij}}{n_{+j}}$ o bien; $f_{ij}^C = \frac{f_{ij}}{f_{+j}}$

cumpliéndose que $\Sigma_{i=1}^{I} f_{ij}^C$=1 dando lugar a la matriz F^C

$$F^C = \begin{bmatrix} \frac{f_{11}}{f_{+1}} & \frac{f_{12}}{f_{+2}} & \cdots & \frac{f_{1j}}{f_{+J}} \\ \frac{f_{21}}{f_{+1}} & \frac{f_{22}}{f_{+2}} & \cdots & \frac{f_{2j}}{f_{+J}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{f_{I1}}{f_{+1}} & \frac{f_{I2}}{f_{+2}} & \cdots & \frac{f_{IJ}}{F_{+J}} \end{bmatrix}$$

En la siguiente tabla se registraron las frecuencias de preferencia de un número de consumidores de cuatro marcas distintas de un producto cualquiera (A, B, C, D) y su segmento (1, 2, 3). Se desea analizar la relación entre las variables marca y segmento (Molina, 2007:3).

Tabla 4: Frecuencias absolutas entre el segmento y la marca

	Segmento			
Marca	1	2	3	**Total fila**
A	30	30	155	215
B	30	130	30	190
C	80	30	30	140
D	80	30	5	115
Total columna	220	220	220	660

Elaborado con datos del Capítulo 3. Análisis de correspondencias. Molina, 2007:3

Tabla 5: Frecuencias relativas entre el segmento y la marca

	Segmento			
Marca	1	2	3	**Total fila**
A	0.045	0.045	0.235	0.326
B	0.045	0.197	0.045	0.288
C	0.121	0.045	0.045	0.212
D	0.121	0.045	0.008	0.174
Total columna	0.333	0.333	0.333	1.000

Tabla 6: Perfiles renglón entre el segmento y la marca

	Segmento			
Marca	1	2	3	**Total fila**
A	0.140	0.140	0.721	1.000
B	0.158	0.684	0.158	1.000
C	0.571	0.214	0.214	1.000
D	0.696	0.261	0.043	1.000
Perfil promedio del renglón	0.333	0.333	0.333	1.000

Tabla 7: Perfiles columna entre el segmento y la marca

	Segmento			
Marca	1	2	3	**Perfil promedio columna**
A	0.136	0.136	0.705	0.326
B	0.136	0.591	0.136	0.288
C	0.364	0.136	0.136	0.212
D	0.364	0.136	0.023	0.174
Total columna	1.000	1.000	1.000	1.000

Al cociente de la división entre los perfiles y los perfiles promedio se le denomina tasa de contingencia. Por ejemplo, para el caso de los renglones, la tasa de consumo de la marca A del segmento 1 es de 0.419 $\left(\frac{0.140}{0.333}\right)$. El mismo resultado se obtiene para los perfiles columna, por ejemplo, la tasa de consumo del segmento 1 de la marca A es de 0.419 $\left(\frac{0.136}{0.326}\right)$. De esta manera, el uso indistinto del perfil por renglón o columna conducirá al mismo resultado (Chiapa, 2015:12).

2.2.4 Prueba Chi- cuadrada

Utilizaremos el test de la Chi-cuadrada de Pearson para contrastar la hipótesis de independencia con el fin de afirmar la existencia o inexistencia de relación entre las variables.

1. Establecer el juego de hipótesis

$H_0 =$ Ambas variables son independientes. El segmento de los individuos no está relacionado con la marca que consumen

$H_1 =$ Existe una relación de dependencia. El segmento de los individuos si está relacionado con la marca que consumen

2. Cálculo de los valores esperados (e_{ij})

La existencia o ausencia de asociación se comprueba mediante el contraste de las frecuencias observadas contra las frecuencias esperadas, teóricamente estas últimas toman el valor presente en caso de no existir asociación. Si al contrastar ambos valores no hay diferencia entre ellos se establece que las variables son independientes por lo tanto no existe asociación. En el caso contrario se debe probar que la diferencia en los valores son lo suficientemente significativas para afirmar que existe asociación entre las variables (López y Fachelli, 2015b:16). La fórmula general para el cálculo de los valores esperados es:

$$n_{ij}^{e} = \frac{n_{i+} \times n_{+j}}{n}$$

$$e_{11} = \frac{f_{1+} * f_{+1}}{f_{++}} = \frac{215 * 220}{660} = 71.67 \qquad e_{31} = \frac{f_{3+} * f_{+1}}{f_{++}} = \frac{140 * 200}{660} = 46.67$$

$$e_{12} = \frac{f_{1+} * f_{+2}}{f_{++}} = \frac{215 * 220}{660} = 71.67 \qquad e_{32} = \frac{f_{3+} * f_{+2}}{f_{++}} = \frac{140 * 220}{660} = 46.67$$

$$e_{13} = \frac{f_{1+} * f_{+3}}{f_{++}} = \frac{215 * 220}{660} = 71.67 \qquad e_{33} = \frac{f_{3+} * f_{+3}}{f_{++}} = \frac{140 * 220}{660} = 46.67$$

$$e_{21} = \frac{f_{2+} * f_{+1}}{f_{++}} = \frac{190 * 220}{660} = 63.33 \qquad e_{41} = \frac{f_{4+} * f_{+1}}{f_{++}} = \frac{115 * 220}{660} = 38.33$$

$$e_{22} = \frac{f_{2+} * f_{+2}}{f_{++}} = \frac{190 * 220}{660} = 63.33 \qquad e_{42} = \frac{f_{4+} * f_{+2}}{f_{++}} = \frac{115 * 220}{660} = 38.33$$

$$e_{23} = \frac{f_{2*} * f_{+3}}{f_{++}} = \frac{190 * 220}{660} = 63.33 \qquad e_{43} = \frac{f_{4+} * f_{+3}}{f_{++}} = \frac{115 * 220}{660} = 38.33$$

3. Cálculo de la Chi-cuadrada (X_C^2)

Las diferencias se evalúan a partir de una distancia media cuadrática (X^2):

$$X_c^2 = \sum_{r=1}^{r} \sum_{j=1}^{j} \frac{(f_{ij} - e_{ij})^2}{e_{ij}}$$

$$X_C^2 = \frac{(30-71.67)^2}{71.67} + \frac{(30-71.67)^2}{71.67} + \frac{(155-71.67)^2}{71.67} + \frac{(30-63.33)^2}{63.33} + \frac{(130-63.33)^2}{63.33} + \frac{(30-63.33)^2}{63.33} + \frac{(80-46.67)^2}{46.67} + \frac{(30-46.67)^2}{46.67} + \frac{(30-46.67)^2}{46.67} + \frac{(80-38.33)^2}{38.33} + \frac{(30-38.33)^2}{38.33} + \frac{(5-38.33)^2}{38.33}$$

$$X_c^2 = 24.22 + 24.22 + 96.90 + 17.54 + 70.18 + 17.54 + 23.81 + 5.95 + 5.95 + 45.29 + 1.81 + 28.99$$

$$X_c^2 = 362.41$$

El valor resultante mide la distancia media entre los valores.

4. Determinación de la probabilidad asociada al estadístico. Chi- cuadrada teórica (X_t^2)

El resultado se contrasta con un valor crítico o valor de tablas (distribución teórica de la (X^2), el valor se establece considerando el nivel de significación (α) que generalmente se establece en 0.05. Los grados de libertad se definen a partir de la multiplicación del número de renglones de la tabla de contingencia menos uno por el número de columnas menos uno. Mientras mayor sea el valor del estadístico X^2, la diferencia entre el valor esperado y el observado es mayor (Chiapa, 2015:17).

$X_{t\ (r-1)(c-1)gl}^{2\ \propto=0.05}$

$X_{t\ (4-1)(3-1)gl}^{2\ \propto=0.05}$

$X_{t\ (6\ gl)}^{\propto=0.05}$

$X_t^2 = 12.5916$

5. Estadística de prueba

$X_c^2 < X_t^2 \approx$ No se rechaza H_0

$X_c^2 > X_t^2 \approx$ Se rechaza H_0

6. Toma de decisión

362.413>12.5916 $\approx$ Se rechaza H_0

7. Coeficiente de contingencia

$$C_c = \sqrt{\frac{X_c^2}{X_t^2 + n}}$$

$$C_c = \sqrt{\frac{362.413}{12.5916 + 660}}$$

$$C_c = 0.7341$$

$$C_{max} = \sqrt{\frac{k-1}{k}}$$

$$C_{max} = \sqrt{\frac{4-1}{4}}$$

$$C_{max} = 0.8660$$

$$C_{corregida} = \frac{C_c}{C_{max}}$$

$$C_{corregida} = \frac{0.7341}{0.8660}$$

$$C_{corregida} = 0.848$$

En un rango de 0 a 1 en donde 0 es igual a la ausencia de asociación y 1 a la máxima asociación, el grado de correlación entre las variables de la tabla de continencia es de 0.848.

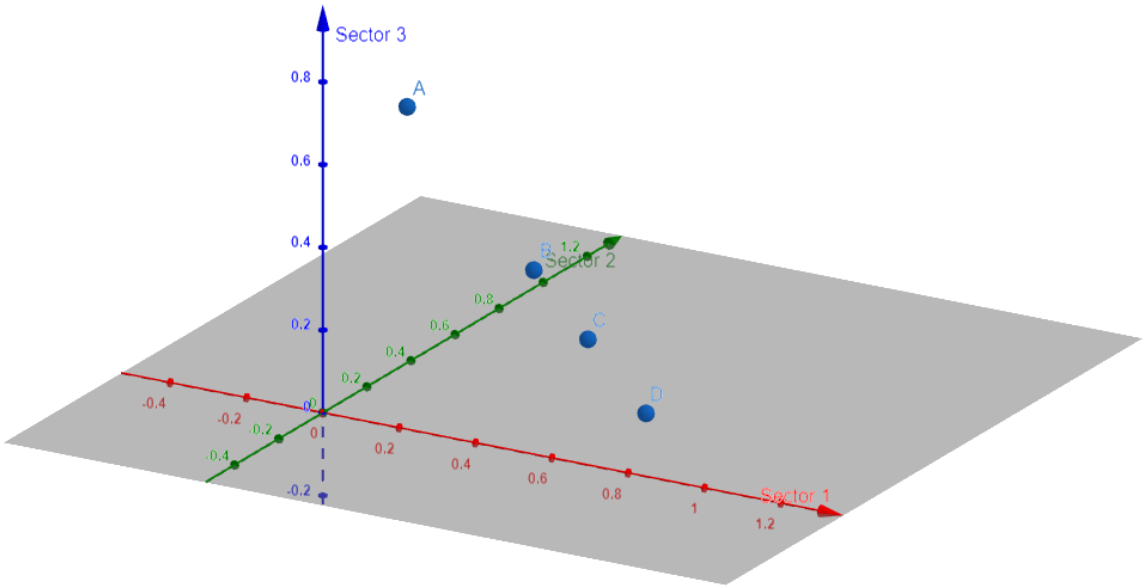
2.2.5 Espacio de perfil

Se entiende por espacio de perfil al plano de m dimensiones. El valor de m es equivalente a la dimensión del plano, por ejemplo, para una matriz de tres por tres, su representación es tridimensional ($m = 3$). En el espacio de perfil se localizan los perfiles conformados por m elementos y cuya suma es igual a la unidad, a su vez, los perfiles ocupan un sub espacio de $m - 1$ dimensiones (Chiapa, 2015:12).

La información de los perfiles de fila o columna se puede representar geométrica y gráficamente, las coordenadas o nubes de puntos se representan en un vector como un diagrama de dispersión. Cada punto está dotado de una masa o inercia y la proximidad entre ellos es un indicador de similitud (López y Fachelli, 2015:81-82).

Siguiendo nuestro ejemplo, el gráfico para la tabla de perfiles renglón ocupa un espacio tridimensional, los ejes x, y, z representan las columnas ($Segmento$ 1,2,3), en el plano se ubican los puntos de las filas ($Marcas\ A, B, C, D$) y de su correspondiente perfil promedio (Gráfico 2.1). Las coordenadas, por ejemplo, para la marca "C" son $(0.571, 0.214, 0.214)$ mientras que para el perfil promedio son $(0.333, 0.333, 0.333)$. Las coordenadas se ubican en un triángulo equilátero unido por los puntos vértice $[0,0,1], [0,1,0], [1,0,0]$ (Chiapa, 2015:12)

Gráfico 1: Representación tridimensional de los perfiles fila

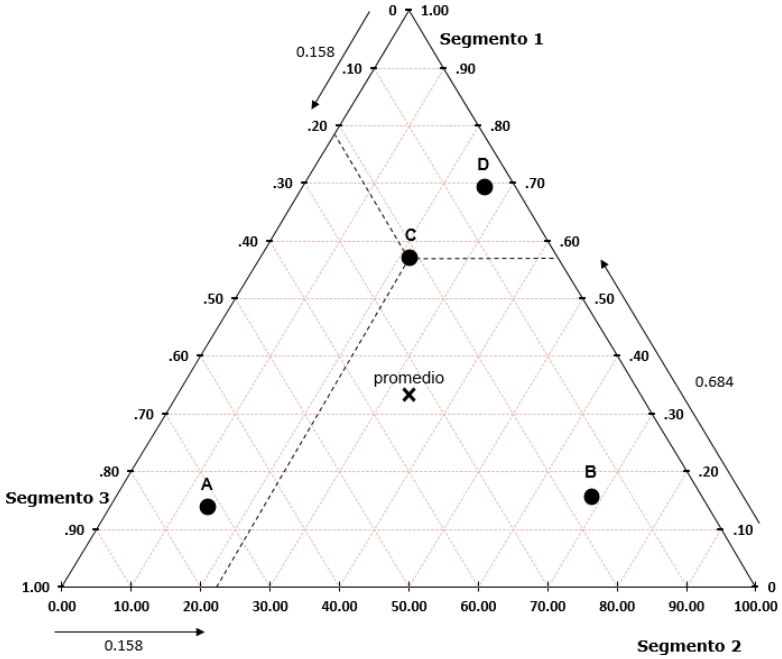


Elaboración propia con base en López y Fachelli, 2015.

Un sistema de coordenadas triangular se utiliza únicamente para perfiles con tres elementos. En caso de $m > 3$ perfiles se usará un sistema coordenado baricéntrico. En el gráfico 2.2, la asociación entre

las variables parece más clara que en la figura anterior, se aprecia una mayor cercanía entre el "segmento 3" y la "marca A", el "segmento 2" y la "marca B" y entre el "segmento 3" y las marcas "C" y "D". Todos los perfiles fila se encuentran cercanos al promedio.

Gráfico 2: Representación bidimensional de los perfiles fila



Elaboración propia con base en López y Fachelli, 2015.

Los puntos de las coordenadas en la gráfica son los promedios ponderados o centroides. Para la marca "A", el centroide es igual a 120.116 y se obtiene de la suma de las frecuencias absolutas de cada categoría de las columnas ($Segmento\ 1{,}2{,}3$) por sus respectivos perfiles de renglón:

$$Centroide\ A = (frecuencia\ del\ segmento\ 1\ de\ la\ marca\ A * perfil\ del\ renglón_{A1}) + (frecuencia\ del\ segmento\ 2\ de\ la\ marca\ A * perfil\ de\ renglón_{A2}) + (frecuencia\ del\ segmento\ 3\ de\ la\ marca\ A * perfil\ del\ renglón_{A3})$$

$$= (30 * 0.140) + (30 * 0.140) + (155 * 0.721)$$

Se procede de la misma manera para el cálculo del centroide o promedio ponderado del perfil promedio del renglón

$$Centroide = (frecuencia\ total\ del\ segmento\ 1 * perfil\ promedio\ del\ renglón_{+1}) + (frecuencia\ del\ segmento\ 2 * perfil\ promedio\ del\ renglón_{+2}) + (frecuencia\ del\ segmento\ 3 * perfil\ promedio\ del\ renglón_{+3})$$

$$= (220 * 0.333) + (220 * 0.333) + (222 * 0.333) = 220$$

Tabla 8: Centroides

Marca	Centroides
A	120.116
B	98.421
C	58.571
D	63.696
Perfil promedio del renglón	220.000

2.2.6 Masa

Son las frecuencias marginales por renglón, resulta de la división entre la frecuencia de consumo de cada marca sobre el total de consumidores

$$r_i = \frac{n_{i+}}{n}$$

Tabla 9: Masas de fila (marca)

	Segmento			
Marca	1	2	3	Masas de renglón
A	0.140	0.140	0.721	0.326
B	0.158	0.684	0.158	0.288
C	0.571	0.214	0.214	0.212
D	0.696	0.261	0.043	0.174
Perfil promedio del renglón	0.333	0.333	0.333	

A su vez, las masas de columna se obtienen mediante:

$$c_i = \frac{n_{+j}}{n}$$

Tabla 10: Masas de columnas (segmento)

	Segmento			
Marca	1	2	3	Perfil promedio columna
A	0.136	0.136	0.705	0.326
B	0.136	0.591	0.136	0.288
C	0.364	0.136	0.136	0.212

D	0.364	0.136	0.023	0.174
Masa de columna	0.333	0.333	0.333	

Las masas de los renglones y las columnas funcionan como pesos ponderados y como promedios (Chiapa, 2015:15). En la tabla 2.6, el perfil promedio de columna es la masa del renglón de la tabla 2.8 y el perfil promedio de renglón en la tabla 2.5 es la masa de columna de la tabla 2.9. Las masas se pueden considerar como el centroide o promedio ponderado. En el ejemplo, para los renglones al multiplicar el perfil del renglón por la masa de la tabla 2.8, se obtiene la masa de columna de la tabla 2.9.

$$[A(0.140 * 0.326) + B(0.158 * 0.288) + C(0.571 * 0.212) + D(0.696 * 0.174)] = 0.333$$

2.2.7 Inercia total (IT)

Mide la varianza de los datos contenidos en una tabla de contingencia. En términos geométricos, mide la distancia (lejanía) entre los perfiles de renglón o de columna y el perfil promedio. Cuando la inercia es baja se dice que existe poca asociación entre los renglones y las columnas ya que los valores de los perfiles de renglón son cercanos a su perfil promedio. Del mismo modo, cuando la inercia es mayor, la asociación entre los perfiles de los renglones y las columnas es alta. No hay correlación cuando el valor de la inercia es cero, en este caso el perfil promedio es igual a los perfiles de renglón o columna (Chiapa, 2015:16).

La inercia total se calcula a partir de la división entre el valor del estadístico Chi-cuadrada entre el número total de casos de la tabla de doble entrada. Para la tabla A, el valor del estadístico es igual a 362.412 dividido entre 660, la inercia total es de 0.549 $\left(\frac{X^2}{660}\right)$.

El estadístico X^2 permite medir la heterogeneidad entre los perfiles.

$$X^2 = \Sigma \frac{(valor\ observado - valor\ esperado)^2}{valor\ esperado}$$

La distancia Chi-cuadrado responde al principio de equivalencia distribucional que establece que dos perfiles iguales pueden transformarse es una sola categoría equivalente a la suma de sus pesos. Este principio permite que se agrupen las categorías con perfiles más semejantes ya sea por renglón o columna (De Fuente, 2011:4).

Al ajustar la formula en términos de los perfiles se tiene que:

$$\text{IT} = \Sigma\, Total\ renglón\ \times \frac{(perfil\ renglón\ observado - perfil\ renglón\ esperado)^2}{perfil\ \ renglón\ esperado}$$

$$0.326 \times \left[\frac{(0.140\times0.333)^2}{0.333} + \frac{(0.140\times0.333)^2}{0.333} + \frac{(0.721\times0.333)^2}{0.333}\right] + 0.288 \times \left[\frac{(0.158\times0.333)^2}{0.333} + \frac{(0.684\times0.333)^2}{0.333} + \frac{(0.158\times0.333)^2}{0.333}\right] + 0.212 \times \left[\frac{(0.571\times0.333)^2}{0.333} + \frac{(0.214\times0.333)^2}{0.333} + \frac{(0.214\times0.333)^2}{0.333}\right] + 0.174\left[\frac{(0.696\times0.333)^2}{0.333} + \frac{(0.261\times0.333)^2}{0.333} + \frac{(0.043\times0.333)^2}{0.333}\right] = 0.549$$

Los conceptos por renglón son equivalentes a las masas multiplicadas por una distancia (X^2). De esta manera la inercia puede expresarse como:

$$IT = \sum (Masa_{i+} \times distancia\ X^2 del\ perfiel\ iésimo\ al\ centroide)^2$$

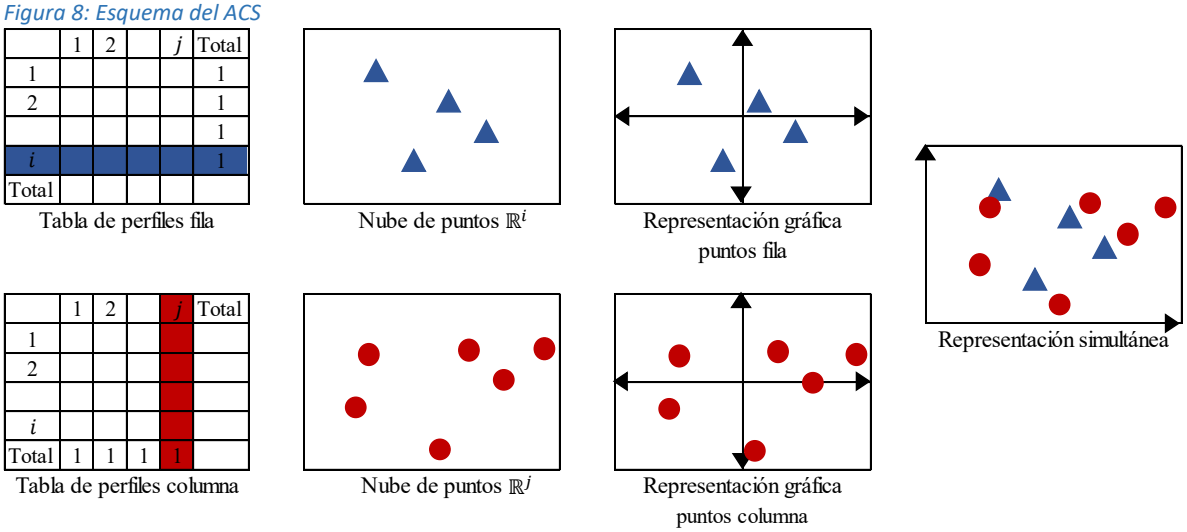
La inercia total figura como un promedio ponderado de las distancias de los renglones fila y columna entre su perfil promedio. En términos geométricos se utiliza la inercia para la reducción de dimensiones y su posterior representación gráfica. Para ello el análisis de correspondencias utiliza la suma ponderada de distancias al cuadrado y la descomposición de los valores singulares equivalente a la disposición matricial de los eigenvalores y los eigenvectores (Chiapa, 2015:18).

En nuestro ejemplo, la inercia total es explicada por dos ejes. El primero explica el 62% de la varianza total y el segundo el 38%, la inercia para el primer eje es de 0.340 y para el segundo de 0.209 $(0.340 + 0.209 = 549)$. A los valores anteriores se les conoce también como eigenvalores.

El procedimiento del ACS se estructura de dos momentos, la construcción de la matriz de varianzas y covarianzas definidas por una métrica y la extracción de las dimensiones o factores. La dispersión de los puntos sobre el eje se mide a través de los valores propios o eigenvalores (López y Fachelli, 2015:83). En términos generales, el ACS trasforma o re-proyecta los datos de los puntos fila y columna con el fin de que estos puedan representarse geométricamente en un mismo espacio conservando la mayor variabilidad posible de los datos originales.

Con base en la representación gráfica de todas las categorías de las variables se esperaría la interpretación del porqué de la posición de los puntos en el plano. La inercia total, de cada dimensión son fundamentales en el análisis del modelo, así mientras mayor sea el porcentaje en las dos primeras dimensiones, éste estará mejor representado en un espacio bidimensional (Chiapa, 2015:18).

Figura 8: Esquema del ACS



Elaboración propia con base en López y Facheli, 2015:84

2.3 INTERPRETACIÓN DEL ACS

El proceso de interpretación del análisis de correspondencias implica la definición del número de factores que resuman la mayor variabilidad posible y la representatividad en cada componente de la asociación entre las categorías de las variables (López y Facheli, 2015:86).

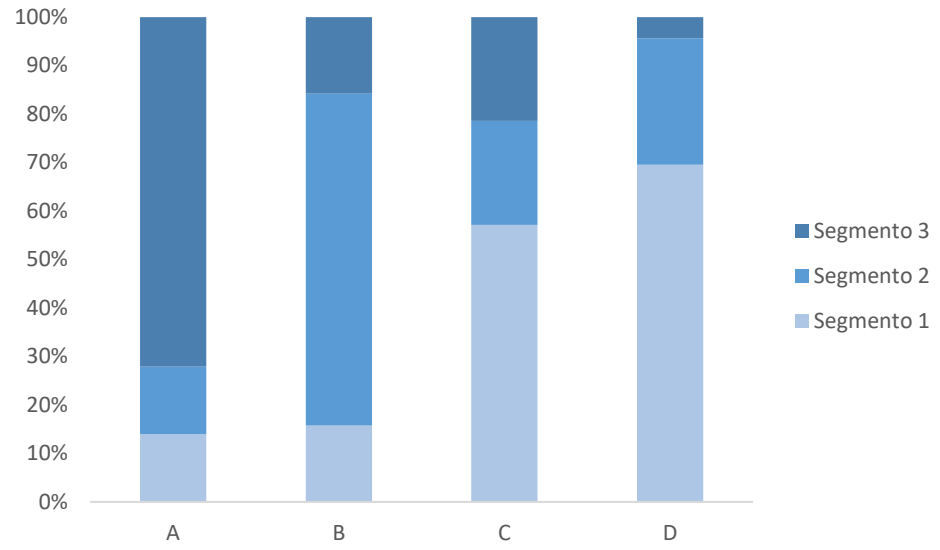
En término generales, en la tabla 2.10 y el gráfico 2.3, se aprecia la asociación entre el consumo de la marca "A" por el segmento "3", una mayor predisposición de consumo de la marca "B" por el segmento "2" y de la marca "C" y "D" por el segmento "1". Sin embargo, no se aprecia claramente una tendencia de consumo de los sectores referente al resto de las marcas, por ejemplo, el consumo de la marca "B" por el segmento 3, disminuye respecto a la marca "A" y "C" pero aumenta respecto a la marca "D".

Tabla11: Distribución de frecuencias. Preferencia de consumo por segmento

			Segmento			**Total**
			Seg_1	Seg_2	Seg_3	
Marca	Marca A	Recuento	30	30	155	215
		% por fila	14.0%	14.0%	72.1%	100.0%
	Marca B	Recuento	30	130	30	190
		% por fila	15.8%	68.4%	15.8%	100.0%
	Marca C	Recuento	80	30	30	140
		% por fila	57.1%	21.4%	21.4%	100.0%
	Marca D	Recuento	80	30	5	115
		% por fila	69.6%	26.1%	4.3%	100.0%

Total	Recuento	220	220	220	660
	% fila	33.3%	33.3%	33.3%	100.0%

Gráfico 4: Consumo de marca según segmento



El número de factores o dimensiones es igual al mínimo de columnas y filas menos 1 $(mín\{I,J\}-1)$. Para nuestra tabla el número de categorías para la variable "Marca" es igual a 4 y para la variable "Segmento" es igual a 3, por lo tanto, el menor número de categorías es igual a tres menos uno, dando como resultado dos dimensiones. En la tabla 2.11 se muestran los valores propios equivalentes a la inercia total o varianza acumulada por cada dimensión o factor.

Tabla 12: Valores propios e inercia total (Segmento –Marca)

	Valores propios		
Factor	Valor	% de la varianza	% acumulado
1	.340	62.00	.620
2	.209	38.00	100
Total	.549	100.00	

El primer factor recoge el 62% de la varianza total y el segundo el 38%, la suma de los valores propios es igual a la inercia total de la nube de puntos. Técnicamente se sugiere que el primer factor acumule la mayor parte de la inercia o varianza, en caso de que el número de dimensiones sea mayor a 2, esta puede ser un buen indicador para definir el número de dimensiones a representar, el valor de referencia en esta situación es de por lo menos el 70% (López y Facheli, 2015:88).

La información contenida en la tabla 2.12 y 2.13 es auxiliar para la interpretación de los factores, los datos se presentan para cada variable en relación a los puntos fila y columna. El peso relativo o masa,

corresponde a la frecuencia relativa en relación al total y evalúa la importancia de cada categoría. Por otra parte, las coordenadas permiten la representación gráfica de las categorías en un mapa perceptual.

Las contribuciones absolutas se refieren a la proporción de varianza total explicada o inercia que aporta cada variable para cada eje factorial. Si cada categoría aportara la misma inercia la contribución absoluta para una de ellas sería de 0.33 en las columnas $(1 \div 3)$ y de 0.25 $(1 \div 4)$en las filas, es por ello que las categorías que aporten por sobre ese porcentaje (influencia superior a la media) serán aquellas que tengan mayor influencia en el eje.

En relación a las columnas, en el primer factor la categoría que más aporta es el segmento "3" con el 66.2% (0.662>0.33). En cuanto a las filas, la marca "A" aporta al eje el 64.2% (0.642> 0.25). El resto de las categorías no hacen una aportación significativa. El primer factor que aporta el 62% de la varianza total de la tabla de contingencia es representativo de la asociación entre la marca "A" y el segmento "3".

Para el segundo factor, en las columnas el segmento "2" aporta el 54.7% (0.547>0.33) y el segmento "1" el 44.7% (.447>0.33). En las filas, la variable que más aporta es la marca "B" con el 62.8% (0.628>0.33) y en menor medida la marca "D" con el 17.4% y la marca "C" con el 19%. El patrón de asociación se identifica con el segmento "2" a la marca "B".

Tabla 13: Coordenadas y contribuciones. Puntos de columna

		Coordenadas		Contribución				
				Del punto en la inercia de dimensión (Absolutas)		De la dimensión en la inercia del punto (Relativas)		
Segmento	Peso relativo o Masa	Eje 1	Eje 2	Eje 1	Eje 2	Eje 1	Eje 2	Total
Seg_1	.333	.619	.783	.219	.447	.444	.556	1.000
Seg_2	.333	.457	-.866	.119	.547	.262	.738	1.000
Seg_3	.333	-1.076	.083	.662	.005	.995	.005	1.000
Total activo	1.000			1.000	1.000			

Tabla 14: Coordenadas y contribuciones. Puntos de fila

Marca	Peso relativo o masa	Coordenadas		Contribución				
				Del punto en la inercia de dimensión (Absolutas)		De la dimensión en la inercia del punto (Relativas)		
		Eje 1	Eje 2	Eje 1	Eje 2	Eje 1	Eje 2	Total
Marca A	.326	-1.072	.106	.642	.008	.992	.008	1.000
Marca B	.288	.412	-.998	.084	.628	.179	.821	1.000
Marca C	.212	.379	.612	.052	.174	.329	.671	1.000
Marca D	.174	.862	.706	.222	.190	.656	.344	1.000
Total activo	1.000			1.000	1.000			

Las contribuciones de la dimensión a la inercia del punto o contribuciones relativas se definen como la proporción de varianza total explicada de la categoría respecto a los ejes. Se interpretan como un coeficiente de correlación y muestra que tan representativa es la categoría de una variable en cada factor (López y Facheli, 2015:90).

Se dice que la correlación es alta cuando una variable acumula la mayor inercia sobre una dimensión y tiende a proyectarse mayormente sobre un eje y en menor medida sobre el resto. Los valores cercanos a 1 indican una mejor representación de la categoría i sobre el eje k. El segmento “3” y la marca “A” acumulan más del 90% de inercia en el primer eje. El segmento “2” acumula el 73% y la marca “B” el 82% de la variabilidad en el segundo eje.

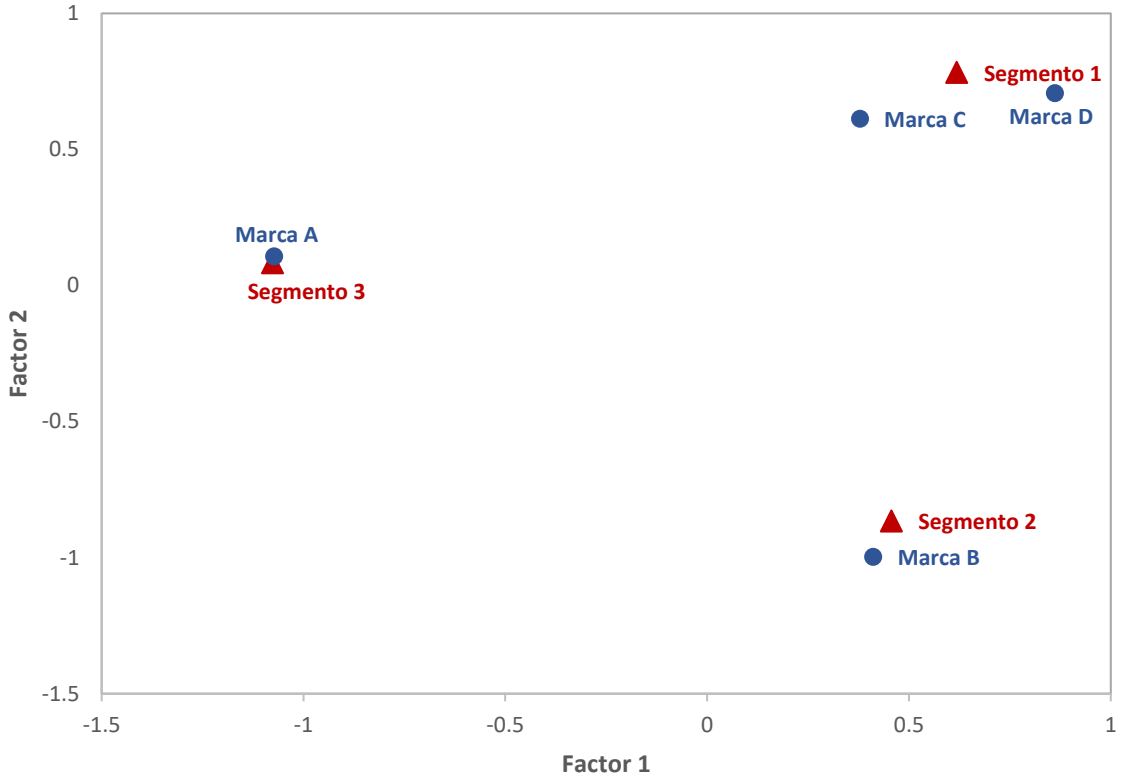
El último elemento en la interpretación del análisis de correspondencias es la representación gráfica. Este aspecto permite analizar de manera visual los patrones que se identificaron en las tablas anteriores. La representación se puede hacer mediante la graficación de los puntos fila ($\mathbb{R}^J$), los puntos columna ($\mathbb{R}^I$) o de manera simultánea. En la interpretación de los gráficos de debe considerar que cada categoría está representada en el espacio vectorial.

Para analizar la estructura y orden de las asociaciones, a partir del razonamiento de tablas, se localizan en el gráfico las categorías con mayor contribución absoluta, de estás mismas, se retoman las contribuciones relativas o la representatividad para corroborar la correlación con los ejes.

Aquellas categorías con perfiles fila o columna cercanas al perfil promedio se ubicarán próximas al origen, mientras más se alejen de este más lejanas se ubican de la media. Así mismo, categorías con perfiles similares se ubican próximas entre sí, sin embargo, está proximidad se analiza solo si se ubica

alejada del origen, de lo contario la proximidad podría asociarse a la lejanía con otros perfiles más que a la cercanía con la variable contigua (López y Facheli, 2015:91).

Gráfico 4: Mapa perceptual. Segmento y marca



Elaboración propia con base en López y Fachelli, 2015.

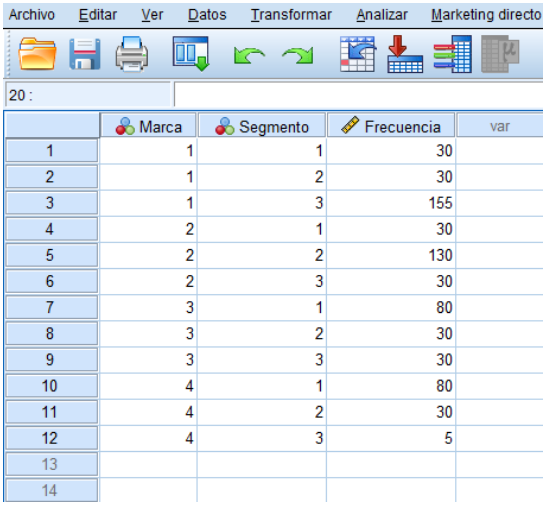
El mapa perceptual para la tabla de contingencia indica un patrón de asociación de tres grupos entre el segmento "3" y la marca "A", el segmento "2" y la marca "B" y el "segmento 3" con la marca "C" y "D". Si bien en las tablas el primer grupo parecía bien definido, a través del análisis gráfico pudieron definirse el resto de los grupos.

2.4 ANÁLISIS DE CORRESPONDENCIAS CON SPSS

2.4.1 Diseño de la vista de datos

Para realizar el análisis de correspondencias en SPSS requerimos organizar los datos contenidos en una tabla de contingencia o de doble entrada. La vista de datos se conforma por tres variables; "Variable 1", "Variable 2" y "Frecuencia", el número de filas es igual al producto del número de categorías de la variable 1 por la variable 2 ($4\ x\ 3 = 12$) quedando de la siguiente manera:

Figura 9: Vista de datos

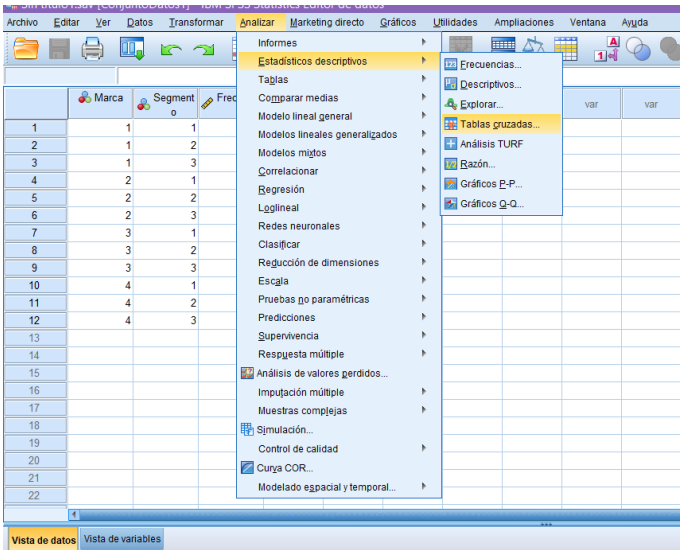


	Marca	Segmento	Frecuencia	var
1	1	1	30	
2	1	2	30	
3	1	3	155	
4	2	1	30	
5	2	2	130	
6	2	3	30	
7	3	1	80	
8	3	2	30	
9	3	3	30	
10	4	1	80	
11	4	2	30	
12	4	3	5	
13				
14				

2.4.2 Ajuste del modelo

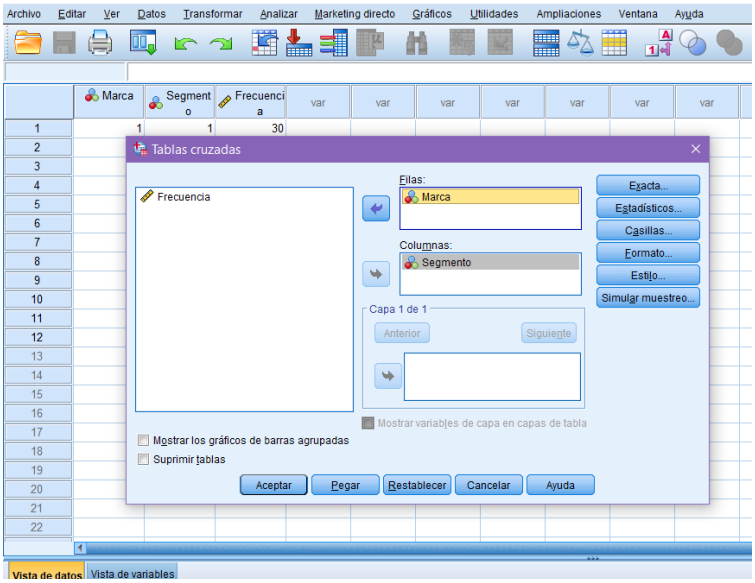
Antes de aplicar el método se debe corroborar que el conjunto de datos se ajuste al modelo. Si el objetivo del AC es la estructura de asociación se debe asegurar la presencia de ésta entre las categorías de las variables. Para ello nos apoyamos en la prueba Chi-cuadrada (Rodríguez y Mora, 2001:48-49)-

Figura 10: Analizar/ Estadísticos descriptivos/ Tablas cruzadas



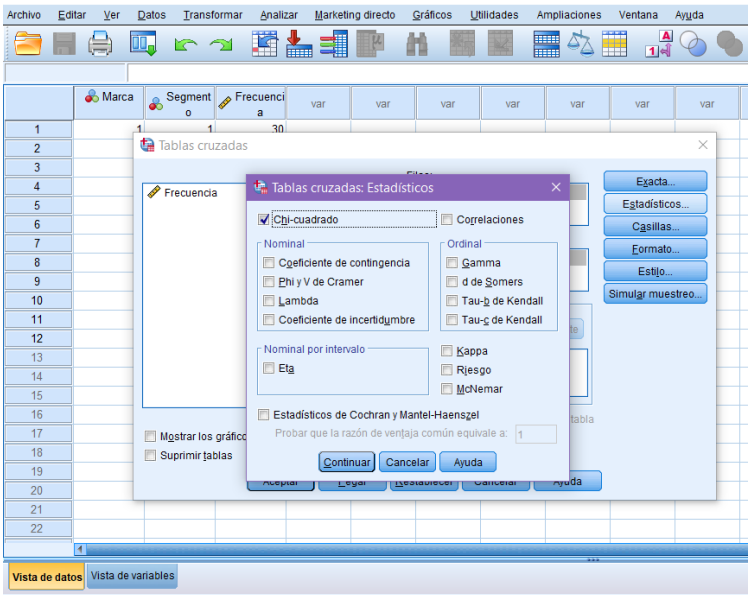
Paso 2. Ingresar los datos de acuerdo a la estructura de la tabla de doble entrada. Se sugiere colocar a la variable dependiente en las filas y a la independiente en las columnas.

Figura 11: Ingreso de variables



Paso 3. Estadísticos/ Chi-cuadrado/ Continuar/ Aceptar

Figura 12: Tablas cruzadas, estadísticas

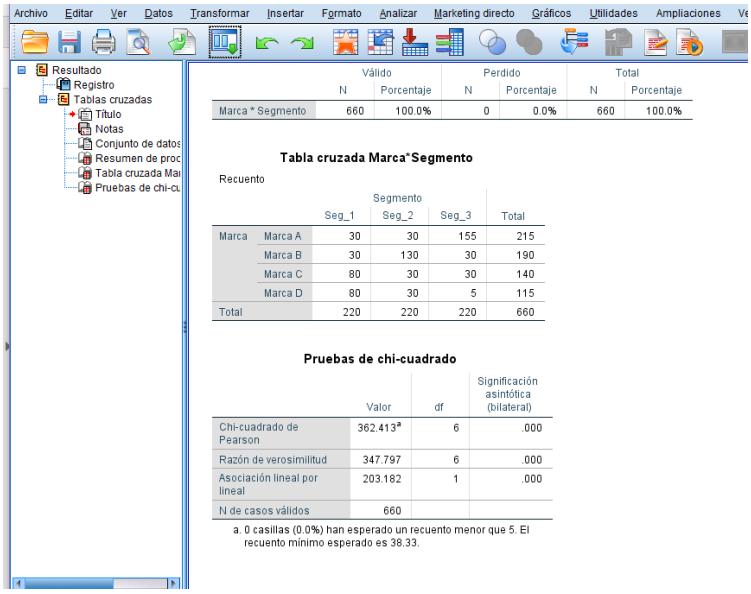


Paso 4. Análisis de la prueba.

La hipótesis de independencia (H_0) establece que no hay asociación entre las variables. Se rechaza la (H_0) si el estadístico es menor o igual al p-valor (= 0.05 para este trabajo).

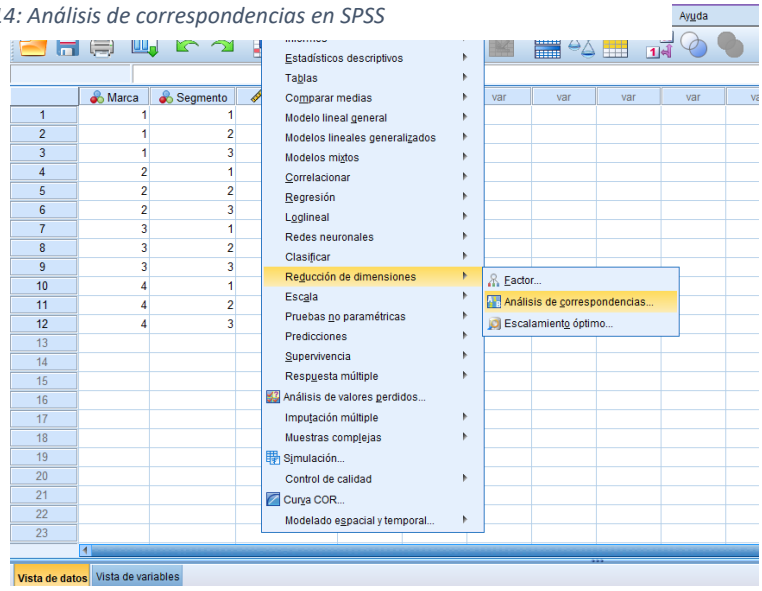
Por ejemplo, el valor del estadístico para el ejemplo es de 362.4, por lo tanto, al ser mayor al p-valor el AC procede.

Figura 13: Salida análisis de la prueba



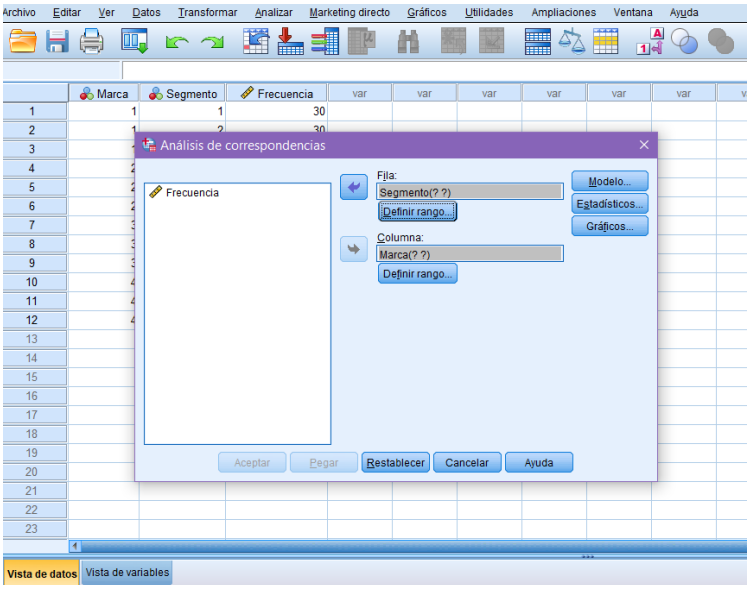
Paso 5. Menú principal/ Analizar/ Reducción de dimensiones/Análisis de correspondencias

Figura 14: Análisis de correspondencias en SPSS



Paso 6. Especificar nuevamente las variables para las columnas y las filas de acuerdo al diseño de la tabla de contingencia.

Figura 15: Especificación de variables

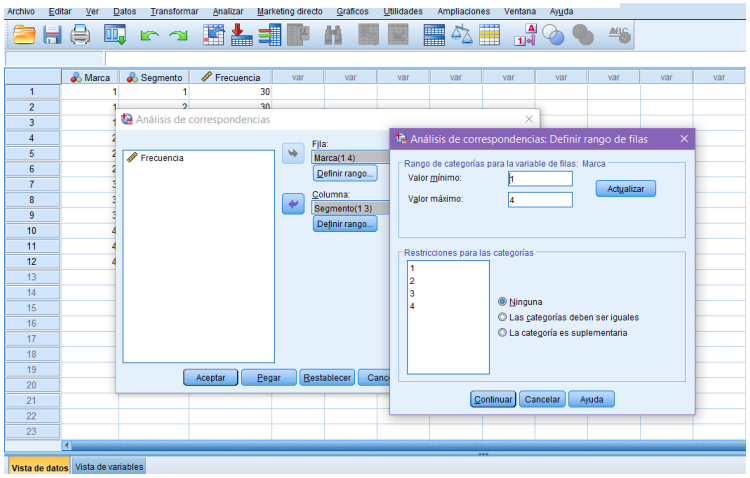


Paso 7. Definir el rango para cada variable. El rango es igual al valor mínimo y máximo del número de categorías las variables, define la amplitud entre las categorías de las variables.

Por ejemplo, la variable “Marca” tiene cuatro categorías, por lo tanto, su valor mínimo es 1 y el máximo 4. La variable “Segmento” tiene 3 categorías, por lo tanto, el valor mínimo es de 1 y el máximo de 3.

Actualizar/ Continuar

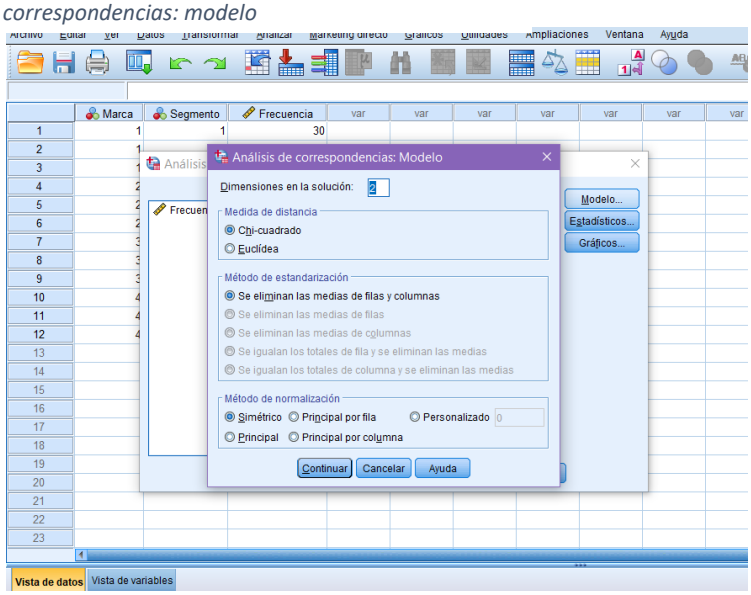
Figura 16: Definición de categorías por variable



Paso 8. Al determinar la especificación del rango, se procede a seleccionar el método de extracción de las dimensiones o factores. En este cuadro de dialogo podemos definir:

a. Número de dimensión de la solución. Para una mejor representación visual del mapa perceptual se seleccionan dos dimensiones. El número máximo de dimensiones es igual al número de categorías (de la variable con el menor número de categorías) menos 1. Por ejemplo, para el conjunto de datos a trabajar el número máximo de dimensiones son 2
b. Medida de distancia. Por el tipo de datos, generalmente se utiliza la Chi-cuadrada para establecer la distancia. De manera alternativa puede utilizarse otra medida como la euclidiana que se basa la raíz cuadrada de la suma de los cuadrados, sin embargo, esta se asocia más a otros métodos estadísticos multivariados como el análisis de componentes principales
c. Método de estandarización. Se determina a partir de la distancia seleccionada, para la Chi-cuadrada el método de estandarización implica una medida de filas y columnas. En el caso de haber seleccionado una distancia euclidiana se consideran cualquiera del resto de las cuatro opciones.
d. Método de normalización. Fija el método de normalización de las filas y las columnas. El método simétrico asigna el mismo valor o peso para las filas y las columnas. El método principal compara la distancia entre las categorías y no entre las variables. El método por fila o columna compara las categorías de las filas o columnas y el método personalizado permite ingresar un valor de -1 a 1 refiriéndose a las columnas o a las filas respectivamente (López y Facheli, 2015:130)

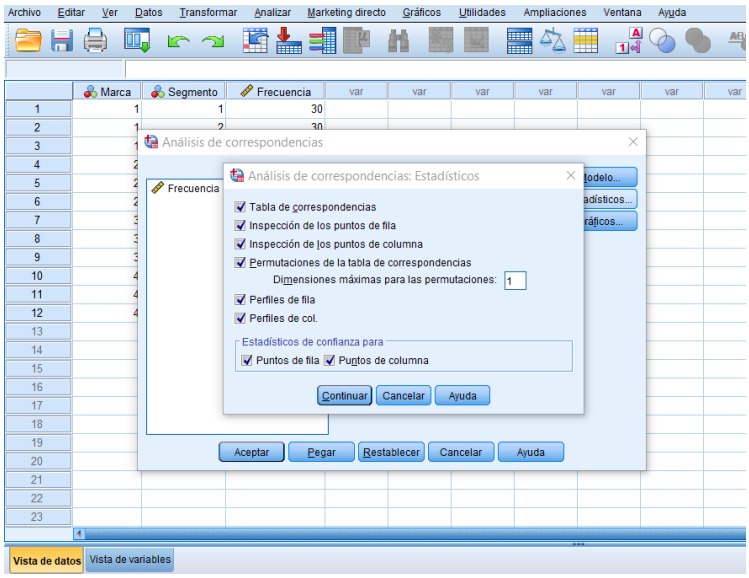
Figura 17: Análisis de correspondencias: modelo



Paso 9. En el cuadro de salida de Estadísticos se especifican las medidas de interés para identificar las similitudes entre categorías:

a. Tabla de correspondencias. Tabla de doble entrada con las frecuencias absolutas entre las dos variables
b. Inspección de los puntos fila o columna. Tabla para cada categoría por fila o columna en donde se registran las masas, las puntuaciones, la inercia y la contribución de la inercia al punto y viceversa. Esta tabla puede ser un auxiliar para el análisis gráfico
c. Permutaciones de la tabla de correspondencias. Se obtiene una tabla permutada a partir de la original con base en las puntuaciones de las filas y las columnas. Para este tipo de análisis se considera únicamente la primera dimensión.
d. Perfiles de fila y de columna. Tabla de frecuencias relativas por fila o por columna
e. Estadísticos de confianza. Se asociación a la desviación y la correlación de los puntos fila y los puntos columna (López y Facheli, 2015:130)

Figura 18: Análisis de correspondencias, estadísticas



Paso 10. En el cuadro de dialogó “Gráficos” se disponen de las siguientes opciones:

a. Diagrama de dispersión biespacial. Es el mapa perceptual que representa las categorías de las variables de las dos primeras dimensiones. Cuando se trata con más de dos

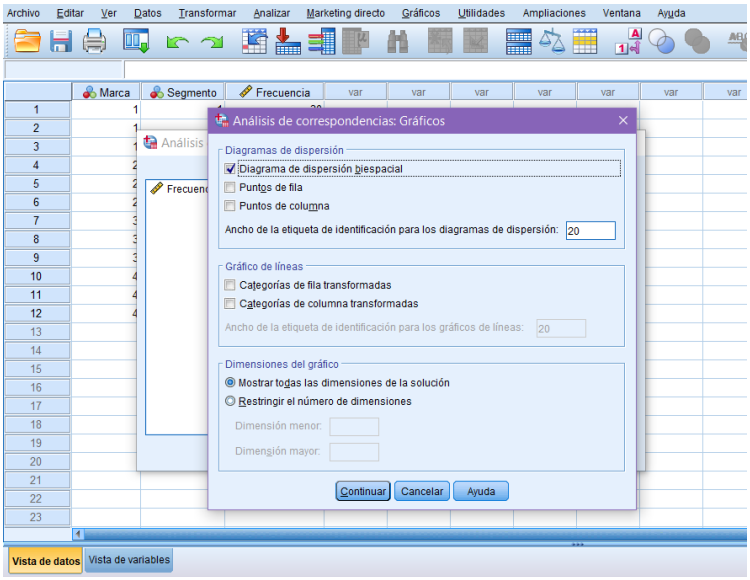
dimensiones, los mapas perceptuales se presentan separado. También se pueden graficar los puntos fila y columna

b. Gráficos de línea. Representan las puntuaciones y las categorías de las variables por fila o por columna

c. Dimensiones del gráfico. Se pueden seleccionar el número máximo de dimensiones que se deseen representar en el grafico

Continuar/ Aceptar

Figura 19: Análisis de correspondencias, gráficos



Si se desea seguir la rutina anterior en SPSS, la sintaxis de las indicaciones anteriores es la siguiente:

Archivo/ Nuevo/ Sintaxis/ Seleccionar /Ejecutar selección

```
CORRESPONDENCE TABLE=Marca (1 4) BY Segmento (1 3)
/DIMENSIONS=2
/MEASURE=CHISQ
/STANDARDIZE=RCMEAN
/NORMALIZATION=SYMMETRICAL
/PRINT=TABLE RPOINTS CPOINTS PERMUTATION (1) RPROFILES CPROFILES
RCONF CCONF
```

2.4.3 Resultados

De la ejecución del procedimiento anterior se desprenden los siguientes resultados:

Tabla de correspondencias, de contingencia o de doble entrada. Se muestra la distribución de las frecuencias relativas

Tabla 15: Correspondencias

	Segmento			
Marca	Seg_1	Seg_2	Seg_3	Margen activo
Marca A	30	30	155	215
Marca B	30	130	30	190
Marca C	80	30	30	140
Marca D	80	30	5	115
Margen activo	220	220	220	660

Tabla de perfiles de fila y de perfiles de columna. Muestra las proporciones por fila o columna

Tabla16: Perfiles de fila, SPSS

	Segmento			
Marca	Seg_1	Seg_2	Seg_3	Margen activo
Marca A	.140	.140	.721	1.000
Marca B	.158	.684	.158	1.000
Marca C	.571	.214	.214	1.000
Marca D	.696	.261	.043	1.000
Masa	.333	.333	.333	

Tabla 17: Perfiles de columna, SPSS

	Segmento			
Marca	Seg_1	Seg_2	Seg_3	Masa
Marca A	.136	.136	.705	.326
Marca B	.136	.591	.136	.288
Marca C	.364	.136	.136	.212
Marca D	.364	.136	.023	.174
Margen activo	1.000	1.000	1.000	

Tabla resumen. Para nuestro conjunto de datos el número de dimensiones que acumula el total de la varianza explicada es igual a 2 ($mín\{I,J\} - 1 = min\{4,3\} - 1 = 2$). En la tabla se muestra el valor singular que es igual a la raíz cuadrada del valor propio o inercia de cada dimensión. El total de la inercia es igual al valor de la Chi- cuadrada entre el número total de casos, mide la dispersión de la

nube de puntos. ($362.413 \div 660 = .549$). En la tabla se muestra la proporción de varianza explicada para cada dimensión (la primera dimensión explica el 62% de la varianza total y la dimensión dos el 38%).

Por último, el valor singular de confianza muestra la desviación estándar y la correlación entre las dimensiones, estas medidas son un indicador de la precisión de las dimensiones, valores cercanos a cero indican una mayor confiabilidad en los datos (López y Facheli, 2015:133).

Tabla 18: Resumen, SPSS

Dimensión	Valor singular	Inercia	Chi cuadrado	Sig.	Proporción de inercia		Valor singular de confianza	
					Contabilizado para	Acumulado	Desviación estándar	Correlación 2
1	.584	.340			.620	.620	.032	.155
2	.457	.209			.380	1.000	.038	
Total		.549	362.413	.000[a]	1.000	1.000		

a. 6 grados de libertad

1. Puntos de fila o de columna general

En las siguientes tablas se muestran las masas o pesos, las puntuaciones en dimensión o coordenadas del gráfico perceptual, la inercia equivalente a la varianza explicada, las contribuciones absolutas a la inercia de la dimensión y las contribuciones relativas de la dimensión a la inercia del punto de cada categoría de las variables por fila y columna.

Tabla 19: Puntos de fila generales, SPSS

Marca	Masa	Puntuación en dimensión		Inercia	Contribución				
					Del punto en la inercia de dimensión		De la dimensión en la inercia del punto		
		1	2		1	2	1	2	Total
Marca A	.326	-1.072	.106	.220	.642	.008	.992	.008	1.000
Marca B	.288	.412	-.998	.159	.084	.628	.179	.821	1.000
Marca C	.212	.379	.612	.054	.052	.174	.329	.671	1.000
Marca D	.174	.862	.706	.115	.222	.190	.656	.344	1.000
Total activo	1.000			.549	1.000	1.000			

a. Normalización simétrica

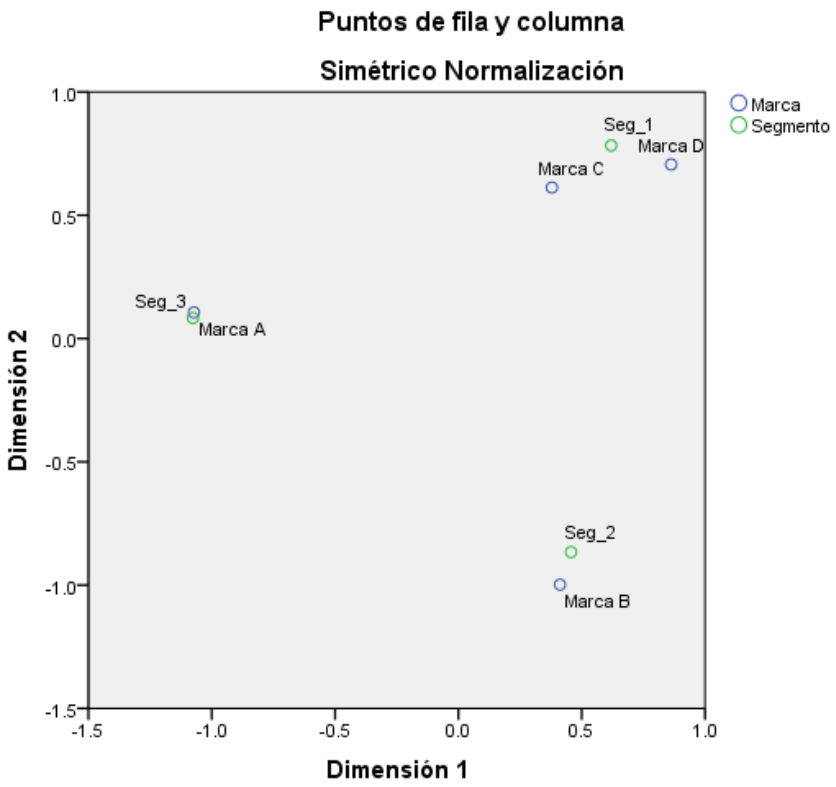
Tabla 20: Puntos de columnas generales, SPSS

		Puntuación en dimensión			Contribución				
					Del punto en la inercia de dimensión		De la dimensión en la inercia del punto		
Segmento	Masa	1	2	Inercia	1	2	1	2	Total
Seg_1	.333	.619	.783	.168	.219	.447	.444	.556	1.000
Seg_2	.333	.457	-.866	.155	.119	.547	.262	.738	1.000
Seg_3	.333	-1.076	.083	.226	.662	.005	.995	.005	1.000
Total activo	1.000			.549	1.000	1.000			

a. Normalización simétrica

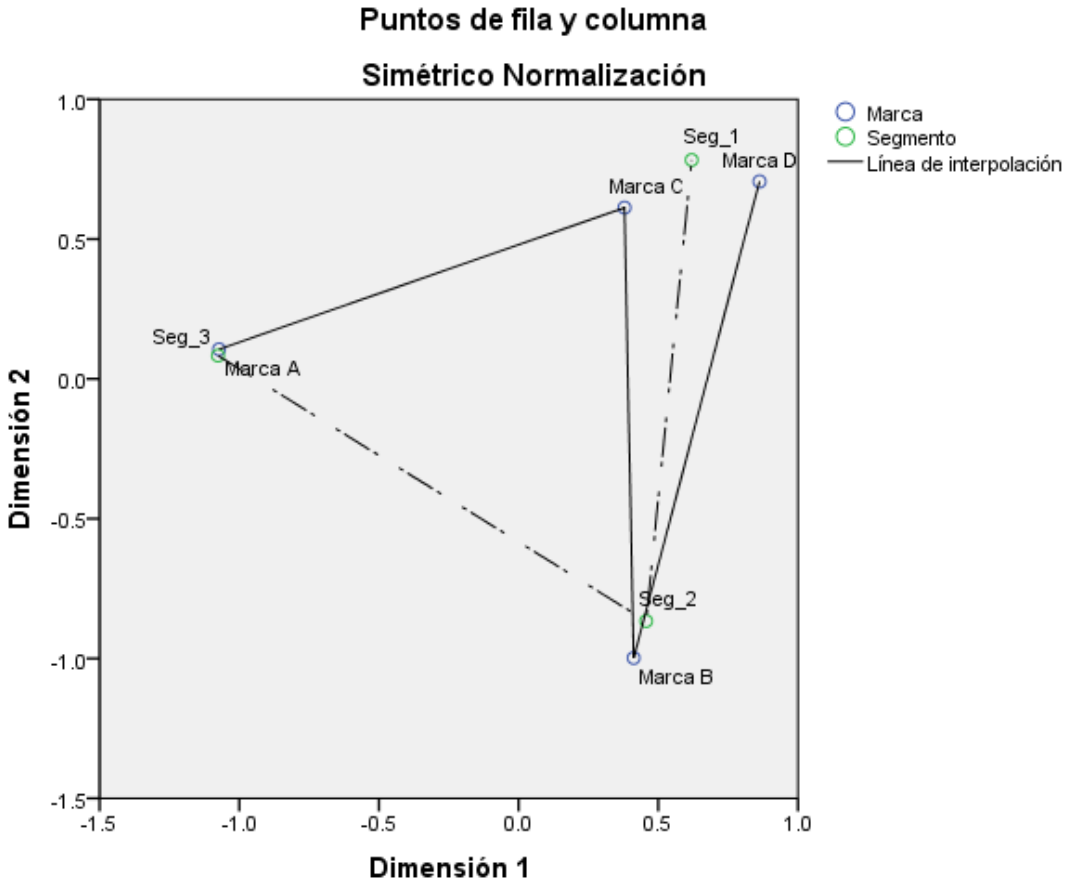
2. Mapa perceptual. Gráfica el conjunto de los puntos fila y los puntos columna

Gráfico 6: Puntos de fila y columna



La unión de los puntos se puede representar a través de una línea de interpolación

Gráfico 6: Puntos de fila y columna. Línea de interpretación



2.5 Validación del procedimiento

Para verificar los resultados obtenidos de un análisis de correspondencia basta con aplicar otra técnica para determinar la interdependencia entre los datos originales. Una técnica sugerida es el escalamiento multidimensional que en términos generales consiste en representar geométricamente las proximidades (de semejanza o desemejanza) de un conjunto de estímulos medidos en un número de variables (Aldás, 2000:11).

CAPITULO 3: APLICACIONES DE LA MINERÍA DE DATOS EN GEOGRAFÍA

La minería de datos para la extracción de conocimiento puede aplicarse a distintas áreas de la geografía, cómo la geografía física, económica, política o de la percepción. A continuación, se desarrollan tres aplicaciones relacionadas con el asesinato de periodistas, la migración de jornales agrícolas y la ocurrencia de fenómenos naturales en México.

3.1. GEOGRAFÍA SOCIAL.

"LOS PAÍSES MÁS PELIGROSOS PARA EJERCER EL PERIODISMO"

3.1.1. Planteamiento.

La segunda búsqueda más frecuente en Google si se introduce la frase: México, país peligroso para, corresponde, solo después de la palabra "mujeres", a "periodistas"; México, país peligroso para periodistas. Una búsqueda rápida sobre el tema da cuenta del peligro que representa el quehacer periodístico en México. Tan solo a mediados de marzo del presente año circuló entre los medios nacionales el encabezado: "México, el país más peligroso para ejercer el periodismo: RSF" en donde de acuerdo con la organización Reporteros Sin Fronteras, México se sitúa en el lugar 144 de 180 países en el ranking mundial sobre libertad de prensa, las cifras reportadas indican que en 2018 fueron asesinados 10 periodistas habiendo una tendencia al alza y en un contexto de impunidad (Proceso, 2019a).

De acuerdo con Arribas (2016) y Ríos (2012), existe una lógica espacial que explica la muerte de periodistas en México, de acuerdo con ambos autores, coexiste una relación entre las zonas rojas que registran el mayor número de asesinatos y la presencia del crimen organizado, principalmente asociado con el narcotráfico. Los periodistas que cubren áreas estratégicas en donde operan grupos criminales, como son los estados de Guerrero, Chihuahua y Veracruz, tienen mayor probabilidad de ser víctima de un ataque. Arribas (2016), concluye que este tipo específico de inseguridad está asociado con la muerte de periodistas, a manera de comparación un profesional puede cubrir zonas con altos niveles de homicidios cometidos por ciudadanos comunes sin que esto suponga un riesgo significativo.

El ejercicio periodístico se convierte en una herramienta que permite garantizar el acceso de la población a información en temas estratégicos de interés nacional. De acuerdo con la Declaración de los Derechos Humanos, el artículo 19 establece que cualquier individuo puede gozar de la libertad de opinar y expresarse sin que ello implique ser molestado. El ambiente generalizado de inseguridad e ilegalidad que impera en el país y que se vio encrudecido a partir de la "Guerra contra el narcotráfico" impulsada en el sexenio de Felipe Calderón Hinojosa, ha trasgredido dos derechos constitucionales y humanos fundamentales: el derecho a la libertad de expresión y el derecho a la vida. Una prensa crítica y con autonomía garantiza el acceso de la ciudadanía a la información y a su vez, permite que ésta se involucre en asuntos prioritarios para el país (Arribas, 2016).

Violentar la labor periodística no es un asunto menor, este tipo de violencia resquebraja el sistema democrático al no ser este capaz de garantizar el respeto y la promoción de los derechos humanos. Así también, debilita la estructura institucional y pone en duda la capacidad del estado en la impartición de justicia poniendo de manifiesto la simulación de este (Ramírez, 2018). El asesinato de periodistas representa un mecanismo de control sobre la información, coartar la libre expresión promueve un ambiente de desinformación entre la ciudadanía y obstaculiza el libre ejercicio de la labor periodística en temas estratégicos como la seguridad nacional, el narcotráfico o la corrupción creando así, vacíos de información (Arribas 2016).

De acuerdo con cifras del Comité para la Protección de los Periodistas (CPJ por sus siglas en inglés), México ocupa el lugar número 11 de esta lista, sin embargo, su posición varía de acuerdo al año de referencia, por ejemplo, para el año 2017 México se ubicó en el tercer lugar, solo por debajo de Irak y Siria y superó a Afganistán. De 2015 a la fecha, México se ubicó entré los 10 países con mayor número de muertes a periodistas. (Gráfico 7).

De 1994 a enero de 2019, se han registrado 1,337 asesinatos de periodistas en 103 países de todo el mundo. Irak encabeza la lista con 186 muertes, Siria con 126 y Filipinas con 80. Tan solo el 70% de las muertes se concentra en 15 países (Figura 20).

Gráfico 7: Países con mayor registro de muertes a periodistas (2015 - 2019)

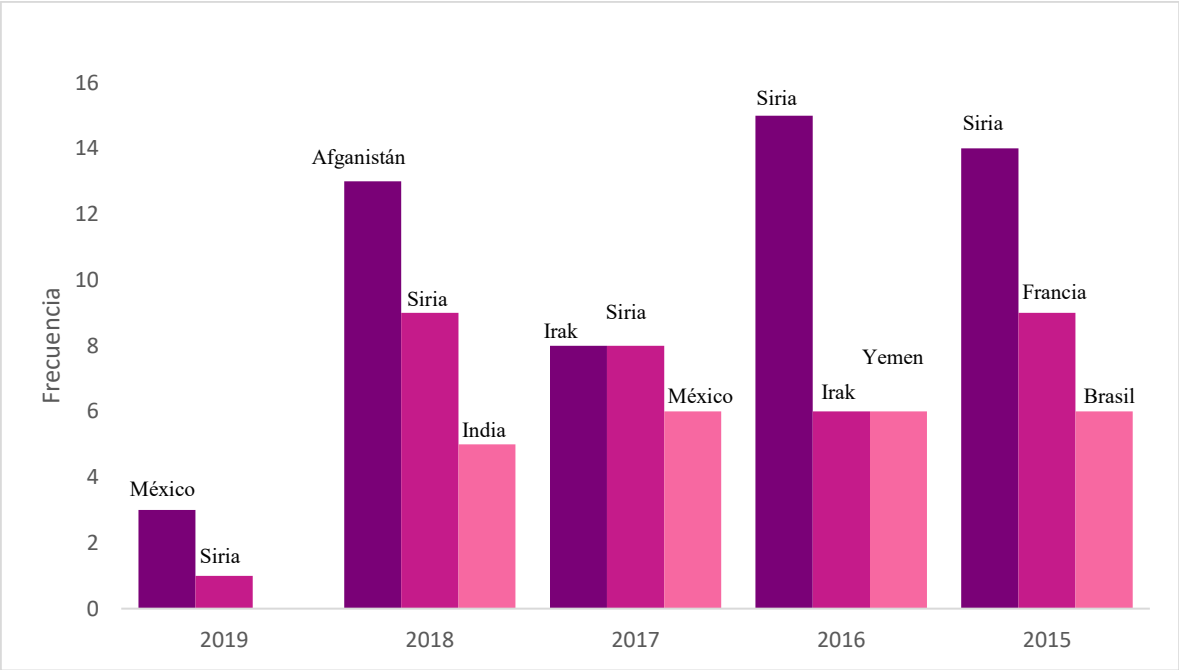
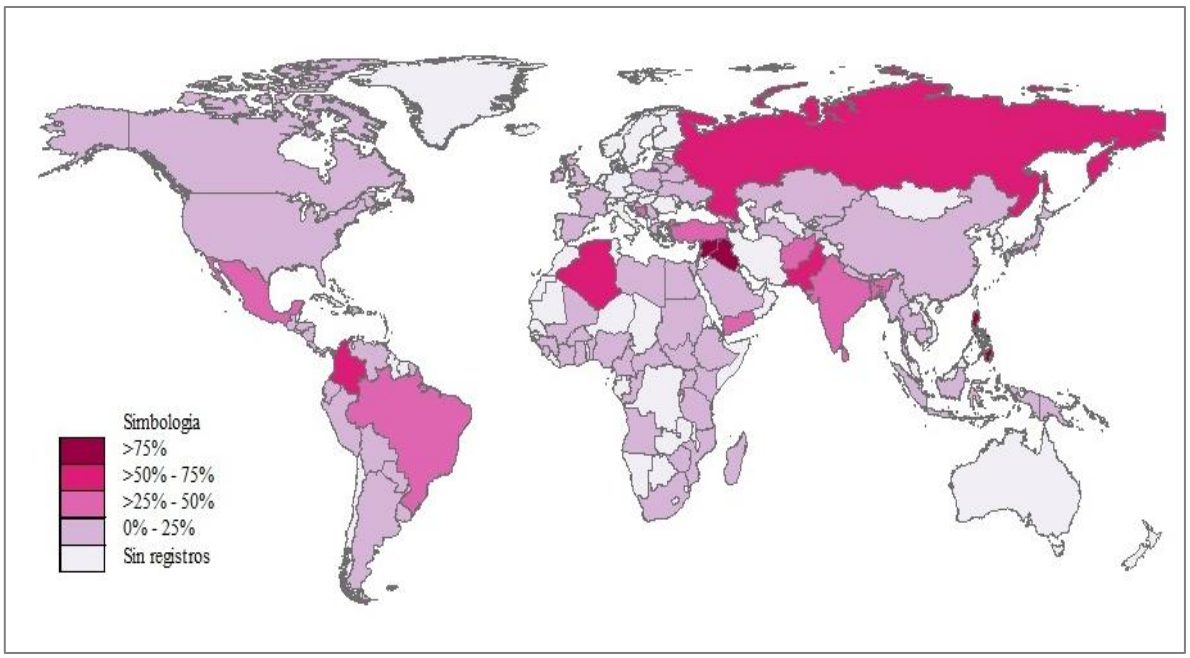


Figura 20: Registro de muertes de periodistas en el mundo



Tales cifras han llevado a afirmar que México es comparable con países en guerra, cuenta de ello son las siguientes notas que tuvieron fecha el 17 de diciembre de 2018: “México es igual de peligroso que Afganistán para ejercer el periodismo” (Infobae, 2018) y “México, Afganistán y Siria, los países más letales para los periodistas en 2018” (Proceso 2018). El análisis de las cifras reportadas por distintas fuentes privilegia los datos en términos cuantitativos, sin embargo, en términos de cualidad, ¿podríamos afirmar que México es comparable con Siria y Afganistán? ¿Las causas de muerte son las mismas para estos países? De acuerdo con las aseveraciones anteriores, se propone analizar las causas de muerte de periodistas en todo el mundo con el objetivo de identificar el conjunto de países que son comparables con México (Godina, 2017).

3.1.2. Fuente de los datos.

Existen diferentes fuentes de información sobre el número de muertes a periodistas, sin embargo, en ninguna de ellas existe un consenso general (Arribas 2016). Para el presente trabajo se recogen los datos del Comité para la Protección de los Periodistas cuyos registros van del año de 1994 hasta la actualidad, la información que se presenta se basa en evidencia judicial y puede desagregarse por fecha y motivo de muerte.

El CPJ (2019), define a un periodista como aquel sujeto encargado de cubrir noticias de sucesos públicos o hechos coyunturales en diferentes plataformas de comunicación como el radio, la televisión, la fotografía, en medios impresos o a través de la Internet. La base de datos utilizada en el presente análisis considera únicamente aquellos casos confirmados en donde la muerte del periodista fue motivada por la labor desempeñada.

Los registros muestran el nombre del periodista, la organización empleadora, la fecha del deceso, el año, la ubicación y el tipo de muerte. Esta última variable indica la causa directa de la muerte y tiene tres categorías;

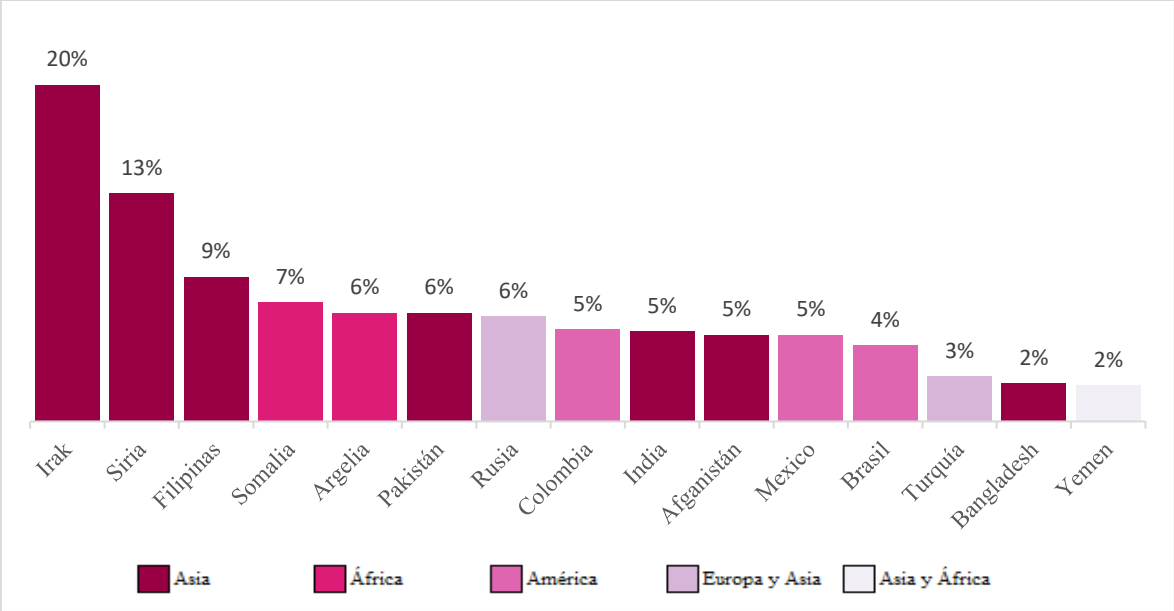
a. Fuego cruzado o combate. Identifica la causa de la muerte en un contexto militar o de guerra.
b. Asignación peligrosa. Refiere a toda muerte cuya causa es motivada por cubrir eventos que por su naturaleza representan un riesgo para el periodista como pueden ser, por mencionar algunos, disturbios, enfrentamientos entre grupos o manifestaciones.
c. Asesinato. Se consideran dentro de esta categoría las muertes cuyo móvil de la acción fue la labor del periodista, pudiendo ser el asesinato premeditado o no.

3.1.3. Análisis exploratorio de datos.

Como se mencionó con anterioridad, la base datos cuenta con 1,337 registros. Considerando que en 15 países se registra el 70% de las muertes y a fin de garantizar una buena asociación en la prueba de tablas y representación gráfica, se tomaron en cuenta las 15 primeras posiciones del listado, quedando así un total de 941 registros. Cabe señalar que el resto de los países pueden retomarse dentro del análisis de resultados.

En el Gráfico 8, se puede observar que Irak, Siria y Filipinas, reportan el mayor número de muertes con poco más del 40%. Turquía, Bangladesh y Yemen se ubican en los últimos tres lugares y representan casi el 10%. El mayor número de países se ubican en el continente asiático con casi el 60% de las muertes. El continente americano representa poco más del 15% y los tres países de la región: Colombia, México y Brasil, pertenecen a países latinoamericanos.

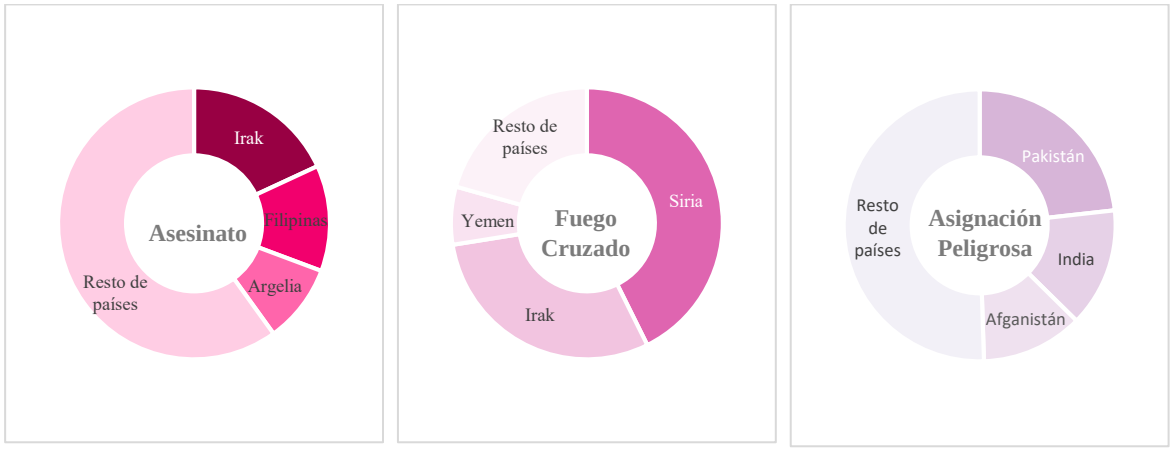
Gráfico 8. Países con mayor número de muertes de periodistas (1994 - 2019)



Del total de muertes por asesinatos, Irak, Filipinas y Argelia, representan el 40%, en cuanto a las muertes a causa de fuego cruzado, Siria representa el 43% de todas ellas. Finalmente, en muertes por asignación peligrosa Pakistán, India y Afganistán suman el 49% (Gráfico 9). Para el caso particular,

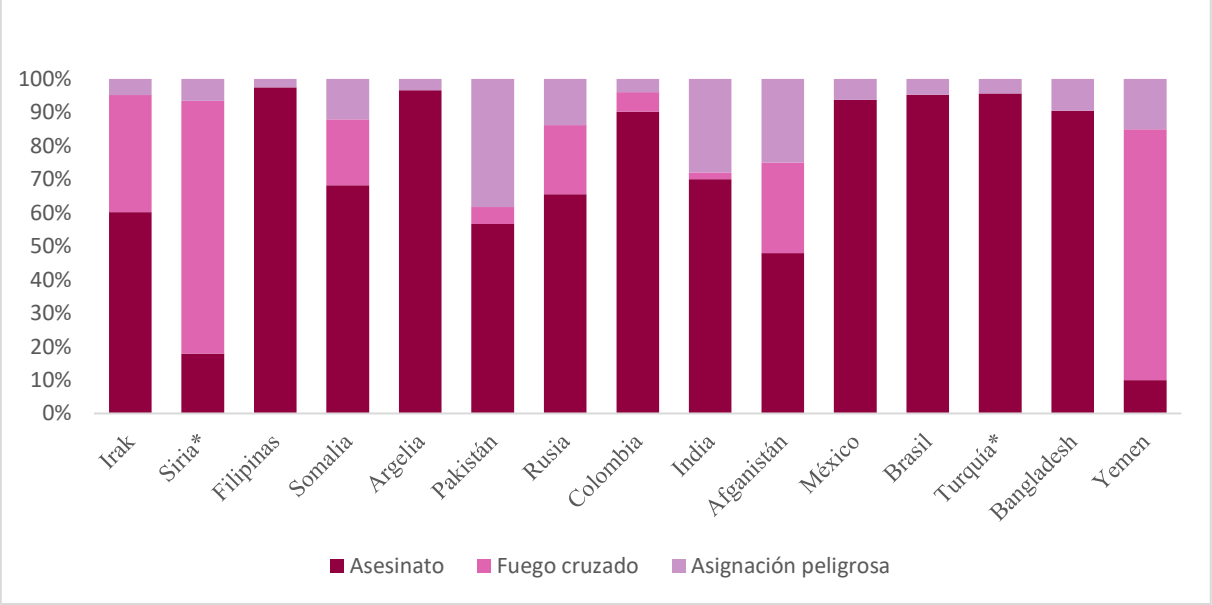
México contribuye con 7% a muertes por asesinato, 3% a muertes por asignación peligrosa y carece de registros en cuanto a muertes por fuego cruzado.

Gráfico 9: Causas de muerte



De cada 100 muertes de periodistas, 60 son por asesinatos, 20 a causa de fuego cruzado y 10 por asignación peligrosa. En casi todos los países el mayor porcentaje de muerte es por asesinato, a excepción de Siria y Yemen, cuyo porcentaje (>70%) se concentra en fuego cruzado. En países como Filipinas, Argelia, Colombia, México, Brasil y Bangladesh, el 90% o más de las causas de muerte son por asesinatos, en otros, como es el caso de Somalia, Pakistán, Rusia y Afganistán, el porcentaje de las causas se distribuye de manera más equitativa (Figura 51).

Gráfico 70: Causa de muerte por país



* El porcentaje restante corresponde a no especificado

3.1.4. Análisis de tablas previo al ACS.

Para el análisis se consideraron dos variables, país y causa de muerte, la primera consideró 15 registros correspondientes a los principales países con mayor frecuencia de muertes, la muestra seleccionada corresponde al 70% de los casos reportados por parte de la CPJ. En la variable causa de muerte, sus categorías están representadas por "Fuego cruzado o combate", "Asignación peligrosa" y "Asesinato". El análisis de los datos pretende identificar el grupo de países que pueden ser comparados con México a partir de la cualidad de la variable. Los datos presentados en la tabla de contingencia tienen un referente espacial asociado al lugar y la cualidad de ocurrencia del objeto bajo estudio.

La tabla 21, está conformada por i =15 renglones y j = 3 columnas, para nuestra matriz, una representación apropiada asociada con el total de varianza explicada tendría que hacerse en un espacio de dos dimensiones ($mín\{I, J\} - 1 = min\{15,3\} - 1 = 2$).

Tabla 21: Periodistas asesinados por causa de muerte y país de ocurrencia (1994 - 2019)

	Causa de muerte			
País	Asesinato	Asignación peligrosa	Fuego cruzado	Total general
Irak	112	9	65	**186**
Siria	22	8	93	**123**
Filipinas	78	2	0	**80**
Somalia	45	8	13	**66**
Argelia	58	2	0	**60**
Pakistán	34	23	3	**60**
Rusia	38	8	12	**58**
Colombia	46	2	3	**51**
India	35	14	1	**50**
Afganistán	23	12	13	**48**
México	45	3	0	**48**
Brasil	40	2	0	**42**
Turquía	22	1	0	**23**
Bangladesh	19	2	0	**21**
Yemen	2	3	15	20
Total general	619	99	218	**936**

El juego de hipótesis a probar establece que:

$H_0 =$ La causa de muerte de periodistas no está relacionada con el país de ocurrencia.

$H_1 =$ La causa de muerte de periodistas si está relacionado con el país de ocurrencia.

El valor de la prueba Chi- cuadrada de Pearson es de X^2 =435.870 con 28 grados de libertad. El valor teórico de la Chi- cuadrada utilizando un $\propto = 0.05$ se establece en 41.337, por lo que se puede afirmar con un 95% de confianza que la causa de muerte de periodistas si está relacionada con el país de ocurrencia (Tabla 22). Adicional a ello, los valores del coeficiente de contingencia C y la V de Cramer son de .564 y .483 los cuales indican un grado de asociación entre las variables. El 74.3% de la varianza explicada se concentra en la primera dimensión y el 25.7% en la segunda, de manera que la varianza se explica fielmente en las dos primeras dimensiones. El valor de la inercia de .466 refiere a una asociación significativamente mayor entre los perfiles de los renglones y las columnas.

Tabla 22: Tabla resumen periodistas

Dimensión	Valor singular	Inercia	Chi cuadrado	Sig.	Proporción de inercia		Valor singular de confianza	
					Contabilizado para	Acumulado	Desviación estándar	Correlación 2
1	.588	.346			.743	.743	.025	.048
2	.346	.120			.257	1.000	.041	
Total		.466	435.870	.000[a]	1.000	1.000		

a. 28 grados de libertad

3.1.5. Resultados del ACS.

En el gráfico de puntos fila (Gráfico 11) se puede apreciar que en la medida en que se avanza hacia la izquierda sobre la dimensión 1, se agrupa un conjunto bien definido de países. Siria es quien mayor aporta a la dimensión 1 y Yemen y Pakistán a la dimensión 2. En cuanto al gráfico de puntos columna se puede observar que la causa de muerte por fuego cruzado aporta mayor inercia a la dimensión 1 y por asignación peligrosa a la dimensión 2 (Gráfico 12).

Gráfico 11: Puntos de fila para país

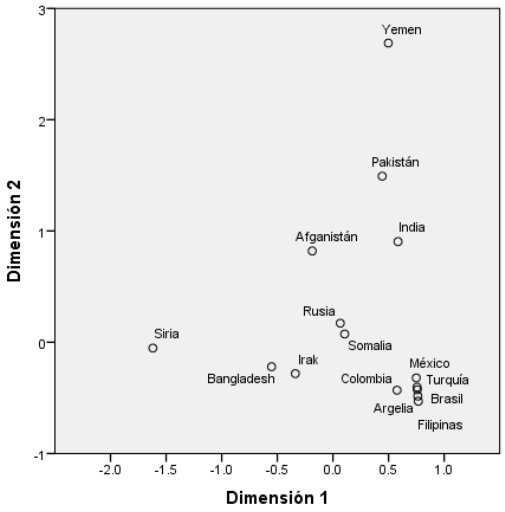
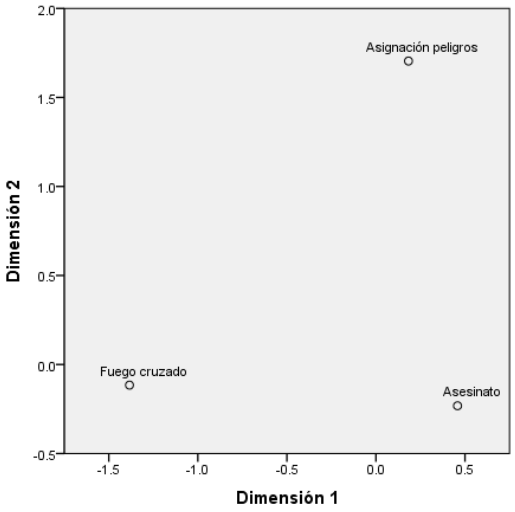


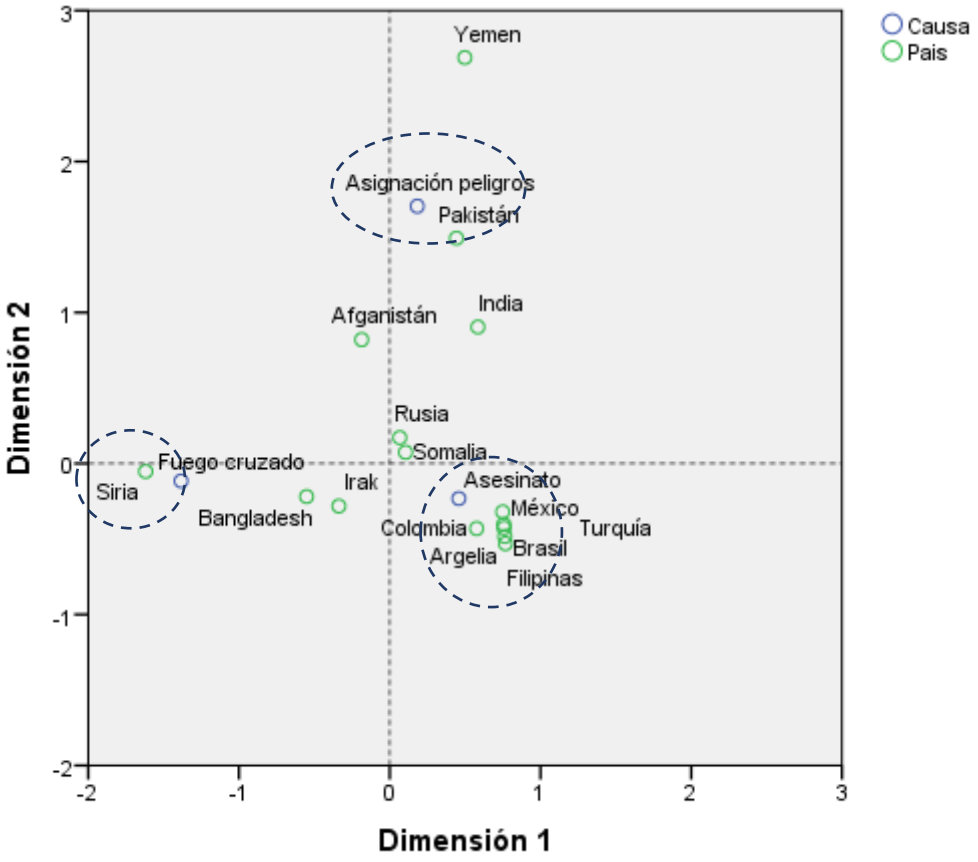
Gráfico 12: Puntos de columna para causas



En el Gráfico 13 se muestra que Siria se relaciona con fuego cruzado, Pakistán con asignación peligrosa y Colombia, México, Brasil, Argelia, Turquía y Filipinas con asesinato. Entre los tres grupos, hay mayor distancia entre asignación peligrosa y fuego cruzado. Afganistán se encuentra entre fuego cruzado y asignación peligrosa y Rusia entre fuego cruzado y asesinato. Se puede apreciar que en el tercer conglomerado se agrupan de manera más cercana un conjunto de países.

En el mapa perceptual Afganistán y Siria se alejan de México, partiendo del razonamiento inicial, se puede concluir que, tomando como referencia la causa de muerte de los periodistas, México no es comparable en término cualitativos con estos dos países, sin embargo, lo anterior no desestima la afirmación de que en términos cuantitativos, las cifras de periodistas asesinados en México son comparables con países en Guerra y que la muerte de periodistas en México está asociada con un tipo particular de violencia vinculada al narcotráfico y al crimen organizado.

Gráfico 13: Gráfico de puntos fila y puntos columna, periodistas



3.2. GEOGRAFÍA DE LA POBLACIÓN.

"MIGRACIÓN DE JORNALEROS AGRÍCOLAS EN MÉXICO"

3.2.1. Planteamiento.

La política de sustitución de importaciones inspiró un patrón de desarrollo heterogéneo y desigual en el espacio rural, entre 1940 y 1960 la industrialización de la agricultura se dio de manera diferenciada y en el territorio, éste contraste se expresó en el desarrollo de algunas zonas y el estancamiento de otras. Para 1970 y ante el desgaste del modelo de sustitución de importaciones, se comenzó a gestar una fase caracterizada por la apertura global y el libre mercado. Los grandes capitales agrícolas, localizados principalmente el norte de país y cuya ubicación resultaba estratégica debido a la cercanía con los grandes centros de consumo, fueron favorecidos por la inversión estatal y la inyección de capital internacional. La fase agroexportadora neoliberal se caracterizó por la apertura de las fronteras

nacionales, la producción de productos agrícolas para la exportación, la incorporación de nuevas tecnologías, la transición de una agricultura temporal hacía una de riego y la demanda masiva de mano de obra (Velazco, 2014, Rojas,2009).

Rubio (2002) identifica en esta fase una forma particular de subordinación excluyente de los pequeños campesinos y medianos empresarios. Es subordinante, porque obstaculiza la reproducción de las formas tradicionales de la agricultura y, excluyente, porque habiéndose transferido la gestión agrícola responsabilidad del Estado hacía las agroindustrias, un sector específico de la población queda relegado del proceso de producción de alimentos básicos. Para el sector excluido, las consecuencias se expresan en la pauperización de la actividad agrícola en pequeña escala, la disminución de la producción nacional, el debilitamiento de la soberanía alimentaria, la profundización de la dependencia alimentaria y el empobrecimiento en general de las comunidades rurales campesinas.

A partir de la función de cada región en la dinámica agrícola global, la integración de cada una de ellas se hace de manera diferenciada. Las asimetrías geográficas definidas por el modelo económico neoliberal colocan en desventaja a las economías agrícolas tradicionales, la profundización de la precariedad en algunas regiones y la demanda de mano de obra para satisfacer los requerimientos globales en otras, propicia la migración de la población campesina hacía regiones de mayor desarrollo económico (Tlachinollan, 2015). En este proceso, los trabajadores rurales fungen como reserva móvil de capital humano y son sometidos a estructuras específicas de explotación laboral, Rojas (2017) define este tipo de migración como un síntoma de la demanda de acumulación de capital del modelo neoliberal (Velazco, 2014).

En México la migración de jornaleros agrícolas cobró importancia a comienzos de 1970, a partir de entonces, se comenzó a configurar un patrón migratorio específico. La configuración de este respondió a las demandas productivas del sector agrícola producto de la oferta y demanda de la fuerza de trabajo, es decir, por una parte, se presentó la necesidad imperante de mano de obra masiva que fuera capaz de atender las exigencias del mercado global y, de manera paralela, la exacerbada necesidad de empleo de la población migrante, de esta manera, las regiones involucradas en el proceso se caracterizaron por especificidades definidas por la dinámica y por el tipo de producción de las unidades agro productivas (Rojas, 2017). La demanda de mano de obra migrante varía de acuerdo a la región, esta puede ser local, regional o intra regional. En la primera, los trabajadores provienen de la misma región productiva, en la segunda, son originarios de regiones colindantes y su estancia es estacionaria, finalmente, en la tercera, la población proviene de otros estados y el periodo de estancia se prolonga por meses.

En un intento por definir un mapa de la migración de jornaleros agrícolas en México, Rojas (2017) identifica tres regiones cuyo desarrollo agrícola permite su diferenciación:

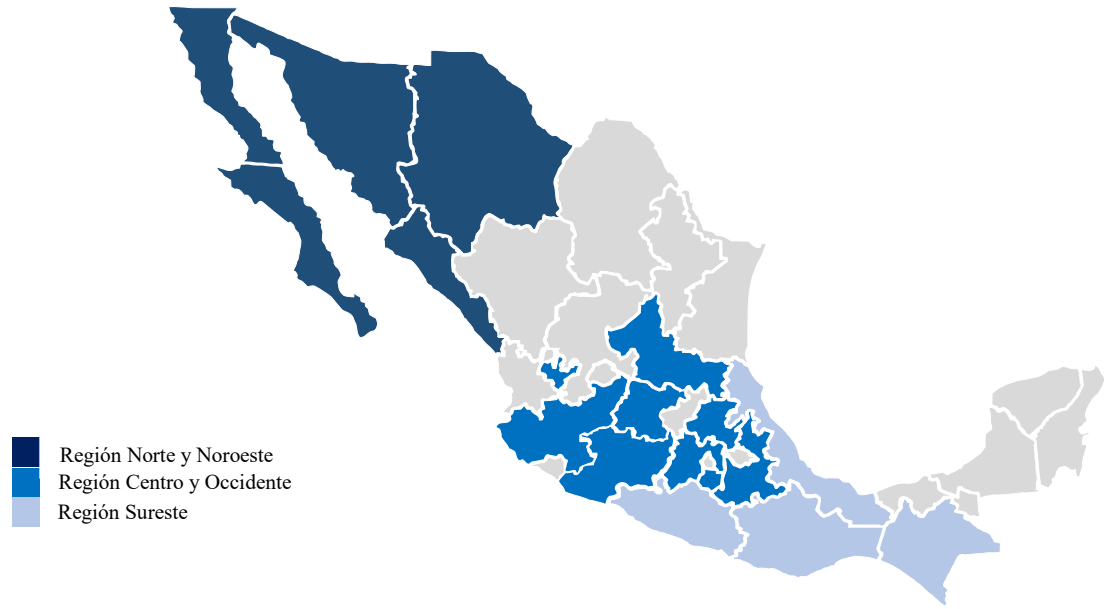
1. Norte y Noroeste. En ella se localizan los principales polos productivo agrícolas, cuenta con condiciones propicias para el desarrollo de la agricultura, la producción que aquí se desarrolla se vincula a la exportación, son espacios de inversión rentable de capital, cuenta con infraestructura, tecnificación de los procesos productivos, demanda de mano de obra barata y flexible, son zonas de atracción y receptoras de migrantes. Los estados más representativos de esta región son Sinaloa, Sonora, Baja California, Baja California Sur y Chihuahua.
2. Centro y Occidente. En esta zona se desarrollan principalmente actividades ligadas a la industria y al sector terciario. En las principales zonas productoras a nivel nacional, como Michoacán y Jalisco, la agricultura se desarrolla principalmente en unidades agrícolas de superficie limitada y son medianos o pequeños productores quienes realizan la actividad. El mercado de los productos de esta región se destina principalmente al nacional y al regional. En términos migratorios, es una zona intermedia o mixta pues se caracteriza por atraer y expulsar mano de obra cuya demanda se cubre con población migrante local y regional. Los estados más representativos son Michoacán, Guanajuato, San Luis Potosí, Jalisco, Puebla, Hidalgo, Morelos y el Estado de México.
3. Sureste. La agricultura que en esta región se desarrolla es de subsistencia o autoconsumo, su producción se dirige principalmente al mercado local. Las características del relieve y orográficas de esta región son poco aptas para el desarrollo de esta actividad, una importante proporción de los suelos se encuentra erosionada y no se cuenta con una adecuada infraestructura ni red de insumos. Es una zona de expulsión o de origen de jornaleros agrícolas. Los estados más representativos son Guerrero, Oaxaca, Chiapas y Veracruz.

A partir de los movimientos recurrentes de la población, la Secretaria de Desarrollo Social (Sedesol, 2010) y Rojas (2017), han definido cuatro rutas migratorias.

1. Ruta del pacífico. Los jornaleros agrícolas se desplazan desde los estados de Oaxaca, Guerrero y Puebla hacía los estados de Sinaloa, Sonora, Baja California, Baja California Sur, Jalisco y Nayarit
2. Ruta del golfo: La población se desplazada de estados con altos índices de pobreza y bajos niveles de desarrollo como Veracruz, Hidalgo y Puebla y se dirige hacía Tamaulipas y Nuevo León
3. Ruta centro. Los movimientos son inter regionales y se presentan en los estados de San Luis Potosí, Guanajuato, Zacatecas, Durango, Coahuila y Chihuahua.

4. Ruta del sureste. Al igual que la anterior, los movimientos son inter regionales y se presentan en los estados Oaxaca, Chiapas, Tabasco, Campeche y Yucatán.

Figura 21: Regiones agrícolas de México



Elaboración propia con base en los resultados obtenidos

En los lugares de origen, la migración de jornaleros agrícolas impacta de distinta manera la estructura social y territorial, los movimientos constantes de la población entre el lugar de origen y los centros de trabajo, transforman el tejido comunitario y familiar, erosionan la cohesión social y propician el desarraigo; en cuanto al territorio se refiere, este fenómeno moldea patrones específicos de asentamientos tanto en los lugares de partida como en los de destino. En estos últimos, las condiciones laborales en las que se emplea la población migrante suelen ser precarias, es común el trabajo infantil, las largas y extenuantes jornadas de trabajo, el endeudamiento en las "tiendas de raya", las enfermedades por la exposición prolongada a sustancias toxicas, el hacinamiento y la falta de servicios en los campamentos (Rojas,2017).

Los migrantes agrícolas son, dentro de la dinámica global, un sujeto desechable que puede reemplazarse con facilidad, la migración se convierte en una extensión de la pobreza que viven los jornaleros en sus lugares de origen y que, en los de destino, las condiciones de precariedad de vida y laboral se profundizan. El discurso de la pobreza coloca a la explotación de los trabajadores rurales en segundo término y sobrepone el "alivio" que la migración representa para subsanarla negando que, sin ellos la producción masiva de productos agrícolas no sería posible (Rubio, 2002, Salinas, 2012).

Mirar la migración desde una perspectiva regional permite identificar las problemáticas, necesidades y áreas de oportunidad e intervención tanto en los lugares de destino como en los de origen; un entendimiento adecuado del fenómeno nos permite atender todas las aristas que del fenómeno se desprenden. Como punto de partida se pretende identificar el patrón espacial entre los estados expulsores y las zonas receptoras de la migración agrícola.

3.2.2. Fuente de los datos.

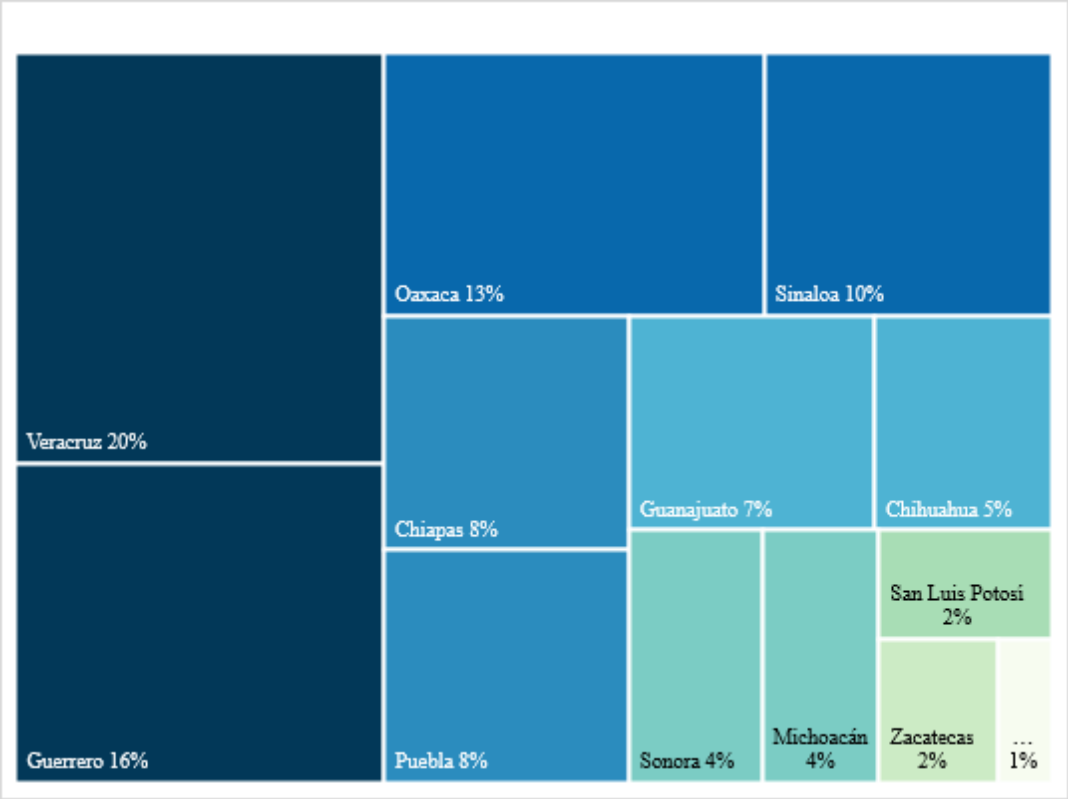
Las estadísticas agrícolas en México, en comparación con otras como las económicas o las sociales, presentan cierto rezago en su avance. Los datos se obtuvieron de la Encuesta Nacional de Jornaleros Agrícolas (ENJO) realizada en 2009 por la entonces Secretaría de Desarrollo Social (Sedesol) y la Universidad Autónoma Chapingo (UACh), cabe destacar que desde la presentación de la encuesta al desarrollo de éste documento, no existe otro ejercicio similar, de ahí que, sin bien los datos tienen una antigüedad de 10 años, por su nivel de desagregación, estos nos permiten mostrar un panorama del patrón espacial de la migración en ese periodo de tiempo.

La ENJO define a un jornalero agrícola como el individuo que se emplea en actividades vinculadas con la agricultura de manera temporal o permanente y por lo cual recibe una remuneración a cambio. Una unidad empleadora refiere al sujeto físico o moral que emplea al jornalero agrícola y cuyo giro puede dirigirse hacia la producción, comercialización, empaque o contratación. La presente encuesta se estructura en dos cuestionarios, el primero tiene como unidad de análisis a los jornaleros agrícolas y el segundo, a las unidades empleadoras, a partir de la disponibilidad de la información, los datos se tomaron del segundo cuestionario. Las variables que se consideraron para el análisis son; lugar de origen de los jornaleros temporales y ubicación de las unidades empleadoras. La unidad de medida de ambas variables fue el número de jornaleros agrícolas por entidad federativa. Se contó con 134,454 registros, sin embargo, para la primera variable solo se consideraron los 13 principales estados expulsores de migrantes y, para la segunda, los 17 estados que conforman las tres regiones establecidas por Rojas (2017. De esta manera, se conservó una muestra de 115,195 registros, lo cual representa el 82% del total de la información.

3.2.3. Análisis exploratorio de los datos.

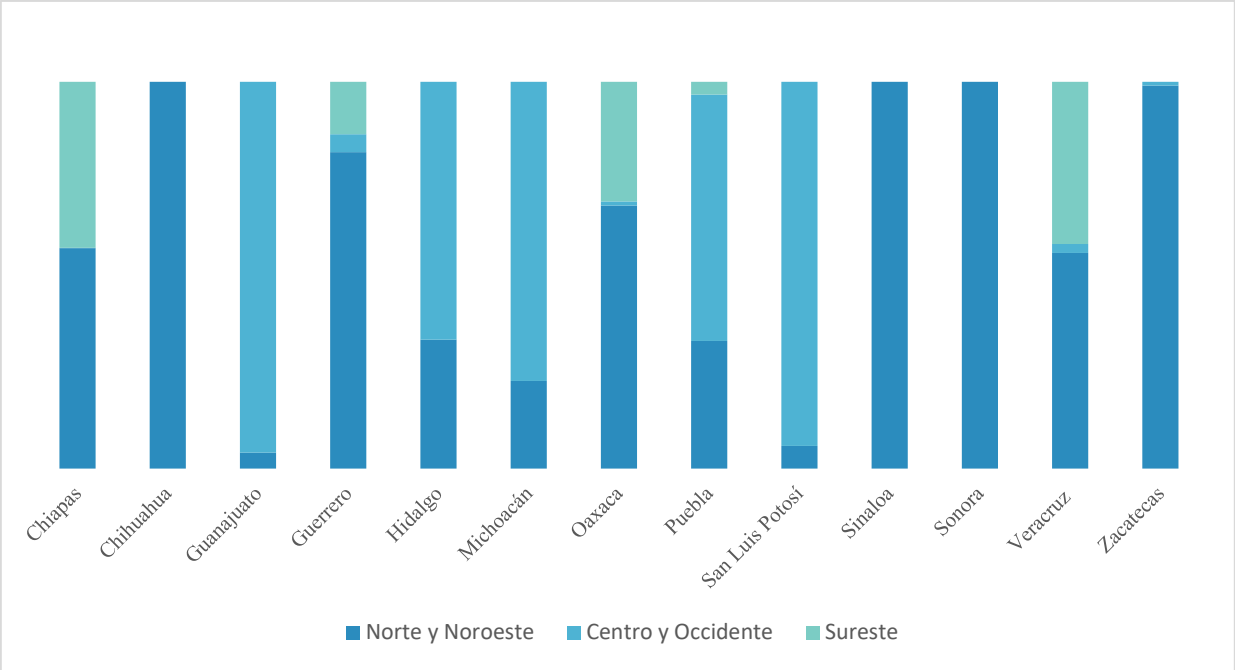
Casi el 50% de los jornaleros agrícolas que se ocupan en las distintas unidades empleadoras del país provienen de los estados de Veracruz, Guerrero y Oaxaca, el 51.2% restante se distribuye entre las otras diez entidades (Gráfico 14).

Gráfico 14: Porcentaje de procedencia de los jornaleros agrícolas



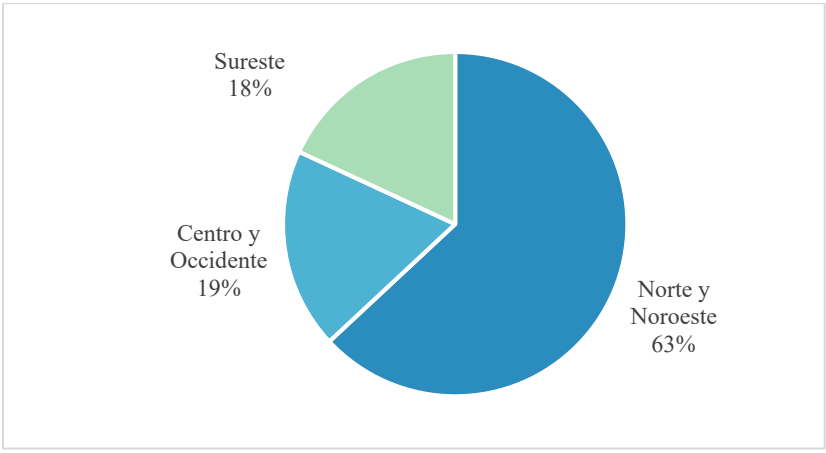
De acuerdo con la dinámica interna del fenómeno, los migrantes tienen a conformar un patrón que se establece a partir del lugar de origen y la región de destino. Por ejemplo, los estados de Sinaloa, Sonora, Chihuahua, Zacatecas y Guerrero tienden a migran hacía la zona Norte y Noroeste; los estados de Guanajuato, Hidalgo y Michoacán hacía la zona centro y occidente y las entidades de Chiapas y Veracruz lo hacen hacía la zona norte y el sureste del país (Gráfico 15).

Gráfico 15: Región de destino según entidad de origen



De acuerdo con la frecuencia con que ocurre la migración, el 63.1% se dirige hacia la región Norte y Noroeste y, en proporción similar, hacía la zona Centro y Occidente y Sureste con 18.9% y 18.1% respectivamente.

Gráfico 16: Región de destino según entidad de origen



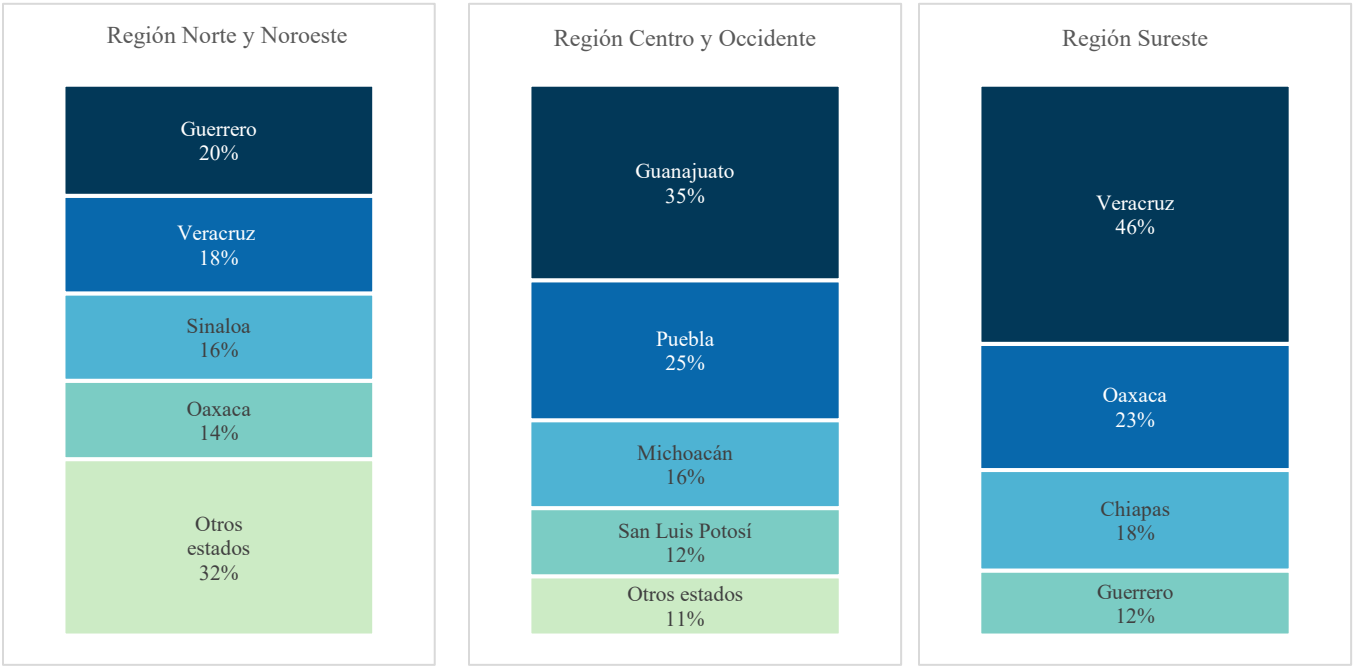
Los datos muestran un patrón en la dinámica migratoria que marca la relación entre el lugar de origen y la región de destino. Casi el 70% de la población migrante jornalera en la región Norte y Noreste proviene de los estados de Guerrero, Veracruz, Sinaloa y Oaxaca. La presencia de jornaleros cuyo lugar de origen se encuentra en los lugares del occidente o centro como Puebla, Zacatecas,

Michoacán, Hidalgo, Guanajuato y San Luis Potosí solo representa el 10%. Para esta región se identifican al menos 13 entidades de procedencia de los migrantes.

En lo que refiere a la región Centro y Occidente, casi el 90% de los jornaleros proviene de los estados de Guanajuato, Puebla, Michoacán y San Luis Potosí. Cabe destacar que no hay ningún jornalero procedente de los estados de Chihuahua, Sinaloa y Sonora pertenecientes a la región norte - noroeste. En comparación con la región anterior, en esta se identifican 9 estados de procedencia de los migrantes.

Finalmente, el 98% de la población migrante jornalera de la región sureste proviene de Veracruz, Chiapas, Oaxaca y Guerrero, en un porcentaje mínimo de Puebla y sin registros para el resto de los estados del norte como Sonora o Sinaloa o del centro y occidente como Guanajuato o San Luis Potosí. En comparación con las dos regiones anteriores, en ésta solo se identifican 4 estados de origen de los migrantes. Los datos anteriores comienzan a dar un panorama del patrón de la migración jornalera en México que podrá verse más claramente en el análisis de correspondencias.

Gráfico 17: Principales estados de procedencia por región



Nota. El porcentaje restante para la región Sureste corresponde al estado de Puebla

3.2.4. Análisis de tablas previo al ACS.

El análisis consideró las variables; lugar de origen y región de destino de los jornaleros, la primera consideró 13 entidades correspondientes al 82% de la base de datos. La variable región de destino se conforma por tres categorías; Región Norte y Noroeste, Región Centro y Occidente y Región Sureste.

El análisis de los datos tiene por objetivo identificar el patrón espacial entre los estados expulsores y las regiones receptoras. Ambas variables tienen un referente espacial que se vincula a la dinámica migratoria del origen y destino de los migrantes.

La tabla de contingencia está conformada por i= 13 renglones y j=3 columnas. Siguiendo la lógica trazada, el total de varianza explicada queda bien representado en un espacio de dos dimensiones (mín{I,J}-1=min{13,3}-1=2).

Tabla 23: Jornaleros agrícolas temporales en unidades empleadoras según estado de origen y región de destino para el año 2009

	Región de destino			
Estado de origen	Norte y Noroeste	Centro y Occidente	Sureste	**Total General**
Chiapas	4,742	3,569	5	**8,316**
Chihuahua	5,529	0	0	**5,529**
Guanajuato	315	0	7,349	**7,664**
Guerrero	13,979	2,316	794	**17,089**
Hidalgo	379	0	757	**1,136**
Michoacán	966	0	3,299	**4,265**
Oaxaca	9,869	4,491	146	**14,506**
Puebla	2,724	269	5,262	**8,255**
San Luis Potosí	160	0	2,560	**2,720**
Sinaloa	10,975	0	0	**10,975**
Sonora	4,888	0	0	**4,888**
Veracruz	12,260	9,216	505	**21,981**
Zacatecas	2,450	0	25	**2,475**
Total General	**69,236**	**19,861**	**20,702**	**109,799**

El juego de hipótesis que se establece a partir de la tabla anterior es el siguiente:

$H_0 =$ El estado de origen de los migrantes jornaleros no está relacionado con la región de destino.

$H_1 =$ El estado de origen de los migrantes jornaleros si está relacionado con la región de destino.

El valor de la prueba Chi- cuadrada de Pearson es de X^2 =97, 615.30 con 24 grados de libertad. El valor teórico de la Chi- cuadrada con un $\propto = 0.05$ se establece en 36.415. Los datos anteriores son evidencia suficiente para afirmar que el estado de origen de los migrantes jornaleros si está relacionado con la región de destino (Tabla 24). Los valores del coeficiente de contingencia C= .686 y la V de Cramer= 667 indican una buena asociación entre las variables. El 80.7% de la varianza se

explica en la primera dimensión y el resto en la segunda. Destaca que la inercia = .889 indica una asociación significativa entre los perfiles de renglón y los de columna.

Tabla 24: Tabla resumen

Dimensión	Valor singular	Inercia	Chi cuadrado	Sig.	Proporción de inercia		Valor singular de confianza	
					Contabilizado para	Acumulado	Desviación estándar	Correlación 2
1	.847	.717			.807	.807	.002	.039
2	.415	.172			.193	1.000	.002	
Total		.889	97615.300	.000[a]	1.000	1.000		

a 24 grados de libertad

3.2.5. Resultados del ACS.

En la Gráfico 18 se puede observar que sobre la dimensión 2 se agrupa un conglomerado definido por San Luis Potosí, Guanajuato, Michoacán, Hidalgo y Puebla. Sobre la dimensión 1 se distingue más claramente la agrupación de los estados de Sinaloa, Sonora, Chihuahua y Zacatecas; Chiapas, Oaxaca y Veracruz conforman otro grupo más disperso, finalmente, se aprecia que Guerrero se ubica entre el grupo dos y tres. En la Gráfico 19, se puede observar que las tres regiones se distribuyen de manera dispersa sobre el espacio, la región centro y occidente aporta mayor inercia a la dimensión 1 y la región sureste a la dimensión 2.

Gráfico 18: Puntos de fila jornaleros

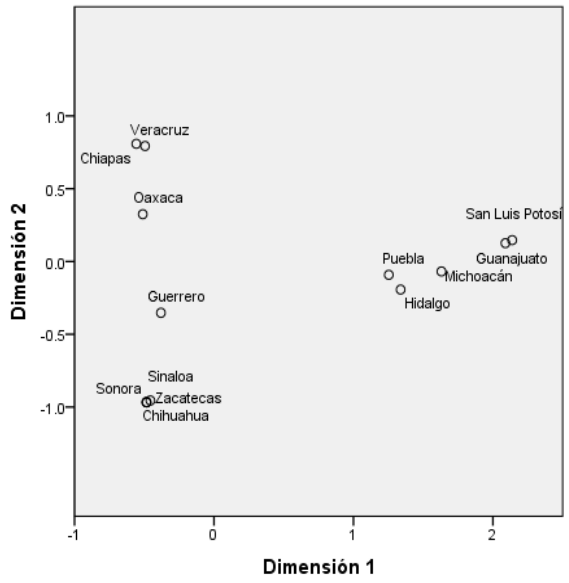


Gráfico 19: Puntos de columna jornaleros

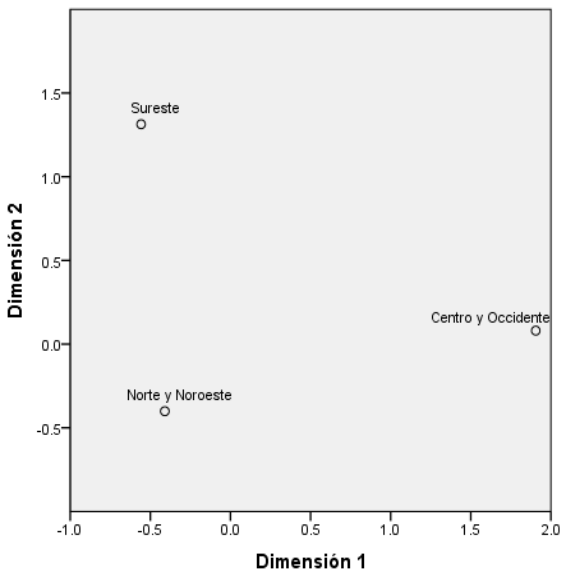
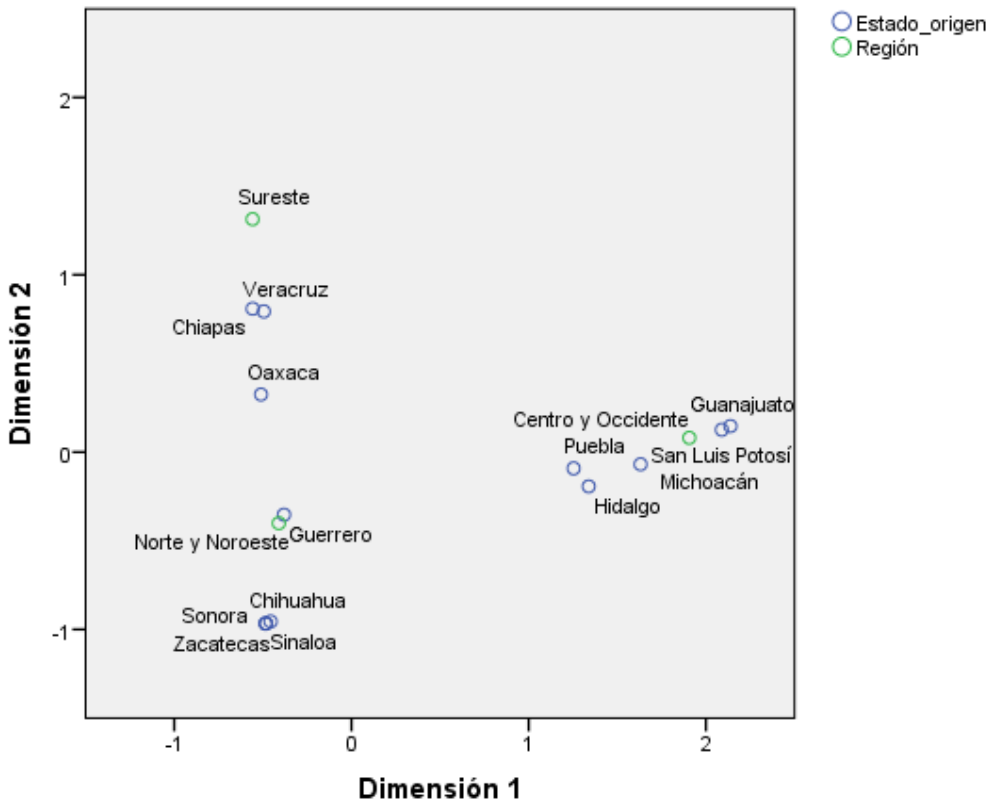


Gráfico 20: Mapa perceptual, jornaleros

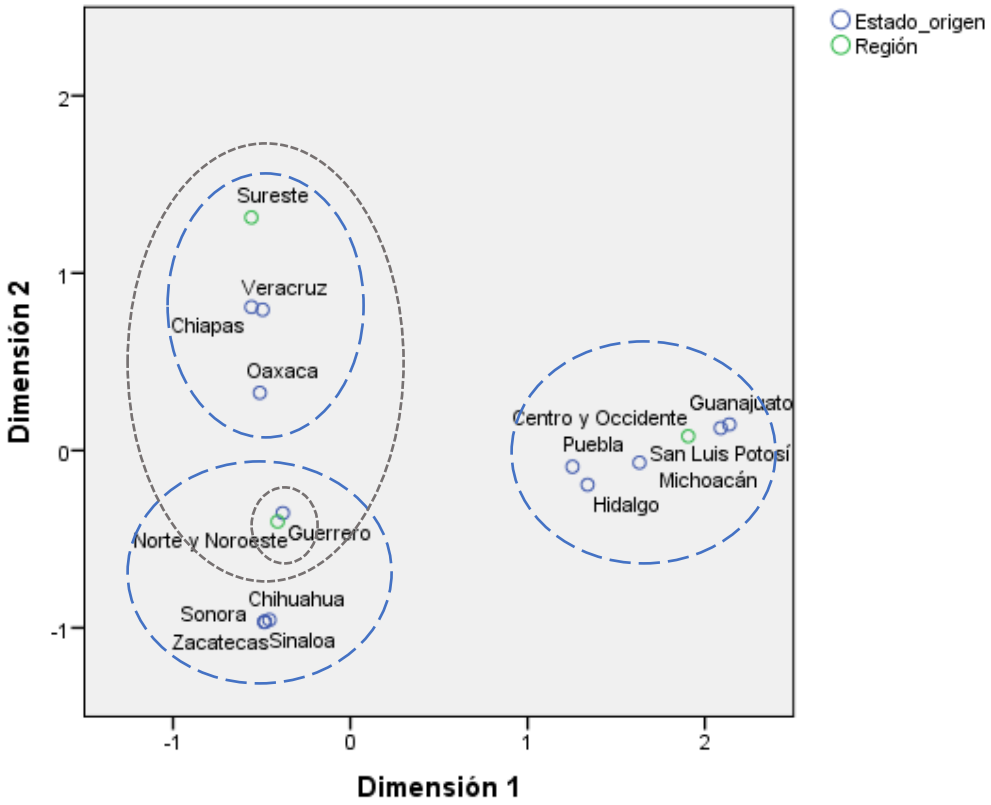


En el gráfico del ACS (Gráfico 21) se puede apreciar que la región Centro y Occidente atrae en su mayoría a población de la misma región: Guanajuato, Michoacán, San Luis Potosí, Hidalgo y Puebla. La región sureste atrae población de Veracruz, Chiapas y Oaxaca, a diferencia con la región anterior, las entidades se encuentran más dispersas. Esta dispersión se explica por la tendencia de atracción de estas entidades, hacía la región Norte y Noroeste. Por ejemplo, en el gráfico podemos observar que la población originaria de Veracruz y Chiapas tiene un patrón de migración que se dirige principalmente hacía la región Sureste y en posterior, a la región Norte y Noroeste, por otro lado, la población de jornaleros agrícolas originarios de Oaxaca, tienden a migrar de manera equilibrada, hacía la región Sureste y Norte y Noroeste.

Para la región Norte y Noroeste se observa que la mayor población de jornaleros agrícolas proviene del estado de Guerrero y, en segundo lugar, de un grupo de entidades conformado por Sonora, Chihuahua, Sinaloa y Sonora. La presencia de población originaria de dos estados de sureste del país: Guerrero y Oaxaca, resalta un tipo de migración interregional, en donde, la demanda de mano de obra no es cubierta por la población local, estos movimientos de población se dirigen desde estados con altos índices de pobreza hacía regiones con mayor desarrollo económico vinculadas al mercado agrícola internacional. A comparación con las regiones Centro Occidente y Sureste, se aprecia que en éstas, la población de la región es suficiente para cubrir la demanda agro productiva, destacándose

que, en la primera, el número de entidades da cuenta de una migración regional y en la segunda, de una migración local.

Gráfico 21: Mapa perceptual



A diez años de la publicación de la Encuesta Nacional de Jornaleros Agrícolas (ENJO), en un reportaje especial publicado por el semanario Proceso; "Los Jornaleros agrícolas, invisibles para el gobierno", la titular de la recién creada Secretaría del Bienestar, declara que se desconoce quiénes son y de dónde vienen los jornaleros agrícolas. Diferentes sectores de la sociedad civil y la academia, así como la Comisión Nacional de los Derechos Humanos, coinciden en la necesidad de actualizar los datos y elaborar un diagnóstico sobre la situación de los jornaleros agrícolas. De acuerdo con un grupo de especialistas, la política social de la administración actual, ha excluido a este grupo de población que, por años, ha vivido en condiciones de vulnerabilidad y explotación laboral (Proceso, 2019b). El ejercicio realizado, permite identificar, de manera parcial, el patrón de la migración de los jornaleros agrícolas, un conocimiento profundo de quienes son los jornaleros, de dónde vienen, hacía donde van y las causas que motivan el proceso, es un eslabón fundamental para atender las problemáticas que del fenómeno se desprenden.

3.3. GEOGRAFÍA DE LOS RIESGOS.

OCURRENCIA DE FENÓMENOS HIDROMETEOROLÓGICOS Y GEOLÓGICOS EN MÉXICO.

3.3.1. Planteamiento.

El martes 19 de septiembre de 2017 a las 13:14 horas un sismo de magnitud 7.1 con epicentro en Axochiapan, Morelos sacudió el centro del país (Servicio Sismológico Nacional:2017). Las pérdidas materiales y económicas se contabilizaron en millones de pesos y las afectaciones humanas se registraron en miles de damnificados y más de un centenar de muertos. ¿Fue este movimiento telúrico el causante de todas estas pérdidas? En entrevista con un periódico local, Carlos Morales Cienfuegos, fundador de topos México USAR-BREC, declaró que; "En los desastres las estructuras se colapsan por corrupción y pobreza, no hay más. Más del 80% de lo que se derrumbó lo construyó el gobierno: hospitales, escuelas, edificios, etcétera." (Publimetro, 2017).

Desde fines del siglo anterior, Andrew Maskey (1993) afirmaba que los desastres no son naturales sino un producto de la construcción social del riesgo, los desastres son causados por fenómenos los cuales pueden ser de origen natural o antrópico y cuya ocurrencia sobre un sistema endeble puede causar afectaciones. Comúnmente se asocia un desastre con fuerzas naturales extraordinarias, sin embargo, éste tiene lugar por la ocurrencia de un fenómeno natural en determinadas condiciones de vulnerabilidad (físicas, socioeconómicas, o institucionales).

Los fenómenos naturales son originados por las variaciones en las condiciones ambientales o en la actividad geológica y por su tipo estos pueden ser (CENAPRED 2014):

a. Hidrometeorológicos. Se presentan con mayor frecuencia y son los fenómenos que causan mayores daños. De este tipo podemos encontrar ciclones tropicales, precipitaciones, tormentas, heladas, sequías, tornados y ondas cálidas y gélidas.
b. Geológicos: Se producen por movimientos internos en la corteza terrestre, aunque su nivel de ocurrencia puede ser menor, sus efectos pueden ser muy destructivos. Entre los fenómenos geológicos se encuentran los sismos, maremotos, vulcanismo, inestabilidad de laderas, hundimientos y agrietamientos.

En cuanto a los desastres originados por la actividad humana, éstos se clasifican en:

a. Químicos. Son provocados por fugas o derrames de sustancias químicas dañinas.

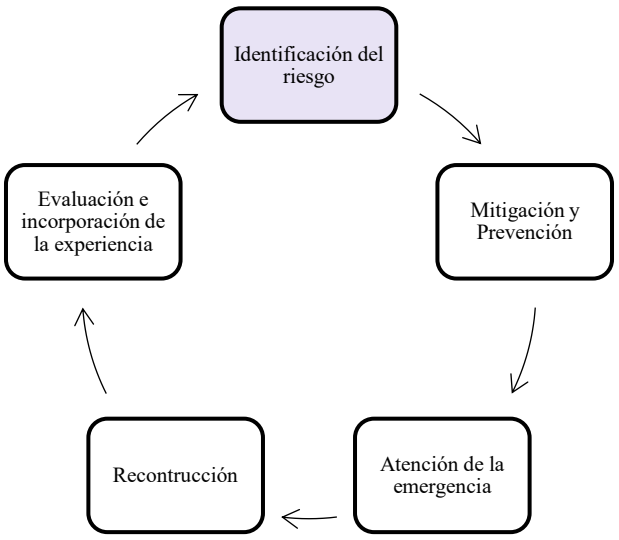
b. Sanitarios ecológicos. Son provocados por la contaminación del suelo, agua o aire y afectan el ambiente y la salud de los seres vivos.
c. Socio organizativos. Tienen su origen en errores humanos o en decisiones predeterminadas.

Cuando un fenómeno se presenta en situaciones de vulnerabilidad ocurre un desastre, estos según su impacto, puede causar pérdidas económicas o humanas, afectaciones en la estructura o alterar el equilibrio ambiental. El riesgo de que ocurra un desastre implica la probabilidad de ocurrencia de un fenómeno previsto o fortuito, en un sistema; se es vulnerable cuando se es susceptible a sufrir un daño, o bien, cuando el sistema no cuenta con los elementos necesarios para su recuperación. La vulnerabilidad puede construirse socialmente y, esta a su vez, crea las condiciones de riesgo, son las sociedades mismas quienes, por ignorancia o aberración, crean entornos pocos seguros para establecerse; solo a través de un conocimiento profundo sobre el funcionamiento del medio natural, y ante el hecho inminente de que las sociedades no tienen control sobre los fenómenos naturales, se puede reducir el riesgo de que ocurra un desastre (Maskrey, 1993).

La referencia espacial de un riesgo y del lugar de ocurrencia donde puede presentarse un fenómeno o desencadenarse un desastre, es fundamental para diseñar planes de protección civil y resarcir posibles afectaciones al medio natural y a la población en general. Por su referencia geográfica, México se encuentra ubicado en zonas de alta actividad sísmica y volcánica, así mismo, es susceptible a la presencia de fenómenos hidrometeorológicos que, según la época del año y las condiciones del sistema donde se presenten, pueden provocar severas afectaciones como inundaciones o deslaves (Guevara, Quaas y Fernández, 2006).

La prevención es un mecanismo adecuado para reducir el impacto de una amenaza. El ciclo de la prevención se compone de cinco eslabones. En el primer eslabón del ciclo, es de vital importancia conocer con certeza los peligros y amenazas bajo los cuales se encuentra una población y, de manera paralela, tener un conocimiento profundo sobre los niveles del riesgo, la identificación del sitio de ocurrencia y el impacto y alcance que pudiese tener un fenómeno (Guevara, Quaas y Fernández, 2006).

Figura 22: Ciclo de prevención

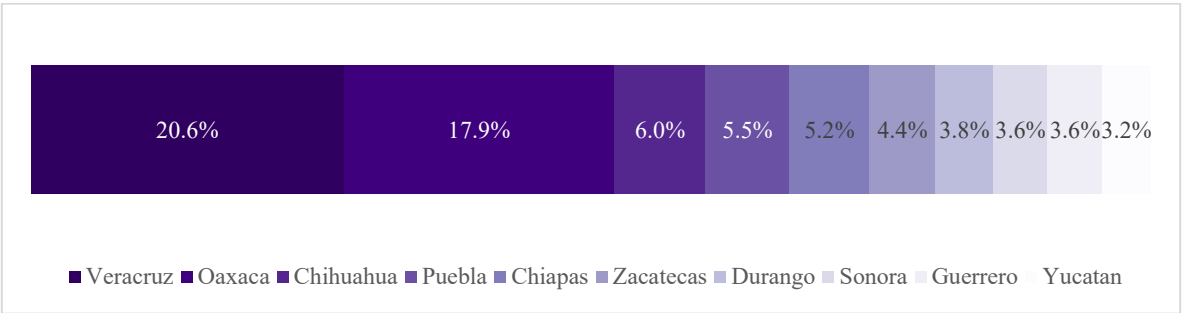


3.3.2. Fuente de la información.

Los datos que se utilizaron para el ACS se tomaron de la información tabular sobre declaratorias registradas en el periodo de 2000 y 2016 publicada por el Centro Nacional de Prevención de Desastres (CENAPRED, 2017a). La base de datos se conforma por 10 variables: estado, clave de la contingencia, municipio, fecha de publicación, fecha de inicio, fecha de término, tipo de declaratoria, tipo de fenómeno, clasificación y observaciones. Para el presente análisis se registró el conteo sobre las variables estado y tipo de fenómeno. Se cuentan con 22,786 registros.

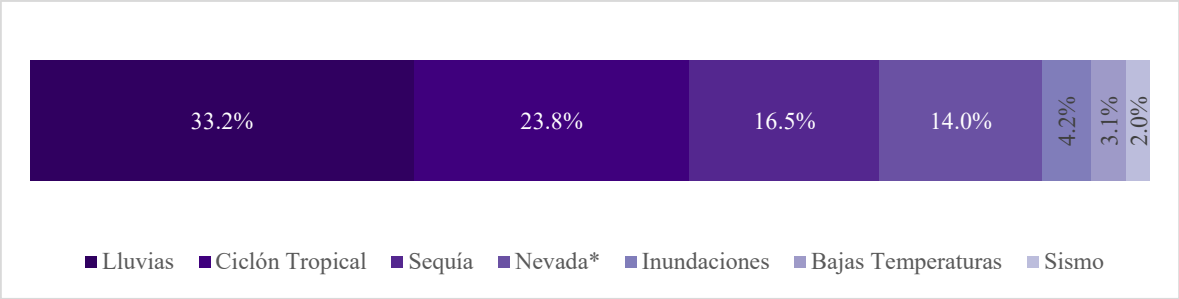
Siete de cada diez fenómenos que ocurren en el país se presentan en 10 entidades y nueve de cada diez fenómenos corresponden a 7 de los 15 fenómenos identificados en la base de datos (Gráfico 22). Por lo que, para garantizar una buena representación gráfica de los datos, se consideraron los 10 estados y los siete fenómenos principales quedando un total de 14,461 registros correspondiente al 63.46% del universo.

Gráfico 22: Principales estados



Nota. El 26.1 restante corresponde al resto de los estados

Gráfico 23: Principales fenómenos



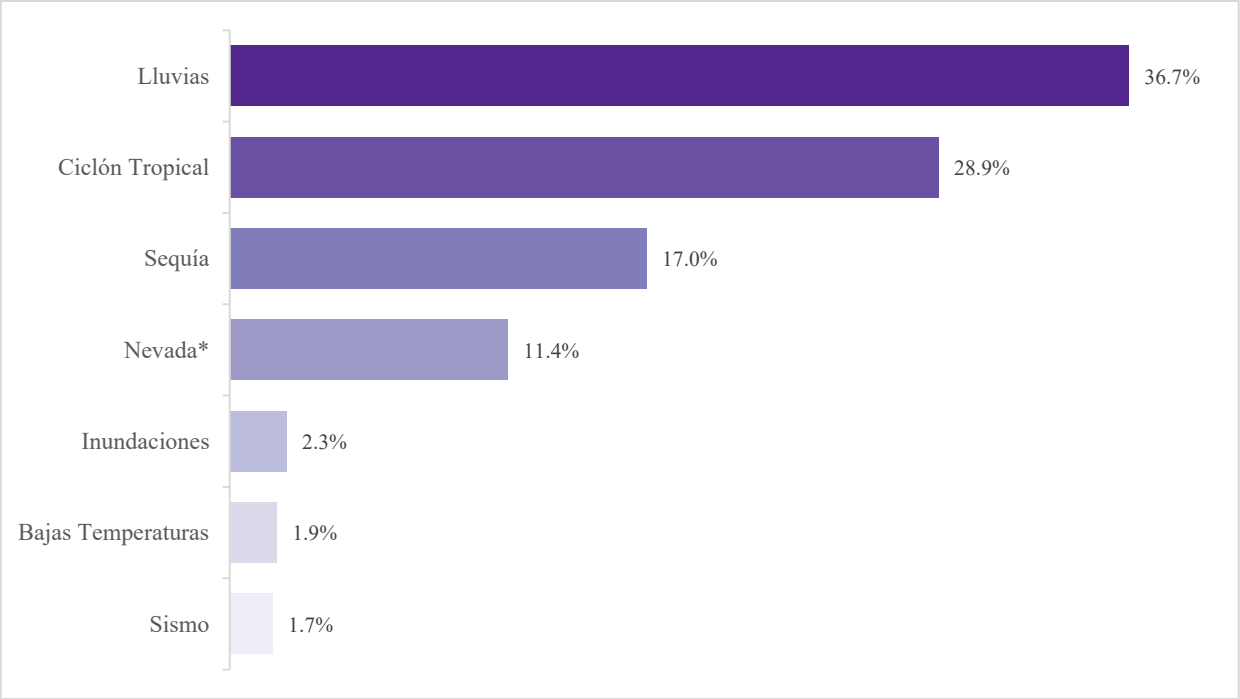
* Incluye Heladas y Granizadas

Nota. 3.2% restante corresponde a Deslave, Incendios forestales, Vientos, Marea Roja, Tormentas severas, Tornado, Erupción volcánica

3.3.3. Análisis exploratorio de los datos.

En la base de datos original se registran cerca de 15 fenómenos, de los cuales 7 de ellos representa el 96.8% de la frecuencia de ocurrencia. De este subconjunto, los principales fenómenos que tienen lugar en el territorio nacional son las lluvias con el 37% y los ciclones tropicales con el 29%, ambos registros, representan poco más del 65% del total; en lo que refiere a los fenómenos con menor porcentaje de ocurrencia, las inundaciones, bajas temperaturas y los sismos representan 6% del total de ocurrencia.

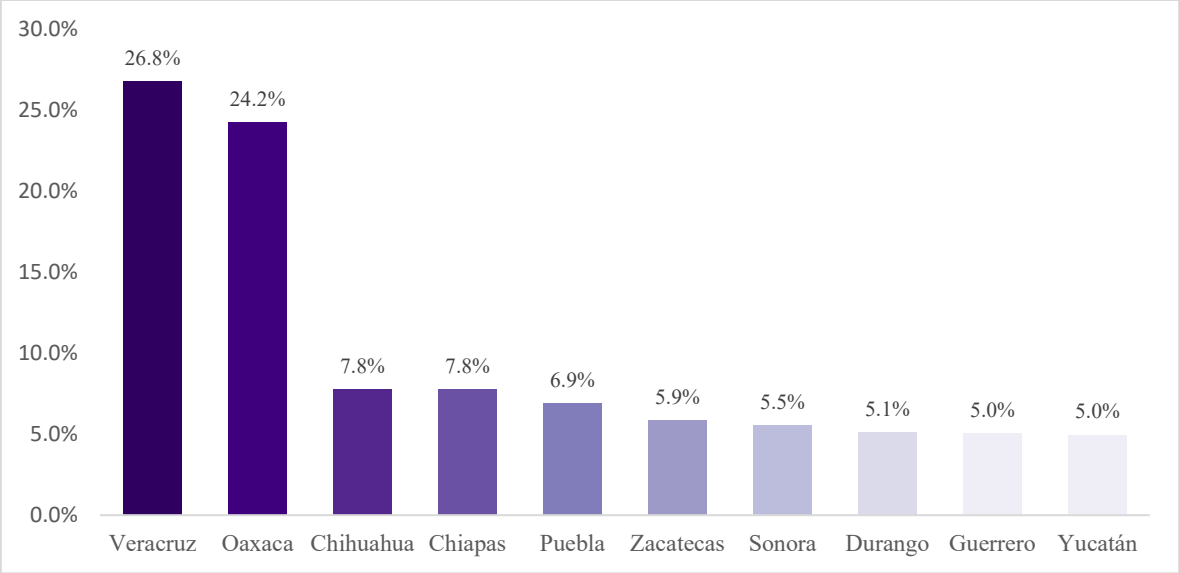
Gráfico 24: Porcentaje de ocurrencia de los fenómenos



*Incluye helada y granizada

Referente a su distribución geográfica, poco más del 50% de los fenómenos del subconjunto tienen lugar en Veracruz y Oaxaca, le sigue Chihuahua y Chiapas con el 15.6% y Durango, Guerrero y Yucatán con el 15% de ocurrencia.

Gráfico 25: Porcentaje de ocurrencia de los fenómenos por entidad federativa



De acuerdo con sus características geográficas, en cada entidad federativa se presentan particularidades en cuanto al tipo de fenómeno. Por ejemplo, en Oaxaca y Puebla se observa que el porcentaje de ocurrencia para ciclones tropicales, sequías y lluvias se presenta de manera equilibrada; en Yucatán el porcentaje de ocurrencia se concentra mayoritariamente en ciclones tropicales (78%). En Durango, Guerrero, Oaxaca y Veracruz, las lluvias tienen mayor porcentaje de ocurrencia, en Chihuahua y Sonora las nevadas, en Puebla y Zacatecas las sequías y en Chiapas y Yucatán los ciclones tropicales.

Gráfico 26: frecuencia de ocurrencia de acuerdo al tipo de fenómeno por entidad

Nota. Nevada incluye helada y granizada

3.3.4. Análisis de tablas previo al ACS.

El ACS consideró dos variables; tipo de fenómeno y entidad federativa de ocurrencia. Las categorías de la primera variable son: bajas temperaturas, ciclón tropical, inundaciones, lluvias, nevada (incluye helada y granizada), sequía y sismo. Para la segunda son: Chiapas, Chihuahua, Durango, Guerrero, Oaxaca, Puebla, Sonora, Veracruz, Yucatán y Zacatecas. Con el fin de contar con elementos para desarrollar un diagnostico en la primera etapa del ciclo de prevención, el análisis de los datos tiene por objetivo identificar las asociaciones entre los fenómenos y su lugar de ocurrencia.

La Tabla 25, está conformada por i =10 renglones y j = 7 columnas, para nuestra matriz, una representación apropiada asociada con el total de varianza explicada, tendría que hacerse en un espacio de seis dimensiones ($mín\{I, J\} - 1 = min\{10,7\} - 1 = 6$).

Tabla 25: Frecuencia de fenómenos por tipo y entidad federativa de ocurrencias

	Total de eventos registrados (Fenómenos)							
Entidad **Federativa**	Lluvias	Ciclón Tropical	Inundaciones	Sequía	Nevadas	Bajas Temperaturas	Sismo	**Total General**
Veracruz	2147	1390	335	0	0	0	0	3872
Oaxaca	1379	1009	0	1117	0	0	0	3505
Chihuahua	0	0	0	293	684	149	0	1126
Puebla	330	301	0	369	0	0	0	1000
Chiapas	500	550	0	0	0	0	76	1126
Zacatecas	0	0	0	371	347	128	0	846
Durango	360	0	0	151	231	0	0	742
Sonora	131	289	0	0	381	0	0	801
Guerrero	443	107	0	0	0	0	177	727
Yucatán	18	539	0	159	0	0	0	716
Total General	5308	4185	335	2460	1643	277	253	14461

*Incluye heladas o granizada

El juego de hipótesis a probar establece que:

$H_0 =$ El tipo de fenómeno no está relacionado con la entidad federativa de ocurrencia.

$H_1 =$ El tipo de fenómeno está relacionado con la entidad federativa de ocurrencia.

El valor de la prueba Chi- cuadrada de Pearson es de X^2 =16,090.456 con 54 grados de libertad. El valor teórico de la Chi- cuadrada con un $\propto = 0.05$ se establece en 73.3115. Existe evidencia suficiente para afirmar que el tipo de fenómeno si está relacionado con la entidad federativa de ocurrencia (Tabla 26). Adicional a ello, los valores del coeficiente de contingencia C y la V de Cramer son de .726 y .431 los cuales indican un grado de asociación representativo entre las variables. El 55.6% de la varianza explicada se concentra en la primera dimensión, el 18.5% en la segunda, el 13.8% en la tercera, el 7.7% en la cuarta, el 4.0% en la quinta y el 0.4% en la sexta.

Tabla 26: Tabla resumen, fenómenos

Dimensión	Valor singular	Inercia	Chi cuadrado	Sig.	Proporción de inercia		Valor singular de confianza	
					Contabilizado para	Acumulado	Desviación estándar	Correlación 2
1	.787	.619			.556	.556	.004	.043
2	.454	.206			.185	.742	.009	
3	.392	.154			.138	.880		
4	.293	.086			.077	.957		
5	.210	.044			.040	.996		
6	.063	.004			.004	1.000		
Total		1.113	16090.456	.000[a]	1.000	1.000		

a 54 grados de libertad

3.3.5. Resultados del ACS.

En el gráfico de puntos fila (Gráfico 27), se observa un patrón de distribución que, conforme a su cercanía, agrupa a las siguientes entidades: Zacatecas y Chihuahua; Durango y Sonora; Oaxaca, Yucatán y Puebla; Chiapas y Veracruz. Chihuahua y Zacatecas aportan mayor inercia a la dimensión 1 y Guerrero a la dimensión 2.

En el gráfico de puntos columna se observar la asociación entre nevadas, heladas o granizadas y bajas temperaturas; inundaciones, lluvias y ciclón tropical; la sequía se ubica entre ambos grupos y los sismos, se alejan considerablemente del conglomerado de puntos. Como es de suponer, este último aporta mayor inercia a la dimensión 2 y nevadas, heladas o granizadas y bajas temperaturas a la dimensión 1.

De acuerdo con los umbrales de temperatura o precipitación, se puede observar la asociación entre fenómenos que comparten características. De todos ellos, los sismos son el único tipo que corresponde a los fenómenos geológicos, razón por la cual se aleja de la nube de puntos.

Gráfico 27: Puntos fila para entidad, fenómenos

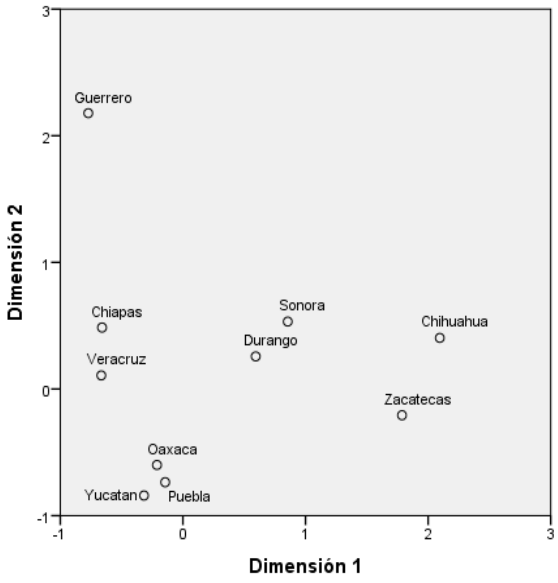
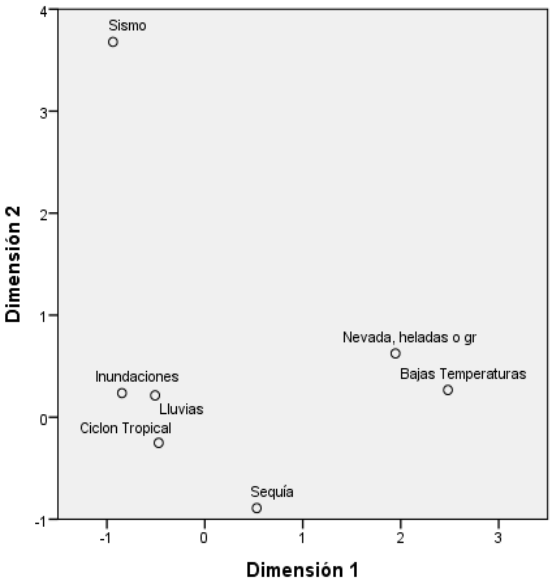


Gráfico 28: Puntos columnas, fenómenos



En el mapa perceptual (Gráfico 29), se observa que la asociación más clara se establece entre sismos y el estado de Guerrero, las lluvias e inundaciones con los estados de Chiapas y Veracruz, las entidades de Oaxaca, Yucatán y Puebla se encuentran entre los ciclones tropicales y las sequias, Zacatecas se asocia con bajas temperaturas y Chihuahua con nevadas, heladas o granizadas y bajas temperaturas.

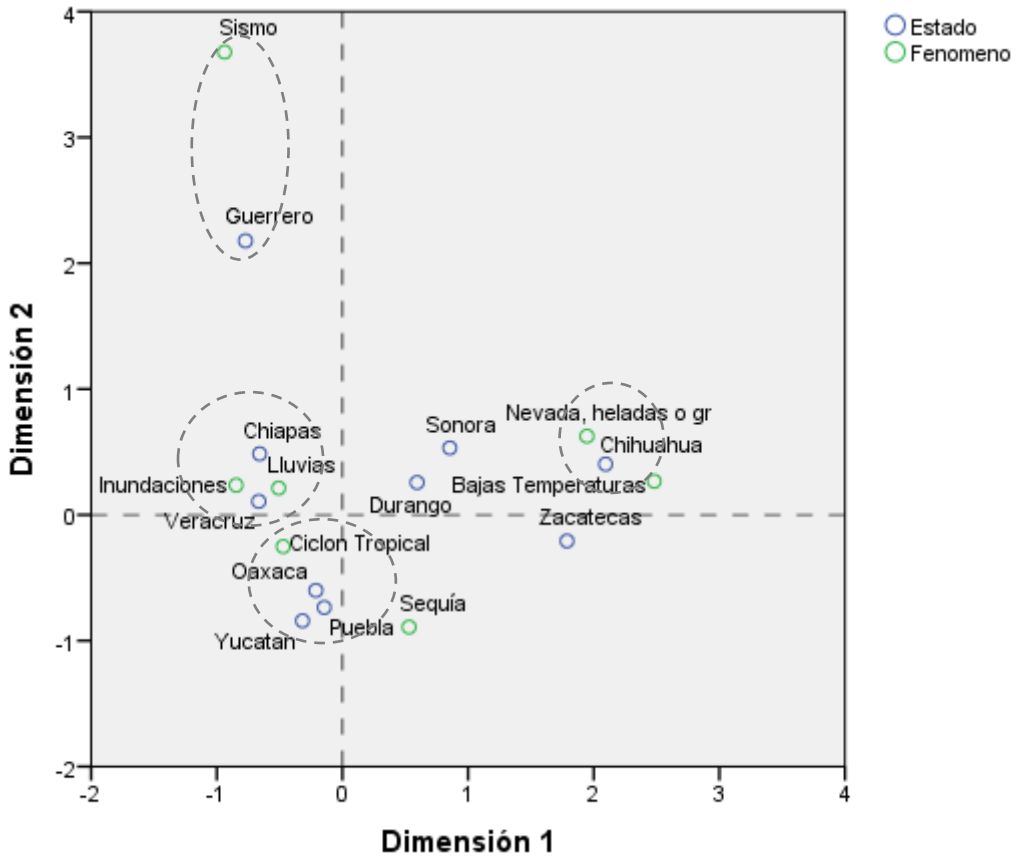
Figura 23. Identificación de las principales asociaciones

Elaboración propia con base en los resultados obtenidos

De acuerdo con información del Centro Nacional de Prevención de Desastres (CENAPRED, 2017b) entre el 2010 y 2015 de acuerdo con el fenómeno y el lugar de ocurrencia, se registran las siguientes afectaciones:

- Sismo. En guerrero se registraron 4 defunciones, 153, 916 personas, 27,614 viviendas, 282 escuelas y 40 hospitales afectados; los daños se calcularon en aproximadamente 1, 569 millones de pesos
- Lluvias e inundaciones. En Chiapas y Veracruz se registraron 64 defunciones, 698,068 personas, 45,526 viviendas, 857 escuelas, 1 hospital y 25,543 hectáreas de cultivo afectadas; los daños se calcularon en aproximadamente 20,214 millones de pesos
- Sequía. En Puebla se registraron 6,951 personas y 5,748 hectáreas de cultivo afectadas; los daños calculados fueron de 32 millones de pesos.
- Bajas temperaturas y nevadas. En Chihuahua se registraron 113 defunciones, 338,379 personas, 2,339 viviendas y 1,454 hectáreas de cultivo afectadas; los daños se calcularon en 1,370 millones de pesos.

Gráfico 29: Mapa perceptual, fenómenos



Las afectaciones por la ocurrencia de un fenómeno serán mayores mientras más vulnerable sea un sistema, cualquier mecanismo que tenga por objetivo reducir el riesgo de un desastre, tendría que implementarse a partir de un conocimiento profundo acerca del fenómeno y las características de los lugares en donde se desarrolla. Tan solo en la primera etapa del ciclo de la prevención, el buen desarrollo de un diagnóstico, permitiría, a partir de la propuesta de acciones concretas, reducir las afectaciones económicas, sociales y humanas que un fenómeno pudiera ocasionar.

CONCLUSIONES

La minería de datos es una herramienta que permite abordar distintas problemáticas en las ciencias sociales. Dentro de la geografía, la minería de datos se perfila como un mecanismo que, viable a los cambios globales asociados a la generación masiva de información, permite abordar desde diferentes ópticas la realidad espacial. Las distintas fuentes de información de las que se nutre la minería de datos permiten abordar de manera creativa e innovadora una problemática comprendida desde enfoques tradicionales o fuentes oficiales. Sin embargo, lo anterior implica cierta rigurosidad en el método, ya que los datos obtenidos de registros no oficiales contienen, en menor o gran medida, un sesgo de recolección. Si nos detenemos por un momento a mirar la cantidad de información que se genera diariamente, a partir de ello, seguramente surgiría más de un planteamiento sobre como leer o analizar una problemática.

El vínculo más obvio entre la minería de datos y la geografía se halla en la geografía automatizada, ambas tienen por objetivo la extracción de información de un conjunto de datos apoyándose de la automatización de los procedimientos. En cierta medida, la minería de datos, exigiría de la geografía automatizada la incorporación de mecanismos que permitan en sus métodos, el tratamiento de grandes cantidades de datos. Sin embargo, analizar estas grandes bases de datos, requiere fundamentalmente de un sólido conocimiento teórico para interpretar la problemática por lo que es importante anotar que la automatización de los procedimientos, no incluye el razonamiento humano de la información.

Como se expuso en el primer capítulo, existe un abanico amplio de métodos en minería de datos, elegir el método adecuado es un paso crucial para obtener resultados satisfactorios. El analista de información requiere tener conocimiento sobre el conjunto de métodos, parámetros, tipos de variable, objetivos e interpretación de resultados. El conocimiento de lo anterior da la posibilidad de elegir el método más adecuado y realizar una interpretación correcta de los resultados. Por ejemplo, el análisis de correspondencias simple, método usado en el presente trabajo para ejemplificar la aplicación de la minería de datos en un conjunto de datos, pertenece a los métodos descriptivos, por lo que plantear una predicción a partir de ellos implicaría resultados imprecisos, así también, el método se establece para variables discretas medidas en escala absoluta o de conteo y relacionadas entre sí.

Los ejemplos seleccionados permitieron mostrar la posible aplicación de la minería de datos en tres áreas distintas del conocimiento en geografía. El primer ejercicio correspondiente al asesinato de periodistas, permitió afinar la afirmación que equipara el asesinato de periodistas en México con países en guerra como Siria concluyendo que, la comparación es válida en términos cuantitativos, sin

embargo, en términos cualitativos, en México el asesinato de periodistas está asociada a una tipología de ejecuciones relacionada estrechamente con el crimen organizado. En el segundo ejemplo, el análisis de los datos permitió observar que, tal como indica la teoría, existe un patrón migratorio que se define por las características estructurales del lugar de origen y las características económicas de los lugares de destino: con esa base se pudo identificar en el gráfico la tradición migratoria de los jornaleros agrícolas en el periodo de referencia de los datos. La antigüedad de los datos oficiales de la migración jornalera en México, pone de manifiesto la necesidad de utilizar como parámetro otras fuentes de información. Por ejemplo, se podría analizar este fenómeno a partir de la información que genera el uso de un aparato móvil, las redes sociales, las remesas o la audiencia radiofónica de la población migrante en relación con su lugar de origen y los centros de trabajo agrícola. Finalmente, en el tercer ejemplo, el método permitió observar el patrón de ocurrencia entre la presencia de un conjunto de fenómenos y el lugar de desarrollo como diagnóstico para la prevención de desastres naturales.

El método seleccionado permitió demostrar la aplicación de la minería de datos en tres áreas distintas de la geografía, a reserva de que las cantidades de datos no fueron sustantivas o las fuentes de información fueron oficiales, se concluye que esta opción metodológica propia de la estadística y noble para abordarse desde la geografía cuantitativa y las geotecnologías de la información, es una herramienta que permite extraer información valiosa e identificar patrones en geografía. Así mismo, se destaca que la información y el conocimiento extraído de un conjunto de datos tienen la bondad de poder contribuir a la resolución de una problemática o al mejoramiento de una situación indeseable.

BIBLIOGRAFÍA

- Aldás, J. (2000). *El análisis de correspondencias simples.* España: Universidad de Valencia. Dpto. de Dirección de Empresas "Juan José Renau Piqueras".
- Arcila, C., Barbosa, E., & Cabezuelo, F. (2016). *Técnicas big data: análisis de textos a gran escala para la investigación científica y periodística.* El profesional de la información, 25(4), 623-631.
- Arribas, A. (2016) "*Ser o no ser periodista en México".* En Infoamérica. Revista Iberoamericana de Comunicación. No. 10. España. pp. 39-49
- Ballesteros, V., Iñiguez, G., & Velasco, M. (2018). *Minería de Datos*. RECIMUNDO: Revista Científica de la Investigación y el Conocimiento, 2(1), 339-349.
- Baxendale, C. (2015). *Ordenar el territorio con base en la Geografía Cuantitativa*. En *Teoría y métodos de la geografía cuantitativa: Libro 1, Por una geografía de lo real* (39-52). Buenos Aires: MCA Libros.
- Beltrán, B. (2003). *Minería de datos*. Tesis de maestría. Benemérita Universidad Autónoma de Puebla.
- Beltrán, G. (2015). *La geolocalización social.* Polígonos. Revista de geografía, (27), 97-118.
- Bosque, J. (2015). *Neogeografía, Big Data y TIG: problemas y nuevas posibilidades*. Polígonos. Revista de Geografía, (27), 165-173.
- Buzai, G & García de León, A. (2015). *Balance y actualidad de la geografía cuantitativa*. En *Geografía, geotecnología y análisis espacial: tendencias, métodos y aplicación* (31-54). Santiago de Chile: Triángulo.
- Buzai, G. (2001). *Paradigma Geotecnológico, Geografía Global y CiberGeografía, la gran explosión de un universo digital en expansión*. GeoFocus. Revista Internacional de Ciencia y Tecnología de la Información Geográfica, (1), 24-48.
- Buzai, G. (2005). *Geografía global. Paradigma geográfico para el análisis socio espacial interdisciplinario*. Junio 30, 2019, de Researche Gate Sitio web: https://www.researchgate.net/publication/299398136_Geografia_Global_Paradigma_geografico_para_el_analisis_socioespacial_interdisciplinario.
- Buzai, G. *et. al.* (2015). *Teoría y métodos de la geografía cuantitativa: Libro 1, Por una geografía de lo real*. Buenos Aires: MCA Libros.
- Buzai, G. (2015). Evolución del pensamiento geográfico hacia la Geografía Global y la Neogeografía. En *Geografía, geotecnología y análisis espacial: tendencias, métodos y aplicación* (4-16). Santiago de Chile: Triángulo.

- Buzai, G., & Ruiz, E. (2012). *Geotecnósfera. Tecnologías de la información geográfica en el contexto global del sistema mundo*. Anekumene, (4), 88-106.
- Capula, E. (2004). *Imagen y posicionamiento: Aplicación del análisis de correspondencia a la investigación de mercados*. Tesis de licenciatura. Universidad Nacional Autónoma de México.
- Castillo, J. (2008). *Aplicaciones de escalamiento multidimensional y el análisis de correspondencias a los seguros de salud y pensiones*. Tesis de licenciatura. Universidad Nacional Autónoma de México.
- CENAPRED. (2014). *Manual de protección civil.* México: Sistema Nacional de Protección Civil.
- Chiapa, A. (2015). *Aplicación de análisis de correspondencias a empresas innovadoras de México.* Tesina para optar por el grado de especialista en estadística aplicada. Universidad Nacional Autónoma de México.
- De la Fuente, S. (2011). *Análisis de correspondencias simples y múltiples*. Facultad de Ciencias Económicas y Empresariales. Madrid-España.
- Dueñas, M. (2009). *Minería de datos espaciales en búsqueda de la verdadera información*. Ingeniería y universidad, 13(1), 137-156.
- Fuenzalizada, M., Buzai, G. D., Moreno, A. & García de León, A. (2015). *Geografía, geotecnología y análisis espacial: tendencias, métodos y aplicación*. Santiago de Chile: Triángulo.
- Guevara, E., Quaas, R. & Fernández, G. (2006). *Guía Básica para la elaboración de atlas estatales y municipales de peligros y riesgos*. México: CENAPRED- SEGOB.
- Gutiérrez, J. (2018). *Big Data y nuevas geografías: la huella digital de las actividades humanas*. Documents d'anàlisi geogràfica, 64(2), 195-217.
- Gutiérrez, J., García, C., & Salas, H. (2016). *Big (Geo) Data en Ciencias Sociales: Retos y Oportunidades*. Revista de estudios andaluces, 33(1), 1-23.
- Hernández, J., Ramírez, M.J. & Ferri, C. (2004). *Introducción a la minería de datos*. Madrid: Pearson educación.
- Infobae (2018). *México es igual de peligroso que Afganistán para ejercer el periodismo*. Recuperado el 12 de agosto de 2019, de https://www.infobae.com/america/mexico/2018/12/17/mexico-es-igual-de-peligroso-que-afganistan-para-ejercer-el-periodismo/
- Jaramillo, F. (2016). *Introducción a la minería de datos en la plataforma KNIME.* Tesis de licenciatura. Universidad Nacional Autónoma de México.

- Juanes, P. (2014). *La Geografía y la Estadística, dos necesidades para entender BigData.* Tesis de maestría. Universidad de Salamanca.
- López, P & Fachelli, S. (2015). *Metodología de la investigación social cuantitativa. Parte III. Análisis*. España: Universidad Autónoma de Barcelona.
- Maskrey, A (compilador), (1993). *Los desastres no son naturales*. Bogotá: LA RED-Tercer Mundo Editores.
- Molina, I. (2007). *Análisis de correspondencia simple*. Universidad Carlos III de Madrid. Apuntes recuperados el 12 de agosto de 2019, de http://halweb.uc3m.es/esp/Personal/personas/imolina/MiDocencia/TecnicasInvestigacion/TecnicasDeInvestigacion.htm
- Molina, L. (2002). *Data Mining: Torturando a los datos hasta que confiesen*. Barcelona España: Universidad politécnica de Catalunya FUOC.
- Montes, E. (2015). *Estructura diacrónica de los procesos de investigación aplicada a la geografía cuantitativa*. En *Teoría y métodos de la geografía cuantitativa: Libro 1, Por una geografía de lo real* (53-72). Buenos Aíres: MCA Libros.
- Morales, J. (2004). *Aplicación e interpretación de técnicas de reducción de datos según escalamiento óptimo. (Análisis de correspondencia múltiple y análisis de componentes principales categórico).* Tesis de licenciatura. Universidad de Chile. Facultad de Ciencias Sociales. Departamento de Sociología.
- Moreno, A. (2015). *Singularidades gnoseológicas de la praxis geotecnológica en la ciencia geográfica*. En *Geografía, geotecnología y análisis espacial: tendencias, métodos y aplicación* (17-30). Santiago de Chile: Triángulo.
- Olvera, J, *et.al.* (2014). *Infraestructura de datos espaciales y normatividad geográfica en México: una perspectiva actual*. Instituto de Geografía. México.
- Oropeza, M., & Díaz, N. (2007). *La geotecnología y su inserción en el pensamiento geográfico*. Terra Nueva Etapa, 23(34), 71-95.
- Pérez, C & Satín, D. (2007). *Minería de datos. Técnicas y Herramientas*. Madrid, España: Paraninfo.
- Pineda, I. (2003). *Análisis de correspondencias para el estudio de asociación entre variables categóricas: un enfoque aplicado*. Tesis de licenciatura. Universidad Nacional Autónoma de México.
- Proceso (2019a). *México, el país más peligroso de América para ejercer el periodismo: RSF*. Recuperado el 12 de agosto de 2019 de https://www.proceso.com.mx/580351/mexico-el-pais-mas-peligroso-de-america-para-ejercer-el-periodismo-rsf

- Proceso. (2019b). *Los jornaleros agrícolas, "invisibles" para el gobierno*. Recuperado el 12 de agosto de 2019 de https://www.proceso.com.mx/591958/los-jornaleros-agricolas-invisibles-para-el-gobierno.
- Proceso. (2018). *México, Afganistán y Siria, los países más letales para los periodistas en 2018: PEC*. Recuperado el 12 de agosto de 2019 de https://www.proceso.com.mx/564189/mexico-afganistan-y-siria-los-paises-mas-letales-para-los-periodistas-en-2018-pec
- Publimetro. (2017). *80% de los derrumbes de casas por corrupción y no por sismos*. Recuperado el 12 de agosto de 2019 de https://www.publimetro.com.mx/mx/nacional/2017/09/17/80-los-derrumbes-casas-corrupcion-no-sismos.html
- Ramírez, D. (2008). *La libertad de expresión en México amenazada por las agresiones a periodistas y la concentración de medios*. El Cotidiano, 23 (150), 47-52.
- Ramírez, L. & Claret, R. (2015). *Modelos multicriterio basados en funciones de utilidad*. En *Teoría y métodos de la geografía cuantitativa: Libro 1, Por una geografía de lo real* (105-122). Buenos Aíres: MCA Libros.
- Ríos, V. (2012). *El asesinato de periodistas y alcaldes en México y su relación con el crimen organizado*. En J. A. Aguilar (Coord.), Las bases sociales del crimen organizado y la violencia en México (275-308). México, D. F.: CIES
- Riquelme, J. C., Ruiz, R., & Gilbert, K. (2006). *Minería de datos: Conceptos y tendencias.* Inteligencia Artificial: Revista Iberoamericana de Inteligencia Artificial, 10 (29), 11-18.
- Rodríguez, M., & Mora, R. (2001). *Estadística informática: casos y ejemplos con el SPSS*. Alicante: Universidad de Alicante.
- Rojas, T. (2017). *Migración rural jornalera en México: La circularidad de la pobreza*. Iberoforum. Revista de Ciencias Sociales de la Universidad Iberoamericana, XII (23), 1-35.
- Rojas, T. (2009). *La crisis del sector rural y el coste migratorio en México.* Iberoforum. Revista de Ciencias Sociales de la Universidad Iberoamericana, IV (8), 40-81.
- Rubio, B. (2002). *La exclusión de los campesinos y las nuevas corrientes teóricas de interpretación*. Nueva sociedad, 182, 21-33.
- Saavedra, R. (2012). *El análisis de correspondencias conjunto y múltiple ajustado.* Tesis de maestría. Pontifica Universidad Católica de Perú.
- Saldaña, J.P. (2005). *Análisis de la percepción de inseguridad por medio del modelo de correspondencias.* Tesis de licenciatura. Universidad Nacional Autónoma de México.

- Salinas, S. (2012). *Jornaleros agrícolas: invisibilización deliberada*. La Jornada del Campo, No.54. Recuperado el 12 de agosto de 2015, de http://www.jornada.unam.mx/2012/03/17/cam-jornaleros.html
- SEDESOL. (2010). *Diagnóstico del Programa de Atención a Jornaleros Agrícolas*. México: SEDESOL
- Tlachinollan (2015). *La montaña de Guerrero. Tierra de mujeres migrantes*. México: Centro de Derechos Humanos de la Montaña.
- Tulla, A. (1993). *Métodos y técnicas cuantitativas. Valoración y aplicaciones en Geografía Rural.* En V Coloquio de Geografía Cuantitativa. Universidad de Zaragoza.
- Velasco, L. (2014). *De jornaleros a colonos: residencia, trabajo e identidad en el Valle de San Quintín*, Tijuana, El Colegio de la Frontera Norte: México.
- Zaldivar, A., *et al*. (2011). C*omparativa entre los Métodos de Regresión Lineal y Minería de Datos para la Identificación de Variables Asociadas al Éxito Académico en Estudiantes de Educación Superior*. En XV Congreso Nacional y I Internacional de Modelos de Investigación Educativa. Universidad Nacional de Educación a Distancia.

Bases de datos

- Comité para la Protección de los Periodistas. (2019). *Asesinatos desde 1992* [base de datos]. Recuperado de https://cpj.org/data/killed/?status=Killed&motiveConfirmed%5B%5D=Confirmed&type%5B%5D=Journalist&start_year=1992&end_year=2019&group_by=year
- Centro Nacional de Prevención de Desastres (CENAPRED). (2017a). *Declaratorias sobre emergencia, desastre y contingencia climatológica* [base de datos]. Recuperado de https://datos.gob.mx/busca/dataset/declaratorias-sobre-emergencia-desastre-y-contingencia-climatologica
- Centro Nacional de Prevención de Desastres (CENAPRED). (2017b). *Impacto socioeconómico de desastres de 2000 a 2015* [base de datos]. Recuperado de https://datos.gob.mx/busca/dataset/impacto-socioeconomico-de-desastres-de-2000-a-2015
- Secretaría de Desarrollo Social & Universidad Autónoma Chapingo. (2009). E*ncuesta Nacional de Jornaleros 2009* [base de datos]. Recuperado de http://www.cipet.gob.mx/jornaleros/

Video

- Godina, M. (2017, mayo 31). Análisis de correspondencias. Recuperado de https://www.youtube.com/watch?v=A5gtHx2fHJE&t=17s